Introduction

Les sociétés d'assurance vie et mixte gèrent environ 2 000 milliards d'encours d'assurance vie. Sur cette somme, près de 1 600 milliards euros sont investis dans des contrats dont le capital est garanti par l'assureur sur toute la durée des engagements quelles que soient les conditions économiques. En cas de déséquilibre de marché prolongé, cette garantie fait peser un risque systémique sur l'industrie. Ainsi, depuis plusieurs années, l'environnement de taux d'intérêt bas voire négatifs persistant met en danger ce modèle économique en limitant le rendement des actifs investis au regard des engagements pris. La baisse régulière des taux affecte de façon significative et durable la rentabilité, mais aussi les fonds propres économiques disponibles, pouvant conduire à une diminution marquée du ratio de solvabilité moyen du secteur.

La directive dite « Solvabilité II » impose aux sociétés d'assurance le maintien d'un niveau de fonds propres économiques suffisants pour résister à un choc bicentenaire à horizon d'un an. Le calcul repose sur une évaluation du bilan en valeur économique, selon des modèles stochastiques en général pour tenir compte des nombreuses interactions entre les actifs et les passifs d'assurance dans divers scénarios économiques, ainsi que de la valeur économique des options ou garanties financières en faveur des assurés. L'objectif du présent mémoire est de proposer une méthode alternative de détermination des fonds propres économiques basée sur l'observation dynamique des mouvements historiques des bilans économiques des assureurs tels qu'ils sont déclarés dans les remises réglementaires. Cette approche doit permettre de mieux appréhender les déterminants des fonds propres économiques des sociétés d'assurance vie et mixte et d'anticiper les impacts de changements économiques sur les fonds propres disponibles et donc sur leur solvabilité.

Afin de modéliser l'évolution des fonds propres économiques, nous allons étudier dans un premier temps la nature de leurs principales composantes, puis leur influence respective sur le calcul des fonds propres économiques. Nous étudierons ensuite différents choix de modèles de prédiction automatisés et nous retiendrons celui qui restitue au mieux la variation des fonds propres économiques. Enfin, dans un dernier temps, une fois les composantes déterminées et le modèle validé, l'objectif sera de tester la sensibilité des fonds propres économiques à différents scénarios pour comprendre comment les fonds propres sont affectés par différents chocs économiques.

1. Chapitre 1 : présentation générale

1.1. Objectifs de l'étude

Dans le cadre du présent mémoire, nous nous intéressons au calcul des fonds propres économiques des sociétés d'assurance vie et mixtes. Ces sociétés sont les seules habilitées à proposer des engagements d'assurance vie conformément au principe de spécialisation de l'article L.321-1 du code des assurances. Ces engagements présentent des caractéristiques relativement proches en terme de fonctionnement actuariel, en tout cas pour la partie liée à l'assurance vie à capital garanti. Cette homogénéité présente l'avantage d'un fonctionnement comparable des interactions entre variables (notamment d'actif et de passif) qui devrait permettre à notre algorithme de prédiction des fonds propres économiques de prévoir fidèlement l'impact des évolutions des variables d'entrée sur les fonds propres économiques.

1.1.1. Enjeu du mémoire

Sur le plan réglementaire, les fonds propres économiques sont issus de la différence entre la valeur de marché des actifs (notamment des placements) et la meilleure estimation des passifs (notamment des engagements d'assurance dits « BEL » pour Best Estimate Liabilities). Pour déterminer les BEL sur des contrats d'épargne, l'assureur doit modéliser et actualiser l'ensemble des flux relatifs aux contrats d'assurance dans différents scénarios possibles (approche stochastique s'appuyant le plus souvent sur des générateurs de scénarios économiques) en tenant compte des nombreuses interactions actif-passif (via des modèles de gestion actif-passif dits ALM), de la politique de participation aux bénéfices ou du niveau de TMG des contrats par exemple.

La réglementation n'impose pas les hypothèses de calcul des modèles ALM, qui sont déterminées par les sociétés elles-mêmes en fonction de leur appétence au risque et des caractéristiques de leurs contrats. C'est pourquoi il est intéressant de construire un outil qui, sur la base des observations historiques d'évolution des fonds propres économiques de ces assureurs, est capable d'en prédire les évolutions à horizon d'une année, de manière simplifiée et rapide à mettre en œuvre, afin d'anticiper les mouvements et de comprendre les principaux déterminants de la richesse de ce type de sociétés, hors facteurs exogènes telle qu'une recapitalisation.

1.1.2. Utilisation des données réglementaires et déroulement de l'étude

Afin de calibrer notre algorithme de prédiction des fonds propres économiques des sociétés d'assurance vie et mixtes, nous avons recours aux données communiquées par les sociétés dans leurs reporting quantitatifs imposés par le pilier III de Solvabilité II en place depuis le 1er janvier 2016. Ces données sont disponibles pour 197 sociétés vie et mixte, sur la période du 31 décembre 2016 au 30 juin 2021 pour des arrêtés trimestriels et annuels. Nous disposons ainsi de 2 771 observations sur lesquelles nous allons calibrer l'algorithme de prédiction des fonds propres économiques.

