Mémoire présenté devant le jury de l'EURIA et de l'INSA Rennes en vue de l'obtention du diplôme d'actuaire EURIA, du diplôme d'ingénieur de l'INSA Rennes, et de l'admission à l'Institut des Actuaires

le 24 Septembre 2021

Par : Robin MIRALLES
Titre : Analyse des grandeurs explicatives des arbitrages des contrats d'assurance vie en mode de gestion libre par méthodes d'apprentissage statistique

Confidentialité : Non

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

Membres présents du jury de l'Institut des Actuaires :
LE BERRE Anaëlle
BIESSY Guillaume
Signatures :

Entreprise :
ALLIANZ IARD
Signature :

Membres présents du jury de l'EURIA ou de l'INSA Rennes :
VERMET Franck
Signature :

Directeur de mémoire en entreprise :
METGE Guillaume
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

Les contrats d'épargne-retraite en assurance vie permettent aux assurés de se constituer une épargne à partir d'un capital que l'assureur investit sur les marchés financiers. Les assurés ayant souscrit un contrat multi-supports en gestion libre disposent du droit et de l'initiative, à tout moment, de transférer leurs capitaux entre les différents supports du contrat. On parle alors d'arbitrages. Ces arbitrages peuvent être réalisés soit depuis les supports dits euros, soit depuis les supports dits en unités de compte (UC). Les supports euros sont composés majoritairement d'obligations, mais aussi d'immobilier, d'actions et de monétaire : ils sont peu risqués, présentent un rendement faible, en décroissance ces dernières années, et font porter le risque par l'assureur. Les supports en UC peuvent être des parts d'actions, de sociétés, ou de fonds de placements, qui présentent généralement de bien meilleurs rendements pour un risque plus élevé et qui font porter le risque sur l'assuré.

Le contexte réglementaire dans lequel se situent les assureurs vis à vis du risque conduit alors à anticiper, et donc modéliser, les arbitrages afin d'avoir une bonne estimation du niveau de risque auquel les assureurs sont soumis. En effet, les placements sur les fonds euros garantissent le capital investi à l'assuré, et donc augmentent les besoins en fonds propres de l'assureur, tandis que les placements sur les supports en UC ne sont garantis qu'en nombre de parts indépendamment de leur valeur liquidative.

Le contexte actuel des taux bas, ainsi que l'accès à l'information en temps réel via internet, viennent perturber et complexifier les dynamiques d'arbitrages. Des recherches sur le sujet ont été menées et il semblerait que les dynamiques d'arbitrages soient soumises à des phénomènes de contagion et sensibles à la conjecture économique. Partant de ces constats, une modélisation par modèle de machine learning est proposée afin de caractériser les mouvements d'arbitrages à la maille produit. Les données utilisées sont issues d'un croisement de bases de données internes portant sur les contrats d'assurance vie individuelle multi-supports en mode de gestion libre des produits phares d'Allianz, ainsi que d'autres données externes à Allianz relatives à la conjecture économique. L'information sur les UC sera utilisé.

Mots clefs: Assurance vie, gestion libre, arbitrages, analyse de donnée

Abstract

Life insurance retirement savings contracts allow policyholders to build up savings from a capital sum that the insurer invests in the financial markets. Policyholders who have subscribed to a multi-support contract with free management mode have the right and the initiative, at any time, to transfer their capital between the different supports of the contract. This is called arbitration. These arbitrations can be carried out either from the so-called euro supports, or from the so-called unit-linked supports (UL). The euro supports are mainly composed of bonds, but also of real estate, shares and money market : they are not very risky, have a low yield which has been decreasing in the last few years, and make the insurer bear the risk. Unit-linked products can be shares, companies, or investment funds, which generally offer much better returns for a higher risk and shift the risk to the insured.

The regulatory context in which insurers find themselves with regard to risk leads to anticipating, and therefore modeling, arbitrages in order to have a good estimate of the level of risk to which insurers are subject. Indeed, investments in euro funds guarantee the capital invested to the insured, and therefore increase the insurer's equity requirements, whereas investments in unit-linked products are only guaranteed in terms of the number of units, regardless of their net asset value.

The current context of low interest rates, as well as access to real-time information via the Internet, disrupt and complicate the dynamics of arbitration. Research on the subject has been conducted and it seems that arbitrage dynamics are subject to contagion phenomena and sensitive to economic conditions. Based on these findings, a machine learning model is proposed to characterize arbitrage movements at the product level. The data used are derived from a cross-referencing of internal databases on Allianz's flagship individual multi-support life insurance contracts, as well as other data external to Allianz relating to the economic outlook. Information on the ULs will be used.

Keywords: Life insurance, self asset management, switches, data analysis

Remerciements

Je remercie Guillaume METGE, manager du Reporting MVBS/NBV, de m'avoir accepté dans son équipe, ainsi que Yannis AMAMOU, mon encadrant d'alternance, pour son professionnalisme. Leur encadrement a été remarquable malgré la situation sanitaire. Je remercie également Nicolas BOURE, directeur de l'actuariat vie d'Allianz France de m'avoir donné l'opportunité de travailler au sein de ses équipes.

Je tiens à remercier toute l'équipe du Reporting MVBS/NBV de Allianz France avec laquelle j'ai passé ma première année d'expérience professionnelle dans les meilleures conditions. Je me souviendrai de leur bienveillance, mais aussi de leur bonne humeur ainsi que de nos conversations.

Je souhaite exprimer mes remerciements à Marine HABART, directrice groupe de l'Actuariat Vie d'AXA mais aussi mon tuteur de mémoire EURIA, pour ses conseils avisés et sa prise de recul sur mon mémoire qui n'ont pas manqués de me challenger.

Je souhaite témoigner ma plus grande reconnaissance à Dominique ABGRALL et le remercier pour son altruisme : il m'a beaucoup donné et a définitivement eu un impact positif et non négligeable sur mon parcours.

Je souhaite également remercier toute l'équipe enseignante de l'EURIA et de l'INSA de Rennes pour l'enseignement de qualité que j'ai eu la chance de recevoir durant ces 6 années, avec une pensée toute particulière pour Franck VERMET, directeur de l'EURIA, mais aussi mon tuteur EURIA d'alternance.

Enfin, merci à ma famille ainsi qu'à tous mes amis qui ont tenu leur rôle de pilier dans ma vie durant cette période enrichissante et dense.

Note de synthèse

Le contrat d'assurance vie multi-supports est un contrat bien défini par la législation française, disposant d'une fiscalité propre. Il s'agit du produit d'épargne préféré des français. Celui-ci présente des mécanismes et options qui lui sont propres : l'option d'arbitrage et de rachat, le mécanisme de participation aux bénéfices, un éventuel TMG, un mode de gestion, une périodicité de paiement de primes etc...

Son principe est simple : l'assuré répartit les sommes investies sur un panel de fonds UC et euros que lui propose l'assureur (souvent fonction du produit d'assurance vie). Ces fonds sont composés de différents actifs sous-jacents. Les fonds euros ont une forte composition en titres de créances et d'obligations, tandis que les fonds UC présentent une forte composition en valeurs mobilières et actifs financiers. Les fonds UC présentent un rendement et un risque plus élevé que les fonds euros. Ils sont exprimés en parts.

L'assureur se doit de garantir le capital investi sur les fonds euros, tandis qu'il n'est tenu de garantir que la part d'UC sur les fonds UC. Dès lors, le risque associé aux capitaux sur les fonds euros est principalement supporté par l'assureur tandis que le risque associé aux capitaux sur les fonds UC est supporté par l'assuré.

Si l'assuré possède un contrat en mode de gestion libre, il peut décider à tout moment de transférer tout ou une partie des capitaux d'un fond à un autre : il exerce son option d'arbitrage. Il y a donc un transfert de risque, avantageux ou non pour l'assureur. En mode de gestion libre, l'assureur n'a pas le contrôle sur ce transfert de risque, qui peut être ponctuellement soudain, massif, et désavantageux pour l'assureur en terme d'engagement auprès de ses assurés. Les recherches menées sur le sujet montrent que ces phénomènes sont soumis à des variables endogènes et exogènes au contrat d'assurance vie.

Ces phénomènes incitent donc les assureurs à modéliser les arbitrages.

En vue d'étudier les arbitrages dans les contrats d'assurance vie en mode de gestion libre, nous construisons une base de données dont une ligne donne toutes les informations (provision mathématique et éventuels mouvements d'arbitrages entre autres) d'un contrat pour un mois, entre janvier 2015 et mai 2021. Par la suite, dans le chapitre 4 nous décrivons comment nous prenons en compte l'information sur les fonds à la maille contrat.

De cette manière nous sommes en mesure par la suite de constituer une base de données à la maille $produit \times mois$ nous indiquant les taux d'arbitrages mensuels, tout en disposant des informations à la maille contrat.

Pour ce faire de nombreuses bases de données sont requises (bases de provisions mathématiques, de mouvements d'arbitrages, d'informations sur les contrats, des valeurs liquidatives des fonds UC et euros etc...) ainsi que de nombreux retraitements afin de capturer l'information propre à chaque contrat. À ce titre un effort particulier sur la méthode de calcul des taux d'arbitrages, mais aussi sur la qualité de la donnée, a été fourni à chaque étape de la construction de la base de données.

Nous décidons également d'ajouter des variables exogènes classiques comme l'évolution du TME et du CAC40, la volatilité du CAC40, mais aussi moins classiques comme les Google trends associés à certains mots clefs, afin de tenir compte de l'impact de la conjecture économique sur la dynamique des arbitrages.

De cette manière nous dressons un profil temporel, au sens des variables explicatives construites, pour chaque produit en tenant compte de l'information à la maille contrat. Ces profils temporels de produit seront injectés dans les algorithmes d'apprentissage statistique supervisé.

Les nombreux algorithmes de ML utilisés tout au long du mémoire sont les suivants :

A. *Apprentissage statistique supervisé* :

   (a) GLM : modèle de régression généralisé
   (b) SVM : Support Vector Machine
   (c) RF : Random Forest (mode supervisé)
   (d) XGboost : eXtreme Gradient boosting

B. *Apprentissage statistique non supervisé* :

   (a) ACP : Analyse en Composantes Principales
   (b) K-means : K-moyennes
   (c) CAH : Classification Ascendante Hiérarchique
   (d) RF : Random Forest (mode non-supervisé)
   (e) MMG : Modèle de Mélange Gaussien

Après la constitution des données, dans un premier temps plusieurs triplets méthode/distance/algorithme ont été testés en vue de former des groupes homogènes de fonds UC et euros. En effet, la forte granularité des fonds nécessite de les regrouper afin de réaliser des analyses pertinentes par la suite. Ces clustering s'attachent à regrouper les fonds présentant les mêmes caractéristiques statistiques , financières, ou bien la même dynamique d'évolution sur leur historique. L'objectif est de cartographier les supports des contrats d'assurance vie en vue de les prendre en compte dans notre modélisation.

Les groupes de fonds UC et euros qui s'en dégagent sont donc caractérisés différemment et apportent des informations différentes selon un angle d'analyse différent, selon la méthode employée. Nous utiliserons les trois clusterings présentés ci-dessous :

A. *Temporal-proximity based clustering* :

cluster 1 : les "fonds à tendance haussière et insensible à la conjecture économique"

cluster 2 : les "fonds à dynamique euros"

cluster 3 : les "fonds risqués"

cluster 4 : les "fonds dynamiques sensibles à la conjecture économique"

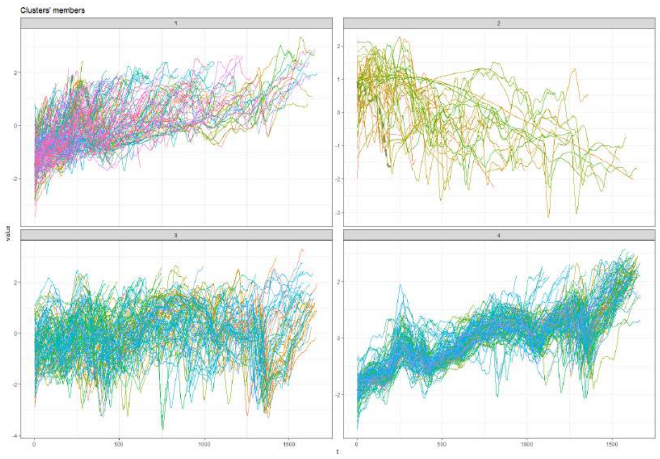


FIGURE 1 – Clustering retenu pour la dynamique d'évolution des fonds euros et UC

B. *Characteristic based clustering* :

i. Indicateurs statistiques :

cluster 1 : les fonds "investissement risque modéré" caractérisés par une légère tendance haussière mais instable.

cluster 2 : les fonds "investissement risqué" uniquement caractérisés par une volatilité élevée.

cluster 3 : les fonds "investissement peu risqué" caractérisés par une stabilité et peu de dynamisme, mais avec une tendance haussière marquée.

ii. Indicateurs financiers :

cluster 1 : les "fonds dynamiques" caractérisés par une volatilité forte, et une élasticité par rapport au CAC40, qui présentent donc une tendance générale haussière

cluster 2 : les "fonds modérés" caractérisés comme le cluster 1, mais avec des valeurs plus modérées.

cluster 3 : les "fonds de sécurisation" caractérisés uniquement par une tendance faible voir baissière.

Pour rappel, nous avons accès à la PM par fonds UC et euros à la maille *contrat* $\times$ *mois*. En "injectant" les différents clusters construits dans la base de données des arbitrages, nous avons donc accès à la proportion de PM correspondant à chaque groupe (selon les différentes méthodes et indicateurs) à la maille *contrat* $\times$ *mois*. Cette information sera injectée dans les modèles de prédictions des taux d'arbitrages en vue de quantifier l'importance ou non de la dynamique des fonds UC et euros sur les taux d'arbitrages en gestion libre.

Dans un second temps des statistiques descriptives et une analyse univariée sont entreprises. Celles-ci confrontent directement qualitativement et quantitativement les taux d'arbitrages aux variables explicatives retenues. Celles-ci mettent clairement en évidence :

1. la structure du portefeuille étudié, de par les caractéristiques des contrats,
2. les structures linéaires reliant les variables explicatives aux taux d'arbitrages,
3. les structures non-linéaires reliant les variables explicatives aux taux d'arbitrages,
4. l'impact asymétrique sur les différentes origines/destinations d'arbitrage (UC_EUR, EUR_UC, et UC_UC) des variables explicatives.
5. des relations déjà établies (experts et mémoires) liant les taux d'arbitrages à des variables explicatives tant endogènes qu'exogènes, mais aussi de nouvelles. De cette manière, des comportements d'arbitrages sont identifiés et quantifiés.

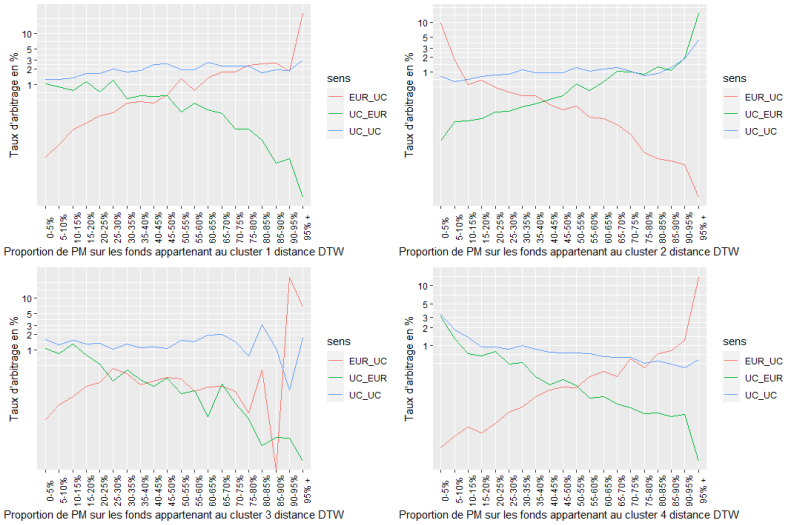


FIGURE 2 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la proportion de PM allouée à chaque cluster formés par leur dynamique d'évolution (distance DTW + K-means)

À partir de ces constats, une découpe des variables explicatives en classes est faite de manière à en tenir compte dans une base de données à la maille *produit* × *mois* qui servira de base aux modèles de machine learning. En effet, les statistiques descriptives et l'analyse univariée ne suffisent pas à modéliser les taux d'arbitrages : il faut prendre en compte tous ces phénomènes simultanément et c'est pourquoi une approche machine learning est adoptée avec les algorithmes d'apprentissage statistiques supervisés cités plus haut.

Deux approches de modélisation des taux par ML sont considérées : une approche globale et une approche spécifique. L'approche globale permet une interprétation directe du modèle via la méthode SHAP mais présente une moins bonne modélisation que l'approche spécifique. L'approche spécifique permet une meilleure modélisation, au détriment d'une certaine opacité d'interprétation du modèle. Ce compromis interprétabilité / performance est vérifié quantitativement et qualitativement.

Origine/destination	Approche	Algorithme	RMSE a	RMSE t	MAE a	MAE t	Critère a	Critère t
UC_EUR	*globale*	*Reg.lin*	2.267	5.682	1.268	2.934	19735	11350
		SVM	1.373	4.576	0.5762	2.409	8845	6989
		RF	1.755	4.701	0.7356	2.079	12733	7948
		XGboost	0.3443	4.338	0.2323	1.886	432	6883
	spécifique	*17 SVM, 6 XGboost, 1 RF, 1 Reg.lin*	1.233	3.669	0.4205	1.451	6698	4779
EUR_UC	*globale*	*Reg.lin*	0.1206	0.1504	0.0515	0.0629	32.88	3.55
		SVM	0.083	0.157	0.019	0.0575	12.32	3.31
		RF	0.1371	0.1351	0.0505	0.0536	50.85	2.97
		XGboost	0.0732	0.1243	0.0383	0.0477	14.65	2.25
	spécifique	*18 SVM, 3 XGboost, 3 RF, 1 Reg.lin*	0.0824	0.1497	0.0251	0.0431	13.651	3.157
UC_UC	*globale*	*Reg.lin*	3.508	2.163	1.65	1.534	x	x
		SVM	2.703	1.798	0.7484	1.226	x	x
		RF	3.213	2.263	1.193	1.454	x	x
		XGboost	2.729	1.219	1.093	0.7678	x	x
	spécifique	*14 XGboost, 7 SVM, 4 RF*	2.243	1.013	0.5733	0.5271	x	x

FIGURE 3 – Tableau récapitulatif des meilleurs modèles obtenus (approche globale et spécifique)

L'analyse critique des deux approches de modélisation des taux, notamment avec la méthode SHAP, conduit à quelques résultats généraux :

1. Les XGboost, et les SVM sont performants sur cette problématique. Il conviendrait de pousser plus loin la paramétrisation de ceux-ci et de les alimenter avec d'autres variables explicatives pertinentes non prises en compte dans ce mémoire, comme par exemple la valeur du taux d'arbitrage le mois précédant. Les taux UC_EUR posent un problème de modélisation.

2. La nature non linéaire des comportements/phénomènes d'arbitrages est vérifiée par le fait que seuls des modèles complexes ("boites noires") fournissent des résultats convenables. La régression linéaire est à oublier si l'on souhaite dans le futur créer un outil de prédictions des taux.

3. La méthode SHAP nous indique que les XGboost arrivent à apprendre certains comportements décelés dans les statistiques descriptives. Cette capacité d'apprentissage est donc très certainement la cause de la performance des XGboost. Les

variables explicatives possédant les pouvoirs explicatifs les plus forts sont les variables des PM de contrats, de nombre de supports, de CSP, de mois de l'année, et d'âge actuariel. De plus, nous montrons que l'ajout des variables de clustering de fonds (indicateurs statistiques et distance DTW majoritairement) présentent un pouvoir explicatif fort. Ces variables de clustering viennent éclipser les variables dites exogènes comme le rendement du CAC40. Elles viennent de plus suggérer que seule une partie (certains fonds) du portefeuille pourrait être sujette à des phénomènes dynamiques d'arbitrages.

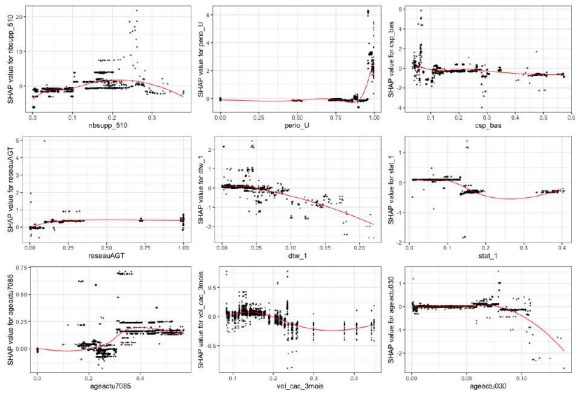


FIGURE 4 – *SHAP plot* de quelques variables explicatives de l'XGboost approche globale présentant une relation avec les taux d'arbitrages UC_EUR

4. Les XGboost semblent détecter assez bien les taux élevés selon certaines variables : ils peuvent alors être entraînés et utilisés afin de créer un outil de monitoring des risques d'arbitrages massifs.

FIGURE 5 – *Break Down profile plot* de l'XGboost approche globale associé aux quatre taux d'arbitrages UC_UC les plus élevés : identification des variables à l'origine d'un taux élevé.

5. Le fait que l'approche spécifique présente une meilleure modélisation que l'approche globale, sachant que l'analyse de l'approche globale nous informe que celle-ci ne tient pas réellement compte des produits, nous indique que les produits présentent chacun une dynamique différente vis à vis des taux d'arbitrages (dû à la structure spécifique des contrats le composant) et que les modèles complexes sont capables d'en cerner les spécificités. Néanmoins, l'approche spécifique serait elle aussi à améliorer afin de constater une réelle distinction notable avec l'approche globale.

De nombreux autres résultats très locaux sont obtenus et sont donc détaillés dans le rapport.

Suite à ces travaux exploratoires sur la qualité de donnée et sur un certain nombre de modèles d'apprentissage, nous pouvons envisager par la suite une modélisation des arbitrages dynamiques sur les sous-portefeuilles pertinents, ou bien établir dans un premier temps des outils de monitoring du risque d'arbitrage. La méthode d'analyse proposée peut être utilisée dans d'autres situations où la donnée est complexe. Malgré la promesse de ces modèles, il y a toujours un fort besoin d'analyse critique de l'actuaire.

Executive summary

The multisupport life insurance contract is a contract well defined by the French legislation, with its own taxation. It is the preferred savings product of French people. It has its own mechanisms and options : the option of arbitration and buyback, the profit-sharing mechanism, a possible MGR, a management mode, a periodicity of payment of premiums etc...

Its principle is simple : the insured distributes the sums invested over a panel of unit linked and euro funds offered by the insurer (often depending on the life insurance product). These funds are composed of different underlying assets. The euro funds have a strong composition in debt securities and bonds, while the UL funds have a strong composition in securities and financial assets. UC funds have a higher return and risk than euro funds. They are expressed in units.

The insurer has to guarantee the capital invested in the euro funds, while he is only obliged to guarantee the share of the UL funds. Therefore, the risk associated with the capital on the euro funds is mainly borne by the insurer while the risk associated with the capital on the unit-linked funds is borne by the insured.

If the insured has a contract in free management mode, he can decide at any time to transfer all or part of the capital from one fund to another : he exercises his arbitrage option. There is thus a transfer of risk, advantageous or not for the insurer. In free management mode, the insurer does not have control over this transfer of risk, which can be sudden, massive and disadvantageous for the insurer in terms of commitment to its policyholders. Research on the subject shows that these phenomena are subject to variables endogenous and exogenous to the life insurance contract.

These phenomena therefore encourage insurers to model arbitrations.

In order to study the arbitrations in the life insurance contracts in free management mode, we build a database where a line gives all the information (mathematical provision and possible arbitration movement among others) of a contract for one month, between January 2015 and May 2021. Then, in chapter 4, we describe how we take into account the information on the funds at the contract level.

In this way, we are able to create a database at the $product \times month$ level indicating the monthly arbitration rates, while having information at the contract level.

To do this, many databases are required (mathematical reserves MR, arbitration movements, information on contracts, net asset values of unit-linked and euro funds, etc.) as well as numerous restatements in order to capture the information specific to each contract. In this respect, a special effort was made at each stage of the database construction, not only on the method of calculating the arbitration rates, but also on the quality of the data.

We also decide to add classical exogenous variables such as the evolution of the MLR and the CAC40, the volatility of the CAC40, but also less classical ones such as the Google trends associated to some keywords, in order to take into account the impact of the economic conjecture on the dynamics of arbitrage.

In this way, we draw up a temporal profile, in the sense of the explanatory variables constructed, for each product, taking into account the information at the contract mesh. These product temporal profiles will be injected into the supervised statistical learning algorithms.

The many ML algorithms used throughout the dissertation are as follows :

A. *Supervised statistical learning* :

   (a) GLM : Generalized Linear Model
   (b) SVM : Support Vector Machine
   (c) RF : Random Forest (supervised mode)
   (d) XGboost : eXtreme Gradient boosting

B. *Unsupervised statistical learning* :

   (a) PCA : Principal Component Analysis
   (b) K-means
   (c) CAH : Hierarchical Ascending Classification
   (d) RF : Random Forest (unsupervised mode)
   (e) GMM : Gaussian Mixture Model

After the constitution of the data, several method/distance/algorithm triples were first tested in order to form homogeneous groups of UL and euro funds. Indeed, the high granularity of the funds requires them to be grouped together in order to carry out relevant analyses later on. This clustering aims at grouping funds with the same statistical and financial characteristics, or with the same historical evolution dynamics. The objective is to map the supports of life insurance contracts in order to take them into account in our modeling.

The groups of unit-linked and euro funds that emerge are therefore characterized differently and provide different information from a different angle of analysis, depending on the method used. We will use the three clusterings presented below :

A. *Temporal-proximity based clustering* :

cluster 1 : funds with an "upward trend and insensitive to economic conditions"

cluster 2 : the "euro's dynamic funds"

cluster 3 : the "risky funds"

cluster 4 : the "dynamic funds sensitive to economic conditions"

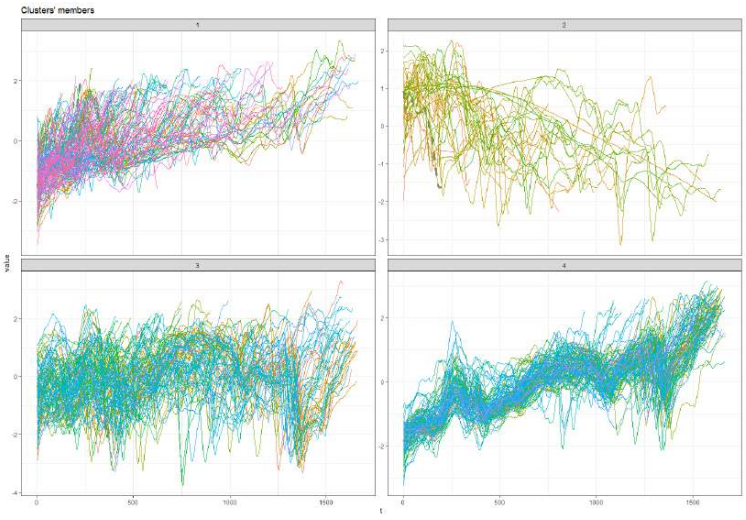


FIGURE 6 – Clustering retained for the evolution dynamics of euro and UL funds

B. *Characteristic based clustering* :

i. Statistical indicators :

cluster 1 : the "moderate risk investment funds with a slight but unstable upward trend".

cluster 2 : the "risky investment funds only characterized by high volatility".

cluster 3 :the "low-risk investment funds characterized by stability and little dynamism, but with a marked upward trend".

ii. Financial indicators :

cluster 1 : the "dynamic funds" characterized by high volatility and elasticity in relation to the CAC40, which therefore have a general upward trend.

cluster 2 : the "moderate funds" characterized as cluster 1, but with more moderate values.

cluster 3 : the "security funds" characterized only by a weak or even downward trend.

As a reminder, we have access to the MR per unit trust and euro fund at the *contract*× *months* grid. By "injecting" the different clusters constructed in the arbitrage database, we have access to the proportion of MP corresponding to each group (according to the different methods and indicators) at the *contract* × *months* grid. This information will be injected into the arbitrage rate prediction models with a view to quantifying the importance or otherwise of the dynamics of the unit-linked and euro funds on the arbitrage rates for free management.

In a second step, descriptive statistics and a univariate analysis are undertaken. These directly compare qualitatively and quantitatively the arbitration rates with the explanatory variables selected. These clearly highlight :

1. the structure of the portfolio studied, by the characteristics of the contracts,
2. the linear structures linking the explanatory variables to the arbitrage rates,
3. the non-linear structures linking the explanatory variables to the arbitrage rates,
4. the asymmetric impact on the different arbitrage origins/destinations (UL_EUR, EUR_UL, and UL_UL) of the explanatory variables.
5. the already established relationships (experts and briefs) linking arbitrage rates to both endogenous and exogenous explanatory variables, but also new ones. In this way, arbitrage behaviors are identified and quantified.

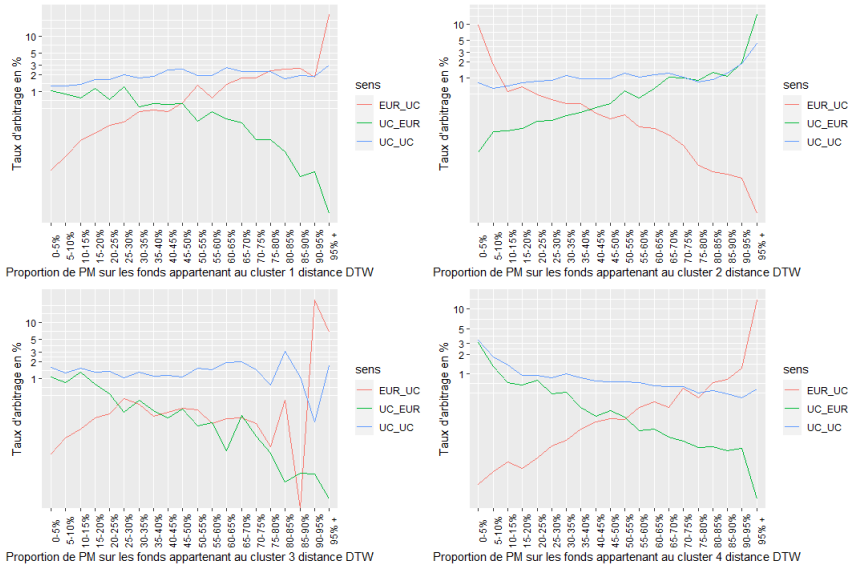


FIGURE 7 – Evolution on the unrestricted managmeent mode portfolio of arbitrage rates as a function of the proportion of MR allocated to each cluster formed by their evolution dynamics (DTW distance + K-means)

Based on these observations, the explanatory variables are divided into classes in order to take into account the discoveries in a *product* × *month* database, which will be used as a basis for the machine learning models. Indeed, descriptive statistics and univariate analysis are not enough to model arbitrage rates : all these phenomena must be taken into account simultaneously, which is why a machine learning approach is adopted with the supervised statistical learning algorithms mentioned above.

Two approaches to modeling rates by ML are considered : a global approach and a specific approach. The global approach allows a direct interpretation of the model via the SHAP method but presents a less good modeling than the specific approach. The specific approach allows a better modeling, at the expense of a certain opacity of interpretation of the model. This trade-off between interpretability and performance is verified quantitatively and qualitatively.

Origin/destination	Approach	Algorithm	RMSE a	RMSE t	MAE a	MAE t	Criteria a	Criteria t
UL_EUR	*global*	*Reg.lin*	2.267	5.682	1.268	2.934	19735	11350
		SVM	1.373	4.576	0.5762	2.409	8845	6989
		RF	1.755	4.701	0.7356	2.079	12733	7948
		XGboost	0.3443	4.338	0.2323	1.886	432	6883
	spécifique	*17 SVM, 6 XGboost, 1 RF, 1 Reg.lin*	1.233	3.669	0.4205	1.451	6698	4779
EUR_UL	*global*	*Reg.lin*	0.1206	0.1504	0.0515	0.0629	32.88	3.55
		SVM	0.083	0.157	0.019	0.0575	12.32	3.31
		RF	0.1371	0.1351	0.0505	0.0536	50.85	2.97
		XGboost	0.0732	0.1243	0.0383	0.0477	14.65	2.25
	specific	*18 SVM, 3 XGboost, 3 RF, 1 Reg.lin*	0.0824	0.1497	0.0251	0.0431	13.651	3.157
UL_UL	*global*	*Reg.lin*	3.508	2.163	1.65	1.534	x	x
		SVM	2.703	1.798	0.7484	1.226	x	x
		RF	3.213	2.263	1.193	1.454	x	x
		XGboost	2.729	1.219	1.093	0.7678	x	x
	specific	*14 XGboost, 7 SVM, 4 RF*	2.243	1.013	0.5733	0.5271	x	x

FIGURE 8 – Summary table of the best models obtained

The critical analysis of the two rate modeling approaches, especially with the SHAP method, leads to some general results :

1. XGboost and SVM are efficient on this problem. It would be advisable to further parameterize them and to feed them with other relevant explanatory variables not taken into account in this thesis, such as the value of the arbitrage rate the previous month. UL_EUR rates pose a modeling problem.

2. The non-linear nature of arbitrage behavior/phenomena is verified by the fact that only complex models ("black boxes") provide suitable results. Linear regression is to be forgotten if one wishes in the future to create a tool for predicting rates.

3. The SHAP method indicates that XGboost can learn certain behaviors detected in the descriptive statistics. This learning capacity is therefore most certainly the cause of the performance of the XGboost. The explanatory variables with the strongest explanatory power are the fund clustering variables (statistical indicators and DTW distance), contract MR, number of supports, CSP, month, and actuarial age. Moreover, we show that the addition of fund clustering variables (mainly statistical

indicators and DTW distance) has a strong explanatory power. These clustering variables overshadow the so-called exogenous variables such as the CAC40 return. They also suggest that only a part (certain funds) of the portfolio could be subject to dynamic arbitrage phenomena.

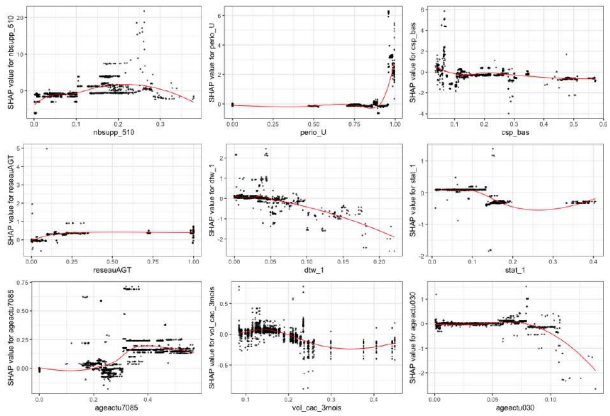


FIGURE 9 – *SHAP plot* of some explanatory variables of the XGboost global approach presenting a relationship with the UL_EUR arbitrage rates

4. XGboost seems to detect quite well the high rates according to certain variables : they can then be trained and used to create a tool for monitoring the risks of massive arbitrage.

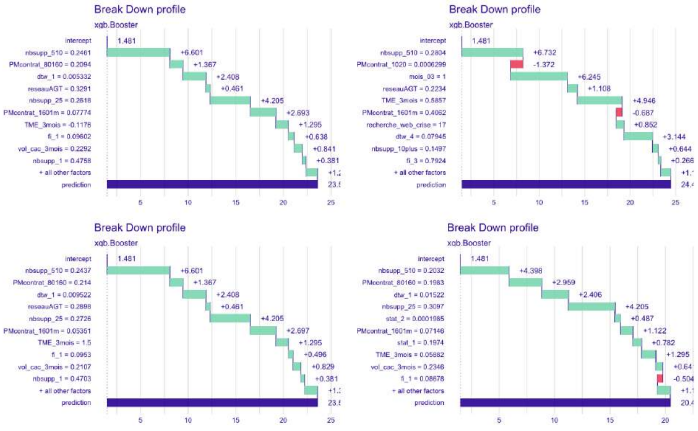


FIGURE 10 – *Break Down profile plot* associated with the four highest UL arbitration rates : identification of the variables causing a high rate

5. The fact that the specific approach presents a better modeling than the global approach, knowing that the analysis of the global approach informs us that it does not really take into account the products, indicates that the products present each

one a different dynamic with respect to the arbitrage rates (due to the specific structure of the contracts composing it) and that the complex models are able to identify their specificities. Nevertheless, the specific approach should also be improved in order to see a real distinction with the global approach.

Many other very local results are obtained and are therefore detailed in the report.

Following this exploratory work on data quality and on a certain number of learning models, we can then consider modeling dynamic arbitrages on the relevant sub-portfolios, or establish tools for the arbitrage risk monitoring. The proposed analysis method can be used in other situations where the data is complex. Despite the promise of these models, there is still a strong need of critical analysis by the actuary.

Table des matières

Remerciements **v**

Note de synthèse **vii**

Introduction **1**

1 Contexte de travail **3**
1.1 Contexte de l'assurance vie en France 3
1.1.1 Généralités sur l'assurance et le contrat d'assurance 3
1.1.2 Zoom sur l'assurance vie en France 5
1.2 Le contrat d'assurance vie en France 8
1.2.1 Définition et différents types de contrats d'assurance vie 8
1.2.2 Briques élémentaires d'un contrat d'assurance vie : les fonds UC et les fonds euros 9
1.2.3 Les modes de gestions d'un contrat d'assurance vie 13
1.2.4 L'option d'arbitrage en assurance vie 14
1.2.5 L'option de rachat en assurance vie 15
1.2.6 Fiscalité des contrats d'assurance vie 15
1.3 Enjeux de la modélisation des arbitrages 17
1.4 Bilan sur le contexte de travail du mémoire 19

2 Données de travail **21**
2.1 Périmètre d'étude 21
2.2 Construction de la base de données : informations endogènes 22
2.2.1 Les bases de données élémentaires 22
2.2.2 Construction de la base de données de l'étude 25
2.2.3 Calcul des taux d'arbitrages 29
2.3 Construction de la base de donnée : informations exogènes 30
2.4 Qualité de la donnée 34
2.5 Bilan sur les données de travail 35

3 Principaux outils utilisés **37**
3.1 Apprentissage statistique supervisé 37
3.1.1 Généralités 37

3.1.2 Quelques notions élémentaires de ML . . . 37
3.1.3 Modèles de régression généralisé (GLM) . . . 40
3.1.4 Séparateurs à Vastes Marges . . . 41
3.1.5 Arbre de décision CART . . . 44
3.1.6 Random Forest (RF) . . . 45
3.1.7 XGBoost . . . 47
3.1.8 Interprétabilité des modèles "boîtes noires" : la méthode SHAP . . . 50
3.2 Apprentissage statistique non-supervisé . . . 52
3.2.1 ACP : Analyse en composantes principales . . . 52
3.2.2 K-means . . . 54
3.2.3 Classification ascendante hiérarchique (CAH) . . . 57
3.2.4 Clustering avec des forêts aléatoires . . . 58
3.2.5 Modèle de mélange Gaussien et apprentissage non supervisé . . . 59
3.2.6 Clustering de séries temporelles . . . 61
3.3 Bilan sur les principaux outils utilisés . . . 69

4 Clustering en groupe homogènes des fonds UC et euros 71
4.1 Clustering des séries temporelles des fonds . . . 71
4.2 Characteristic based clustering . . . 73
4.2.1 Indicateurs "statistiques" . . . 73
4.2.2 Indicateurs "financiers" . . . 89
4.3 Temporal-proximity based clustering . . . 102
4.4 Bilan sur le clustering des fonds UC et euros . . . 109

5 Statistiques descriptives et modèles de prédictions 111
5.1 Statistiques descriptives . . . 111
5.1.1 Caractérisation du portefeuille étudié . . . 111
5.1.2 Relations taux d'arbitrages VS variables explicatives . . . 119
5.1.3 Court résumé de l'apport des statistiques descriptives . . . 149
5.2 Maille produit . . . 150
5.2.1 Base de donnée . . . 150
5.2.2 Matrice de corrélation des variables explicatives . . . 152
5.2.3 Approches considérées . . . 154
5.3 Approche globale . . . 155
5.3.1 Régression linéaire . . . 155
5.3.2 SVM . . . 156
5.3.3 Forêt aléatoire . . . 157
5.3.4 XGboost . . . 158
5.3.5 Bilan : approche globale . . . 159
5.3.6 Analyse des modèles : approche globale . . . 160
5.4 Approche spécifique . . . 175
5.4.1 taux UC_EUR . . . 176
5.4.2 taux EUR_UC . . . 176
5.4.3 taux UC_UC . . . 176

5.4.4 Analyse des modèles : approche spécifique 177
5.5 Bilan sur les statistiques descriptives et les modèles de prédictions 181

Conclusion **183**

Annexes **185**

Table des figures

1 Clustering retenu pour la dynamique d'évolution des fonds euros et UC . . ix
2 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la proportion de PM allouée à chaque cluster formés par leur dynamique d'évolution (distance DTW + K-means) x
3 Tableau récapitulatif des meilleurs modèles obtenus (approche globale et spécifique) xi
4 *SHAP plot* de quelques variables explicatives de l'XGboost approche globale présentant une relation avec les taux d'arbitrages UC_EUR xii
5 *Break Down profile plot* de l'XGboost approche globale associé aux quatre taux d'arbitrages UC_UC les plus élevés : identification des variables à l'origine d'un taux élevé. xii
6 Clustering retained for the evolution dynamics of euro and UL funds . . . xvii
7 Evolution on the unrestricted managmeent mode portfolio of arbitrage rates as a function of the proportion of MR allocated to each cluster formed by their evolution dynamics (DTW distance + K-means) xviii
8 Summary table of the best models obtained xix
9 *SHAP plot* of some explanatory variables of the XGboost global approach presenting a relationship with the UL_EUR arbitrage rates xx
10 *Break Down profile plot* associated with the four highest UL arbitration rates : identification of the variables causing a high rate xx

1.1 Assureur-souscripteur-assuré-bénéficiaire, cas de l'assurance vie. 4
1.2 Visuel provenant du journal Les Échos du 17/08/2019 [32] 7
1.3 Évolution du taux de l'OAT 10 ans France [36] 9
1.4 Caractéristiques principale de la SICAV Allianz Foncier C/D [9] 11
1.5 Valeur liquidative et composition de la SICAV Allianz Foncier C/D [9] . . 12
1.6 Répartition de la PM d'assurance vie et part des supports en UC en France [3] 13
1.7 Représentation des arbitrages possibles entre deux fonds UC et un fond UC d'un contrat fictif à une date donnée. 14

2.1 Extrait de la base MTT-MTO (pour des raisons de place, l'ensemble des colonnes n'est pas affiché) 23

2.2 Extrait de la base des valeurs liquidatives 23
2.3 Extrait de la base Garprinc (pour des raisons de place, l'ensemble des colonnes n'est pas affiché, notamment la PM en euros) 24
2.4 Extrait de la base des informations sur les contrats 24
2.5 Représentation des arbitrages possibles entre le fond euros et les fonds UC 26
2.6 Sens des arbitrages dans le cas A des mouvements d'arbitrages 26
2.7 Sens des arbitrages dans le cas B des mouvements d'arbitrages 27
2.8 Sens des arbitrages dans le cas C des mouvements d'arbitrages 27
2.9 Un exemple de mouvements et arbitrages au sein d'un contrat durant un mois 29
2.10 Capture d'écran de la courbe Google Trends associé à la recherche "assurance vie" 32
2.11 Capture d'écran de la courbe Google Trends associé à la recherche "Allianz" 33

3.1 Métaphore du concept de biais/variance en ML 38
3.2 Schéma de la méthode de validation croisée. Ici k = 5. 39
3.3 Représentation de l'équation de l'hyperplan à marge optimale 42
3.4 Idée principale des séparateurs non linéaires : ici l'astuce du noyau. Les individus ne sont pas "séparables" en dimension 2, mais le sont en dimension 3 en appliquant une fonction (noyau) adéquate. 43
3.5 Chronologie : de l'arbre CART à l'XGboost 47
3.6 De gauche à droite, de haut en bas, un exemple de représentation itérative de l'algorithme des k-means avec $k = 2$ et $p = 2$. [27] 56
3.7 Exemple de dendogramme 58
3.8 Représentation de la différence entre la distance euclidienne et la distance DTW 62
3.9 Illustration de l'apha et du beta d'un actif, ici un fond UC 68

4.1 Représentation d'un échantillon de 9 fonds UC et euros 72
4.2 Extrait de la base de donnée contenant les indicateurs statistiques des fonds UC et euros 73
4.3 Dendogramme (a), comportement des valeurs liquidatives des fonds par cluster (b), et comportement moyen des clusters (c), pour une distance euclidienne + CAH et 3 clusters, sur les séries temporelles résumées en indicateurs statistiques 74
4.4 Pourcentage de variance expliqué par les dimensions de l'ACP sur les indicateurs statistiques 75
4.5 Contribution à la construction des axes de l'ACP des indicateurs statistiques 75
4.6 Biplot de l'ACP avec indicateurs statistiques, et groupes constitués par une CAH avec k = 3 76
4.7 Pourcentage de variance expliqué par les dimensions de l'ACP sur les indicateurs statistiques sélectionnés 77

4.8 Biplot de l'ACP sur la sélection d'indicateurs statistiques des séries temporelles et représentation des cluster indentifié par une CAH à $k = 3$ . . . 78
4.9 Dendogramme (a), comportement des valeurs liquidatives des fonds par cluster (b), et comportement moyen des clusters (c), pour une distance euclidienne + CAH et 3 clusters, sur les séries temporelles résumées en une sélection d'indicateurs statistiques 79
4.10 Silhouette plot des K-means pour des nombres de clusters allant de 2 à 10, sur les indicateurs statistiques sélectionnés 80
4.11 Silhouette plot pour $k = 5$, sur les indicateurs statistiques sélectionnés . . 80
4.12 Représentation des comportements des fonds UC et euros selon les groupes obtenue par un K-means $k = 5$ sur les indicateurs statistiques sélectionnés 81
4.13 Biplot des axes 1 et 2 81
4.14 Biplot des axes 3 et 4 82
4.15 Dendogramme (a) et représentation du comportements des clusters (b) de la CAH associé au clustering par RF non supervisée pour $k = 2$ 83
4.16 Biplot des axes 1 et 2 et représentation des groupes formés par la RF non supervisée 83
4.17 Graphique BIC = f(k,géométrie MMG) du package Mclust sur le logiciel R 84
4.18 Comportement des groupes formés par un MMG de géométrie VEV avec $k = 4$ 85
4.19 Biplot des deux premier axes de l'ACP sur les indicateurs statistiques, avec des groupes formés par un MMG de géométrie VEV à $k = 4$ 85
4.20 Biplot des axes 3 et 4 de l'ACP sur les indicateurs statistiques, avec des groupes formés par un MMG de géométrie VEV à $k = 4$ 86
4.21 Biplot des axes 4 et 5 de l'ACP sur les indicateurs statistiques, avec des groupes formés par un MMG de géométrie VEV à $k = 4$ 86
4.22 Résumé des clusterings des séries temporelles des fonds UC et euros réalisés sur les indicateurs statistiques proposés par Wang et al (2006) [50] 87
4.23 Extrait de la base de donnée contenant les indicateurs financiers des fonds UC et euros. Toutes les variables ne sont pas visibles 89
4.24 Dendogramme (a), comportement des valeurs liquidatives des fonds par cluster (b), et comportement moyen des clusters (c), pour une distance euclidienne + CAH et 3 clusters, sur les séries temporelles résumées en indicateurs financiers. 90
4.25 Pourcentage de variance expliqué par les dimensions de l'ACP sur les indicateurs financiers 91
4.26 Contribution des variables à la construction des dimensions de l'ACP sur les indicateurs financiers 91
4.27 Biplot des deux premiers axes de l'ACP sur les indicateur financiers des séries temporelles des valeurs liquidatives des fonds UC et euros 92
4.28 Silhouette plot des K-means pour des nombres de cluster allant de 2 à 10, sur les indicateurs financiers sélectionnés 93
4.29 Silhouette plot pour $k = 4$, sur les indicateurs financiers sélectionnés . . . 93

4.30 Représentation des comportements des fonds UC et euros par un clustering obtenu par un K-means $k = 4$ sur les indicateurs financiers sélectionnés . . 94
4.31 Biplot de l'ACP sur une sélection d'indicateurs financiers, avec un clustering K-means $k = 4$ et distance euclidienne 95
4.32 Dendogramme (a), comportement des valeurs liquidatives des fonds par cluster (b), et comportement moyen des clusters (c), pour une matrice de similarité obtenue par RF non supervisée + CAH et 2 clusters, sur les séries temporelles résumées sur une sélection d'indicateurs financiers. . . . 96
4.33 Biplot de l'ACP (axes 1 et 2) sur une sélection d'indicateurs financiers, avec un clustering CAH $k = 2$ et une matrice de similarité calculée avec une RF non supervisée. 97
4.34 Biplot de l'ACP (axes 3 et 4) sur une sélection d'indicateurs financiers, avec un clustering CAH $k = 2$ et une matrice de similarité calculée avec une RF non supervisée. 97
4.35 Graphique BIC = f(k,géométrie MMG) du package Mclust sur le logiciel R 98
4.36 Comportement des groupes formés par un MMG de géométrie VVV avec $k = 3$ 99
4.37 Biplot des deux premier axes de l'ACP sur une sélection d'indicateurs financiers, avec des groupes formés par un MMG de géométrie VVV à $k = 3$100
4.38 Biplot des axes 3 et 4 de l'ACP sur une sélection d'indicateurs financiers, avec des groupes formés par un MMG de géométrie VVV à $k = 3$ 100
4.39 Graphique des différents CVI sur des K-means appliqués sur les distances DTW des séries temporelles 103
4.40 Clustering $k = 4$ avec distance DTW couplée à un K-means, réalisé sur les séries temporelles lissées des valeurs liquidatives des fonds UC et euros 104
4.41 Graphiques des différents CVI sur des CAH appliquées sur les distances DTW des séries temporelles 105
4.42 Clustering $k = 4$ avec distance DTW couplée à une CAH, réalisé sur les séries temporelles lissées des valeurs liquidatives des fonds UC et euros . . 106
4.43 Tableau récapitulatif des clustering avec la distance DTW 107

5.1 Evolution de la PM et du nombre de contrats sur l'historique d'observation du portefeuille en GL étudié 112
5.2 Evolution du taux d'arbitrage sur l'historique d'observation du portefeuille en GL étudié 113
5.3 Evolution du taux d'arbitrage sur l'historique d'observation du portefeuille en GL étudié, par origine et destination 114
5.4 Proportion de PM des clusters formés par les indicateurs statistiques(a), indicateurs financiers (b), et distance DTW (c) 115
5.5 Evolution sur le portefeuille en GL de l'âge à la souscription moyen, de l'âge actuariel moyen et de l'ancienneté moyenne des contrats 116
5.6 Evolution sur le portefeuille en GL de la proportion de femmes 117

5.7 Evolution sur le portefeuille en GL de la proportion des différentes périodicités de paiement des primes . . . 118
5.8 Evolution sur le portefeuille en GL de la proportion des différentes CSP . 119
5.9 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de l'âge actuariel de l'assuré . . . 120
5.10 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de l'ancienneté du contrat . . . 121
5.11 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de l'âge à la souscription de l'assuré . . . 122
5.12 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de l'âge à la souscription de l'assuré . . . 123
5.13 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la CSP de l'assuré . . . 124
5.14 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction du sexe de l'assuré . . . 125
5.15 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction du mois d'observation . . . 125
5.16 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la proportion de PM allouée à chaque cluster formé par les indicateurs statistiques . . . 126
5.17 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la proportion de PM allouée à chaque cluster formé par les indicateurs financiers . . . 128
5.18 Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la proportion de PM allouée à chaque cluster formé par la distance DTW + K-means . . . 129
5.19 Nombre de contrats et proportion d'UC en fonction du produit . . . 131
5.20 PM ventilée par type de fond en fonction du produit . . . 132
5.21 PM moyenne par contrat en fonction du produit . . . 133
5.22 Taux d'arbitrages moyens en fonction du produit . . . 134
5.23 Taux d'arbitrages moyens en fonction du nombre de fonds différents . . . 135
5.24 Taux d'arbitrages moyens en fonction du réseau . . . 136
5.25 Taux d'arbitrages global moyens en fonction du réseau . . . 137
5.26 Taux d'arbitrages globaux moyens en fonction de la PM totale des contrats 138
5.27 Taux d'arbitrages moyens en fonction de la périodicité de paiement de prime du contrat . . . 139
5.28 Taux d'arbitrages moyens en fonction de la part d'UC des contrats . . . 140
5.29 Taux d'arbitrages moyens en fonction du rendement du CAC40 à 1,3,6, et 12 mois . . . 141
5.30 Taux d'arbitrages moyens en fonction de la volatilité du CAC40 sur 1,3,6, et 12 mois . . . 143
5.31 Evolution du TME ces dernières années . . . 144
5.32 Taux d'arbitrages en fonction de l'évolution du TME à 1, 3, 6 et 12 mois . 145

5.33 Taux d'arbitrages moyens en fonction de l'importance de la recherche Google associée au mot "assurance vie" . . . 147
5.34 Taux d'arbitrages moyens en fonction de la de l'importance de la recherche Google associée au mot "crise économique" . . . 147
5.35 Matrice de corrélation de Pearson des variables à la maille produit . . . 152
5.36 Corrélation de Pearson des taux d'arbitrages avec les variables explicatives 154
5.37 Tableau récapitulatif des meilleures régressions linéaires obtenues en approche globale . . . 156
5.38 Tableau récapitulatif des meilleurs SVM obtenus en approche globale . . . 157
5.39 Tableau récapitulatif des meilleures forêts aléatoires obtenus en approche globale . . . 157
5.40 Tableau récapitulatif des meilleurs XGboost obtenus en approche globale . 158
5.41 Tableau récapitulatif des meilleurs modèles obtenus en approche globale . 159
5.42 Quelques éléments d'analyses des résultats de l'XGboost champion modélisant les taux UC_EUR en approche globale . . . 160
5.43 Importance des variables de l'XGboost champion modélisant les taux UC_EUR en approche globale . . . 161
5.44 *SHAP summary plot* de l'XGboost en approche globale pour les taux UC_EUR . . . 162
5.45 *SHAP plot* de quelques variables explicatives de l'XGboost présentant une relation avec les taux d'arbitrages UC_EUR . . . 163
5.46 *Break down profile plot* associé aux quatre taux d'arbitrages UC_EUR les plus élevés . . . 165
5.47 Quelques éléments d'analyses des résultats de l'XGboost champion modélisant les taux EUR_UC en approche globale . . . 167
5.48 Importance des variables de l'XGboost champion modélisant les taux EUR_UC en approche globale . . . 168
5.49 *SHAP summary plot* (a) de l'XGboost en approche globale pour les taux EUR_UC et *SHAP plot* de quelques variables explicatives correspondantes . . . 169
5.50 *Break Down profile plot* associé aux quatre taux d'arbitrages EUR_UC les plus élevés . . . 170
5.51 Quelques éléments d'analyses des résultats de l'XGboost champion modélisant les taux UC_UC en approche globale . . . 171
5.52 Importance des variables de l'XGboost champion modélisant les taux UC_UC en approche globale . . . 172
5.53 *SHAP summary plot* (a) de l'XGboost en approche globale pour les taux UC_UC et *SHAP plot* de quelques variables explicatives correspondantes 173
5.54 *Break Down profile plot* associé aux quatre taux d'arbitrages UC_UC les plus élevés . . . 174
5.55 Tableau récapitulatif des résultats obtenus en approche spécifique sur les taux UC_EUR . . . 176

5.56 Tableau récapitulatif des résultats obtenus en approche spécifique sur les taux EUR_UC 176
5.57 Tableau récapitulatif des résultats obtenus en approche spécifique sur les taux UC_UC 176
5.58 Quelques éléments d'analyses des résultats de la modélisation les taux UC_EUR en approche spécifique 177
5.59 Quelques éléments d'analyses des résultats de la modélisation les taux EUR_UC en approche spécifique 179
5.60 Quelques éléments d'analyses des résultats de la modélisation les taux UC_UC en approche spécifique 180

A.1 Contribution d'une sélection d'indicateurs statistiques à la construction des axes de l'ACP 185
A.2 Contributions des indicateurs statistiques à la construction des axes de l'ACP 186
A.3 Dendogramme (a), et comportement des valeurs liquidatives des fonds par cluster (b), pour une distance euclidienne + CAH et 4 clusters, sur les séries temporelles résumées en indicateurs financiers. 187
A.4 Variance des données expliquée par les axes de l'ACP (a), et contribution de la sélection d'indicateurs financiers à la construction des axes de l'ACP (b) 187
A.5 Biplot de l'ACP sur une sélection d'indicateurs financiers 188
A.6 Variance des données expliquée par les axes de l'ACP (a), et contribution de la sélection d'indicateurs financiers à la construction des axes de l'ACP (b), sans la variables "proportion augmentation" 189

B.1 Itérations de la procédure stepwise sur les régressions linéaires des taux d'arbitrages UC_EUR 191
B.2 Valeurs des régresseurs de la régression linéaire retenue par procédure stepwise sur les taux d'arbitrages UC_EUR 192
B.3 R2 ajusté, statistique et test de Fisher, de la régression linéaire obtenue par procédure stepwise sur les taux d'arbitrages UC_EUR 192
B.4 Itérations de la procédure stepwise sur les régressions linéaires des taux d'arbitrages EUR_UC 193
B.5 Valeurs des régresseurs de la régression linéaire retenue par procédure stepwise sur les taux d'arbitrages EUR_UC 193
B.6 R2 ajusté, statistique et test de Fisher, de la régression linéaire obtenue par procédure stepwise sur les taux d'arbitrages EUR_UC 194
B.7 Itérations de la procédure stepwise sur les régressions linéaires des taux d'arbitrages UC_UC 194
B.8 Valeurs des régresseurs de la régression linéaire retenue par procédure stepwise sur les taux d'arbitrages UC_UC 195
B.9 R2 ajusté, statistique et test de Fisher, de la régression linéaire obtenue par procédure stepwise sur les taux d'arbitrages UC_UC 195

B.10 Importance des variables dans les forêts aléatoires modélisant les taux UC_EUR (a), EUR_UC (b) , et UC_UC (c), en approche globale, servant à la sélection des variables 196
B.11 Importance des variables dans les XGboost modélisant les taux UC_EUR (a), EUR_UC (b) , et UC_UC (c), en approche globale, servant à la sélection des variables 197
‘

Liste des sigles et acronymes

ACAV Assurance à Capital Variable

ACPR Autorité de Contrôle Prudentiel et de Résolution

AEC Allianz Expertise Conseil (noté AFC dans les bases de données)

AGT Agent Général

BE Best Estimate

CAH Classification Ascendante Hiérarchique

CSP Catégorie Socio-Professionnelle

CRT Courtier

CVI Cluster Validity Indices

DTW Dynamic Time Warping

EIOPA European Insurance and Ocuppational Pensions Authority

EURIA Euro Institut d'Actuariat

FCP Fond Commun de Placement

FFA Fédération Française de l'Assurance

GL Gestion libre

GLM Generalized Linear Model

GP Gestion Profilée

GSM Gestion Sous Mandat

ISIN International Securities Identification Number

MAE Mean Absolute Error

MMG Modèle de Mélange Gaussien

ML Machine Learning

OAT Obligation Assimilable au Trésor

OPCVM Organisme de Placement Collectif en Valeurs Mobilières

PB Participation aux Bénéfices

PM Provision Mathématique

RF Random Forest

RMSE Root Mean Squared Error

S2 Solvabilité II

SCPI Société Civile de Placement Immobilier

SCI Société Civile Immobilière

SICAV Société d'Investissement à Capital Variable

SVM Support Vector Machine

TME Taux Moyen des Emprunts d'états

TMG Taux Minimum Garanti

UC Unité de Compte

VL Valeur Liquidative

Introduction

Les rendements de fonds euros ne cessent de décroître ces dernières années, conséquence d'une décroissance des rendements obligataires. Ces fonds restent une valeur de refuge pour les ménages français, historiquement averses au risque, puisqu'ils représentent en moyenne 80% de l'encours des contrats d'assurance vie. Les assureurs doivent, contractuellement, garantir cet encours, et même très souvent garantir un taux minimum de rendement du contrat d'assurance vie. Il est alors de plus en plus difficile pour les assureurs d'honorer leur engagements envers leurs assurés. La réglementation impose également aux assureurs un niveau de fonds propres plus élevé lorsque l'encours sur les fonds euros augmente. En effet, l'inadéquation de l'engagement de l'assureur en terme de rendement des contrats d'assurance vie avec la chute des rendements des fonds euros est fortement pénalisée par la directive européenne Solvabilité 2.

Parallèlement, les contrats d'assurance vie en gestion libre présentent l'option d'arbitrage. Cette option permet à l'assuré de transférer tout ou une partie de ses capitaux disponibles sur son contrat d'assurance vie, d'un fond à un autre, que ce soit un fond euros ou un fond UC. De plus, il est constaté que les mouvements d'arbitrages sont soumis à la conjecture économique et aux caractéristiques du contrat.

Dès lors, des pics d'arbitrages massifs sont observés ponctuellement et présentent donc un risque majeur de solvabilité pour l'assureur, au sens de la norme Solvabilité 2.

Cependant, la détermination des taux d'arbitrages, actuellement, ne tient pas compte de la conjecture économique et de l'information disponible à la maille contrat. Il existe donc une asymétrie entre l'actuelle modélisation des taux d'arbitrages et le risque que ceux-ci présentent pour l'assureur.

L'objectif de ce mémoire est donc de proposer une modélisation, tenant compte de l'information de chaque support UC, des taux d'arbitrages des contrats d'assurance vie d'Allianz France en mode de gestion libre en vue d'expliquer les phénomènes d'arbitrages à l'échelle produit.

Nous allons pour se faire commencer par construire une base de données représentant 13.6 milliards d'euros de provisions mathématiques fin décembre 2020 sur 25 produits phares d'Allianz Vie France, à l'échelle $contrat \times mois$. La qualité des données sera un point d'attention non négligeable.

Des variables explicatives classiques endogènes et exogènes au contrat d'assurance vie seront incorporées, mais aussi des variables moins communes comme la typologie des fonds UC composant le contrat ou bien des Goole trend. À partir de ces éléments

nous pourrons étudier le profil de risque des taux d'arbitrages en fonction des variables explicatives en vue de caractériser le portefeuille étudié et d'en dégager quelques résultats sur les éléments impactant les taux d'arbitrages en gestion libre. Cette analyse servira de base à la constitution d'une base de données pertinente vis à vis des taux d'arbitrages, à la maille $produit \times mois$.

Sur cette dernière base de données nous serons alors en mesure d'entraîner des modèles de machine learning (régression linéaire, SVM, RF et XGboost) selon deux approches : soit nous considérons le produit comme une variable explicative d'un modèle global, soit nous considérons un modèle spécifique par produit. Nous montrerons que l'approche globale est légèrement moins performante en terme de modélisation que l'approche spécifique, mais qu'elle présente l'avantage d'être interprétable avec les méthodes d'interprétabilité des modèles "boites noires" .

De l'interprétation de l'approche spécifique, nous conclurons sur la pertinence de la modélisation des taux d'arbitrages avec les variables explicatives sélectionnées, et plus généralement sur la pertinence de la modélisation des taux d'arbitrages par modèles de machine learning.

Ce mémoire est donc à forte composante data, fouille de données, et apprentissage statistique. La taille du rapport étant conséquent du fait de très nombreux graphiques, à chaque fin de chapitre un bilan est dressé.

Chapitre 1

Contexte de travail

1.1 Contexte de l'assurance vie en France

1.1.1 Généralités sur l'assurance et le contrat d'assurance

Le concept d'assurance repose sur plusieurs principes de bases [17] :

- **Présence d'aléa :** la prestation financière ou la prestation de services de l'assureur dépend de la réalisation ou non du risque sur le lequel porte le contrat d'assurance.
- **Cycle de production inversé :** le coût d'un produit d'assurance est connu après sa vente. L'assuré paye immédiatement une prime faible pour avoir une éventuelle prestation élevée. En assurance vie, la prestation sera forfaitaire (prestation fixée à l'avance) tandis qu'en assurance non vie elle sera indemnitaire (prestation égale à la valeur du sinistre)
- **Mutualisation :** l'assureur regroupe les risques de même nature. Ainsi, pour un portefeuille d'assurés portant sur un risque homogène, la prime des uns paie les sinistres, bien moins nombreux, des autres.
- **Protection :** Le métier d'assureur consiste à accepter, contre rémunération, le transfert de risque de l'assuré sur l'assureur. Le transfert du risque à l'assureur donne à l'assuré la garantie immédiate d'une prestation en cas de sinistre. L'enjeu se situe alors dans la quantification et la gestion de ce risque afin d'en dégager des profits.

Les personnes intéressées au contrat d'assurance sont au nombre de quatre :

- **l'assureur :** il s'agit d'une personne morale
- **le souscripteur :** c'est le contractant, personne physique ou morale qui va s'engager avec l'assureur. Il est propriétaire du contrat ; sur lui repose la charge du paiement des primes. C'est lui qui peut décider de modifier ou d'annuler l'assurance.

- **l'assuré :** c'est toujours une personne physique, c'est sur sa tête que repose le risque. C'est de sa survie ou de son décès que dépendra le versement de la prestation par l'assureur. Le souscripteur et l'assuré sont souvent la même personne.
- **le bénéficiaire :** c'est la personne désignée par le souscripteur, pour recevoir la prestation (capital décès, rente ...) garantie par le contrat.

Schématiquement pour une assurance vie :

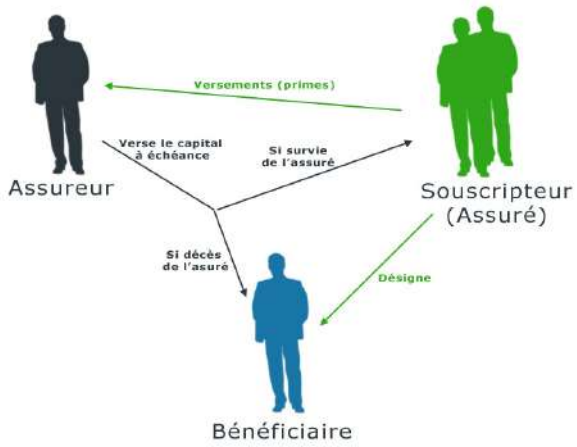


FIGURE 1.1 – Assureur-souscripteur-assuré-bénéficiaire, cas de l'assurance vie.

Le contrat d'assurance est caractérisé par les conditions générales, clauses communes à tous les souscripteurs d'un même type de contrat (définition des termes utilisés dans le document, descriptions des garanties accordées et règles de base de l'assurance comme la formation et la résiliation du contrat), et les conditions particulières, partie du contrat d'assurance qui adapte les CG au cas particulier du souscripteur, en prenant en compte ses coordonnées, le risque, le montant de la cotisation et encore la durée du contrat. [30]. Le contrat d'assurance vient définir par écrit les engagements des parties : primes, prestations, risques couverts, conditions etc... L'assuré a alors l'obligation de payer ses primes contre quoi l'assureur a, lui, l'obligation de payer la prestation au bénéficiaire en cas de réalisation du sinistre.

Si l'assuré ne paie pas ses primes, l'assureur peut se voir suspendre les garanties puis résilier son contrat. Le principe de bonne information est appliqué : un assureur accepte un risque bien défini (clauses d'exclusions, de résiliation, de nullité, de déchéance ou de suspension de garantie) et l'assuré s'engage à lui communiquer toutes les informations susceptibles d'altérer son profil de risques.

Finalement, la fin du contrat peut provenir de diverse raisons : perte de la chose assurée, fin de contrat explicite dans la police d'assurance, liquidation judiciaire, retrait d'agrément de l'assureur par les autorités de contrôle, résiliation.

De plus, l'assureur peut décider, pour diverses raisons, de contracter un traité de réassurance [1] :

- Céder du risque sur son portefeuille d'assuré
- Financer son nouveau produit d'assurance
- Arbitrer une norme (S2, IFRS)
- Profiter de la mutualisation très élevée des réassureurs en vue d'une meilleure tarification technique

.

Une fois la prime versée à l'assureur, celui-ci la place sur les marchés financiers en vue de faire des plus-values. Les bénéfices lui permettent par la suite de couvrir les sinistres en constituant des provisions mais également de se rémunérer, en scénario économique normal du moins. C'est pour cela que l'on dit que le bilan d'un assureur se lit de "droite à gauche", "du passif vers l'actif".

1.1.2 Zoom sur l'assurance vie en France

1.1.2.1 Naissance de l'assurance vie

Dès la première moitié du XVème siècle, sont conclus des contrats d'assurance sur la vie de l'épouse, ou des parents ou de tiers, qui garantissent le contractant à l'égard des pertes éventuelles que le décès de l'un ou des autres aurait pu entraîner. Ces premières formes d'assurance ressemblaient souvent à des paris sur la mort ou la survie d'hommes illustres comme le pape, les rois, empereurs, et étaient alors de pratique courante. Aussi, les grandes lois sur l'assurance, comme l'Ordonnance de Barcelone ou de Colbert en 1681, interdisent l'assurance sur la vie compte tenu de ce caractère de spéculation sur la vie et d'un prix attribué à la vie humaine. Toutefois, les tontines dont les fonds sont confiés à l'État, restent autorisées. En 1762 fut créée à Londres, la société EQUITABLE, première société à pratiquer un tarif en fonction de l'âge, à partir des travaux de PRICE. Jusqu'alors, les sociétés pratiquaient un tarif uniforme, indépendant de l'âge. En 1774, Le GAMBLING ACT fonde l'assurance vie sur des bases plus rationnelles et scientifiques [2].

1.1.2.2 Développement de l'assurance vie et législation afférente

Au XXème siècle, les progrès techniques et économiques ont permis le développement de l'assurance vie. La révolution industrielle a considérablement augmenté le nombre et la gravité des risques et donc le besoin en assurance vie. La loi de 1905 institua un contrôle strict de l'Etat et la loi de 1930 donna une réglementation précise aux contrats d'assurance.

Les activités des organismes assureurs sont régies par des codes distincts selon la forme juridique de leur société : le Code assurances pour les compagnies d'assurances, le Code de la Mutualité pour les mutuelles, et le Code de la Sécurité Sociale pour les institutions de prévoyance.

D'après ces codes, afin de pouvoir exercer leurs opérations d'assurance, ces différentes entreprises doivent en avoir le droit en obtenant les agréments par l'ACPR, l'organe régulateur en France. Il existe un agrément pour chaque branche d'assurance. L'attribution de ces agréments repose sur **trois principes** [8] :

- **Principe de spécialité :** un organisme d'assurance ne peut pratiquer que les opérations pour lesquelles il a obtenu l'agrément. Toutefois, il peut être autorisé, dans certaines conditions, à présenter des garanties pour le compte d'autres organismes agréés avec lesquels il a conclu un accord à cet effet.

- **Principe de spécialisation :** les organismes sont agréés pour exercer des activités exclusivement en assurance vie ou en assurance non-vie. Néanmoins ce principe peut être atténué pour couvrir l'ensemble des risques liés à la personne, les organismes agréés en assurance vie peuvent être également agréés pour couvrir les risques maladie et accident.

- **Principe de l'agrément par branche :** les branches sont définies au niveau communautaire. Il existe 18 branches communautaires en non vie, et 7 branches en assurance vie en France. Les entreprises d'assurance sont autorisées à exercer des activités parmi l'ensemble de ces branches, le champ d'activité est plus restreint pour les mutuelles et les institutions de prévoyance.

Le code des assurances regroupe les opérations d'assurance en 25 sous-branches (pas de sous-branche 19) qui peuvent être rassemblées en 2 groupes : les assurances de dommages et les assurances de personnes.

L'assurance vie est une partie intégrante de l'assurance de personnes (sans les dommages corporels) et correspond aux produits d'épargnes/retraites et santé/prévoyances, c'est à dire les branches 20 à 26 [5] :

- 20. Vie-décès : opérations comportant des engagements dont l'exécution dépend de la durée de la vie humaine, autres que les activités visées aux branches 22,23 et 26

- 21. Nuptialité-Natalité : opérations ayant pour objet le versement d'un capital en cas de mariage ou de naissance d'enfants

- 22. Assurances liées à des fonds d'investissement : opérations comportant des engagements dont l'exécution dépend de la durée de la vie humaine et liées à un fond d'investissement

- 23. Opérations tontinières : opérations comportant la constitution d'associations réunissant des adhérents en vue de capitaliser en commun leurs cotisations et de répartir l'avoir ainsi constitué soit entre les survivants, soit entre les ayants-droit des décédés

- 24. Capitalisation : opération d'appel à l'épargne en vue de la capitalisation et comportant, en échange de versements uniques ou périodiques, directs ou indirects, des engagements déterminés quant à leur durée et à leur montant.

- 25. Gestion de fonds collectifs : opérations consistant à gérer les placements et notamment les actifs représentatifs des réserves d'entreprises et qui fournissent des prestations en cas de vie, en cas de décès ou en cas de cessation ou de réduction d'activité.
- 26.Prévoyance collective : opérations à caractère collectif

Ce mémoire portera sur des produits d'épargne/retraite de la sous-branche n°20, 22 et 24 : les contrats d'assurance vie mono et multi-supports individuels, que propose Allianz plus particulièrement.

1.1.2.3 Quelques chiffres

Le marché de l'assurance-vie représente 38% de l'épargne en France, pour un encours de 2 103 milliards d'euros en 2020. 44,3% des ménages dont la personne de référence a 60 ans ou plus détiennent un produit d'assurance vie en 2018, contre 23,7% des moins de 30 ans. L'épargne des français sur les contrats d'assurance vie est globalement tournée vers les fonds euros. Sachant qu'une bonne partie des fonds euros sont en réalité des obligations d'états, les assureurs sont alors les acteurs majoritaires du financement de la dette française.

Le marché de l'assurance vie est concentré. Les cinq premiers contributeurs représentent 50% des primes en 2020 et les quinze premiers 80%. [20]

La loi Pacte de 2019, vient conforter cette dynamique, puisqu'elle vient faciliter d'un point de vue fiscal le transfert de l'assurance-vie vers un PER ou l'assurance-vie d'une même compagnie.

Les produits d'assurance vie sont relativement liquides, nécessitent peu de temps de gestion de la part du souscripteur, présentent un statut fiscal avantageux, proposent une indemnisation en cas de décès au bénéficiaire et donc la transmission d'un patrimoine, et la constitution d'une épargne présentant un meilleur rendement (1,10% de taux servi moyen en 2020) que le livret A par exemple (0,50% en 2020).