La détermination des variables d'entrée de l'algorithme revêt un caractère fondamental puisque ce sont ces variables qui vont permettre de prédire la variable de sortie. L'ensemble des variables constitutives du bilan prudentiel des sociétés seront retenues dans un premier temps pour estimer les corrélations entre les variables. Ensuite, une sélection sera faite afin de retenir les variables qui permettront d'assurer la meilleure capacité prédictive des modèles. Enfin, après comparaison de la performance de plusieurs modèles, l'algorithme présentant la meilleure capacité de prédiction sera retenu pour réaliser les tests de sensibilité sur les fonds propres économiques.

1.1.3. Choix des modèles d'apprentissage et application aux tests de sensibilité

Le choix de l'algorithme permet de définir la fonction de transformation des variables. Dans notre étude, nous visons la prédiction d'une variable de sortie réelles, en nous appuyant sur des données étiquetées, il nous faut donc recourir à un algorithme d'apprentissage supervisé de régression qui répond à nos besoins à la fois de continuité et d'apprentissage en mode supervisé (c'est-à-dire lorsqu'on connait la valeur de la prédiction dans notre base d'apprentissage).

L'apprentissage automatique supervisé consiste à prédire une variable Y à partir d'un vecteur X de variables observées contenant la variable à prédire. Pour ce faire, on calibre l'algorithme sur la base de données préalablement scindée en une base d'apprentissage et en base de test, afin de déterminer dès la conception de l'algorithme la qualité de la prédiction en fonction de l'erreur constatée sur la base de test.

Les différents algorithmes étudiés dans le présent mémoire sont la régression linéaire multiple, les arbres de régression, les forêts aléatoires et le gradient boosting. En fonction de la qualité de la prédiction, le modèle le plus pertinent sera retenu.

Dans un second temps, les fonds propres économiques feront l'objet de tests de sensibilité afin d'observer leur comportement en cas de modification de telle ou telle variable d'entrée.

1.2. Engagements d'assurance vie

1.2.1. Généralités sur l'assurance vie

L'assurance vie est le premier moyen d'épargne en France. Elle totalise, en valeur économique, plus de 2 000 milliards d'euros d'engagements au 31 décembre 2020, selon les données issues des remises réglementaires transmises à l'ACPR.

De manière générale, l'assurance vie est un contrat par lequel l'assureur s'engage, en contrepartie de la perception de primes, à verser une rente ou un capital à une ou des personnes déterminées. Il existe trois types de contrats d'assurance vie : l'assurance en cas de vie, l'assurance en cas de décès et un contrat mixte de vie et décès. Les assurances-vie garantissent le versement d'un capital ou d'une rente au souscripteur ou au bénéficiaire désigné dans le contrat. L'assurance en cas de décès constitue une garantie pour les proches de l'assuré, alors que l'assurance en cas de vie est davantage utilisée comme placement, l'assuré pouvant être lui-même le bénéficiaire du contrat.

De nombreuses formules d'assurance vie existent sur le marché, et varient selon la durée choisie et les options de sortie (versement d'une rente ou d'un capital). Les risques encourus par l'assuré varient également selon le support choisi : les contrats souscrits en euros bénéficient d'un capital garanti, alors que le capital des contrats en unité de compte ou en action varie en fonction des marchés.

Les contrats d'assurance vie sont régis par le code des assurances (notamment les articles L131-1 et L132-1 et suivants). La fiscalité de l'assurance vie est différente selon les contrats et selon les conditions de sortie. Ainsi, les bénéficiaires de contrats liquidés au moment du décès bénéficient d'une exonération de droits de succession dans les conditions précisées par la documentation fiscale.

1.2.2. Focus sur l'assurance vie en euros

L'assurance vie en euros est un support d'investissement spécifique aux contrats d'assurance vie et de capitalisation. Il s'adresse généralement aux épargnants qui recherchent la sécurité pour leur investissement ou qui ont un horizon de placement à moyen terme.

Les garanties d'un contrat d'assurance vie en euros prennent la forme suivante :

- Au terme, si l'assuré est en vie : paiement d'un capital ou d'une rente à l'assuré.
- En cas de décès de l'assuré : paiement d'un capital ou d'une rente au(x) bénéficiaire(s) désigné(s).

Le contrat comporte une garantie en capital qui est au moins égale aux sommes versées, nettes ou brutes de frais selon les contrats. Il peut être prévu le versement de participation aux bénéfices contractuelle. Les conditions d'affectation des bénéfices techniques et financiers de chacun des fonds en euros doivent être conformes aux conditions réglementaires de participations aux bénéfices. Le contrat d'assurance vie comporte une faculté de rachat à tout instant.

Le contrat d'épargne en euro est un contrat individuel ou collectif, souscrit par auprès d'un assureur. Ce contrat est régi par le Code des assurances et relève de la branche 20 « Vie-Décès » pour les supports en euros ou 22 « Assurances liées à des fonds d'investissement » pour les supports en unités de compte.

Le contrat est alimenté par des versements et rachats libres et/ou libres programmés. L'assuré détermine librement la durée de son adhésion (viagère ou déterminée) en fonction de l'orientation patrimoniale qu'il souhaite lui donner.

La durée de l'adhésion peut être viagère, elle prend fin en cas de rachat total ou en cas de décès de l'assuré. Elle peut être d'une durée déterminée et prendre fin avant le terme, en cas de rachat total ou en cas de décès de l'assuré, ou au terme que l'assuré aura fixé sous réserve d'une demande de règlement de la valeur atteinte de l'adhésion ou de service d'une rente viagère.

Les frais applicables au titre du contrat sont les suivants :

- Frais à l'entrée et sur versements : frais sur les versements initial, libre et libres programmés.
- Frais en cours de vie du contrat : frais de gestion sur les supports en euros.
- Frais de sortie le cas échéant.

Les supports utilisés pour le placement des fonds euros sont constitués d'actifs diversifiés (obligations, actions, prêts, immobiliers et trésorerie). Ils sont investis, conformément au Code des assurances, sur les marchés financiers et immobiliers.

Concernant l'attribution des bénéfices, les bases réglementaires de tarification et provisionnement sont prudentes et impliquent (en espérance) des bénéfices techniques et financiers probables à partager avec les assurés. Il existe deux types de participation aux bénéfices (PB) :

- La PB contractuelle : les clauses contractuelles peuvent préciser le mécanisme de calcul et d'affectation de la participation aux bénéfices

- La PB réglementaire : la réglementation définit un niveau minimum de participation aux bénéfices. Son calcul se fait au niveau de l'entreprise d'assurance, et non contrat par contrat ou fonds par fonds (exception des cantons réglementaires). La rémunération globale de la compagnie est au maximum égale à 10 % du résultat technique (i.e., le résultat lié à la mortalité et à la gestion), et 15 % du résultat financier (hors produits financiers sur fonds propres). En cas de déficit technique ou financier, les pertes sont supportées par l'assureur.

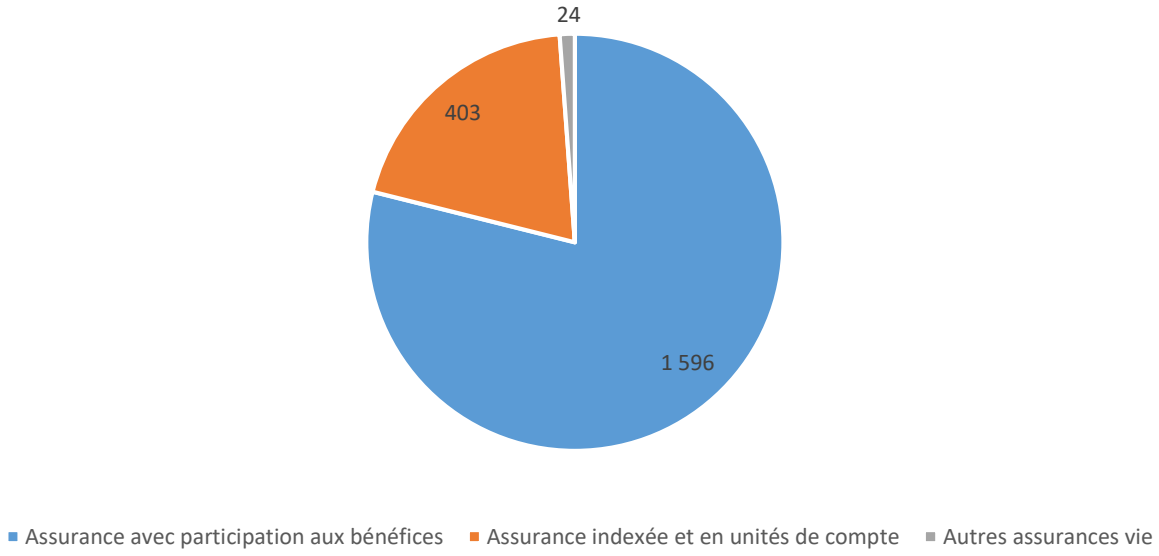
Selon les dispositions de l'article A. 132-16, le montant des participations aux bénéfices peut être affecté directement aux provisions mathématiques ou porté, partiellement ou totalement, à la provision pour participation aux bénéfices mentionnée à l'article R. 343-3. Les sommes portées à cette dernière provision sont affectées à la provision mathématique ou versées aux souscripteurs au cours des huit exercices suivant celui au titre duquel elles ont été portées à la provision pour participation aux bénéfices.