À la vue de ces éléments, l'assurance vie est le produit d'épargne préféré des français :

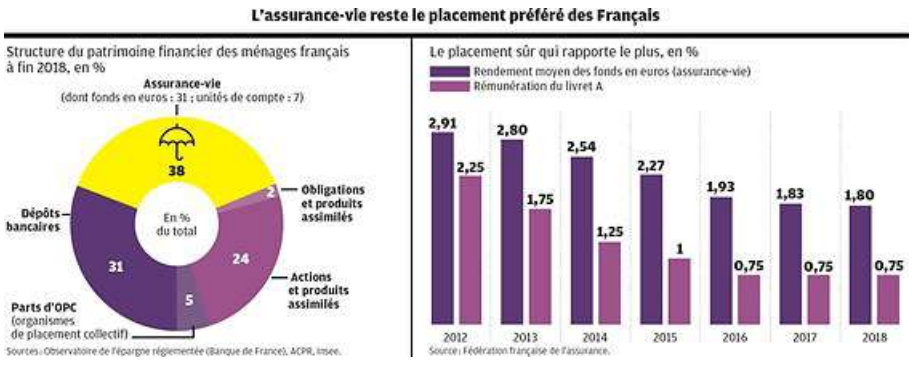


FIGURE 1.2 – Visuel provenant du journal Les Échos du 17/08/2019 [32]

1.2 Le contrat d'assurance vie en France

1.2.1 Définition et différents types de contrats d'assurance vie

D'un point de vue juridique, le contrat d'assurance sur la vie se définit comme celui par lequel, en échange d'une ou de plusieurs primes, l'assureur s'engage à verser au souscripteur ou au tiers désigné, une somme déterminée en cas de survenance d'un événement défini, lié à la durée de la vie humaine. La prestation est donc forfaitaire.

Le cumul d'assurances sur la vie est possible, contrairement aux assurances dommages.

Ces contrats permettent à l'assuré d'épargner en réalisant des versements libres ou programmés. L'épargne reste disponible : l'assuré peut à tout moment demander le rachat de tout ou une partie de son épargne via l'option de rachat.

Le contrat d'épargne peut prendre fin pour des raisons diverses : la demande de rachat total de l'encours par l'assuré, la transformation en rente par exemple au départ de la retraite, le décès de l'assuré ou l'arrivée à terme du contrat.

Ces contrats peuvent correspondre à plusieurs attentes de la part de l'assuré : constituer une épargne qui sera débloquable à terme en capital ou en rente, transmettre un patrimoine en cas de décès, constituer une épargne relativement liquide, piloter une épargne via notamment l'option d'arbitrage, avoir un meilleur rendement que d'autres placements comme le livret A.

Il existe différents types de contrat d'assurance vie avec des caractéristiques et des supports financiers différents. Le choix de ces contrats s'effectue en fonction des objectifs de placement du client [4].

- Les contrats assurance vie **mono-support en euros** : ce type de contrat d'assurance vie ne comporte qu'un seul support : le fond euros. Il est composé de titres de créance et d'obligations du secteur public ou privé. Ces contrats sont donc peu volatiles mais présentent un rendement faible.

- Les contrats assurance vie **multi-supports** : les capitaux sur ces contrats peuvent être investis sur des unités de compte (UC). Ce sont des parts d'OPC : FCP, SICAV, parts de SCI ou de SCPI. Les capitaux peuvent également être investis sur un fond euros. La proportion d'UC peut varier de 0 à 100% selon le profil d'investissement que souhaite adopter le souscripteur. Plus la proportion d'UC est élevée, plus le risque et donc le rendement sont élevés pour le souscripteur.

- Les contrats assurance vie dits **croissance ou Eurocroissance** : ces contrats comportent des supports spécifiques pour lesquels les engagements donnent lieu à une provision de diversification. Ils ont été conçus en vue de réellement offrir une alternative aux contrats mono-support et multi-supports et bénéficient à l'échéance choisie par l'adhérent, d'un minimum de 8 ans, d'une garantie en capital partielle ou totale des sommes investies. Leur création a eu pour objectif d'orienter une partie des encours de l'assurance-vie vers le financement des entreprises comme les PME.

À noter que depuis 2005, il est possible de convertir le contrat d'assurance vie mono-support, en contrat multi-supports afin de dynamiser l'épargne de l'assuré. Cette possibilité de transformation a été introduite par l'amendement Fourgous. Ce transfert s'effectue en conservant les avantages fiscaux associés à l'ancienneté du contrat [47]

1.2.2 Briques élémentaires d'un contrat d'assurance vie : les fonds UC et les fonds euros

Les investissements sur les contrats d'épargne peuvent être réalisés sur deux grandes catégories de supports : les supports euros et les supports en unité de compte (UC).

A. Les supports euros

Ils sont composés majoritairement de titres de créances et d'obligations, mais aussi d'actions. Ils comportent une garantie en capital, voire une garantie de rendement minimum (le TMG : taux minimum garanti annuel, réglementé par l'article A132-1 du code des assurances et défini par rapport au taux moyen des emprunts d'états, le TME).

Le risque financier est donc porté par la compagnie d'assurance qui s'est engagée à garantir le capital injecté, net de frais de gestion, par l'assuré.

Depuis plusieurs années, la diminution des taux de rendement des obligations et titres de créances ont engendré une baisse des rendements des fonds euros. Le graphique ci-dessous illustre cette baisse avec l'évolution des taux d'obligations OAT 10 ans France long terme :

FIGURE 1.3 – Évolution du taux de l'OAT 10 ans France [36]

Les taux n'ont cessé de diminuer, jusqu'à devenir négatifs milieu 2019, et ne sont repassés que récemment positifs en mars 2020 avec la crise de la COVID-19. Cette évolution impacte fortement les placements des assureurs car ils possèdent, d'une part historiquement, beaucoup d'encours sur les fonds euros de par leurs assurés mais aussi, d'autre

part, parce qu'ils investissent eux-mêmes beaucoup dans des titres de créances réputés stables comme les OAT, pour avoir des cash-flow entrant futurs quasiment certains.

Prenons l'exemple simplifié d'un contrat mono-support euros où l'assuré verse une prime unique de 100 €. Des frais de gestion de 2%, pour payer le réseau de distribution par exemple, amputent donc de 2 € la somme investie, soit 98 €. Situons nous à la période où les taux étaient les plus bas en mi-décembre 2020 à -0.38% : placer 98 € revient donc à perdre 0.37 € et donc d'avoir un encours de 97.63 €. Il n'y a pas de gain financier et donc moins de marge financière pour l'assureur, et donc mécaniquement pas de PB financière. Dans le cas fictif où il n'y a pas de frais de gestion (ce qui viendrait alourdir le bilan négatif, mais qui usuellement se situe aux alentours des 0.6%), et où le TMG serait à 0%, l'assureur qui doit garantir les 100 € de primes initiales ferait une moins values de 2.37€ sur ce contrat seul.

Ce problème est bien connu et c'est pourquoi certains annoncent la "mort" des fonds euros : très pratiques pour les assurés mais trop coûteux pour les assureurs actuellement. Les fonds dits Eurocroissances ont été créés en 2014 et modernisés en 2020 dans le but d'améliorer les rendements des fonds en euros et ne garantissent le capital versé que partiellement à partir d'une date future seulement : différence majeure n'obligeant pas l'assureur à disposer de la somme réglementaire à tout moment. Ils ne sont cependant pas encore très répandus à cause de leur complexité.

Une autre approche, plus commerciale, est envisageable pour les assureurs pour éviter les encours trop élevés sur les fonds euros, et permet de se démarquer dans un marché concurrentiel : une prime de bienvenue ou bien une réduction de frais si le contrat souscrit dépasse une certaine proportion de fonds UC, un rendement bonifié (échelonné et croissant avec la part d'UC du contrat) sur le fond euros du contrat si l'épargne investie dépasse là encore un seuil de proportion d'UC, exonération des frais d'arbitrage de l'euro vers l'UC, politique de PB conçue pour inciter la ré-allocation de l'épargne vers les fonds UC (sujet du mémoire de K.Lyoubi [26]) etc...Par exemple, le nouveau produit d'Allianz Vie, nommé AVF, redistribue la PB sur les fonds UC du contrat.

C'est pourquoi peu de nouveaux contrats d'épargne actuellement commercialisés proposent des taux minimum garantis supérieurs à 0%. Cependant, les anciens contrats sur lesquels des TMG , aujourd'hui considérés comme très élevés, ont été garantis et continuent donc à être servis. Pour ces contrats, les assureurs ont aujourd'hui bien du mal à servir les taux contractuels et incitent donc les assurés à changer de produits avec des mécanismes de participation aux bénéfices (PB) présentes en CG du contrat. En effet, la PB oblige l'assureur à redistribuer au minimum 85% du résultat financier aux assurés ayant un contrat d'assurance vie : libre à l'assureur ensuite de ventiler cette redistribution (fonction de sa politique commerciale, avantager/désavantager un produit ancien à fort TMG etc, récompenser la fidélité de certains clients etc...). L'assureur a la possibilité de verser immédiatement la PB en l'affectant aux provisions mathématiques du contrat, ou de la porter à la provision pour participation aux bénéfices (il aura ensuite huit années pour la reverser). Du point de vue de l'assureur, elle permet également de lisser les taux servis années après années en constituant la PPB (provision pour participation aux bénéfices). Les années où l'assureur aura du mal à servir les taux contractuels (années où

les taux sont bas voir négatifs par exemple), il pourra utiliser cette PPB afin d'honorer ses engagements.

Le TMG est donc un véritable outil commercial à manier avec prudence dans le contexte actuel de taux bas voir négatif, et la PB un levier de pilotage des risques de la compagnie d'assurance.

B. Les supports en unités de comptes

Dits supports UC pour lesquels les montants ne sont plus exprimés en euros mais en parts d'UC détenus. **Le risque financier est alors totalement porté par l'assuré** : l'assureur s'engage à garantir la part et non la valeur liquidative des positions de l'assuré. En effet si le fond UC chute de 20%, la part que détient l'assuré sur cette UC est inchangée et l'assureur possède toujours la même part de cette UC sur les marchés financiers, indépendamment de sa valeur. L'encours valorisé est alors obtenu en multipliant le nombre d'unités de comptes détenues par la valeur liquidative de l'UC.

Selon le produit d'assurance vie dont dispose l'assuré, et selon le type de gestion qu'il a choisie, il peut choisir d'investir les sommes sur différents panels de fonds UC, dans le cas de contrats multi-supports.

Les fonds UC comme les fonds euros sont caractérisés par leur code ISIN. Par exemple, l'ISIN FR0000945503 correspond au support Allianz Foncier C/D (SICAV) et dont le site Boursorama, entre autres, nous donne toutes les informations en vue d'un éventuel investissement : *"La SICAV a pour objectif de gestion de permettre une dynamisation des investissements, effectués sur le marché des actions des pays de la zone euro, orientés principalement vers les sociétés foncières et immobilières, afin de rechercher une valorisation du capital à long terme. L'indice FTSE EPRA/NAREIT Eurozone Capped Net Return Index EUR pourra constituer un élément d'appréciation."*[9]

La valeur liquidative ainsi que la composition de cette ISIN est affiché sur la figure ci dessous :

FIGURE 1.4 – Caractéristiques principale de la SICAV Allianz Foncier C/D [9]

Il s'agit bien d'un mixte pondéré d'action de sociétés européennes immobilières et foncières qui présente une très bonne tendance haussière depuis 2013. L'objectif d'investissement affiché par les gestionnaires de ce fond est donc respecté :

FIGURE 1.5 – Valeur liquidative et composition de la SICAV Allianz Foncier C/D [9]

Les UC ont l'avantage d'être divisibles : un assuré souhaitant investir une somme précise sur un fond UC donné va en acquérir la part correspondant à sa valeur liquidative, et inversement lors d'une vente. Ils présentent également des rendements plus élevés que les fonds euros, au prix d'une volatilité et donc d'un risque plus élevé.

Un contrat d'assurance vie en GL dont l'assuré est averse au risque aura donc une proportion conséquente de son capital sur le fond euros tandis qu'un assuré risquophile favorisera les fonds UC plus dynamiques. Il est également commun d'observer une proportion sur les fonds euros croissante avec l'âge de l'assuré : plus celui-ci s'approche de la retraite, plus il souhaite sécuriser son épargne en vue de préparer sa retraite.

C. La provision mathématique (PM) d'un contrat d'assurance vie

À tout instant, un contrat d'assurance vie possède une PM, qui peut être définie comme la différence entre l'engagement de l'assureur avec celui de l'assuré.

L'assureur, selon les garanties souscrites par l'assuré, s'est engagé auprès de celui-ci à lui garantir son capital investi, c'est à dire la somme des primes versées à l'assureur sur les fonds euros et ceux en UC, nette des frais prélevés par l'assureur, et après revalorisation (mécanismes de TMG et PB) et plus-values dégagées par le placement des primes.

À l'émission de la première prime d'assurance, en début d'année disons, la PM est égale juste après ce versement à la prime. C'est la PM d'ouverture de l'année. À la fin de l'année, la PM est égale à la PM d'ouverture sommée avec les différents flux nets de l'année observés sur le contrat : des ajustements ACAV, des intérêts et participation aux bénéfices versés.

La PM représente donc la valeur du contrat. Cette PM peut-être décomposer en PM sur le fond euros et en PM sur les fonds UC, et également en PM sur chaque fond distinct.

Hormis les contrats Eurocroissance, l'assureur doit être en mesure de faire face à ses engagements à tout moment, c'est à dire qu'il doit être en mesure de régler la PM du contrat à l'assuré. Cette PM, qui est donc une provision technique, est donc inscrite au passif de l'assureur et représente la très grande majorité, en valeur, du bilan passif d'un assureur vie.

Historiquement, les français, plutôt averses au risque, favorisent les supports euros avec des parts en fonds UC faibles, comme l'illustre le graphique suivant :

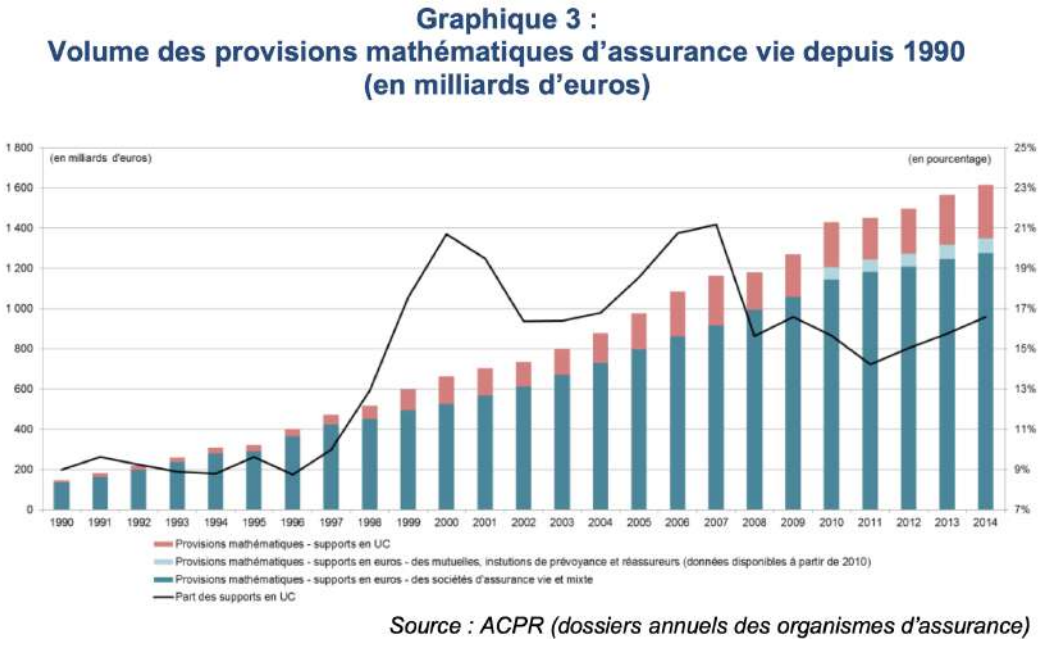


FIGURE 1.6 – Répartition de la PM d'assurance vie et part des supports en UC en France [3]

1.2.3 Les modes de gestions d'un contrat d'assurance vie

Plusieurs types de gestion sont proposés au sein des contrats : la gestion libre, la gestion profilée et la gestion sous mandat. Généralement, l'assuré peut à tout moment basculer d'un mode à l'autre :

- La gestion libre (GL) : l'assuré est autonome et gère son contrat lui-même, c'est-à-

dire qu'il décide des supports d'investissements sur lesquels son épargne est investie, et donne les ordres d'arbitrages.

- La gestion profilée (GP) : l'assuré décide du profil de risque qu'il est prêt à prendre et laisse à l'assureur le soin de répartir son capital sur les différents supports.
- La gestion sous mandat (GSM) : l'assureur agit comme gestionnaire du contrat en fonction du profil de l'assuré, générant les ordres de répartition entre les supports et les arbitrages pour le compte de l'assuré.

Ce mémoire s'attachera à étudier les contrats en GL car il s'agit des contrats sur lesquelles les mouvements ne sont pas prévisibles par l'assureur puisqu'ils sont effectués uniquement par l'assuré.

1.2.4 L'option d'arbitrage en assurance vie

L'arbitrage est l'opération qui consiste à réorienter tout ou une partie du capital constitué sur un ou plusieurs supports (euros ou UC) vers un ou plusieurs autres supports disponibles, et ce, en conservant l'antériorité fiscale (contrairement aux rachats qui ont pour but de réorienter l'épargne vers un autre produit). Schématiquement, pour un contrat avec un fond euros et deux fonds UC, les arbitrages à une date donnée peuvent être représentés comme cela :

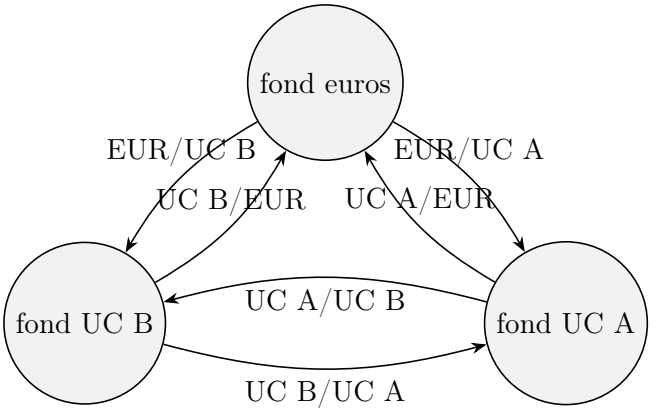


FIGURE 1.7 – Représentation des arbitrages possibles entre deux fonds UC et un fond UC d'un contrat fictif à une date donnée.

Les raisons qui incitent les épargnants à exercer leur option d'arbitrage sont diverses et peuvent être divisées en deux groupes qui interagissent entre eux :

- **Raisons structurelles** : les caractéristiques micro du contrat, comme l'âge et le profil d'investissement, influencent l'exercice ou non de l'option d'arbitrage
- **Raisons conjoncturelles** : les caractéristiques macro-économiques, comme les conditions de marché, une annonce sur de nouvelles mesures fiscales concernant

les produits d'assurance vie, ou bien des phénomènes de contagion entre assurés, comme le met en exergue le mémoire d'actuariat de C.Nicolas [12].

Des études sur le sujet ont déjà étés menées dans le cadre de mémoire d'actuariat et serviront de base à ce mémoire. [12] [11] [7] [40] [6]

Les contrats d'assurance vie multi-supports peuvent imposer (ou non, pour les produits d'assurance vie en ligne) des frais d'arbitrage dont le montant maximal est stipulé dans les CG du contrat. L'assureur peut prélever une commission, variable d'un contrat à l'autre, proportionnelle au montant de l'épargne transférée ou bien fixe. L'assureur peut également, sur certains produits, imposer des frais d'arbitrage à partir d'un certain nombre d'arbitrage : le premier arbitrage de l'année peut être gratuit puis les autres à 25 € par exemple.

En GL, cette option peut présenter un risque pour l'assuré non initié aux mécanismes sous-jacents des assurances vie. En effet, la fuite d'un support en chute occasionnerait des moins-values à la vente. Si de plus, il choisissait un support en augmentation, il prendrait des positions à un prix élevé sans garantie que sa nouvelle position ait cette même dynamique dans le futur.

1.2.5 L'option de rachat en assurance vie

Le contrat d'assurance vie prévoit l'option de racheter son contrat, c'est à dire de récupérer tout ou une partie de l'épargne disponible à tout moment. On parle alors, de rachat total ou partiel. Le rachat d'un contrat peut être lié à de nombreuses raisons que l'on peut distinguer en rachats structurels et conjoncturels (au même titre que l'option d'arbitrage)

La valeur de rachat brute est égale à la provision mathématique (PM) du contrat qui est la somme des investissements augmentés des intérêts contractuels.

Le fait que cette option est déclenchable à n'importe quel moment oblige l'assureur à avoir cette PM disponible. C'est donc un facteur de risque que l'assureur est amené à modéliser (sous-module "lapse" de la pieuvre S2)

Cette option ne sera pas étudiée dans le cadre de ce mémoire.

1.2.6 Fiscalité des contrats d'assurance vie

Le régime fiscal de l'assurance vie impose que les capitaux investis ne soient soumis qu'à la sortie du contrat, et que seule la fraction du capital récupéré au terme ou lors d'un rachat est fiscalisée.

Plus en détail, et selon fédération française de l'assurance [19] :

En cas de rachat :

- Versements effectués à compter du 27 septembre 2017 : Si le montant des versements effectués est égal ou supérieur à 150 000 : les produits sont imposés au taux de 12.8% et en cas de rachat du contrat après 8 ans, la fraction des produits correspondant aux versements effectués en dessous de 150 000 euros est imposée au taux de 7.5%. Si le montant des versements effectués est inférieur à 150 000 euros : les produits

correspondants à ces versements sont imposés au taux de 12.8% et en cas de rachat du contrat après 8 ans, les produits sont imposés au taux de 7.5%. Par ailleurs, en cas de rachat du contrat après 8 ans, le souscripteur du contrat bénéficie d'un abattement annuel sur les produits de 4 600 euros pour une personne seule.

- Versements effectués jusqu'au 26 septembre 2017 : En cas de rachat du contrat avant 8 ans, les produits sont imposés au taux : de 35% en cas de rachat avant quatre ans et de 15% en cas de rachat entre quatre et huit ans et de 7.5% sur les produits en cas de rachat après huit ans. En cas de rachat du contrat après 8 ans, le souscripteur du contrat bénéficie d'un abattement annuel sur les produits de 4 600 euros pour une personne seule.

Les produits sont également soumis à des prélèvements sociaux de 17.2%, quelle que soit la date de versement.

Cela explique les pics de rachats qui sont observés à la huitième année d'ancienneté. La fiscalité est donc une éventuelle raison structurelle de rachat.

En cas de décès de l'assuré :

Le capital ou la rente versée au bénéficiaire lors du décès de l'assuré n'entre pas dans la succession de ce dernier.

- Pour les cotisations versées après 70 ans : Les contrats inférieurs à 30 500 euros sont exonérés des droits de succession. Pour les contrats d'assurance vie dépassant 30 500 euros et souscrits depuis le 20 novembre 1991, les cotisations payées après les 70 ans de l'assuré donnent lieu au règlement de droits de succession, pour la seule partie supérieure à 30 500 euros, selon le degré de parenté entre le bénéficiaire et l'assuré. Les intérêts capitalisés ne sont pas imposables.

- Pour les cotisations versées avant 70 ans : Les contrats inférieurs à 152 500 euros dont les cotisations ont été versées avant les 70 ans de l'assuré sont exonérés de droits de succession. Un prélèvement est dû par chaque bénéficiaire lorsque la part de capital décès qui lui revient excède 152 500 euros et fonctionne par système de tranches et d'abattements que nous ne détaillerons pas ici.

D'autres régimes fiscaux existent également pour les contrats Vie génération et Eurocroissance.

Le montant des prélèvements et contributions sociales s'élève à 15,5% lors du dénouement du contrat par décès.

Impôt sur la fortune immobilière :

Dans le cas où le souscripteur du contrat est soumis à l'impôt sur la fortune immobilière, la valeur de rachat correspondante aux UC immobilières des contrats d'assurance vie est incluse dans le patrimoine du souscripteur.

1.3 Enjeux de la modélisation des arbitrages

Pourquoi modéliser les arbitrages ?

Comme expliqué plus haut, plus l'encours sur les fonds euros est élevé plus l'assureur s'expose au risque. Dans le contexte actuel de taux bas voir négatif, le rendement des supports est très faible : l'assureur va avoir des difficultés à dégager un gain financier suffisant pour, à la fois honorer ses engagements (TMG, PB etc) et à la fois, se rémunérer (marge financière, qui actuellement est très largement réduite par les frais de gestion et de versement). Cela va donc également favoriser les rachats de la part des assurés, qui peuvent se dire que les taux servis sont trop faibles. L'assureur peut également réaliser des pertes en cas de désinvestissement du fond en euros : si l'arbitrage s'effectue du fond en euros vers les fonds UC et que l'opération est réalisée pour l'assureur en moins-value, cela l'oblige à puiser dans la réserve de capitalisation.

Ces phénomènes interagissant entre eux vont donc nécessiter un besoin en fonds propres élevé, une augmentation du coût d'immobilisation du capital, et un risque accru de liquidité pour l'assureur. En effet, sous S2, les assureurs doivent adapter le niveau d'exigence minimale de fonds propres aux risques réels auxquels ils sont exposés : plus il y a de risques, plus l'assureur doit mobiliser des fonds propres en conséquence. L'assureur doit donc constituer des provisions en connaissance des risques qu'il prend en investissant l'épargne des assurés sur les marchés et en prenant en compte l'évolution des marchés. Ceci se traduit par l'application de chocs sur les différents actifs, tels qu'une baisse subite du cours des actions ou des obligations, une hausse importante du nombre de rachats ou de décès dans le portefeuille. Les chocs retenus par l'EIOPA sont issus d'un raisonnement de Value at Risk de niveau 99.5% (VaR 99.5%) : l'assureur doit déterminer le montant de fonds propres réglementaire dont elle doit disposer à horizon 1 an pour ne pas faire faillite avec une probabilité de 99,5%. Autrement dit l'assureur doit calculer le montant de ses fonds propres tel qu'il ne fasse faillite qu'une fois tout les 200 ans en moyenne, compte tenu de son profil de risque.

De plus, à travers le mode de gestion sous mandat et profilé, l'assureur, en tant que gestionnaire, peut limiter et relativement bien appréhender ces risques en effectuant les opérations d'arbitrages qui lui convient : il conserve un certain degré de liberté et une capacité à agir sur ces risques. Ce n'est pas le cas de la gestion libre, pour laquelle seul l'assuré est maître de ses choix d'arbitrages. Dès lors, comment prévoir les grosses masses de mouvements d'arbitrages des fonds UC vers les fonds euros, que pourraient être amenés à réaliser un portefeuille d'assurés en GL pour diverses raisons, et qui seraient susceptible de mettre en péril la solvabilité de la compagnie d'assurance ? Ce risque est-il même quantifiable, obéit-il à certaines règles sous-jacentes ?

On comprend donc aisément par ces mécanismes la nécessité pour les assureurs de modéliser les arbitrages en gestion libre sortant des fonds UC entrant dans les fonds euros, et d'une manière plus générale de comprendre les ressorts sous-jacents liés à ces arbitrages (arbitrages des fonds UC vers les fonds euros, des fonds UC vers d'autres fonds UC, et des fonds euros vers les fonds UC).

Pourtant, l'option d'arbitrage, bien qu'à présent reconnue par les compagnies d'assu-

rance vie, est pour le moment modélisée par avis d'expert. Pour Allianz, les lois d'arbitrages sont calculées en faisant la moyenne pondérée des taux des trois dernières années. Cela a le mérite d'être facile à mettre en place d'un point de vue opérationnel mais montre vite ses limites en périodes économiques anormales à forte volatilité où le comportement des assurés devient irrationnel.

Cela est contradictoire avec les résultats d'études récentes menées sur le sujet : l'option d'arbitrage est soumise à des facteurs conjoncturels et structurels complexes.

D'un côté, la modélisation de l'aspect conjoncturel des arbitrages est complexe et difficile à mettre en place, car elle nécessite de prendre en compte tant des variables exogènes que des variables endogènes, qui nécessitent un retraitement de la donnée conséquent. Mais de l'autre côté, jamais la donnée n'a été si facile à obtenir, et ce, en quantité. Il doit donc exister un juste milieu.

Objectif du mémoire

L'objectif de ce mémoire est alors de chercher à modéliser les taux d'arbitrage des assurés d'Allianz ayant un contrat d'assurance vie mono et multi-supports individuel en mode de gestion libre, en intégrant des informations disponibles dans leurs contrats, ainsi que des données externes reflétant les conditions macro-économiques perçues par les assurés. L'analyse sera menée à la maille produit, en intégrant la donnée disponible à l'échelle ISIN et en utilisant diverses méthodes de machine learning.

1.4 Bilan sur le contexte de travail du mémoire

Le contrat d'assurance vie est un contrat bien défini par la législation française, disposant d'une fiscalité propre. Il s'agit du produit d'épargne préféré des français. Celui-ci présente des mécanismes et options qui lui sont propres : l'option d'arbitrage et de rachat, le mécanisme de participation aux bénéfices, un éventuel TMG, un mode de gestion, une périodicité de paiement de primes etc...

Son principe est simple : l'assuré réparti les sommes investies sur un panel de fonds UC et euros que lui propose l'assureur (souvent fonction du produit d'assurance vie). Ces fonds sont composés de différents actifs sous-jacents. Les fonds euros ont une forte composition en titres de créances et d'obligations, tandis que les fonds UC présentent une forte composition en valeurs mobilières et actifs financiers. Les fonds UC présentent un rendement et un risque plus élevés que les fonds euros. Ils sont exprimés en parts.

L'assureur se doit de garantir le capital investi sur les fonds euros, tandis qu'il n'est tenu de garantir que la part d'UC sur les fonds UC. Dès lors, le risque associé aux capitaux sur les fonds euros est supporté par l'assureur tandis que le risque associé aux capitaux sur les fonds UC est supporté par l'assuré.

Si l'assuré possède un contrat en mode de gestion libre, il peut décider à tout moment de transférer tout ou une partie des capitaux d'un fond à un autre : il exerce son option d'arbitrage. Il y a donc un transfert de risque, avantageux ou non pour l'assureur. En mode de gestion libre, l'assureur n'a pas le contrôle sur ce transfert de risque, qui peut être ponctuellement soudain, massif, et désavantageux pour l'assureur en terme d'engagement auprès de ses assurés. Les recherche menées sur le sujet montrent que ces phénomènes sont soumis à des variables endogènes et exogènes au contrat d'assurance vie.

Ces phénomènes incitent donc les assureurs à modéliser les arbitrages. Ce mémoire s'attachera donc à effectuer ce travail avec une approche machine learning sur un portefeuille de contrats d'assurance vie d'Allianz Vie France en mode de gestion libre.

Chapitre 2

Données de travail

La donnée est la pièce fondamentale de toute étude statistique. Dans le cadre d'une étude statistique sur les arbitrages en assurance vie, la donnée n'est pas "directement" exploitable. Un travail conséquent est alors à fournir pour cibler, extraire, et calculer les variables d'intérêt du périmètre d'étude.

La qualité de la donnée est également un enjeu majeur : avoir confiance en la véracité et la représentativité des données, c'est se permettre d'accorder de l'importance et de faire confiance aux statistiques et modèles développés, et donc de mieux comprendre et piloter les risques en toute connaissance de cause.

Un intérêt tout particulier a été donné à ces deux points et représente 80% du temps de travail consacré à cette étude relatée dans ce mémoire.

2.1 Périmètre d'étude

L'étude menée dans ce mémoire concerne une partie des contrats mono et multi-support d'assurance vie individuelle, en épargne et en retraite d'Allianz France.

Les contrats en mode de gestion libre sont étudiés. Il est possible que, durant la vie d'un contrat, celui-ci passe d'un mode de gestion à un autres : GSM puis GL ou GL puis GSM. Dans ce cas de figure, nous conservons uniquement la partie en GL du contrat.

Nous nous intéressons également uniquement aux contrats appartenant aux réseaux de distribution suivants : AFC (Allianz expertise conseil), AGT (agent général) et CRT (courtage). Le réseau partenariat n'est donc pas pris en compte, mais représente peu de volume.

Nous considérons également un nombre restreint de produits : les produits phares d'Allianz. Ceux-ci sont représentatifs en volume de PM et en nombre de contrats de la totalité des contrats d'assurance vie du portefeuille d'assuré. Les autres produits étant considérés comme "exotiques" et présentant des comportements bien spécifiques selon les experts, nous les excluons afin de ne pas ajouter un éventuel "bruit" dans nos données. Nous décidons également d'étudier les arbitrages sur les garanties principales des contrats.

Enfin, seuls les contrats du 1er janvier 2015 au 31 mai 2021, c'est à dire les contrats ayant été souscrits avant 2015 et toujours en vigueur sur cette période ainsi que les

contrats souscrits durant cette période, sont considérés.

2.2 Construction de la base de données : informations endogènes

Les données exploitées proviennent de l'infocentre GCP et du système Inventaire pour la Vie Individuelle d'Allianz. Elles vont servir à récupérer les information intrinsèques à un contrat afin de modéliser les facteurs structurels influençant la décision pour l'assuré de réaliser un arbitrage ou non.

Le logiciel SAS guide enterprise est utilisé en vue de se connecter aux serveurs d'Allianz et ainsi pouvoir récupérer plusieurs bases de données. Ce logiciel est très efficace pour le traitement de bases de données volumineuses comme celles traitées dans ce mémoire.

Nous utilisons majoritairement 4 bases de données d'Allianz : les "MTT-MTO" qui nous donnent tous les mouvements effectués, les "garprinc" nous donnent les PM, des fichiers excels nous donnent les informations micro des contrats, et les "VL-quotidiennes" nous donnent la valeur liquidative quotidienne des fonds UC et euros. Ces bases permettent la construction de la base de données cible nécessaire à l'étude des arbitrages.

2.2.1 Les bases de données élémentaires

Les bases "MTT-MTO" :

Ces bases des mouvements enregistrent l'ensemble des mouvements effectués sur un contrat à une date donnée et sur un support donné. Ces mouvements sont comptés en euros pour les supports euros, et en parts d'UC et montants en euros correspondant pour les supports UC. Il est important de noter que pour un mouvement, nous avons accès avec cette base uniquement au support d'origine du mouvement : aucune information sur le support éventuel de destination n'est fourni. Néanmoins, pour un mouvement d'arbitrage sortant nous avons un flux sortant d'un fond (une ligne, montant négatif car correspondant à un débit), et pour un mouvement d'arbitrage entrant la somme des flux entrant dans le fond (une autre ligne, montant positif correspondant à un crédit). Cela aura certaines conséquences dans la suite de la constitution des données. De plus, un mouvement à une date donnée pour un contrat n'est associé à aucun mode de gestion. Ainsi pour les contrats concernés par des changements de mode de gestion, la distinction entre un arbitrage effectué en GL et un autre effectué en GSM a nécessité de lourds calculs de retraitements qui ne seront pas détaillés .

Nous prenons les MTT-MTO de fin d'année car elles regroupent tous les mouvements de l'année. Les mouvements sont identifiés par le numéro de compte de plan alternatif qui correspond au plan comptable des assurances (PCA) : afin de récupérer le libellé de ces mouvements et de pouvoir identifier les arbitrages nous croisons cette base avec un référentiel du PCA. Ces mouvements peuvent être des arbitrages, des ajustements ACAV, charges de prestations comme les rachats, participations aux bénéfices, pertes et produits de réalisations, primes émises , provisions etc... Un croisement de ces bases de données

avec un référentiel de réseau de distribution est également réalisé pour pouvoir identifier correctement le réseau de distribution auquel appartient un contrat, ainsi qu'avec un référentiel d'ISIN afin d'identifier le code ISIN d'un code support :

	DATE_COMPTABLE_G...	DATE_COMPTABLE	TRIPTYQUE	NATURE	ID_CONTRAT	Code_ISIN	SUPPORT	TYPE_SUPPORT	QUANTITE_UC	MT_DEVISE_ORIG
1	20151100	28NOV2015	791 UC XXX	791	[redacted]	FR0010135103	XQ0000	U	-40.70482	-25727.89
2	20151100	28NOV2015	791 UC XXX	791		FR0007051040	9N0000	U	70.16633	25727.89
3	20170100	23JAN2017	791 UC XXX	791		ALZ000000314	S40000	U	37.25457	19646.94
4	20170100	23JAN2017	791 UC XXX	791		FR0007051040	9N0000	U	-34.64681	-12668.26
5	20170100	23JAN2017	791 EUR XXX	791		EURO	3S0000	H	0	-6978.68
6	20190600	22JUN2019	791 UC XXX	791		FR0007051040	9N0000	U	-33.87501	-12668.24
7	20190600	22JUN2019	791 EUR XXX	791		EURO	3S0000	H	0	12668.24
8	20160800	13AUG2016	791 UC XXX	791		FR0000449423	3G0000	U	-24.53479	-2327.18
9	20160800	13AUG2016	791 UC XXX	791		LU1228143191	V60000	U	22.58354	2327.18
10	20160800	13AUG2016	791 UC XXX	791		FR0000449423	3G0000	U	-7.85255	-744.83
11	20160800	13AUG2016	791 UC XXX	791		LU1228143191	V60000	U	7.22804	744.83
12	20160800	19AUG2016	791 UC XXX	791		LU1228143191	V60000	U	7.22804	744.83
13	20160800	19AUG2016	791 UC XXX	791		LU1228143191	V60000	U	-7.22804	-744.83

FIGURE 2.1 – Extrait de la base MTT-MTO (pour des raisons de place, l'ensemble des colonnes n'est pas affiché)

Finalement, la concaténation verticale de ces bases de données est nommée la base des arbitrages et comptabilise 46 millions de lignes.

La base "VL-quotidiennes" :

Cette base, qui est une concaténation de plusieurs bases de données de valeurs liquidatives, donne les valeurs liquidatives quotidiennes des fonds, c'est à dire leur cotation quotidienne sur les marchés financiers, et comptabilise 530 000 lignes :

	SUPPORT	date	vl
1	170000	31AUG2012	295.1300
2	180000	31AUG2012	239.2200
3	190000	31AUG2012	314.5500
4	210000	31AUG2012	211.3300
5	220000	31AUG2012	182.1300
6	250000	31AUG2012	185.2000
7	2Z0000	31AUG2012	228.4600
8	300000	31AUG2012	154.1772
9	320000	31AUG2012	239.2200

FIGURE 2.2 – Extrait de la base des valeurs liquidatives

À partir de celle-ci, pour chaque support et pour chaque mois, est extraite sa valeur liquidative en fin de mois, car par la suite, le travail sera effectué avec un pas de temps mensuel avec des PM données en fin de mois. Cette base de données comptabilise 30 000 lignes.

La base des "Garprinc" :

Garprinc (pour garantie principale) contient la valeur des PM en euros pour les fonds euros et la PM en quantité d'UC pour les fonds UC, pour un contrat et un support et à la fin d'un mois donné, que le contrat ait arbitré sur ce support ce mois-ci ou non, de 2015 à 2021. Elle est issue de la concaténation des "garprincs mensuelles" qui donnent la PM pour un contrat, support, et date donnée, à la fin d'un mois. L'accès à la chronique des PM, mois après mois, par contrat et par fond est alors possible.

Cette base est donc très volumineuse et présente 62 millions de lignes environ.

Nous sommes obligés de la croiser avec des référentiels d'ISIN, de réseau de distribution et de valeurs liquidatives mensuelles afin d'obtenir correctement et respectivement le code ISIN associé à un code support, le réseau de distribution, et la valeur en euros des PM des fonds UC en multipliant la quantité d'UC par sa valeur liquidative :

DATE_INVENTAIRE	TRIPTYQUE	NATURE	ID_CONTRAT	TYPE_UC	CODE_OPTION_GESTI..	PM	Montant_Euro	SUPPORT
20200100	281_UC_XXX	281	[redacted]	U	0	7.30486	453.8100	BT0000
20200200	281_UC_XXX	281		U	0	7.30486	442.3100	BT0000
20200300	281_UC_XXX	281		U	0	7.30486	420.5200	BT0000
20200400	281_UC_XXX	281		U	0	7.28678	437.8100	BT0000
20200500	281_UC_XXX	281		U	0	7.28678	446.7300	BT0000
20200600	281_UC_XXX	281		U	0	7.28678	453.5400	BT0000
20200700	281_UC_XXX	281		U	0	7.26875	462.4100	BT0000
20200800	281_UC_XXX	281		U	0	7.26875	470.1200	BT0000
20200900	281_UC_XXX	281		U	0	7.26875	467.1000	BT0000
20201000	281_UC_XXX	281		U	0	7.25076	468.5100	BT0000
20201100	281_UC_XXX	281		U	0	7.25076	485.4300	BT0000
20201200	281_UC_XXX	281		U	0	7.25076	491.1600	BT0000
20210100	281_UC_XXX	281		U	0	7.23281	494.7400	BT0000
20210200	281_UC_XXX	281		U	0	7.23281	495.9100	BT0000
20210300	281_UC_XXX	281		U	0	7.23281	495.0900	BT0000
20210400	281_UC_XXX	281		U	0	7.21491	501.1000	BT0000
20210500	281_UC_XXX	281		U	0	7.21491	501.2700	BT0000
20150100	281_UC_XXX	281		U	0	9.25398	247.4400	XP0000
20150200	281_UC_XXX	281		U	0	9.35009	255.2600	XP0000
20150300	281_UC_XXX	281		U	0	9.4425	258.4000	XP0000

FIGURE 2.3 – Extrait de la base Garprinc (pour des raisons de place, l'ensemble des colonnes n'est pas affiché, notamment la PM en euros)

Les Fichiers excels avec informations des contrats :

Cette base de données, en format excel, donne les informations de l'assuré : numéro de contrat, date d'effet du contrat, le sexe, la date de naissance, l'âge à la souscription et la classe d'âge à la souscription, la périodicité de paiement des primes, ainsi que la catégorie socio-professionnelle.

Ces informations sont également disponibles dans les bases de données MTT-MTO (bases des mouvements) et Garprinc (base des PM), mais avec une qualité de données très médiocre. Il a alors été décidé d'utiliser ces fichiers excels, présentant une bien meilleure qualité de données, provenant d'autres bases de données en vue d'obtenir les informations micros des assurés.

Au final voici un extrait de cette base de donnée :

	ID_CONTRAT	SEXE	DATE_NAISSANCE	AGE_SOUS	DATE_EFFET	CLASSE_AGE	CSP	CSP_LIBELLE	CSP_NUM	DATE_INVENTAIRE	PERIODICITE
198	[redacted]	M	25Mar1943	42	14NOV1985	40-44 ans	4700	technicien	47	20150100	U
199		M	15Aug1954	31	17DEC1985	moins de 35 ans	5500	employ de com..	55	20150100	U
200		F	03Feb1938	47	20DEC1985	45-49 ans	6200	ouvrier qualifie	62	20150100	U
201		M	30Nov1943	42	17DEC1985	40-44 ans	7500	ancien cadre p..	75	20150100	U
202		M	11Feb1955	30	07JAN1986	moins de 35 ans	6200	ouvrier qualifie	62	20150100	U
203		M	15Oct1952	33	02JAN1986	moins de 35 ans	10	agriculteur	10	20150100	U
204		M	15Oct1952	33	02JAN1986	moins de 35 ans	10	agriculteur	10	20150100	U
205		M	15Oct1952	33	02JAN1986	moins de 35 ans	10	agriculteur	10	20150100	U
206		M	25Feb1940	45	24JAN1986	45-49 ans	2130	artisan	21	20150100	U
207		F	23Jul1956	29	20FEB1986	moins de 35 ans	2250	commerçant	22	20150100	U
208		M	19Jul1949	36	11FEB1986	35-39 ans	4200	prof intermedia..	42	20150100	U
209		M	09May1933	52	11FEB1986	50-54 ans	1220	agriculteur	12	20150100	U
210		F	24Apr1943	42	06FEB1986	40-44 ans	7700	ancien ouvrier..	77	20150100	U
211		M	01Apr1971	14	19FEB1986	moins de 35 ans	5400	employ adm e..	54	20150100	U
212		M	09Jun1929	56	03MAR1986	55-59 ans	7210	ancien artisan..	72	20150100	U

FIGURE 2.4 – Extrait de la base des informations sur les contrats

2.2.2 Construction de la base de données de l'étude

Une fois la première étape de la section précédente réalisée, il est alors possible de réaliser une jointure gauche de la base de PM mensuel des contrats x supports avec celle des mouvements sur ces contrats x supports afin d'obtenir la chronique mensuelle des PM, mouvements d'arbitrages, et autres mouvements qui ne sont pas des arbitrages, d'un contrat x support.

Lorsqu'un mouvement d'arbitrage, pour un contrat x support n'est lié à aucune PM dans la base des PM, cela signifie que l'arbitrage constitue un désinvestissement total du fond au profit d'un autre, soit un taux d'arbitrage sortant de 100% et que donc la PM en fin de mois sur ce contrat x support est égale à 0.

Comme dit plus haut, pour un flux d'arbitrage sortant d'un fond (dont on n'a pas la destination), nous avons symétriquement dans notre base de donnée un flux d'arbitrage entrant codé dans une autre ligne (dont on n'a pas l'information sur le fond émetteur). Cependant nous avons accès pour chaque support tant UC que euros, et sur chaque ligne, accès à son type de support (euros ou UC). Ainsi, pour un contrat x support x mois, nous sommes capables de calculer la PM en fin de mois des fonds euros et UC, ainsi qu'aux flux d'arbitrages euros vers UC, euros vers euros, UC vers euros et UC vers UC. La méthodologie adoptée est la même que celle du mémoire de C.Nicolas dans son mémoire [12], à savoir :

notons :

- EUR_in : mouvements d'arbitrages cumulés entrants dans le fond euros
- EUR_out : mouvements d'arbitrages cumulés sortants du fond euros
- UC_out : mouvements d'arbitrages cumulés sortants des fonds UC
- UC_in : mouvements d'arbitrages cumulés entrants sur les fonds UC

et

- EUR/EUR : arbitrages cumulés du fond euros vers le fond euros
- EUR/UC : arbitrages cumulés du fond euros vers les fonds UC
- UC/UC : arbitrages cumulés des fonds UC vers les fonds UC
- UC/EUR : arbitrages cumulés des fonds UC vers le fond euros

Alors on a :

$$EUR\_in + UC\_in = EUR\_out + UC\_out \tag{2.1}$$

c'est à dire que "tout flux entrant est sorti de quelque part, pour une date donnée"

Grâce à cette relation nous pouvons relier les mouvements d'arbitrages sans destination EUR_in, UC_in, EUR_out et UC_out aux mouvements d'arbitrages EUR/EUR, EUR/UC, UC/UC et UC/EUR possédant une origine et une destination.

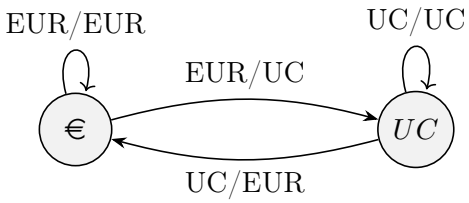


FIGURE 2.5 – Représentation des arbitrages possibles entre le fond euros et les fonds UC

C'est à dire relier les mouvements d'arbitrages sans destination à :

Ainsi, en se positionnant du point de vue du fond euros, déterminer les flux d'arbitrages, avec fond d'origine et fond de destination (EUR/EUR, EUR/UC, UC/UC et UC/EUR) revient à distinguer trois cas :

cas A : si $EUR\_in > EUR\_out$: les mouvements d'arbitrages sont globalement orientés vers le fond euros et alors $EUR/EUR = EUR\_out$, $EUR/UC = 0$, $UC/EUR = UC\_out - UC\_in$ et $UC/UC = UC\_in$

exemple A : $EUR\_in = 20$, $EUR\_out = 10$, $UC\_in = 80$ et $UC\_out = 90$

L'équation 2.1 est respectée.

Alors les arbitrages "rentrent" plus qu'ils ne "sortent" de l'euro et les arbitrages "sortent" plus qu'ils de "rentrent" de l'UC. Cela se traduit par des arbitrages EUR/EUR, UC/EUR et UC/UC mais pas EUR/UC comme le montre la figure ci-dessous :

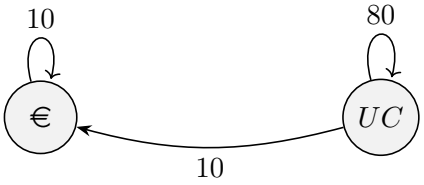


FIGURE 2.6 – Sens des arbitrages dans le cas A des mouvements d'arbitrages

cas B : $EUR\_in < EUR\_out$: les mouvements d'arbitrages sont globalement orientés vers les fonds UC et alors $EUR/EUR = EUR\_in$, $EUR/UC = EUR\_out - EUR\_in$, $UC/EUR = 0$ et $UC/UC = UC\_out$

exemple B : $EUR\_in = 10$, $EUR\_out = 40$, $UC\_in = 90$ et $UC\_out = 60$

L'équation 2.1 est respectée.

Alors les arbitrages "rentrent" plus qu'ils ne "sortent" de l'UC et les arbitrages "sortent" plus qu'ils de "rentrent" de l'euro. Cela se traduit par des arbitrages EUR/UC, EUR/EUR et UC/UC mais pas UC/EUR comme le montre la figure ci-dessous :

cas C : $EUR\_in = EUR\_out$: les mouvements d'arbitrages entre les différents types de fonds sont globalement centrés sur eux mêmes et alors $EUR/EUR = EUR\_in = EUR\_out$, $EUR/UC = 0$, $UC/EUR = 0$ et $UC/UC = UC\_in = UC\_out$

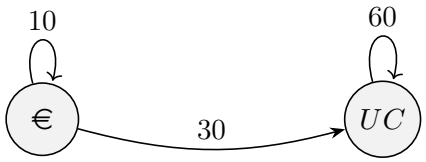


FIGURE 2.7 – Sens des arbitrages dans le cas B des mouvements d'arbitrages

exemple C : $EUR\_in = 20$, $EUR\_out = 20$, $UC\_in = 80$ et $UC\_out = 80$
L'équation 2.1 est respectée.

Alors les arbitrages "rentrent" autant qu'ils ne "sortent" de l'UC et les arbitrages "entrent" autant qu'ils de "sortent" de l'euro. Cela se traduit par des arbitrages EUR/EUR, UC/UC mais pas UC/EUR ni EUR/UC comme le montre la figure ci-dessous :

FIGURE 2.8 – Sens des arbitrages dans le cas C des mouvements d'arbitrages

En appliquant cette méthodologie, il a alors été possible d'avoir pour les contrats x mois x type de fond la PM du fond euros, la PM du fond UC, et les arbitrages EUR/EUR, EUR/UC, UC/UC et UC/EUR.

Enfin, une jointure gauche est effectuée entre cette base et celle des informations sur les contrats.

Ainsi, une ligne décrit les mouvements d'arbitrages (ou non) d'un contrat pour un mois donné. Si un contrat possède 12 lignes, cela veut dire que celui-ci possède une chronique sur 12 mois, à la maille type de support (fond euros, fond UC).

Cette base de donnée contient 27 680 000 lignes.

Au final, la base de données obtenue comporte les variables endogènes suivantes :

- **Date comptable** : par exemple mars 2020 est codé 20200100
- **Numéro de contrat** : identifiant unique du contrat
- **EUR_in, EUR_out, UC_in, UC_out** : somme des mouvements d'arbitrages par sens et par type de fond d'origine durant le mois
- **EUR_in_autres, EUR_out_autres, UC_in_autres, UC_out_autres** : somme des mouvements autres que ceux d'arbitrages, et dont la date est supérieure à la date minimum d'arbitrage sur le mois, par sens et par type de fond d'origine. Ces variables sont décrites dans la sous-section suivante

- **EUR_EUR, EUR_UC, UC_UC, UC_EUR** : somme des arbitrages par type de support d'origine et type de support de destination durant le mois, recalculés selon la méthodologie décrite plus haut
- **taux_EUR_EUR, taux_EUR_UC, taux_UC_UC, taux_UC_EUR** : taux d'arbitrage par type de support d'origine et type de support de destination, calculé sur le mois. Le calcul de ces taux est expliqué dans la sous section suivante
- **PM_EUR** : le montant en euros de PM du fond euros à la fin du mois
- **PM_UC** : le montant en euros de la PM des fonds UC à la fin du mois
- **check_somme_arbitrage** : variable calculant la somme des arbitrages. Elle doit être égale à 0 selon l'équation 2.1
- **nb_support** : compte le nombre de supports distincts présents sur le contrat à la fin du mois
- **Nature** : produit d'appartenance du contrat
- **Reseau_CDG** : réseau de distribution d'appartenance du contrat.
- **Sexe** : sexe de l'assuré
- **date_naissance** : date de naissance de l'assuré
- **age_souscription** : âge à la souscription du contrat
- **age_actuariel** : âge actuariel de l'assuré à la fin du mois
- **annciennete** : ancienneté du contrat à la fin du mois
- **classe_age** : classe d'âge à la souscription du contrat de l'assuré. Un assuré de 25 ans appartiendra à la classe "moins de 35 ans", et un assuré de 42 ans appartiendra à la classe "40-44ans" par exemple
- **date_effet** : date de souscription du contrat
- **CSP_libelle** : libellé de la catégorie socio-professionnelle de l'assuré
- **Periodicite** : périodicité de l'émission des primes

Il a également été décidé de conserver, pour chaque contrat x mois, le détail des fonds UC et euros en présence et de leur proportion en terme de PM. En effet, dans la suite de l'étude nous souhaitons conserver ces informations en vue d'expliquer les arbitrages. L'idée est de considérer, mois après mois, les fonds plus ou moins risqués au sens de l'option d'arbitrage, en prenant en compte leur proportion en terme de PM dans le contrat de l'assuré. Il s'agit là de l'apport majeur de ce mémoire. Cela nécessite donc en amont

de construire des groupes de risques homogènes d'ISIN UC et euros. Cette étape sera décrite plus loin.

Cette base de donnée servira d'élément de base pour calculer les taux d'arbitrages et agréger les informations à la maille produit. En effet, usuellement, le pilotage des risques sur les contrats d'assurance vie ne se font généralement pas à des mailles très fines (réseaux x mode de gestion).

2.2.3 Calcul des taux d'arbitrages

Une fois les arbitrages regroupés par type de support d'origine et type de support de destination, il faut prendre en compte les autres mouvements qui ont eu lieu sur le contrat entre l'arbitrage et la fin du mois car la PM de fin de mois est accessible. Par exemple, dans cette situation :

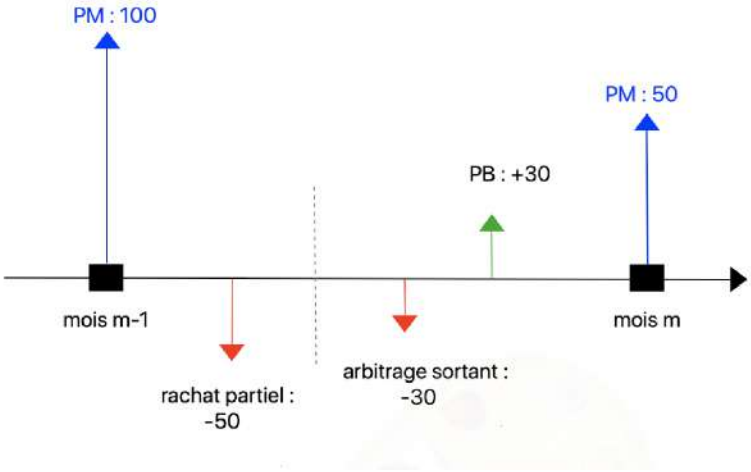


FIGURE 2.9 – Un exemple de mouvements et arbitrages au sein d'un contrat durant un mois

Trois calculs de taux d'arbitrages sont possibles dans cette situation (la valeur absolue est prise pour plus de clarté) :

- A. Ignorer les mouvements autres que les arbitrages :

$$taux = \frac{arbitrage_{sortant}}{PM_m + arbitrage_{sortant}} = \frac{30}{50+30} = 37.5\%$$

- B. Prendre en compte tous les mouvements autres que les arbitrages :

$$taux = \frac{arbitrage_{sortant}}{PM_m + (rachat_{partiel} + arbitrage_{sortant} - PB)} = \frac{arbitrage_{sortant}}{PM_{m-1}} = \frac{30}{100} = 30\%$$

- C. Prendre en compte uniquement les mouvements intervenant après l'arbitrage :

$$taux = \frac{arbitrage_{sortant}}{PM_m + (arbitrage_{sortant} - PB)} = \frac{30}{50 + (30 - 30)} = 60\%$$

Évidemment, c'est cette dernière qui est la meilleure approximation puisque un taux d'arbitrage se calcule en faisant le rapport de la somme arbitrée sortante divisée par la PM juste avant cet arbitrage. Cet exemple bien que simpliste, illustre bien que la manière de calculer les taux peut impacter grandement le taux obtenu. Nous utilisons donc la méthode C.

Dans le cas où plusieurs arbitrages ont lieu et plusieurs mouvements autres que des arbitrages ont également lieu au sein d'un mois, la date minimum d'arbitrage du mois est retenue. Seuls les mouvements qui ne sont pas des arbitrages et dont la date est supérieure à cette date minimum sont considérés. De cette manière le calcul du taux d'arbitrage se rapproche du taux théorique, et ce, même lors de l'agrégation à la maille $contrat \times mois \times typedesupport$.

C'est pourquoi les calculs des taux d'arbitrages sont calculés de la manière suivante (dans la base de données, tous les montants sont positifs) :

$$taux\_EUR\_UC = \frac{EUR\_UC}{PM\_EUR\_m + EUR\_UC - EUR\_in\_autres + EUR\_out\_autres}$$

$$taux\_UC\_UC = \frac{UC\_UC}{PM\_UC\_m - UC\_in\_autres + UC\_out\_autres}$$

$$taux\_UC\_EUR = \frac{UC\_EUR}{PM\_UC\_m + UC\_EUR - EUR\_in\_autres + EUR\_out\_autres}$$

2.3 Construction de la base de donnée : informations exogènes

Il est également souhaitable d'intégrer des variables exogènes aux contrats afin de modéliser les facteurs conjecturaux influençant la décision pour l'assuré d'effectuer ou non un arbitrage.

Les experts le constatent, et des recherches sur le sujet, notamment des mémoires d'actuariat, l'ont déjà démontré.

Pour ce faire, nous choisissons d'une part d'ajouter des variables classiques ayant déjà fait leur preuve comme l'évolution du CAC40 et du TME (les données sont accessibles sur différents sites comme yahoo finance au format .csv), et d'autre part d'ajouter de nouvelles variables moins communes.

Le CAC40

Le CAC 40, indice phare de la Bourse de Paris, est composé de 40 valeurs françaises représentant l'ensemble des secteurs d'activité, sélectionnées parmi les 100 plus fortes

capitalisations. Cet indice est donc représentatif de l'évolution des cours des actions et de la tendance globale de l'économie des grandes entreprises françaises. Il est donc naturel de se pencher vers cet indice comme variable afin de prendre en compte la dynamique économique d'une période, en lien avec des contrats d'assurance vie françaises.

Nous nous intéressons donc aux taux d'évolution du CAC40 en considérant :

$$return\_CAC_{x\ mois} = \left(\frac{CAC_m}{CAC_{m-x}}\right)^{\frac{1}{x}} - 1$$

On pourra prendre $x \in \{1, 3, 6, 12, 24, 48\}$ qui représente le delta de temps souhaité.

La volatilité annualisée du CAC40 est également considérée afin d'éventuellement prendre en compte l'aversion au risque des assurés :

$$volatilite\_CAC_{xmois} = \sqrt{\frac{255}{N_x - 1} \times \sum_{i=1}^{N_x} (R(i) - \bar{R})^2}$$

N_x est le nombre de variations journalières observées sur la période de temps considérée, $R(i)$ la i^{eme} variation journalière, et $\bar{R}$ la moyenne des variations journalières sur la période de temps considérée. La volatilité annualisée du CAC40 sur 1, 3, 6, et 12 mois est alors calculée.

Le TME

Le TME correspond au taux moyen de rendement des emprunts d'état et des OAT émises par l'état français, à taux fixe, et d'une durée supérieure à 7 ans. Il sert de référence aux banques et aux assurances pour déterminer le niveau des taux d'intérêts fixes. Or Le TMG est est égal (au minimum) à 60% des 6 derniers TME. Prendre en compte le TME dans les variables exogènes, c'est donc prendre en compte le TMG, qui si il est très faible peut provoquer des rachats, ou bien inciter l'assuré à dynamiser son épargne en réalisant des arbitrages vers le fond UC. Il est alors proposé de prendre l'évolution du TME avec la formule suivante :

$$evo\_TME_{x\ mois} = \left(\frac{TME_m - TME_{m-x}}{TME_{m-x}}\right),\ x = 1, 3, 6, 12$$

Les Google trends

Google Trends est un outil mis en place par Google en 2006 pour identifier le nombre de fois qu'un terme a fait l'objet d'une requête dans son moteur de recherche : il permet donc d'analyser la popularité d'un terme sur le moteur de recherche, dans une période de temps déterminée. Une "courbe Google Trends" est donc une courbe représentant l'évolution du nombre de recherches d'un terme en fonction du temps. La courbe n'indique pas un nombre de recherches mais une proportion entre 0 et 100, où 100 représente la quantité maximale d'utilisation du terme dans la période et le lieu défini. Évidemment, l'algorithme caché derrière demeure secret.

L'idée est de se dire que les assurés, qui ne sont pas tous sensibilisés aux marchés financiers, ne suivent pas tous l'actualité économique au jour le jour, mais qu'ils sont

plus ou moins soumis à des mouvements de tendances (bouche à oreille, informations télévisées etc...). Ainsi, obtenir une mesure de popularité de recherche internet peut amener à modéliser ce phénomène, surtout si il s'agit du moteur de recherche internet le plus utilisé au monde.

Les google trends associés aux mots clefs "assurance vie", "crise économique" sont choisis.

Voici par exemple la courbe Google Trends associée à la recherche "assurance vie" en France. Elle met en évidence un effet de saisonnalité (pics en début d'année, période à laquelle l'assuré est informé de sa PB et de son taux servi) assez marquant :

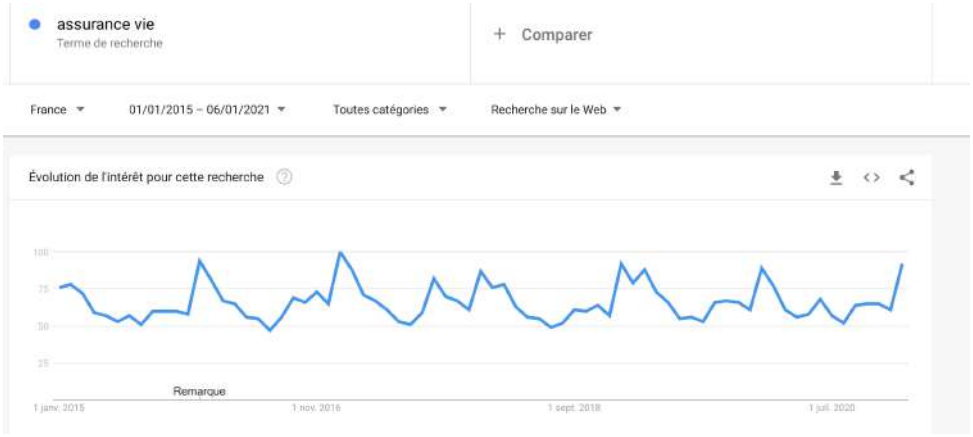


FIGURE 2.10 – Capture d'écran de la courbe Google Trends associé à la recherche "assurance vie"

Un autre exemple : la courbe Google Trends associée à la recherche "Allianz" en France qui possède une légère tendance haussière au fil du temps, mais qui présente une très forte chute de popularité durant le premier confinement en mars et avril 2020 :

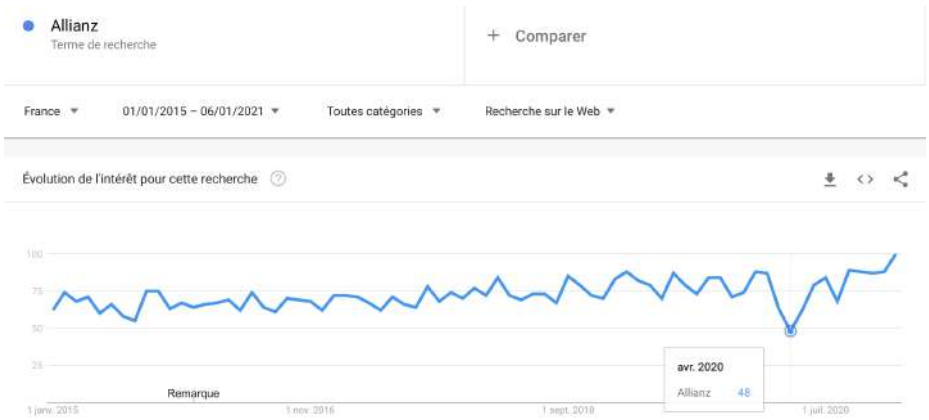


FIGURE 2.11 – Capture d'écran de la courbe Google Trends associé à la recherche "Allianz"

Il est important de remarquer que les Google trends sont constatées à posteriori. Elle n'ont donc pas de valeurs prédictives possibles (bien que Google mette à disposition un outil de prédiction des recherches), mais seulement une valeur explicative.

2.4 Qualité de la donnée

À chaque étape de la construction, des actions sont mises en place afin de garantir une certaine qualité des données. En effet, la proximité de ces données avec le système de gestion fait que des actions sont à mener en vue d'avoir une certaine confiance, à savoir :

- **La bonne réconciliation des bases de données** : sur les bases de mouvements et les bases de PM par exemple, les contrats présentant des valeurs différentes entre ces deux bases pour des variables ayant une valeur unique par contrat (par exemple, le réseau de distribution ou le produit qui sont censés être uniques) sont supprimés. On vérifie que tous les contrats dans la base d'arbitrage sont également présents dans la base des PM mensuelles. On vérifie également que la PM recalculée avec nos bases correspond bien à celle qui est remontée dans d'autres outils du groupe.
- **Des contrôles systématiques sont effectués** : par exemple, le respect de l'équation 2.1 pour un contrat x mois est contrôlée via la variable *is_ check_ somme_ arbitrage*. Celle-ci est bien égale à 0 en majorité sur nos 27 millions de lignes. L'écart le plus élevé observé par cette variable est de 32 € seulement.
- **La PM mensuelle est recalculée de proche en proche** selon une méthode itérative de rétro propagation : la PM à la fin du mois m-1 est égale à la PM de la fin du mois m à laquelle on enlève tous les mouvements observés sur ce contrat, et ainsi de suite. Si l'écart relatif constaté est de plus de 5%, alors le contrat est supprimé.
- **Lors des jointures gauches** : veiller à ce qu'il n'y ait aucun doublon ou aucune ligne créée.
- **Pour la base des valeurs liquidatives quotidiennes des UC**, les valeurs aberrantes sont repérées et supprimées.
- **Lors de suppressions de contrats** : la perte en PM est regardée et celle-ci n'est jamais suffisamment conséquente pour reconsidérer la suppression des contrats.

Le parti pris de supprimer de nombreuses lignes est un choix assez naturel dans la mesure où les bases de données sont de tailles très conséquentes : la perte d'information est alors très faible.

2.5 Bilan sur les données de travail

En vue d'étudier les arbitrages dans les contrats d'assurance vie en mode de gestion libre, nous construisons une base de données dont une ligne donne toutes les informations (provision mathématique et éventuel mouvement d'arbitrage entre autres) d'un contrat pour un mois, entre janvier 2015 et mai 2021. Par la suite, dans le chapitre 4 nous décrivons comment nous prenons en compte l'information sur les ISIN à la maille contrat.

De cette manière nous sommes en mesure par la suite de constituer une base de données à la maille $produit \times mois$ nous indiquant les taux d'arbitrages mensuels, tout en disposant des informations à la maille contrat.

Pour ce faire, de nombreuses bases de données sont requises (bases de provisions mathématiques, de mouvements d'arbitrages, d'informations sur les contrats, de valeurs liquidatives des fonds UC et euros etc...) ainsi que de nombreux retraitements afin de capturer l'information propre à chaque contrat. À ce titre un effort particulier sur la méthode de calcul des taux d'arbitrages, mais aussi sur la qualité de la donnée, a été fourni à chaque étape de la construction de la base de données.

Nous décidons également d'ajouter des variables exogènes classiques comme l'évolution du TME et du CAC40, la volatilité du CAC40, mais aussi moins classiques, comme les Google trends associés à certains mots clefs, afin de tenir compte de l'impact de la conjecture économique sur la dynamique des arbitrages.

De cette manière, nous dressons un profil temporel, au sens des variables explicatives construites, pour chaque produit en tenant compte de l'information à la maille contrat. Ces profils temporels de produits serons injectés dans les algorithmes d'apprentissage statistique supervisé.

Chapitre 3

Principaux outils utilisés

Ce chapitre va s'attacher à apporter les éléments mathématiques nécessaires portant sur les outils utilisés durant le travail développé dans ce mémoire.

3.1 Apprentissage statistique supervisé

3.1.1 Généralités

L'apprentissage statistique supervisé désigne l'ensemble des méthodes pour lesquelles on utilise un ensemble de données dont on connaît la valeur de la variable-cible afin de construire un modèle. Il est dit "supervisé" car on force le modèle à converger vers des valeurs cibles Y : il apprend. L'apprentissage statistique, supervisé ou non, est plus communément appelé machine learning (ML)

De manière générale, nous sommes en présence de données avec X la matrice des variables explicatives (usuellement, une ligne représente l'un des n individus décrit par p variables en colonnes), et Y le vecteur de la variable à expliquer. Y peut être de plusieurs natures selon la nature du problème soumis à l'algorithme d'apprentissage statistique : $Y \in \mathbb{R}$ pour une régression, et $Y \in \{1, 2, ..., K\}$ dans le cas d'une classification à K-classes.

Il est donc question de la recherche d'une fonction f liant les variables explicatives à la variable à expliquer : $Y = f(X) + \epsilon$ avec ϵ un bruit centré. Usuellement, $\hat{Y}$ désigne la prédiction faite par un modèle : $\hat{Y} = f(X)$. La difficulté pour l'algorithme d'apprentissage va donc être d'approximer la fonction liante f sans pour autant apprendre le bruit ϵ.

La description mathématique des algorithmes qui suivent sont pour la plupart tirés du cours de Master 1 de l'EURIA d'apprentissage statistique donnée par F.Vermet [18] et P.Aillot [37], les articles originaux les ayant présentés, ainsi que différents articles et sites synthétisant l'information .

3.1.2 Quelques notions élémentaires de ML

Le compromis biais/variance

Il existe deux sources d'erreurs qui empêchent les algorithmes d'apprentissage statistique de parfaitement bien apprendre la mécanique oeuvrante sous-jacente :

- Le biais : c'est l'erreur globale du modèle. Il s'agit de l'erreur quantifiant l'incapacité d'un algorithme à capter la relation entre les variables explicatives et la variable cible.
- La variance : elle correspond à l'erreur de sensibilité aux données utilisées.

Le biais et la variance d'un modèle sont deux vases communiquants : agir sur l'un en le diminuant entraînera systématiquement une augmentation de l'autre. C'est pour cela que l'on parle de compromis.

Usuellement, l'utilisation du graphique ci-dessous est employé pour décrire ce concept :

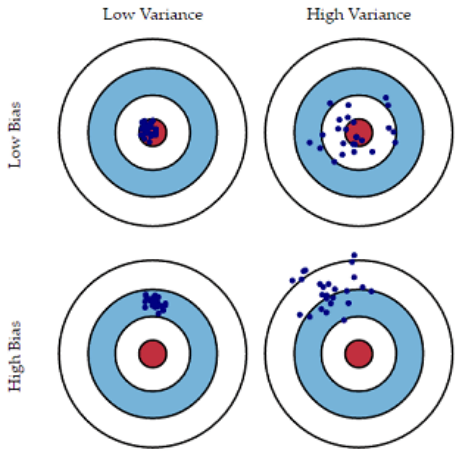


FIGURE 3.1 – Métaphore du concept de biais/variance en ML

Le sur-apprentissage

Lorsqu'un algorithme présente un biais très faible, alors mécaniquement de par le concept de vases communiquant décrit plus haut, la variance est très élevée. Autrement dit, l'erreur sur la base d'apprentissage va être très faible, contre quoi les erreurs vont être très élevées sur une nouvelle base n'ayant pas servi à l'apprentissage. Cela motive, dans un projet orienté ML, le découpage des bases de données en base d'apprentissage et base de validation, et plus généralement à l'utilisation de la méthode de validation croisée décrite plus bas.

Les métriques de performance

Le choix de la métrique de performance, parfois appelée score, va avoir un impact sur le modèle en apprentissage supervisé : les paramètres de l'algorithme sont choisis de manière à minimiser une certaine métrique.

En régression les métriques de performances usuelles sont :

- Erreur absolue moyenne : $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- Erreur de biais moyen : $MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$

- Erreur quadratique moyenne : $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Ces métriques d'évaluations doivent être choisies dès le début du projet de ML car elles sont étroitement liées à la problématique métier soulevée. Par exemple, nous pourrions être amenés à pénaliser les prédictions sur-évaluant les valeurs cibles avec :

$$E_{sur-mesure} = \frac{1}{n}\sum_{i=1}^{n}(2.|y_i - \hat{y}_i|.\mathbf{1}_{y_i<\hat{y}_i} + |y_i - \hat{y}_i|.\mathbf{1}_{y_i\geq\hat{y}_i})$$

ou bien partitionner par "gravité" les erreurs en régression. En classification $K = 2$, nous pourrions affecter un poids trois fois supérieur aux faux négatifs pour un algorithme détectant si oui ou non un e-mail est un spam dans le cadre d'un client particulièrement averse à la réception de spams, etc...

La validation croisée

L'objectif principal de la validation croisée, ou k-fold cross validation, est d'éviter le sur-apprentissage.

Cette méthode consiste à subdiviser la base de données en k blocs de tailles identique. Éventuellement, $k = n$ avec n le nombre lignes/individus. Sur chaque bloc de données est effectuée une prédiction grâce à un modèle entraîné sur les $k - 1$ autres blocs de données. De cette manière, une meilleure généralisation (diminution de la variance donc) est possible (au prix d'un biais un peu plus élevé). L'erreur du modèle est calculée sur chacun des k blocs de validation utilisés et l'erreur globale est alors calculée comme la moyenne de ces erreurs.

Schématiquement :

FIGURE 3.2 – Schéma de la méthode de validation croisée. Ici k = 5.

Évidemment, plus k sera élevé, plus la méthode sera coûteuse en temps.

3.1.3 Modèles de régression généralisé (GLM)

Le GLM généralise la bien connue régression linéaire en permettant au modèle linéaire d'être relié à la variable réponse via une fonction lien (link function en anglais). De cette manière, il est possible de transformer une prédiction dans $\mathbb{R}$ à une prédiction dans $\{0,1\}$ dans le cas de la régression logistique ou bien dans $\mathbb{R}^+$, ou même dans $\mathbb{N}$.

Les n individus décrits par p variables explicatives sont représentés par la matrice X en ligne. Soit β un vecteur de $p+1$ paramètres correspondant aux p variables explicatives et à un "intercept". C'est pour cela que l'on rajoute une colonne de 1 à la matrice X. On note alors le prédicteur linéaire $\eta = X\beta$.

Prenons le cas de la régression logistique :

Dans ce cas, $Y_i \in \{0,1\}$, on utilise la loi de Bernoulli et on suppose que $Y_i \sim Ber(\pi_i)$ avec $\pi_i = \pi(x_{i,1}, x_{i,2}, ..., x_{i,p})$, c'est à dire que $\mathbb{P}(Y_i = 1) = \pi$ et $\mathbb{E}(Y_i) = \pi$.

On suppose que $g(\mathbb{E}(Y_i)) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}$ avec g la fonction lien. Pour la régression logistique, on prend :

$$g(x) = logit(x) = \left(\frac{exp(x)}{1+exp(x)}\right)^{-1} = ln\left(\frac{x}{1-x}\right),$$

c'est à dire une bijection telle que $g :]\,0,1[\to \mathbb{R}$ et donc $g^{-1} : \mathbb{R} \to]\,0,1[$ (injection d'une prédiction dans $\mathbb{R}$ à une prédiction dans $]\,0,1[$).

On obtient donc le modèle de régression logistique avec :

$$\begin{cases} Y_i \sim Ber(\pi(x_{i,1}, x_{i,2}, ..., x_{i,p})) \\ ln(\frac{\pi(x_{i,1},x_{i,2},...,x_{i,p})}{1-\pi(x_{i,1},x_{i,2},...,x_{i,p})}) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p} \end{cases}$$

et donc $\pi(x_{i,1}, x_{i,2}, ..., x_{i,p}) = \frac{exp(\beta_0+\beta_1 x_{i,1}+...+\beta_p x_{i,p})}{1+exp(\beta_0+\beta_1 x_{i,1}+...+\beta_p x_{i,p})},\ i = 1, ..., n.$

Cette dernière prédiction de la variable cible pour l'individu i est donc à valeur dans $]\,0,1[$. En choisissant un seuil s entre $]\,0,1[$, nous sommes alors capables d'affecter pour chaque individu i une valeur de prédiction $y_i \in \{0,1\}$: si $\pi_i > s$ alors $y_i = 1$. Ce seuil s est en pratique choisi de manière à optimiser l'erreur au sens de la métrique de performance utilisée.

Pour une régression linéaire, $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}, \sigma^2)$ et $Y_i \in \mathbb{R}$, $g(x) = \mathcal{I}\lceil(x)$. Pour une régression de poisson, $Y_i \sim Poisson(\lambda(x_{i,1}, ..., x_{i,p}))$ et $Y_i \in \mathbb{N}$, $g(x) = ln(x)$. Pour une régression gamma, $Y_i \sim Gamma(\alpha, \beta(x_{i,1}, ..., x_{i,p}))$ et $Y_i \in \mathbb{R}^+$, $g(x) = \frac{1}{x}$ ou $g(x) = ln(x)$.

Pour estimer les paramètres inconnus θ (les β_i pour une régression linéaire, logistique ou poisson, α et β pour une régression gamma), la méthode de l'estimateur du maximum de vraisemblance est employé (EMV). Soit la fonction de vraisemblance définie par :

$L(\theta) = f(y_1, ..., y_n; \theta) \overset{Y_i \overset{iid}{=}}{=} \Pi_{i=1}^n f_{Yi}(\theta)$ où f_{Yi} est la fonction densité de la loi suivie par Y_i (donné par le modèle GLM choisi donc). Par exemple, dans le cas de la régression logistique :

$$\mathrm{f}_{Yi}(\theta) = \pi_\theta(x_{i,1}, ..., x_{i,p})\mathbf{1}_{\{y_i=1\}} + (1 - \pi_\theta(x_{i,1}, ..., x_{i,p}))\mathbf{1}_{\{y_i=0\}} = \pi_\theta(x_{i,1}, ..., x_{i,p})^{y_i}(1 - \pi_\theta(x_{i,1}, ..., x_{i,p}))^{1-y_i}$$

avec $\pi(x_{i,1}, x_{i,2}, ..., x_{i,p}) = \frac{exp(\beta_0+\beta_1 x_{i,1}+...+\beta_p x_{i,p})}{1+exp(\beta_0+\beta_1 x_{i,1}+...+\beta_p x_{i,p})}$ et $\theta = (\beta_0, ..., \beta_p)'$.

L'EMV est alors $\hat{\theta} = arg\max_\theta L(\theta)$ et est calculé avec des méthodes numériques de descente de gradient, ou de type Newton-Raphson.

3.1.4 Séparateurs à Vastes Marges

Les séparateurs à vastes marges (SVM, ou support vector machine en anglais) constituent une classe d'algorithmes d'apprentissage et découlent des travaux de V.Vapnik et C.Cortes (1995). Ils sont à l'origine, conçus pour la classification à $K = 2$, mais ont depuis été généralisés pour $K \geq 2$ et également pour la prédiction de variables quantitatives (on parle alors de SVR). Dans le cas de la classification à deux états, ils sont basés sur la recherche de l'équation de l'hyperplan à marge optimale séparant au mieux les deux classes d'individus. [18]

Dans le cas où les observations sont linéairement séparables :

Soit $B_a = \{(x^r, y^r) \in \mathbb{R}^n \times \{-1, +1\}, r = 1, ..., p\}$ une base d'apprentissage. La valeur $y_i^r \in \{-1, +1\}$ indique la classe à laquelle l'individu i appartient. L'espace $\mathbb{R}^n$ est muni du produit scalaire ".". Un hyperplan H de $\mathbb{R}^n$ est défini par l'équation $w.x - b = 0$ où w est un vecteur orthogonal à H. L'ensemble des éléments de B_a est dit linéairement séparable si il existe un vecteur $w \in \mathbb{R}^n$ et un scalaire $b \in \mathbb{R}$ tels que les inégalités ci-dessous soient vérifiées pour $r = 1, ..., p$:

$$\begin{cases} w.x^r - b \geq 1 & \text{si} \quad y^r = +1, \\ w.x^r - b \leq -1 & \text{si} \quad y^r = -1 \end{cases}$$

Ceci équivaut à $y^r(w.x^r - b) \geq 1, \ \forall r$. Dans ce cas, le signe de $f(x) = w.x - b$ indique de quel côté de l'hyperplan se situe le point x, et donc à quelle classe il appartient. Le choix de la constante 1 est arbitraire puisque un hyperplan est défini à une constante multiplicative près. Il y a donc une infinité de solutions : nous cherchons celle de la marge maximale, c'est à dire, l'hyperplan qui se trouve le plus loin possible de tous les exemples. Nous appelons la marge la distance entre les deux hyperplans d'équations $w.x - b = 1$ et $w.x - b = -1$.

La distance d'un point x à un hyperplan H d'équation $w.x - b = 0$ est :

$$d(x, H) = (x - w_1).w_0 = x.w_0 - w_1.w_0,$$

où w_1 est l'unique point de H tel que le vecteur w_1 soit colinéaire au vecteur w et $w_0 = \dfrac{w}{||w||}$.

Puisque $w_1 \in H$, alors $w_1.w_0 = \dfrac{b}{||w||}$. En reprenant l'équation d'avant il vient alors que :

$$d(x, H) = x.\frac{w}{||w||} - \frac{b}{||w||} = \frac{x.w - b}{||w||} = \frac{f(x)}{||w||}$$

.

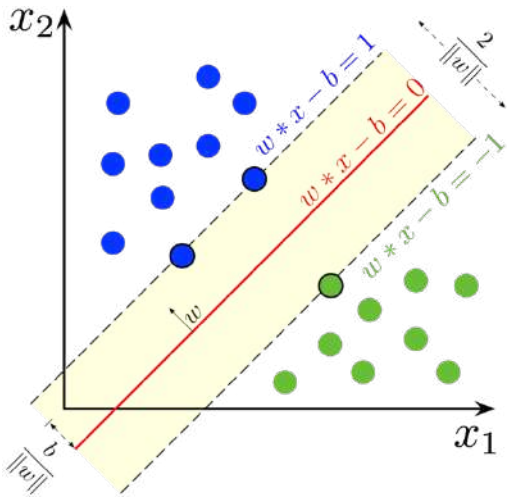


FIGURE 3.3 – Représentation de l'équation de l'hyperplan à marge optimale

La fonction f valant -1 et 1 sur les bords, nous en déduisons que la marge associée à l'hyperplan H vaut $\frac{2}{||w||}$. Pour obtenir la marge maximale, nous sommes donc amenés à minimiser la fonction $\Phi(w, b) = ||w||$ sous les contraintes $y^r(w.x^r + b) \geq 1, \forall r$.

En utilisant la méthode standard des multiplicateurs de Lagrange, nous pouvons montrer que l'équation de l'hyperplan optimal est donnée par :

$$\sum_{r=1}^{p} \alpha_r^* y^r x.x^r - b^* = 0 \; et \; b^* = \frac{1}{2}(x^*.x^+ + w^*.x^-),$$

où x^+ et x^- sont des vecteurs supports de chaque classe.

Pour une nouvelle observation x non apprise présentée au modèle, le signe de la fonction :

$$f(x) = \sum_{r=1}^{p} \alpha_r^* y^r x.x^r - b^r,$$

donne la classe à lui attribuer.

Dans le cas où les observations ne sont pas séparables :

En général, il n'est pas possible de trouver une section linéaire de l'espace satisfaisante. Il se peut aussi que l'hyperplan séparateur ne soit pas la meilleure solution au problème de classification. C'est pourquoi la technique de "marge souple" a été proposée en 1995 par C.Cortes et V.Vapnik : les contraintes sont assouplies par l'introduction de termes d'erreurs ξ^r qui en contrôlent le dépassement (cherche l'hyperplan minimisant les erreurs donc) : $y^r(w.x^r - b) \geq 1 - \xi^r, r = 1, ..., p$. Le modèle attribue donc ainsi une réponse fausse à un vecteur x^r si le ξ^r correspondant est supérieur à 1. Le problème d'optimisation est alors ramené à :

$$\Psi(w, b, \xi) = \frac{1}{2}||w||^2 + \delta \sum_{r=1}^{p} \xi^r,$$

sous les contraintes $y^r(w.x^r - b) \geq 1 - \xi^r$ et $\xi^r \geq 0$ pour $r = 1, ..., p$. La méthode de résolution dans ce cas n'est pas décrite ici, mais l'est par C.Cortes et V.Vapnik en 1995 dans leur article [49].

Séparateur non linéaire

Lorsque les observations ne sont pas linéairement séparables, il est parfois envisageable de trouver une frontière non linéaire. Il suffit alors de "transformer" l'espace de travail. En effet, en considérant l'image des observations par une application non linéaire Φ dans un espace H de dimension supérieure à celui d'origine $F = \mathbb{R}^n$. L'idée est donc de choisir Φ intelligemment afin de rendre les observations linéairement séparables comme le montre la figure ci-dessous :

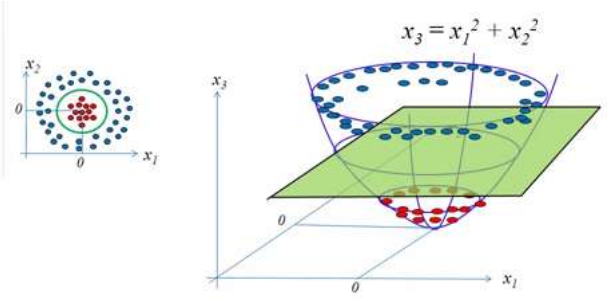


FIGURE 3.4 – Idée principale des séparateurs non linéaires : ici l'astuce du noyau. Les individus ne sont pas "séparables" en dimension 2, mais le sont en dimension 3 en appliquant une fonction (noyau) adéquate.

Il n'est pas nécessaire pour autant de connaître Φ puisque la méthode ne fait intervenir que des produits scalaires $y.y'$ pour des y de la forme $y = \Phi(x)$. Il est alors considéré une fonction noyau (kerneml en anglais) k :$F \times F \to \mathbb{R}$ symétrique appelée noyau telle que : $k(x, x') = \Phi(x).\Phi(x')$

La condition de Mercer assure qu'une telle fonction, symétrique, est un noyau si pour tous les x_i possibles, la matrice $(k(x_i, x_j)))_{i,j}$ est définie positive, c'est à dire qu'elle définit un produit scalaire. Dans ce cas, il existe un espace H et une fonction Φ tels que $k(x, x') = \Phi(x).\Phi(x')$

Les noyaux employés usuellement sont :

- le noyau polynomial : $k(x, x') = c.x.x'$, $c \in \mathbb{R}$
- le noyau gaussien/radial : $k(x, x') = exp(-\frac{||x-x'||^2}{2\sigma^2})$, $\sigma > 0$
- le noyau tangente hyperbolique : $k(x, x') = tanh(c_1.x.x' + c_2)$, $c_1, c_2 \in \mathbb{R}$

Les SVM présentent quelques avantages non négligeables comme leur capacité à travailler avec des données en grandes dimensions, le faible nombre d'hyperparamètres, la théorie développée qui est éprouvée, et leur bons résultats en pratique.

3.1.5 Arbre de décision CART

Un arbre de décision désigne une méthode d'apprentissage statistique supervisé basée sur l'utilisation d'un arbre de décision comme modèle prédictif.

L'idée est de partitionner l'espace des entrées en sous parties homogènes/variables à prédire, de manière itérative.

Les arbres de décisions CART ont étés introduits par L.Breiman en 1984 [28].

Voici les élément de compréhension principaux [18] :

- **Étape 0** :

  Au début, tous les individus appartiennent à la même classe.

- **Étape 1** :

  Partitionnement des individus en deux classes suivant $B_a = \{(X_i, Y_i), i = 1, ..., n\}$ avec $X_i = (X_i^1, ..., X_i^p) \in \mathbb{R}^p$ en opérant une section/coupure s de la forme $\{X_j \leq s\}$ et $\{X_j > s\}$ pour un $j \in \{1, .., p\}$ et la section $s \in \mathbb{R}$ (une section suivant l'axe d'une seule variable).

  L'objectif est donc de choisir la meilleure section possible.

  Pour se faire, pour une section s donnée notons les deux sous-ensembles $n_{i,-}(j, s) = \{i \in \{1, ..., n\} : X_i^j \leq s\}$ et $n_{i,+}(j, s) = \{i \in \{1, ..., n\} : X_i^j > s\}$. À ces deux sous-ensembles il est possible d'attribuer les valeurs $\bar{y_-}$ et $\bar{y_+}$ qui sont les moyennes respectives des ensembles $n_{i,-}(j, s)$ et $n_{i,+}(j, s)$.

  La section s retenue est alors celle qui minimise, dans le cas de la régression, la variance des deux classes obtenues :

$$C(j, s) = \sum_{i \in n_{i,-}(j,s)} (Y_i - \bar{y}_-)^2 + \sum_{i \in n_{i,+}(j,s)} (Y_i - \bar{y}_+)^2$$

- **Étape 2** :

  La procédure est réitérée, sur chaque sous arbres formés. Pour chaque feuille de l'arbre, on attribue la moyenne des y_i (régression) ou bien l'issue du vote majoritaire (classification). Elle s'arrête selon deux critères : soit les feuilles de l'arbre obtenu après sectionnement sont trop "pauvres" (les Y_i sont égaux pour tous les individus de la classe), soit le cardinal (le nombre d'individu "tombant" dans cette feuille) de la feuille (de la classe) est inférieur à un nombre m fixé en paramètre de l'algorithme.

- **Étape 3** :

  Étape d'élagage de l'arbre maximal obtenu possédant T_{max} feuilles afin de trouver le meilleur sous-arbre possible. Tout d'abord une suite de sous-arbres élagués les uns des autres est construite. Pour cela, on minimise un critère défini par tout sous-arbres T de l'arbre maximale T_{max} et tout $\alpha \geq 0$:

$$Critere_\alpha(T) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{T,i})^2 + \alpha.|T|,$$

où T est le nombre de feuilles de l'arbre T et $\hat{y}_{T,i}$ la prédiction de l'individu i faite par l'arbre T. En augmentant progressivement α, on obtient avec ce critère une suite de sous-arbres $(T_j)_{1\leq j\leq J}$ ayant de moins en moins de feuilles. Finalement, le meilleur sous-arbre de cette suite est choisi suivant deux critères proposés par L.Breiman : validation croisée et minimisation de l'erreur sur une base de test (base non utilisée pour la calibration du modèle, par définition).

Un arbre "profond" (beaucoup de sections) aura un biais faible, mais une variance élevée. Un arbre "peu profond" (peu de sections) présentera une faible variance mais un biais fort. Tout l'enjeu du paramétrage d'un arbre de décision se situe donc dans l'optimisation des hyper-paramètres en vue d'obtenir un bon compromis biais/variance.

Un arbre de décision CART présente quelques avantages :

- Facilité d'interprétation : les règles de décisions sont explicites et intelligibles
- Prise en compte de tout type de variables (discrète, continue, catégorielle)
- Ne nécessite de pré-traitement des données
- Performant sur les grands jeux de données
- Prend en compte des variables linéairement liées (contrairement au modèle linéaire)
- Fonctionne pour des relations non linéaires entre Y et les X^j

Mais aussi des inconvénients conséquents :

- Faible robustesse : l'arbre obtenu peut varier fortement si la base d'apprentissage est modifiée et est donc par conséquent sujet au sur-apprentissage. Il sera plutôt utilisé comme brique élémentaire d'algorithme de "bagging" comme les forêts aléatoires ou les méthodes de boosting.
- Dans le cas où de nombreuses variables explicatives sont présentes : l'arbre de décision devient alors difficile à interpréter

3.1.6 Random Forest (RF)

Les forêts d'arbres décisionnels (random forest en anglais) ont été formellement proposées en 2001 par L.Breiman et A.Cutler [29], bien que l'idée à l'origine soit de TK.Ho en 1995.

Cet algorithme appartient à la famille des algorithmes de "bagging" : algorithmes ensemblistes, qui agrègent des modèles de bases. Dans le cas de la forêt aléatoire, les modèles de bases sont des arbres de décisions CART et sont construit indépendamment les uns des autres, contrairement aux algorithmes de boosting.

Le bagging est pertinent si les modèles de bases sont des modèles à forte variance (comme les arbres de décisions dans le cas de la RF). Pour mettre en évidence cela : soit B_n un échantillon d'apprentissage de n observations : $B_n = \{(x_i, y_i), i = 1, ..., n\}$. Dans

le cas d'une régression $y_i \in \mathbb{R}$, dans le cas d'une classification $y_i \in \{C_1, ..., C_k\}$. Pour construire les B modèles de bases, on tire au hasard B échantillons : $B_{n,1}, ..., B_{n,B}$. Sur chaque échantillon de base $B_{n,i}$, le calcul d'un modèle de base, un arbre de décision, $\hat{\Phi}_i$ est effectué. L'algorithme de bagging, comme une random forest, est donné alors par les les B prédictions $\hat{\Phi}_1, \hat{\Phi}_2, ..., \hat{\Phi}_B$ par :

- $\hat{\Phi}(x) = \frac{1}{B} \sum_{i=1}^{B} \hat{\Phi}_i(x)$, dans le cas d'une régression (moyenne des modèles de bases)
- $\hat{\Phi}(x) = argmax_{j=1,...,K}\ card\{i = 1, ..., B : \hat{\Phi}_i(x) = c_j\}$, dans le d'une classification à K classes (vote majoritaire des modèles de base)

Voici un pseudo-code possible pour les RF [18] :

Algorithm 1: Forêt aléatoire

initialisation : choisir le nombre d'arbre B et le nombre de variables m (usuellement, $sqrt(p)$ en classification , $\frac{p}{3}$ en régression) ;

for $i \leftarrow 1$ **to** B **do**

- tirer un échantillon bootstrap $B_{n,i}$ dans B_n.
- Avec celui-ci, calibrer un arbre de décision CART $\hat{\Phi}_i$. Chaque arbre est déterminé en se restreignant à m variables possibles parmi les p. La performance de cet arbre est mesurée par l'erreur "out of the bag", c'est à dire par l'erreur calculée sur les observations n'ayant pas participé à l'apprentissage de cet arbre.

end

Agregation : On pose $\hat{\Phi}(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{\Phi}_i(x)$ (moyenne, en régression) ou $argmax_{j=1,...,K}\ card\{i = 1, ..., B : \hat{\Phi}_i(x) = c_j\}$ (vote majoritaire, en classification).

De manière simplifiée, supposons que chaque prédicteur de base a une variance σ^2 et que deux prédicteurs de base aient une corrélation de ρ. Alors la variance obtenue par bagging est égale à :

$$\rho\sigma^2 + \frac{(1-\rho)\sigma^2}{B}$$

Si B est grand, alors la variance de la RF diminue, et devient plus faible que chaque arbre pris individuellement. Cette méthode n'a d'intérêt que si les modèles de bases sont peu corrélés, c'est à dire qu'ils sont très sensibles à la base d'apprentissage ($|\rho|$ petit).

Les modèles de bagging comme les RF sont complexes à interpréter : le terme de boîte noire est souvent employé. Elles sont également plus lentes à entraîner. Cependant, pour une RF, chaque variable peut être mesurée par ordre d'importance, en mesurant la variation de l'erreur en retirant du modèle chaque variable : l'idée est de se dire que si en retirant une variable l'erreur n'évolue pas, alors cette variable n'est pas importante. Les RF présentent cependant de nombreux avantages : pas de sur-apprentissage possible par construction même de l'algorithme de la RF, pas besoin de validation croisée grâce aux échantillons "out of the bag" qui remplissent déjà ce rôle, et hyper-paramètres faciles à calibrer (nombre d'arbres, nombre d'observations minimum dans une feuille, la profondeur maximale de chaque arbre, etc...).

3.1.7 XGBoost

Les méthodes dites de "Boosting" reposent sur une agrégation d'arbres. Comme énoncé plus haut, les RF vont s'attacher à construire des arbres CART indépendamment les uns des autres (en "parallèle", ce qui donne lieu par exemple à la notion de prédiction en mode "vote majoritaire" par exemple), là où les méthodes de Boosting construisent des classifieurs dits "faibles" en série (ils possèdent donc une "mémoire"). Itérativement, un arbre CART comme classifieur faible, dans le cadre d'une méthode de boosting, construit à l'itération i+1 favorisera son apprentissage vers les individus mal prédits de l'arbre i, qui lui même aura précédemment appris de l'arbre i-1, etc...

Parmi ces méthodes de Boosting ceux-ci sont notables : *Adaboost* (Y.Freund et R.Shapire en 1997 [51] et sa variation en régression de H.Drucker en 1997 [21]), le *Gradient Boosting* (JH.Friedman,1999 [25]) et le *XGboost* (pour eXtreme Gradient Boosting, T.Chen en 2016 [45])

Le Xgboost a été utilisé dans ce mémoire. Cet algorithme est considéré comme la star des algorithmes de machine learning ; il remporte régulièrement des compétitions depuis sa mise au point. Il est de plus en plus utilisé en actuariat comme le montre les mémoires de R.Gauville [39] et D.Delcaillau [13]

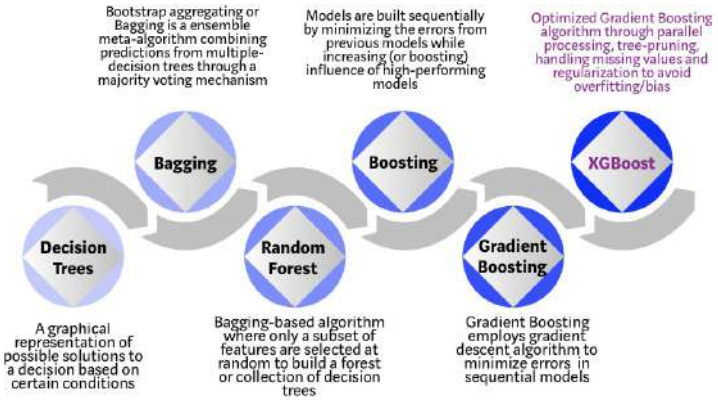


FIGURE 3.5 – Chronologie : de l'arbre CART à l'XGboost

Néanmoins, afin de comprendre le XGboost, il faut comprendre le Gradient Boosting.

Le Gradient Boosting

Nous allons présenter le principe de la construction d'un modèle de Gradient Boosting. Cette méthode combine une séquence de modèles de base combinés à une descente de gradient. Ici nous allons nous intéresser à une séquence d'arbres, dans le cas d'une régression ($Y \in \mathbb{R}^n$).

On note $B_a := \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R},\ i = 1, ..., n\}$ et $x_i = (x_{i,1}, ..., x_{i,n})' \in \mathbb{R}^p$ les observations des p variables explicatives de l'individu i. y_i est la variable à prédire de l'individu i.

On cherche une fonction f (un modèle) telle que $f(x_i) \approx y_i$. Pour ce faire on dispose d'une fonction l de perte/d'erreur (loss function en anglais). On peut prendre

$l(y_i, f(x_i)) = (y_i - f(x_i))^2$ la fonction d'erreur quadratique par exemple, qui est une fonction convexe. L'erreur globale est alors $L(y, f) = \sum_{i=1}^{n} l(y_i, f(x_i))$ et l'objectif est donc de la minimiser sur B_a.

Une méthode employée pour minimiser une fonction convexe est celle de la descente de gradient. En voici le principe : soit $g : \mathbb{R}^p \Leftrightarrow \mathbb{R}$ et telle que g est deux fois différentiable, alors le gradient de g est défini par :

$$\nabla g = (\frac{\partial g}{\partial z_1}, ..., \frac{\partial g}{\partial z_p})$$

. Pour un $\epsilon > 0$, tant que $||\nabla g(x_k)|| > \epsilon$: on calcul $x_{k+1} = x_k - \alpha_k \nabla g(x_k)$. α_k peut-être constante ou bien peut varier en fonction de k.

On applique la méthode de descente de gradient à la fonction d'erreur globale $L(y, f)$:

- 1. Initialiser $f_0(x) = arg\min_\theta \sum_{i=1}^{n} l(y_i, \theta) = arg\min_\theta \sum_{i=1}^{n} (y_i - \theta)^2 = arg\min_\theta L(y, \theta)$ avec $\theta = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$
- 2. Pour $m = 1, ..., M$ (M est le nombre de modèles en séquence, le nombre de "classifieurs faibles") :
  - Calcul des pseudo-résidus : $\sigma_{i,m} = - \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\Big|_{f_{x_i} = f_{m-1}(x_i)}$, $i = 1, ..., n$
  - On souhaite construire un modèle h_m capable de calculer les pseudo-résidus pour n'importe quel x. On calibre donc un classifieur $h_m(x)$ (un arbre CART par exemple) sur les données $(x_i, r_{i,m})$, $i = 1, ..., n$
  - Le classifieur à l'étape m est alors : $f_m(x) = f_{m-1}(x) + \theta_m h_m(x)$. θ_m peut être choisi en en prenant : $\theta_m = arg\min_\theta \sum_{i=1}^{n} l(y_i, f_{m-1}(x_i) + \theta h_m(x_i))$
- Le modèle final est $f_M(x) = \theta_M h_M(x) + \theta_{M-1} h_{M-1}(x) + ...$

Le XGboost

Le XGboost est une version extrême du Gradient Boosting. Celui-ci permet au Gradient Boosting d'ajouter (ou pas) d'autres modèles aux modèles de bases, comme des régressions linéaires par exemple dans le cas d'un XGboost en régression. Le XGBoost construit des modèles de bases h_m en série tout en contrôlant leur complexité.

À l'étape m, on cherche h_m minimisant :

$$\sum_{i=1}^{n} l(y_i, f_{m-1}(x_i) + h_m(x_i)) + \sum_{k=1}^{m} \Omega(h_k)$$

Ici on a choisi $\theta_m = 1$ pour simplifier, et on remarque que le terme $\sum_{k=1}^{m} \Omega(h_k)$ est destiné à prendre en compte la complexité des m modèles dans la minimisation. Le terme peut être vu comme : $\sum_{k=1}^{m} \Omega(h_k) = \Omega(h_m) + cste$ puisque la complexité du modèle h_m n'a rien à voir avec la complexité des autres classifieurs. Dans le cas où $l(x, y) = (x - y)^2$ (fonction d'erreur quadratique), on cherche h_m minimisant :

$$\sum_{i=1}^{n} (y_i - (f_{m-1}(x_i) + h_m(x_i)))^2 + \Omega(h_m)$$

Cela revient à minimiser :

$$\sum_{i=1}^{n} 2(y_i - f_{m-1}(x_i))h_m(x_i) + h_m(x_i)^2 + \Omega(h_m)$$

En effet, $\sum_{i=1}^{n} y_i^2$ ne dépend pas de h_m. Dans le cas général où l est deux fois dérivable, on h_m en minimisant :

$$\sum_{i=1}^{n} l(y_i, f_{m-1}(x_i)) + \alpha_i^m h_m(x_i) + \frac{1}{2}\beta_i^m h_m(x_i)^2 + \Omega(h_m)$$

où $\alpha_i^m = \frac{\partial}{\partial y} l(y_i, y)\Big|_{y=h_{m-1}(x_i)}$ et $\beta_i^m = \frac{\partial^2}{\partial^2 y} l(y_i, y)\Big|_{y=h_{m-1}(x_i)}$

Dans le cas du XGboost, si on considère des h_m comme étant des arbres CART : soit T le nombre de feuille de l'arbre et $(W_t)_{t=1,...,T}$ le "score" prédit par la feuille T. On a $h_m(x) = W_s(x)$ où $s : \mathbb{R}^p \to \{1, ..., T\}$ est une section. s est la feuille dans laquelle se trouve l'individu i. Usuellement, la fonction de complexité $\Omega(h_m)$ s'écrit :

$$\Omega(h_m) = \partial T + \alpha \sum_{t=1}^{T} |W_t| + \frac{\lambda}{2} \sum_{i=1}^{T} W_t^2$$

où γ, α et λ sont les paramètres de régularisation de complexité du modèle à choisir. On peut choisir pour simplifier que $\alpha = 0$.

Finalement, on cherche h_m minimisant :

$$\sum_{r=1}^{n} \alpha_i(m)h_m(x_i) + \frac{1}{2}\beta_i^m h_m(x_i)^2 + \gamma T + \frac{\lambda}{2} \sum_{t=1}^{T} W_t^2$$

qui est égale à :

$$\sum_{t=1}^{T} (\sum_{i \in I_t} \alpha_i^{(m)}) W_t + \frac{1}{2}(\sum_{i \in I_t} \beta_i^{(m)} + \lambda W_t^2) + \gamma T$$

où $I_t = \{i \in \{1, ..., n\}, s(x_i) = t\}$ De plus, en notant $A_t = \sum_{i \in I_t} \alpha_i^{(m)}$ et $B_t = \sum_{i \in I_t} \beta_i^{(m)}$ on obtient :

$$\sum_{t=1}^{T} (A_t W_t + \frac{1}{2}(B_t + \lambda) W_t^2)$$

Pour une fonction $s : \mathbb{R}^p \to \{1, ..., T\}$ donnée (structure de l'arbre), on peut trouver les (W_t) minimisant la fonction d'erreur précédente avec $W_t = -\frac{A_t}{B_t + \lambda}$, $\forall t \in 1, ..., T$. Alors la fonction objectif à minimiser devient :

$$-\frac{1}{2} \sum_{j=1}^{T} \frac{A_j^2}{B_j + \lambda} + \gamma T$$

qui mesure la précision et la complexité du modèle.

Cependant, il est impossible de tester toutes les sections s possibles en pratique. Une manière de procéder est alors de regarder "étage par étage" : on ajoute de nouvelles feuilles (à gauche ou à droite)à l'arbre CART et on regarde si le gain G est positif :

$$G = \frac{1}{2}(\frac{A_g^2}{B_g + \lambda} + \frac{A_d^2}{B_d + \lambda} - \frac{(A_g + A_d)^2}{B_g + B_d + \lambda}) - \gamma$$

avec $\frac{A_g^2}{B_g+\lambda}$ et $\frac{A_d^2}{B_d+\lambda}$ le score sur la partie gauche et la partie droite de la section, $\frac{(A_g+A_d)^2}{B_g+B_d+\lambda}$ le score sur la feuille avant section, et γ la complexité ajoutée lors du rajout de feuilles.

Au vu des éléments précédents, le XGboost présente de nombreux paramètres à optimiser lors de son entraînement sur la base d'apprentissage :

- le nombre d'arbre
- le taux d'apprentissage (learning rate en anglais) θ_m
- γ : paramètre de baisse minimale de la fonction objectif pour déclencher une subdivision supplémentaire
- la profondeur maximale d'un arbre
- le nombre minimal d'individus par feuille
- λ et α : paramètres de régularisation de la complexité du modèle
- les paramètres optionnels similaires à ceux d'une RF : ratio d'échantillonage des individus, le ratio de variables explicatives par arbres.

3.1.8 Interprétabilité des modèles "boîtes noires" : la méthode SHAP

Cette partie est issue du mémoire de D.Delcaillau [13] intitulé *Contrôle et Transparence des modèles complexes en actuariat* qui vise à faire un état de l'art des différentes techniques d'interprétabilité de modèles complexes de machine learning. Dans ce mémoire nous n'utiliserons que la méthode de SHAP (SHapley Additive exPlanations), bien qu'il existe également les méthodes PDP (partial dependance plot), ALE (Accumulated Local Effects) et LIME (Local Interpretable Model-agnostic Explanations). La théorie mathématique associée à cette méthode n'est pas détaillée ici car dépasse le cadre de ce mémoire.

Nous choisissons cette méthode pour intertpréter nos modèles "boites noires" car elle présente l'avantage d'être facile d'interprétation et de fournir une analyse des variables explicatives. Elle présente cependant des temps de calculs importants.

La méthode SHAP consiste à calculer la contribution de chaque variable explicative dans la prédiction faite par un modèle complexe comme un XGboost ou une forêt aléatoire. Il s'agit de l'unique méthode reposant sur une théorie mathématique : les valeurs de Shapley issues de la théorie des jeux. Elle serait donc une bonne candidate à l'explication d'un modèle complexe dans un cadre RGPD exigeant une transparence des modèles.

Principe :

Dans un modèle, chaque variable explicative a sa contribution à la prédiction finale pour une observation. Cependant, Il est possible qu'un léger changement dans une variable explicative modifie considérablement la prédiction, alors que pour d'autres variables non.

La méthode SHAP mesure cet impact via les valeurs de Shapley en tenant compte de l'interaction entre variables explicatives.

Les valeurs de Shapley indiquent l'importance d'une variable en comparant ce qu'un modèle prédit avec et sans cette variable. Cependant, étant donné que l'ordre dans lequel un modèle "voit" les variables peut affecter ses prédictions, cette comparaison est effectuée dans tous les ordres possibles, afin que les variables soient comparées équitablement.

3.2 Apprentissage statistique non-supervisé

L'apprentissage non-supervisé consiste à extraire d'un jeu de données des groupes d'individus présentant des caractéristiques communes. L'algorithme ne "tend" pas à minimiser une fonction d'erreur à partir d'un objectif cible, il vise plutôt à découvrir et mettre en évidence des motifs sous-jacents : l'apprentissage par l'algorithme se fait de manière indépendante. Il permet, entre autres, le partitionnement des données ou bien la réduction de dimensions. Ainsi, nous sommes en présence d'une matrice X de p variables (colonnes) et n individus (lignes), sans présence de variables à expliquer Y.

3.2.1 ACP : Analyse en composantes principales

L'ACP est une méthode d'apprentissage non supervisée de réduction de dimensions : elle permet de synthétiser les variables quantitatives décrivant des individus en axes appelés composantes principales. Ces axes sont une interprétation de une ou plusieurs variables (réduction de dimension).

soit X la matrice des observations (n individus décrits sur p variables) :

$$R = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,p} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{i,1} & \cdots & r_{i,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & r_{n,p} \end{pmatrix} = (r_{i,j})\ i = 1, 2, ..., n\ et\ j = 1, 2, ..., p$$

L'individu i est identifié par un vecteur ligne $e_i' = (r_{i1}, r_{i2}, ..., r_{ip})$ et la variable j par le vecteur colonne $r_j^{'} = (r_{1j}, r_{2j}, ..., r_{nj})$.

Le centre de gravité de l'ensemble des observations est alors : $g = R\mathbf{1}_n = (\bar{r}_1, \bar{r}_2, ..., \bar{r}_p)'$ avec $\bar{r}_j = \sum_{i=1}^{n} r_{ij}$

Afin d'éliminer l'effet de certaines variables, les données sont centrées. La matrice des données centrées est alors :

$$X = R - \mathbf{1}_n g' = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} = (x_{i,j})\ i = 1, 2, ..., n\ et\ j = 1, 2, ..., p$$

avec $x_{i,j} = r_{ij} - \bar{r}_j$

Le regroupement des individus se base sur l'analyse des distances. En cas de choix d'une distance euclidienne, la distance au carré entre deux individus i_1 et i_2 est déterminée par : $d(i_1, i_2) = \sum_{j=1}^{p} (x_{i_1 j} - x_{i_2 j})$

Dans un espace à deux dimensions, l'ensemble des individus i=1,...,n auxquels correspond le couple d'observations (x_1(i),x_2(i))=(x_{i1},x_{i2}), forme un nuage de points dans $\mathbb{R}$. Nous cherchons à approximer ce nuage par une droite D de vecteur directeur $u' = (\alpha_1, \alpha_2)$ qui pourra être déterminé en minimisant la somme des carrés des distances mesurées perpendiculairement entre les points observés et la droite D. En faisant une projection orthogonale de point ci en Pi (projection de l'espace des individus), l'idée est donc de minimiser la somme des longueurs de ces projections, c'est-à-dire : $min \sum_{i=1}^{n} c_i P_i$

Ce problème de minimisation revient à une équation à deux inconnues : $X'Xu = \lambda u \Leftrightarrow (V - \lambda I)u = 0$ *avec* $V = X'X$ (matrice de variance/covariance) Pour trouver une solution il faut donc que $(V - \lambda I)$ soit singulière, et que donc son déterminant soit nul. Cela correspond aux valeurs propres de V (et donc le vecteur u correspond au vecteur propre associé). Plus généralement, pour p variables la poursuite de la procédure p fois nous amène à constituer une nouvelle base orthonormée de $\mathbb{R}_p$: la base $u_1, u_2, ..., u_p$ formée de p vecteurs propres de la matrice $X'X$ associés aux valeurs propres $\lambda_1, \lambda_2, ..., \lambda_p$. Nous pouvons écrire : $X'Xu_p = \lambda_p u_p$ Avec : u_p vecteur propre associé à la valeur propre λ_p. Ainsi, avoir les vecteurs propres ainsi que les valeurs propres de V, c'est pouvoir résoudre le problème d'optimisation, et donc c'est trouver la droite D.

Ici, nous effectuons un changement de repère dans $\mathbb{R}_p$ de façon à se placer dans un nouveau système de représentation où le premier axe apporte le plus possible de l'inertie totale des nuages, le deuxième axe le plus possible de l'inertie non prise en compte par le premier axe, et ainsi de suite. Cette technique s'appuie sur la diagonalisation de la matrice de variances-covariances ou de corrélation.

Les valeurs propres représentent les inerties ou variances expliquées par les axes. La valeur λ_1 représente l'inertie expliquée par le premier axe, la valeur λ_2 celle expliquée par le second axe, ainsi de suite. La somme des valeurs propres est égale à la variance totale. Lorsque les données sont centrées et réduites, l'inertie totale est égale au nombre p des variables. : $\lambda = \sum_{k=1}^{n} \lambda_k = p$

Les facteurs, u_k pour $k = 1, 2, ..., p$, ou axes principaux, sont les vecteurs propres associés à la matrice de covariance. Les composantes principales, $c_k \in \mathbb{R}^n$, sont constituées des facteurs, par la relation : $c_k = Xu_k$

La composante principale c_k est le vecteur contenant les coordonnées des projections orthogonales des individus sur l'axe défini par u_k, $\mathbb{V}\text{ar}(c_k) = \lambda_k$ et les composantes principales sont non corrélées entre elles.

A partir des facteurs principaux et des composantes principales, il est possible de reconstituer le tableau initial avec la relation : $c_k = Xu'_k$ De la même manière que pour l'espace des individus, nous pouvons projeter une variable j dont les coordonnées sont X_j (projection de l'espace des variables) qui est un élément de $\mathbb{R}^n$ et représente la colonne j de la matrice des données X . La projection de la variable j vérifie : $v'X^j = \sum_{i=1}^{n} z_i x_{ij}$ avec $v' = [z_1, z_2, ..., z_n]$. Comme minimiser la somme des carrés de projections revient à maximiser : $\max \sum_{j=1}^{p} v'X_j v'X_j = v'XX'v$ cela revient à résoudre : $X'Xv_k = \mu_k v_k$ avec v_k le vecteur propre associé à la valeur propre μ_k . Nous constituons une nouvelle base orthonormée de $\mathbb{R}^n$: la base $v_1, v_2, ..., v_n$ formée de n vecteurs propres de la matrice de covariance $V = X'X$ associés aux valeurs propres $\mu_1, \mu_2, ..., \mu_n$.

Les espaces des individus et des variables sont donc reliés par les relations : Dans $\mathbb{R}^p$ par $u_k = \frac{X' v_k}{\sqrt{\lambda_k}}$ et dans $\mathbb{R}^p$ par $v_k = \frac{X' u_k}{\sqrt{\lambda_k}}$

Dans une étude ACP, nous cherchons à répondre à des questions de ce type : quels sont les individus qui se ressemblent ? Quels sont les individus qui sont différents ? Plus généralement, nous souhaitons décrire la variabilité des individus. Pour cela, nous cherchons à mettre en évidence des groupes homogènes d'individus dans le cadre d'une typologie des individus. Nous cherchons également une typologie des variables. Quelles sont les variables qui sont positivement corrélées ou qui s'opposent ?

Le choix du nombre d'axes principaux ou de dimension dépend de la part de l'inertie expliquée par la réduction de dimension. Si l'on choisit q axes principaux alors la part d'inertie expliquée par les q premiers axes est : $r = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$. Cet indice détermine la qualité globale de la réduction de dimension par les q premiers axes principaux.

La contribution relative de la j^{eme} variable à la variance de l'axe k est : $contrib_{jk} = u_{kj}^2$ Les variables peuvent être classées par ordre d'importance de leur contribution à la construction de l'axe. Les variables sont classées par contribution décroissante pour chaque axe. Le calcul de contribution est également possible au niveau des individus.

Le cercle des corrélations représente le lien entre les composantes principales. Ce cercle est constitué par les coefficients de corrélation. Les variables proches des extrémités du cercle sont très corrélées avec les axes et celles éloignées sont moins corrélées. Les coefficients de corrélation sont déterminés par la relation suivante : $cor(c_k, x_j) = (\lambda_k)^{\frac{1}{2}} u_{kj}$

Le cos^2 d'une variable représente le cosinus carré de l'angle formé par le point et l'axe (en %) : $cos^2 = \lambda_k u_{kj}^2$. Il permet de représenter la bonne représentation de la variable sur les différents axes. La somme sur l'ensemble des axes est égale à 100%. Pour un axe donné et une variable donnée, plus celle-ci possède un cos^2 proche de 1, plus cette variable est bien représentée sur cet axe. Ce calcul de cos^2 est également possible au niveau des individus.

3.2.2 K-means

Le partitionnement en k-moyennes (ou k-means en anglais) est un algorithme d'apprentissage non-supervisé. Il s'agit d'une méthode de partitionnement de données et un problème d'optimisation combinatoire. L'idée originale a été proposée par Hugo Steinhaus en 1957 [43].

Étant donnés des points et un entier k, le problème est de diviser les points en k groupes, appelés clusters, de façon à minimiser une certaine fonction de distance.

La méthode des k-means, qui est un algorithme, est donc caractérisée par le nombre de clusters souhaités k, la mesure de distance d'un point (individu) à la moyenne des points du cluster d'appartenance, et de la somme des carrés de ces distances. Il existe des mesures de distances différentes, et des variantes plus complexes englobant cet algorithme (nuées dynamiques par exemple).

Soit $X = (x_1, x_2, ..., x_n)$, avec $x_j = (x_{l1}, x_{l2}, ..., x_{lp})'$, le l^{eme} individu décrit par p variables. On cherche à partitionner les n points en k ensembles $S = \{C_1, C_2, \ldots, C_k\}$, $(k \leq n)$, en minimisant la distance entre les points à l'intérieur de chaque partition, avec

k fixé :

$$arg\min_{S} \sum_{i=1}^{k} \sum_{x_l \in S_j} \|x_l - g_i\|^2$$

où g_i est le centre de gravité des points dans C_i.

L'algorithme utilise la notion d'inertie intra-classe, d'inertie inter-classe, et d'inertie globale.

Supposons que chaque individu x_l avec $l = 1, ..., n$, possède un poids/une importance p_l tel que $\sum_{l=1}^{n} p_l = 1$. Alors l'inertie totale du nuage des n individus est

$$\mathcal{I}_g = \sum_{l=1}^{n} p_l.d(g, x_l)^2$$

où $g = \sum_{l=1}^{n} p_l$.

Il est également possible d'attribuer à chaque classe $C_j,\ j = 1, ..., k$ une inertie de classe $\mathcal{I}_{C_j}$ définie par :

$$\mathcal{I}_{C_j} = \sum_{x_l \in C_j} = \frac{p_l}{m_j} d(g_j, x_l)^2$$

où $m_j = \sum_{x_l \in C_j} p_l$ et $g_j = \sum_{x_l \in C_j} \frac{p_l}{m_j}$ est le centre de gravité de C_j.

Il est alors possible de définir l'inertie intraclasse $\mathcal{I}_{intra}$ et interclasse par :

$$\mathcal{I}_{intra}(S) = \sum_{i=1}^{k} m_i \mathcal{I}_{C_i},\ et\ \mathcal{I}_{inter}(S) = \sum_{i=1}^{k} m_i d(g_i, g)^2$$

Une faible valeur de l'inertie intraclasse indique une homogénéité des classes, et une grande valeur d'inertie interclasse indique une bonne séparation des classes. Il est facile de se rendre compte que l'inertie intraclasse et interclasse sont deux vases communiquants et que la somme des deux est égale à l'inertie globale. Par conséquent, maximiser l'inertie interclasse c'est minimiser l'inertie intraclasse.

L'algorithme classique de cette méthode est très utilisé en pratique et est considéré comme efficace bien que cette méthode ne garantisse pas l'optimalité, ni un temps de calcul polynomial.

En voici un pseudo code possible (point de vue minimisation de l'inertie intraclasse) :

Algorithm 2: K-means

initialisation : Choisir une partition initiale $S_1 = \{C_1^{(1)}, ..., C_k^{(1)}\}$ avec k fixé, une distance d, un maximum d'itération max_iter, et une valeur ϵ petite contrôlant la convergence de l'algorithme ;
Déterminer les centres de gravité $g_i^{(1)}$ des classes $C_i^{(1)}$, *pour* $i = 1, ..., k$;
$j = 1$;
while *L'inertie intraclasse diminue plus que* ϵ *ou que* $j < max\_iter$ **do**
- Effectuer une nouvelle partition $S_{j+1} = \{C_1^{(j+1)}, ..., C_k^{(j+1)}\}$ en regroupant les individus les plus proches autour des points $g_i^{(j)}$: x est affecté à $C_i^{(j+1)}$ si il est plus proche, au sens de la distance choisie, du centre de gravité $g_i^{(j)}$ de cette classe parmi toutes les autres. ;
- mettre à jour le centre de gravité de chaque cluster $C_i^{(j+1)}$, $i = 1, ..., k$;
- $j = j + 1$,

end

Visuellement et itérativement :

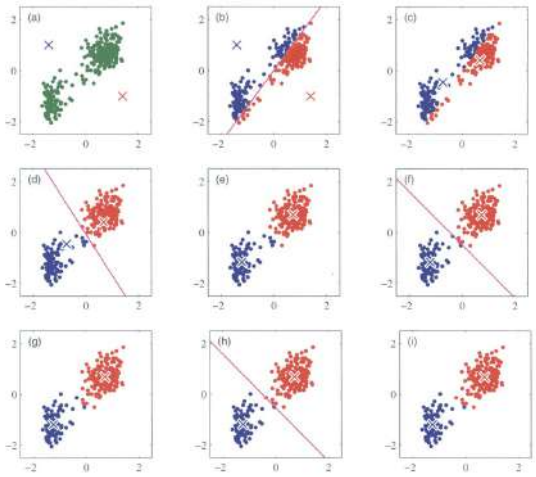


FIGURE 3.6 – De gauche à droite, de haut en bas, un exemple de représentation itérative de l'algorithme des k-means avec $k = 2$ et $p = 2$. [27]

L'initialisation et la mesure de distance sont des facteurs déterminants dans la qualité des résultats : un minimum local est garanti, mais pas le minimum global. Ce sujet fait donc l'objet de nombreuses recherches. Le nombre k de clusters voulu peut-être vu comme un avantage ou un inconvénient selon la nature du problème ou bien selon le degré de "non-supervisation" désiré.

3.2.3 Classification ascendante hiérarchique (CAH)

La CAH est une procédure d'apprentissage non supervisée. Il s'agit d'une méthode de classification ascendante : elle part d'une situation où tous les individus sont seuls dans une classe, puis sont rassemblés en classes de plus en plus grandes.

La CAH est caractérisée par la mesure de dissimilarité employée pour quantifier la dissimilitude entre individus : dans un espace euclidien, la distance euclidienne peut-être utilisée par exemple.

On note Ω l'ensemble des n individus à classifier, ω un individu et H une hiérarchie. L'algorithme de la CAH vérifie les propriétés suivantes :

- $\Omega \in H$: au sommet de la hiérarchie, lorsqu'on groupe de manière à obtenir une seule classe, tous les individus sont regroupés
- $\forall \omega \in \Omega : \{\omega\} \in H$: en bas de la hiérarchie, tous les individus se trouvent seuls
- $\forall (h, h') \in H^2,\ h \cap h' = O \ ou\ h \subset h' \ ou\ h' \subset h$: si l'on considère deux classes du regroupement, alors soit elles n'ont pas d'individu en commun, soit l'une est incluse dans l'autre.

Initialement, chaque individu forme une classe. Le but est donc de réduire ce nombre de classes. Cela se fait itérativement : à chaque itération, deux classes sont fusionnées. Les deux classes choisies pour être fusionnées sont celles présentant une dissimilarité entre elles minimales. Cette procédure est réalisée jusqu'à avoir une classe comportant l'ensemble des individus.

À une itération i donnée, une classe est composée de plusieurs individus. Afin de mesurer la dissimilarité entre deux classes, il existe plusieurs approches : calculer le minimum des distances entre les individus des deux classes (le "saut minimum"), calculer la dissimilarité entre les individus des deux classes les plus éloignées (le "saut maximum"), calculer la moyenne des distances entre les individus des deux classes (le "lien moyen"), ou bien maximiser l'inertie inter-classe ("distance de Ward"). C'est cette dernière approche qui sera retenue dans ce mémoire et est définie par : $dissim(C_1, C_2) = \frac{n1*n2}{n_1+n_2} * dissim(G_1, G_2)$, avec n_i effectif de la classe C_i et G_i centre de gravité de la classe C_i

En voici un pseudo code possible :

Algorithm 3: CAH

initialisation : attribuer à chaque individus sa propre classe ;
while *le nombre de classe > 1* **do**
- Calculer les dissimilarités entre classes ;
- Recherche des dissimilarités minimum deux à deux ;
- Fusionner les classes minimisant leur dissimilarité, et conserver les informations des classes précédentes ;
- Recalculer le nombre de classes ;

end

Visuellement, une représentation graphique de la CAH est possible via son dendrogramme. Il s'agit de la représentation d'une classification ascendante hiérarchique. Il se

présente comme un arbre binaire dont les feuilles sont les individus alignés sur l'axe des abscisses :

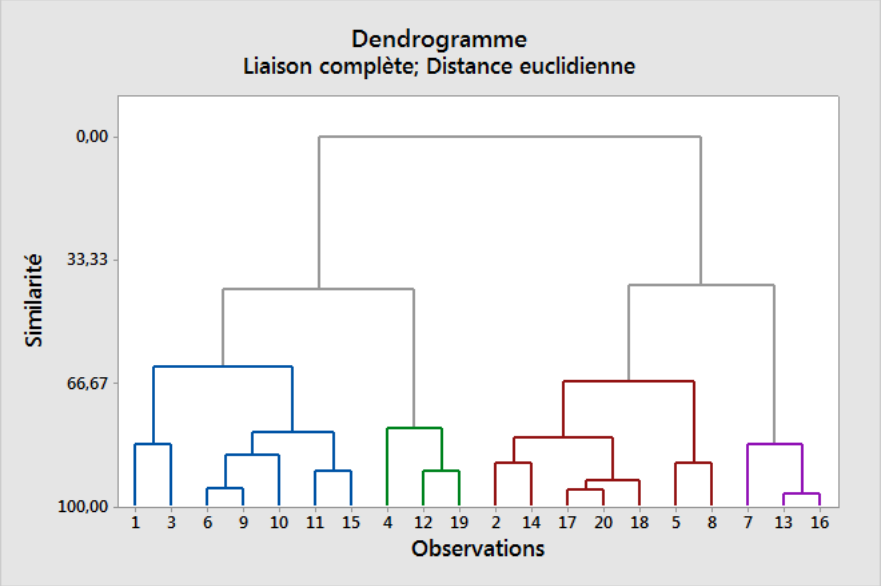


FIGURE 3.7 – Exemple de dendogramme

En ordonnée se lit la mesure de dissimilarité/similarité (dépend du sens de lecture) : au niveau du noeud du groupe bleu, le groupe composé des observations $\{1, 3\}$ et celui composé des observations $\{6, 9, 10, 11, 15\}$ possèdent une similarité de 66.67 environ au sens de la mesure de dissimilarité employée.

3.2.4 Clustering avec des forêts aléatoires

Dans le cas du RF non supervisé, nous n'avons pas de variable cible vers laquelle il est demandé à l'algortihme de tendre.

Pour une RF non supervisée le principe reste le même que dans une RF supervisée : un échantillon aléatoire de m variables explicatives est tiré parmi les p disponibles pour chaque arbre entraîné. A l'origine cet algorithme a été utilisé dans le cadre de travaux en biologie : L.Breiman et A.Cutler en 2003 l'on appliqué à du clustering sur des données ADN, Allen et al. en 2003 à des séquences génomiques, et T.Shi et al. en 2004 sur des données de marqueurs tumorales.

Dans le cas d'une RF non supervisée, une distribution conjointe des variables explicatives est construite et des tirages sont effectués à partir de cette distribution pour créer des données synthétiques. Dans la plupart des cas, le même nombre de tirages que dans l'ensemble de données initial sera effectué. Ensuite, les données initiales et synthétiques sont combinées et une nouvelle variable est créée. Cette variable vaudra par exemple 1 pour les données initiales et 0 pour les données synthétiques.

La RF non supervisée fonctionne ensuite de la même manière qu'une RF supervisée : elle construit un ensemble d'arbres CART (*weak learners*) et en détermine si l'individu i (n individus) appartient aux données initiales ou aux données synthétiques créées (classification supervisée $K = 2$)

Il est alors possible de calculer une matrice $n \times n$ de similarité (ou de dissimilarité) M. Il s'agit d'une matrice où chaque valeur représente la proportion de fois où l'observation i et j se trouvent dans le même nœud terminal. Par exemple, si 100 arbres ont été ajustés et que $M_{ij} = 0.9$, cela signifie que 90 fois sur 100 l'observation i et j étaient dans le même nœud terminal. Ainsi, la similarité de i et de j est de 90%

La RF non supervisée mesure donc une distance et n'est pas un algorithme de clustering à proprement parler.

Enfin, avec la matrice de similarité (ou dissimilarité), il est alors possible d'effectuer une procédure classique de clustering comme un K-means ou une CAH etc...

Synthétiquement, voici une procédure pour effectuer un clustering dont la base est une RF non supervisée :

- 1. Étiqueter les données initiales comme appartenant à la classe "donnée initiale"
- 2. Générer des données synthétiques, connaissant les données initiales, selon une des deux méthodes suivantes :
  - 2.a Échantillonnage indépendant de chacune des distributions univariées de chaque variables
  - 3.b Échantillonnage indépendant à partir de lois uniformes, de telle sorte à ce que chaque loi uniforme ait une étendue égale à l'étendue de la variable qu'elle modélise
- 3. Etiqueter les données générées comme appartenant à la classe "donnée générée"
- 4. Construction d'une RF supervisée de classification à $K = 2$ qui prédit si chaque individu appartient à la classe "donnée initiale" ou bien à la classe "donnée générée"
- 5. Utiliser la matrice de similarité/dissimilarité dans un algorithme de clustering

3.2.5 Modèle de mélange Gaussien et apprentissage non supervisé

Il s'agit d'une approche de clustering par modélisation (versus clustering par partitionnement type K-means et clustering hiérarchique type CAH)

Considérons un échantillon de n individus $(X_1, ..., X_n)$ caractérisés par d variables continues. Supposons que chacun de ces n individus appartiennent à 2 groupes différents pour simplifier et que chacun des de ces 2 groupes suivent une loi normale $\mathcal{N}(\mu_j, \Sigma_j)$. Dans la classe 1, les données X suivent une loi normale $\mathcal{N}(\mu_1, \Sigma_1)$ et dans la classe 2 les données X suivent une loi normale $\mathcal{N}(\mu_2, \Sigma_2)$.

Alors la loi marginale de X est :

$$f(X_i) = \pi_1 f(X_i|Z = 1) + \pi_2 f(X_i|Z = 2)$$

avec π_j probabilité a priori que l'individu appartienne à la classe j (la somme de ces probabilités vaut 1), $f(X_i|Z=j)$ densité conditionnelle suivant une loi normale $\mathcal{N}(\mu_j, \Sigma_j)$ et Z variable aléatoire indiquant la classe de X_i. $f(X_i)$, appelé le mélange de densités, est totalement déterminée si les paramètres π_j, μ_j et Σ_j sont connus ($j = 1, 2$ ici). Ces paramètres sont regroupés dans θ l'ensemble des paramètres.

Si on connaît le modèle de mélange, alors on connaît les probabilités à priori π_j et les lois conditionnelles et alors on peut en déduire les probabilités à posteriori d'appartenance de X_i à la classe j : $\mathbb{P}(Z=j|X_i) = \frac{\pi_j f(X_i|Z=j)}{f(X_i)}$ et $\sum_j \mathbb{P}(Z=j|X_i) = 1$.

X_i est alors affecté à la classe 1 si la probabilité à posteriori d'appartenance à la classe 1 est plus grande que celle de la classe 2 (cela se généralise facilement pour $K > 2$ cluster)

Pour estimer les paramètres $\theta = (\pi_1, \mu_1, \Sigma_1, \pi_2, \mu_2, \Sigma_2)$, on utilise la méthode de la maximisation de la log-vraissemblance :

$$L(\theta; x_1, ..., x_n) = \sum_{i=1}^{n} log\left(\pi_1 f_{\mathcal{N}(x_i;\mu_1,\Sigma_1)} + \pi_2 f_{\mathcal{N}(x_i;\mu_2,\Sigma_2)}\right)$$

qui ne possède pas de solution analytique à cause du logarithme de la somme. Cependant, si l'on donne arbitrairement l'appartenance z_i d'un cluster pour chaque individu i, alors la log-vraissemblance complétée est :

$$L(\theta, x_1, ..., x_n, z_1, ...z_n) = \sum_{i=1}^{n} log(z_i f_{\mathcal{N}(x_i,\mu_1,\Sigma_1)} + (1-z_i) f_{\mathcal{N}(x_i,\mu_2,\Sigma_2)})$$

en considérant $z_i = 1$ si x_i est dans la classe 1 et $z_i = 0$ si x_i est dans la classe 2. Il est alors possible de déterminer $\hat{\pi_j}$, $\hat{\mu_j}$, et $\hat{\Sigma_j}$

Cependant, les z_i ont étés fixés de manière arbitraire. Pour contourner ce problème, prenons l'espérance conditionnelle de L :

$$\mathbb{E}(L|x_1, ..., x_n) = \sum_{i=1}^{n} \mathbb{E}(z_i|x_i) log\left(\pi_1 f_{\mathcal{N}(x_i;\mu_1,\Sigma_1)}\right) + (1 - \mathbb{E}(z_i|x_i)) log\left(\pi_2 f_{\mathcal{N}(x_i;\mu_2,\Sigma_2)}\right)$$

avec $\mathbb{E}(z_i|x_i) = \mathbb{P}(z_i = 1|x_i)$ la probabilité à posteriori que x_i soit dans le groupe 1 Ce nombre indique le nombre de fois où x_i appartiendrait à la classe 1 si on tirait aléatoirement un grand nombre de fois la classe de x_i.

Puis par la loi de Bayes :

$$\mathbb{P}(z_i = 1|x_i) = \frac{\mathbb{P}(z_i = 1) f(x_i|z_i = 1)}{f(x_i)} = \frac{\pi_1 f_{\mathcal{N}(x_i;\mu_1,\Sigma_1)}}{\pi_1 f_{\mathcal{N}(x_i;\mu_1,\Sigma_1)} + \pi_2 f_{\mathcal{N}(x_i;\mu_2,\Sigma_2)}}$$

Connaissant ces paramètres, il est alors possible de calculer les probabilités à posteriori d'appartenance à un groupe en utilisant l'algorithme EM (Espérance Maximisation, D.Rubin en 1977 [14]) qui peut être synthétisé comme ci-dessous tant que la convergence n'est pas constatée :

- étape Estimation : calcul des probabilités à posteriori :

$$\mathbb{P}(z_i = 1|x_i) = \gamma_i^{(1)} = \frac{\pi_1 f_{\mathcal{N}(x_i;\mu_1,\Sigma_1)}}{\pi_1 f_{\mathcal{N}(x_i;\mu_1,\Sigma_1)} + \pi_2 f_{\mathcal{N}(x_i;\mu_2,\Sigma_2)}} = 1 - \gamma_i^{(2)},\ \forall i = 1, ..., n$$

- étape Maximisation : calcul des paramètres :

$$\mu_j = \frac{\sum_{i=1}^n \gamma_i^{(j)} x_i}{\sum_{i=1}^n \gamma_i^{(j)}}\ ;\ \pi_j = \frac{\sum_{i=1}^n \gamma_i^{(j)}}{n}\ ;\ \Sigma_j = \frac{\sum_{i=1}^n \gamma_i^{(j)}(x_i - \mu_j)(x_i - \mu_j)'}{\sum_{i=1}^n \gamma_i^{(j)}}$$

La convergence s'observe en terme de variation de la log-vraissemblance et la convergence se fait vers un extremum local. L'initialisation peut se faire avec un k-means ou aléatoirement.

Chaque composante d'une densité de mélange est associée à un cluster. La plupart des applications supposent que toutes les densités de composants proviennent de la même famille de distribution paramétrique. Le modèle de mélange gaussien (GMM) suppose une distribution gaussienne (multivariée) pour chaque composante, c'est-à-dire $f_k(x, \theta_k)$ suit une loi normal $\mathcal{N}(\mu_k, \Sigma_k)$. Ainsi, les clusters sont ellipsoïdaux, centrés sur le vecteur moyen μ_k, et avec d'autres caractéristiques géométriques, telles que le volume, la forme et l'orientation, qui sont déterminés par la matrice de covariance Σ_k. Des paramétrages des matrices de covariance peuvent être obtenus au moyen d'une décomposition de la forme $\Sigma_k = \lambda_k D_k A_k D_k^T$, où λ_k est un scalaire contrôlant le volume de l'ellipsoïde, A_k une matrice diagonale spécifiant la forme des contours de densité, et D_k une matrice déterminant l'orientation de l'ellipsoïde correspondante. Dans une dimension, il n'y a que deux modèles : le modèle E pour la variance égale ou le modèle V pour la variance variable. Dans le cadre multivarié, le volume, la forme et l'orientation des covariances peuvent être contraints à être égaux ou variables entre les groupes : il y a donc 14 modèles possibles avec des caractéristiques géométriques différentes (EEE, EVE, VEE, VEV, VVV etc...). [34]

Le package *mclust* dans *R* permet de tester ces 14 modèles pour des nombres de cluster k différents. Le critère de sélection des paramètres utilisé est le BIC qui est défini par $BIC = 2.ll(\hat{\Psi}) - v.log(n)$ où $ll(\hat{\Psi})$ est la log-vraissemblance du modèle $\hat{\Psi}$, v le nombre de paramètres estimés, et n le nombre d'individus. Ce critère est donc à maximiser tel que défini.

3.2.6 Clustering de séries temporelles

Pour regrouper des séries temporelles similaires, il est possible d'utiliser les méthodes classiques de clustering hérarchiques ou non décrites plus haut (CAH, K-means...). Cependant, dans le cas de séries temporelles de types financières comme l'évolution de la valeur liquidative de nos fonds, cela présentent des problèmes : décalages temporels, longueur de séries temporelles différentes, bruit, différence d'échelles, etc... Pour la différence d'échelle (les valeurs liquidatives vont de 2€ à 35 000€), un simple centrage/réduction permet de mettre à l'échelle toutes les séries temporelles afin de les comparer. Mais les

autres problèmes soulevés par la nature même en deux dimensions des séries temporelles vont nécessiter des outils plus puissants [16].

Il existe pléthore de littérature sur le sujet, et ici nous retiendrons les articles de Rani [42] ainsi que celui de Thomas Lampert et al. [46], ainsi que de la page internet NCBI [35] qui donnent une bonne vue d'ensemble des méthodes possibles. Ainsi quatre approches sont possibles :

- Sur les séries brutes (*temporal-proximity based clustering*) : application directe des méthodes de clustering sur les séries brutes. Il faut dans ce cas utiliser une mesure de distance adéquate car des signaux similaires décalés dans le temps pourraient être mal interprétés par une mesure classique comme la distance euclidienne. Par exemple la distance DTW (Dynamic Time Warping) qui considère un axe temporel élastique non linéaire permet elle de prendre en compte comme similarité les motifs décalés dans le temps entre deux séries temporelles. Il s'agira de la méthode A. Schématiquement :

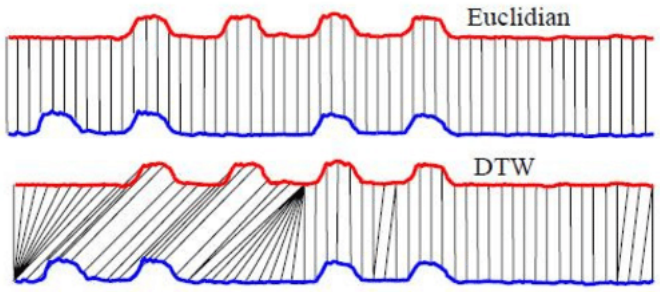


FIGURE 3.8 – Représentation de la différence entre la distance euclidienne et la distance DTW

- Sur une représentation simplifiée des séries temporelles (*representation based clustering*) : les séries subissent un traitement préliminaire, par exemple, une réduction de dimension afin d'extraire des séries leurs caractéristiques principales afin de les regrouper par caractéristiques similaires.

- Sur la modélisation des séries temporelles (*model based clustering*) : par des modèles types $ARIMA$ par exemple. Deux séries sont considérées comme similaires si leur modélisation est similaire (par exemple, les résidus de leur modélisation par un processus $ARMA$ sont similaires). N'importe quel couple distance/algorithme de clustering peut-être utilisé dès lors.

- Sur des métriques d'évaluations des séries (*characteristic based clustering*) : chaque série temporelle est caractérisée par des indicateurs bien choisis, par exemple la tendance, la saisonnalité, le kurtosis, la non-linéarité etc... Par la suite, n'importe quel couple distance/algorithme de clustering est utilisé pour former les groupes

et une ACP est utilisée par exemple pour caractériser les groupes. Il s'agira de la méthode B.

- Sur de courts échantillons de chaque série temporelle (*Shappelet based clustering*) : l'enjeu est de trouver de courts patterns définissant un cluster. Une série temporelle appartient donc à un cluster si elle présente, peu importe à quel moment dans sa "vie", les mêmes pattern représentatifs de ce même cluster.

Il est important d'essayer plusieurs couples distance/algorithme pour chacune de ces approches. En effet, chacun des triplets méthode/distance/algorithme va potentiellement "capturer" une dynamique sous-jacente différente de nos séries temporelles. De plus, la littérature est unanime concernant le meilleur algorithme de clustering de série temporelle : il n'y en a pas car cela dépend de beaucoup de paramètres liés à la nature et à la structure des séries temporelles étudiées.

Dans ce mémoire sera étudié la *temporal-proximity based clustering* (méthode A) et la *characteristic based clustering* (méthode B)

3.2.6.1 Clustering de séries temporelles : méthode A

L'approche *temporal-proximity based clustering* s'applique directement sur les séries brutes. Il faut dans ce cas utiliser une mesure de distance adéquate car des signaux similaires décalés dans le temps pourraient être mal interprétés par une mesure classique comme la distance euclidienne.

Ici, la distance DTW (Dynamic Time Warping) sera utilisée et a été introduite par H.Sakoe et S.Chiba en 1971 [22].

Cette distance considère un axe temporel élastique non linéaire et permet de prendre en compte comme similarité les motifs décalés dans le temps entre deux séries temporelles. Elle tente d'"appareiller" les observations de deux séries temporelles. Voici certaines règles et contraintes de l'algorithme :

- Chaque observation de la première série temporelle doit être associée avec une ou plusieurs observations de la seconde série temporelle avec laquelle elle est comparée, et ceci réciproquement.

- La première et dernière observation de la première série doivent être associées (pas forcément uniquement) respectivement à la première et dernière observation de la seconde série.

- Le mapping de l'appareillement des observations de la première série aux observations de la seconde doit être croissant monotone, c'est à dire que si $j > i$ sont les indices de la première série alors il ne doit pas y avoir deux observations $l > k$ telles que i est associé avec l et j, et ceux réciproquement entre les deux séries.

- le coût, c'est à dire la distance entre les deux séries temporelles, est calculé comme la somme des différences au carré, pour chaque paire d'indices appariés, entre leurs valeurs.

Le mapping optimal est donc le mapping qui satisfait toutes les restrictions et règles et qui présente le coût minimal.

Formellement, soient deux séries temporelles $a = (a_1, a_2, ..., a_m)$, $b = (b_1, b_2, ..., b_m)$ (on suppose ici que les deux séries ont la même taille), et $M(a, b)$ la matrice de distance $m \times m$ entre les deux séries a et b où $M_{i,j} = (a_i - b_j)^2$. Le "chemin de distorsion" $P = ((e_1, f_1), (e_2, f_2), ..., (e_s, f_s))$ est la série de points (couples d'indices) traduisant un parcours de la matrice M.

Par exemple, pour la distance euclidienne $d_E(a, b) = \sum_{i=1}^{m}(a_i - b_i)^2$ est le parcours de la matrice M dans sa diagonale.

Les conditions d'éligibilité à un chemin de distorsion citées plus haut sont traduites par $(e_1, f_1) = (1, 1)$, $(e_s, f_s) = (m, m)$, et $\forall i < m$ on a $0 \leq e_{i+1} - e_i \leq 1$ et $0 \leq f_{i+1} - f_i \leq 1$. La distance DTW est alors définie comme étant le chemin de distorsion sur M qui minimise la distance totale. Soit $p_i = M_{e_i,f_i}$ la distance entre une position e_i de a et une position f_i de b pour le i^{eme} couple de points d'un chemin de distorsion P. La distance d'un chemin P est alors $D_P(a, b) = \sum_{i=1}^{s} p_i$.

Si $\mathcal{P}$ is l'espace de l'ensemble des chemins de distorsions possibles, le chemin DTW $\mathcal{P}^*$ correspond à celui minimisant la distance :

$$\mathcal{P}^* = \min_{P \in \mathcal{P}} D_P(a, b)$$

Il est possible d'ajouter une contrainte de nombre de "distorsion" maximale r effectuée, c'est à dire le nombre maximal de couple d'indice pour lesquelles $(e_k, f_k) \neq (a_l, f_l)$. La contrainte s'écrit : $|e_i - f_i| \leq r.m \ \forall (e_i, f_i) \in P^*$

Voici un pseudo-code possible [48] :

Algorithm 4: Distance DTW

initialisation : poser n la longueur de la série temporelle a, et m celle de b ;
Poser $DTW(i, j)$ la matrice des distances entre les éléments $a\,[1 : i]$ et $b\,[1 : j]$ de a et b
for $i = 1$ **to** n **do**
  for $j = 1$ **to** m **do**
    $DTW(i, j) = \infty$
  end
end
$DTW(0, 0) = 0$
for $i = 1$ **to** n **do**
  for $j = 1$ **to** m **do**
    $cout = distance(a_i, b_j)$;
    $DTW(i, j) = cout + min(DTW(i - 1, j),$
    DTW(i,j-1),
    DTW(i-1,j-1))**end**
  end
  retourner($DTW(n, m)$)

De manière non exhaustive, Il existe des variantes :

- Weigthed DTW (2011) : l'idée est d'appliquer un poids à la mesure de distance telle que $M_{i,j} = w_{|i-j|}.(a_i - b_j)^2$ afin de pénaliser les trop grosses distorsions.
- TWE (Time Warp Edit, 2009) : assez similaire au WDTW, qui consiste à créer une mesure de distance élastique à l'aide de paramètres de raideurs.
- MSM (Move-split-Merge, 2013) : la mesure de similarité intervient dans le cadre d'opérations élémentaires autorisées : substitution (move), insertion (split), et suppression (merge).
- SDTW (Soft DTW, 2017) : cette fois-ci la distance DTW peut renvoyer des valeurs négatives et SDTW(x,x) peut être différent de 0.

3.2.6.2 Clustering de séries temporelles : méthode B

Il s'agit de la *characteristic based clustering* où les séries sont résumées en indicateurs, à choisir. Un ensemble de n séries temporelles est donc représentée par p indicateurs dans une matrice X $n \times p$. Un algorithme de classification couplé à une distance est utilisé sur la matrice X des indicateurs pour réaliser la formation des groupes. Ensuite, en vue de caractériser les groupes, une réduction de dimension, comme une ACP par exemple, est effectuée afin d'extraire des groupes leurs caractéristiques principales.

Dans ce mémoire il est proposé deux types d'ensembles d'indicateurs : les indicateurs "statistiques" et les indicateurs "financiers". Les indicateurs statistiques sont les indicateurs proposés par Wang et al. (2006) [50] et dont RJ.Hyndman fournit le code R [41]. Les indicateurs financiers sont des indicateurs classiques utilisés en finance que nous proposons d'utiliser dans ce mémoire sur les séries temporelles des valeurs liquidatives des fonds UC et euros.