1.2.3. Données du marché de l'assurance vie

Les données présentées dans cette partie sont issues des remises réglementaires au 31 décembre 2020 des sociétés vie et mixte françaises. Elles sont exprimées en valeur économique. Le marché de l'assurance vie est quasi-exclusivement porté par les sociétés d'assurance vie et les sociétés mixte. Les sociétés d'assurance non vie ne peuvent pratiquer de l'assurance vie qu'à titre accessoire, selon le principe de spécialisation défini par le code des assurances à l'article L. 321-1.

Parmi les engagements d'assurance vie des sociétés vie et mixte, la part relative des contrats qui disposent d'une participation bénéficiaire est très majoritaire. Comme illustré à la figure 1, les données issues des remises réglementaires Solvabilité II à fin 2020 montrent que ces contrats représentent 1 596 milliards d'euros d'engagements calculés conformément aux exigences quantitatives du pilier 1. Les contrats en unités de compte représentent 403 milliards d'euros d'engagement. La meilleure estimation des autres contrats d'assurance vie s'élève enfin à 24 milliards d'euros.

Figure 1 : Répartition de la meilleure estimation des provisions techniques vie au 31 décembre 2020 en milliards d'euros



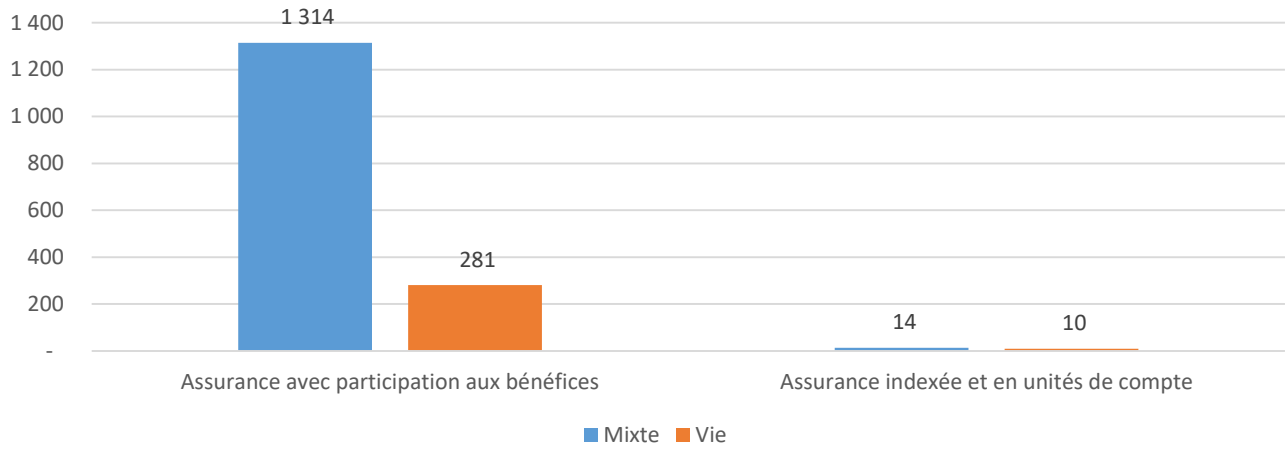
Dans le cadre du présent mémoire, nous nous intéressons plus particulièrement aux sociétés qui détiennent des engagements d'assurance vie avec participation aux bénéfices car les engagements liés à ces contrats sont calculés selon des approches comparables (souvent selon des modèles stochastiques en vision économique) tout en obéissant aux même contraintes de gestion actif/passif.

Les contrats en unité de compte ne comportent pas de risque assurantiel pour la société puisque c'est le seul assuré qui porte le risque (sauf si l'assureur commercialise des garanties plancher). Au bilan de la société d'assurance, les engagements UC au passif se compensent presque totalement avec les actifs comptabilisés en représentation de ces engagements UC. Ainsi ils n'ont quasiment aucune influence sur la constitution des fonds propres économiques de la société. Nous les excluons donc de notre étude.

Les autres engagements d'assurance vie ne fonctionnent pas avec des contraintes de gestion actif passif et l'influent pas sur la constitution des fonds propres prudentiels de la même manière que les contrats d'assurance vie avec PB. Il faudra donc garder ce point à l'esprit pour interpréter la capacité de prédiction des algorithmes sur les différents périmètres.