Les indicateurs "statistiques" :

- Ils sont au nombre de 13, et mis sur une échelle $[0, 1]$ -

- **"frequency"** : la période de la série temporelle.
- **"trend"** : la tendance de la série temporelle.
- **"seasonal"** : la saisonnalité de la série temporelle.
- **"autocorrelation"** : la mesure de l'autocorrélation de la série temporelle en utilisant un test de Box-Pierce. Plus particulièrement, la statistique de test du test de Box-Pierce : plus la statistique est élevée, plus la probabilité pour que celle-ci soit la réalisation de la variable aléatoire cible est faible (p-value faible), plus la probabilité de rejeter l'hypothèse nulle d'indépendance augmente. Synthétiquement, plus la statistique de test est élevée, plus la série temporelle est autocoréllée.
- **"non-linear"** : la mesure de la non-linéarité de la série en utilisant un test de Teraesvirta basé sur des réseaux de neurones. Le même mécanisme que pour l'autocorrélation est employé : l'indicateur "non-linear" est la statistique de test du

test de Teraesvita. L'hypothèse nulle ici est la "linéarité en moyenne" de la série temporelle. Plus "non-linear" sera élevé, plus la statistique de test est élevée, plus la probabilité de rejeter l'hypothèse nulle est grande, c'est à dire plus la série temporelle sera "non linéaire en moyenne".

- **"skewness"** : le skewness de la série temporelle, c'est à dire le coefficient d'asymétrie de la série lorsque celle-ci est vue comme le tirage successif d'une variable aléatoire X. Le skewness est défini comme $skewness = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ avec μ la moyenne et σ l'écart type. Un skewness positif caratérisera des séries à queues inférieures lourdes, et un skewness négatif une queue droite lourde.

- **"kurtosis"** : le kurtosis de la série temporelle, c'est à dire le coefficient d'aplatissement de la série temporelle lorsque celle-ci est vue comme un tirage successif d'une variable aléatoire X. Le kurtosis est défini comme $kurtosis = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$ avec μ la moyenne et σ l'écart type. Un kurtosis élevé indique que la série est plutôt pointue en sa moyenne (centre),alors qu'un kurtosis faible indique une distributions aplatie.

- **"Hurst"** : le paramètre de "mémoire à long terme" de la série temporelle représenté par l'exposant de Hurst noté H, qui est définit par : $\mathbb{E}\left[\frac{\mathcal{R}(n)}{\mathcal{S}(n)}\right] = Cn^H$ où $\mathcal{R}(n)$ est la plage du premier n écarts cumulé par rapport à la moyenne, $\mathcal{S}(n)$ la série des n premiers écarts types, n la durée d'observation et C une constante. L'exposant de Hurst quantifie la tendance relative d'une série temporelle à régresser vers sa moyenne. Si $H \in [0, 1]$: la série présente une autocorrélation positive à long terme. Si $H \in [0, 0.5]$: la série temporelle a tendance à alterner les valeurs hautes et basses. Le calcul de celui-ci n'est pas détaillé ici car dépasse le cadre de ce mémoire.

- **"Lyapounov"** : le coefficient de Lyapounov qui quantifie l'instabilité d'une série temporelle (à valeur dans $\mathbb{R}$). Le calcul de celui-ci n'est pas détaillé ici car dépasse le cadre de ce mémoire. Une série instable présente un exposant de Liapounov positif, une série stable présente elle à l'inverse un exposant de Liapounov négatif.

- **"dc autocorrelation", "dc non-linear", "dc skewness", "dc kurtosis"** : les mêmes indicateurs sans le "dc" calculés sur les séries temporelles ajustées par une transformation de Box-Cox. La transformation de Box-cox n'est pas détaillée ici car dépasse le cadre de ce mémoire.

Les indicateurs "financiers" :
- Ils sont au nombre de 20 -

- **"moy return daily", "moy return weekly", "moy return monthly", "moy return yearly"** : moyenne des rendements journaliers, hebdomadaires, mensuels et annuels de la série temporelle de la valeur liquidative du fond

- **"sigma daily", "sigma weekly", "sigma monthly", "sigma yearly"** : écart-type des rendements journaliers, hebdomadaires, mensuels et annuels de la série temporelle de la valeur liquidative du fond.

- **"ratio sharpe TME" et "ratio sharpe 1"** : le ratio de Sharpe dont le placement de comparaison serait, respectivement, le TME sur la même période et un placement avec 1% de retour annuel. Le ratio de Sharpe mesure l'écart de rendement d'un actif avec un autre placement (souvent un placement sans risque comme une obligation OAT). Le ratio de Sharpe d'un actif S se définit comme : $S = \frac{R-r}{\sigma}$ où R est l'espérance de rendement de l'actif, r le rendement du placement de comparaison, et σ l'écart type de l'actif. Plus le ratio de Sharpe d'un actif est élevé, plus celui-ci sur-performe le placement de comparaison. La courbe du TME est facilement accessible au format excel sur internet, notamment sur le site de la banque de France.

- **"proportion augmentation"** : la proportion des jours clôturant à la hausse pour le fond considéré.

- **"return ancien", "return mid" et "return recent"** : le rendement sur une découpe temporelle en trois morceaux de la série temporelle des valeurs liquidatives du fond considéré. La découpe est faite de manière à avoir autant d'observations dans chaque morceaux et ainsi le morceaux "ancien" correspond aux valeurs liquidatives les plus anciennes.

- **"sd ancien", "sd mid" et "sd recent"** : sur la même découpe temporelle , il s'agit de l'écart type des rendements observés sur ces portions.

- **"alpha" et "beta"** : il s'agit de l'alpha et du beta du fond considéré. Si l'on trace le nuage de points (x_i, y_i) avec x_i le rendement journalier à la date i d'un marché de référence (par exemple le CAC40 dont nous récupérons son évolution sur le site finance.yahoo) et y_i le rendement journalier du fond UC à la date i alors nous obtenons un nuage de points. Effectuer une régression linéaire sur ce nuage de points permet donc de quantifier la sensibilité du fond UC à ce marché de référence : l'alpha est alors défini comme l'intercept de la régression linéaire et le beta comme le régresseur associé au marché de référence. Autrement dit, l'alpha mesure le rendement moyen de l'UC lorsque le marché performe de 0% tandis que le beta mesure le nombre de points de pourcentage de rendements moyen que l'UC aura lorsque le marché de référence augmentera de 1 point de pourcentage. Un beta égal à deux signifie que si le marché de référence présente un rendement journalier de 3%, alors l'UC aura en moyenne un rendement de 6% ce même jour. Le beta et l'alpha mesurent donc l'élasticité d'un actif par rapport à un marché de référence :

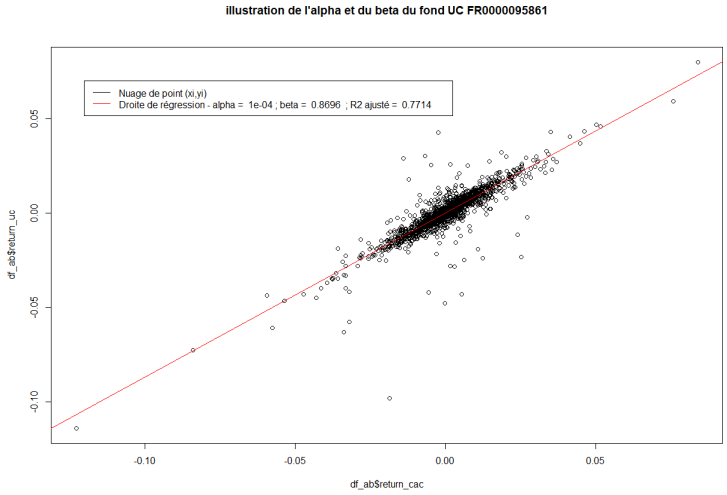


FIGURE 3.9 – Illustration de l'apha et du beta d'un actif, ici un fond UC

Différents ensembles d'indicateurs ont étés considérés : le but est de capturer des informations selon l'ensemble choisi. Ainsi les indicateurs "financiers" auront vocations à capturer les informations tirées d'un clustering construit avec un "oeil" d'investisseur (rendement, volatilité), tandis que les indicateurs "statistiques" auront vocation à capturer les informations tirées d'un clustering construit avec un "oeil" sensible à la série temporelle perçue comme un objet mathématique.

Pour chaque ensemble d'indicateurs et pour chaque série temporelle nous avons des valeurs quantitatives : il est alors possible, dans cette méthode B dans le cadre de ce mémoire, de réaliser une ACP (réduction de dimension) en vue de caractériser "à la main" les éventuels groupes distincts notables. En effet, les variables de l'ACP seront les indicateurs, et les individus seront les séries temporelles (résumées en indicateurs) : il suffit alors d'effectuer une analyse classique d'ACP (contribution, cos^2, caractérisation des dimensions de l'ACP) pour caractériser les cluster.

3.3 Bilan sur les principaux outils utilisés

Dans ce chapitre nous avons présenté les différents concepts de base de ML, ainsi que les algorithmes utilisés tout au long du mémoire. Ces algorithmes sont listés ci-dessous :

A. *Apprentissage statistique supervisé* :

(a) GLM : modèle de régression généralisé

(b) SVM : Support Vector Machine

(c) RF : Random Forest (mode supervisé)

(d) XGboost : eXtreme Gradient boosting

B. *Apprentissage statistique non supervisé* :

(a) ACP : Analyse en Composantes Principales

(b) K-means : K-moyennes

(c) CAH : Classification Ascendante Hiérarchique

(d) RF : Random Forest (mode non-supervisé)

(e) MMG : Modèle de Mélange Gaussien

Dans ce mémoire intervient l'apprentissage statistique non supervisé dans le cas de clustering de séries temporelles (valeur liquidative des fonds UC). Un focus est donc fait sur deux méthodes de clustering de séries temporelles utilisant les algorithmes ci-dessus : le *temporal-proximity based clustering* et le *characteristic based clustering*. Le *characteristic based clustering* nécessitant de qualifier nos séries temporelles par des indicateurs, nous retenons et détaillons deux ensembles d'indicateurs : les indicateurs statistiques et les indicateurs financiers. Enfin, une courte partie est dédiée à la méthode que nous utliserons dans l'interprétation de nos modèles complexes : la méthode SHAP.

Chapitre 4

Clustering en groupe homogènes des fonds UC et euros

4.1 Clustering des séries temporelles des fonds

La première étape, avant la prédiction des taux d'arbitrages, est la constitution des groupes homogènes de fonds en réalisant un clustering des séries temporelles des valeurs liquidatives des fonds UC et euros. En effet, pour prendre en compte l'information des ISIN, nous ne pouvons pas créer une variable explicative par fond : cela représenterait plus de 400 variables, sûrement vides de sens prises séparément. C'est pourquoi il est préférable de regrouper les fonds euros et UC en groupes homogènes. Cette homogénéité sera une homogénéité au sens du panel d'indicateurs statistiques (présenté au chapitre 3), du panel d'indicateurs financiers (présenté au chapitre 3) et au sens de la dynamique des fonds (distance DTW).

Pour rappel, nous avons des indicateurs "statistiques" et des indicateurs "financiers" sur les séries temporelles "résumées" sur lesquelles nous pouvons appliquer différents couples distance/algorithme de clustering.

Pour le clustering sur les séries temporelles brutes, une seule mesure est proposée et il s'agit de la distance DTW, il suffit de trouver un nombre de clusters convenable selon certains critères (CVI : Cluster Validity Indices).

Pour résumer, voici les clusterings que nous allons tester :

Approches considérées	
Temporal-proximity based clustering	*Characteristic based clustering*
distance DTW + K-means/CAH	Distance euclidienne + CAH Distance euclidienne + K-means Matrice de similarité par RF non-supervisée + CAH Modèles de mélange Gaussiens à différentes géométries

Nous disposons de 473 supports UC et euros différents, présents dans nos bases de données d'arbitrages et de PM :

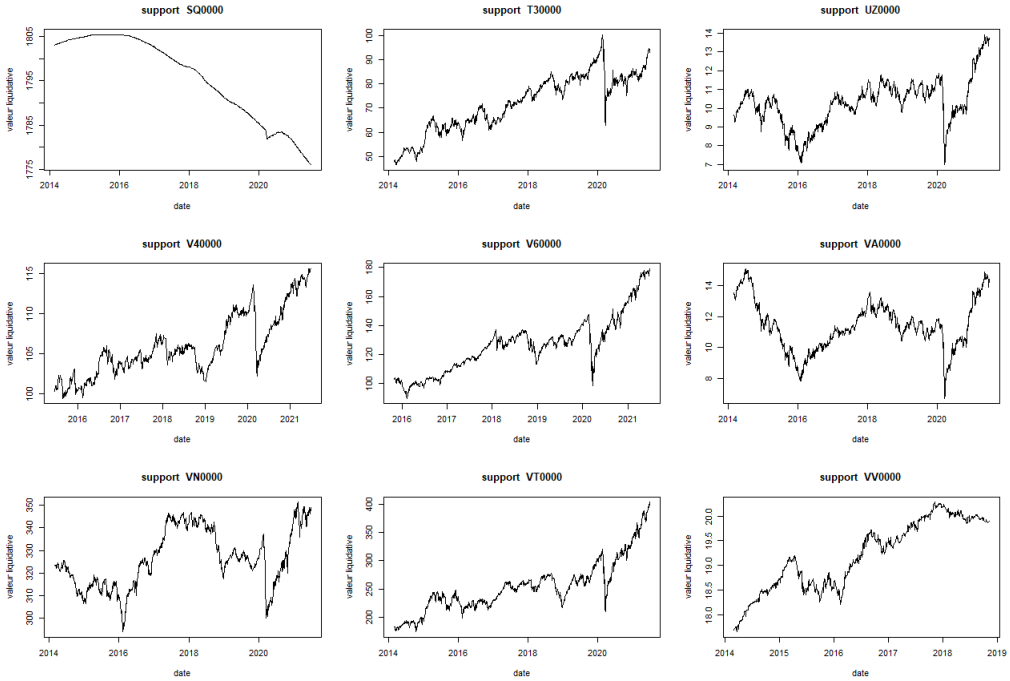


FIGURE 4.1 – Représentation d'un échantillon de 9 fonds UC et euros

Le support SQ0000 (haut à gauche) est un support euros : il présente une dynamique baissière typique de ces fonds. Les autres sont des fonds UC, dont l'impact du premier confinement français lié à la Covid19 est très visible en mars-avril 2020 puisqu'il s'agit très généralement de la plus forte baisse historique de ces supports. Néanmoins, certains UC présentent des différences majeures. Par exemple le support VT0000 (dynamique haussière constante) et le support VA0000 (périodes de baisses et de hausses alternées).

L'intérêt du clustering, si celui-ci parvient à différencier les UC entre elles (comme dans l'exemple précédant), dans la modélisation des taux d'arbitrages, va alors consister à constater (ou non) si les dynamiques d'arbitrages sont différentes selon les cluster d'UC. L'analyse des taux d'arbitrages ne sera alors pas limitée à la proportion d'UC des contrats, mais aux proportions d'UC des contrats appartenant aux différentes typologies des fonds UC.

4.2 Characteristic based clustering

4.2.1 Indicateurs "statistiques"

voici un extrait de la base de données des indicateurs statistiques :

	frequency	trend	seasonal	autocorrelation	non-linear	skewness	kurtosis	hurst	lyapunov	dc autocorrelation	dc non-linear	dc skewness	dc kurtosis
5F0000	0.1683414625	0.570482617	-0.118012308	0.18560782	-0.749164011	0.957172986	0.007926430	0.12385081	-0.91990036	-1.25964978	1.9423323341	0.101464439	0.606693681
5L0000	0.3372188658	0.571607105	-0.132812049	0.18502998	-0.232693234	0.287533393	-0.101407303	0.12385081	-0.78983766	-1.20556399	-0.6055121188	-0.140716773	0.606687161
5N0000	-0.3387637144	-5.315402452	-0.159953431	0.18809760	0.650126650	0.012463619	-0.583957219	0.12741699	-1.13489070	0.97897220	-0.5504048287	1.189059657	0.605779245
5O0000	0.3372188658	0.442448512	0.393410721	0.04472774	0.010146992	0.315967133	0.114492082	0.07004209	-1.43609740	-1.20237636	1.6984331360	-0.865273703	0.606693681

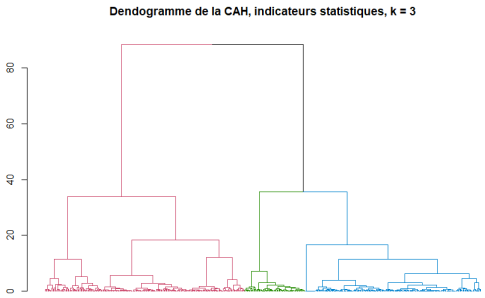
FIGURE 4.2 – Extrait de la base de donnée contenant les indicateurs statistiques des fonds UC et euros

Par la suite, pour chaque couple distance/algorithme, nous retenons le nombre k de clusters qui maximise "l'interprétabilité" en terme :

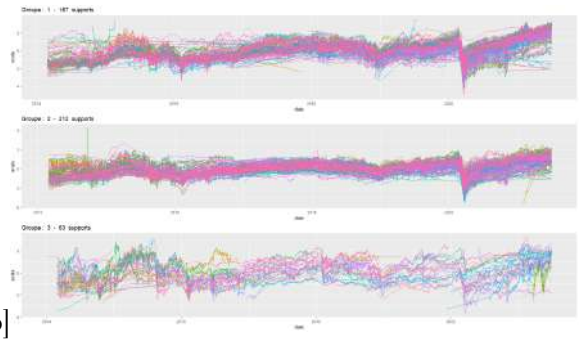
- de forme des clusters intra-groupe,
- de la bonne distinction inter-groupe,
- de la distinction des groupes sur la représentation de ceux-ci sur les axes d'une ACP

Distance euclidienne + CAH

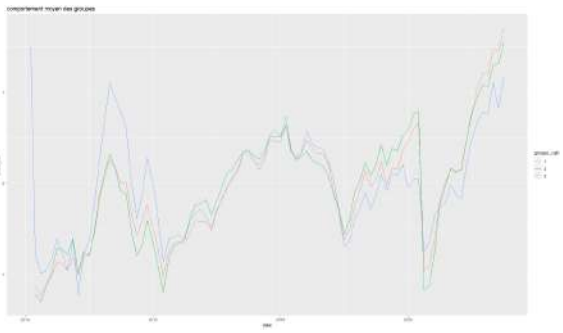
Les séries temporelles des supports sont affichées centrées-réduites afin de pouvoir les comparer. La Méthode de Ward est utilisée pour le calcul des distances inter-groupes. Enfin, une CAH est appliquée :



[a]



[b]



[c]

FIGURE 4.3 – Dendogramme (a), comportement des valeurs liquidatives des fonds par cluster (b), et comportement moyen des clusters (c), pour une distance euclidienne + CAH et 3 clusters, sur les séries temporelles résumées en indicateurs statistiques

Les nombres k de clusters testés vont de 2 à 6 : pas trop pour conserver un minimum d'individu dans chaque groupe et tenter d'opérer une distinction ayant du sens, mais pas trop peu non plus pour ne pas faire des clusters trop "généralisés" faisant état d'une différenciation triviale.

Ici nous choisissons $k = 3$ car c'est le moins mauvais en terme de différenciation intra et inter groupes. Les groupes ne sont pour autant pas bien différenciés en terme de forme, mis à part le groupe 3 qui regroupe des séries plus volatiles. Le reste est trop similaire : juste un effet de niveau/shift vertical selon les périodes temporelles observées. L'allure générale est la même, avec un creux en mars 2020 correspondant à la crise de la COVID19.

Il est possible de caractériser, pas sur la forme mais sur les propriétés statistiques des séries temporelles composant chaque groupe, les groupes en effectuant une ACP.

Les deux premiers axes sont les plus importants et expliquent à eux seuls 38.8% de la variance des données :

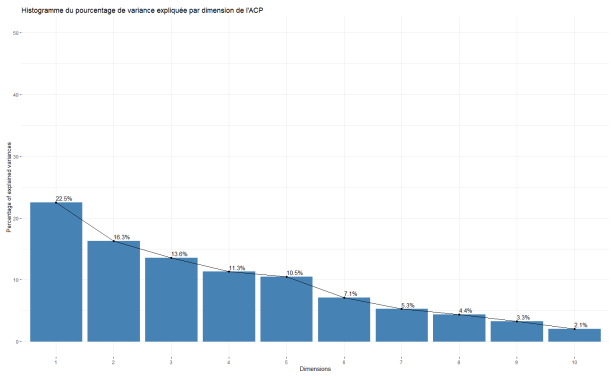


FIGURE 4.4 – Pourcentage de variance expliqué par les dimensions de l'ACP sur les indicateurs statistiques

Le premier axe de l'ACP est construit par les indicateurs "Lyapounov" (contribution de 26%), "dc autocorrelation" (contribution de 23%), "frequency" (contribution de 15%) et "dc non linear" (contribution de 14%), et le second par les indicateurs "autocorrelation" (contribution de 34%), "hurst" (contribution de 28%) :

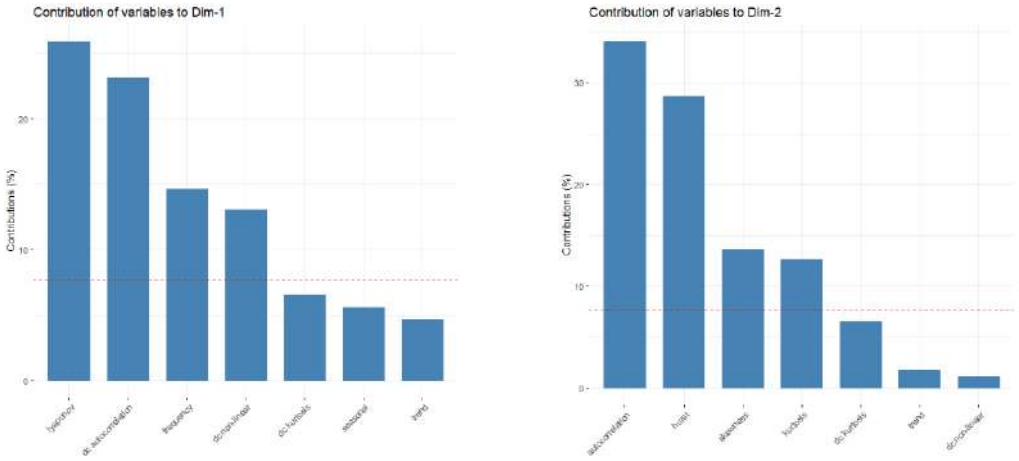


FIGURE 4.5 – Contribution à la construction des axes de l'ACP des indicateurs statistiques

En affichant les séries temporelles (individus caractérisés par les indicateurs statistiques) et les indicateurs sur le même graphique :

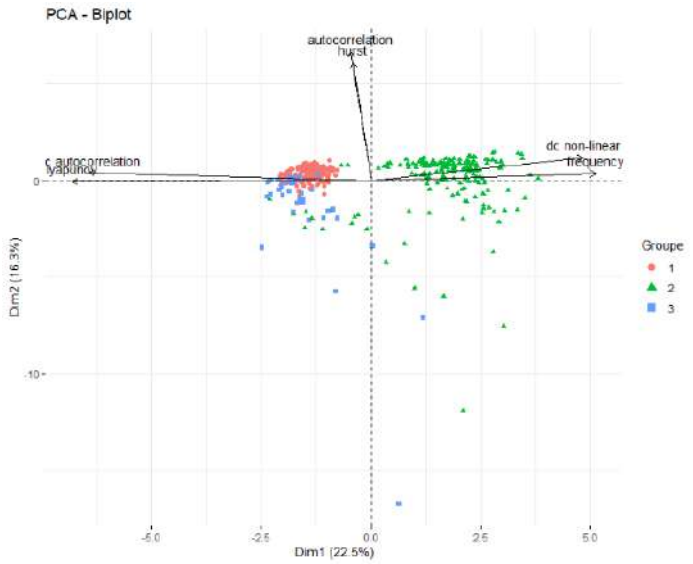


FIGURE 4.6 – Biplot de l'ACP avec indicateurs statistiques, et groupes constitués par une CAH avec k = 3

Le groupe 2 est bien différencié des deux autres : il est caractérisé par des valeurs de "frequency" et de "dc non linear" élevées et des valeurs de "Lyapounov" et "dc autocorrelation" faible : c'est à dire des séries temporelles avec de longues périodes, peu linéaires, assez stable en moyenne et avec une indépendance temporelle (faible valeur de la statistique de test du test de Box-Pierce) Les groupes 2 et 3 semblent être les opposés du groupe 1 : des séries temporelles avec de plus courtes périodes, plus linéaires mais instables en moyenne, et auto corrélées dans le temps.

Manifestement, ce clustering n'est pas satisfaisant : pas de réelle homogénéité de forme intra-classe ni d'hétérogénéité de forme inter-classe. De plus l'analyse des groupes via une ACP n'est pas satisfaisante. Par exemple, les variables "dc autocorrelation" et "autocorrelation" sont perpendiculaires et ne sont pas représentées sur le même axe, les individus sont pour la majorité mal représentés sur les axes cos^2 et contribution faible, certains individus à eux seuls contribuent à construire un axe (l'axe 2 par exemple).

Le clustering par une CAH en méthode de Ward avec distance euclidienne à 3 clusters est donc rejeté sur ce panel d'indicateurs statistiques.

Il est alors décidé de ne garder que les variables en "dc" ("dc autocorrelation" vs "autocorrelation") et de supprimer quelques individus dans l'ACP présentant une contribution trop élevée. Pour rappel le "dc" signifie que l'indicateur a été calculé sur la série temporelle ayant subi une transformation de Box Cox. Il s'agit donc de s'affranchir de la redondance et du bruit présent dans les indicateurs. Les indicateurs "trend", "frequency", "seasonal", "hurst", "lyapounov", "dc autocorrelation", "dc non linear", "dc skwewness" et "dc kurtosis" sont donc conservés et une CAH à $k = 3$ cluster est retenue.

Cela a pour effet d'améliorer la distinction des groupes en terme de caractérisation des axes de l'ACP puisque ceux-ci présentent une plus grande distance inter-groupe et une

meilleure représentation des séries temporelles (individus) sur les axes : l'interprétation est donc plus facile et fiable.

Les deux premiers axes expliquent 61.8% de la variance totale des données et l'axe 1 à lui seul 46.7% :

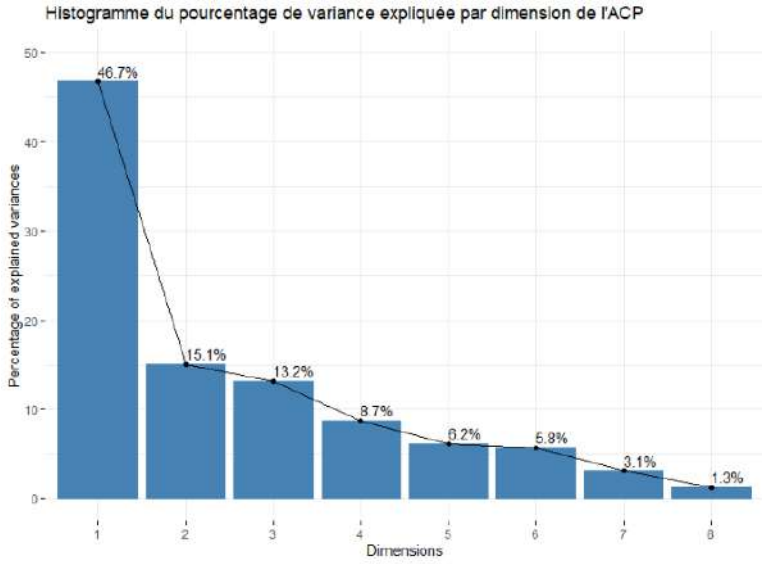


FIGURE 4.7 – Pourcentage de variance expliqué par les dimensions de l'ACP sur les indicateurs statistiques sélectionnés

le premier axe de l'ACP est construit (voir annexe "Contribution à la construction des axes de l'ACP ayant servi à la représentation de du clustering par CAH sur une sélection d'indicateurs statistiques") par les indicateurs "Lyapounov" (contribution de 22%) et "dc autocorrelation" (contribution de 19%) opposés (négativement corrélés) à "frequency" (contribution de 18%) et "trend" (contribution de 12%). Le second axe est caractérisé par les indicateurs "dc skewness" (contribution de 59%) et "dc kurtosis (contribution de 26%)

Le biplot(représentation sur le même graphique des indicateurs et des individus d'une ACP) des deux premiers axes est alors obtenu :

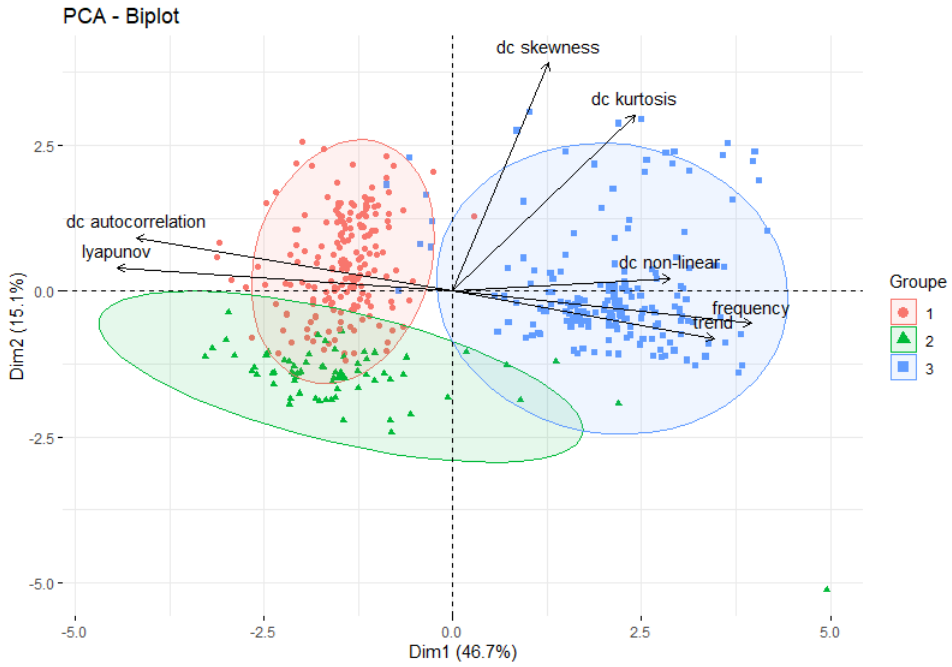


FIGURE 4.8 – Biplot de l'ACP sur la sélection d'indicateurs statistiques des séries temporelles et représentation des cluster indentifié par une CAH à $k = 3$

L'axe 1 oppose les série temporelles instables et autocoréllées (à gauche) aux séries temporelles présentant une tendance et une période élevée. L'axe 2, lui, caractérise les séries présentant des skewness et kurtosis élevés, c'est à dire des distributions plutôt pointues au centre et avec une queue inférieure lourde

Les trois groupes sont bien distincts. Le groupe 3 est opposé aux groupe 1 et 2 sur l'axe 1. Le groupe 2 se différencie du groupe 1 sur l'axe deux où ils sont en opposition assez légèrement.

Le cluster 3 caractérise donc des fonds UC et euros avec une tendance marquée, une période élevé de saisonalité, stable et peu autocoréllée. Les cluster 1 et 2 sont eux représentés par des séries plus instables et autocoréllées, avec une tendance faible voire baissière et une période de saisonalité plus faible. La différence entre le cluster 1 et 2 est que le cluster 2 présente des skewness et kurtosis (en moindre mesure) plus faibles que le cluster 1, c'est à dire que les séries temporelles composant le cluster 2 présentent des distributions plus pointues au centre et une queue inférieure moins lourde, voir une queue supérieure (valeur liquidatives élevée) plus lourde.

La différence de forme entre chaque classe reste cependant assez insatisfaisante car se joue sur des détails : la classe 1 qui présente effectivement sur les valeurs récentes une certaine tendance haussière, la classe 2 qui présente des fonds légèrement plus volatiles (les queues de distributions sont plus lourdes), et la classe 3 qui présente des fonds à dynamique plus monotone (période élevée). Cela se remarque assez bien sur les graphiques suivants représentant les fonds par groupe :

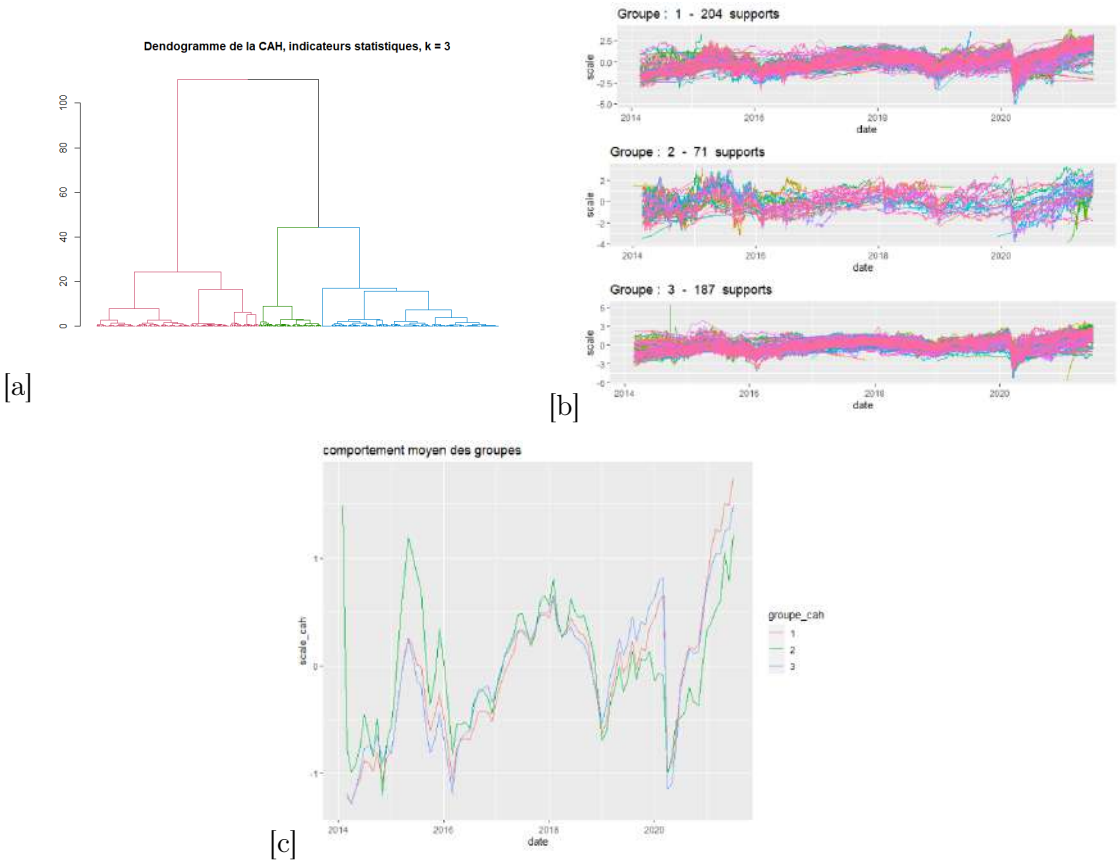


FIGURE 4.9 – Dendogramme (a), comportement des valeurs liquidatives des fonds par cluster (b), et comportement moyen des clusters (c), pour une distance euclidienne + CAH et 3 clusters, sur les séries temporelles résumées en une sélection d'indicateurs statistiques

Distance euclidienne + K-means

Les séries temporelles des supports sont affichées centrées-réduites afin de pouvoir les comparer.

Les nombres k de clusters testés vont de 2 à 10. Pour choisir le nombre k optimal, le graphique des silhouettes moyennes est tracé et nous choisissons k tel que la silhouette moyenne est maximisée :

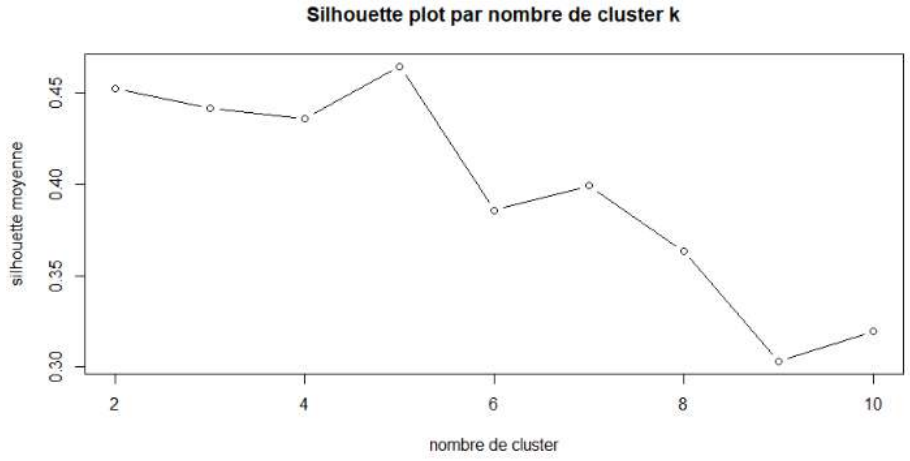


FIGURE 4.10 – Silhouette plot des K-means pour des nombres de clusters allant de 2 à 10, sur les indicateurs statistiques sélectionnés

La silhouette d'un individu i est définit par : $sil_i = \frac{a_i - b_i}{\max(a_i, b_i)}$ avec a_i la distance moyenne de l'individu i à son groupe et b_i la distance moyenne de l'individu i à son groupe voisin.

Il est alors retenu 5 clusters pour lesquelles la silhouette moyenne est de 0.46 :

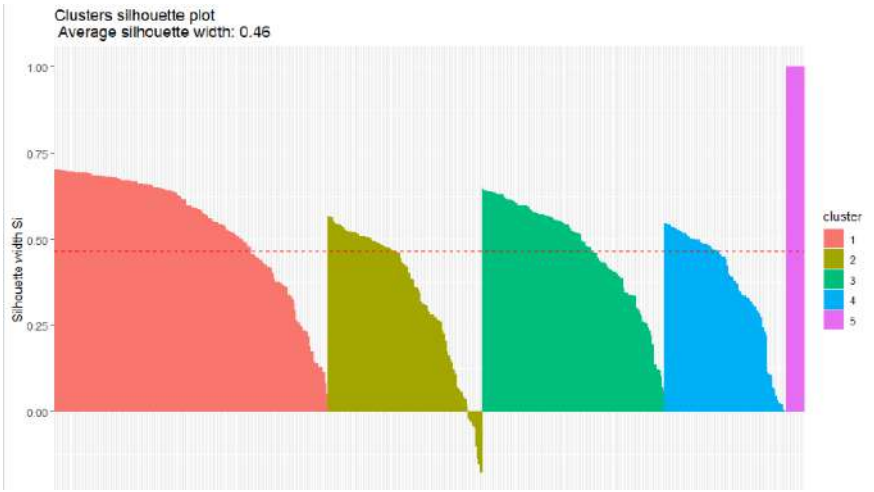


FIGURE 4.11 – Silhouette plot pour $k = 5$, sur les indicateurs statistiques sélectionnés

Le groupe 2 comprend des séries temporelles dont la silhouette est négative : cela signifie qu'elles sont en moyenne plus proches d'un groupe voisin. Cela ne veut pas pour autant dire que la classification est mauvaise : peut être qu'il s'agit là de séries temporelles présentant à elles seules une certaine caractéristique du groupe.

Le groupe 5 est composé de 11 supports identiques (la silhouette du cluster vaut donc

1) : un même ISIN peut correspondre à plusieurs codes supports différents (produits et donc tarification différente etc...)

Voici les séries temporelles centrées-réduites affichées par groupe résultant de ce clustering :

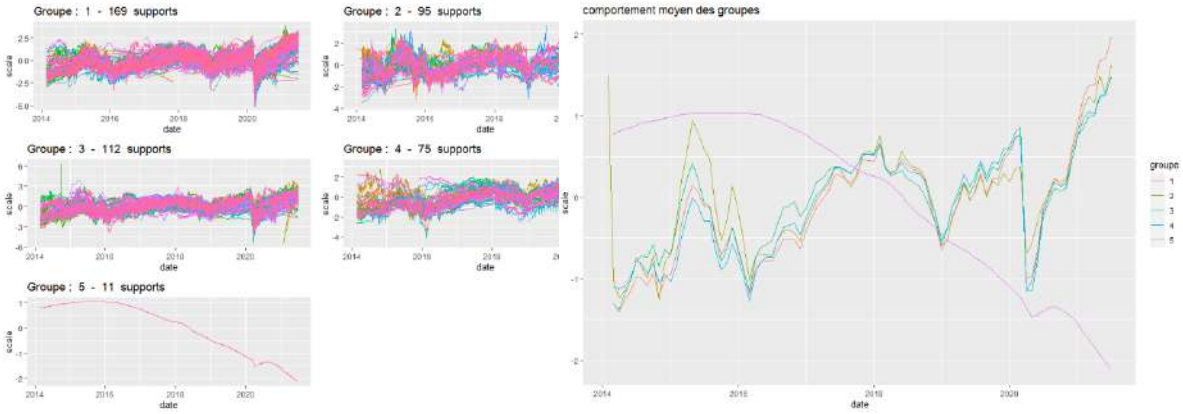


FIGURE 4.12 – Représentation des comportements des fonds UC et euros selon les groupes obtenue par un K-means $k = 5$ sur les indicateurs statistiques sélectionnés

Seul le groupe 5, qui est composé du fond euros Allianz sécurité C, est bien distingué des autres groupes en terme de forme car présente une tendance baissière (la baisse des taux entraîne la baisse des fonds dont une partie est en monétaire).

La caractérisation des groupes va donc se faire sur l'ACP (dont la contribution des indicateurs à la construction des axes se situe en annexe "Résultats de l'ACP ayant servi à la représentation du clustering par K-means sur une sélection d'indicateurs statistiques"). Voici les biplot obtenus :

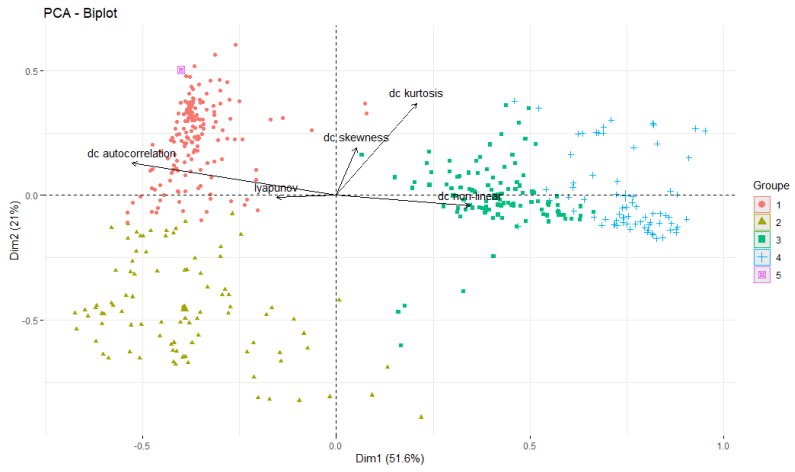


FIGURE 4.13 – Biplot des axes 1 et 2

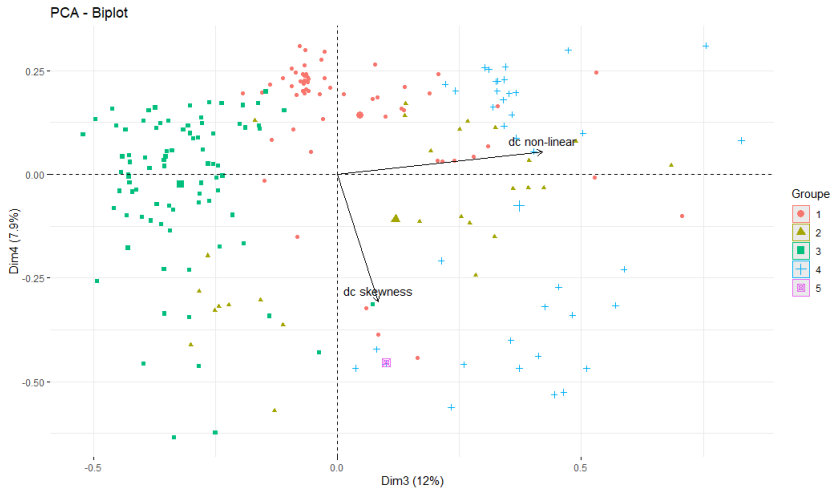


FIGURE 4.14 – Biplot des axes 3 et 4

l'axe 1 est caractérisé par "dc autocorrélation" (58% de contribution). L'axe 2 lui est caractérisé par "dc kurtosis" (70% de contribution) . L'axe 3 est lui caractérisé par "dc non linear" (72% de contribution). Enfin, l'axe 4 est défini par "dc skewness" (59% de contribution).

Le cluster 1 est identifié comme appartenant au cadran supérieur gauche des 2 premiers axes et bien représenté sur l'axe 4 : il s'agit de fonds autocoréllés (axe 1), avec un kurtosis plus élevé (axe 2) et un skewness faible (axe 4), c'est à dire des fonds dépendant de leurs valeurs liquidatives précédentes qui présentent une distribution plutôt aplatie et une queue supérieure de distribution des valeurs liquidatives plutôt élevée.

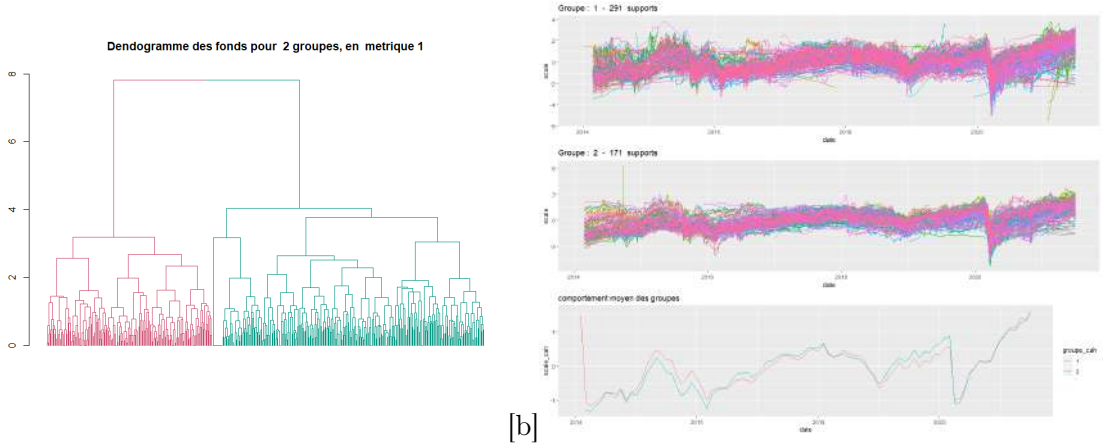
En effectuant le même raisonnement sur les autres groupes :

- le cluster 2 présente des fonds dépendant de leurs valeurs liquidatives précédentes (axe 1) mais avec une distribution plus pointue en son centre (axe 2).
- le cluster 3 présente des fonds dépendants moins de leurs valeurs précédentes (axe 1) et présentant une linéarité marquée en moyenne.
- le cluster 4 présente des fonds indépendants de leurs valeurs précédentes (axe 1) et une non linéarité marquée en moyenne(axe 3).
- le cluster 5 (le fond monétaire Allianz sécurité C), qui se distingue par sa forme, présente des fonds dépendant de leur valeurs précédentes (tendance baissière) et une queue inférieure plutôt lourde (axe 2) au regard de sa distribution assez aplatie (axe 3).

Ici les groupes sur l'ACP sont biens établis, mais la forme générale des groupes est trop similaire (sauf pour le groupe 5, qui est composé de seulement 11 supports qui se trouvent être le même ISIN)

RF non supervisée + CAH

Une RF non supervisée possédant 1000 arbres CART est entraînée sur les indicateurs statistiques sélectionnés afin de calculer une matrice de similarité. Cette matrice de similarité est injectée dans une CAH en vue de former des clusters. Le nombre de cluster retenus est alors $k = 2$, car pour tous les $k > 2$ les groupes formés sur les axes de l'ACP étaient entremêlés :



[a] [b]

FIGURE 4.15 – Dendogramme (a) et représentation du comportements des clusters (b) de la CAH associé au clustering par RF non supervisée pour $k = 2$

Ces deux groupes ne sont pas distingués par leur allure, mais sur l'ACP :

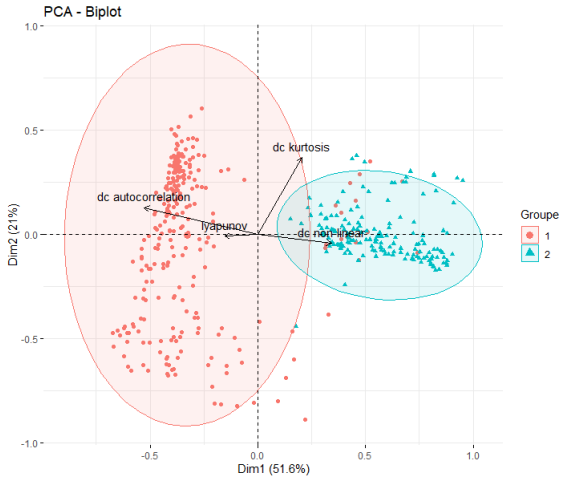


FIGURE 4.16 – Biplot des axes 1 et 2 et représentation des groupes formés par la RF non supervisée

Les axes sont les mêmes que ceux du K-means précédent (voir annexe "Résultats de l'ACP ayant servi à la représentation du clustering par K-means sur une sélection

d'indicateurs statistiques"). Ici la RF ne fait bien la distinction de deux groupes que sur l'axe 1. Le groupe 1 est caractérisé par des fonds autocoréllés , le groupe 2 par des fonds peu autocoréllés. L'analyse de ce clustering n'est pas poussée car celui-ci ne convient pas : il présente trop peu de clusters à la vue du sens qu'il donne à chaque cluster.

Modèle de mélange Gaussiens (MMG)

La fonction $Mclust()$ du package $mclust$ du logiciel R sélectionne le modèle de mélange gaussien optimal selon le critère BIC, en utilisant l'algorithme EM. L'attribution initiale des groupes aux individus se fait par clustering hiérarchique par défaut.

Nous choisissons le modèle et le nombre de clusters avec le BIC (Bayesian Information Criteria) le plus grand.

Nous regardons un nombre k de cluster allant de 2 à 10 pour les 14 algorithmes proposés par la fonction $Mclust()$:

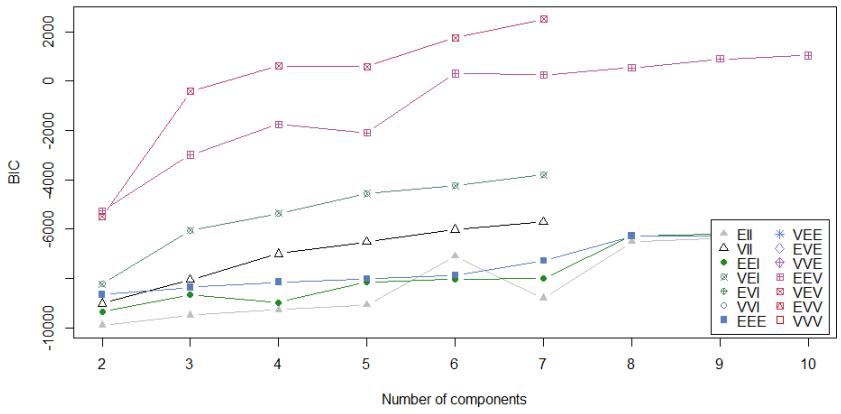


FIGURE 4.17 – Graphique BIC = f(k,géométrie MMG) du package Mclust sur le logiciel R

Nous souhaitons maximiser le BIC ici car la fonction R calcul le BIC suivant la formule : $BIC = 2.ll(\hat{\Psi}) - v.log(n)$ (maximisation de la log-vraissemblance tout en pénalisant les modèles présentant trop de paramètres). Il faudrait prendre les caractéristiques géométriques VEV (ellipsoïdale, de forme égale, volume et orientation variables) avec $k = 7$ mais cela ferait trop de classes et une confusion dans l'analyse de celles-ci. Nous choisissons plutôt $k = 4$ avec les mêmes caractéristiques géométriques qui possèdent tout de même un plus grand BIC que toutes les autres géométries peu importe le nombre de clusters $k < 7$.

Nous obtenons alors :

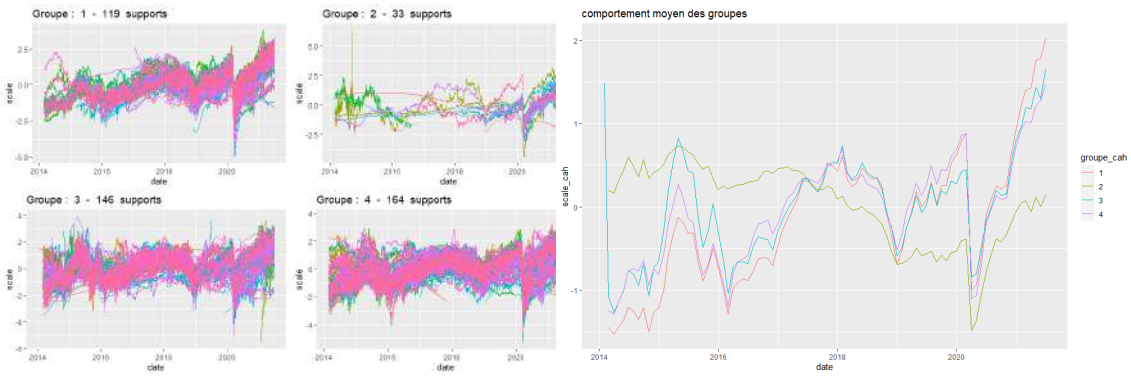


FIGURE 4.18 – Comportement des groupes formés par un MMG de géométrie VEV avec $k = 4$

L'allure générale des groupes, comme les autres clusterings proposés, est toujours très similaire : un pic à la baisse en mars 2020 (Covid19) suivi d'un rebond dépassant le niveau ante-Covid19, et une bosse centrée en 2018. Le groupe 2 se distingue par une sur performance entre 2014 et 2017 et une sous performance entre 2019 et 2021 comparé aux autres groupes. Ce dernier présente une tendance baissière sur l'historique dont nous disposons (d'ailleurs il est composé par le fonds monétaire Allianz sécurité C représentant 11 supports qui est décroissant monotone). Le groupe 3 est caractérisé par les fonds les plus volatiles. Les groupes 1 et 4 sont assez similaires en formes, mais la distinction se fera au niveau de l'ACP.

Si l'on regarde l'ACP (dont la contribution des indicateurs à la construction des axes est la même que pour le K-means et la RF, voir annexe "Résultats de l'ACP ayant servi à la représentation du clustering par K-means sur une sélection d'indicateurs statistiques") :

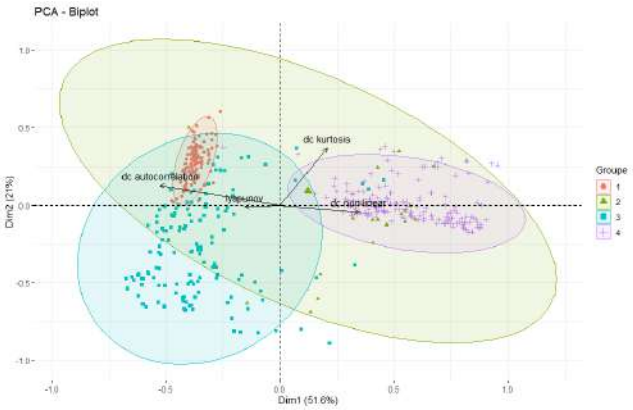


FIGURE 4.19 – Biplot des deux premier axes de l'ACP sur les indicateurs statistiques, avec des groupes formés par un MMG de géométrie VEV à $k = 4$

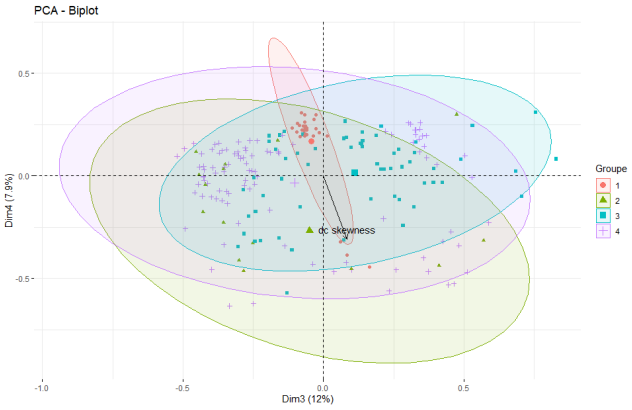


FIGURE 4.20 – Biplot des axes 3 et 4 de l'ACP sur les indicateurs statistiques, avec des groupes formés par un MMG de géométrie VEV à $k = 4$

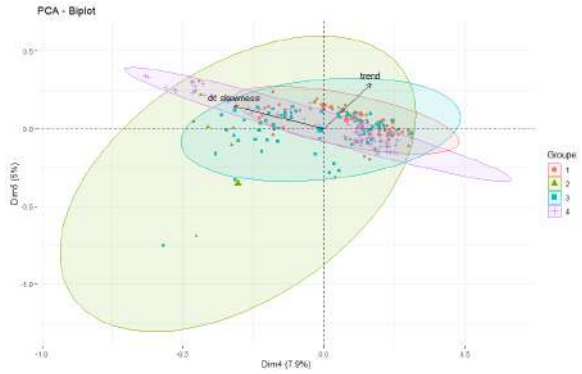


FIGURE 4.21 – Biplot des axes 4 et 5 de l'ACP sur les indicateurs statistiques, avec des groupes formés par un MMG de géométrie VEV à $k = 4$

L'analyse des groupes est assez similaire à celle des groupes formés par le K-means à $k = 5$. Pour comprendre la tendance baissière du groupe 2 ici, il faut "remonter" jusqu'aux axes 3, 4 et 5 pour lesquels ce groupe est bien représenté et présente des valeurs faibles de "trend" et des valeurs hautes de "dc skewness" : c'est à dire une tendance décroissante et des queues inférieures de distributions lourdes. Ainsi :

- le groupe 1 caractérise des fonds présentant une autocorrélation temporelle (axe 1), une distribution plutôt aplatie (axe 2) et une queue supérieure plutôt lourde (axe 3)
- le groupe 2 se caractérise par des fonds à tendance baissière comme expliqué plus haut (axes 3, 4 et 5)
- le groupe 3 se caractérise par des fonds présentant une autocorrélation temporelle (axe 1) et une distribution plutôt pointue en son centre (axe 2)

- le groupe 4 correspond à des fonds plutôt non linéaire et indépendants dans le temps.

Bilan du clustering sur indicateurs statistiques

De manière générale, le clustering de séries temporelles de nature financières réduites à des indicateurs statistiques semble être peu discriminant en terme de forme.

Une explication peut être que la représentation se fait sur les données centrées réduites sur de grandes amplitudes temporelles : par exemple on constate systématiquement une chute des cours en mars 2020 suivi d'une remontée plus grande post premier confinement pour la majorité des fonds. Le minimum et le maximum des cours sont donc souvent constatés sur cette période (le skewness et le kurtosis en sont donc affectés) et donc la représentation en terme de centrage réduction s'en retrouve biaisée, C'est pourquoi les indicateurs en "dc" ont fourni de meilleurs résultats. Nous ne pouvons pas pour autant supprimer cette période : tous les fonds n'ont pas chutés puis remontés, certains ont simplement chutés sans remonter par la suite par exemple, et l'enjeu du clustering revient alors à distinguer ces cas. Une représentation en base 100 des cours des fonds à été effectuée également en parallèle de celle en centrage-réduction de manière à essayer de s'affranchir de cette limite, mais celle ci ne parvenait pas non plus à gommer cet effet. De manière générale, les indicateurs statistiques, résumant la structure de la série temporelle, ont été choisis initialement [50] sur des bases de données de séries temporelles de natures financières et non-financières : le choix de ces indicateurs n'a pas été optimisé pour des séries temporelles de nature financière. Les particularités des séries temporelles financières sont qu'elles sont très longues et assez chaotiques : la notion même de similarité/dissimilarité devient alors problématique dans les espaces en grandes dimensions et certains algorithmes de clustering peuvent alors échouer (comme la RF précédemment par exemple)

Le tableau ci-dessous résume les différents clusterings effectués sur les indicateurs statistiques :

Algorithme	Nombre de cluster	Caractérisation de la forme	Caractérisation sur l'ACP
CAH	3	1. légère tendance haussière	1. Instable, autocoréllation,
		2. volatilité plus élevée	2. Instable, autocoréllation, volatile
		3. Dynamique faible	3. Tendance plus élevée, période élevée , stabilité, peu autocoréllé
K-means	5	1. NA	1. Autocoréllation, distribution aplatie, queue supérieure lourde
		2. NA	2. Autocoréllation, distribution pointue en son centre
		3. NA	3. Indépendance temporelle, mais linéaire en moyenne
		4. NA	4. Indépendance temporelle et non-linéaire en moyenne
		5. décroissance lente mais sûre	5. Autocorrélation, distribution aplatie, queue inférieure lourde
RF	2	1. NA	1. Autocoréllé et linéaire
		2. NA	2. Indépendance temporelle et non linéaire
MMG	4	1. NA	1. Autocoréllation, distribution aplatie, queue supérieure lourde
		2. Tendance baissière historique	2. Tendance nulle voire décroissante, queue inférieure lourde
		3. Fonds les plus volatiles	3. Autocoréllation et distribution pointue en son centre
		4. NA	4. Indépendance temporelle et non-linéaire en moyenne

FIGURE 4.22 – Résumé des clusterings des séries temporelles des fonds UC et euros réalisés sur les indicateurs statistiques proposés par Wang et al (2006) [50]

Le couple CAH/distance euclidienne est donc choisi pour représenter d'un point de vue statistique les groupes homogènes de fonds UC et euros : il présente la meilleure

caractérisation en terme de forme et en terme de caractérisation sur les axes de l'ACP. Elle présente l'avantage de fournir un nombre réduit ($k = 3$) de groupes distincts :

- cluster 1 : les fonds à légère tendance haussière sur l'historique mais instable. Nous nommons ce cluster "investissement risque modéré"
- cluster 2 : les fonds à volatilité plus élevée sur l'historique. Nous nommons ce cluster "investissement risqué"
- cluster 3 : les fonds stables peu dynamiques mais à avec une tendance haussière marquée sur l'historique. Nous nommons ce cluster "investissement peu risqué"

On remarque également que "l'angle de classification" est différent selon les algorithmes employés : nombre de clusters différents, analyse différente (et parfois similaire aussi), découpe des groupes différentes dans l'espace de l'ACP etc... Cela prouve la nécessité d'essayer des algorithmes différents en apprentissage non supervisé : les différents algorithmes ayant par définition des structures différentes, cela va induire un angle de classification selon des indicateurs discriminants différents.

4.2.2 Indicateurs "financiers"

voici un extrait de la base de données des indicateurs financiers :

	moy_return_daily	moy_return_monthly	moy_return_quaterly	moy_return_yearly	sigma_daily	sigma_weekly	sigma_monthly	sigma_quaterly	sigma_yearly	ratio_sharpe_TME	ratio_sharpe_1	proportion_augmentation
080000	5.651093e-04	9.028094e-03	2.739206e-02	1.011429e-01	1.222071e-02	2.626274e-02	0.0432263039	0.0815984074	0.131671103	9.608975e-03	0.0775876589	0.5494839
0D0000	1.044407e-04	1.668368e-03	4.786712e-03	1.419114e-02	2.247517e-03	6.637821e-03	0.0138658905	0.0242277614	0.038753848	-1.349402e-02	0.0140229597	0.5822021
0E0000	1.964330e-03	2.217812e-02	6.544047e-02	1.961288e-01	2.633642e-02	4.985128e-02	0.0818848591	0.1399339622	0.188823118	3.666149e-02	0.0558252288	0.5767045
0F0000	1.200424e-04	1.819164e-03	5.444803e-03	1.391053e-02	3.713089e-03	1.021975e-02	0.0221572489	0.0442308313	0.029830442	-6.321451e-03	0.0095472557	0.6184874
0H0000	8.245199e-04	3.686198e-03	1.050682e-02	3.183458e-02	2.509287e-03	2.425892e-03	0.0042735191	0.0107737172	0.019787807	8.710377e-01	0.9098903140	0.8223684

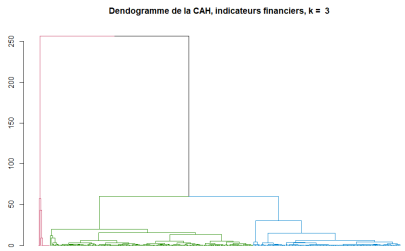
FIGURE 4.23 – Extrait de la base de donnée contenant les indicateurs financiers des fonds UC et euros. Toutes les variables ne sont pas visibles

Par la suite, pour chaque couple distance/algorithme, nous retenons le nombre k de clusters qui maximise "l'interprétabilité" en terme :

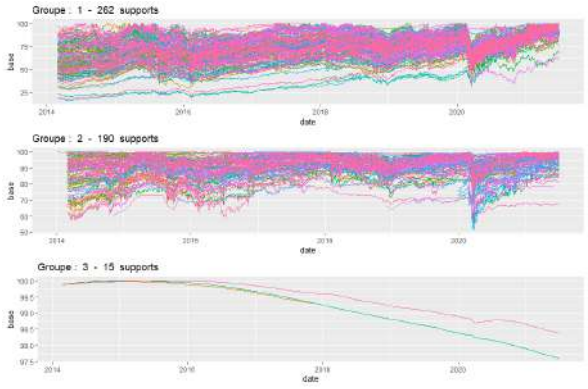
- de forme des clusters intra-groupe,
- de la bonne distinction inter-groupe,
- de la distinction des groupes sur la représentation de ceux-ci sur les axes d'une ACP

Distance euclidienne + CAH

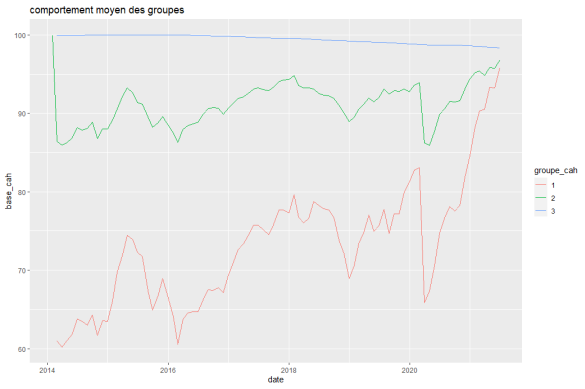
Les séries temporelles des supports sont affichées cette fois-ci en pourcentage relatif par rapport à la valeur liquidative maximum observée pour chaque série temporelle afin de pouvoir les comparer. La Méthode de Ward pour le calcul inter-classe est utilisée et nous retenons un nombre de cluster $k = 3$:



[a]



[b]



[c]

FIGURE 4.24 – Dendogramme (a), comportement des valeurs liquidatives des fonds par cluster (b), et comportement moyen des clusters (c), pour une distance euclidienne + CAH et 3 clusters, sur les séries temporelles résumées en indicateurs financiers.

Les indicateur financiers semblent fournir une meilleure distinction que les indicateurs statistiques en terme d'allure générale des clusters formés. Le cluster 1 présente une forte tendance haussière, le cluster 2 une tendance quasiment nulle, et le cluster 3 une légère tendance baissière. L'impact Covid19 est visible sur les clusters 1 et 2. Le cluster 1 semble plus volatile que le cluster 2 car présente de plus fortes baisses/hausses relatives aux mêmes dates.

Plus finement, avec une ACP :

Les deux premiers axes expliquent respectivement 47.8% et 15.6% de la variance, soit un total de 63.4% de la variance des données :

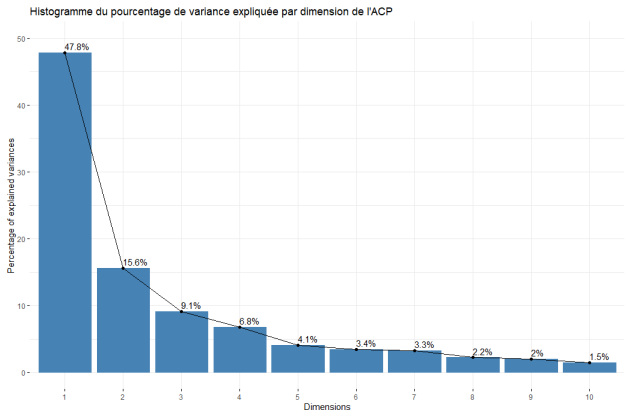


FIGURE 4.25 – Pourcentage de variance expliqué par les dimensions de l'ACP sur les indicateurs financiers

La contribution des variables (indicateurs financiers) à la construction des axes est donnée par le graphique suivant :

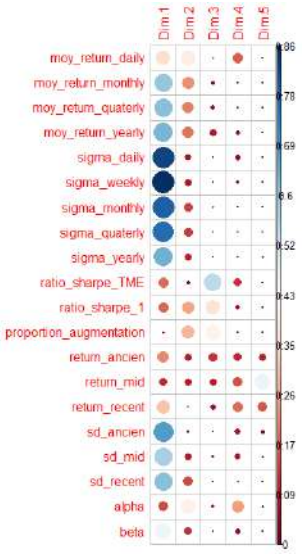


FIGURE 4.26 – Contribution des variables à la construction des dimensions de l'ACP sur les indicateurs financiers

Le premier axe est donc construit par les indicateurs de volatilités et de rendements, et le second par les indicateurs de performance ("proportion augmentation", "alpha", "ratio de Sharpe 1" et "moy return daily").

En représentant les individus et les variables sur le même graphique :

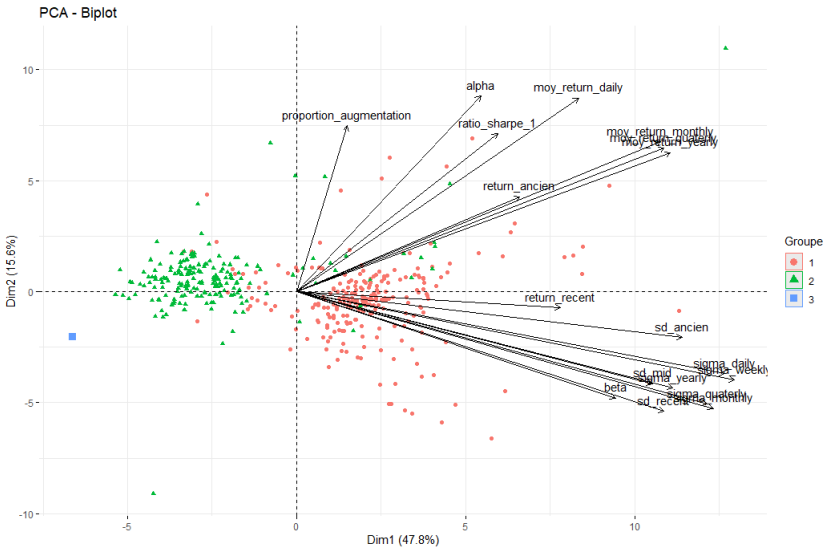


FIGURE 4.27 – Biplot des deux premiers axes de l'ACP sur les indicateur financiers des séries temporelles des valeurs liquidatives des fonds UC et euros

Le cluster 1 est opposé aux cluster 2 et 3 sur le premier axe. Le cluster 1 présente donc des fonds ayant une plus forte tendance haussière et une plus forte volatilité que les cluster 2 et 3. Le cluster 3 est totalement caractérisé par une tendance faible voire négative. Ces résultats sont vérifiés par les graphiques de comportements des clusters obtenus plus haut. L'axe 2 n'apporte pas plus d'informations sur une différenciation entre les cluster 1 et 2.

Le clustering distance euclidienne + CAH différencie les fonds uniquement sur le couple rendement/risque des fonds UC et sépare les fonds UC des fonds euros (cette information est déja disponible dans nos données d'arbitrages car nous avons la proportion d'UC et d'euros pour chaque pas de temps pour chaque contrat). Il ne tient pas compte des performances relatives comme le ratio de Sharpe car les clusters ne sont pas bien distincts sur l'axe 2, ni même du couple rendement/risque des périodes "récentes", "mid" et "ancien" pour la même raison. Ce clustering proposé n'est donc pas pleinement satisfaisant.

Nous avons donc essayé de forcer la CAH a détecter des différences entre individus sur ces indicateurs. Sur les 20 indicateurs financiers, les indicateurs "ratio Sharpe TME", "ratio Sharpe 1", "proportion augmentation", "alpha", "beta", "return ancien", "return mid", "return recent", "sd ancien", "sd mid", et "sd recent" ont donc été retenus. Ils n'ont cependant pas fourni de meilleurs résultats (voir annexe "Résultats du clustering CAH + distance Euclidienne sur une sélection d'indicateurs financiers") et le clustering obtenu ne sera pas détaillé.

Distance euclidienne + K-means

Les séries temporelles des supports sont affichées en pourcentage relatif par rapport à la valeur liquidative maximum observée pour chaque série temporelle afin de pouvoir les comparer.

Les nombres k de clusters testés pour les K-means vont de 2 à 10. Pour choisir le nombre k optimal, le graphique des silhouettes moyennes est tracé et nous choisissons k tel que la silhouette moyenne est maximisée :

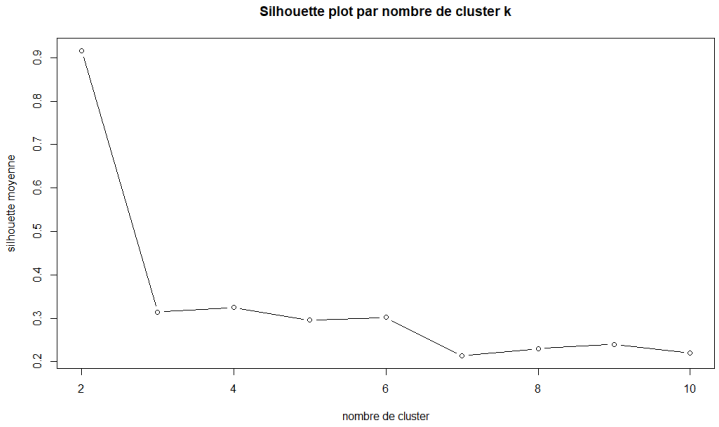


FIGURE 4.28 – Silhouette plot des K-means pour des nombres de cluster allant de 2 à 10, sur les indicateurs financiers sélectionnés

Choisir deux groupes n'est pas intéressant pour une analyse. Il est alors retenu 4 clusters pour lesquels la silhouette moyenne est de 0.46 :

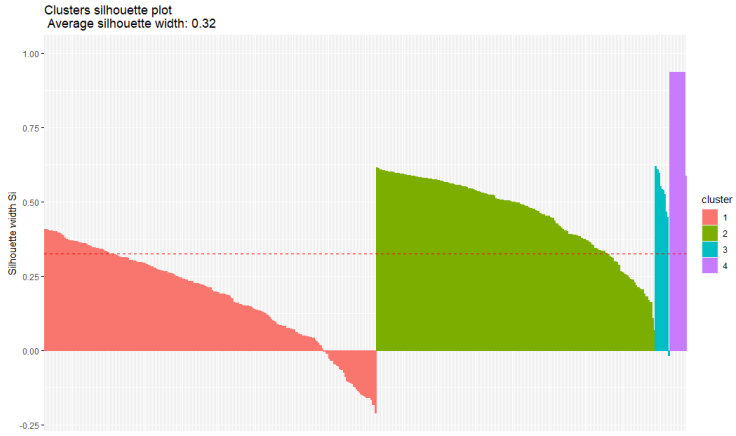


FIGURE 4.29 – Silhouette plot pour $k = 4$, sur les indicateurs financiers sélectionnés

La silhouette moyenne du groupe 1 est de 0.18 : cela est insatisfaisant d'autant plus

qu'elle présente des silhouettes négatives. Les silhouettes des groupes 2, 3 et 4 sont de 0.46, 0.50 et 0.91 : une analyse est possible.

Voici les séries temporelles affichées en pourcentage relatif par rapport à la valeur liquidative maximum observée pour chaque série temporelle par groupe résultant de ce clustering :

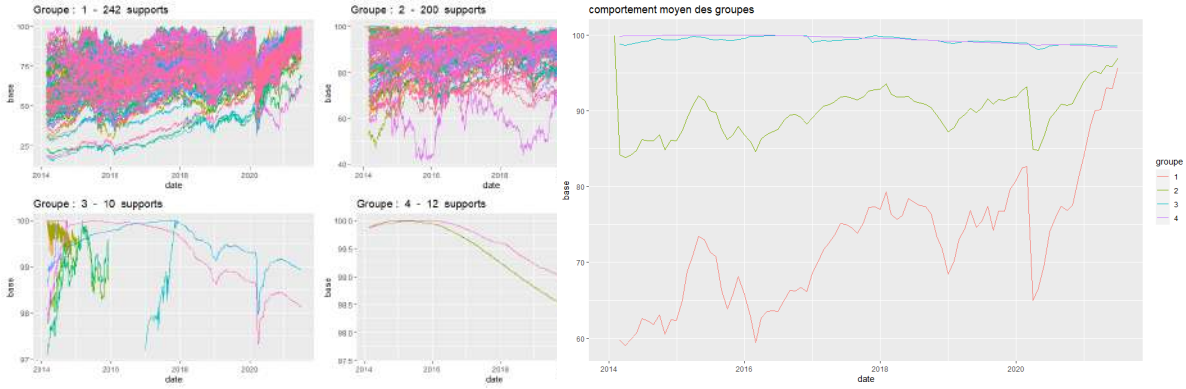


FIGURE 4.30 – Représentation des comportements des fonds UC et euros par un clustering obtenu par un K-means $k = 4$ sur les indicateurs financiers sélectionnés

En terme d'allure générale des clusters, ce clustering n'est pas satisfaisant : les groupes 4 et 3 seraient à fusionner (ils restent cependant distincts dans le cas $k = 3$ qui n'est pas montré ici pour des raisons de taille de mémoire) car ils semblent regrouper des séries temporelles en forme de "cloche". De plus, les fonds ne sont pas répartis en nombre de séries temporelles par groupe de manière homogène.

Les groupes 1 et 2 se distinguent respectivement pour leur tendance nulle et haussière. Le groupe 2 semble présenter plus de volatilité que le groupe 1.

Regardons l'ACP pour essayer d'expliquer la distinction entre les groupes qui reste assez obscure pour certains d'entre eux (la construction des axes est visible en annexe "Résultats du clustering CAH + distance Euclidienne sur une sélection d'indicateurs financiers", figure 5.3) :

La seule différence avec le clustering précédant (mise à part la restriction à un nombre réduit d'indicateurs financiers qui n'impacte pas la caractérisation des axes) s'observe à l'extrême "ouest" de l'axe 1 : le K-means à $k = 4$ cluster fait une distinction entre les fonds avec une volatilité et un rendement très faibles voir négatifs (cluster 3) et ceux avec une volatilité et un rendement faible (cluster 2). Cette ACP isole le groupe 4 qui est caractérisé par une proportion d'augmentation très faible, une sous-performance marquée par rapport au TME, et des rendements faibles voir négatifs (cela est vérifié par les graphiques précédents). Le cluster 1, plus dynamique et présentant un meilleur rendement car bien représenté positivement sur l'axe 1 vient vérifier les graphiques précédents. Enfin, la distinction entre le groupe 3 et 4 qui présentent visuellement une tendance baissière se fait sur la variable "proportion augmentation" : le cluster 4 ne fait que chuter tandis que le cluster 3 présente une forme en cloche plus marquée qui induit donc une proportion

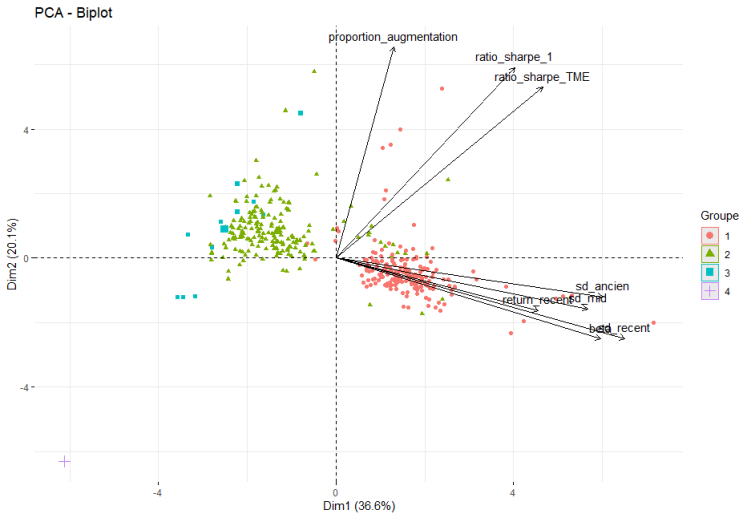


FIGURE 4.31 – Biplot de l'ACP sur une sélection d'indicateurs financiers, avec un clustering K-means $k = 4$ et distance euclidienne

d'augmentation plus élevée (partie gauche de la cloche haussière) que le cluster 3. Nous décidons donc d'enlever la variable "proportion augmentation" qui conduit à séparer les séries temporelles en clusters peu pertinents (axe 2 peu discriminant) sur une partie minoritaire en nombre des séries temporelles.

RF non supervisée + CAH

Une RF non supervisée possédant 1000 arbres CART est entraînée sur les indicateurs financiers sélectionnés afin de calculer une matrice de similarité. Cette matrice de similarité est injectée dans une CAH en vue de former des clusters. Le nombre de clusters retenu est alors $k = 2$, car pour tous les $k \neq 2$ les groupes formés sur les axes de l'ACP n'apportaient que de la confusion en plus en terme d'interprétation :

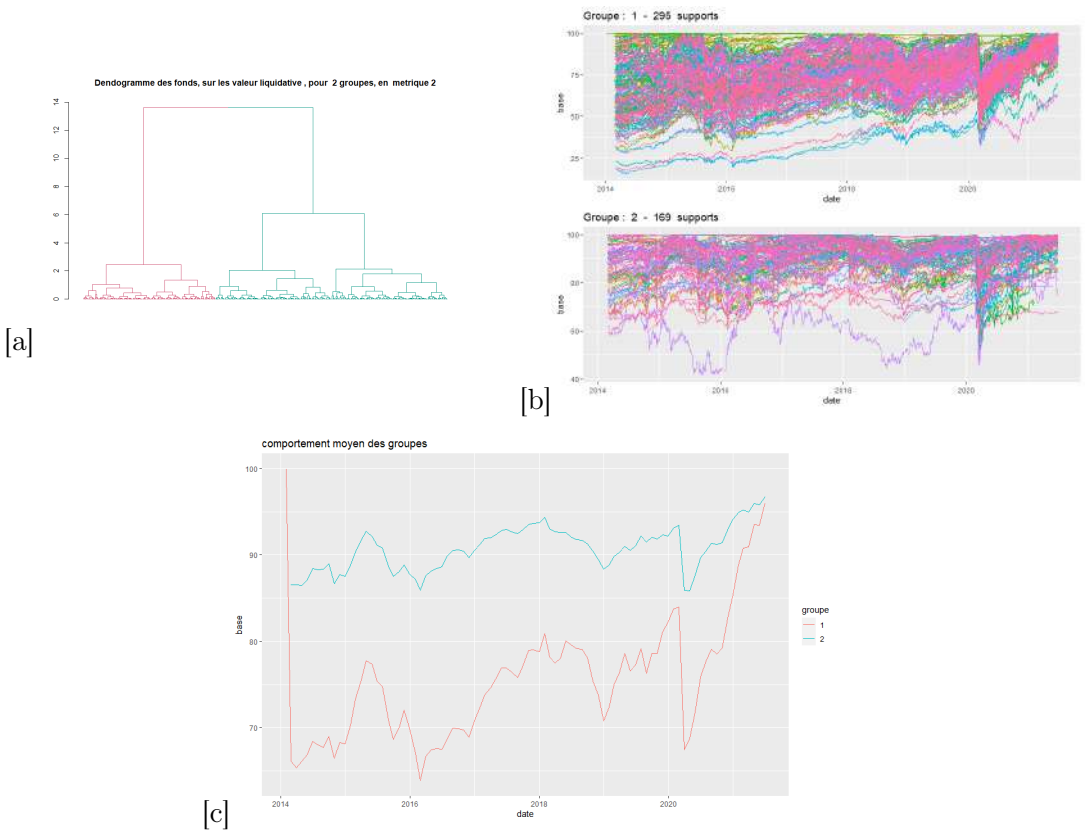


FIGURE 4.32 – Dendogramme (a), comportement des valeurs liquidatives des fonds par cluster (b), et comportement moyen des clusters (c), pour une matrice de similarité obtenue par RF non supervisée + CAH et 2 clusters, sur les séries temporelles résumées sur une sélection d'indicateurs financiers.

Puisque nous avons supprimé la variable "proportion augmentation", l'ACP construite est légèrement différente (voir annexe "Résultats de l'ACP sur une sélection d'indicateurs financiers ne comprenant pas la variable "proportion augmentation"") : les quatre premiers axes expliquent 79% de la variance des données. Les biplot obtenus de l'ACP sont les suivants :

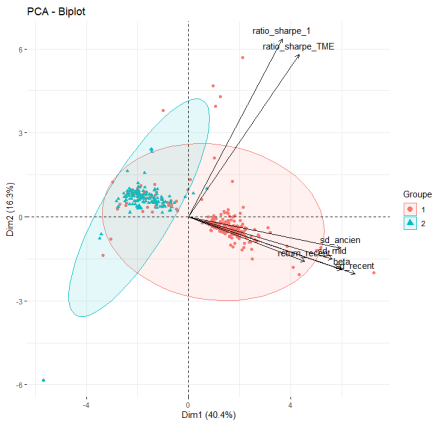


FIGURE 4.33 – Biplot de l'ACP (axes 1 et 2) sur une sélection d'indicateurs financiers, avec un clustering CAH $k = 2$ et une matrice de similarité calculée avec une RF non supervisée.

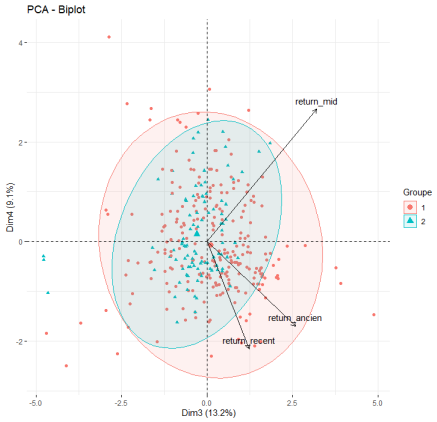


FIGURE 4.34 – Biplot de l'ACP (axes 3 et 4) sur une sélection d'indicateurs financiers, avec un clustering CAH $k = 2$ et une matrice de similarité calculée avec une RF non supervisée.

L'axe 1 caractérise, comme toujours, les fonds élastiques ("beta") avec un rendement et une volatilité élevé. L'axe 2 caractérise un peu plus les fonds sur-performant le TME et/ou un rendement de 1% annuel (les variables "ratio sharpe TME" et "ratio sharpe 1" sont d'ailleurs positivement corrélées avec les indicateurs de rendement et de volatilité sur l'axe 1 et participent à la contribution de l'axe 1). L'axe 3 réalise un focus les rendements tandis que l'axe 4 oppose les fonds présentant un rendement élevé en milieu de vie ("return mid") aux fonds présentant un rendement élevé ancien et récent ("return ancien" et "return recent").

Sur l'axe 1, les clusters 1 et 2 sont plutôt bien séparés, bien que manifestement certaines séries temporelles aient été mal classées au sens de l'axe 1. Sur tous les autres

axes, la distinction n'est pas claire. L'unique remarque peut se faire sur le biplot des axes 3 et 4, sur lequel on remarque que le cluster 1 englobe le cluster 2, et que donc les fonds du cluster 2 présentent des valeurs de rendement plus élevées en moyenne.

Ce clustering reste insatisfaisant. De manière générale, une RF non supervisée associée à une CAH sur les indicateurs testés (financiers et statistiques) de séries temporelles de nature financière semble fournir de mauvais résultats en terme d'interprétation (peu de différenciation inter groupes).

Modèle de mélange Gaussien

Comme réalisé précédemment avec les indicateurs statistiques, nous sélectionnons le modèle de mélange gaussien optimal selon le critère BIC, en utilisant l'algorithme EM. Nous choisissons le modèle et le nombre de clusters avec le BIC (Bayesian Information-Criteria) le plus grand. Nous regardons un nombre de clusters allant de 2 à 10 pour les 14 algorithmes proposés par la fonction Mclust() :

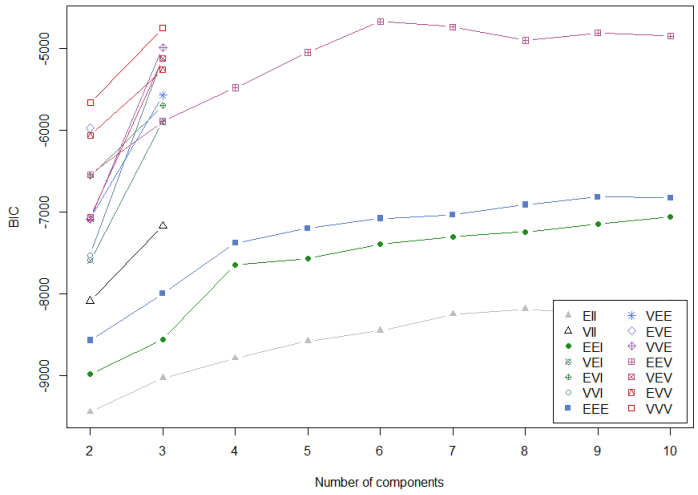


FIGURE 4.35 – Graphique BIC = f(k,géométrie MMG) du package Mclust sur le logiciel R

Il faudrait prendre les caractéristiques géométriques EEV (ellipsoïdale, de forme égale, volume égale et orientations variables) avec $k = 6$ mais cela ferait trop de classes et une confusion dans l'analyse de celles-ci. L'analyse a été faite avec ce clustering mais les groupes formés étaient trop confus au niveau de la représentation sur les axes de l'ACP Nous préférerons prendre plutôt $k = 3$ avec des caractéristiques géométriques VVV (ellipsoïdale, de forme, volume et orientations variables) qui possèdent tout de même un plus grand BIC que toutes les autres géométries peu importe le nombre de clusters $k < 6$.

Nous obtenons alors :

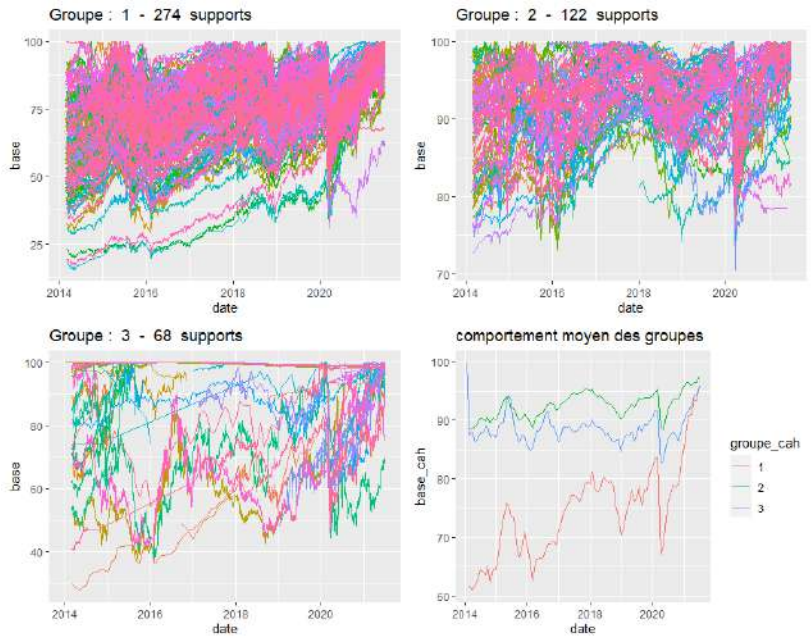


FIGURE 4.36 – Comportement des groupes formés par un MMG de géométrie VVV avec $k = 3$

L'allure générale des groupes est assez similaire : un pic à la baisse en mars 2020 (confinement Covid19) suivi d'un rebond dépassant le niveau ante-covid, et une bosse centrée en 2018. Le groupe 3 est assez chaotique.

Si l'on regarde l'ACP (dont la contribution des indicateurs à la construction des axes est la même que pour la RF précédente, voir en annexe" Résultats de l'ACP sur une sélection d'indicateurs financiers ne comprenant pas la variable "proportion augmentation"") :

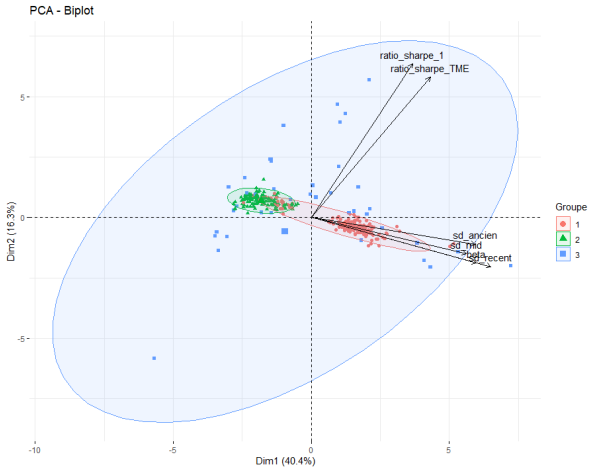


FIGURE 4.37 – Biplot des deux premier axes de l'ACP sur une sélection d'indicateurs financiers, avec des groupes formés par un MMG de géométrie VVV à $k = 3$

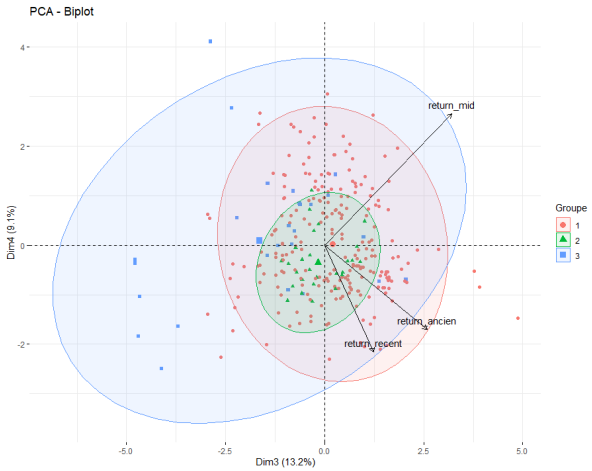


FIGURE 4.38 – Biplot des axes 3 et 4 de l'ACP sur une sélection d'indicateurs financiers, avec des groupes formés par un MMG de géométrie VVV à $k = 3$

Le cluster 1 caractérise les fonds à forte volatilité et élasticité car ce groupe est bien représenté positivement sur l'axe 1. Le cluster 2 caractérise à l'inverse les fonds à moindre volatilité et élasticité car ce groupe est représenté négativement sur l'axe 1. Le cluster 3 est lui caractérisé par une représentation globalement négative sur l'axe 3 et présente donc des fonds à rendements faibles sur leur milieu temporel de leur historique.

Ce clustering ne convient donc pas : il y a trop de confusions dans l'identification des groupes (ACP et allure générale)

Bilan du clustering sur indicateurs financiers

De manière générale, le clustering de séries temporelles de nature financières réduites

à des indicateurs financiers est plus discriminant en terme d'allure générale des clusters, mais moins interprétable avec les ACP. Seul le premier clustering proposé est convenable, les autres étant soit non interprétables soit dénués de sens , mais se limite à une interprétation simpliste :

- cluster 1 : fonds à une tendance haussière et à une volatilité forte, et élastique par rapport au CAC40, qui présentent donc une allure générale à la hausse. Nous les nommerons les "fonds dynamiques".
- cluster 2 : des fonds présentant des caractéristiques plus modérés que ceux du cluster 1, présentant donc une tendance plus aplatie. Nous les nommerons les "fonds modérés".
- cluster 3 : les fonds à tendance faibles voir négatives. Nous les nommerons les "fonds de sécurisation".

Au final, les cluster obtenus ne se distinguent principalement que par le couple rendement/risque, leur élasticité par rapport au marché de référence du CAC40, et leur tendance globale.

4.3 Temporal-proximity based clustering

L'objectif avec cette méthode est de regrouper les séries temporelles avec les mêmes dynamiques/formes à l'aide de la distance DTW sur les séries temporelles des valeurs liquidatives des fonds UC et euros, c'est à dire de regrouper avec un "oeil géométrique". L'idée est de se dire qu'un assuré ne voit pas une courbe avec des chiffres et des concept mathématiques plus ou moins compliqués (rendement journalier, kurtosis, autocorrélation etc...), mais plutôt en terme de niveau de risque perçu et associé à une dynamique/forme de série temporelle des valeurs liquidatives des supports de son contrat d'assurance vie.

Il est alors décidé de réaliser cette approche de clustering sur les série brutes en lissant légèrement les courbes des valeurs liquidatives des fonds UC et euros en prenant la moyenne des valeurs liquidatives sur les 30 derniers jours en chaque date : l'idée consiste à gommer les "pics" sucessifs liés à la nature financière de ces séries temporelles afin d'en dégager une allure générale plus représentative (comme si l'on devait "résumer" une courbe en un coup d'oeil).

Distance DTW + K-means

Nous calculons la distance entre chaque série temporelle à l'aide de la distance DTW. Une fois ces distances obtenues pour chaque couple de série temporelle, nous appliquons un K-means pour un nombre de groupe k allant de 2 à 8. Pour chaque k nous obtenons les indicateurs de validité des clusters (CVI : Cluster Validation Indices) résumés sur le graphique suivant :

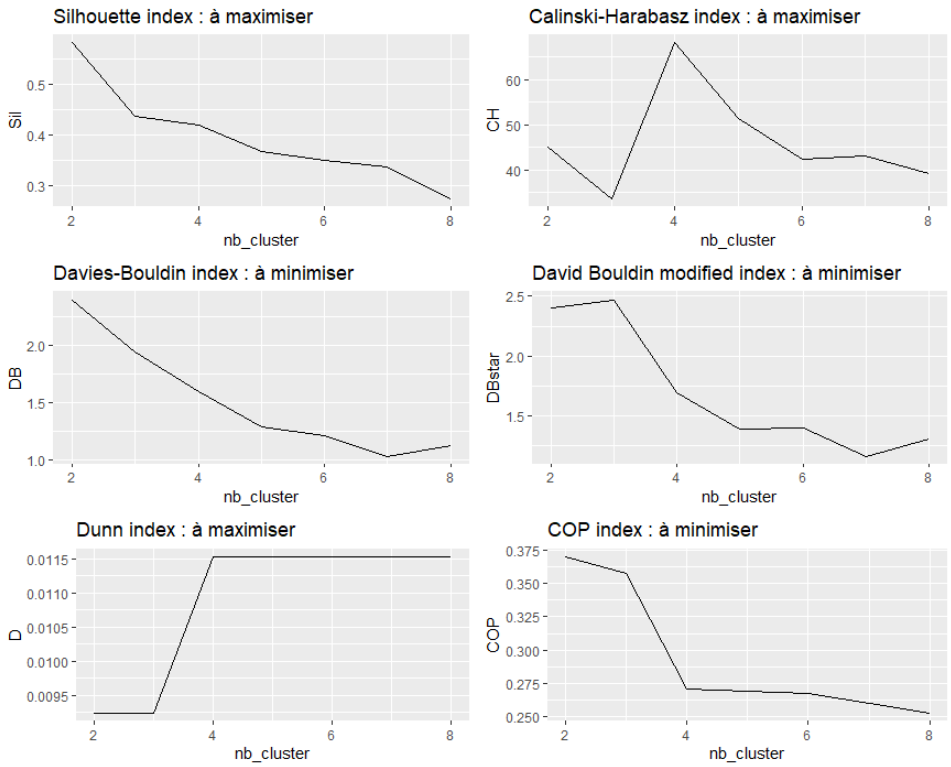


FIGURE 4.39 – Graphique des différents CVI sur des K-means appliqués sur les distances DTW des séries temporelles

Les CVI sont des indicateurs de la qualité de la formation des clusters. Ils ne seront pas détaillés mathématiquement dans ce mémoire. Les CVI utilisés sont : l'indice de silhouette [38], de Calinski-Harabasz [44], de Davies Bouldin [15], de Davies Bouldin modifié [33], de Dunn [24], et de COP [23].

Nous retenons donc 4 clusters à la vue des CVI, et cela nous donne les clusters suivants :

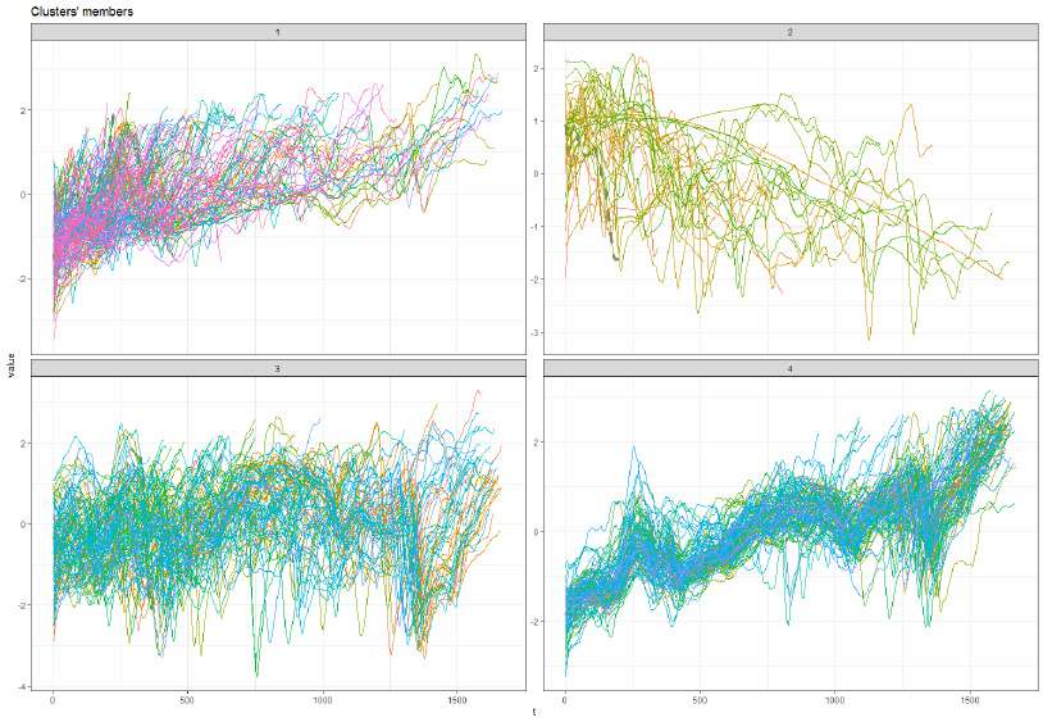


FIGURE 4.40 – Clustering $k = 4$ avec distance DTW couplée à un K-means, réalisé sur les séries temporelles lissées des valeurs liquidatives des fonds UC et euros

Les cluster 1 à 4 sont composés respectivement de 170, 53, 117, et 122 fonds UC et euros. Ils présentent une distance DTW intra groupe moyenne de 99.8, 191.9, 271.1 et 154.2.

La forme de chaque groupe est bien distinct :

- cluster 1 : tendance globale haussière, avec une certaine concavité. Impact Covid quasiment nul sur la tendance.
- cluster 2 : tendance générale baissière et impact Covid nul. Plus grosse densité de fonds dans les dates anciennes : ce fond possède beaucoup de fonds clôturés/absorbés avant 2017 qui présente une forme en "∩"
- cluster 3 : tendance nulle mais grosses fluctuations autour de celle-ci. Impact Covid fort. Il s'agit du groupe possédant la moins grande inertie intra-classe du fait de la nature volatile des fonds le composant.
- cluster 4 : tendance générale haussière. Impact Covid fort. L'inertie intra classe est bonne car les fonds composant ce groupe fluctuent de la même manière (densité en la moyenne forte) : creux au niveau du premier confinement Covid19, rebond post Covid19 immédiatement après, pic suivi d'une chute en mi-2015 (le "krach chinois" [31]), chute début 2019 (guerre commerciale entre les US et la Chine [10]).

Ces fonds sont donc *drivés* essentiellement par l'actualité économique : tendance haussière des marchés avec des "singularités" de temps à autres liées à l'actualité économique.

Distance DTW + CAH

Nous calculons la distance entre chaque série temporelle à l'aide de la distance DTW. Une fois ces distances obtenues pour chaque couple de série temporelle, nous appliquons dessus une CAH pour un nombre de groupe k allant de 2 à 8. Pour chaque k nous obtenons les indicateurs de validité des clusters (CVI : Cluster Validation Indices) résumés sur le graphique suivant :

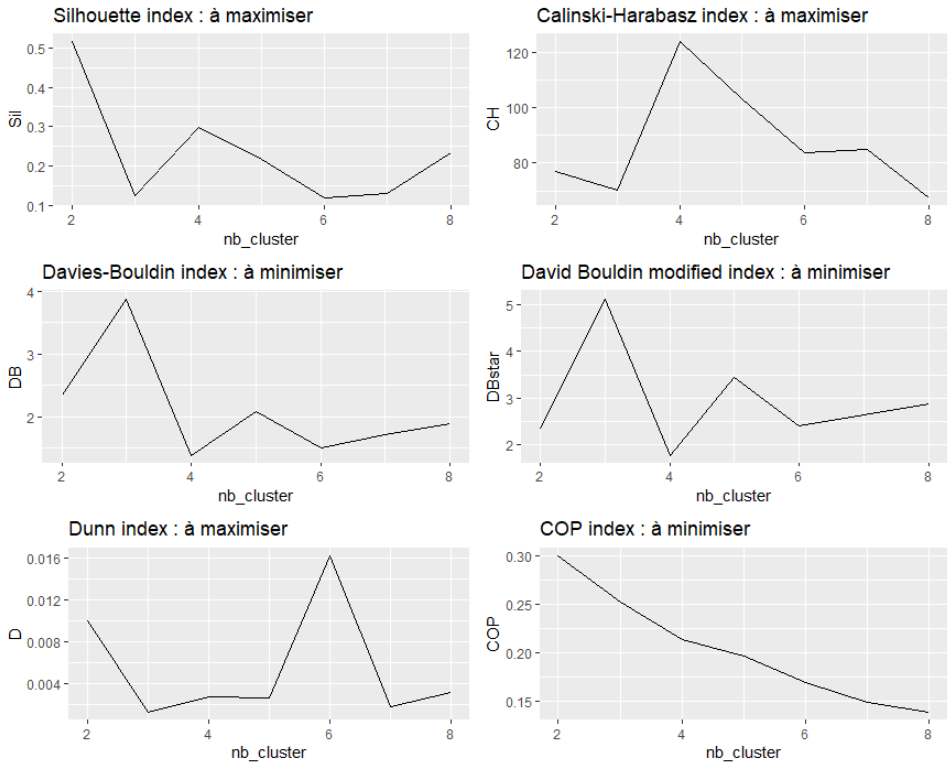


FIGURE 4.41 – Graphiques des différents CVI sur des CAH appliquées sur les distances DTW des séries temporelles

Nous retenons donc 4 clusters à la vue des CVI, et cela nous donne les clusters suivants :

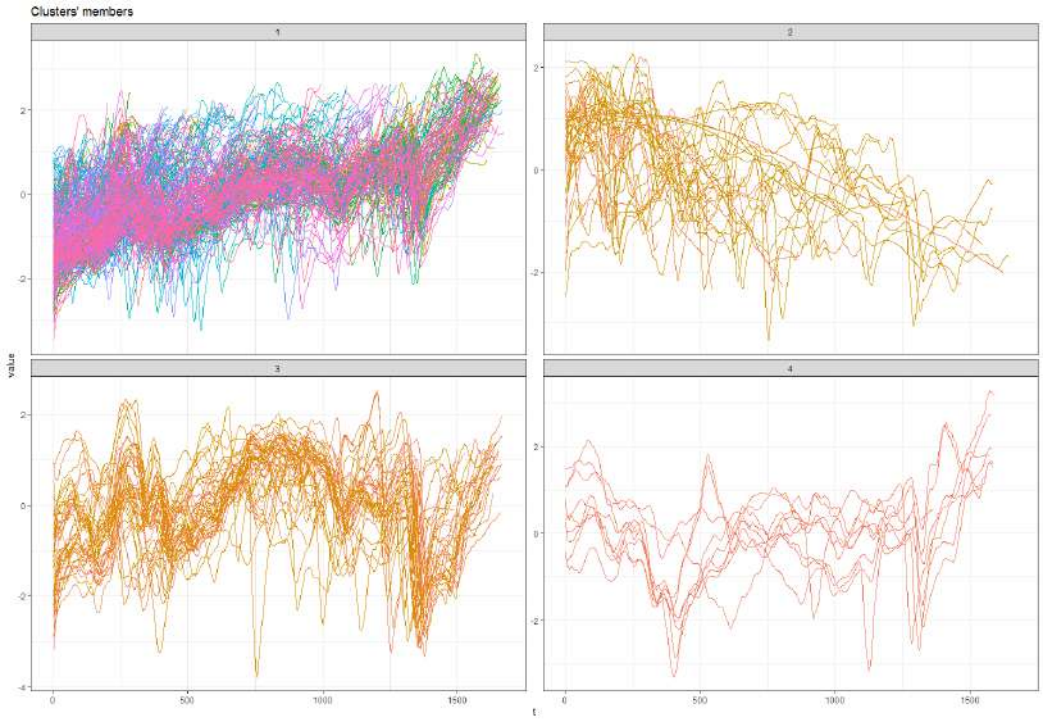


FIGURE 4.42 – Clustering $k = 4$ avec distance DTW couplée à une CAH, réalisé sur les séries temporelles lissées des valeurs liquidatives des fonds UC et euros

Clairement, les groupes formés sont moins homogènes que ceux réalisés en DTW + K-means. Les cluster 1 à 4 comportent respectivement 355, 55, 43, et 9 fonds UC et euros et présentent une distance DTW moyenne intra-classe de 182.9, 219.7, 352.6 et 368.8. Le cluster 1 compte trop de fonds en son sein par rapport aux autres cluster et celui-ci perd en inertie intra classe et en interprétation par conséquence. Les autres clusters formés possèdent des formes assez similaires mais moins bonnes que pour ceux en DTW + K-means. Les allures générales restent néanmoins bien meilleures que celles obtenues par approche *characteristic based clustering*.

Bilan du clustering sur les séries brutes

Voici un tableau récapitulatif des deux clusterings sur les séries brutes proposés :

Couple distance/algorithme	Nombre de cluster	Taille des cluster	Distance intra-groupe	Somme des distances
DTW + CAH	4	355	182.9	1124
		55	219.7	
		43	352.6	
		9	368.8	
DTW + K-means	4	170	99.8	717
		53	191.9	
		117	271.1	
		122	154.2	

FIGURE 4.43 – Tableau récapitulatif des clustering avec la distance DTW

Nous retenons le couple distance DTW + K-means pour l'obtention des cluster en méthodologie *temporal-proximity based clustering* : il présente la somme de distances intra-groupe la plus faible et une meilleure répartition des fonds. Le clustering obtenu est le suivant :

- cluster 1 : les fonds à tendance haussière peu impactés par la conjecture économique. Nous les nommerons "fonds dynamiques insensibles à la conjecture économique"
- cluster 2 : les fonds à tendance baissière impactés majoritairement par le contexte de taux décroissant. Nous les nommerons "fonds à dynamique euros".
- cluster 3 : les fonds à forte volatilité. Nous les nommerons "fonds risqués"
- cluster 4 : les fonds à tendance haussière mais impactés par la conjecture économique. Nous les nommerons "fonds dynamiques sensible à la conjecture économique"

Le désavantage principal de cette méthode est le temps de calcul élevé : 2h45 de temps de calcul pour plus de 450 séries temporelles des fonds contre un temps de calcul instantané pour les algorithmes en *characteristic based clustering* (en ne comptant pas le temps de calcul nécessaire au calcul des indicateurs).

4.4 Bilan sur le clustering des fonds UC et euros

Plusieurs triplets méthode/distance/algorithme ont étés testés en vue de former des groupes homogènes de fonds UC et euros. En effet, la forte granularité des fonds nécessite de les regrouper afin de réaliser des analyses pertinentes par la suite. Ces clustering s'attachent à regrouper les fonds présentant les mêmes caractéristiques statistiques, financières, ou bien la même dynamique d'évolution sur leur historique. Les groupes de fonds UC et euros qui s'en dégagent sont donc caractérisés différemment et apportent des informations différentes selon un angle d'analyse différent :

A. *Temporal-proximity based clustering* :

   cluster 1 : les "fonds à tendance haussière et insensibles à la conjecture économique"

   cluster 2 : les "fonds à dynamique euros"

   cluster 3 : les "fonds risqués"

   cluster 4 : les "fonds dynamiques sensibles à la conjecture économique"

B. *Characteristic based clustering* :

   i. Indicateurs statistiques :

      cluster 1 : les fonds "investissement risque modéré" caractérisés par une légère tendance haussière mais instable.

      cluster 2 : les fonds "investissement risqué" uniquement caractérisés par une volatilité élevée.

      cluster 3 : les fonds "investissement peu risqué" caractérisés par une stabilité et peu de dynamisme, mais avec une tendance haussière marquée.

   ii. Indicateurs financiers :

      cluster 1 : les "fonds dynamiques" caractérisés par une volatilité forte, et une élasticité par rapport au CAC40, qui présentent donc une tendance générale haussière

      cluster 2 : les "fonds modérés" caractérisés comme le cluster 1, mais avec des valeurs plus modérées.

      cluster 3 : les "fonds de sécurisation" caractérisés uniquement par une tendance faible voire baissière.

Pour rappel, nous avons accès à la PM par fonds UC et euros à la maille $contrat \times mois$. En "injectant" les différents clusters construits dans la base de données des arbitrages, nous avons donc accès à la proportion de PM correspondant à chaque groupe (selon les différentes méthodes et indicateurs) à la maille $contrat \times mois$. Cette information sera injectée dans les modèles de prédictions des taux d'arbitrages en vue de quantifier l'importance ou non de la dynamique des fonds UC et euros sur les taux d'arbitrages en gestion libre.

Chapitre 5

Statistiques descriptives et modèles de prédictions

5.1 Statistiques descriptives

Dans ce chapitre, la base de données des arbitrages enrichie des variables explicatives décrites plus haut va être étudiée en vue d'appréhender les caractéristiques du portefeuille de contrats d'assurance vie en mode de gestion libre. Dans un premier temps, des statistiques descriptives sont présentées. Dans un second temps, les différentes variables explicatives sont mises en face des taux d'arbitrages observés (les taux EUR_UC, UC_UC, et UC_EUR).

5.1.1 Caractérisation du portefeuille étudié

Le graphique ci-dessous présente la répartition en PM des fonds euros et UC, ainsi que le nombre de contrats, avec un pas de temps mensuel :

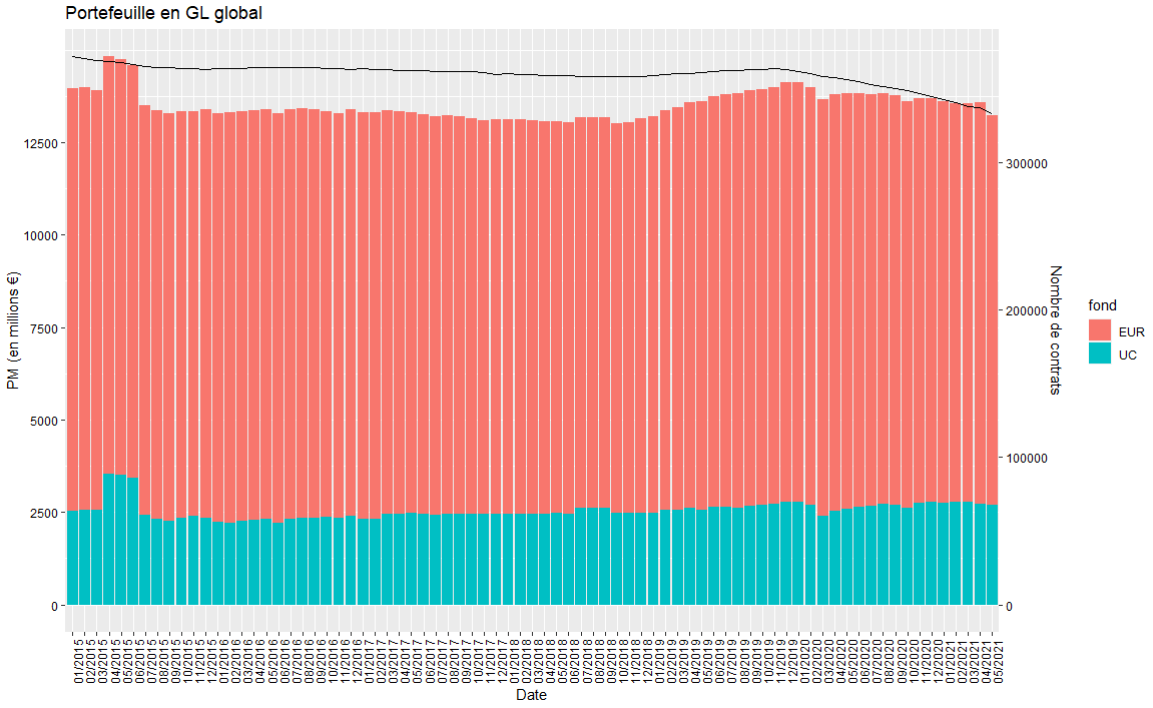


FIGURE 5.1 – Evolution de la PM et du nombre de contrats sur l'historique d'observation du portefeuille en GL étudié

La PM est relativement stable au fil du temps et se situe aux alentours des 13.4 milliards d'euros. La PM est de 13.698 milliards d'euros fin décembre 2020. Le nombre de contrats est en constante baisse sur l'historique de temps, passant de 371 365 contrats en janvier 2015 à 332 697 en mai 2021 : cela est la conséquence d'un changement de stratégie de la compagnie qui privilégie la rentabilité des contrats au nombre de contrats souscrits. La moyenne de proportion d'UC sur ce portefeuille est de 18.84%, et passe de 18.08% en janvier 2015 à 20.43% en mai 2021, dû à une recherche de rendements plus élevés de la part des assurés sur les fonds UC dans un contexte de taux décroissants, mais aussi à la mise en place de mécanismes d'incitations de re-direction de l'épargne vers les fonds UC.

Le graphique ci-dessous présente les taux d'arbitrages mensuels, toutes origines et destinations entre types de fonds confondues :

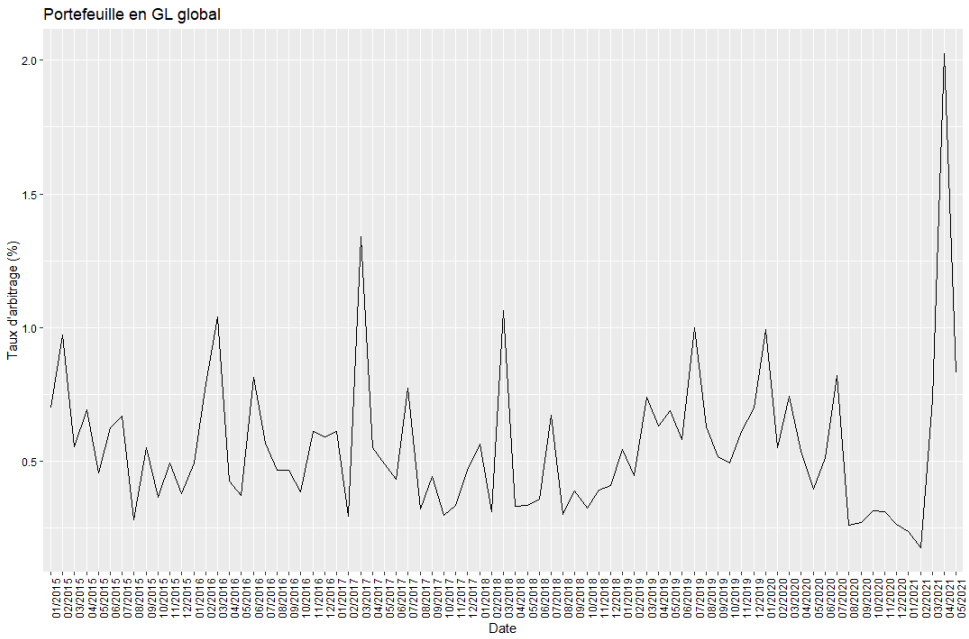


FIGURE 5.2 – Evolution du taux d'arbitrage sur l'historique d'observation du portefeuille en GL étudié

Le taux moyen d'arbitrage mensuel est de 0.5597%. Des pics et des creux d'arbitrages sont bien visibles avec un maximum de 2.0234% en avril 2021 et un minimum de 0.17702% en février 2021. De manière générale, un pic est observé entre janvier et mars : à cette période de l'année la PB est versée à l'assuré qui ensuite peut décider d'arbitrer vers un ou plusieurs fonds. De plus, ce phénomène est accentué par le fait que le premier arbitrage de l'année est gratuit pour l'assuré et sur certains produits. Ces phénomènes de taux d'arbitrages plus élevés sont donc structurels. D'autres pics sont également visibles systématiquement en juillet. Il va donc être utile de prendre en compte les mois comme variable explicative.

Le graphique ci-dessous présente les taux d'arbitrages mensuels, par origine et destination de fonds euros et UC. Un changement d'échelle de l'axe des ordonnées est opéré afin de mieux appréhender la dynamique des taux d'arbitrages EUR_UC :

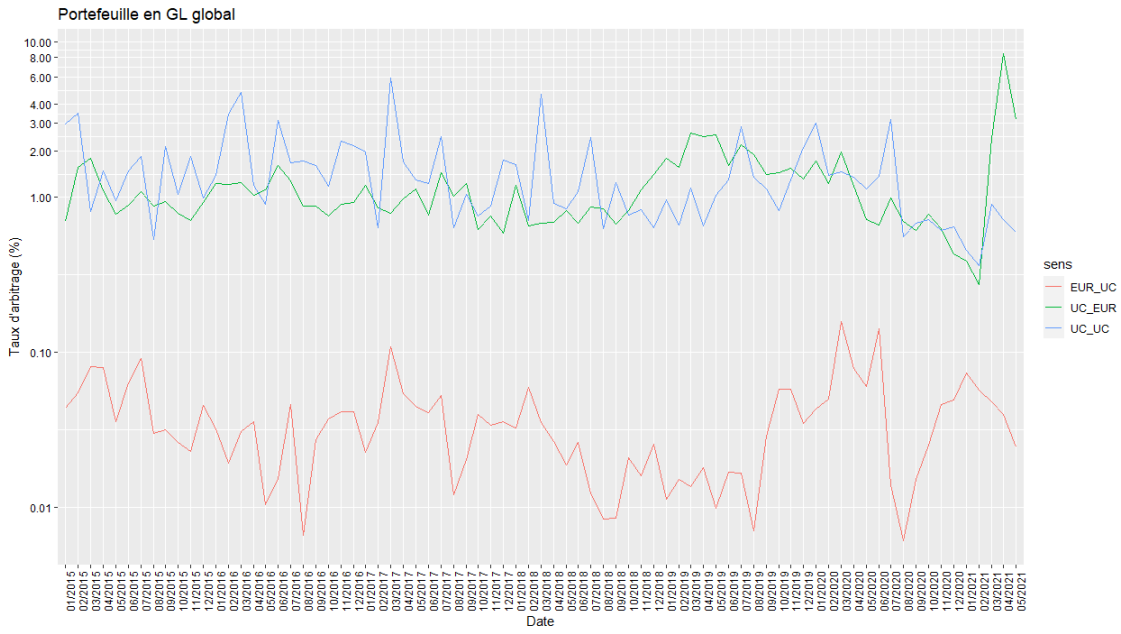


FIGURE 5.3 – Evolution du taux d'arbitrage sur l'historique d'observation du portefeuille en GL étudié, par origine et destination

Les taux EUR vers UC sont plus faibles que les autres origines/destinations, car par structure même du portefeuille il y a plus de PM sur les fonds euros. Les phénomènes de pics d'arbitrages entre janvier et mars, et juillet sont bien visibles également ici. Il est également possible d'expliquer le pic en avril 2021 grâce à ce graphique : ce pic est dû à des arbitrages massifs de l'UC vers l'euro (courbe verte), conséquence d'un retour très récent en territoire positif des taux, et donc mécaniquement un meilleur rendement des fonds euros qui deviennent donc plus attractifs. Une zone temporelle de pics d'arbitrages de l'euro vers l'UC (courbe rouge) est observée de mars à juin 2020, suivie d'un creux (le plus bas historique, à 0.006%). Cette période correspond au premier confinement de la Covid19 : les assurés voyant les cours s'effondrer ont voulu profiter de cette baisse afin d'acheter des parts d'UC peu chères avec le capital disponible sur leur fond euros, et ont stoppé dès que les cours ont rebondi post premier confinement (août 2020). Il est également intéressant de remarquer que lorsqu'il y a un pic de l'euro vers l'UC, cela se traduit généralement par un creux de l'UC vers l'euro, et inversement (vases communiquants). De même, un pic de l'euro vers l'UC correspond souvent à un pic également de l'UC vers l'UC. Ces phénomènes mettent en évidence l'effet de la conjecture économique sur les taux d'arbitrages : les assurés souhaitent maximiser leur rendement financier en désinvestissant les fonds les moins intéressants.

Les graphiques suivants montrent la répartition en PM sur les fonds des différents clusters formés dans le chapitre précédant :

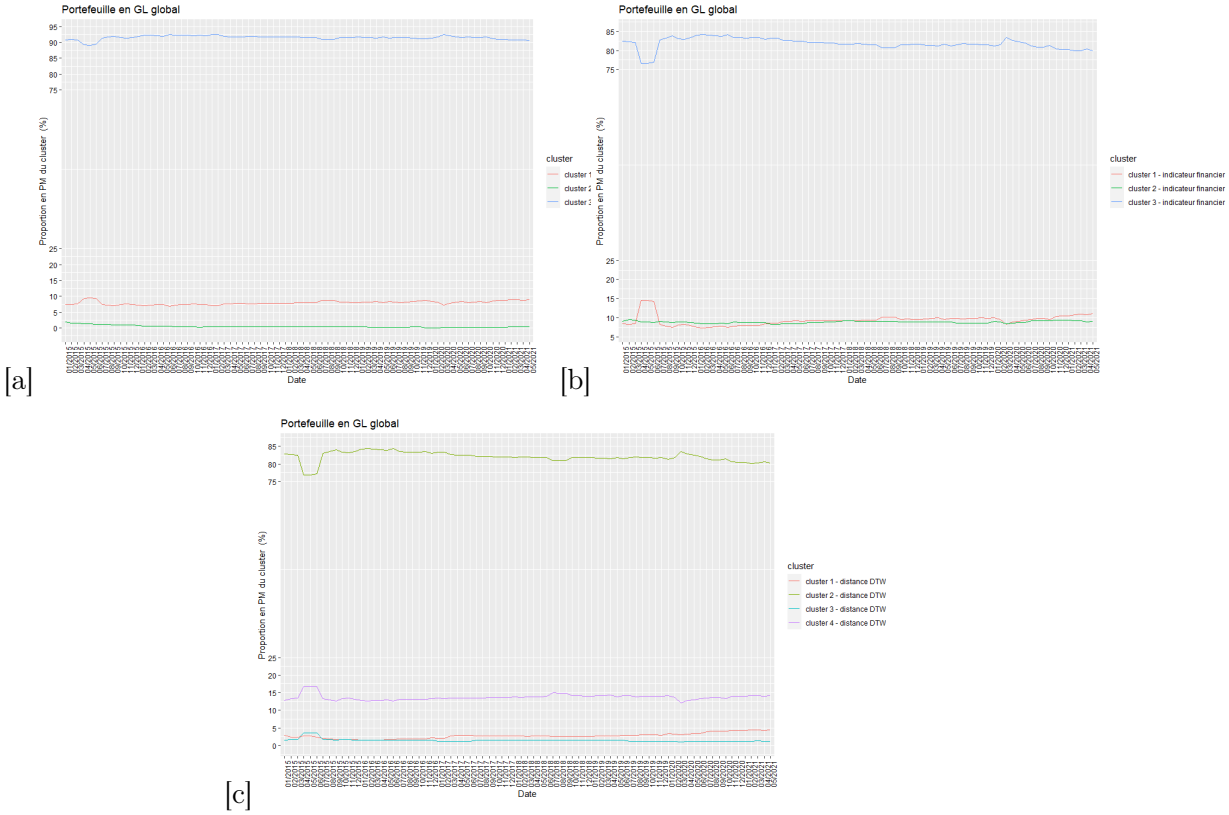


FIGURE 5.4 – Proportion de PM des clusters formés par les indicateurs statistiques(a), indicateurs financiers (b), et distance DTW (c)

Les proportions sont plutôt stables dans le temps mais présentent des aspérités. Par exemple de mars à juillet 2015 on peut remarquer que certains clusters sont désinvestis à la faveur d'autres :

- **Indicateurs statistiques** : les fonds appartenant au cluster 3 sont désinvestis à la faveur du cluster 1 uniquement. Autrement dit, les fonds caractérisés comme "peu risqués" sont désinvestis en faveur des fonds "investissements risque modéré". Cela caractérise la notion de frontière de risque accepté par l'assuré.

- **Indicateurs financiers** : les fonds appartenant au cluster 3 sont désinvestis à la faveur du cluster 1 uniquement. Autrement dit, les fonds caractérisés comme présentant une "tendance faibles voir négatives" sont désinvestis en faveur des fonds à "volatilité forte, élastique par rapport au cac40, avec tendance haussière".

- **distance DTW** : les fonds appartenant au cluster 2 sont désinvestis à la faveur du cluster 4, et en moindre mesure du cluster 3. Autrement dit, les fonds à "tendance

générale faible voir négative" sont désinvestis à la faveur des fonds à "tendance générale haussière et impactés par la conjecture économique" et en moindre mesure à la faveur des fonds "volatiles".

L'effet inverse par rapport à 2015 est constaté en mars 2020, et correspond à un phénomène de sécurisation de l'épargne. On constate également que dans chaque clustering, une majorité de la PM est investie sur un cluster (entre 80 et 93 %), ce qui correspond majoritairement à la proportion des fonds en euros.

Sur le graphiques *c*, le cluster 1 devient petit à petit plus important en PM que le cluster 3, du fait du désinvestissement progressif du cluster 2. Autrement dit, les fonds à tendance haussière sans impact conjoncturel deviennent plus présents à la défaveur des fonds à tendance baissière ou bien des fonds volatiles (spéculatifs). Les assurés ne tiennent donc pas compte, inconsciemment ou non, uniquement du rendement, mais également de la volatilité des fonds euros et UC et donc du risque perçu associé.

L'intérêt du clustering est ici visible : il permet de vérifier la cohérence des données notamment avec la conjecture économique, de constater l'interaction entre typologie de fonds, de quantifier les phénomènes de mouvements d'arbitrages entre typologies de fonds, le tout sur un axe temporel, et donc de caractériser plus finement les comportements des assurés en mode de gestion libre grâce à l'analyse précédente de ces clusters.

Le graphique suivant montre l'évolution de l'âge et de l'ancienneté du portefeuille d'assuré :

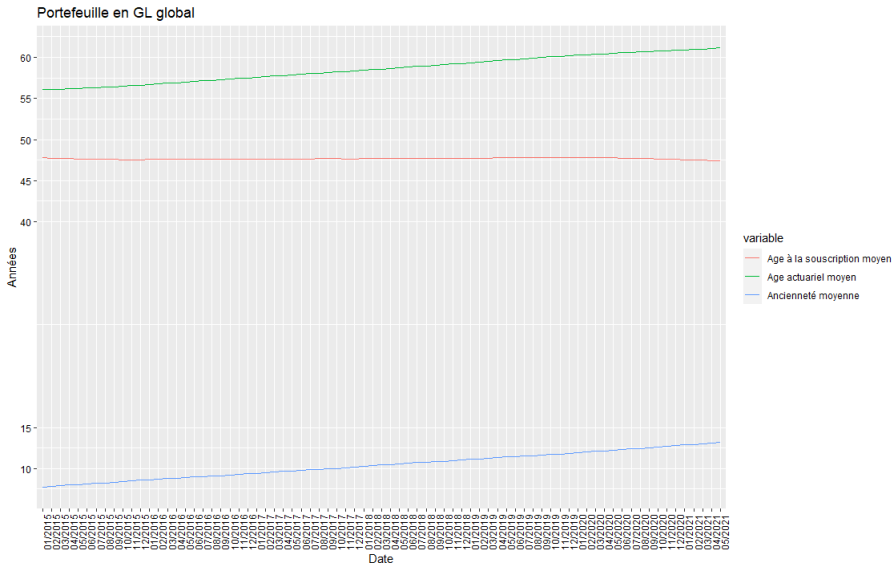


FIGURE 5.5 – Evolution sur le portefeuille en GL de l'âge à la souscription moyen, de l'âge actuariel moyen et de l'ancienneté moyenne des contrats

L'âge à la souscription reste stable aux alentours de 47-48 ans, tandis que l'ancienneté et l'âge actuariel augmentent. Ce graphique illustre donc le fait que peu de souscriptions

sont faites dans ce portefeuille : la rentabilité des contrats plutôt que le nombre est privilégié par la stratégie de la compagnie.

Le graphique suivant montre la proportion de contrat du portefeuille d'assuré dont l'assuré est une femme :

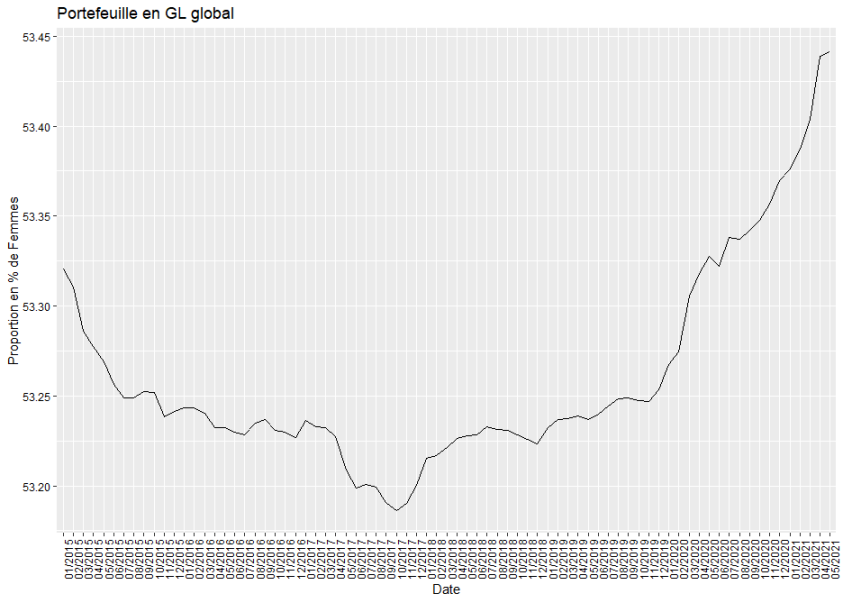


FIGURE 5.6 – Evolution sur le portefeuille en GL de la proportion de femmes

Globalement, il s'agit d'un portefeuille féminin (53% de Femmes), qui le devient de plus en plus avec le temps : les hommes ayant tendance à mourir plus tôt et le portefeuille ne se renouvelant pas beaucoup, les femmes se retrouvent de plus plus représentées dans ce portefeuille.

Le graphique suivant montre la répartition de la périodicité du paiement des primes du portefeuille d'assuré :

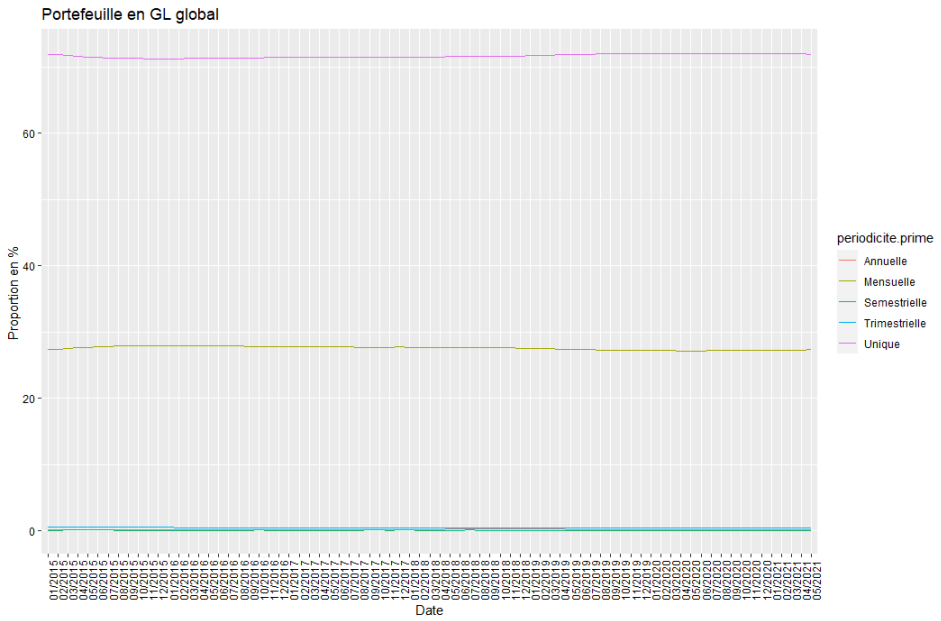


FIGURE 5.7 – Evolution sur le portefeuille en GL de la proportion des différentes périodicités de paiement des primes

Le portefeuille est composé essentiellement de contrats à prime unique (environ 71%), et mensuelle (environ 27%). Les paiements semestriels, trimestriels et annuels sont beaucoup moins répandus.

Le graphique suivant montre la répartition de la CSP du portefeuille d'assuré :

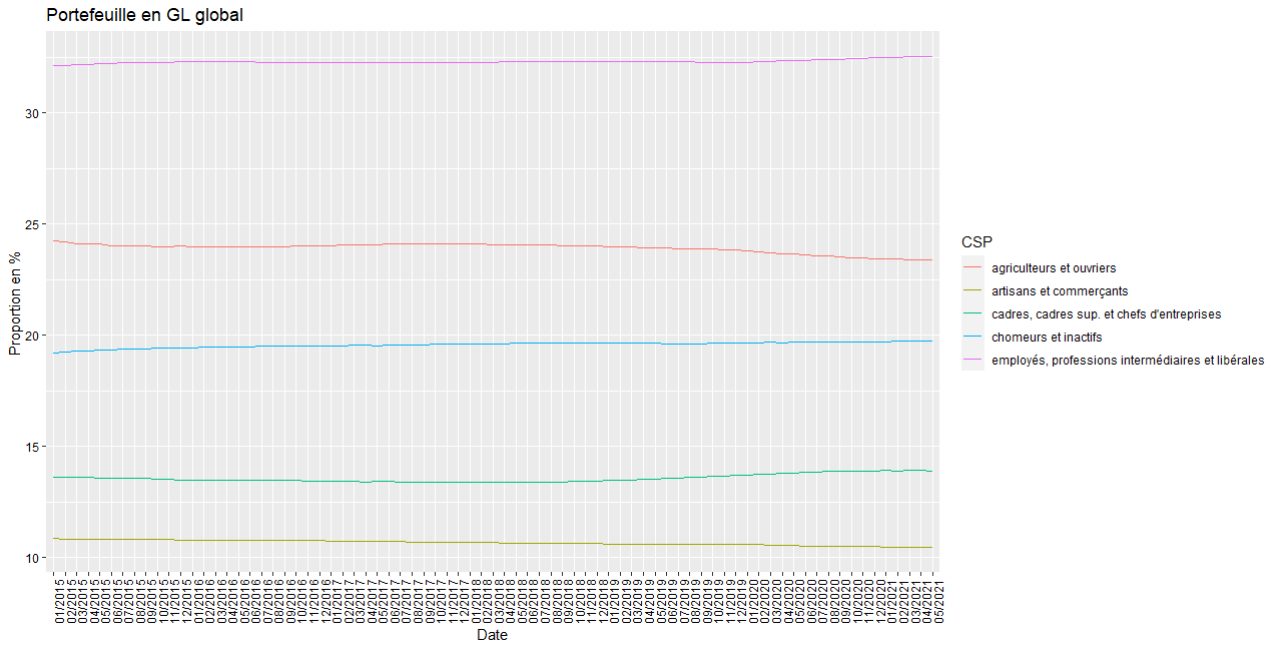


FIGURE 5.8 – Evolution sur le portefeuille en GL de la proportion des différentes CSP

Les catégories de CSP ont étés regroupées en groupes de manière à avoir une lecture plus claire. Les différentes CSP sont stables dans le temps en répartition. Le portefeuille est constitué en majorité d'employés, professions intermédiaires et libérales (32%), d'agriculteurs et ouvriers (24%), d'inactifs et chômeurs (19%), de cadres (13.5%), et d'artisans/commerçants (11%).

5.1.2 Relations taux d'arbitrages VS variables explicatives

Dans cette partie sera étudiée l'influence des variables annoncées aux paragraphes 2.2.2, 2.3 et au chapitre 4 via une analyse univariée. L'impact de chaque variable sur les taux d'arbitrages possédant une origine et une destination EUR_UC, UC_UC, UC_EUR est étudié.

Les taux d'arbitrages sont représentés sur l'axe des ordonnées de gauche. Le nombre de contrats en présence ou la proportion d'UC en présence sont parfois affichés sur l'axe des ordonnées de droite par une courbe ou un histogramme noir afin de mieux se représenter la répartition du portefeuille. Une échelle logarithmique est effectuée sur l'axe des ordonnées afin de pouvoir apprécier sur un même graphique l'impact des variables explicatives sur les trois origines/destinations d'arbitrage différents.

influence de l'âge actuariel

Le graphique suivant montre l'évolution des taux d'arbitrages en fonction de l'âge actuariel de l'assuré au moment de l'arbitrage :

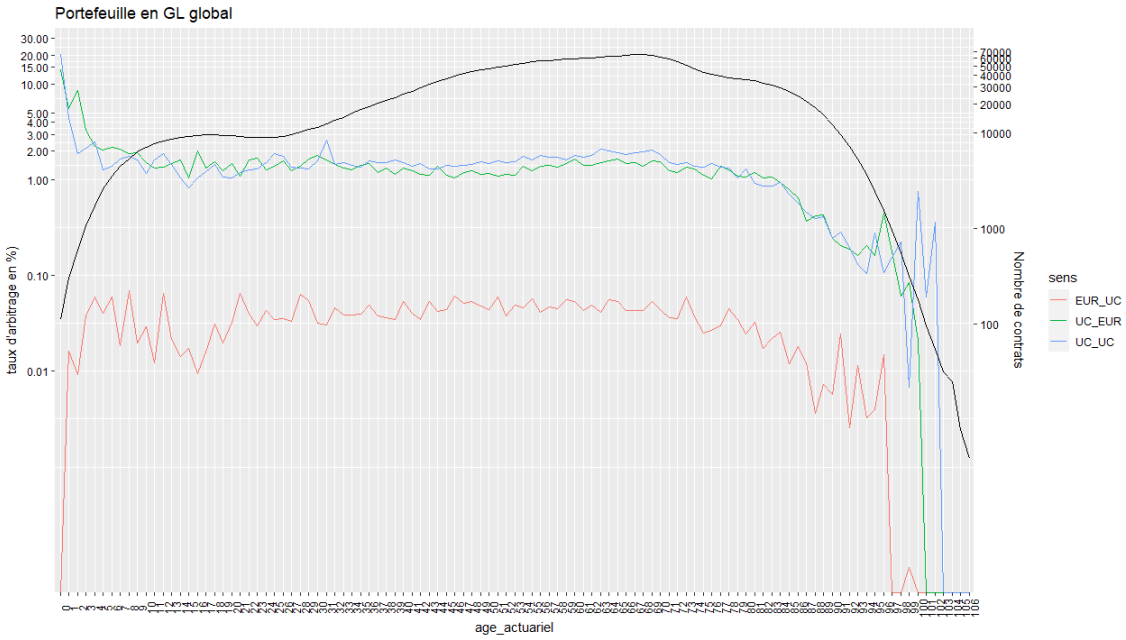


FIGURE 5.9 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de l'âge actuariel de l'assuré

Plus les assurés sont vieux, moins ils arbitrent : leur profil d'investissement est stable et ils ne touchent plus vraiment à la dynamique de celui-ci. Ce constat semble débuter vers 70-75 ans.

Avant les 70 ans, les taux d'arbitrages sont en moyenne stables, avec un léger décalage vers le haut pour les taux UC_EUR et UC_UC sur la tranche 50-70 ans : avant de sécuriser leur épargne sur le fond euros à partir des 70 ans, les assurés veulent au maximum profiter des rendements des différents fonds en vue d'en dégager un meilleur rendement financier.

La tranche d'âge inférieure à 10 ans (les enfants dont les parents constituent une épargne), qui semble dynamique en terme de taux d'arbitrage, est difficilement analysable du fait du plus faible nombre de contrats en présence.

Par la suite dans notre modélisation, nous retiendrons les classes d'âges actuariel 0-30 ans, 31-69 ans, 70-85ans et 86 ans et plus par les variables *ageactu030*, *ageactu3169*, *ageactu7085*, et *ageactu86plus*.

influence de l'ancienneté

Le graphique suivant montre l'évolution des taux d'arbitrages en fonction de l'ancienneté des contrats :

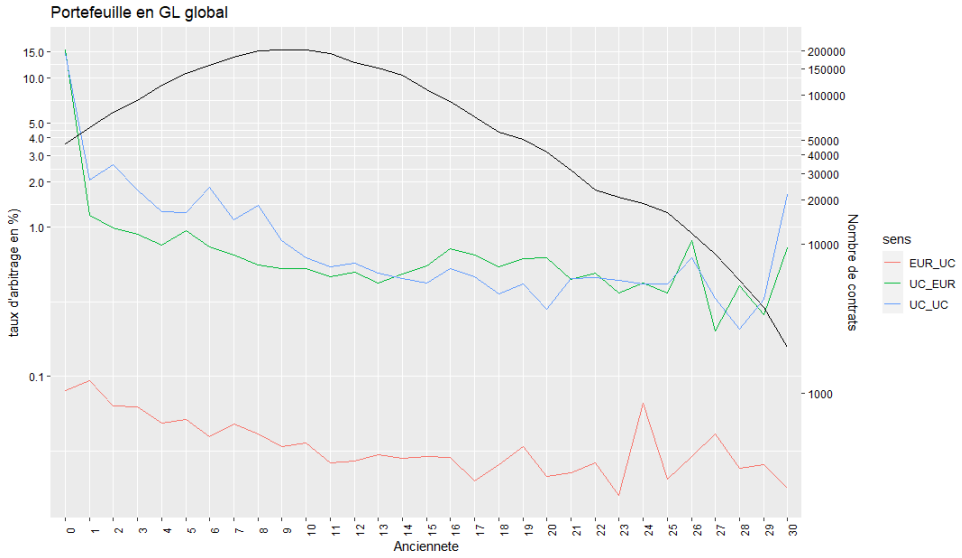


FIGURE 5.10 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de l'ancienneté du contrat

Plus l'ancienneté est grande, moins l'assuré arbitre de manière générale. Le pic de nombre de contrats est atteint entre 8 et 10 ans d'ancienneté puis diminue : les assurés ont tendance à racheter leur contrat une fois la période des 8 ans révolue, pour des raisons fiscales avantageuses sur les rachats après 8 ans. On constate également qu'à partir d'une ancienneté de 14 ans, les taux UC_EUR deviennent plus forts pour la première fois que les taux UC_UC : l'assuré va commencer à sécuriser l' épargne et les gains acquis à partir de 14 ans d'ancienneté.