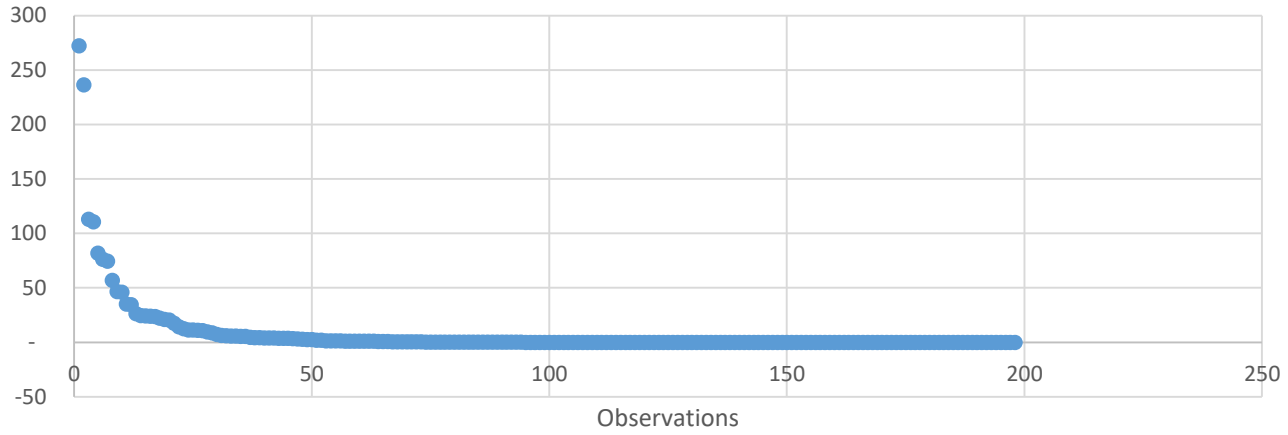
Dans le détail, les contrats avec participation aux bénéfices sont principalement détenus par les sociétés d'assurance mixte (1 314 milliards d'euros, soit 82% du total) et, dans une moindre mesure, par les sociétés vie (281 milliards d'euros soit 18% du total) comme illustré dans la figure 2. La part des autres assurance vie dans le total des engagements diffère sensiblement entre les 2 type d'organisme : seulement 1% pour les sociétés mixtes contre 3% pour les sociétés vie, selon les données issues de la figure ci-dessous. Cette différence pourra expliquer la différence de capacité de prédiction des modèles présentés.

Figure 2 : Répartition des provisions techniques (hors UC) par type d'organisme au 31 décembre 2020 en milliards d'euros



La répartition des contrats d'assurance vie avec PB diffère de manière importante. En moyenne, en 2020, une société du panel détient 30 milliards d'encours. Cependant, les sociétés d'assurance mixte détiennent en moyenne 34 milliards, alors que les sociétés vie détiennent 8 milliards d'engagements en moyenne. Les 10 sociétés détenant le plus d'encours d'assurance vie avec PB concentrent 69% des engagements, et sont toutes des sociétés mixtes. Comme l'illustre la figure 3, les 50 sociétés les plus importantes concentrent la grande majorité des engagements.

Figure 3 : Distribution des provisions techniques des contrats avec participation aux bénéfices au 31 décembre 2020 en milliards d'euros



En conclusion, étant donné la diversité du poids des engagements portés par les sociétés vie et mixte, le mémoire va s'attacher à décrire 4 périmètres différents : toutes sociétés vie et mixte, uniquement mixte, uniquement vie, et enfin toutes sociétés avec majorité de contrats avec PB. L'objectif est de voir si le modèle de prévision des fonds propres économiques peut être amélioré en sélectionnant des sociétés dont les engagements se comportent de manière plus homogène (tout périmètre, vie, mixte ou majorité de PB).

1.3. Bilan économique sous Solvabilité 2

1.3.1. Fonds propres économiques

1.3.1.1. Composition des fonds propres

Les fonds propres se décomposent en fonds propres de base et en fonds propres auxiliaires :

- Les fonds propres de base sont constitués d'une part de l'excédent des actifs sur les passifs (valorisés selon l'article L.351-1 du Code des assurances), et d'autre part des passifs subordonnés.
- Les fonds propres auxiliaires (qui font partie du hors-bilan) comprennent des éléments de passifs, autres que les fonds propres de base, pouvant être appelés et utilisés pour absorber des pertes. Ils peuvent prendre des formes très diverses (fraction de capital non appelée et/ou non versée, lettres de crédits et garanties, ou « tout autre engagement juridiquement contraignant reçu par les entreprises d'assurance et de réassurance ») ; ils doivent néanmoins être approuvés par le superviseur.

Dans le cadre du présent mémoire, nous excluons de notre analyse les fonds propres auxiliaires qui ne dépendent par des données bilancielles disponibles pour notre étude. Nous parlerons donc uniquement de « fonds propres économiques » pour désigner en réalité les seuls fonds propres de base dont le calcul sera modélisé à l'exclusion des fonds propres auxiliaires.

1.3.1.2. Classement des fonds propres

Les fonds propres sont classés selon leur niveau de qualité. Pour effectuer ce classement, la directive s'appuie (article R.351-22 du Codes des assurances) sur différents critères : la disponibilité permanente (pour absorber complètement les pertes), la subordination (en cas de liquidation, disponibilité du montant total sans remboursement possible avant que tous les autres engagements ne soient honorés), la durée suffisante de l'élément de fonds propres, l'absence d'incitation à rembourser, l'absence de charges fixes obligatoires et l'absence de contrainte.