Par la suite dans notre modélisation, nous retiendrons les classes d'anciennetés 0-8 ans, 9-13 ans, 14 ans et plus par les variables *anc08*, *anc913*, et *anc14plus*.

influence de l'âge à la souscription

Les deux graphiques suivant indiquent la dynamique des taux d'arbitrages en fonction de l'âge à la souscription :

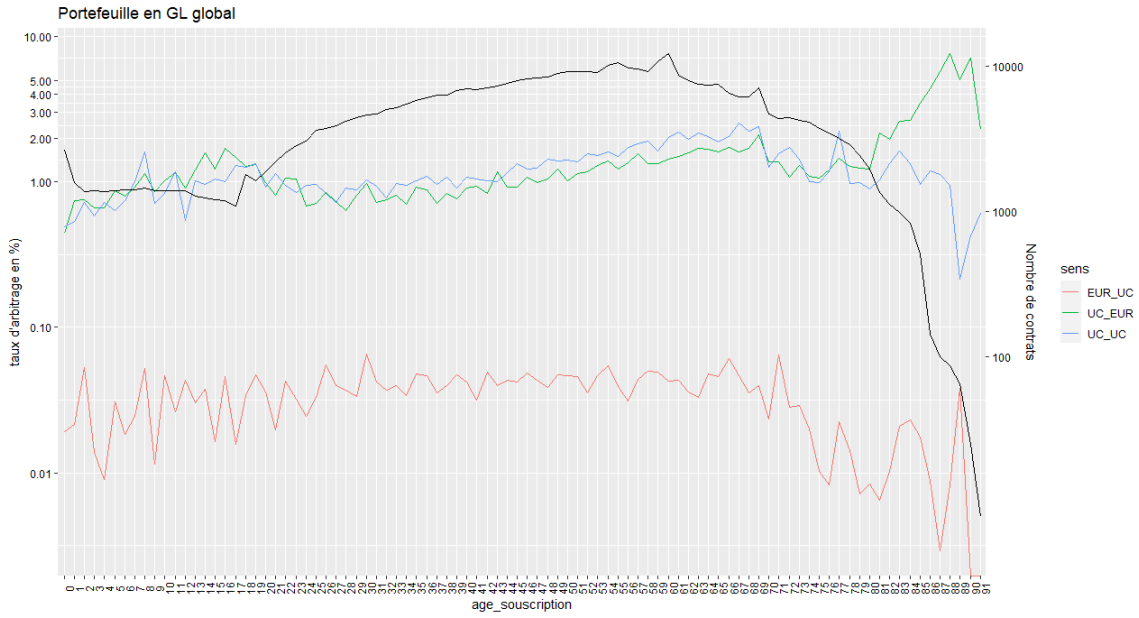


FIGURE 5.11 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de l'âge à la souscription de l'assuré

Si l'on regarde par tranche de 5 ans :

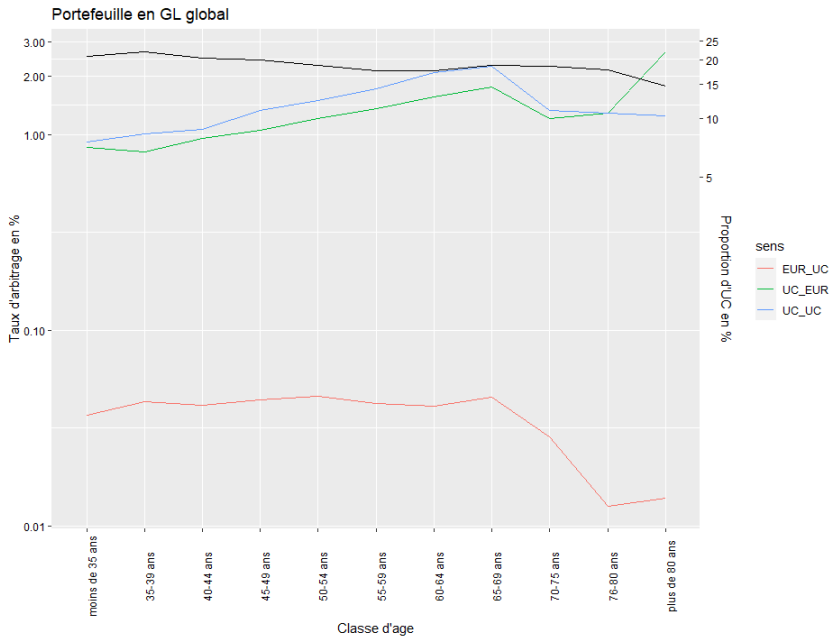


FIGURE 5.12 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de l'âge à la souscription de l'assuré

Toutes les classes d'âge, hormis celle de plus de 85 ans, présentent un nombre suffisant de contrats pour une analyse.

La part d'UC décroît avec l'âge à la souscription : plus l'assuré souscrit vieux, plus il le fait dans une optique d'épargne sans risque. Ainsi, une part d'UC de 20-23% est observée pour les moins de 40 ans, puis décroît jusqu'à atteindre un peu moins de 15% pour les plus de 80 ans. Une "bosse" de part d'UC est observée pour les 65-75 ans (même analyse que l'influence de l'âge actuariel).

Le taux d'arbitrage EUR_UC chute très fortement à partir des 70 ans, alors qu'auparavant il était très stable. Les taux UC_EUR et UC_UC eux ne font que croître pour des âges de souscription inférieurs à 69 ans Le taux UC_EUR connaît son maximum pour les plus de 80 ans et correspond une fois encore à la sécurisation de l'épargne.

Par la suite dans notre modélisation, nous retiendrons les classes d'âge à la souscription 0-20 ans, 21-69 ans et 70 ans et plus par les variables *agesous020*, *agesous2169*, et *agesous70plus*.

influence de la CSP

Le graphique suivant indique la dynamique des taux d'arbitrages en fonction de la CSP :

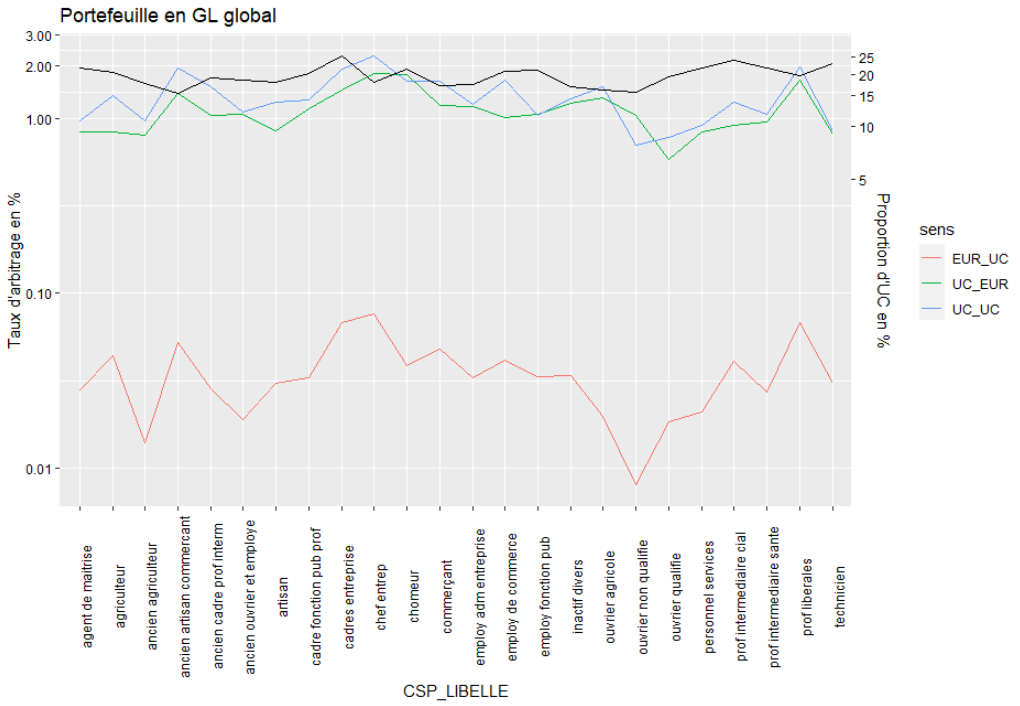


FIGURE 5.13 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la CSP de l'assuré

Les professions possédant les proportions d'UC les plus élevées sont les chefs d'entreprises, les professions intermédiaire et les techniciens. Celles qui arbitrent le moins sont les ouvriers et agriculteurs, et celles qui arbitrent le plus sont les chefs d'entreprises, cadres et professions libérales. Les CSP les mieux représentées en terme de nombre de contrats sont, par ordre décroissant, les inactifs divers (60680 contrats), les employés d'entreprises (41620 contrats), les anciens ouvriers et employés (35496 contrats), les anciens cadres de professions intermédiaires (33230 contrats), les cadres d'entreprises (30079 contrats). Le reste des CSP possèdent un nombre de contrats compris entre 2000 et 20000 contrats.

Par la suite dans notre modélisation, nous constituerons trois classes (trois variables explicatives) de CSP vis à vis des taux d'arbitrages : *csp_bas*, *csp_mid*, *csp_haut* pour regrouper les CSP arbitrant respectivement peu, de manière modérée, et beaucoup. La variable *csp_bas* est composée des CSP : agent de maîtrise, ancien agriculteur, ancien ouvrier, ancien ouvrier et employé, ouvrier non qualifié, ouvrier qualifié, agriculteur, personnel de services, ouvrier agricole. La variable *csp_mid* est composée des CSP : artisan, cadre fonction pub prof, ancien cadre prof interm, chômeur, inactif divers, profession intermédiaire de santé, employé d'entreprise, employé de commerce, employé de fonc-

tion publique, prof intermediaire cial, et technicien. La variable *csp_haut* est composée des CSP : commerçant, cadre d'entreprise, chef d'entreprise, ancien artisan commerçant, professions libérales.

influence du sexe

Le tableau suivant indique la dynamique des taux d'arbitrages en fonction du sexe :

Sexe	%UC	EUR_UC	UC_UC	UC_EUR	nombre de contrats
F	17.73%	0.0292%	1.33%	1.25%	220 765
M	20.23%	0.0506%	1.65%	1.27%	192 766

FIGURE 5.14 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction du sexe de l'assuré

Les femmes sont plus averses au risque que les hommes : proportion d'UC plus faible, arbitrent moins de l'euro vers l'UC, arbitrent moins de l'UC à l'UC, pour un arbitrage de l'UC vers l'euro équivalent.

Par la suite dans notre modélisation, nous retiendrons l'influence du sexe avec la variable *sexeF*.

influence de la date

Le graphique suivant indique la dynamique des taux d'arbitrages en fonction du mois de l'année :

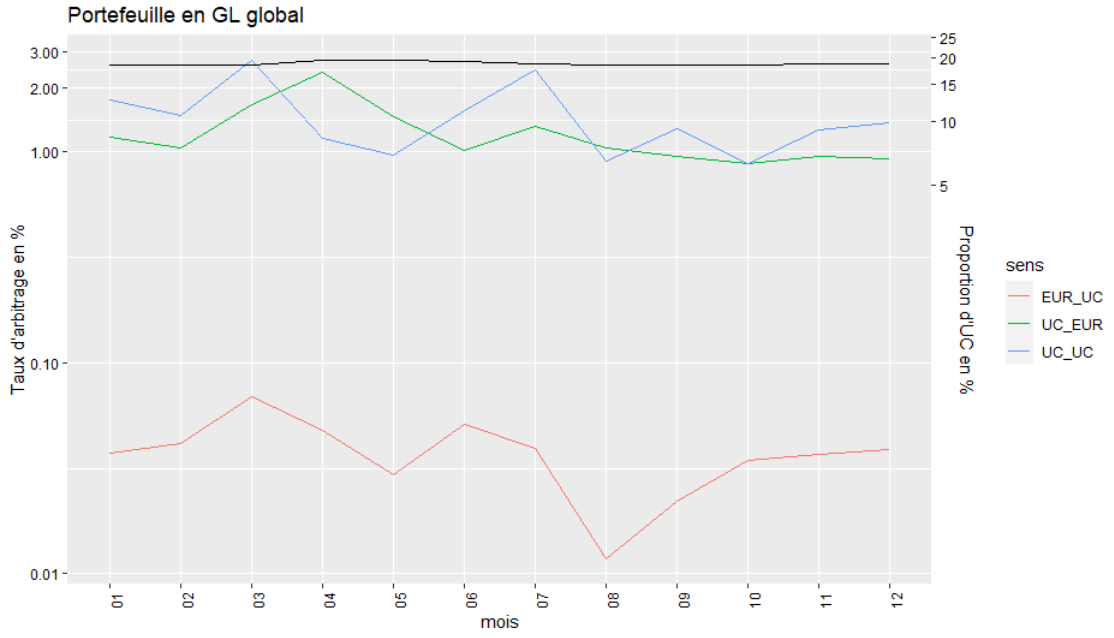


FIGURE 5.15 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction du mois d'observation

Il existe une disparité entre mois d'observation concernant les taux d'arbitrages. Lorsqu'un pic/chute est détecté un mois donné, l'inverse est observé le mois suivant. Les mois arbitrant le plus sont mars et juillet. Le mois où le moins d'arbitrages sont constatés est celui d'août (UC_UC surtout). Cela est lié au fait qu'en début d'année l'assuré est contacté par l'assureur afin de lui communiquer sa PB et de l'inciter à placer ses capitaux sur les UC. En juillet, des taux élevés sont constatés également. En août, les taux bas peuvent être expliqués du fait des taux hauts du mois de juillet : l'assuré ayant effectué ses mouvements d'arbitrages en juillet, il effectue en moyenne moins d'arbitrage en août afin de voir si sa stratégie paie ou non. Il y a donc une certaine inertie suite à un arbitrage (attente de résultats, frais d'arbitrages peut être aussi etc...).

Par la suite dans notre modélisation, nous retiendrons la variable catégorielle *mois* venant spécifier le mois d'observation de l'observation.

influence des clusters formés au chapitre 4

Le graphique suivant indique la dynamique des taux d'arbitrages en fonction de la proportion de PM du contrat affectée à chacun des 3 clusters formés par les indicateurs statistiques :

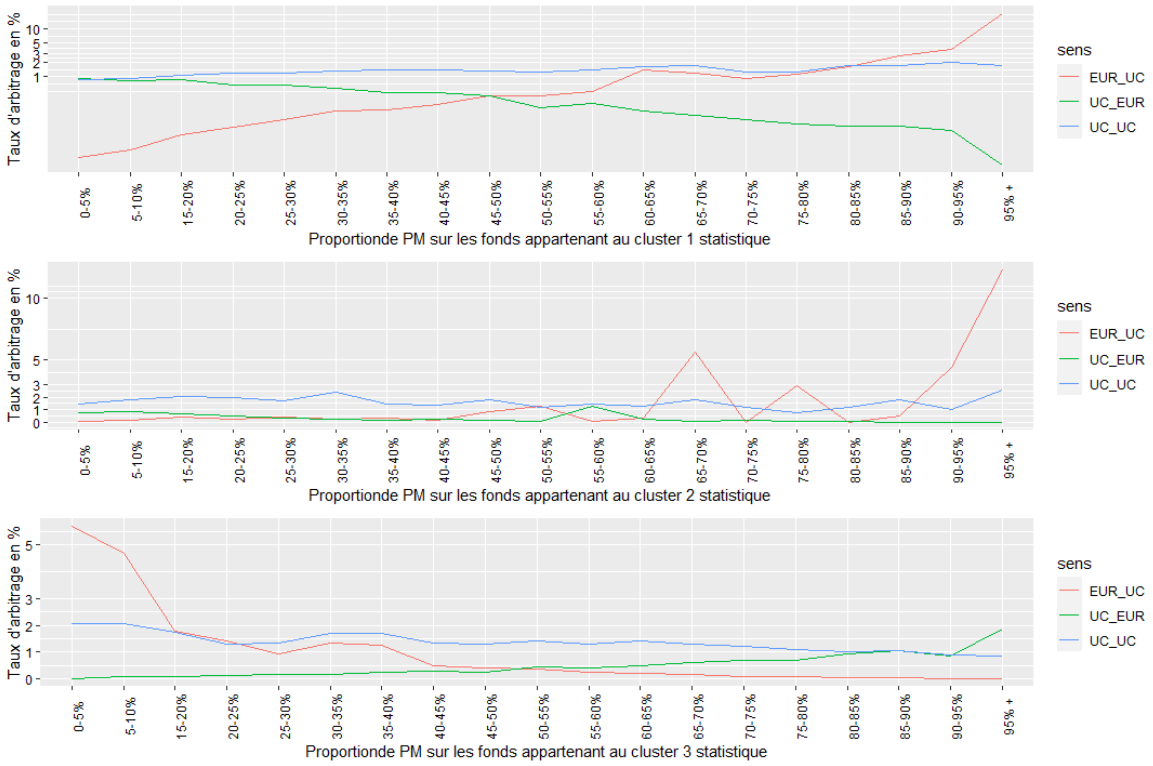


FIGURE 5.16 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la proportion de PM allouée à chaque cluster formé par les indicateurs statistiques

Voici l'analyse que nous pouvons faire :

- cluster 1 / "investissement risque modéré" : le taux EUR_UC et UC_UC sont croissant (relation linéaire) avec la proportion de PM sur ce cluster. Le taux UC_EUR lui est décroissant avec la proportion. Plus la proportion en fonds "investissement risque modéré" est élevée plus l'assuré a tendance à orienter son épargne vers l'UC et à dynamiser celle-ci avec des arbitrages entre fonds UC afin d'en dégager un rendement satisfaisant à ses yeux.

- cluster 2 / "investissement risqué" : des taux EUR_UC d'arbitrages élevés sont observés à partir d'une proportion de PM investi sur ces fonds de 45%. Les taux UC_UC restent constants peu importe la proportion de PM allouée aux investissements risqués. Les taux UC_EUR sont décroissants avec la proportion. Ainsi, plus la proportion en investissement risqué augmente (et donc plus la PM en UC est élevée puisque les fonds UC sont de nature plus risqués que les fonds euros historiquement) plus l'assuré arbitre vers l'UC. Ce phénomène s'observe pour un niveau de PM sur ce cluster de 45% et plus. Cependant, celui-ci maintient un niveau d'arbitrage entre fonds UC constant : la nature risquée de la proportion de ses investissements ne provoque pas d'effet de panique chez l'assuré.

- cluster 3 / "investissement peu risqué" : les taux d'arbitrages EUR_UC et UC_UC sont décroissants, tandis que les taux UC_EUR sont croissants avec la proportion en PM allouée aux investissements peu risqués. Plus la PM en investissement peu risqué augmente, plus l'assuré oriente donc son épargne vers l'euro : il s'agit du profil d'assuré se contentant d'un rendement très modéré mais stable (averse au risque). Pour 80% de PM sur ces fonds et moins, les taux UC_UC sont plus forts que les taux UC_EUR. Ainsi les contrats avec moins de 80% de PM sur ces fonds caractérisent les contrats averses aux risques mais dynamisant leur épargne en arbitrant entre UC, tandis que les contrats possédants plus de 80% de leur PM sur ces fonds dynamisent à la marge leur rendement.

Le graphique suivant indique la dynamique des taux d'arbitrages en fonction de la proportion de PM du contrat affectée à chacun des 3 clusters formés par les indicateurs financiers :

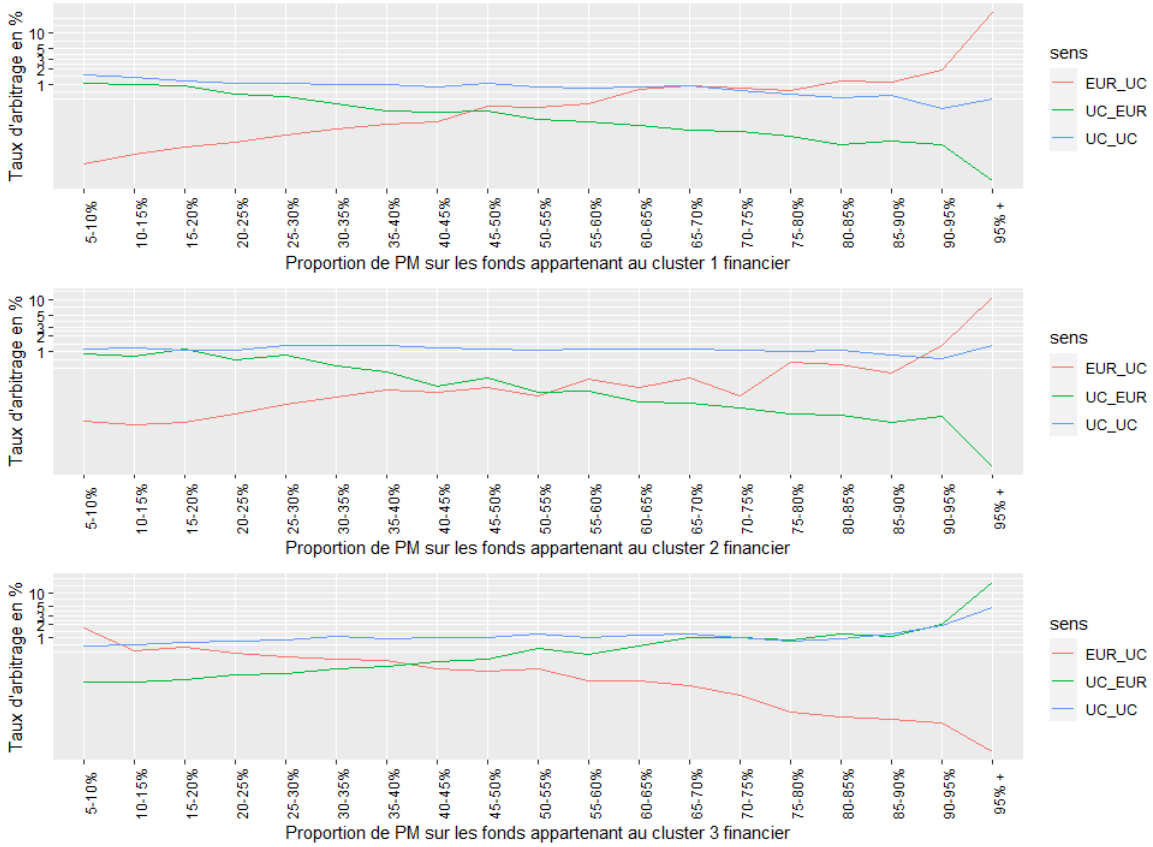


FIGURE 5.17 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la proportion de PM allouée à chaque cluster formé par les indicateurs financiers

Voici l'analyse que nous pouvons faire :

- cluster 1 / "fonds dynamiques" : plus la proportion en PM augmente, plus les taux UC_UC et UC_EUR diminuent, tandis que les taux EUR_UC augmentent. Le constat est logique : plus une proportion en fonds dynamiques augmente, plus l'assuré possède une PM élevée sur les UC. L'ajout ici est de voir que le les taux UC_UC diminuent : les assurés laissent "travailler" leurs fonds dynamiques en arbitrant de moins en moins vers l'euro ou bien entre UC lorsqu'ils présentent une PM élevée sur les fonds dynamiques. Les assurés ont donc soit confiance dans les marchés, soit hésitent à arbitrer leur fonds volatiles par manque de connaissance ou par peur de mal arbitrer (paralysie).

- cluster 2 / "fonds modérés" : la dynamique des arbitrages par origine/destination est la même que pour les fonds dynamiques. La différence réside dans les taux d'arbitrages UC_UC qui restent constants : les fonds étant moins volatiles, les assurés poursuivent un comportement normal d'arbitrage

- cluster 3 / "fonds de sécurisation" : lorsque la proportion de PM allouée aux fonds de sécurisation augmente, les taux EUR_UC diminuent. Cela s'explique par le fait que les fonds de sécurisations sont des fonds euros ou bien des fonds UC à faible dynamique. Les taux UC_EUR et UC_UC augmentent avec la proportion : les assurés tirent leurs rendements en arbitrant entre fonds UC et sécurisent ensuite leurs gains sur les fonds euros.

Le graphique suivant indique la dynamique des taux d'arbitrages en fonction de la proportion de PM du contrat affecté à chacun des 4 clusters formés par le clustering avec distance DTW :

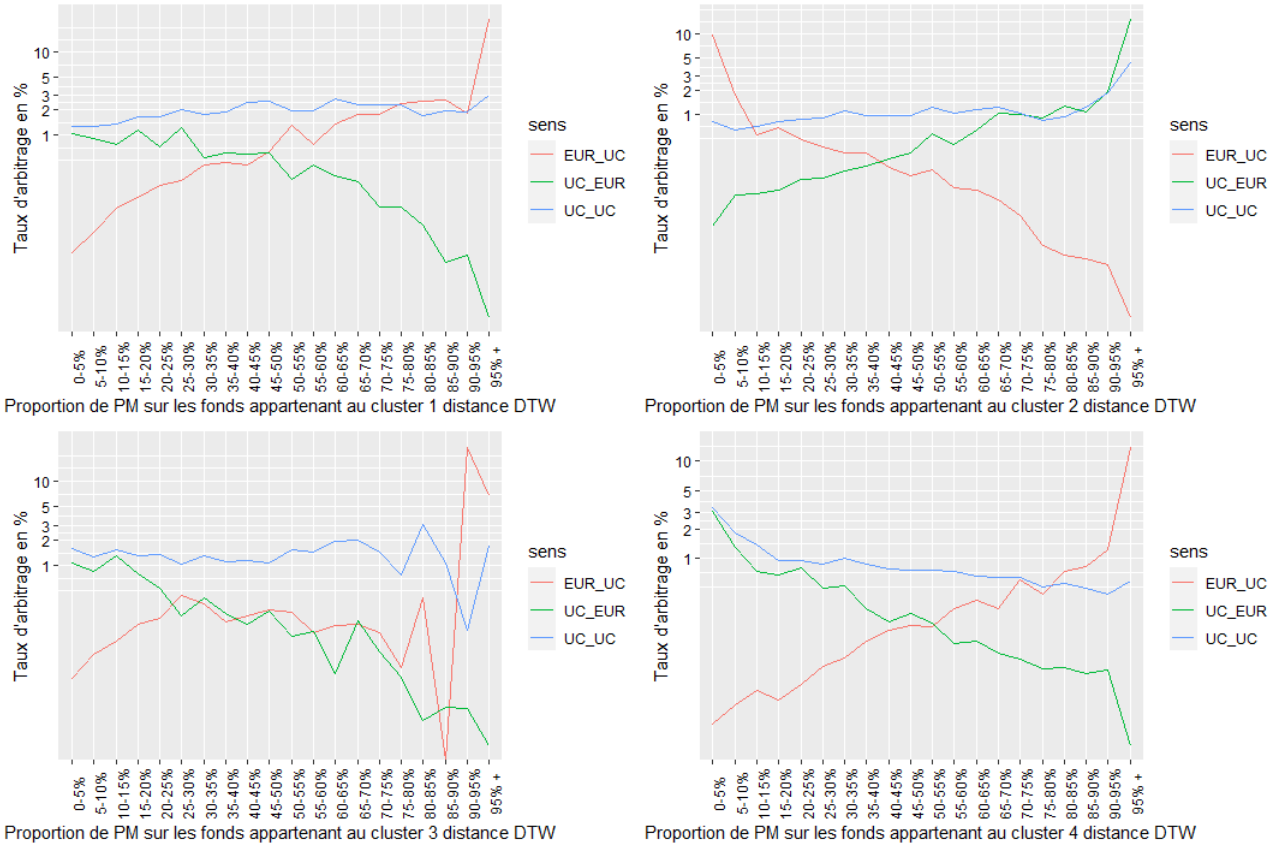


FIGURE 5.18 – Evolution sur le portefeuille en GL des taux d'arbitrages en fonction de la proportion de PM allouée à chaque cluster formé par la distance DTW + K-means

Voici l'analyse que nous pouvons faire :

- cluster 1 / "fonds dynamiques insensibles à la conjecture économique" : les taux d'arbitrages EUR_UC et UC_UC sont croissants avec la proportion en PM. Les taux d'arbitrages UC_EUR sont à l'inverse décroissants avec la proportion en PM. L'assuré oriente donc son épargne vers les fonds UC quand la proportion en PM

des fonds dynamiques et insensible à la conjecture économique augmente. Cela est logique puisque ce cluster est formé majoritairement de fonds UC. Ici il faut remarquer le pic de taux d'arbitrage EUR_UC pour des proportions de plus de 90% : l'assuré se rend compte de la robustesse de ces fonds et par conséquence délaisse complètement les fonds euros. De plus les taux UC_UC augmentent : l'assuré souhaite investir sur ses meilleures UC pour maximiser son rendement.

- cluster 2 / "fonds à dynamique euros" : La dynamique des taux est inverse à celle du cluster précédent, à la différence près qu'il conserve des taux d'arbitrages UC_UC croissant avec la proportion en PM : l'assuré possède une PM élevée sur les fonds euros ou UC à dynamique euros et par conséquent essaie d'obtenir un rendement sur ses autres fonds UC.

- cluster 3 / "fonds risqués" : les taux d'arbitrages présentent des extremum plus marqués sur ce cluster et reflètent donc bien la nature volatile des fonds le composant. Il n'y a pas de réelle relation linéaire, sauf pour les taux d'arbitrages UC_EUR qui sont décroissants avec la proportion en PM : ce cluster est composé en majorité de fonds UC et il est alors naturel d'observer des taux UC_EUR plus bas à mesure que la proportion de PM allouée à ce cluster augmente. À partir d'une proportion de 75% de PM allouée à ces fonds risqués, les taux UC_UC et EUR_UC alternent entre valeurs extrêmes : cela illustre la volonté de l'assuré de tirer profit de la nature volatile des fonds composant son contrat.

- cluster 4 / "fonds dynamiques sensibles à la conjecture économique" : la dynamique des taux d'arbitrage est la même que celle des fonds dynamiques insensibles à la conjecture économique, à la différence près que ici les taux UC_UC sont décroissants lorsque la PM allouée à ces fonds augmente. L'assuré oriente donc son épargne vers les fonds UC quand la proportion en PM des fonds dynamiques et sensibles à la conjecture économique augmente. Cependant, ici, il hésite de plus en plus à arbitrer entre UC, par manque de confiance dans ces fonds sensibles à la conjecture économique, ou bien par peur de réaliser un mauvais arbitrage vers un fond encore plus sensible à la conjecture économique.

À la vue des éléments d'analyses de clusters formés par les différentes méthodes de clustering, nous comprenons que la prise en compte de l'information à la maille ISIN (via le clustering), au delà de l'information à la maille fonds euros/UC, dans les contrats d'assurance vie est une source riche d'information sur le comportement des arbitrages des contrats en mode de gestion libre, puisqu'elle permet de nuancer les comportements par typologie de fonds.

Par la suite dans notre modélisation, nous retiendrons les variables *stat_1*, *stat_2*, *stat_3*, *fi_1*, *fi_2*, *fi_3*, *dtw_1*, *dtw_2*, *dtw_3*, et *dtw_4* (le chiffre indiquant le cluster d'appartenance) pour tenir compte des cluster formés.

influence du produit

Le graphique suivant indique le nombre de contrats et de proportion d'UC en fonction du produit :

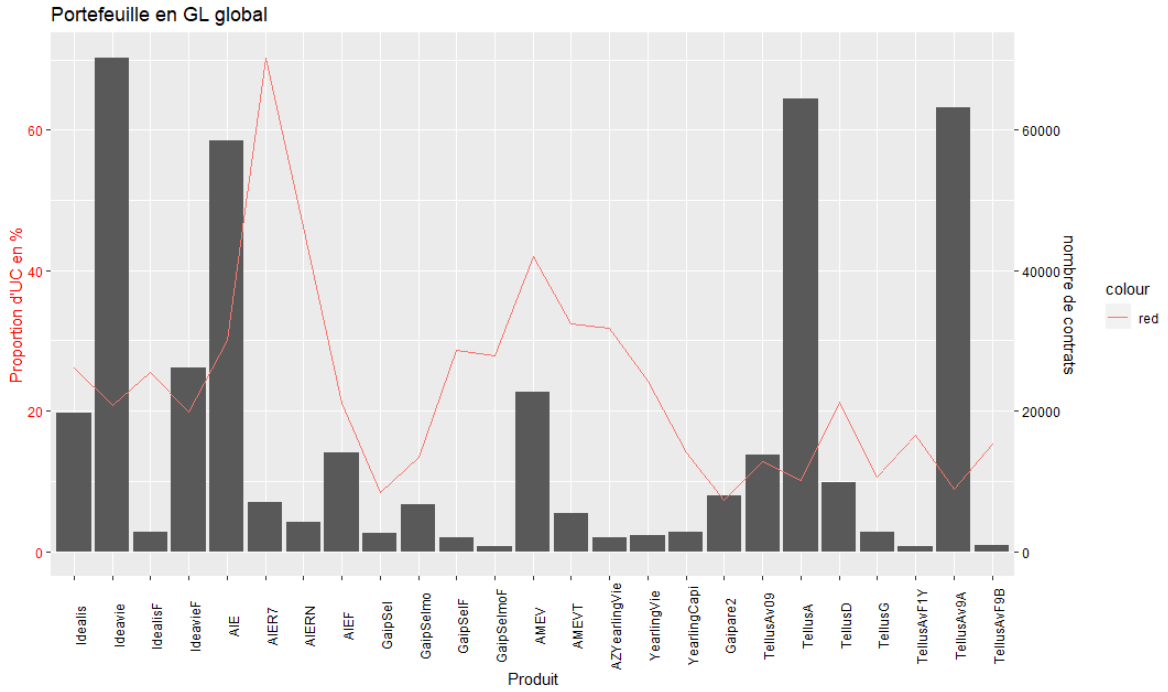


FIGURE 5.19 – Nombre de contrats et proportion d'UC en fonction du produit

Les produits possédant le plus de contrats sont : Ideavie, TellusA, TellusAv9A, AIE, IdeavieF, AMEV, Idealis, AIEF, TellusAv09 avec respectivement 70337, 64486, 63227, 58523, 26138, 22733, 19686, 14076, et 13775 contrats. Les autres produits possèdent entre 800 et 10000 contrats. La proportion d'UC au sein de chaque produit varie très fortement : de 70% pour AIER7 à 8.46% pour GaipSel.‘

Si l'on regarde les PM des produits au 31/12/2020 :

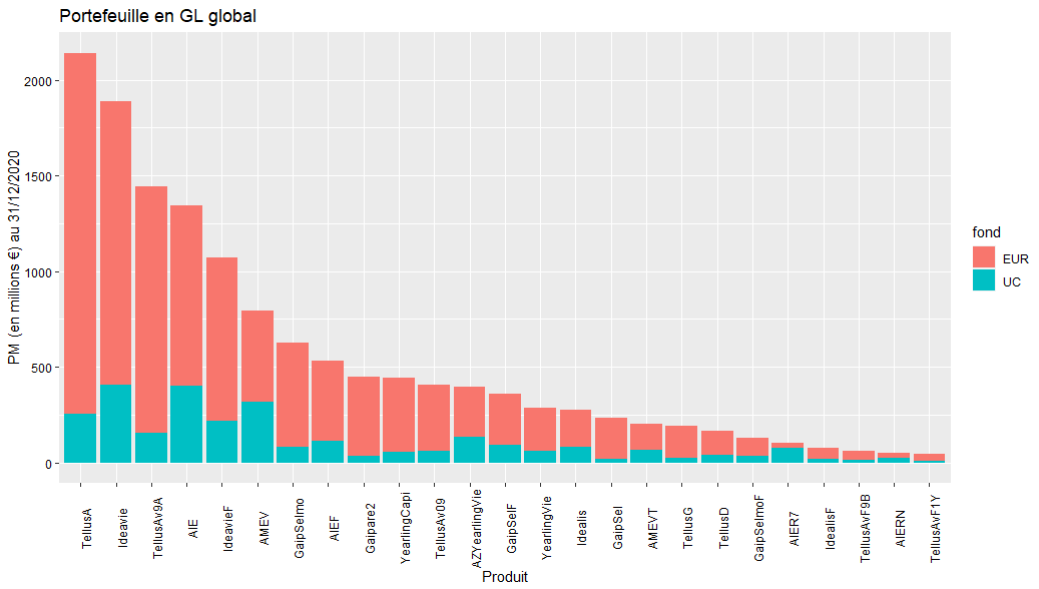


FIGURE 5.20 – PM ventilée par type de fond en fonction du produit

Les produits les plus volumineux en PM sont ceux qui sont également les plus volumineux en nombre de contrats. Les produits les moins volumineux en PM font au minimum 45 millions d'euros de PM.

Si l'on regarde la PM moyenne par contrat des produits :

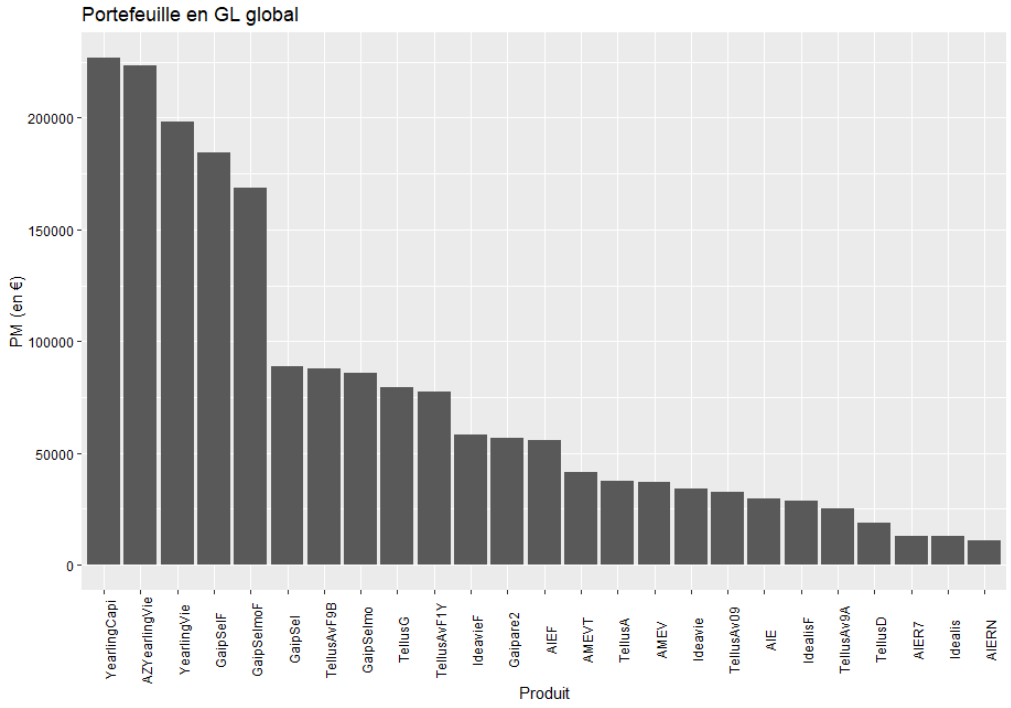


FIGURE 5.21 – PM moyenne par contrat en fonction du produit

Nous pouvons distinguer deux catégories de produits en terme de PM moyenne par contrat : les produits avec PM moyenne supérieure à 150 000€ (qui représentent un peu plus de 2 milliards de PM au 31/12/2020) et ceux à PM moyenne inférieur à 100 000€ (qui représentent environ de 11.6 milliards de PM au 31/12/2020)

Le graphique suivant indique les taux d'arbitrages en fonction du produit :

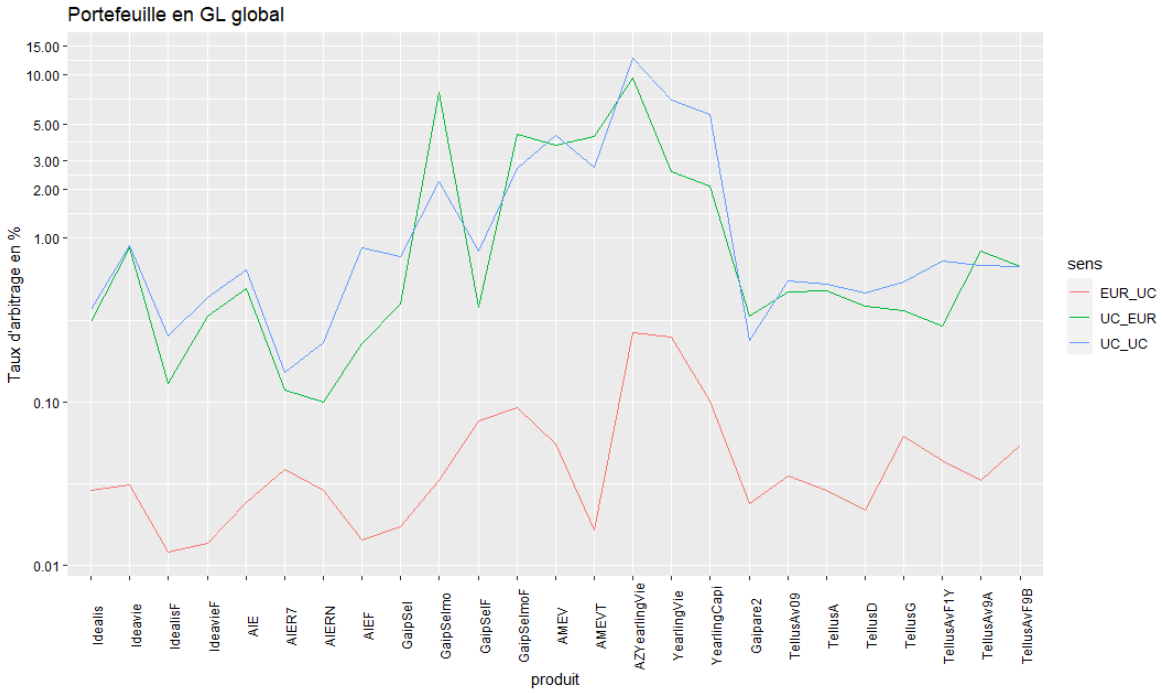


FIGURE 5.22 – Taux d'arbitrages moyens en fonction du produit

En général, un pic/creux est constaté sur l'ensemble des trois origines/destinations pour un produit donné. Par exemple, le produit AZYearlingVie présente les taux d'arbitrages les plus élevés toutes origines/destinations confondues : cela est sûrement dû au fait qu'il s'agit d'un produit premium, composé d'une grande majorité de CSP arbitrant plus (cadres, chefs d'entreprise). À l'opposé, par exemple le produit AIER7, présente lui un taux élevé EUR_UC au regard de ses taux UC_UC et UC_EUR qui sont bas, ce qui peut expliquer le taux élevé de 70% d'UC sur ce produit.

On constate que certains produits sont potentiellement plus risqués en terme d'arbitrage de l'UC vers l'euros : GaipSelmo, GaipSelmoF, AZYearlingVie, TellusAv9A (qui est très volumineux en PM en euros)

influence du nombre de support

Le graphique suivant indique les taux d'arbitrages en fonction du nombre de fonds UC et euros différents dans le contrat à un mois donné :

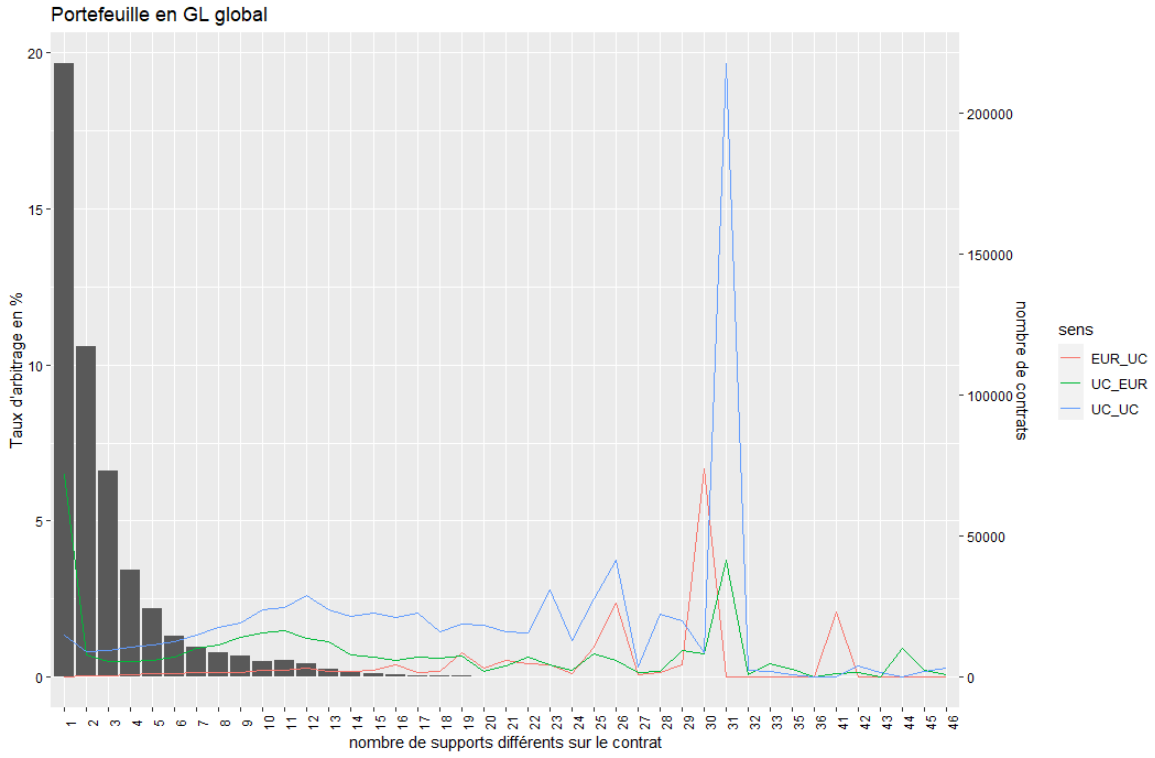


FIGURE 5.23 – Taux d'arbitrages moyens en fonction du nombre de fonds différents

Les taux d'arbitrages toutes origine/destination confondus augmentent avec le nombre de support dont dispose l'assuré. À partir de 3 fonds différents, les taux d'arbitrages augmentent. L'analyse est impossible à partir d'un nombre de support supérieur ou égal à 22, puisque seulement 200 contrats sont concernés.

Par la suite dans notre modélisation, nous retiendrons les classes de nombre de supports 1 support, 2 à 5 supports, 5 à 10 supports, et 10 supports et plus, par les variables *nbsupp_1*, *nbsupp_25*, *nbsupp_510* et *nbsupp_10plus*.

influence du réseau

Le graphique suivant indique les taux d'arbitrages en fonction du réseau de distribution :

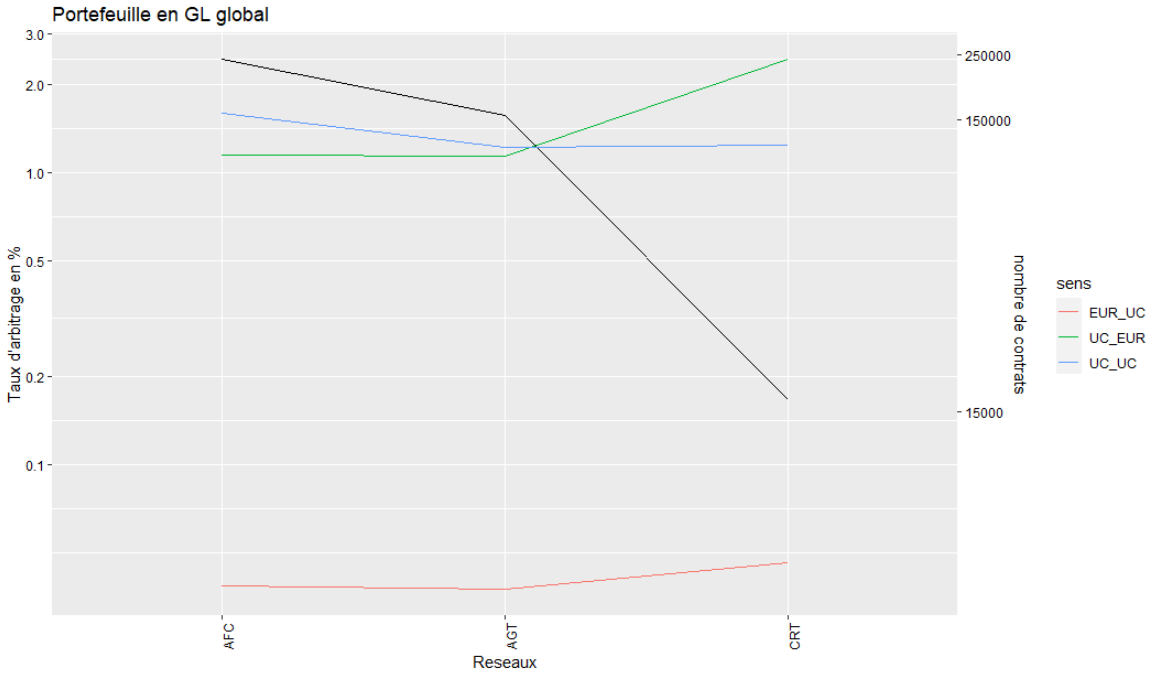


FIGURE 5.24 – Taux d'arbitrages moyens en fonction du réseau

Le réseau de distribution de courtage est le moins bien représenté puisqu'il concerne seulement 16527 contrats, contre 241963 pour le réseau AFC et 155043 pour le réseau AGT.

Le réseau de distribution de courtage possède un plus grand taux d'arbitrage EUR_UC et UC_EUR que les deux autres : il s'agit donc du réseau avec la plus grande dynamique en terme d'arbitrage entre les poches UC et euros. La différence entre le réseau AFC et AGT est que le réseau de distribution AFC présente un plus grand taux d'arbitrage UC_UC.

Si l'on regarde avec un pas de temps mensuel cette différence entre réseaux sur le taux d'arbitrage moyen :

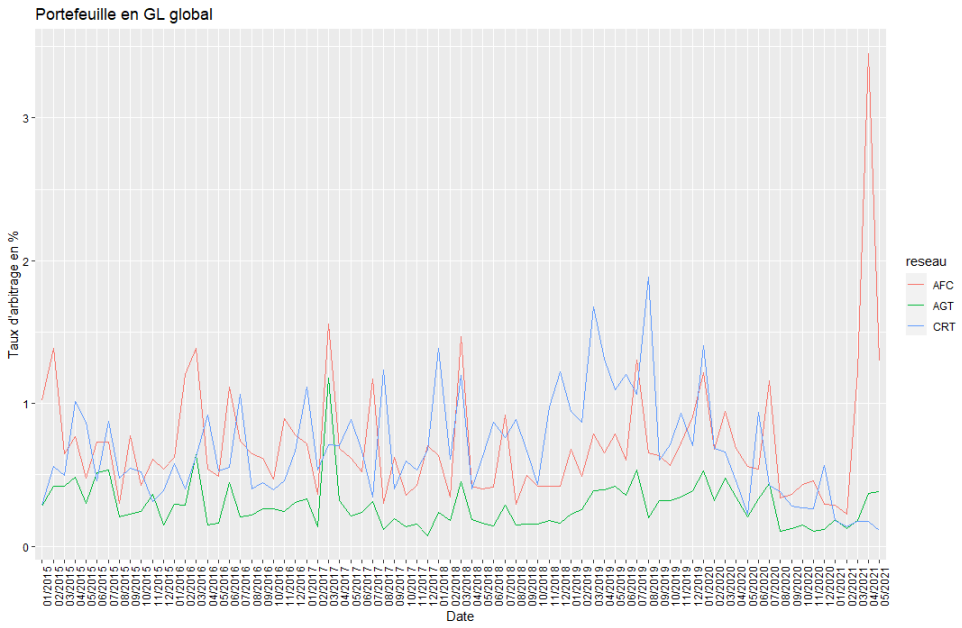


FIGURE 5.25 – Taux d'arbitrages global moyens en fonction du réseau

Le réseau de distribution AGT semble être le plus stable avec une tendance nulle et des taux plus bas, présentant un pic en mars 2017. Les réseaux AFC et CRT eux semblent plus volatiles, avec des amplitudes plus vastes de valeurs extrêmes de taux d'arbitrages. Le réseau AFC présente une tendance nulle tandis que le réseau CRT à lui connu durant la période avril 2018 à aujourd'hui une période de tendance haussière jusqu'à juillet 2019 suivi d'une tendance baissière jusqu'à aujourd'hui.

Par la suite dans notre modélisation, nous retiendrons réseaux de distribution comme variables explicatives par les variables $reseauAFC$, $reseauAGT$, et $reseauCRT$.

influence de la PM des contrats

Le graphique suivant indique les taux d'arbitrages en fonction de la PM des contrats :

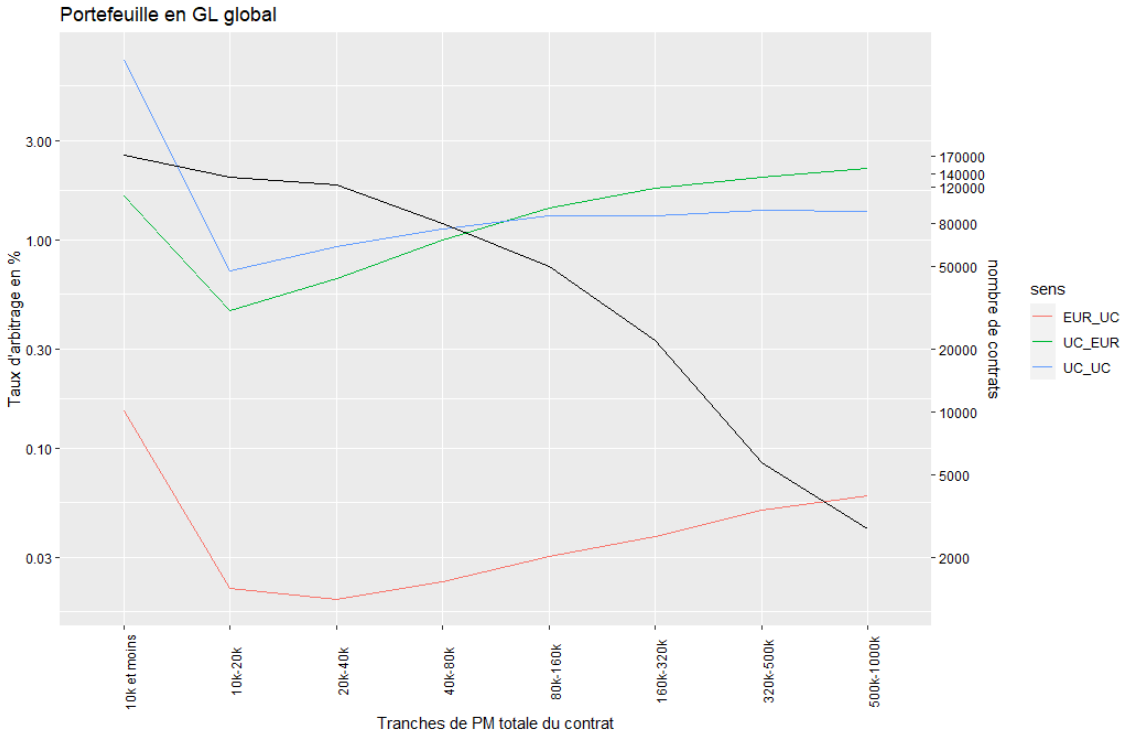


FIGURE 5.26 – Taux d'arbitrages globaux moyens en fonction de la PM totale des contrats

Les contrats arbitrant le plus, toutes origines destinations confondus, sont les contrats possédant une PM inférieure à 10 000 €. Les contrats arbitrant le moins, toutes origines destinations confondus, sont les contrats possédant une PM inférieure à comprise entre 10 000 et 20 000€. Ensuite, pour chaque classe de PM, les taux d'arbitrages augmentent avec la PM. Les taux UC_EUR dépassent les taux UC_UC pour la première fois au niveau de la catégorie "80k-160k" : les assurés possédant des PM élevés de plus de 80 k€ ont tendance à plus sécuriser leur gains que de les réinjecter dans l'UC que les catégories de PM de contrat inférieures. Cependant les contrats possédant une PM élevée arbitrent globalement plus. Ce sont des contrats dont l'assuré est souvent plus fortuné, et donc plus conseillé.

Par la suite dans notre modélisation, nous retiendrons les classes de PM de contrat 10k€ et moins, 10-20k€, 20-40k€, 40-80k€, 80-160k€ et 160-1000k€ , par les variables *PMcontrat_10*, *PMcontrat_1020*, *PMcontrat_2040*, *PMcontrat_4080*, *PMcontrat_80160*, et *PMcontrat_1601m*.

influence de la périodicité de paiement des primes des contrats

Le graphique suivant indique les taux d'arbitrages en fonction de la périodicité de paiement des primes du contrat :

Prime	%UC	EUR_UC	UC_UC	UC_EUR
Annuelle	20.41%	0.076%	1.158%	0.767%
Mensuelle	20.23%	0.0506%	1.65%	1.27%
Semestrielle	19.09%	0.022%	0.61%	0.347%.
Trimestrielle	18.5%	0.024%	0.54%	0.347%
Unique	17.86%	0.042%	1.836%	1.559%

FIGURE 5.27 – Taux d'arbitrages moyens en fonction de la périodicité de paiement de prime du contrat

Les primes annuelles, trimestrielles et semestrielles représentent moins de 1% des contrats, tandis que les primes mensuelles et uniques pèsent respectivement 26% et 73% des contrats. La part d'UC des contrats à primes mensuelles sont plus élevées que les contrats à prime unique. De plus les taux d'arbitrages UC_UC et UC_EUR sont plus élevés pour les contrats à prime unique. Il en ressort donc que les contrats à prime unique semblent correspondre à des profils d'assurés arbitrant vers l'UC en vue de réaliser des plus values et arbitrant ensuite vers l'euro les plus values réalisées. En terme de taux d'arbitrages, les contrats à primes annuelles, semestrielles et trimestrielles semblent être plus proches des contrats à primes mensuelles.

Par la suite dans notre modélisation, nous retiendrons la variable *perio_U* pour un paiement de prime unique.

influence de la part d'UC des contrats

Le graphique suivant indique les taux d'arbitrages en fonction de la part d'UC des contrats :

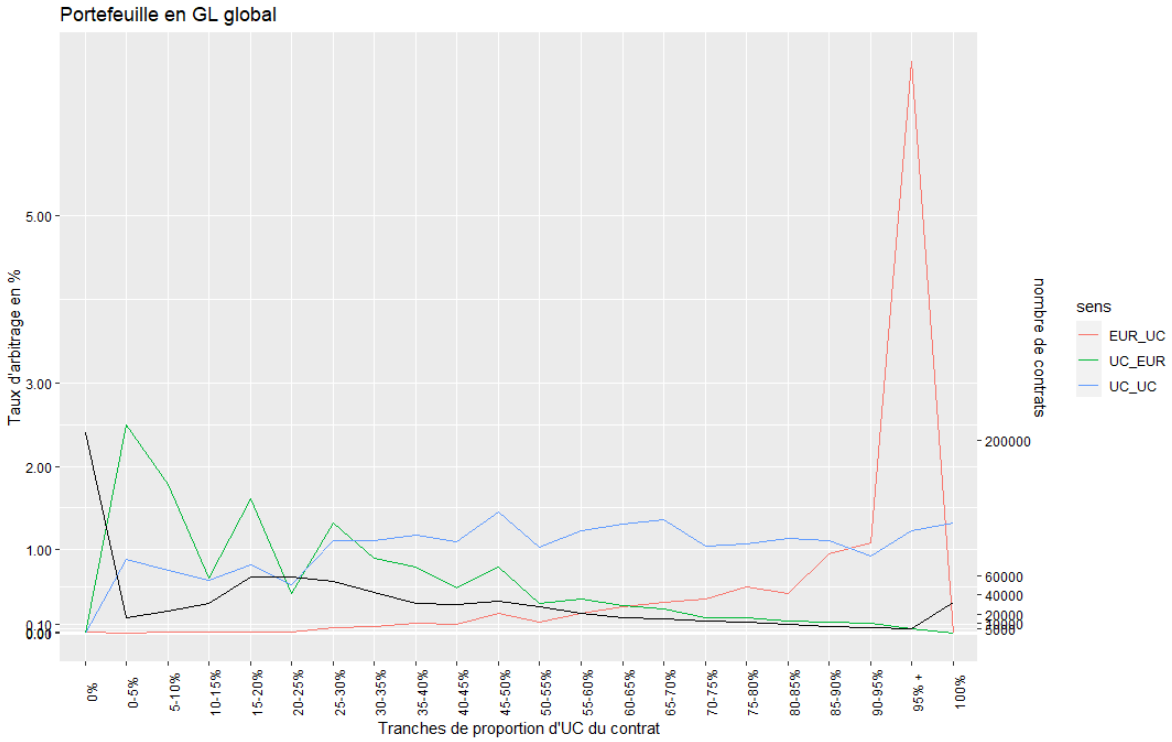


FIGURE 5.28 – Taux d'arbitrages moyens en fonction de la part d'UC des contrats

Plus la proportion d'UC augmente, plus le taux d'arbitrage EUR_UC augmente, ce qui est tout à fait logique, et inversement pour le taux UC_EUR. Pour les taux d'arbitrages UC_UC, les taux d'arbitrages pour une proportion d'UC inférieure à 25% sont en moyenne de 0.65%, mais passée cette barre des 25% les taux d'arbitrages UC_UC doublent presque pour atteindre en moyenne 1.25%. À partir de 25% d'UC dans le contrat, l'assuré à donc tendance à arbitrer plus entre poches UC : une explication est que passé ce seuil, l'assuré possède en général plus de fonds UC différents et arbitre donc plus facilement vers les fonds UC présentant un meilleur rendement. Une autre explication est de se dire que passé ce seuil, un mauvais rendement sur un UC en baisse aura un impact perçu par l'assuré immédiat en terme de valeur de son portefeuille. On remarque également que à partir de 30% d'UC, les taux UC_UC sont plus importants que les taux UC_EUR : à partir de ce seuil les assurés préfèrent réinjecter leur profits dans les UC plutôt que de les sécuriser sur l'euro.

Par la suite dans notre modélisation, nous retiendrons la variable *prop_UC* pour prendre en compte la proportion d'UC.

influence de l'évolution du CAC40

Les graphiques suivant indiquent les taux d'arbitrages en fonction de l'évolution du CAC40 à 1, 3, 6 et 12 mois :

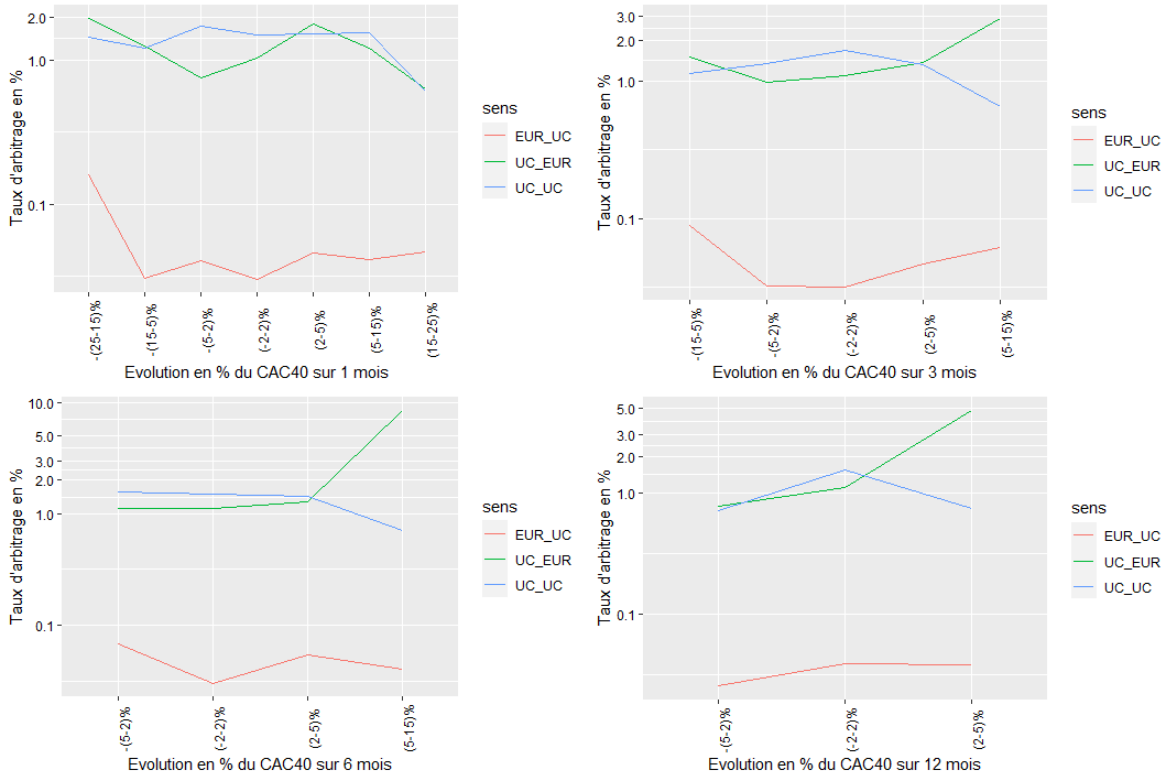


FIGURE 5.29 – Taux d'arbitrages moyens en fonction du rendement du CAC40 à 1,3,6, et 12 mois

Voici l'analyse qu'il est possible d'en tirer :

- taux UC_UC : les taux d'arbitrages sont hauts lorsque le CAC40 évolue de -5% à 5% sur 1, 3 et 6 mois. Les taux d'arbitrages sont plus élevés lorsque le CAC40 évolue de -2 à 2% sur 12 mois.

- taux UC_EUR : les taux augmentent lorsque le CAC40 évolue positivement. Cela correspond à la sécurisation des gains des UC sur le fond euros.

- taux EUR_UC : les taux sont les plus élevés pour les moins bonnes performances du CAC40 sur 1, 3 et 6 mois. Les assurés souhaitent profiter de la dynamique baissière à court et moyen terme pour acheter des parts d'UC moins chères. Les taux sont plus élevés pour les bonnes performances du CAC40 à 12 mois (long terme) : les assurés ont confiance dans la dynamique persistante haussière du marché et arbitrent vers l'UC

Par la suite nous conserverons comme variable explicative l'évolution du CAC40 à 3 mois (Les évolutions du CAC40 à 1, 6 et 12 mois apportent peu d'information en plus) en vue de limiter le nombre de variables explicatives dans nos modèles. Cette variable sera nommée *return_ cac_ 3mois*.

influence de la volatilité du CAC40

Les graphiques suivant indiquent les taux d'arbitrages en fonction de la volatilité annualisée du CAC40 à 1, 3, 6 et 12 mois :

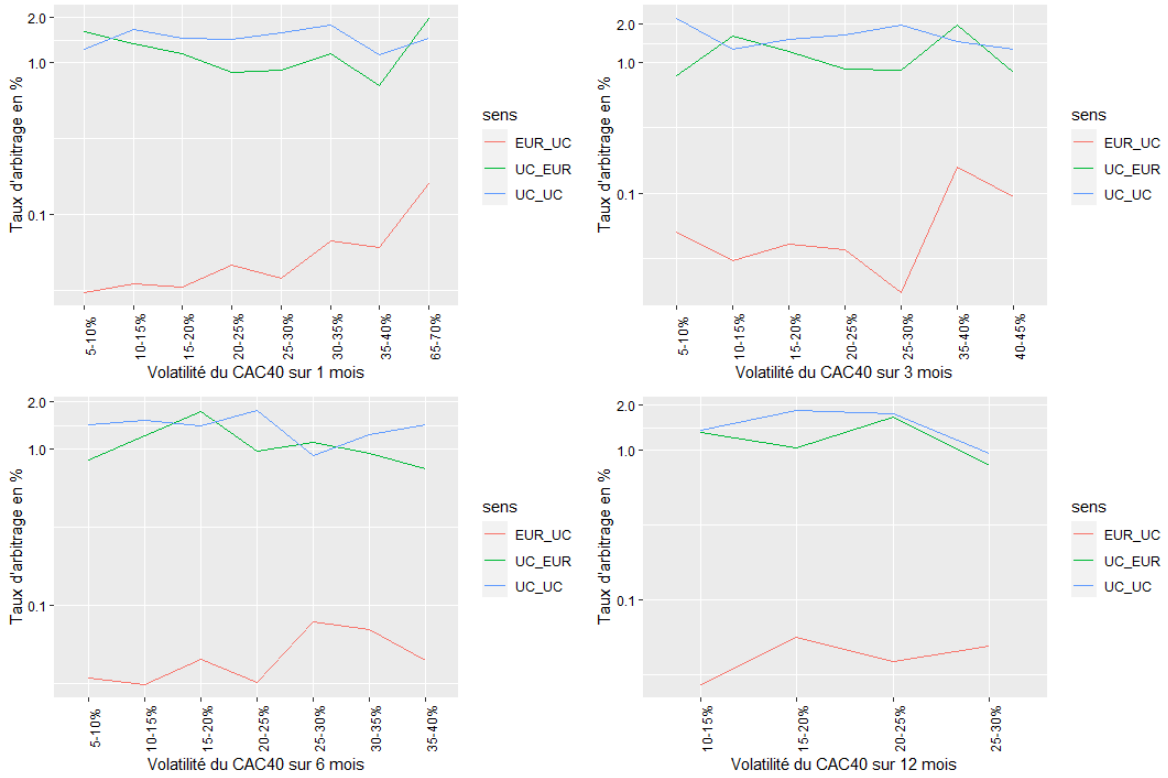


FIGURE 5.30 – Taux d'arbitrages moyens en fonction de la volatilité du CAC40 sur 1,3,6, et 12 mois

Voici l'analyse qu'il est possible d'en tirer :

- volatilité court terme (1 mois) : les taux d'arbitrages UC_UC et EUR_UC augmentent tandis que les taux UC_EUR diminuent lorsque la volatilité augmente : les assurés souhaitent tirer profit de la volatilité récente des marchés.

- volatilité moyen-court terme (3 mois) : les taux UC_UC et UC_EUR ont une dynamique opposée. Pour les valeurs faibles de volatilité à moyen-court terme, l'euro est privilégié, car la dynamique des UC n'est pas suffisante pour présenter un intérêt, tandis que l'UC est privilégié pour les volatilités supérieures à 10%. Pour les arbitrages EUR_UC, des taux élevés sont constatés à partir de 35% de volatilité : les assurés veulent saisir l'opportunité de réaliser des gains avec une volatilité élevée observée sur les trois derniers mois.

- volatilité moyen terme (6 mois) : les taux d'arbitrages EUR_UC augmentent

lorsque la volatilité à moyen terme augmente. Les taux UC_UC et UC_EUR baissent lorsque la volatilité à moyen terme augmente.

- volatilité long terme (12 mois) : les taux d'arbitrages EUR_UC augmentent lorsque la volatilité à long terme augmente. La dynamique des taux UC_UC et UC_EUR sont en opposition, sauf pour les valeurs élevées de volatilité (plus de 25%) pour lesquelles les taux chutent

Les relations taux d'abitrages/volatilité ne sont pas très claires, sauf pour les taux EUR_UC pour lesquelles globalement une augmentation s'opère lorsque la volatilité, sur les différents horizons de temps, augmente.

Par la suite nous conserverons comme variable explicative les volatilités annualisées du CAC40 à 3 mois (Les volatilité annualisées du CAC40 à 1, 6, et 12 mois apportent peu d'information en plus) en vue de limiter le nombre de variables explicatives dans nos modèles. Cette variable sera nommée *vol_cac_3mois*.

influence de l'évolution du TME

Le graphique suivant montre l'évolution du TME de 2011 à aujourd'hui :

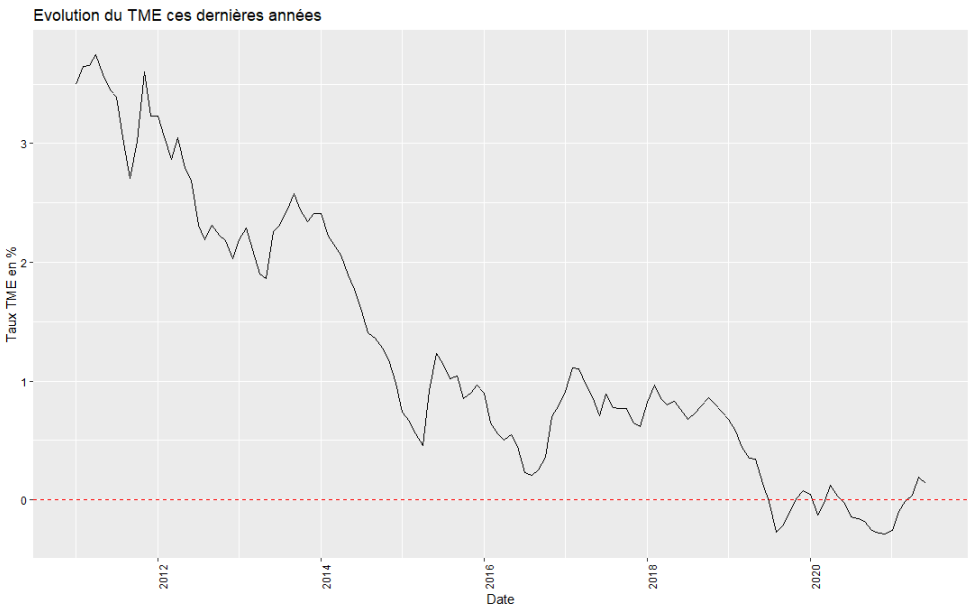


FIGURE 5.31 – Evolution du TME ces dernières années

Le TME ne fait que décroître depuis des années dû au contexte de taux bas/négatifs. Le TME est revenu en territoire positif en mars/avril 2020 avec la COVID19, pour ensuite repasser en territoire négatif. Il est actuellement redevenu positif. La vaste étendue des valeurs du TME nous pousse donc à étudier son évolution plutôt que ses valeurs.

Les graphiques suivant indiquent les taux d'arbitrages en fonction de l'évolution du TME à 1, 3, 6 et 12 mois :

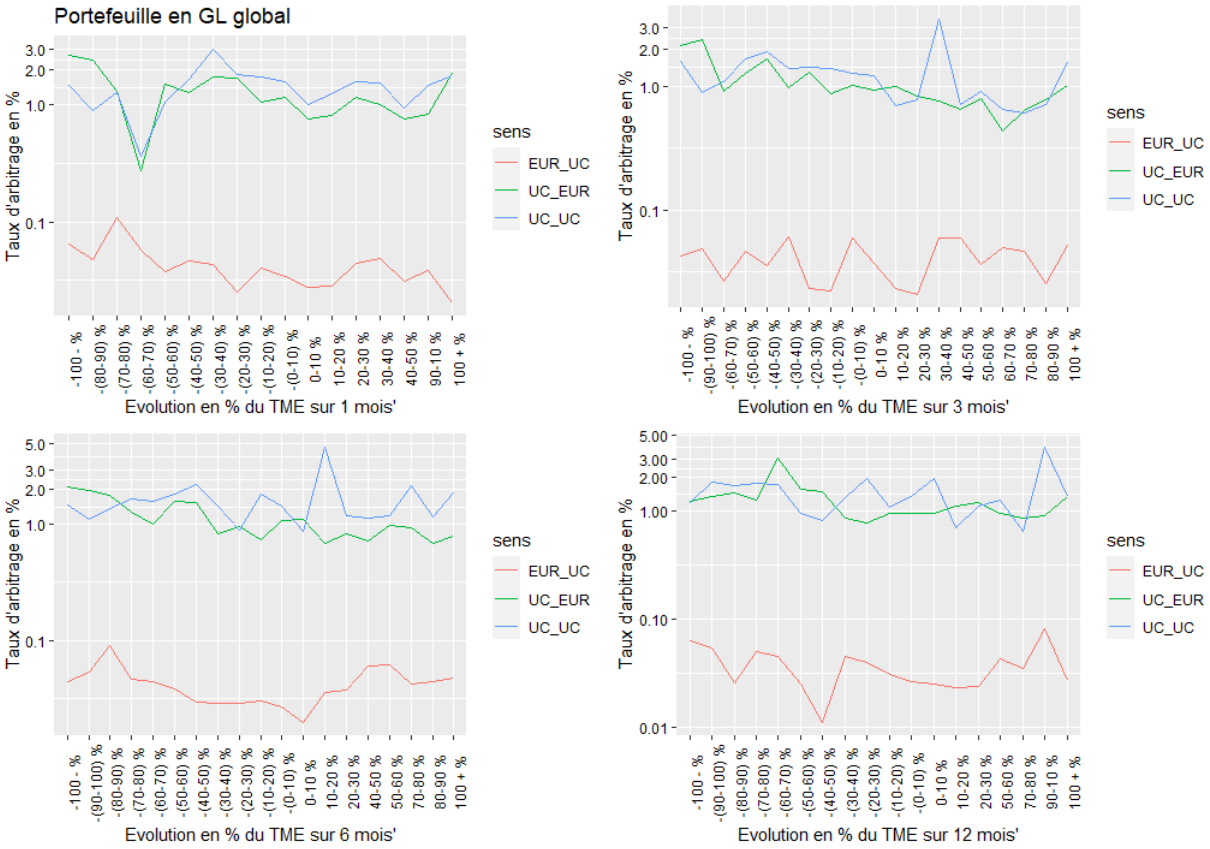


FIGURE 5.32 – Taux d'arbitrages en fonction de l'évolution du TME à 1, 3, 6 et 12 mois

Voici l'analyse qu'il est possible d'en tirer :

- évolution court terme (1 mois) : lorsque le TME augmente, les taux EUR_UC diminuent. Les taux UC_UC sont plus élevés pour des valeurs élevées de TME. Les taux UC_EUR diminuent eux.

- évolution moyen-court terme (3 mois) : lorsque le TME augmente, les taux EUR_UC restent stables en moyenne. Les taux UC_UC et UC_EUR ont une tendance baissière.

- évolution moyen terme (6 mois) : les taux EUR_UC sont plus élevés sur les valeurs extrêmes à la hausse et à la baisse du TME. Les taux UC_EUR évoluent à la baisse lorsque l'évolution du TME est positive. Les taux UC_UC augmentent avec le TME.

- évolution long terme (12 mois) : les taux EUR_UC sont plus élevés sur les valeurs extrêmes à la hausse et à la baisse du TME.

Les relations taux d'arbitrages/évolution du TME ne sont pas très claires, sauf pour les taux EUR_UC pour lesquelles globalement une diminution s'opère à court terme lorsque le TME à court terme augmente (fond euros privilégié pour avoir un meilleur rendement, dans le cadre de proportion en euros élevée). À l'inverse, pour les horizons, moyen-court, moyens et longs termes, un taux d'arbitrage élevé sera observé sur les valeurs extrêmes d'évolution du TME : plus l'évolution du TME est forte, plus les assurés arbitrent. En effet, lorsque que le TME augmente fortement, les placements présentent un meilleur rendement, et donc les plus values sont placées sur les UC. À l'inverse, lorsque le TME diminue fortement, les assurés arbitrent également en faveur de l'UC afin de trouver un moyen de dégager du rendement qui n'est plus possible sur le fond euros.

Par la suite nous conserverons comme variable explicative l'évolution du TME à 3 mois en vue de limiter le nombre de variables explicatives dans nos modèles. Cette variable sera nommée *TME_3mois*.

influence des Google trends

Le graphique suivant indique les taux d'arbitrages en fonction de l'importance de la recherche Google associée au mot "assurance vie" :

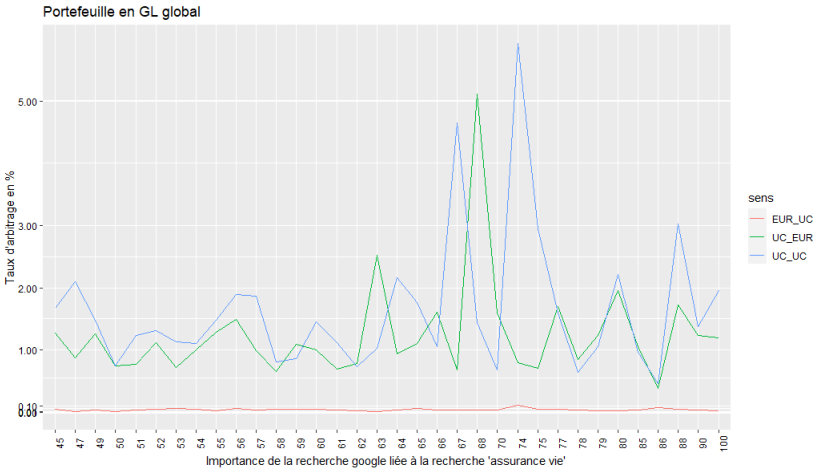


FIGURE 5.33 – Taux d'arbitrages moyens en fonction de l'importance de la recherche Google associée au mot "assurance vie"

Manifestement, plus la recherche est importante ("plus il y a de clics"), plus les taux d'arbitrages présentent des extrêmes.

Le graphique suivant indique les taux d'arbitrages en fonction de l'importance de la recherche Google associée au mot "crise financière" :

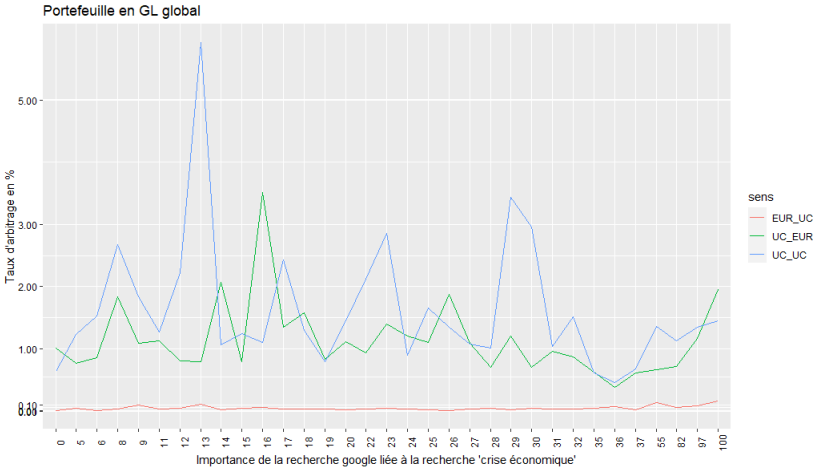


FIGURE 5.34 – Taux d'arbitrages moyens en fonction de la de l'importance de la recherche Google associée au mot "crise économique"

Ici il est important de remarquer que l'axe des abscisses passe d'une valeur de 37 à

55 puis à 82. Ainsi, moins le mot "crise financière" est recherché (importance inférieure à 37) dans la barre de recherche Google, plus les taux présentent des valeurs extrêmes

Enfin, il est important de préciser ici qu'une corrélation n'implique pas forcément un lien de causalité. Les résultats déduits pour les Google trends sont donc à utiliser avec précaution : une analyse trop poussée de ceux-ci n'est en aucun cas justifiable car ils sont soumis à trop de facteurs externes. Cependant, la puissance informative pure de ceux-ci justifie leur prise en compte dans nos modèles.

Les variables prenant en comptes ces google trends seront nommées *recherche_web_assvie* et *recherche_web_crise*.

5.1.3 Court résumé de l'apport des statistiques descriptives

Dans cette partie a été constaté des phénomènes de comportements d'arbitrages classiques, par exemple vis à vis de l'âge de l'assuré ou bien de la PM de son contrat. Ils ne sont donc pas détaillés ici mais sont visibles plus haut dans le corps de rapport.

En revanche, de nouveaux éléments notables sont à mettre en avant ici :

- Il existe une différence de comportement selon la CSP de l'assuré. Si ce constat semble logique, ce phénomène est ici identifié qualitativement et quantitativement. Une segmentation dans une modélisation des taux est donc préconisée.
- Il existe une différence notable de comportement d'arbitrage par produit. Toutefois, le produit est corrélé à l'ancienneté et la CSP. Une segmentation dans une modélisation des taux est donc également préconisée. Nous pourrions par exemple imaginer un clustering de produits par profil de risque au sens de l'arbitrage.
- La variable afférente au nombre de supports, triviale à obtenir, qui modélise ultimement le degré de marche de manoeuvre dont dispose l'assuré (rapatriement de ses capitaux sur le fond euros, ou bien recherche d'un rendement maximum) semble présenter un pouvoir explicatif des taux d'arbitrages. Une segmentation dans une modélisation des taux est donc préconisée.
- Les Google trends semblent retranscrire les possibles périodes de paniques de par leur corrélation avec les taux d'arbitrages en régime extrêmes. Une analyse plus poussée de celles-ci mériterait d'être effectuée afin de déterminer le degré de confiance que nous pourrions leur allouer.
- Le clustering des fonds UC et euros, visant à regrouper les fonds homogènes (selon un critère d'homogénéité statistique, financier et de dynamique d'évolution), fournit beaucoup d'informations. Des relations quasiment linéaires sont observées entre les taux d'arbitrages et la PM allouée à chaque cluster. Le détail des raisonnements sont affichés dans la section correspondante. À titre d'exemple :
  - L'analyse du clustering fait avec le critère d'homogénéité de dynamique d'évolution met en lumière le degré de confiance que peuvent éprouver les assurés selon la dynamique des UC par exemple
  - L'analyse du clustering fait avec des indicateurs financiers nuance notamment les dynamiques d'arbitrages inter UC suivant la PM allouée à des fonds dynamiques ou à des fonds plus modérés.
  - L'analyse du clustering fait avec des indicateurs statistiques montre qu'à partir d'un taux de 45% de PM allouée à des fonds risqués, l'assuré possédant ce profil de contrat n'éprouve pas de panique à la vue de ses taux d'arbitrages normaux.

  Les détails sont données dans la partie "influence des cluster formés au chapitre 4" page 113.

5.2 Maille produit

Nous nous attacherons dans cette partie à modéliser les taux d'arbitrages UC_UC, UC_EUR et EUR_UC à la maille produit, c'est à dire avec une base de donnée dont une ligne décrit les taux d'arbitrages et les valeurs des variables explicatives pour un produit à un mois donné. En effet, comme vu dans les statistiques descriptives, il existe de grandes différences structurelles entre produits qui sont à l'origine de dynamiques de taux d'arbitrages différents. Modéliser convenablement les taux d'arbitrages à cette maille revient donc à appréhender convenablement le risque d'arbitrage en mode de gestion libre de la compagnie et présente donc un intérêt business. En effet cette maille de travail présente l'avantage d'être suffisamment fine pour faire émerger de l'information à forte valeur ajoutée, tout en présentant un niveau d'agrégation suffisamment élevé pour avoir des bases de données de tailles convenables et donc une compléxité de calcul convenable.

Pour se faire, nous aggrégeons notre base de donnée de la maille $contrat \times mois$, décrite au chapitre 2, selon la clef $(produit, mois)$.

5.2.1 Base de donnée

Les statistiques descriptives ayant été construites à la maille $contrat \times mois$ dans la partie précédante, nous tenons compte de l'analyse précédente de celles-ci en prenant la proportion de PM pour le $produit \times mois$ associée aux catégories de valeur de ces variables explicatives. De cette manière nous tenons compte de l'information micro à la maille contrat en vue de modéliser les arbitrages à la maille produit. Concrètement, la variable *PMcontrat_1020* indique la proportion de PM, pour un produit et un mois donné, correspondant à des contrats dont la PM est entre 10000 et 20000€.

De plus l'analyse taux d'arbitrage vs variables explicatives ayant été réalisée avec les données de fin de mois, nous décidons d'associer pour un mois m les valeurs de variables explicatives du mois $m-1$. De cette manière la dimension prédictive est ajoutée à la base de donnée.

La base de donnée à la maille produit, mais tenons compte des informations à la maille $contrat \times fond \times mois$, dotée de 45 variables explicatives est alors obtenue :

- La maille de travail : la date mensuelle et le produit.
- Les différents taux d'arbitrages : UC_UC, UC_EUR et EUR_UC.
- La proportion d'UC à la maille de base, mais aussi la proportion de PM associée à des contrats comprenant plus de 25% d'UC (*prop_UC* et *PMsupp25percent*).
- La proportion de PM associée à chaque typologie de fonds. (statistiques, financier et distance DTW) à la maille de base (*stat_1*, *stat_2*, *stat_3*, *fi_1*, *fi_2*, *fi_3*, *dtw_1*, *dtw_2*, *dtw_3* et *dtw_4*).
- La proportion de PM associée à chaque réseau de distribution (*reseauAFC*, *reseauAGT* et *reseauCRT*).

- La proportion de PM associé à chaque catégorie de nombre de support des contrats (*nbsupp_ 1*, *nbsupp_ 25*, *nbsupp_ 510*, et *nbsupp_ 10plus*).
- La proportion de PM associée à une prime unique (*perio_ U*).
- La proportion de PM associée à chaque catégorie de volume de PM des contrats (*PMcontrat_ 10*, *PMcontrat_ 1020*, *PMcontrat_ 2040*, *PMcontrat_ 4080*, *PMcontrat_ 80160*, et *PMcontrat_ 1601m*).
- La proportion de PM associée à des contrats dont l'assuré est de sexe féminin (*sexeF*).
- La proportion de PM associée à chaque catégorie de CSP (*csp_ bas*, *csp_ mid*, et *csp_ haut*).
- La proportion de PM associée à chaque catégorie d'age actuariel, d'age à la souscription, et d'ancienneté (*ageactu030*, *ageactu3169*, *ageactu7085*, *ageactu86plus*, *agesous020*, *agesous70plus*, *anc08*, *anc913*, *anc14plus*).
- Les valeurs des variables exogènes constatées au mois de la maille de base (*vol_ cac_ 3mois*, *return_ cac_ 3mois*, *recherche_ web_ crise*, *recherche_ web_ assvie* et *TME_ 3mois*).
- Les moyennes à la maille de travail de la part d'UC des contrats, de l'âge à la souscription, de l'âge actuariel, et de l'ancienneté avec les variables *moy_ part_ UC_ contrat*, *moy_ age_ souscription*, *moy_ age_ actuariel*, et *moy_ anciennete*

Il est aussi décidé de supprimer les lignes correspondantes aux début de commercialisation de certains produits. Sur ces lignes, la PM et le nombre de contrats ne sont pas suffisants pour inférer correctement sur les taux d'arbitrages. La base de donnée obtenue présente alors 1854 lignes.

5.2.2 Matrice de corrélation des variables explicatives

Afin de réduire le nombre de variables en vue de conférer à nos modèles une possible interprétation, et de limiter les effets négatifs d'interactions de variables corrélées dans nos modèles, nous étudions la corrélation de Pearson entre chaque variable :

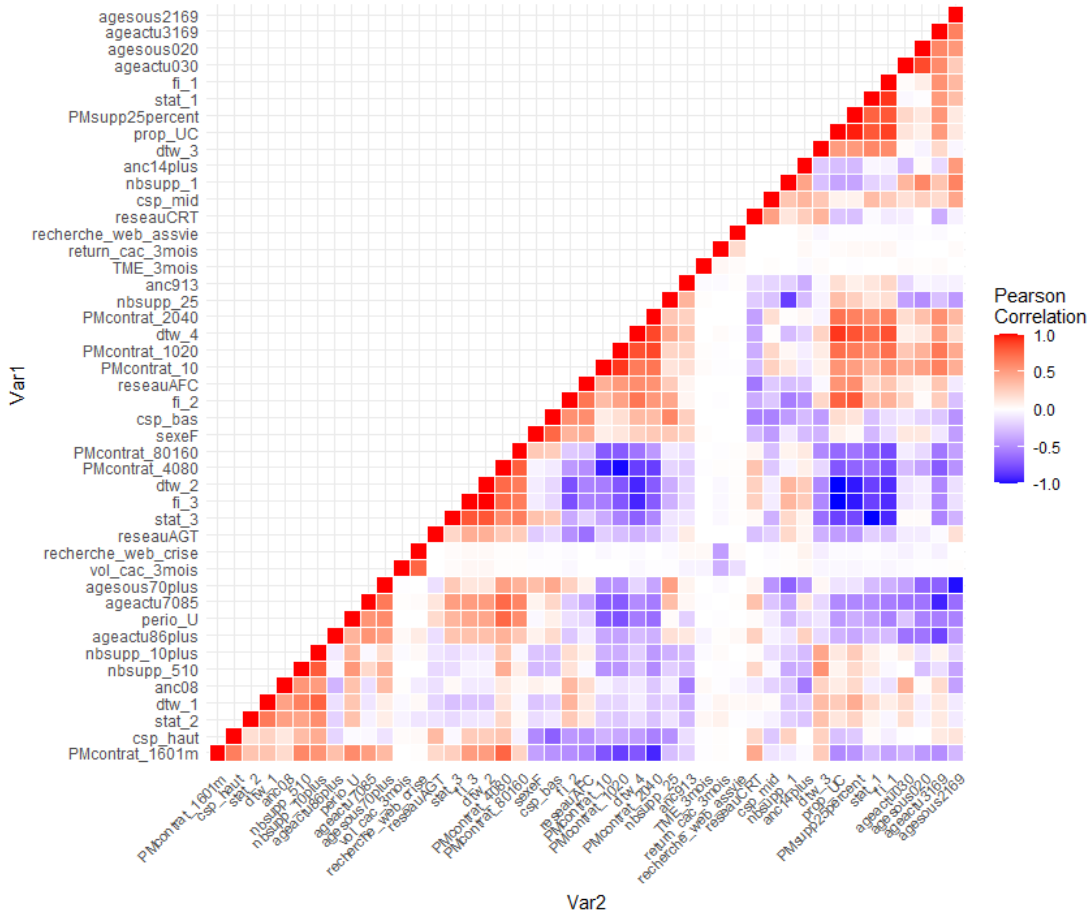


FIGURE 5.35 – Matrice de corrélation de Pearson des variables à la maille produit

Une corrélation négative indique que lorsque la variable 1 augmente, la variable 2 diminue. Une corrélation positive indique que lorsque la variable 1 augmente, la variable 2 augmente elle aussi. La corrélation de Pearson est un indice reflétant une relation linéaire entre deux variables continues. Le coefficient de corrélation varie entre -1 et +1. Une corrélation de 0 indique une relation nulle entre deux variables

Ici, certaines variables sont très corrélées entre elles. Par exemple les proportions de PM du cluster financier 3 et du cluster DTW 2 sont corrélés négativement (-1) avec la proportion d'UC : ces clusters ne possèdent pas de fonds UC et ont donc une proportion d'UC égale à 0. Nous enlevons donc la variable de proportion d'UC (corrélation forte avec les variables de clustering de fonds), les variables d'âge à la souscription (corrélation

avec les variables d'âges actuariels), et la variable de proportion de PM représentée par les contrats possédant un pourcentage d'UC supérieur à 25% (forte corrélation avec les clustering de fonds et la variable de proportion d'UC du produit). Il reste alors 42 variables explicatives des taux d'arbitrages.

De manière générale, les variables séparées en catégories (intervalle de valeurs), comme les variables de clustering de fonds ou bien d'âges, sont corrélés par construction même. Nous nous intéressons donc à des variables provenant d'indicateurs différents. Il en résulte que nous enlevons les variables : proportion d'UC, proportion de PM dont les contrats présentent des proportion d'UC suppérieur à 25%, proportion de PM associées aux catégories d'âge à la souscription.

Cette matrice de corrélation permet aussi de se rendre compte de la qualité de la relation linéaire qui existe entre chaque couple de variable explicative en plus de la cohérence de données. Ainsi certains constats logiques sont fait. Par exemple, la recherche web associée au mot "crise financière" augmente lorsque la volatilité du CAC40 à 3 mois augmente (corrélation de 0.75). D'autres relations, moins évidentes mais déjà misent en lumière précédemment, peuvent être faites : la PM allouée au cluster 1 DTW des fonds haussiers insensibles à la conjecture économique est associée à des contrats possédant plus de 10 supports différents, les femmes sont associées la classe de csp arbitrant le moins (corrélation de 0.75), et les contrats possédant plus de 160k€ de PM présentent une faible proportion de leur PM en fonds à tendance haussières mais dépendant de la conjecture économique (corrélation de 0.5) et font partie de la CSP arbitrant le plus (corrélation avec "csp_haut" de 0.65 et corrélation avec "csp_bas" de -0.73).

Nous pouvons également nous intéresser directement à la corrélation des variables explicatives avec les taux d'arbitrages :

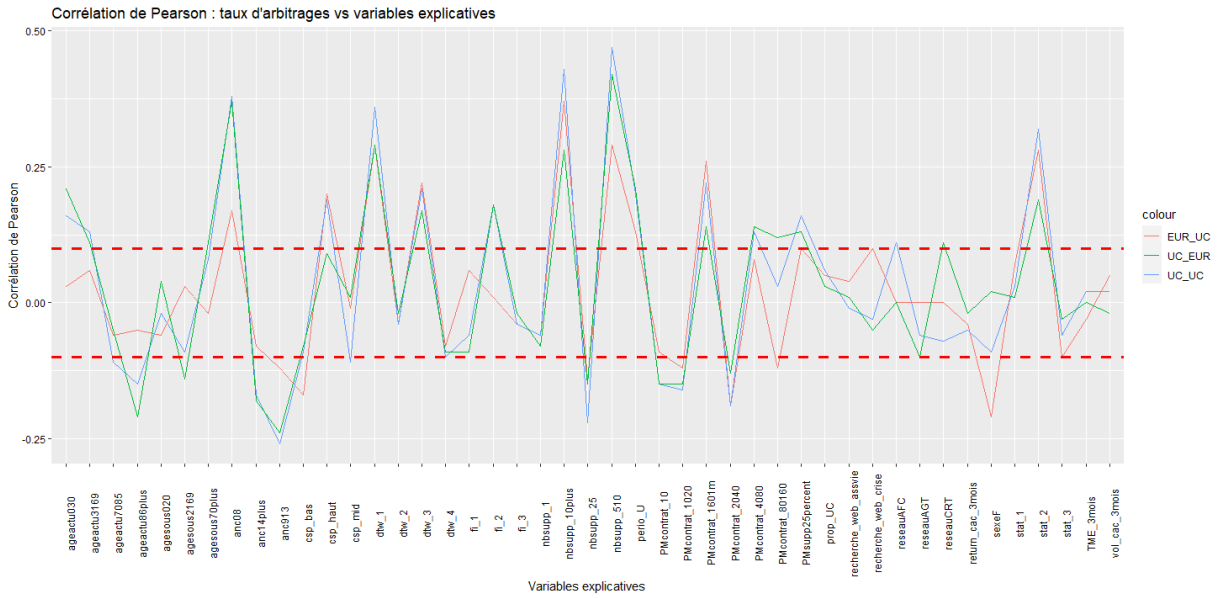


FIGURE 5.36 – Corrélation de Pearson des taux d'arbitrages avec les variables explicatives

Les taux d'arbitrages par origine/destinations sont corrélés (coefficient supérieurs à 10%) avec les variables explicatives globalement. Néanmoins de petites variations existent et les corrélations ne dépassent que rarement 25%, les maximums étant atteints pour les variables "anc08", "dtw_1", "nbsupp10plus", "nbsupp510", et "stat_2".

5.2.3 Approches considérées

Par la suite deux approches seront considérées : une approche globale et une approche spécifique.

L'**approche globale** consistera à modéliser les taux d'arbitrages sur l'ensemble de la base de donnée, c'est à dire que le produit sera considéré comme une variable catégorielle. Les meilleurs modèles de régression linéaires, SVM, forêt aléatoires et XGboost seront mis en compétition en vue de déterminer l'algorithme champion pour chaque origine/destination d'arbitrage. L'avantage de cette méthode est qu'elle présente un coût en temps de calcul modéré, généralise la modélisation des taux, distingue les catégories de risque de produits au sens de l'arbitrage uniquement avec la variable produit (simplification de l'influence du produit sur les taux d'arbitrages), et permet une analyse et une compréhension du modèle et donc de la dynamique d'arbitrage. Le désavantage majeur est que les variables explicatives, autres que la variable produit, sont prises en compte de la même manière pour tous les produits : il n'y a pas de distinctions entre produits de l'influence des variables explicatives et l'analyse à la maille produit devient grossière.

L'**approche spécifique** consistera à séparer la base de données par produit. Pour chaque produit, la meilleure régression linéaire, SVM, forêt aléatoire et XGboost sera retenue et mise en compétition avec les autres en vue de déterminer l'algorithme cham-

pion pour le produit et l'origine/destination de l'arbitrage. L'avantage de cette approche réside dans le fait qu'elle permet la distinction entre produits de l'influences des variables explicatives, et donc permet une analyse plus fine à la maille produit. Le désavantage majeur de cette approche est bien évidemment le coût en temps de calcul bien plus élevé que l'approche précédente et un effort d'analyse multiplié par le nombre de produit, mais aussi un effet "boîte noire" complexifiant l'analyse (25 produits donc 25 modèles différents).

Les **mesures de performances** utilisées seront principalement la RMSE, la MAE, l'AIC (pour la régression linéaire), mais aussi un critère propre à l'origine/destination des taux d'arbitrages regardés. Pour les arbitrages EUR_UC, nous souhaitons pénaliser la sur-estimation des taux, car une sur-estimation de ces taux reviendrait à sous-estimer l'engagement de l'assureur sur les fonds euros. A l'inverse, pour les arbitrages UC_EUR, nous souhaitons pénaliser la sous-estimation de ces taux, car cela reviendrait à sous estimer l'engagement de l'assureur sur les fonds euros. Pour les taux UC_UC, la sur-estimation ou bien la sous-estimation n'a pas de répercussion particulière en terme d'engagement de l'assureur auprès de ses assurés et nous ne proposons donc pas de critère pour cette origine/destination d'arbitrage. Nous proposons donc les deux critères suivant :

$$Crit_{EUR\_UC} = \sum_{i=1}^{n} 3.(y_i - \hat{y_i})^2.\mathbf{1}_{y_i<\hat{y_i}} + (y_i - \hat{y_i})^2.\mathbf{1}_{y_i\geq\hat{y_i}}$$

$$Crit_{UC\_EUR} = \sum_{i=1}^{n} 3.(y_i - \hat{y_i})^2.\mathbf{1}_{y_i>\hat{y_i}} + (y_i - \hat{y_i})^2.\mathbf{1}_{y_i\leq\hat{y_i}}$$

Autrement dit, nous pénalisons trois fois plus les erreurs lorsque celles-ci induisent une sous-estimation de l'engagement de l'assureur auprès de ses assurés sur les fonds euros (sous-estimation des fonds propres réglementaires donc). Ces deux derniers critères serviront à départager les meilleurs algorithmes appartenant à chaque classe d'algorithme entre eux.

5.3 Approche globale

Seuls les modèles champions par origine/destination seront analysés dans cette partie.

5.3.1 Régression linéaire

Si une recherche exhaustive du meilleur modèle au sens d'un critère était effectuée sur les 42 variables explicatives, il faudrait entraîner alors $2^{42} - 1$ modèles : ce qui est en pratique impossible.

La base de donnée comportant beaucoup de variables, il est décidé de sélectionner les variables avec les critères AIC (Akaike Information Criteria, à minimiser) , le RMSE (à minimiser), le R2 ajusté (à faire tendre vers 1). Nous utilisons une procédure "stepwise" :

un modèle complet (comprenant toutes les variables explicatives) est entraîné initialement, puis à chaque itération, une variable est ajoutée ou supprimée si sa p-value est significative ou pas. Les seuils de significativité d'entrée et de sortie d'une variable dans le modèle sont choisis de manière à maximiser le R2 ajusté (le R2 ajusté tiens compte du nombre de variables et ne choisi donc pas le modèle complet). À chaque itération les critère AIC, RMSE, en plus du R2 ajusté, sont calculés. Les résultats de la meilleure régression linéaire pour chaque origine/destination d'arbitrages sont en annexe "Eléments de compréhension des régressions linéaires en approche globale".

Nous obtenons donc les résultats suivants :

Origine/Destination	R2 ajusté	AIC	RMSE a	RMSE t	MAE a	MAE t	Critère a	Critère t	p-value Fisher
UC_EUR	0.754	7789.9	2.267	5.682	1.268	2.934	19735	11350	$<$ 2.2e-16
EUR_UC	0.227	-2353.7	0.1206	0.1504	0.0515	0.0629	32.88	3.551	$<$ 2.2e-16
UC_UC	0.585	9296.6	3.508	2.163	1.650	1.534	x	x	$<$ 2.2e-16

FIGURE 5.37 – Tableau récapitulatif des meilleures régressions linéaires obtenues en approche globale

Le temps de calcul des trois origines/destinations d'arbitrages cumulé est d'environ 20 minutes.

5.3.2 SVM

Les SVM étant capables de tirer parti d'un grand nombre de variables explicatives, nous utilisons les 42 variables explicatives dans ces modèles.

Nous décidons d'optimiser les SVM selon 5 hyper-paramètres :

- *noyau* : ici radial ou polynomial.
- *gamma* : il définit la portée de l'influence d'un seul individu dans l'apprentissage. Les valeurs faibles de gamma signifient "loin" et les valeurs élevées signifient "proche". Nous le faisons varier de 10^{-7} à 10 par puissance de 10.
- *cost* : coût de la violation des contraintes dans le Lagrangien intervenant dans le calcul de la marge maximale. Un coût élevé pénalise fortement la violation de contrainte et le Lagrangien associé s'en retrouve augmenté (minimisation difficile lorsque il y a violation de contrainte donc). Nous le faisons varier de 1 à 3.
- $coef0$: uniquement pour le noyau polynomial par la formule $(gamma * u' * v + coef0)^{degre}$. Nous faisons varier cette constante de -3 à 3 par pas de 0.5.
- *degre* : uniquement pour le noyau polynomial, il s'agit du degré de la régression. Nous le faisons varier de 1 à 5.

Cela revient à entraîner 2178 SVM différents pour chaque origine/destination d'arbitrage. Le temps de calcul des trois origines/destinations d'arbitrages cumulé est d'environ 3h30. Nous retenons comme SVM champions les svm suivants :

Origine/Destination	noyau	degré	gamma	cost	coef0	RMSE a	RMSE t	MAE a	MAE t	Critère a	Critère t
UC_EUR	polynomial	4	0.01	2	2	1.373	4.576	0.576	2.409	8845.5	6989.8
EUR_UC	radial	x	1	3	x	0.083	0.157	0.019	0.0575	12.317	3.311
UC_UC	radial	x	0.1	3	x	2.703	1.798	0.7484	1.226	x	x

FIGURE 5.38 – Tableau récapitulatif des meilleurs SVM obtenus en approche globale

5.3.3 Forêt aléatoire

Nous décidons d'optimiser les forêts aléatoires selon 3 hyper-paramètres :

- *ntree* : le nombre d'arbres de la forêt aléatoire. Nous faisons varier le nombre d'arbres de 10 à 500.
- *mtry* : le nombre de variables échantillonnées aléatoirement comme candidates à chaque fractionnement. Nous faisons varier ce paramètre de 1 à 20.
- *nodesize* : la taille minimale des nœuds terminaux des arbres. Plus ce paramètre augmente, plus les arbres seront plus petits, moins la forêt aléatoire sera sujette au risque de sur-apprentissage. Nous faisons varier ce paramètre de 1 à 10.

Cela représente 4400 forêts aléatoires avec un paramétrage différent pour chaque origine/destination. Le temps de calcul des trois origines/destinations d'arbitrages cumulé est d'environ 4h.

De plus, pour chaque origine/destination, une première étude d'importance des variables est effectuée : nous ne retenons seulement les variables possédant une importance suffisamment élevée afin de diminuer la complexité de la forêt aléatoire.

Les graphiques d'importance des variables des meilleurs modèles servant à la sélection de variable sont en annexe "Importance plot des RF et XGboost en approche globale".

Nous ré-entraînons les modèles sur les nouveaux jeux de variables obtenu. Voici les résultats obtenus par origine/destination :

Origine/Destination	nombre d'arbres	mtry	nodesize	RMSE a	RMSE t	MAE a	MAE t	Critère a	Critère t
UC_EUR	60	3	5	1.755	4.701	0.7356	2.079	12732.8	7948.3
EUR_UC	20	15	4	0.1371	0.1351	0.0505	0.0536	50.85	2.967
UC_UC	70	16	3	3.213	2.263	1.193	1.454	x	x

FIGURE 5.39 – Tableau récapitulatif des meilleures forêts aléatoires obtenus en approche globale

5.3.4 XGboost

Etant basé sur des arbres de décision, les XGBoost capturent tous types de liaisons entre données, y compris les relations non linéaires, à la différence de la régression linéaire.

Nous décidons d'optimiser les XGboost entrainés selon 5 hyperparamètres :

- *nrounds* : le nombre d'itérations de boosting à effectuer, c'est à dire le nombre d'arbres de décisions en chaîne entraînés. Nous faisons varier le nombre d'arbres de 10 à 200.
- *max_depth* : la profondeur d'arbre maximale. Il existe un risque de sur-apprentissage si la profondeur est trop grande, et inversement un risque de sous-apprentissage si la profondeur est trop petite. Nous regardons une profondeur maximale allant de 1 à 20.
- *colsample_bytree* : le pourcentage de variables explicatives prises en compte pour construire un arbre. Un pourcentage de 10% à 80% est étudié ici.
- *learning_rate* : le taux d'apprentissage qui contrôle la vitesse à laquelle la descente du gradient se fait. Nous faisons varier ce taux de 0.1 à 1.
- *gamma* : la diminution minimale de la valeur de la fonction de coût pour prendre la décision de partitionner une feuille d'un arbre ou non. Nous faisons varier ce paramètre de 10^{-3} à 1 par puissance de 10.

Cela représente 11200 XGboost avec un paramétrage différent pour chaque origine/destination. Le temps de calcul des trois origines/destinations d'arbitrages cumulé est d'environ 6h.

De plus, pour chaque origine/destination, une première étude d'importance des variables est effectuée : nous ne retenons seulement les variables possédant une importance suffisamment élevée afin de diminuer la complexité du XGboost. Les graphiques d'importance des variables des meilleurs modèles sont en annexe "Importance plot des RF et XGboost en approche globale".

Nous ré-entraînons les modèles sur les nouveaux jeux de variables obtenus. Voici les résultats obtenus par origine/destination :

Origine/Destination	nombre d'arbres	maxdepth	eta	gamma	colsample_bytree	RMSE a	RMSE t	MAE a	MAE t	Critère a	Critère t
UC_EUR	60	4	1	1	60%	0.3443	4.338	0.2323	1.886	431.9	6883.1
EUR_UC	50	2	1	0.1	50%	0.0732	0.1243	0.0384	0.0477	14.65	2.25
UC_UC	10	8	0.1	1	80%	2.729	1.219	1.093	0.7678	x	x

FIGURE 5.40 – Tableau récapitulatif des meilleurs XGboost obtenus en approche globale

5.3.5 Bilan : approche globale

Le tableau suivant résume les résultats des meilleurs algorithmes en approche globale :

	algorithme	RMSE a	RMSE t	MAE a	MAE t	Critère a	Critère t
UC_EUR	Reg.lin	2.267	5.682	1.268	2.934	19735	11350
	SVM	1.373	4.576	0.5762	2.409	8845	6989
	RF	1.755	4.701	0.7356	2.079	12733	7948
	Xgboost	0.3443	4.338	0.2323	1.886	432	6883
EUR_UC	Reg.lin	0.1206	0.1504	0.0515	0.0629	32.88	3.55
	SVM	0.0830	0.1570	0.0190	0.0575	12.32	3.31
	RF	0.1371	0.1351	0.0505	0.0536	50.85	2.97
	Xgboost	0.0732	0.1243	0.0384	0.0477	14.65	2.25
UC_UC	Reg.lin	3.508	2.163	1.650	1.534	x	x
	SVM	2.703	1.798	0.7484	1.226	x	x
	RF	3.213	2.263	1.193	1.454	x	x
	Xgboost	2.729	1.219	1.093	0.7678	x	x

FIGURE 5.41 – Tableau récapitulatif des meilleurs modèles obtenus en approche globale

UC_EUR : Aucun algorithme n'est vraiment satisfaisant puisqu'ils présentent tous des erreurs sur la base de test largement supérieures à la base d'apprentissage. Nous retenons cependant le XGboost qui reste le plus performant.

EUR_UC : nous retenons le XGboost, qui est meilleur à tout les niveaux.

UC_UC : nous retenons ici encore le XGboost. Sans conteste, ici, il s'agit du meilleur.

5.3.6 Analyse des modèles : approche globale

Rappelons tout dabord que, ici, les résidus sont la différence entre les valeurs observées de taux d'arbitrages et les valeurs prédites. Les QQ-plot confrontent en abscisse les quantiles théoriques d'une loi normale centrée réduite aux quantiles empiriques des résidus en ordonnée. Si le QQ-plot présente des points alignés sur la droites d'équation $y = x$ alors cela signifie que les résidus semblent suivre une loi normale centrée réduite (un bruit blanc donc).

5.3.6.1 UC_EUR : XGboost

Voici quelques éléments d'analyse des résultats fournit par l'approche globale sur les taux UC_EUR :

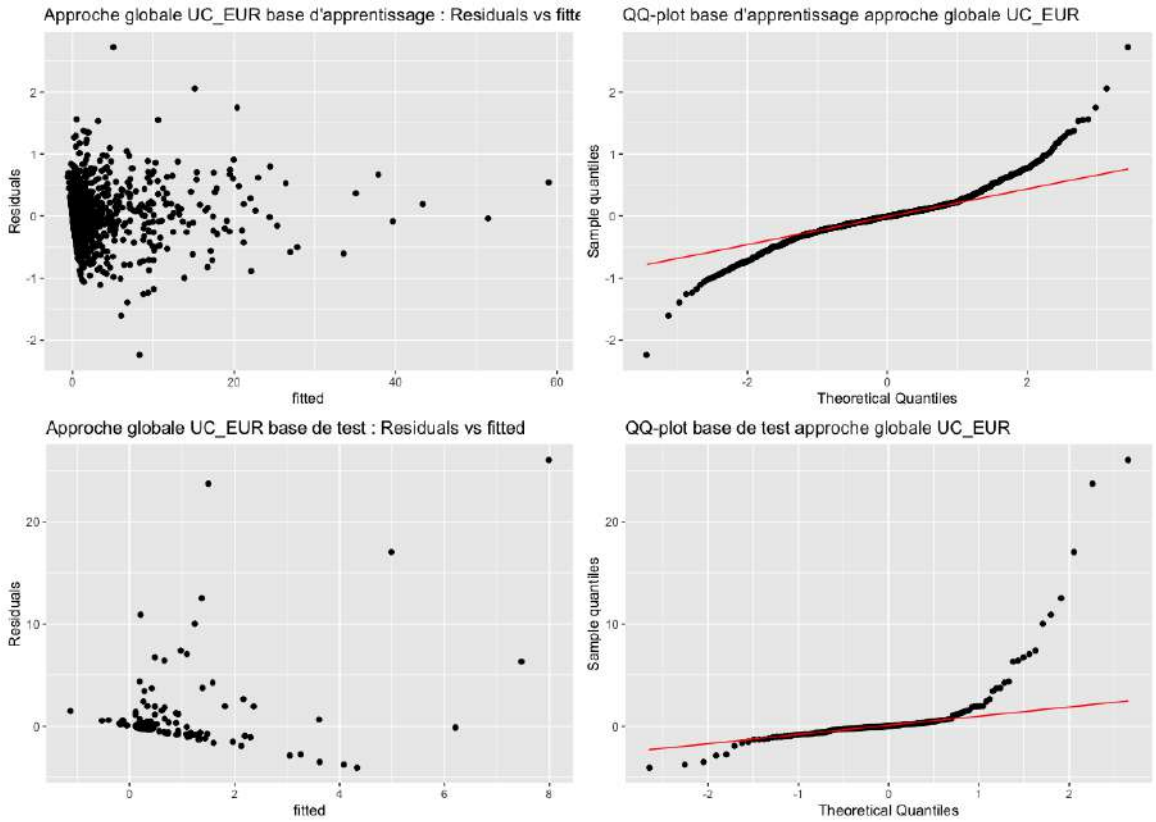


FIGURE 5.42 – Quelques éléments d'analyses des résultats de l'XGboost champion modélisant les taux UC_EUR en approche globale

Globalement les erreurs sont centrées, avec une masse plus importante d'erreurs sur les valeurs prédites faibles : cela est dû au fait que la majorité des observations sont situées dans cet ordre de grandeur. Si les résultats sur la base d'apprentissage sont plutôt

bons, ils le sont moins sur la base de test : l'algorithme, loin d'être parfait, n'arrive pas à généraliser sur la base de test et de grands résidus sont observés. Sur celle-ci, la queue de distribution supérieure plus lourde que l'inférieure nous indique une sous-estimation générale des taux prédits et alors le $Citere_{UC\_EUR}$ ne rempli pas tout à fait son rôle de pénalisation de la sous-estimation des taux. On peut également expliquer cela par le fait que la base d'apprentissage présente peu de taux extrêmes au regard du nombre de taux normaux, et que donc l'algorithme est réticent à prédire des valeurs élevées. De manière générale, et nous le verrons également avec l'approche spécifique, les taux UC_EUR présentent la modélisation la moins satisfaisante. Le graphique ci-dessous indique l'importance des variables explicatives dans l'XGboost au sens du gain (voir partie 3.1.7) :

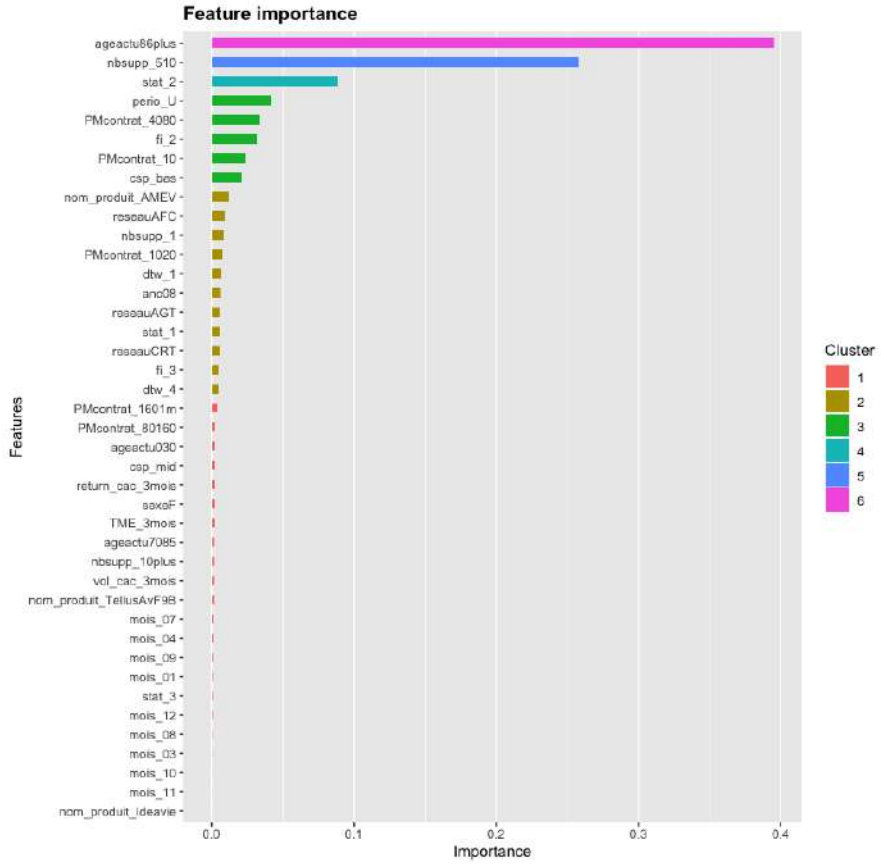


FIGURE 5.43 – Importance des variables de l'XGboost champion modélisant les taux UC_EUR en approche globale

L'XGboost met donc ici en avant les variables *ageactu86plus*, *nbsupp_510*, *stat_2*, *perio_U*, *PMcontrat_4080*, *fi_2*, *PMcontrat_10* et *csp_bas*. Remarquons que les variables relatives aux produits ne sont pas bien représentées : l'algorithme ne fait donc pas de distinction entre produits.

Pour savoir comment l'XGboost utlise ces variables nous appliquons la méthode de SHAP :

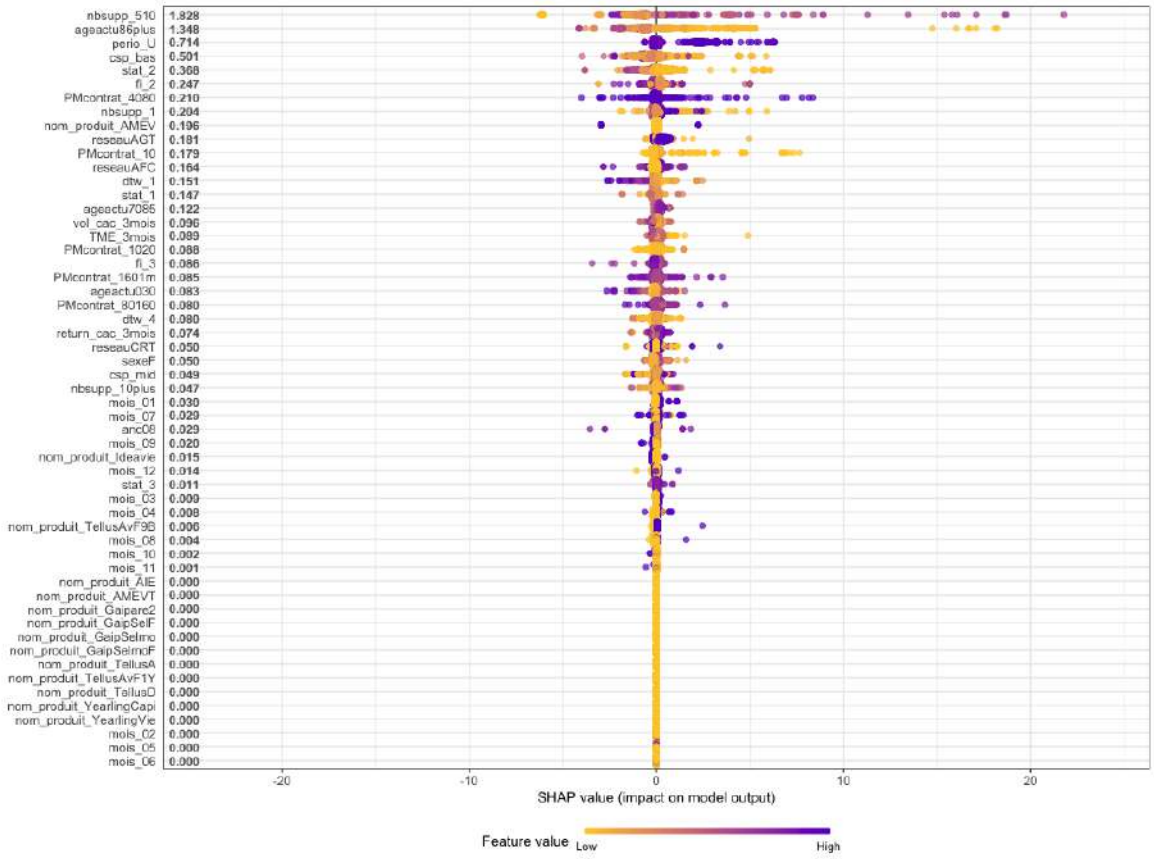


FIGURE 5.44 – *SHAP summary plot* de l'XGboost en approche globale pour les taux UC_EUR

Voici comment interpréter ce graphique :

L'axe des y indique le nom de la variable, par ordre d'importance de haut en bas. La valeur numérique à côté est la valeur SHAP moyenne.

Sur l'axe des x figure la valeur SHAP. A partir de ce nombre, nous pouvons extraire le sens et la force de variation de la prédiction.

La couleur du gradient indique la valeur originale de la variable présente sur l'axe des y. Pour les variables booléennes (*one hot encoder*), il s'agira de deux couleurs, mais dans le cas d'une variable quantitative, il peut contenir tout le spectre de gradient de couleur.

Chaque point représente une ligne de la base d'apprentissage.

Ainsi, pour la variable *dtw_1* (fonds à tendance haussière insensible à la conjecture économique), une valeur élevée de cette variable induit un impact négatif modéré sur les taux d'arbitrages UC_EUR. **Autrement dit, les assurés ont confiance dans ces UC et n'arbitrent pas vers l'euros lorsqu'ils ont en leur possession un fort**

pourcentage de leur PM dessus. Cela induit que les assurés ont une notion de "confiance" dans les UC en fonction de la dynamique d'évolution ("oeil géométrique") de ceux-ci. Dès lors, l'introduction de variables de clustering de fonds UC dans les modèles de prédictions s'avère être pertinente puisque ce pattern sous-jacent est détecté, appris, et considéré comme important en terme de pouvoir explicatif sur les taux d'arbitrages UC_EUR dans l'XGboost.

Ce graphique, exhaustif mais coûteux en effort d'analyse, peut être affichée d'une autre manière, sur une sélection de variables présentant une relation avec les taux d'arbitrages UC_EUR, sur le graphique suivant (qui s'analyse de la même manière, le gradient étant remplacé par l'axe des abscisses) :

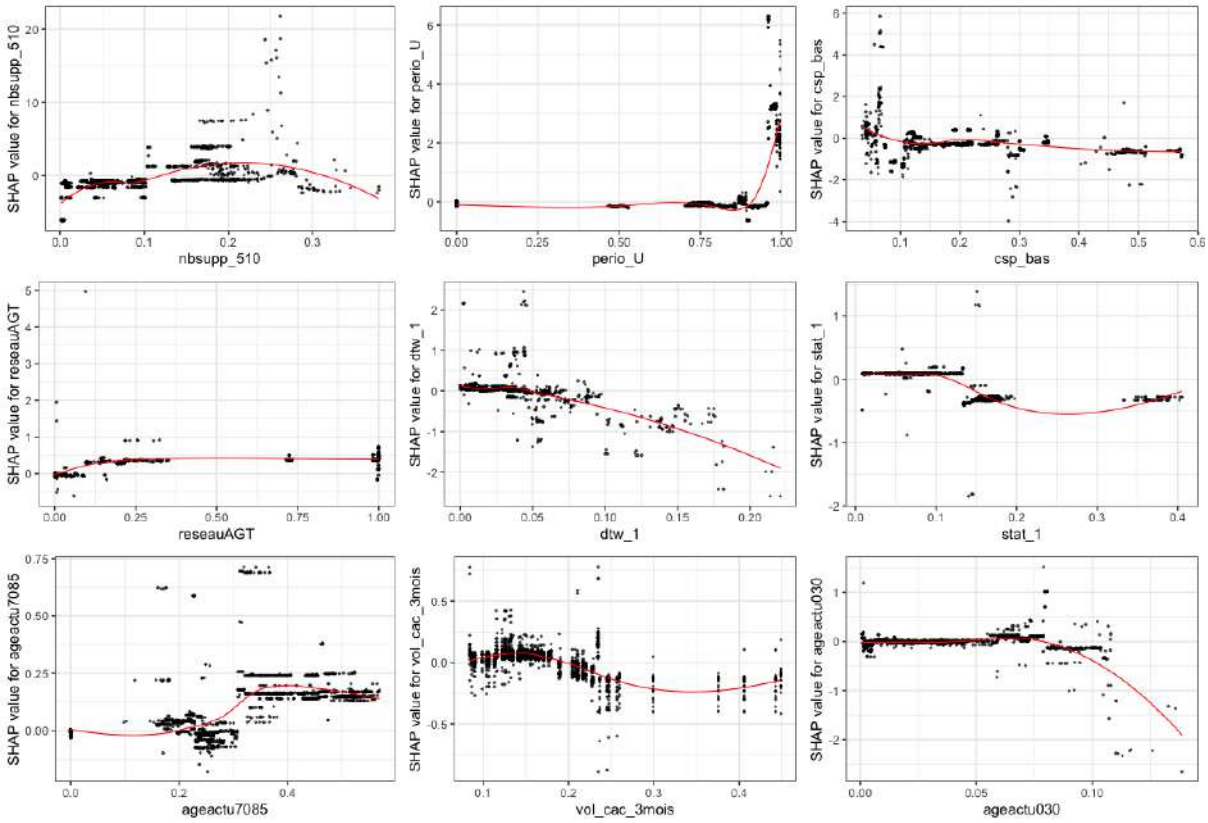


FIGURE 5.45 – *SHAP plot* de quelques variables explicatives de l'XGboost présentant une relation avec les taux d'arbitrages UC_EUR

La largeur satisfaisante du *SHAP summary plot* nous indique qu'il est possible de dégager de l'information avec une analyse de ce qu'a appris l'XGboost :

- *nbsupp_510* : relation non linéaire, en forme de colline. Pour des pourcentages de PM associé à des contrats ayant un nombre de support différents allant de 5 à

10, entre 0-10% et 28%-100%, les taux d'arbitrages prédits sont faibles. Si cette proportion de PM se situe entre 10 et 28%, alors les taux prédits seront plus hauts.

- *perio_U* : à partir de 90% de PM associée à des contrats présentant une prime unique, les taux prédits sont beaucoup plus haut.

- *reseauAGT* : pour des pourcentages faibles de PM associée à des contrats appartenant au réseau de distribution d'agence, les taux prédits diminuent. Les autres réseaux ont donc des taux d'arbitrages perçu par le modèle plus fort.

- *dtw_1* : plus la proportion d'UC est sur les fonds à tendance haussière et insensible à la conjecture économique, moins il y a d'arbitrages de l'UC vers l'euros. Les assurés ont donc confiance en leur insensibilité et en leur tendance haussière afin de réaliser un rendement dans le futur.

- *stat_1* : deux paliers séparés par une proportion de 14% de fonds UC appartenant à des fonds UC à tendance haussière mais instables ("fonds investissement risque modéré"). Cela indique que à partir de 14% de PM sur ces fonds, les assurés auront tendance à moins arbitrer. Cela confirme une certaine paralysie de l'assuré à arbitrer vers l'euros lorsque celui possède une part non négligeable d'UC instable.

- *ageactu7085* : relation non linéaire, en palier. À partir de 30% de PM associé à des assurés ayant un âge actuariel compris entre 70 et 85 ans, les taux d'arbitrages augmentent. Cela correspond à la sécurisation de l'épargne acquise vers les fonds euros lorsque l'âge augmente. Ce constat était également constaté dans la partie statistique descriptive.

- *vol_cac_3mois* : relation en sinusoïde. Globalement, à partir d'une volatilité du CAC40 de 22%, les taux diminuent car les assurés souhaitent bénéficier de l'instabilité des marchés pour faire un meilleur rendement (risque acceptable). À partir d'une volatilité de 35% ces taux remontent : le risque perçu par l'assuré est trop fort et il préfère sécuriser ses capitaux. Ces paliers peuvent donc être associés à la quantification des paliers de risque auxquels l'assuré est disposé à encourir.

- *ageactu030* : à partir de 8% de PM associée à des assurés ayant entre 0 et 30 ans, les taux d'arbitrages vers l'euros chutent. En effet, les moins de 30ans ont une proportion d'UC plus élevé car souhaitent acquérir une épargne.

Ici, la méthode SHAP permet de quantifier l'importance ou non des variables explicatives sur les taux d'arbitrages UC_EUR. Elle met donc en évidence la nature non linéaire régissant les taux UC_EUR, et par conséquent montre en quoi les méthodes classiques, comme les régressions linéaires, sont vouées à l'échec concernant l'estimation des taux d'arbitrages. La méthode SHAP met également en évidence la capacité d'apprentissage du XGboost : celui-ci semble avoir appris certaines mécaniques de comportement d'arbitrage misent en lumière dans la partie statistique descriptive. Cette méthode

met également en évidence les faiblesses de notre modèle : les interactions entre variables ne sont pas tout à fait "comprises". En effet, prenons l'exemple de la variable *vol_ cac_ 3mois* : pour une valeur de variable donnée (axe des abscisses), il n'est pas rare de constater une étendue associée relativement grande sur l'axe des y. Cela met en évidence les phénomènes d'interactions entre variables qui peuvent ajouter du bruit à nos prédictions et compliquer l'interprétabilité du modèle.

Si nous regardons le *break down profile plot* des 4 valeurs extrêmes des taux d'arbitrages :

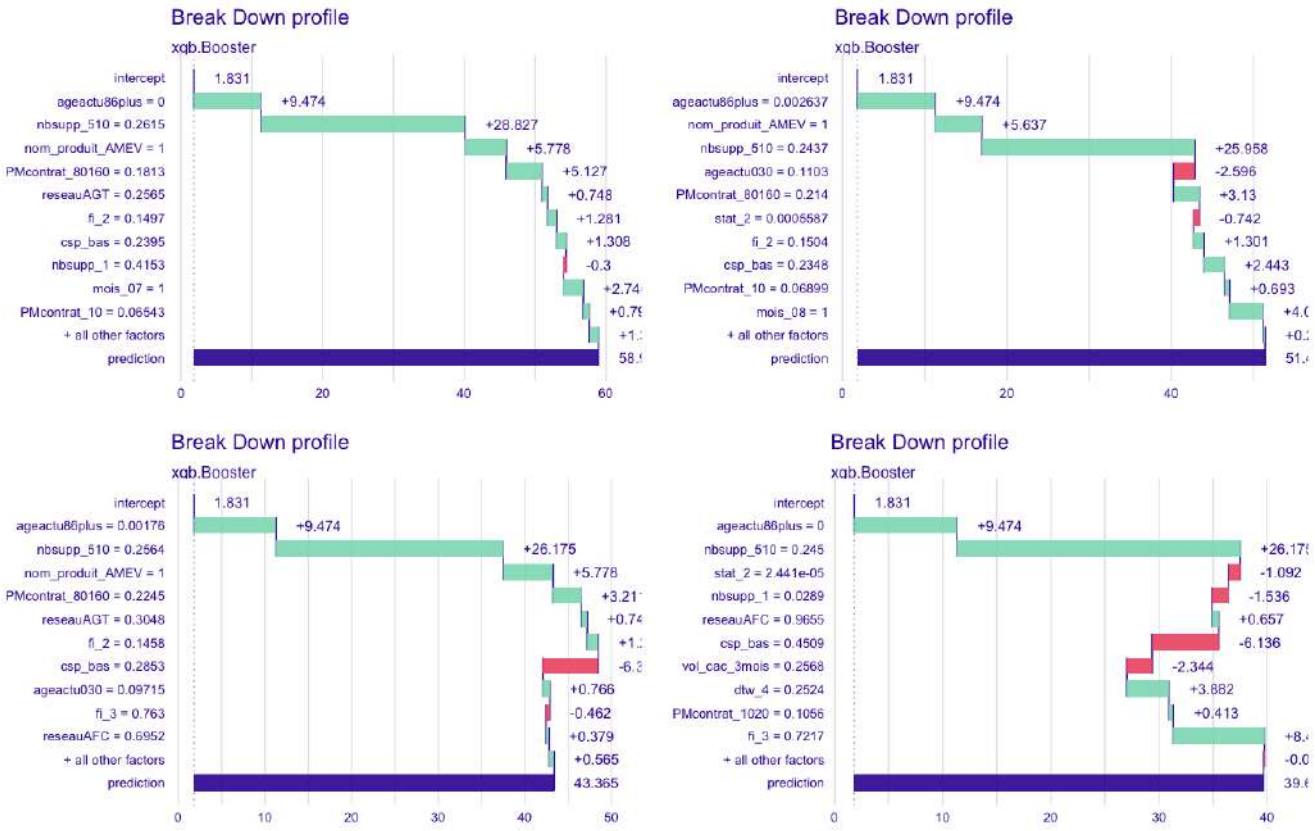


FIGURE 5.46 – *Break down profile plot* associé aux quatre taux d'arbitrages UC_EUR les plus élevés

Ce graphique décompose la prédiction de l'XGboost en attribuant à chaque variable explicative sa "part de responsabilité"/contribution dans la valeur prédite de taux d'arbitrages en fonction de sa valeur pour l'observation d'étude. Pour donner une idée, et par analogie avec la régression linéaire, cela correspondrait au prédicteur linéaire, sauf que ici c'est le "prédicteur linéaire de l'XGboost associé à l'observation". Les détails du calcul de ces contributions sont donnés dans le mémoire de D.Delcaillau [13] , page 96 à

100.

Pour les taux UC_EUR, les variables *nbsupp_510*, *ageactu86plus*, *nom_produit_AMEV* (AMEV étant un produit un plus fort taux d'UC que les autres produits), *PMcontrat_80160*, *mois_07* et *mois_03* semblent avoir un impact particulièrement élevé sur les prédictions de taux hauts, ce qui est cohérent avec l'analyse univariée.

5.3.6.2 EUR_UC : XGboost

Voici quelques éléments d'analyse des résultats fournit par l'approche globale sur les taux EUR_UC :

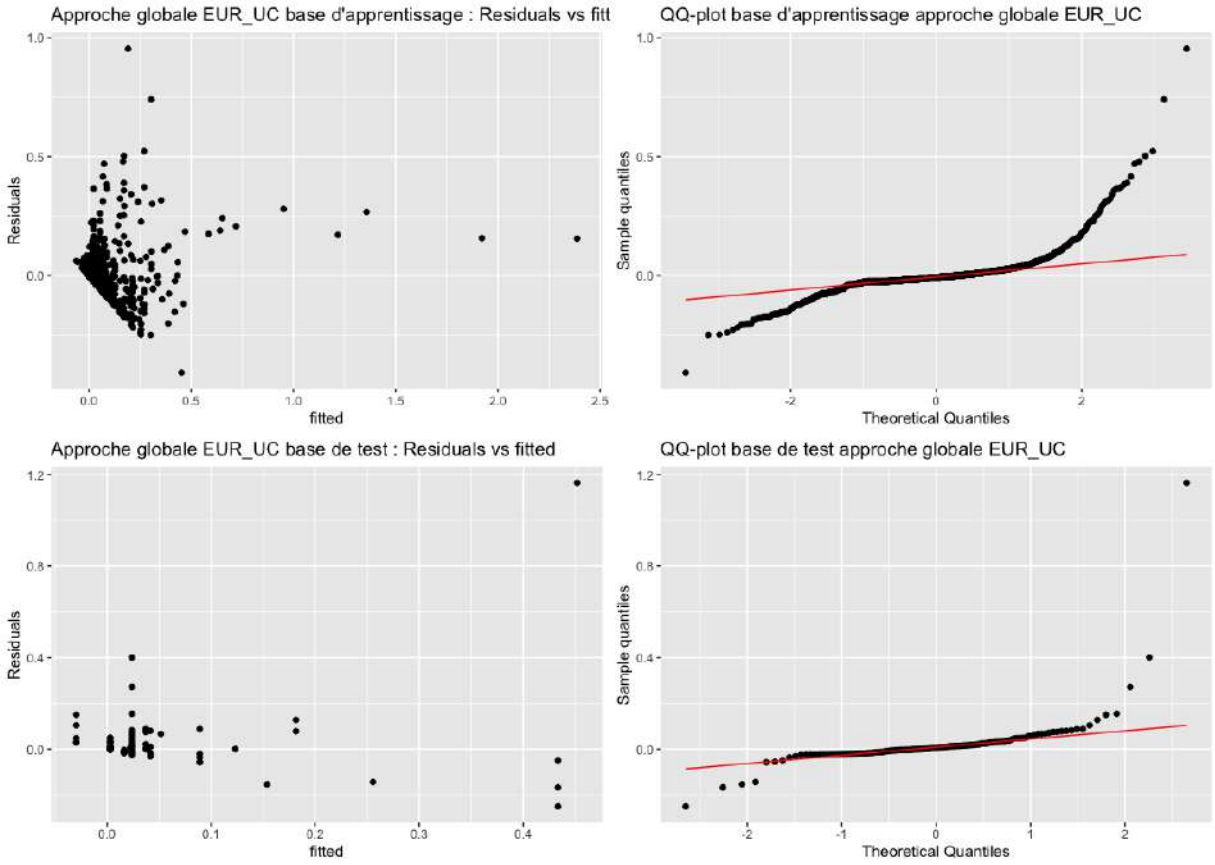


FIGURE 5.47 – Quelques éléments d'analyses des résultats de l'XGboost champion modélisant les taux EUR_UC en approche globale

Ici, les résidus sont centrés mais présentent une structure en forme de cône jusqu'à des valeur prédites de 0.5% sur la base d'apprentissage. Les QQ-plot présentent des queues inférieures moins lourdes que les queues supérieures : il y a donc moins de résidus faibles, c'est à dire moins de sur-prédiction. Cela montre l'impact positif du $Critere_{EUR\_UC}$ qui rempli ici son rôle. Ici le modèle fournit des prédictions peu satisfaisante sur les valeurs faibles et mdérées de taux d'arbitrages, mais plutôt convainquant sur les taux d'arbitrages élevés.

Le graphique ci-dessous indique l'importance des variables explicatives dans l'XGboost au sens du gain :

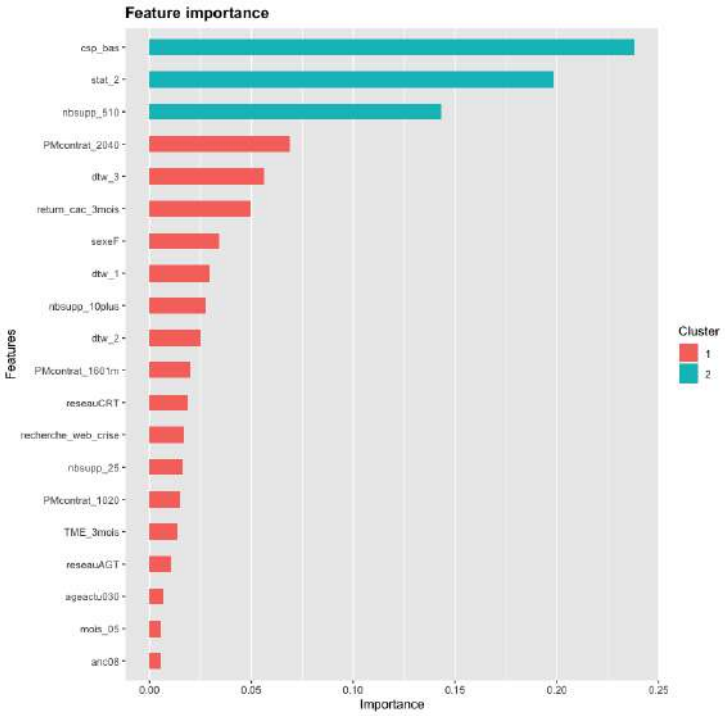
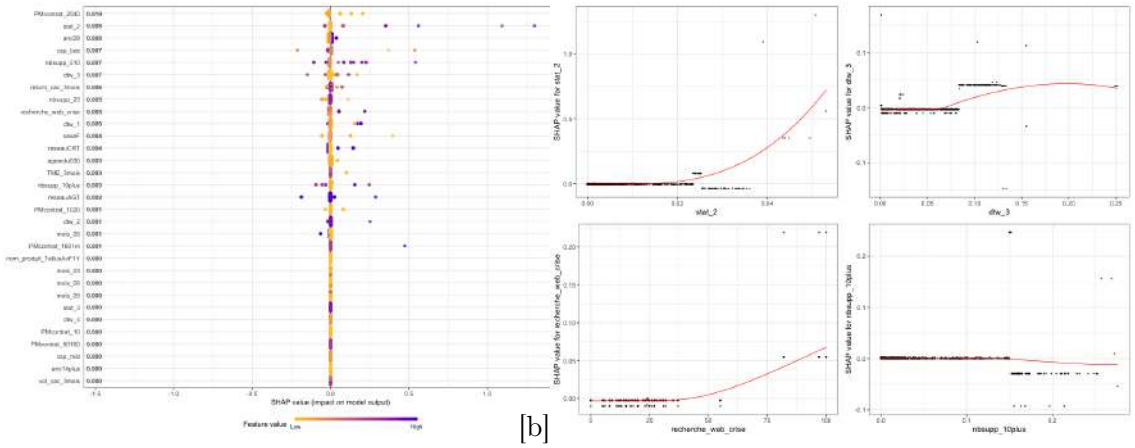


FIGURE 5.48 – Importance des variables de l'XGboost champion modélisant les taux EUR_UC en approche globale

L'XGboost met donc ici en avant les variables *csp_bas*, *stat_2*, *nbsupp_510*, *PMcontrat_4080*, *dtw_3*, et *return_cac_3mois*. Remarquons que les variables relatives aux produits ne sont pas bien représentées : l'algorithme ne fait donc pas de distinction entre produits.

Pour savoir comment l'XGboost utlise ces variables nous appliquons la méthode de SHAP et analysons les réslutats de la même manière que pour les taux EUR_UC :



[a] [b]

FIGURE 5.49 – *SHAP summary plot* (a) de l'XGboost en approche globale pour les taux EUR_UC et *SHAP plot* de quelques variables explicatives correspondantes

Les taux EUR_UC sont globalement assez mal compris : le *SHAP summary plot* est moins "large" (plus centré en $x = 0$) que celui des taux UC_EUR et donc nous dégageons moins d'informations de l'analyse de ce qu'a appris l'XGboost :

- *stat_2* : relation non linéaire en forme de paliers. Pour des proportions de PM correspondantes à des fonds UC à des investissements risqués inférieur allant de 2.5% à 4% les taux prédits sont faibles. À partir de 4% les taux prédits s'envolent et marquent la volonté des assurés de réaliser un rendement avec ces fonds.
- *dtw_3* : relation non linéaire en forme de paliers. À partir de 9% de PM associée à des fonds volatiles sans tendance particulière, les taux d'arbitrages augmentent. Il s'agit du même constat que pour la variable *stat_2*.
- *recherche_web_crise* : lorsque la recherche web Google associée au mot clef "crise financière" est élevée (supérieure à 75%) alors les taux d'arbitrages de l'euros vers l'UC sont très nettements supérieurs.
- *nbsupp_10plus* : plus la proportion de PM associée à des contrats présentant un nombre de support supérieur à 10 est grande, plus les taux prédits de l'euros vers l'UC sont faibles. Cela est le simple constat que plus l'assuré possède d'UC différents, plus il va préférer arbitrer entre fonds UC selon les performances de ceux-ci et considérer le fond euros comme dernier recours.

Si nous regardons le *break down profile plot* des valeurs extrêmes des taux d'arbitrages :

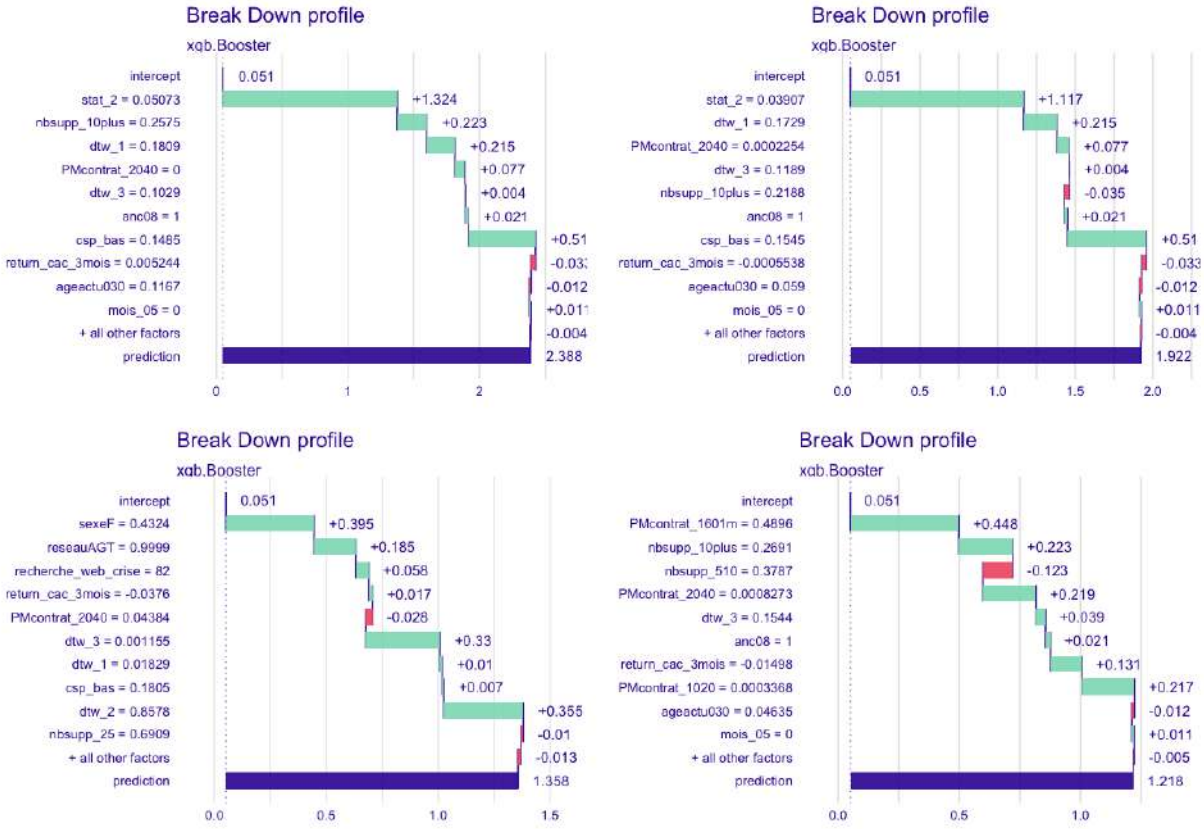


FIGURE 5.50 – *Break Down profile plot* associé aux quatre taux d'arbitrages EUR_UC les plus élevés

Pour les taux EUR_UC, les variables *stat_2*, *dtw_1*, *dtw_2*, *dtw_3*, *csp_bas*, mais aussi quelques variables de PM de contrats, globalement semblent avoir un impact particulièrement élevé sur les prédictions élevée. La dynamique des UC combinée à la PM des contrats est donc prise en compte très fortement dans ce modèle.

5.3.6.3 UC_UC : XGboost

Voici quelques éléments d'analyse des résultats fournit par l'approche globale sur les taux UC_UC :

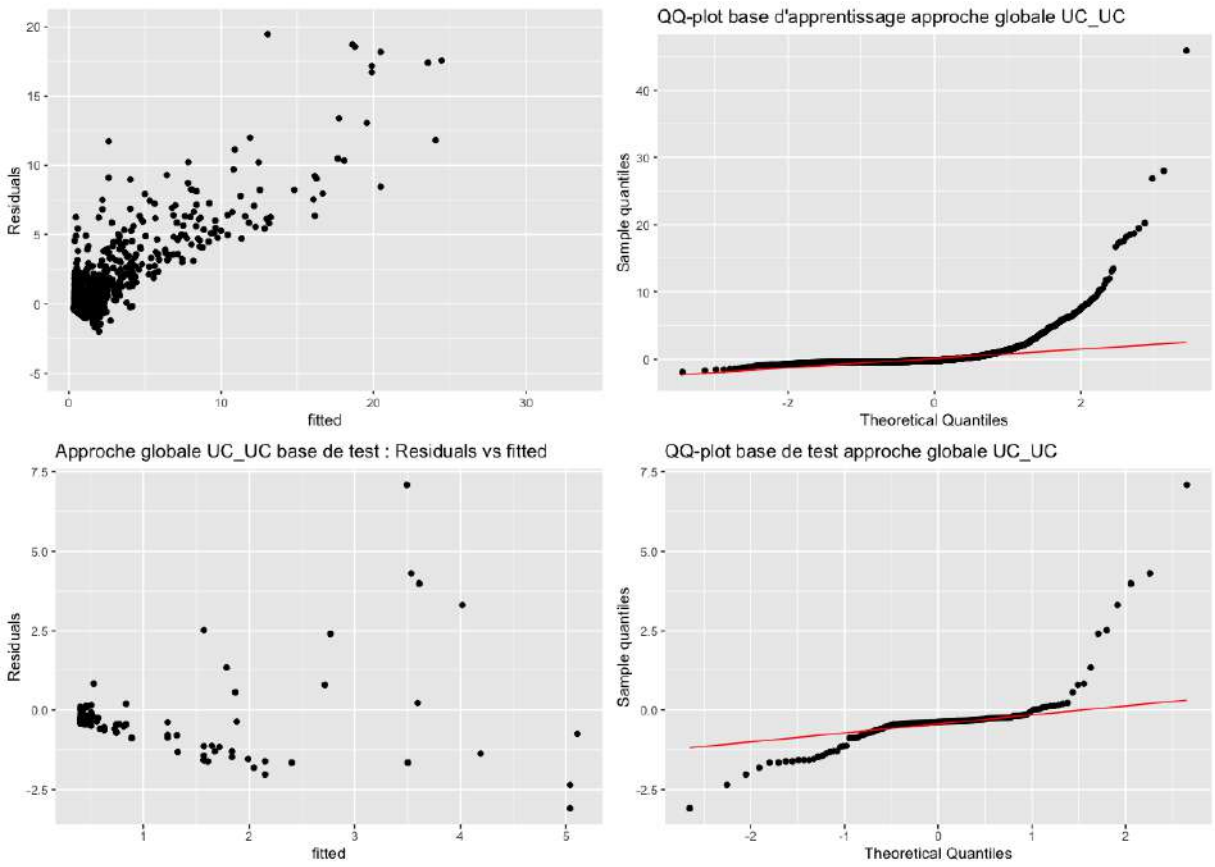


FIGURE 5.51 – Quelques éléments d'analyses des résultats de l'XGboost champion modélisant les taux UC_UC en approche globale

Ici les résidus sur la base d'apprentissage ont une structure linéaire : plus les taux prédits sont hauts, plus les l'erreur est grande (d'où l'existence d'une unique queue supérieure lourde sur le QQ-plot de la base d'apprentissage). Le XGboost présente donc un sous-apprentissage manifeste et passe à côté d'un pattern sous-jacent. Cependant cela n'est pas constaté sur la base de test.