Selon ce classement :

- Le niveau 1 (Tier 1) correspond à la meilleure qualité et ne comprend que des éléments de fonds propres de base continument et immédiatement mobilisables, disponibles en totalité et subordonnés. La réserve de réconciliation en fait partie bien qu'elle ne réponde que partiellement à ces caractéristiques.
- Le niveau 2 (Tier 2) est composé d'éléments de fonds propres de base moins facilement mobilisables, mais dont la totalité est utilisable et subordonnée, ainsi que de fonds propres auxiliaires.
- Le niveau 3 (Tier 3) enfin, comprend les fonds propres de base ne pouvant être classés dans les niveaux précédents ainsi que des fonds propres auxiliaires.

Dans le cadre de ce mémoire, nous n'utiliserons pas le classement des fonds propres ni son impact sur la notion de disponibilité ou d'exigibilité qui en découle pour couvrir l'exigence de capital dit « SCR » pour Solvency Sapital Requirement ou « MCR » pour Mimimum Capital Requirement. Cette distinction n'est pas utile puisque notre étude se limite à la prédiction des fonds propres économiques, et non pas du ratio de couverture de l'exigence de capital qui aurait nécessité de tenir compte des règles d'écrêtement des fonds propres en fonction de leur niveau.

1.3.1.3. Clause transitoire

Afin de lisser en partie l'effet du passage à cette nouvelle classification des fonds propres, une partie des fonds propres admis, sous Solvabilité I, en représentation de l'exigence de marge, sont classés en niveau 1, et dits de « niveau 1 restreint », alors qu'ils ne le seraient pas selon les règles présentées : c'est une clause transitoire relative aux droits acquis (règle du « grandfathering »). Parmi l'ensemble de ces fonds propres disponibles, les éléments dits « éligibles » à la couverture du capital de solvabilité requis (CSR) et du capital minimum requis (MCR) doivent respecter des limites quantitatives assurant que les exigences de solvabilité soient couvertes majoritairement par des fonds propres de la meilleure qualité. En particulier, les fonds propres de niveau 1 doivent couvrir au moins 50 % du CSR et 80 % du MCR.

1.3.1.4. Arrêté PPB

Depuis l'arrêté « PPB » (Provision pour Participation aux Bénéfices) de décembre 2019, les assureurs peuvent comptabiliser une large partie de la provision pour participation aux bénéfices en fonds propres prudentiels, en tant que « fonds propres excédentaires ». L'impact de ce changement de méthode n'est pas pris en compte dans l'étude du fait que les remises réglementaires n'isolent pas la contribution de la PPB dans les fonds propres économiques. Cette PPB est classée en fonds propres excédentaires depuis 2019, comme d'autres éléments de fonds propres, rendant inexact l'éventuel retraitement de cette composante du calcul des fonds propres économiques au titre qu'il serait exclusivement constitué de PPB. Il aurait cependant été souhaitable de faire un tel retraitement si les données avaient été disponibles, afin de maintenir la comparabilité des calculs de fonds propres économiques sur toute la période de l'étude, y compris depuis la mise en œuvre de l'arrêté PPB en vigueur depuis la clôture du 31 décembre 2019. Ce point constitue une limite de la présente étude.

1.3.2. Bilan prudentiel

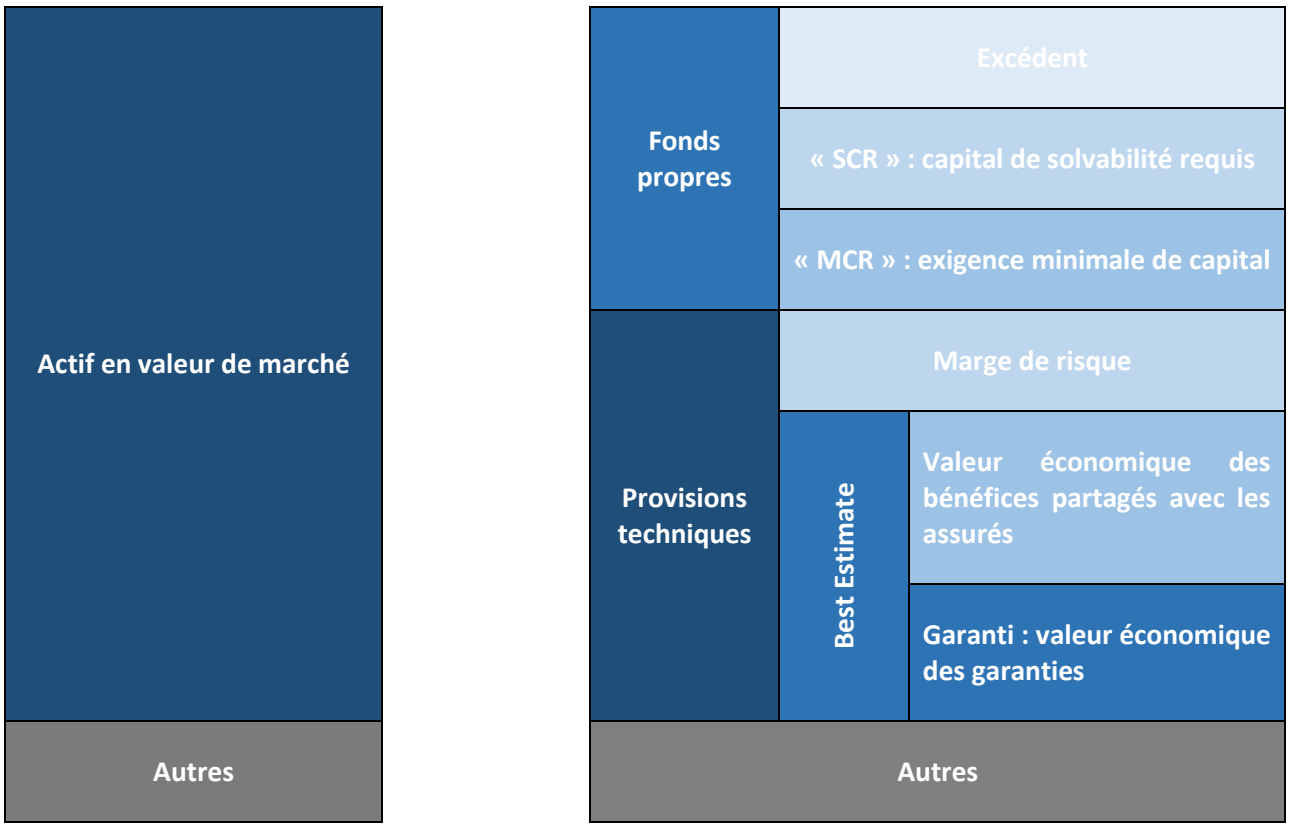
Le bilan Solvabilité 2 est fondé sur des valeurs économiques à l'actif et au passif, comme illustré dans la figure 4 ci-après. Selon l'article 75 de la directive Solvabilité II, les actifs sont valorisés au montant pour lequel ils pourraient être échangés dans le cadre d'une transaction conclue, dans des conditions de concurrence normales, entre des parties informées et consentantes. Les passifs sont valorisés au montant pour lequel ils pourraient être transférés ou réglés dans le cadre d'une transaction conclue, dans des conditions de concurrence normales, entre des parties informées et consentantes.

En outre, conformément à l'article 9 du règlement délégué, les entreprises d'assurance et de réassurance valorisent les actifs et les passifs conformément aux normes comptables internationales IFRS à condition que ces normes prévoient des méthodes de valorisation conformes à la méthode de valorisation prévue à l'article 75 de la directive SII.

Les entreprises d'assurance et de réassurance valorisent les actifs et les passifs en utilisant un prix coté sur un marché actif pour les mêmes actifs ou les mêmes passifs, en se fondant sur l'hypothèse d'une continuité d'exploitation de l'entreprise.

FIGURE 4 : BILAN ECONOMIQUE SOUS SOLVABILITE II

Actif en valeur de marché	**Passif en valeur économique**



Les prix des produits simples (actions, obligations) sont disponibles sur les marchés financiers. Pour les produits financiers complexes, il est nécessaire de recourir à des modèles de valorisation. La valorisation du passif est la plus complexe car il n'existe pas de cotation sur un marché réglementé de produit présentant des caractéristiques et des options similaires.

1.3.2.1. Cadre général de la valorisation du passif

Selon l'article L. 351-2 du code des assurances, les entreprises d'assurance et de réassurance établissent des provisions techniques prudentielles pour tous leurs engagements vis à vis des assurés, des bénéficiaires de contrats et des entreprises réassurées.

La valeur des provisions techniques prudentielles, évaluée conformément à l'article L. 351-1, correspond au montant actuel que les entreprises devraient payer si elles transféraient immédiatement leurs engagements à une autre entité agréée pour pratiquer des opérations d'assurance ou de réassurance.

Le calcul des provisions techniques prudentielles utilise les informations fournies par les marchés financiers et les données généralement disponibles sur les risques de souscription, en cohérence avec ces informations et données. Les provisions techniques prudentielles sont calculées d'une manière prudente, fiable et objective. Ce calcul peut comporter une correction pour volatilité.

La valeur des provisions techniques prudentielles mentionnées à l'article L. 351-2 est égale à la somme de :

- La meilleure estimation : correspond à la moyenne pondérée par leur probabilité des flux de trésorerie futurs compte tenu de la valeur temporelle de l'argent estimée sur la base de la courbe des taux sans risque pertinente, soit la valeur actuelle attendue des flux de trésorerie futurs.
- La marge de risque : calculée de manière à garantir que la valeur des provisions techniques prudentielles mentionnées à l'article L. 351 2 est équivalente au montant qu'une entreprise agréée pour pratiquer les opérations d'assurance ou de réassurance demanderait pour reprendre et honorer les engagements d'assurance et de réassurance.