Le graphique ci-dessous indique l'importance des variables explicatives dans l'XGboost au sens du gain :

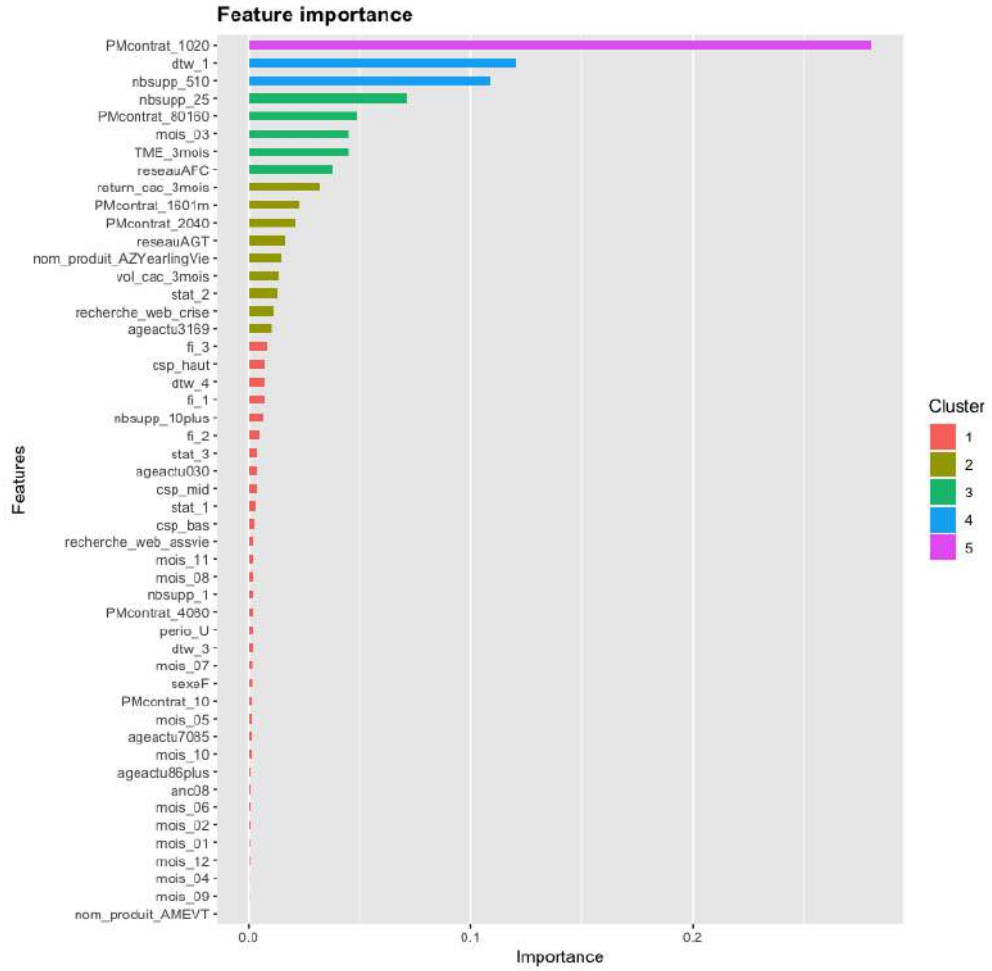
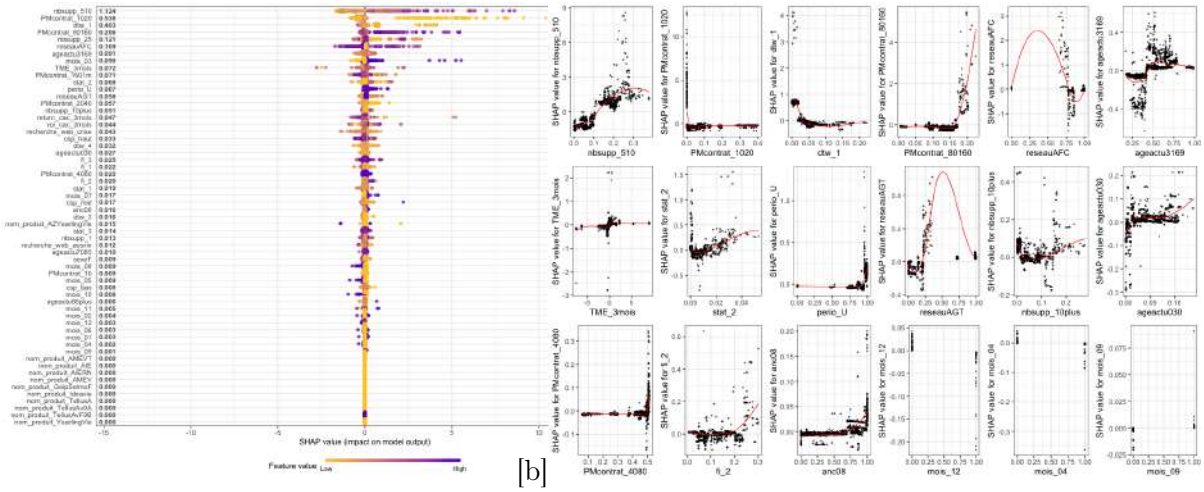


FIGURE 5.52 – Importance des variables de l'XGboost champion modélisant les taux UC_UC en approche globale

L'XGboost met donc ici en avant les variables *PMcontrat_1020*, *dtw_1*, *nbsupp_510*, *nbsupp_25*, *PMcontrat_80160*, *mois_03*, et *TME_3mois*. Remarquons que les variables relatives aux produits ne sont pas bien représentées à l'exception du produit AZYearlingVie : l'algorithme ne fait donc pas de distinction entre produits.

Pour savoir comment l'XGboost utlise ces variables nous appliquons la méthode de SHAP et analysons les réslutats de la même manière que pour les taux UC_UC. La largeur satisfaisante du *SHAP summary plot* nous indique qu'il est possible de dégager de l'information avec une analyse de ce qu'a appris l'XGboost :



[a] [b]

FIGURE 5.53 – *SHAP summary plot* (a) de l'XGboost en approche globale pour les taux UC_UC et *SHAP plot* de quelques variables explicatives correspondantes

Il y a de nombreux constats résultant de la figure de droite, tous cohérents avec l'analyse univariée menée ultérieurement :

- Les mois d'avril et de décembre ont un impact négatifs sur les taux d'arbitrages UC_UC : pour ces mois les valeurs prédites des taux sont faibles. À l'inverse, le mois de septembre est associé à de forts arbitrages.

- Plus la proportion d'UC associée à des contrats dont l'ancienneté est inférieure à 8 ans, plus les taux sont élevés, avec une distinction nette pour un pourcentage de PM de 75%. En effet, la phase 0-8 ans est considérées comme une phase d'acquisition de l'épargne, et donc de bons rendements sont recherchés avec des arbitrages inter UC.

- Pour des proportions supérieures à 3% de PM sur les fonds à tendance haussières insensibles à la conjecture économique (*dtw_1*), les taux d'arbitrages sont faibles : les assurés sont satisfait du rendement que ces fonds présentent et tiennent leur position. Plus la proportion de PM associée à des fonds volatiles à tendance globalement nulle (*stat_2*), plus les assurés arbitrent entre fonds UC : la nature volatile impactant négativement comme positivement le rendement du contrat, les assurés désinvestissent les fonds volatiles sous performant au profit d'autres fonds UC.

- Pour des proportions de PM associées associées à des contrats ayant des PM entre 40 et 160k€, les taux d'arbitrages sont élevés sur les valeurs extrêmes hautes. La classe de contrat dont la PM va de 40k€ à 160k€ est donc celle arbitrant le plus en inter UC.

- Pour une proportion inférieure à 75% de contrats appartenant au réseau AFC les taux sont élevés. Cela est cohérent avec les statistiques descriptives : moins il y

a de contrats appartenant au réseau AFC, plus il y a de contrats appartenant au réseau AGT (qui présente des taux UC_UC plus élevés).

- Concernant la proportion de PM associée à des contrats possédant plus de 5 supports, les taux UC_UC augmentent lorsque cette proportion augmente. Cela illustre le fait que pour les contrats présentant beaucoup de supports différents, les assurés préfèrent arbitrer entre fonds UC lorsqu'un fond UC sous-performe, plutôt que de rapatrier les fonds vers le fond euros.

Si nous regardons le *break down profile plot* des valeurs extrêmes des taux d'arbitrages :

FIGURE 5.54 – *Break Down profile plot* associé aux quatre taux d'arbitrages UC_UC les plus élevés

Pour les taux UC_UC, les variables *nbsupp_510*, *nbsupp_25*, *dtw_1*, *TME_3mois*, *mois_07* et *mois_03* semblent avoir un impact particulièrement élevé sur la prédiction élevée.

5.4 Approche spécifique

Pour chaque produit et pour chaque origine/destination d'arbitrage, nous choisissons la meilleure régression linéaire, le meilleur SVM, la meilleure forêt aléatoire et le meilleur XGboost au sens de la RMSE, et du MAE sur les bases de test et d'entraînement. Ensuite, pour chaque produit et origine destination, l'algorithme champion est choisi. L'algorithme champion est le meilleur algorithme parmi la meilleure régressions linéaire, le meilleur SVM, la meilleure forêt aléatoire et le meilleur XGboost. Afin de déterminer l'algorithme champion, nous regardons la RMSE, la MAE, mais aussi le score au sens du critère propre à l'origine/destination des taux d'arbitrages regardés énoncé plus haut. Les hyperparamètres optimisés sont les mêmes que ceux de l'approche globale, néanmoins nous décidons de réduire le nombre d'hyperparamètres à tester afin de réduire les temps de calculs. Par exemple, pour les XGboost, nous testons un nombre d'arbre allant jusqu'à 100 (contre 200 dans l'approche globale). Il n'est pas non plus nécessaire d'avoir autant d'arbres pour l'approche spécifique puisque la taille de la base de donnée d'apprentissage de chaque produit ne peut excéder 72 lignes ($produit \times mois$).

Ainsi, pour chaque produit nous obtenons l'algorithme "champion" sur la base d'apprentissage et de test. Il ne reste plus qu'à regrouper les algorithmes champions de chaque produit, de compiler leurs prédictions afin de comparer avec l'algorithme champion de l'approche globale, pour chaque origine/destination d'arbitrage.

Tout ces calculs étant effectués en automatique, nous n'avons pas la main sur chaque produit sur le choix de l'algorithme champion. Si nous décidions de prendre l'algorithme minimisant la RMSE, la MAE et le critère spécifique à l'origine/destination de l'algorithme en faisant une somme simple de ces indicateurs, nous nous exposerions au risque de choisir systématiquement un algorithme champion ayant "sur-appris". En effet, la base d'apprentissage étant les données de 2015 à 2020, et la base d'apprentissage étant celle des données jusqu'à mai 2021, si un algorithme minimise la RMSE sur la base d'apprentissage, alors cet algorithme sera favorisé même si ses prédictions sont très mauvaises sur la base de test (les fortes erreurs sur la base de test n'auront que peu d'impact vis à vis de la minimisation sur la base d'entraînement). Nous décidons alors de pénaliser le sur-apprentissage des algorithmes en affectant un poids plus élevé aux mesures de performances réalisées sur la base de test mais aussi en regardant le classement des algorithmes sur chaque mesure de performance plutôt que la valeur. Ainsi l'algorithme champion pour un produit est l'algorithme minimisant $Rank_{RMSE_{B_a}} + 1.5 * Rank_{RMSE_{B_t}} + Rank_{MAE_{B_a}} + 1.5 * Rank_{MAE_{B_t}} + 1.5 * Rank_{Crit_{B_a}} + 3 * Rank_{Crit_{B_t}}$ avec $Rank_i$ le classement de l'algorithme sur la mesure de performance i. De cette manière toutes les mesures de performances peuvent être prises en considération (elles présentent des échelles différentes et donc nous préférons prendre le classement plutôt que la valeur), et nous évitons de choisir un algorithme ayant sur-appris (plus grand poids affecté aux bases de test).

5.4.1 taux UC_EUR

Voici les résultats obtenus en approche spécifique, sur la prédiction des taux UC_EUR :

Approche	RMSE a	RMSE t	MAE a	MAE t	Critère a	Critère t
Spécifique	1.233	3.669	0.4205	1.451	6698	4779
Globale (XGboost)	0.3443	4.338	0.2323	1.886	432	6883

FIGURE 5.55 – Tableau récapitulatif des résultats obtenus en approche spécifique sur les taux UC_EUR

Le temps de calcul s'élève à 8h environ. Ici, l'approche spécifique est plus performante que l'approche globale. L'impact des variables explicatives doit donc être fonction du produit.

5.4.2 taux EUR_UC

Voici les résultats obtenus en approche spécifique, sur la prédiction des taux EUR_UC :

Approche	RMSE a	RMSE t	MAE a	MAE t	Critère a	Critère t
Spécifique	0.0824	0.1497	0.0251	0.0431	13.651	3.157
Globale (XGboost)	0.0732	0.1243	0.0384	0.0477	14.65	2.25

FIGURE 5.56 – Tableau récapitulatif des résultats obtenus en approche spécifique sur les taux EUR_UC

Le temps de calcul s'élève à 8h environ. Ici, l'approche spécifique et l'approche globale sont équivalentes. Cela implique que l'impact des variables explicatives doit être globalement le même quelque soit le produit ou bien que l'XGboost arrive à prendre en compte les différences entre produits dans ses arbres. L'approche globale se distingue donc par sa compléxité et son temp de calcul moindre sur les taux EUR_UC.

5.4.3 taux UC_UC

Voici les résultats obtenus en approche spécifique, sur la prédiction des taux UC_UC :

Approche	RMSE a	RMSE t	MAE a	MAE t	Critère a	Critère t
Spécifique	2.243	1.013	0.5733	0.5271	x	x
Globale (XGboost)	2.729	1.219	1.093	0.7678	x	x

FIGURE 5.57 – Tableau récapitulatif des résultats obtenus en approche spécifique sur les taux UC_UC

Le temps de calcul s'élève à 8h environ. Ici, l'approche spécifique est plus performante que l'approche globale. L'impact des variables explicatives doit donc être fonction du produit.

5.4.4 Analyse des modèles : approche spécifique

5.4.4.1 UC_EUR

17 produits ont des taux d'arbitrages UC_EUR modélisés par un SVM, 6 par des XGboost, 1 par une RF, et 1 par une régression linéaire.

Voici quelques éléments d'analyse des résultats fournit par l'approche spécifique sur les taux UC_EUR :

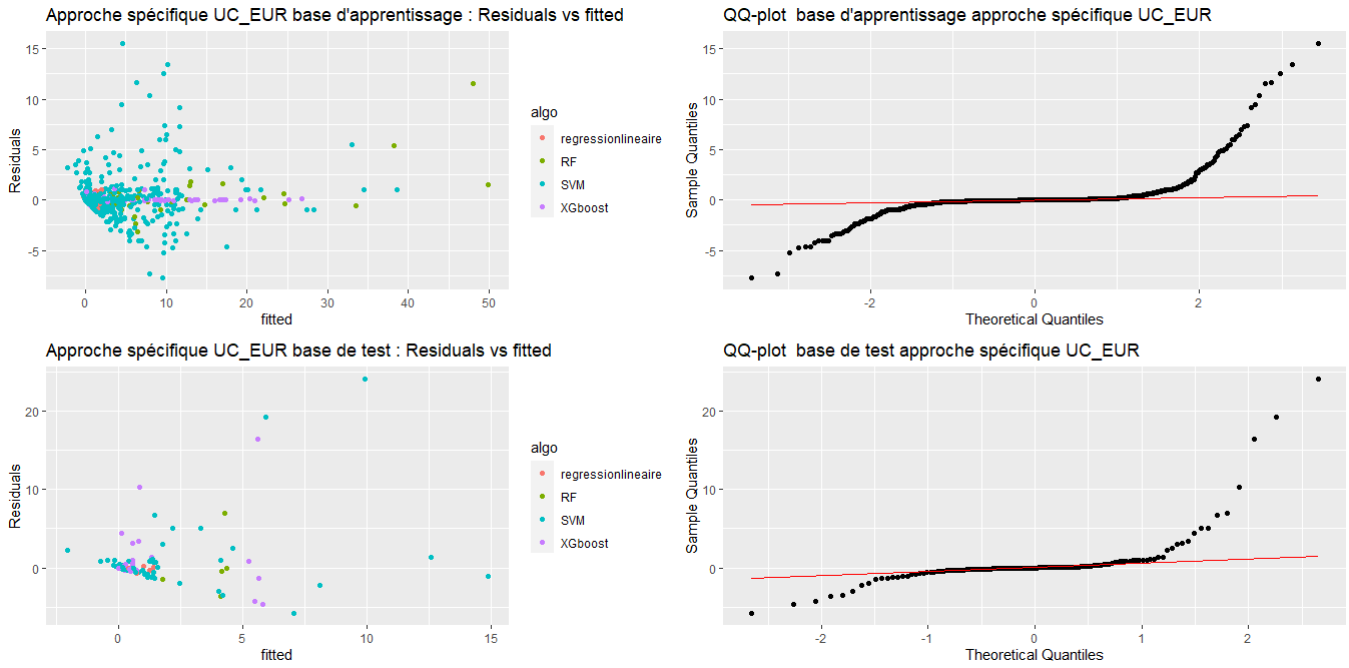


FIGURE 5.58 – Quelques éléments d'analyses des résultats de la modélisation les taux UC_EUR en approche spécifique

Sur la base d'apprentissage comme de test, les résidus sont assez centrés. Sur la base d'apprentissage il semble y avoir de fortes erreurs sur certains individus entre des valeurs prédites allant de 0 à 13% de taux d'arbitrages. Cela n'est pas choquant dans la mesure où de très nombreuses observations sont observées entre ces bornes et on observe tout de même une certaine masse en $y = 0$. Les QQ-plot synthétisent bien ce constat : les queues de distributions étant lourdes, les résidus ne semblent pas être un bruit blanc. Dès lors, l'erreur de prédiction ne peut pas être assimilé à l'apprentissage d'un bruit acceptable par notre algorithme, mais plutôt au fait que notre algorithme "passe à côté de quelque chose". Sur la base de test cela se confirme avec une certaine structure en cône à mesure que les valeurs prédites augmentent. Les SVM étant prédominants (17 produits 25), il héritent donc proportionnellement des produits les plus "dures" à prédire, d'où la sur-représentation des points bleu sur le graphique du haut. Ils présentent néanmoins de bons résultats sur la base de test, à la différence des XGboost qui semblent

avoir légèrement sur-appris (résidus très faibles sur la base d'apprentissage vs résidus plus élevés sur la base de test). Enfin, rappelons que nous avons entraîné ce modèle en choisissant les algorithmes champions pour chaque produit de manière à ce qu'il minimise le $Critere_{UC\_EUR}$, c'est à dire de manière à pénaliser la "sous-prédiction" des taux UC_EUR. Or ici, nous constatons une queue supérieure particulièrement lourde, c'est à dire que notre modèle produit de nombreux résidus élevés, donc que notre modèle ne pénalise pas vraiment cette "sous-prédiction" puisque nous nous retrouvons avec des "sous-prédictions".

Cet algorithme présente cependant quelques avantages. Sur la base d'apprentissage, pour des valeurs prédites supérieures à 20% de taux d'arbitrages, les erreurs semblent convenable vis à vis de l'ordre de grandeur. Par exemple, le point (50,3) signifie que la valeur prédite est un taux de 50% alors que la valeur observée est de 53%. De plus les residus pour des valeurs prédites élevées, sur la base d'apprentissage comme de test, sont soient positif, soient négatif proche de 0 : l'algorithme parvient donc bien à détecter les taux d'arbitrages anormalement élevés (de plus de 20%) au prix d'une moins bonne précision sur les taux situés dans une fourchette normale. Enfin ces résultats restent visuellement meilleurs que ceux de l'approche globale.

5.4.4.2 EUR_UC

18 produits ont des taux d'arbitrages EUR_UC modélisés par un SVM, 3 par des XGboost, 3 par des RF, et 1 par une régression linéaire.

Voici quelques éléments d'analyse des résultats fournit par l'approche spécifique sur les taux EUR_UC :

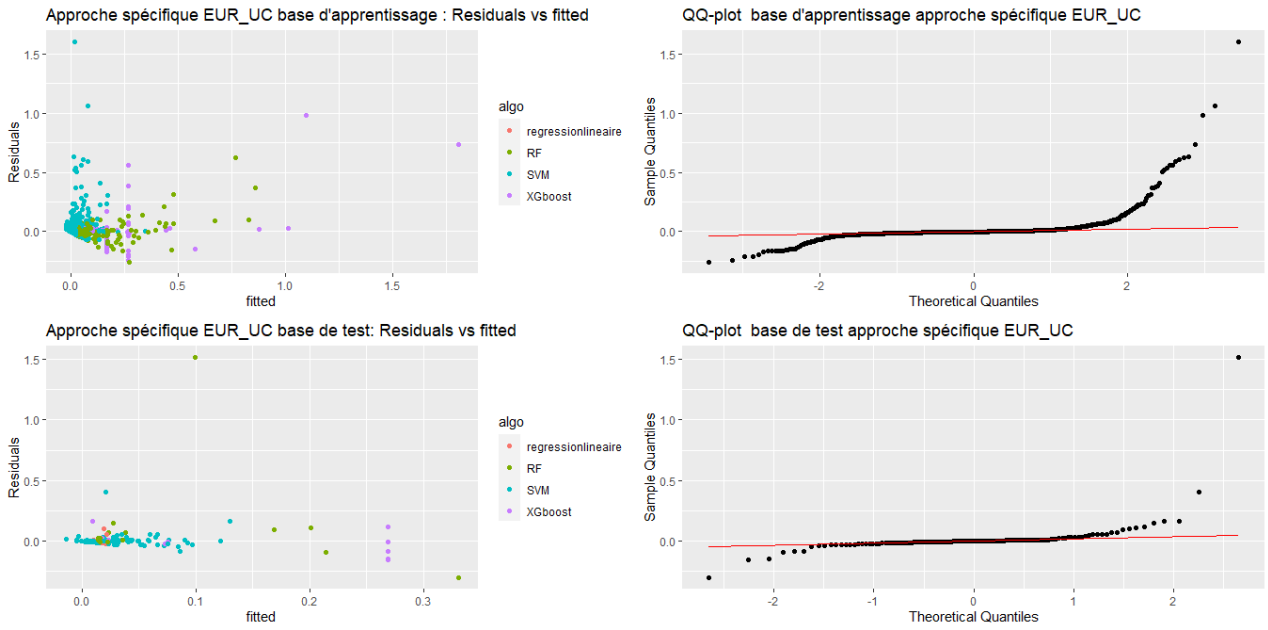


FIGURE 5.59 – Quelques éléments d'analyses des résultats de la modélisation les taux EUR_UC en approche spécifique

Le même constat que pour les taux d'arbitrages UC_EUR est fait et est même plus marqué ici : une bonne détection des taux élevés est faite sur la base d'apprentissage (à partir de taux d'arbitrages de 0.5% sur la base d'apprentissage) . Ici la différence réside dans le fait que les prédictions sur la base de test sont bien meilleures : le modèle arrive à bien généraliser sur la base de test tant sur les taux "normaux" que sur les taux élevés, et nos modèles n'ont donc par conséquent pas sur-appris. Les SVM semblent néanmoins avoir sous-appris (résidus des SVM sur la base d'apprentissage pas convaincants). Nous pouvons également remarquer que les queues de distribution inférieures des QQ-plot sont moins lourdes que les queus supérieures : il y a donc moins de "sur-prédiction" (moins de residus négatifs) des taux. Cela est la conséquence directe du choix du $Critere_{EUR\_UC}$ comme mesure discriminante des algorithmes champions pour chaque produit en approche spécifique qui vise à pénaliser les "sur-pédictions" des mouvements de l'euros vers l'UC.

À la vue de ces résultats nous comprenons la différence de la qualité des valeurs des mesures de performances entre l'approche spécifique des taux UC_EUR avec celles de l'approche spécifique des taux EUR_UC. Le modèle est donc satisfaisant et le $Critere_{EUR\_UC}$ rempli pleinement sa fonction.

5.4.4.3 UC_UC

7 produits ont des taux d'arbitrages UC_UC modélisés par un SVM, 14 par des XGboost, 4 par une RF, et aucun par une régression linéaire.

Voici quelques éléments d'analyse des résultats fournit par l'approche spécifique sur les taux UC_UC :

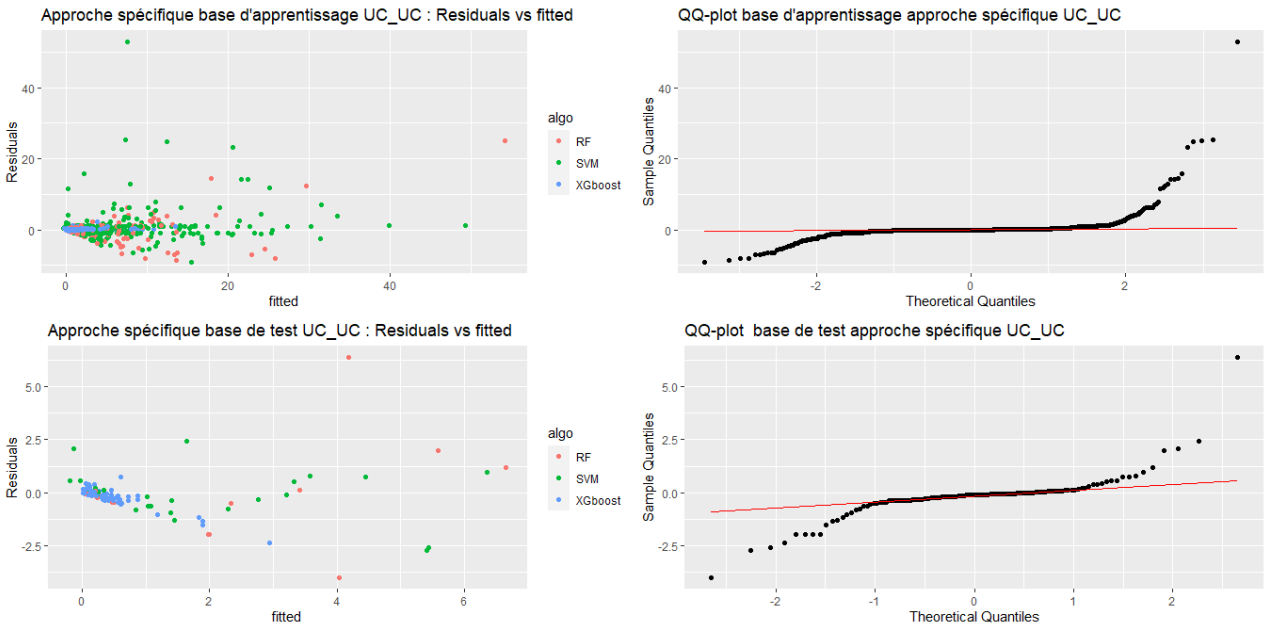


FIGURE 5.60 – Quelques éléments d'analyses des résultats de la modélisation les taux UC_UC en approche spécifique

Nous menons le même raisonnement que pour les origines/destinations d'arbitrages précédents. Ici les résidus sont centrés sur la base d'apprentissage comme sur celle de test. Le modèle a tendance à sous-prédire les taux élevés sur la base d'apprentissage (queue de distribution supérieure plus élevée que l'inférieure sur la base d'apprentissage, visible sur le graphique en haut à gauche également) mais est plus équilibré sur la base de test (QQ-plot sur la base de test symétrique). Le modèle arrive a généraliser assez bien sur la base de test mais présente des faiblesses de prédictions sur les taux prédits aux alentours des 2% : pour ces taux il y a une sur-prédiction dans tous les cas (résidus négatifs).Visuellement, ce modèle en prédiction pure est meilleur qu'en approche globale.

5.5 Bilan sur les statistiques descriptives et les modèles de prédictions

Les statistiques descriptives et l'analyse univariée confrontent directement qualitativement et quantitativement les taux d'arbitrages aux variables explicatives retenues décrites dans le chapitre 2 et 4. Celles-ci mettent clairement en évidence :

1. la structure du portefeuille étudié, de par les caractéristiques des contrats,
2. les structures linéaires reliant les variables explicatives aux taux d'arbitrages,
3. les structures non-linéaires reliant les variables explicatives aux taux d'arbitrages,
4. l'impact asymétrique sur les différentes origines/destinations d'arbitrage (UC_EUR, EUR_UC, et UC_UC) des variables explicatives.
5. des relations déjà établies (experts et mémoires) liant les taux d'arbitrages à des variables explicatives tant endogènes qu'exogènes, mais aussi de nouvelles. De cette manière, des comportements d'arbitrages sont identifiés et quantifiés.

À partir de ces constats, une découpe des variables explicatives en classes est faite de manière à tenir compte des constats dans une base de données à la maille $produit \times mois$ qui servira de base aux modèles de machine learning. En effet, les statistiques descriptives et l'analyse univariée ne suffisent pas à modéliser les taux d'arbitrages : il faut prendre en compte tous ces phénomènes simultanément et c'est pourquoi une approche machine learning est adoptée. Les algorithmes utilisés sont ceux présentés dans le chapitre 3, à savoir les régressions linéaires, les SVM, les forêts aléatoires et les XGboost.

Deux approches de modélisation par ML sont considérées : une approche globale et une approche spécifique. L'approche globale permet une interprétation directe du modèle via la méthode SHAP mais présente une moins bonne modélisation que l'approche spécifique. L'approche spécifique permet une meilleure modélisation, au détriment d'une certaine opacité d'interprétation du modèle. Ce compromis interprétabilité/performance est vérifié quantitativement et qualitativement.
L'analyse des deux approches de modélisation des taux, avec la méthode SHAP, conduit à quelques résultats généraux :

1. Les XGboost, et les SVM sont performants sur cette problématique. Il conviendrait de pousser plus loin la paramétrisation de ceux-ci et de les alimenter avec d'autres variables explicatives pertinentes non prisent en compte dans ce mémoire, comme par exemple la valeur du taux d'arbitrage le mois précédant. Les taux UC_EUR posent un problème particulier de modélisation.
2. La nature non linéaire des comportements/phénomènes d'arbitrages est vérifiée par le fait que seuls des modèles complexes ("boites noires") fournissent des résultats convenables.

La régression linéaire est à oublier si l'on souhaite dans le future créer un outil de prédictions des taux.

3. La méthode SHAP nous indique que les XGboost arrivent à apprendre certains comportements décelés dans les statistiques descriptives. Cette capacité d'apprentissage est donc très certainement la cause de la performance des XGboost. Les variables explicatives possédant les pouvoirs explicatifs les plus forts sont les variables des PM de contrats, de nombre de supports, de CSP, de mois de l'année, et d'âge actuariel. De plus, nous montrons que l'ajout des variables de clustering de fonds (indicateurs statistiques et distance DTW majoritairement) présentent un pouvoir explicatif fort. Ces variables de clustering viennent éclipser les variables dites exogènes comme le rendement du CAC40. Elles viennent de plus suggérer que seule une partie (certains fonds) du portefeuille pourrait être sujette à des phénomènes dynamiques d'arbitrages.

4. Les XGboost semblent détecter assez bien les taux élevés selon certaines variables : ils peuvent alors être entraînés et utilisés afin de créer un outil de monitoring des risques d'arbitrages massifs.

5. Le fait que l'approche spécifique présente une meilleure modélisation que l'approche globale, sachant que l'analyse de l'approche globale nous informe que celle-ci ne tient pas réellement compte des produits, nous indique que les produits présentent chacun une dynamique différente vis à vis des taux d'arbitrages (dû à la structure spécifique des contrats le composant) et que les modèles complexes sont capables d'en cerner les spécificités. Néanmoins, l'approche spécifique serait elle aussi à améliorer afin de constater une réelle distinction notable avec l'approche globale.

De nombreux autres résultats locaux sont détaillés dans le rapport.

Conclusion

La modélisation des arbitrages dans les contrats d'assurance vie pour un assureur vie est importante dans la mesure où les arbitrages forment un risque de solvabilité pour l'assureur.

Cependant, la compréhension et donc l'estimation de ces taux est compliquée : les arbitrages sont soumis à de nombreux facteurs conjoncturels et structurels interagissant entre eux de manière non linéaire, surtout sur les contrats en mode de gestion libre pour lesquels l'assureur n'a pas le contrôle.

Parallèlement, les assureurs disposent d'une formidable quantité d'informations concernant leurs contrats d'assurance vie, la donnée en accès libre sur internet est de plus en plus répandue, et les techniques avancées de machine learning permettant de détecter la non linéarité des phénomènes disposent aujourd'hui d'outils permettant de les analyser de plus en plus finement.

C'est pourquoi une approche machine learning de la modélisation des taux d'arbitrages, en vue de les expliquer, est possible, au prix d'un coût en temps élevé de retraitement de la donnée et d'analyse de celle-ci.

Ce mémoire propose donc une méthodologie pour appréhender, comprendre et modéliser les taux d'arbitrages : de la construction de la donnée à l'interprétation des modèles complexes de ML dits "boites noires" modélisant les taux. L'étude est menée à la maille produit mais tient compte d'informations disponibles à la maille contrat. Une approche est proposée en vue d'expliquer les mécanismes d'arbitrages, et une autre en vue de modéliser les taux le plus fidèlement possible. Il est montré que l'information dont nous disposons à la maille contrat contient un pouvoir explicatif non-négligeable des taux d'arbitrages.

Ces travaux peuvent servir de base pour la construction d'un outil de projection des taux et/ou d'un outil de détection d'arbitrages massifs. En effet, il est mis en évidence que les XGboost semblent bien prendre en compte certains phénomènes non linéaires affectant les arbitrages et dont on connaît une justification "métier", et que certaines variables explicatives présentent un fort pouvoir explicatif. Néanmoins, les résultats pourraient être grandement améliorés. Nous pourrions, dans un premier temps, améliorer exclusivement les XGboost, pour ensuite les analyser plus précisément et en déduire des variables plus pertinentes. De plus, il semble que les algorithmes éprouvent des difficultés à modéliser correctement à la fois les taux extrêmes et les taux normaux. Il serait possible alors de

créer un modèle distinguant ces deux classes de taux pour ensuite appliquer un modèle de prédiction de taux spécifique à la classe. Enfin, d'autres méthodes d'interprétabilité, comme les méthodes LIME ou ALE, pourraient être employées en vue de mieux comprendre nos modèles et d'en cerner les limites plus précisément.

Annexe A

Chapitre 4

A.1 Contribution à la construction des axes de l'ACP ayant servi à la représentation de du clustering par CAH sur une sélection d'indicateurs statistiques

FIGURE A.1 – Contribution d'une sélection d'indicateurs statistiques à la construction des axes de l'ACP

A.2 Résultats de l'ACP ayant servi à la représentation du clustering par K-means sur une sélection d'indicateurs statistiques

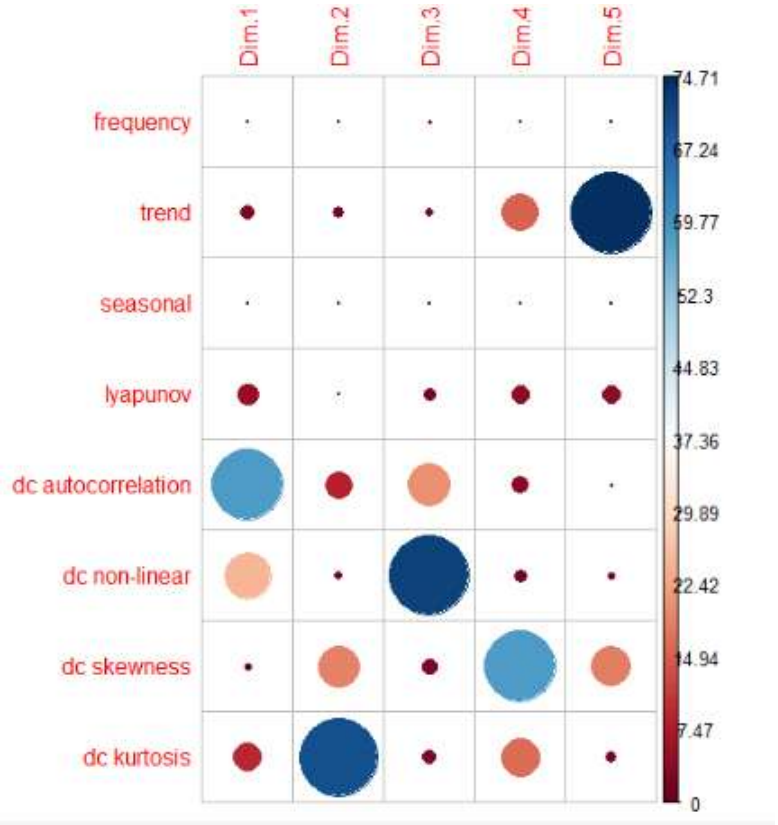
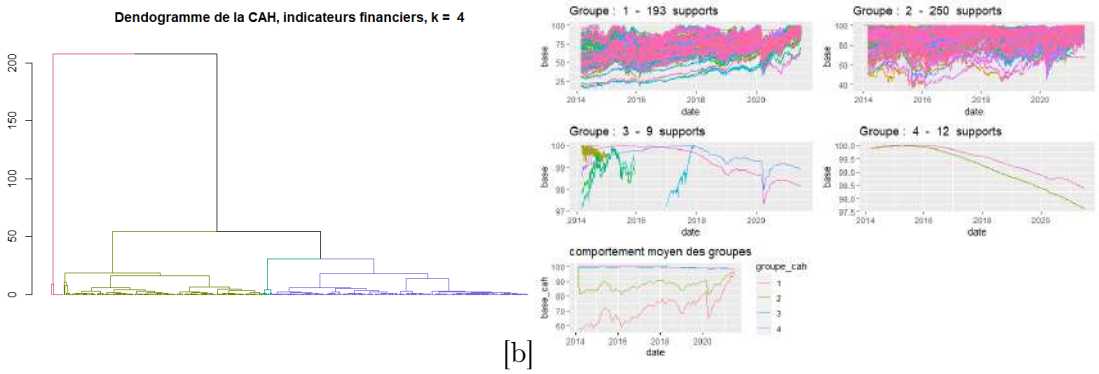


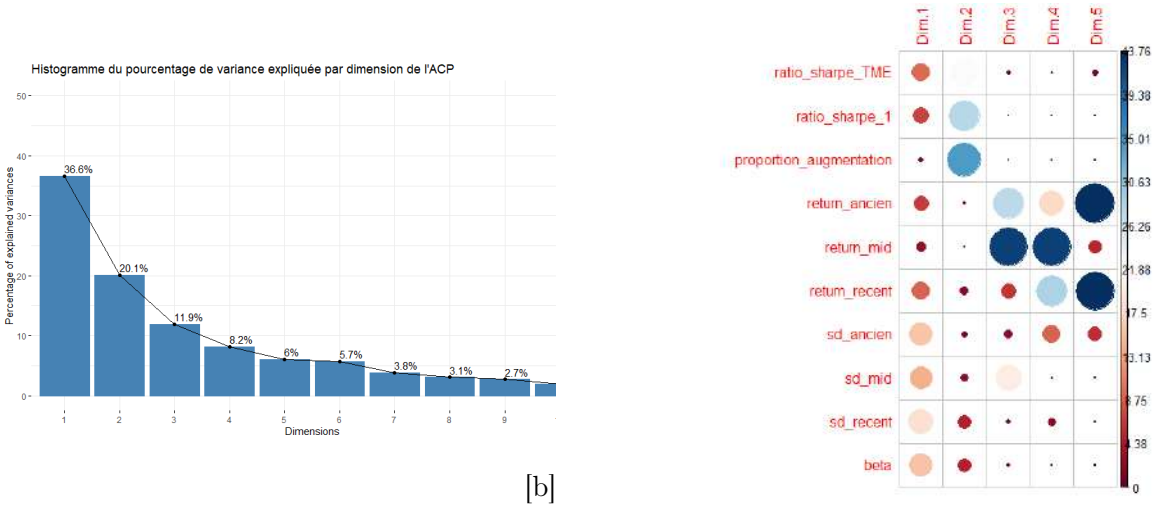
FIGURE A.2 – Contributions des indicateurs statistiques à la construction des axes de l'ACP

A.3 Résultats du clustering CAH + distance Euclidienne sur une sélection d'indicateurs financiers



[a] [b]

FIGURE A.3 – Dendogramme (a), et comportement des valeurs liquidatives des fonds par cluster (b), pour une distance euclidienne + CAH et 4 clusters, sur les séries temporelles résumées en indicateurs financiers.



[a] [b]

FIGURE A.4 – Variance des données expliquée par les axes de l'ACP (a), et contribution de la sélection d'indicateurs financiers à la construction des axes de l'ACP (b)

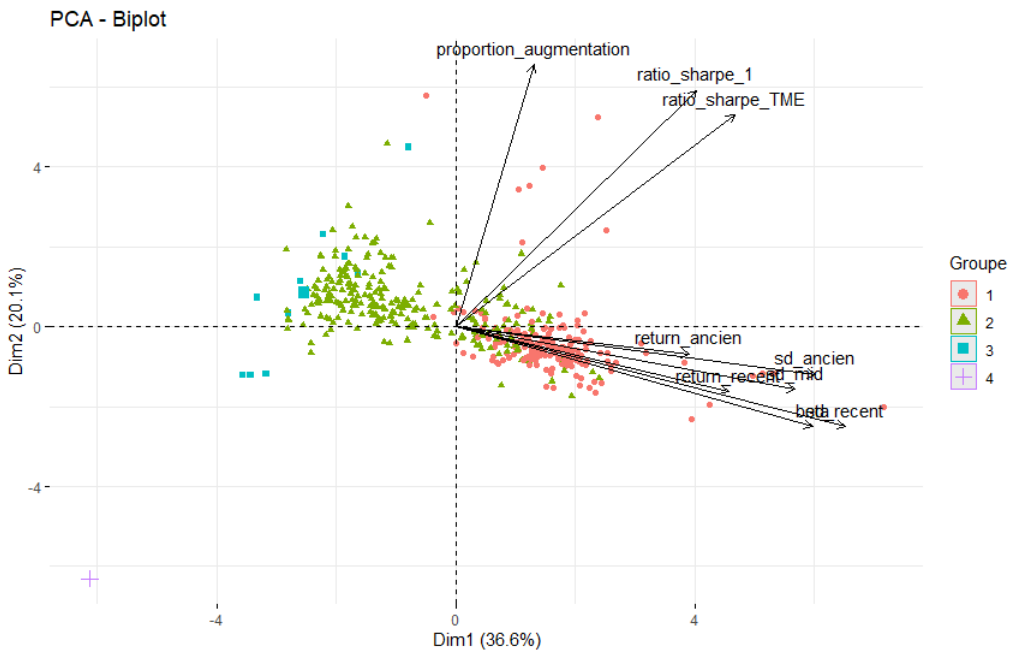
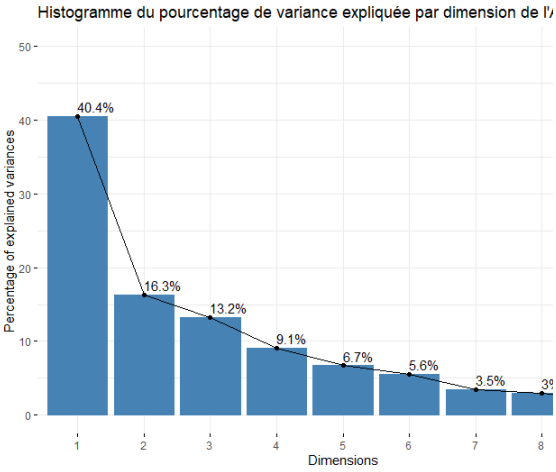
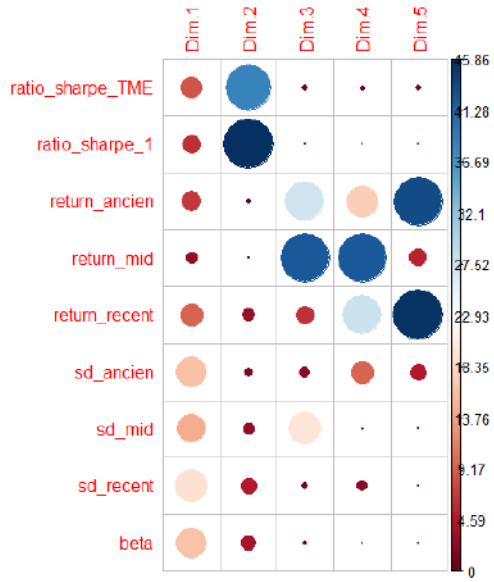


FIGURE A.5 – Biplot de l'ACP sur une sélection d'indicateurs financiers

A.4 Résultats de l'ACP sur une sélection d'indicateurs financiers ne comprenant pas la variable "proportion augmentation"



[a]



[b]

FIGURE A.6 – Variance des données expliquée par les axes de l'ACP (a), et contribution de la sélection d'indicateurs financiers à la construction des axes de l'ACP (b), sans la variables "proportion augmentation"

Annexe B

Chapitre 5

B.1 Eléments de compréhension des régressions linéaires en approche globale

- UC_EUR

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	nom_produit	addition	0.433	0.425	2265.4560	9227.3805	3.4669
2	dtw_1	addition	0.605	0.599	1057.3360	8604.8072	2.8946
3	nbsupp_10plus	addition	0.662	0.657	658.4060	8337.6169	2.6784
4	nbsupp_510	addition	0.692	0.687	446.7300	8176.9517	2.5561
5	fi_2	addition	0.702	0.697	378.6160	8122.0562	2.5151
6	PMcontrat_10	addition	0.709	0.704	333.7170	8084.9265	2.4875
7	PMcontrat_2040	addition	0.716	0.711	288.0490	8046.2254	2.4591
8	reseauAFC	addition	0.719	0.714	264.3120	8025.8450	2.4439
9	dtw_3	addition	0.724	0.718	236.9790	8001.9787	2.4264
10	stat_3	addition	0.729	0.724	200.1980	7969.1758	2.4028
11	ageactu030	addition	0.733	0.728	172.7600	7944.2814	2.3849
12	sexeF	addition	0.740	0.734	130.9040	7905.4228	2.3576
13	fi_2	removal	0.740	0.734	129.0610	7903.5712	2.3570
14	dtw_4	addition	0.743	0.738	106.0690	7881.8489	2.3415
15	csp_haut	addition	0.745	0.740	92.6210	7869.0214	2.3322
16	csp_bas	addition	0.749	0.744	66.9250	7844.1162	2.3148
17	ageactu7085	addition	0.753	0.748	40.3080	7817.8676	2.2966
18	stat_2	addition	0.755	0.749	29.2080	7806.7808	2.2886
19	PMcontrat_80160	addition	0.756	0.751	21.9470	7799.4691	2.2831
20	mois	addition	0.760	0.753	-1.7650	7795.3498	2.2734
21	anc08	addition	0.761	0.753	-3.9130	7793.0992	2.2713
22	anc913	addition	0.761	0.753	-4.3800	7792.5655	2.2703
23	ageactu3169	addition	0.762	0.754	-6.9930	7789.8189	2.2678
24	PMcontrat_1020	addition	0.762	0.754	-6.8460	7789.9089	2.2673
25	nbsupp_25	addition	0.762	0.754	-6.1120	7790.6028	2.2671
26	nbsupp_25	removal	0.762	0.754	-6.8460	7789.9089	2.2673

FIGURE B.1 – Itérations de la procédure stepwise sur les régressions linéaires des taux d'arbitrages UC_EUR

(Intercept)	nom_produitAIEF	nom_produitAIER7	nom_produitAIERN
103.34465050	14.70081198	-48.92772039	-31.58626263
nom_produitAMEV	nom_produitAMEVT	nom_produitAZYearlingVie	nom_produitGaipare2
7.45212247	20.65728179	37.21433754	-53.44989430
nom_produitGaipSel	nom_produitGaipSelF	nom_produitGaipSelmo	nom_produitGaipSelmoF
-42.22989124	-50.71641412	-38.87984080	-42.34149162
nom_produitIdealis	nom_produitIdealisF	nom_produitIdeavie	nom_produitIdeavieF
-9.46245148	8.95091987	7.09573867	15.21262100
nom_produitTellusA	nom_produitTellusAv09	nom_produitTellusAv9A	nom_produitTellusAvF1Y
-59.18246530	-60.01438284	-54.08706117	-49.37213305
nom_produitTellusAvF9B	nom_produitTellusD	nom_produitTellusG	nom_produitYearlingCapi
-48.40168904	-18.76677744	-60.55557600	16.54356940
nom_produitYearlingVie	dtw_1	nbsupp_10plus	nbsupp_510
15.83862013	-108.21713639	-66.31575281	43.69366588
PMcontrat_10	PMcontrat_2040	reseauAFC	dtw_3
96.05364020	39.22581275	-65.76050554	-86.73957257
stat_3	ageactu030	sexeF	dtw_4
-39.89816579	-152.64273523	12.55895380	33.38095735
csp_haut	csp_bas	ageactu7085	stat_2
-39.33653906	-47.28129212	-16.41390103	46.98018377
PMcontrat_80160	mois02	mois03	mois04
43.01066975	-0.33709549	0.18562376	-0.10666965
mois05	mois06	mois07	mois08
-0.02640106	-0.23726341	0.18415140	-0.09292147
mois09	mois10	mois11	mois12
-0.54606661	-0.63148585	-0.55123847	-0.56277906
anc08	anc913	ageactu3169	PMcontrat_1020
-3.18596671	-1.73792797	6.18930868	13.68508405

FIGURE B.2 – Valeurs des régresseurs de la régression linéaire retenue par procédure stepwise sur les taux d'arbitrages UC_EUR

```
Residual standard error: 2.267 on 1672 degrees of freedom
Multiple R-squared:  0.7619,    Adjusted R-squared:  0.7541
F-statistic:  97.3 on 55 and 1672 DF,  p-value: < 2.2e-16
```

FIGURE B.3 – R2 ajusté, statistique et test de Fisher, de la régression linéaire obtenue par procédure stepwise sur les taux d'arbitrages UC_EUR

- EUR_UC

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	nbsupp_10plus	addition	0.136	0.135	199.4410	-2210.3026	0.1275
2	nom_produit	addition	0.175	0.163	113.6650	-2243.0161	0.1254
3	mois	addition	0.198	0.181	63.9150	-2270.4310	0.1241
4	ageactu030	addition	0.207	0.190	46.2400	-2287.5950	0.1234
5	sexeF	addition	0.217	0.199	26.8120	-2306.7109	0.1227
6	PMcontrat_10	addition	0.225	0.207	11.8060	-2321.6552	0.1221
7	nbsupp_510	addition	0.229	0.211	3.8200	-2329.6832	0.1218
8	reseauCRT	addition	0.233	0.214	-3.2490	-2336.8406	0.1215
9	csp_haut	addition	0.236	0.217	-7.5930	-2341.2756	0.1213
10	PMcontrat_4080	addition	0.238	0.218	-9.9320	-2343.6902	0.1212
11	perio_U	addition	0.239	0.219	-10.9780	-2344.7965	0.1211
12	anc913	addition	0.241	0.221	-12.4300	-2346.3234	0.1210
13	dtw_4	addition	0.242	0.221	-12.8740	-2346.8253	0.1210
14	reseauAFC	addition	0.243	0.222	-12.8870	-2346.8881	0.1210
15	dtw_3	addition	0.244	0.222	-13.6880	-2347.7626	0.1209
16	anc08	addition	0.245	0.223	-14.4940	-2348.6471	0.1208
17	dtw_2	addition	0.246	0.224	-14.4970	-2348.7093	0.1208
18	dtw_4	removal	0.246	0.224	-16.2490	-2350.4539	0.1208
19	fi_1	addition	0.247	0.224	-15.9630	-2350.2198	0.1207
20	stat_1	addition	0.247	0.224	-14.6880	-2348.9677	0.1208
21	stat_3	addition	0.250	0.227	-19.2370	-2353.7363	0.1206

FIGURE B.4 – Itérations de la procédure stepwise sur les régressions linéaires des taux d'arbitrages EUR_UC

(Intercept)	nbsupp_10plus	nom_produitAIEF	nom_produitAIER7
2.276185037	0.807035948	0.456115948	-0.607496505
nom_produitAIERN	nom_produitAMEV	nom_produitAMEVT	nom_produitAZYearlingVie
-0.871891436	-0.049131627	0.204175742	0.058273813
nom_produitGaipare2	nom_produitGaipSel	nom_produitGaipSelF	nom_produitGaipSelmo
1.163446319	1.499027068	2.542579736	2.511437256
nom_produitGaipSelmoF	nom_produitIdealis	nom_produitIdealisF	nom_produitIdeavie
2.733909657	-0.348392989	0.402312293	-0.226241413
nom_produitIdeavieF	nom_produitTellusA	nom_produitTellusAv09	nom_produitTellusAv9A
0.421936867	0.683209241	0.732389976	0.632136833
nom_produitTellusAvF1Y	nom_produitTellusAvF9B	nom_produitTellusD	nom_produitTellusG
0.984902753	0.718049506	-0.187086293	2.837687110
nom_produitYearlingCapi	nom_produitYearlingVie	mois02	mois03
0.036796614	-0.150954160	0.025140051	0.046996385
mois04	mois05	mois06	mois07
0.027496929	-0.006649824	0.031551329	0.018434349
mois08	mois09	mois10	mois11
-0.027712570	-0.011494027	0.001773813	0.009839752
mois12	ageactu030	sexeF	PMcontrat_10
0.012918854	5.001461578	-1.770226087	-1.611541612
nbsupp_510	reseauCRT	csp_haut	PMcontrat_4080
0.638323525	-1.919253134	0.713991390	-1.300510702
perio_U	anc913	reseauAFC	dtw_3
-0.920389063	-0.074423584	0.898383967	1.208475015
anc08	dtw_2	fi_1	stat_1
-0.089270317	1.154088302	0.773128350	-2.409995519
stat_3			
-2.087996286			

FIGURE B.5 – Valeurs des régresseurs de la régression linéaire retenue par procédure stepwise sur les taux d'arbitrages EUR_UC

```
Residual standard error: 0.1206 on 1675 degrees of freedom
Multiple R-squared:  0.2502,    Adjusted R-squared:  0.2269
F-statistic: 10.75 on 52 and 1675 DF,  p-value: < 2.2e-16
```

FIGURE B.6 – R2 ajusté, statistique et test de Fisher, de la régression linéaire obtenue par procédure stepwise sur les taux d'arbitrages EUR_UC

- UC_UC

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	nom_produit	addition	0.373	0.364	877.3860	10004.8062	4.3415
2	dtw_1	addition	0.469	0.461	481.7760	9720.1793	3.9971
3	nbsupp_510	addition	0.499	0.491	360.1720	9622.4362	3.8845
4	reseauAGT	addition	0.530	0.523	229.9390	9510.9434	3.7601
5	nbsupp_10plus	addition	0.547	0.539	164.3240	9451.8969	3.6954
6	mois	addition	0.566	0.556	86.1690	9398.6190	3.6275
7	csp_haut	addition	0.570	0.560	71.8230	9384.8569	3.6121
8	dtw_3	addition	0.574	0.563	58.7480	9372.1913	3.5979
9	stat_3	addition	0.581	0.570	31.9620	9345.8372	3.5695
10	ageactu3169	addition	0.585	0.574	16.3540	9330.2616	3.5525
11	PMcontrat_10	addition	0.588	0.578	3.4570	9317.2508	3.5381
12	ageactu030	addition	0.590	0.579	-1.5110	9312.1847	3.5320
13	ageactu7085	addition	0.592	0.580	-5.5860	9307.9998	3.5267
14	perio_U	addition	0.593	0.582	-9.6470	9303.8072	3.5214
15	fi_1	addition	0.594	0.583	-13.5640	9299.7413	3.5163
16	fi_3	addition	0.595	0.583	-14.0340	9299.2020	3.5148
17	nbsupp_25	addition	0.596	0.584	-14.9570	9298.1935	3.5127
18	anc913	addition	0.596	0.584	-14.8190	9298.2734	3.5118
19	anc08	addition	0.597	0.585	-16.9180	9296.0398	3.5086
20	PMcontrat_4080	addition	0.597	0.585	-16.3280	9296.5810	3.5082

FIGURE B.7 – Itérations de la procédure stepwise sur les régressions linéaires des taux d'arbitrages UC_UC

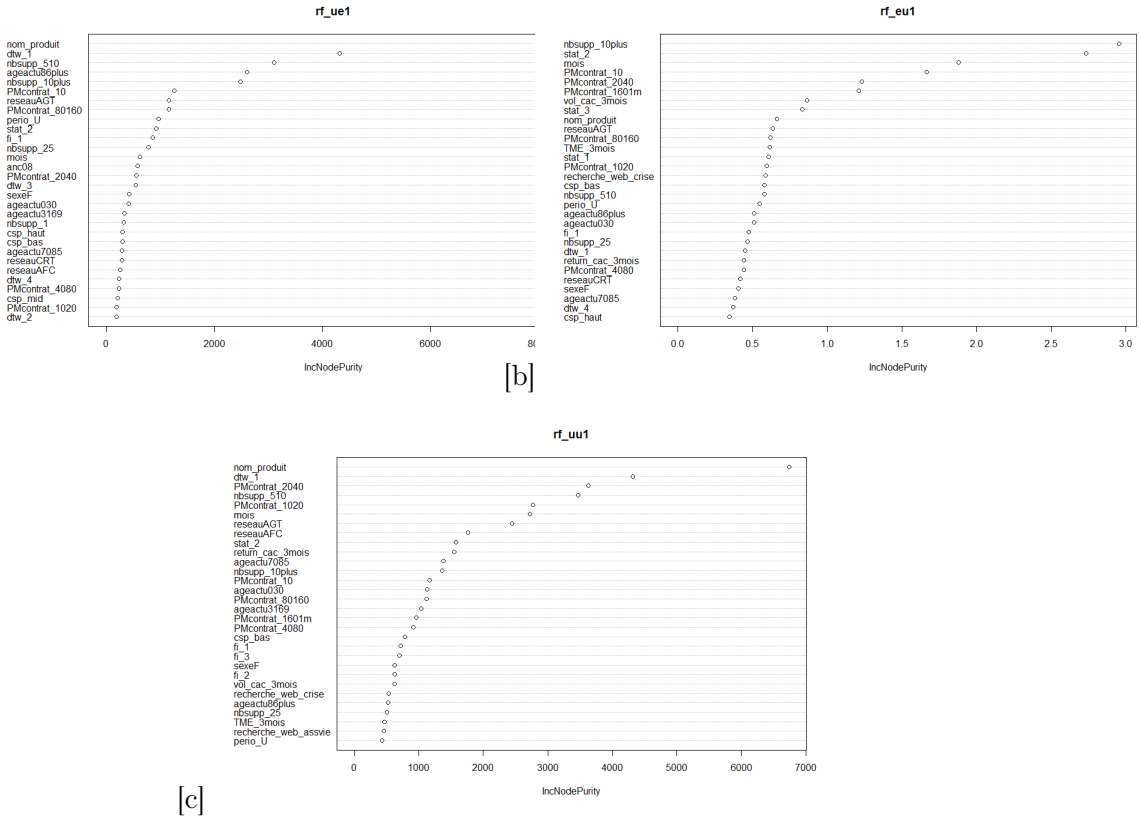
(Intercept)	nom_produitAIEF	nom_produitAIER7	nom_produitAIERN
138.5803853	15.9737309	-67.1694008	-52.1243472
nom_produitAMEV	nom_produitAMEVT	nom_produitAZYearlingVie	nom_produitGaipare2
-5.5184280	10.8443205	18.1521934	-51.3713372
nom_produitGaipSel	nom_produitGaipSelF	nom_produitGaipSelmo	nom_produitGaipSelmoF
-14.0747367	12.0662104	17.0708151	19.7193018
nom_produitIdealis	nom_produitIdealisF	nom_produitIdeavie	nom_produitIdeavieF
-19.4052357	6.0505816	0.1094376	13.3072887
nom_produitTellusA	nom_produitTellusAv09	nom_produitTellusAv9A	nom_produitTellusAvF1Y
-80.8453286	-82.0398447	-79.4386776	-61.5811871
nom_produitTellusAvF9B	nom_produitTellusD	nom_produitTellusG	nom_produitYearlingCapi
-65.1727648	-12.9875723	13.1946168	14.7404543
nom_produitYearlingVie	dtw_1	nbsupp_510	reseauAGT
5.6766867	-106.3808666	54.7453934	83.8675863
nbsupp_10plus	mois02	mois03	mois04
-39.5940646	-0.5832969	1.6420153	-0.4763699
mois05	mois06	mois07	mois08
-0.7914043	0.4365207	0.7125534	-1.0104045
mois09	mois10	mois11	mois12
-0.4294077	-1.1375788	-0.5980574	-0.7324882
csp_haut	dtw_3	stat_3	ageactu3169
-31.5042066	-85.7269616	-71.8847793	16.9460243
PMcontrat_10	ageactu030	ageactu7085	perio_U
28.4386273	-60.2911582	-15.2275862	-41.6306605
fi_1	fi_3	nbsupp_25	anc913
-27.9365892	-23.1883286	-8.0887232	-2.4801199
anc08	PMcontrat_4080		
-3.0030082	-17.7034586		

FIGURE B.8 – Valeurs des régresseurs de la régression linéaire retenue par procédure stepwise sur les taux d'arbitrages UC_UC

```
Residual standard error: 3.508 on 1674 degrees of freedom
Multiple R-squared:  0.5975,    Adjusted R-squared:  0.5848
F-statistic: 46.89 on 53 and 1674 DF,  p-value: < 2.2e-16
```

FIGURE B.9 – R2 ajusté, statistique et test de Fisher, de la régression linéaire obtenue par procédure stepwise sur les taux d'arbitrages UC_UC

B.2 Importance plot des RF et XGboost en approche globale



[a] [b]

[c]

FIGURE B.10 – Importance des variables dans les forêts aléatoires modélisant les taux UC_EUR (a), EUR_UC (b) , et UC_UC (c), en approche globale, servant à la sélection des variables

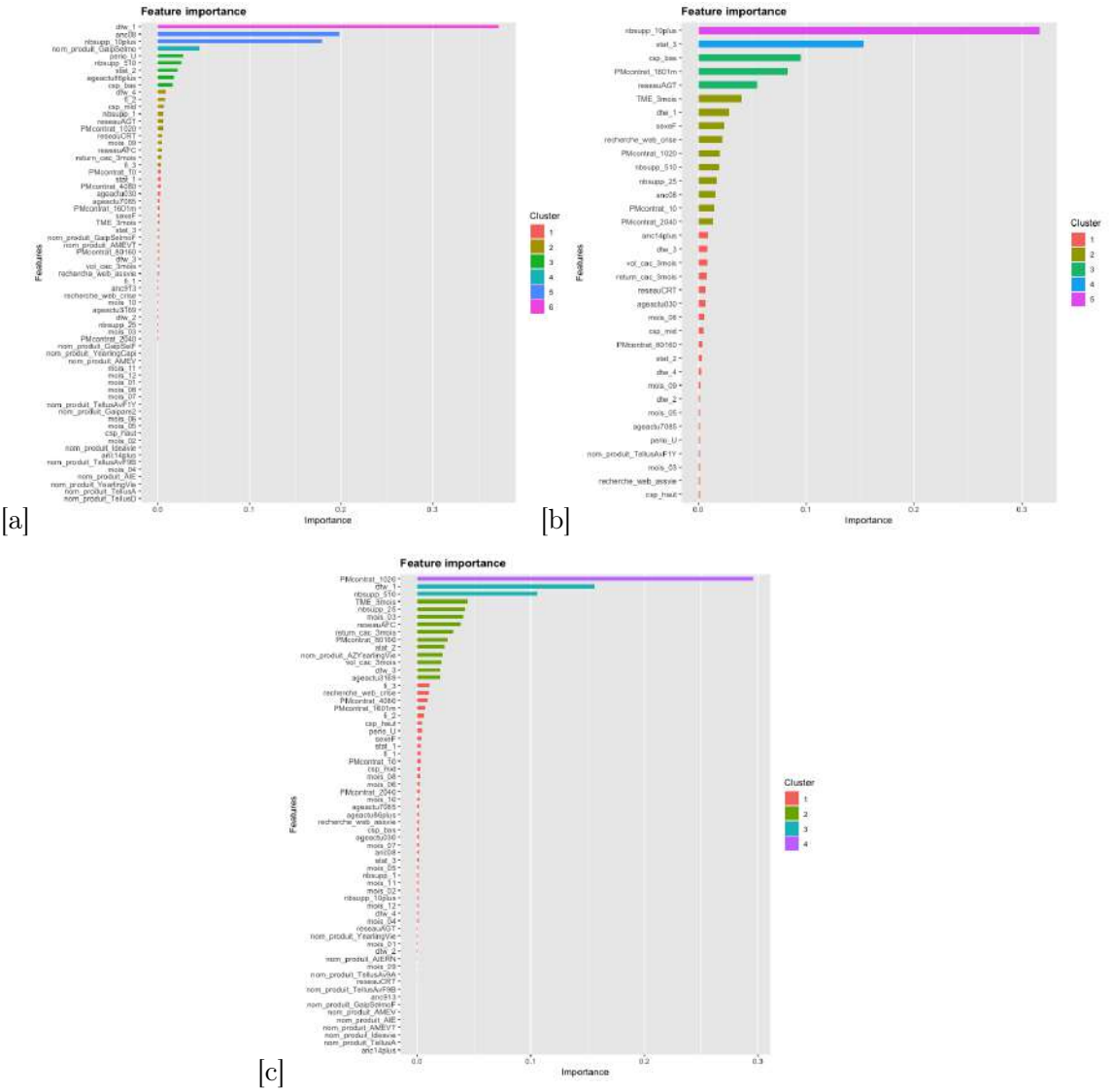


FIGURE B.11 – Importance des variables dans les XGboost modélisant les taux UC_EUR (a), EUR_UC (b) , et UC_UC (c), en approche globale, servant à la sélection des variables

Bibliographie

[1] A.Chevalier et P.Valade. *Cours de Master 2 EURIA, Réassurance*. 2021.

[2] A.Moeglin. *Cours de Master 1 EURIA, Mathématiques des assurances de personnes*. 2020.

[3] ACPR. *Analyses et synthèse - Les différentes composantes de l'assurance vie et leur évolution*. Rapp. tech. ACPR, 2016. URL : https://acpr.banque-france.fr/sites/default/files/medias/documents/201605-as65-differentes-composantes-assurance-vie-evolution.pdf.

[4] AG2R LA MONDIALE. *les différents types d'assurance vie*. URL : https://www.ag2rlamondiale.fr/epargne/assurance-vie/les-differents-types-de-contrat-en-assurance-vie.

[5] ASSURANCE ET MUTUELLE. *les branches d'assurances*. URL : https://www.assurance-et-mutuelle.com/assurance/branches-assurance.html.

[6] B.Amrani. "L'arbitrage dans les contrats d'épargne multi-support". Mémoire d'actuariat. CNAM, 2013.

[7] B.Auffret et S.Bouhzila. "Analyse et pilotage d'un portefeuille de contrats multi-suppports au travers de la logique floue". Mémoire d'actuariat. CEA, 2016.

[8] Banque de France. *Agréments administratif - Principes*. URL : https://acpr.banque-france.fr/autoriser/procedures-secteur-assurance/regime-administratif/agrement-administratif/principes.

[9] Boursorama. *ISIN Allianz Foncier C/D FR0000945503*. URL : https://www.boursorama.com/bourse/opcvm/cours/MP-829263/.

[10] Business Insider. *La Bourse de Paris entame 2019 en forte baisse — et c'est notamment à cause de la Chine*. URL : https://www.businessinsider.fr/bourse-de-paris-cac40-en-forte-baisse-craintes-chine-2-janvier-2019/.

[11] C.Folgoas. "Modélisation des arbitrages sur les contrats d'assurance vie". Mémoire d'actuariat. CEA, 2017.

[12] C.Nicolas. "Les arbitrages en assurance vie sont-ils soumis à un phénomène de contagion ?" Mémoire d'actuariat. EURIA, 2017.

[13] D.Delcaillau. "Contrôle et Transparence des modèles complexes en actuariat". Mémoire d'actuariat. EURIA, 2019.

[14] D.Rubin, NM.Laird et AP.Dempster. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In : *Journal of the Royal Statistical Society* (1977).

[15] DL.Davies et DW.Bouldin. "A cluster separation measure". In : *IEEE Journals and Magazine* (1979).

[16] E.Biernat et M.Lutz. *Data science : fondamentaux et études de cas*. 2017.

[17] F.Ducos. *Cours de master 1 EURIA, Droits financier et des assurances*. 2020.

[18] F.Vermet. *Cours de master 1 EURIA, Apprentissage statistique*. 2019.

[19] FFA. *Le régime fiscal de l'assurance vie*. URL : https://www.ffa-assurance.fr/infos-assures/le-regime-fiscal-de-assurance-vie.

[20] Banque de France. *Analyses et synthèses - Le marché de l'assurance vie pendant la crise sanitaire*. Rapp. tech. Banque de France, 2020. URL : https://acpr.banque-france.fr/sites/default/files/medias/documents/20210401_as_marche_assurance_vie_crise_sanitaire.pdf.

[21] H.Drucker. "Improving Regressors Using Boosting Techniques". In : *Proceedings of the 14th International Conference on Machine Learning* (1997).

[22] H.Sakoe et S.Chiba. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition". In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* (1978).

[23] I.Gurrutxaga et al. "SEP/COP : An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index". In : *Pattern Recognition* (2010).

[24] JC.Dunn. "Well-separated clusters and optimal fuzzy partitions". In : *J. Cybern* (1974).

[25] JH.Friedman. "Stochastic Gradient Boosting". In : *Computational Statistics and Data Analysis* (1999).

[26] K.Lyoubi. "Modélisation de la réponse des assurés à une incitation financière : Arbitrage entre fonds en euros et en unités de compte et politique de participation aux bénéfices". Mémoire d'actuariat. ENSAE, 2020.

[27] Kaggle. *Machine Learning - Non supervisé*. URL : https://www.kaggle.com/zoupet/machine-learning-non-supervis-correction.

[28] L.Breiman. *Classification and regression trees*. 1984.

[29] L.Breiman. "Random Forests". In : *Machine learning 45* (2001).

[30] le comparateur des assurances. *Définitions*. URL : https://www.lecomparateurassurance.com/95089-lexique-toutes-assurances/13-conditions-particulieres.

[31] Les Echos. *Ça s'est passé en 2015 : le krach chinois*. URL : https://www.lesechos.fr/2015/12/ca-sest-passe-en-2015-le-krach-chinois-286059.

[32] Les Echos. *L'assurance-vie : un produit d'épargne idéal pour s'assurer des revenus réguliers*. URL : https://investir.lesechos.fr/placements/vie-pratique/dossiers/comment-devenir-rentier-et-vivre-sans-travailler/un-produit-d-epargne-ideal-pour-s-assurer-des-revenus-reguliers-1867786.php.

[33] M.Kim et RS.Ramakrishna. "New indices for cluster validity assessment". In : *Pattern Recognition Letters* (2005).

[34] NCBI. *Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models*. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/.

[35] NCBI. *The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances*. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6404674/.

[36] OCDE. *Évolution de l'OAT 10ans France*. URL : https://data.oecd.org/fr/interest/taux-d-interet-a-long-terme.html.

[37] P.Aillot et JM.Derien. *Cours de Master 2 EURIA, Séries temporelles*. 2020.

[38] P.Rousseeuw. "Silhouettes : A graphical aid to the interpretation and validation of cluster analysis." In : *J. Comput.Appl. Math.* (1987).

[39] R.Gauville. "Projection du ratio de solvabilité : des méthodes de machine learning pour contourner les contraintes opérationnelles de la méthode des SdS". Mémoire d'actuariat. EURIA, 2017.

[40] RJC.Ndong-Bibang. "Modélisation et calibration du risque d'arbitrage". Mémoire d'actuariat. ISUP, 2020.

[41] robjhyndman. *Measuring time series characteristics*. URL : https://robjhyndman.com/hyndsight/tscharacteristics/.

[42] S.Rani et G.Sikka. "Recent techniques of clustering of time series data : a survey". In : *International Journal of Computer Applications* (2012).

[43] H. Steinhaus. "Sur la division des corps matériels en parties". In : *Bulletin de l'académie polonaise des sciences* (1957).

[44] T.Calinski et J.Harabasz. "A dendrite method for cluster analysis." In : *Commun. Stat. Theory Methods* (1974).

[45] T.Chen et G.Carlos. "XGBoost: A Scalable Tree Boosting System". In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).

[46] T.Lampert et al. "Constrained distance based clustering for time-series: a comparative and experimental study". In : *Data Mining and Knowledge Discovery* (2018).

[47] tout sur mes finances. *transfert Fourgous - passer du monosupport au multisupports sans pénalités*. URL : https://www.toutsurmesfinances.com/placements/assurance-vie-multisupports-comment-ca-marche.html.

[48] TOWARDS DATA SCIENCE. *Dynamic Time Warping : Explanation and Code Implementation.* URL : `https://towardsdatascience.com/dynamic-time-warping-3933f25fcdd`.

[49] V.VAPNIK et C.CORTES. "Support Vectors Networks". In : *Machine Learning 20* (1957).

[50] X.WANG, K.SMITH et R.HYNDMAN. "Characteristic-based clustering for time series data". In : *Data Mining and Knowledge Discovery* (2006).

[51] Y.FREUND et R.SCHAPIRE. "A decision-theoretic generalization of on-line learning and an application to boosting". In : *Journal of Computer and System Sciences* (1997).