1.3.2.2. Calcul de la meilleure estimation en vie

Le calcul de provisions techniques en assurance vie comme une somme actualisée de prestations et frais futurs nécessite de prendre en compte le coût des options et garanties. En pratique, valoriser ces options nécessite :

- En raison de l'interaction actif passif, de faire des simulations dans différents environnements économiques.
- D'estimer les prestations à payer (p.ex. en cas de décès, rachats partiels / totaux).
- Mais aussi le comportement sur la base des données passées et futures, tant de l'assuré (qui détient l'option) que de l'assureur qui influe sur la « valeur » de celle-ci
- En faisant des hypothèses à très long terme, notamment sur le taux cible servi.

L'assureur dispose de plusieurs leviers pour gérer la richesse et piloter le rendement :

- PPB (Provision pour Participation aux Bénéfices) ;
- Réalisation de plus ou moins-values latentes ;
- Dotations de PDD (Provisions pour Dépréciation Durable) ;

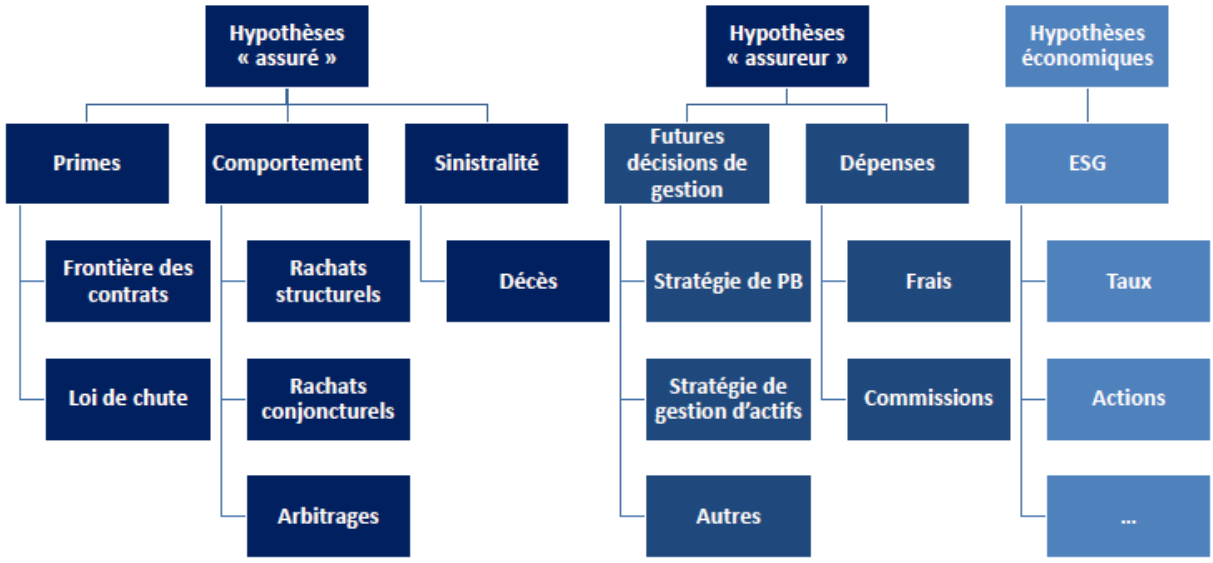
Du fait de l'asymétrie de la répartition des profits ou des pertes entre l'assureur et l'assuré, il est nécessaire de réaliser un calcul stochastique : les contrats d'assurance vie qui donnent lieu à des prestations discrétionnaires dépendent de rendements d'investissements ou comportent des garanties financières et des options contractuelles qui nécessitent de recourir à des méthodes par simulation. La somme actualisée des flux de trésorerie futurs $Flux_{i,t}$ sur N trajectoires économiques équiprobables (de i allant de 1 à N) se présente comme suit :

$$BE = \sum_i \frac{1}{N} \sum_t Flux_{i,t} \times D_{i,t}$$

Avec $D_{i,t} = \frac{1}{\prod_t (1+r_{i,t})}$ le facteur d'actualisation (déflateur)

La modélisation stochastique par scénarios dépend d'hypothèses liées au comportement des assurés, de l'assureur mais également d'hypothèses économiques, comme le montre la figure 5 ci-dessous.

FIGURE 5 : HYPOTHESES DU MODELE DE CALCUL DES ENGAGEMENTS D'ASSURANCE VIE



HYPOTHESES ASSURE

En pratique pour les contrats d'épargne en France, on projette très rarement des primes car l'essentiel des encours présente des garanties « non discernables » à 0% net de chargements. Pour mémoire, une option « discernable » présente une valeur pour l'assuré supérieure à ce qu'il pourrait obtenir sur le marché à un instant donné. Le comportement des assurés nécessite notamment la modélisation des rachats conformément à leur nature :

- Rachats structurels : les comportements dits « structurels » sont en partie liés aux avantages fiscaux des contrats d'assurance vie. Ils dépendent ainsi du nombre d'années de détention du contrat.
- Rachats conjoncturels : faire dépendre les rachats de l'écart entre un taux servi et un taux « attendu » (taux concurrentiel). Par exemple, si le taux servi est inférieur de plus de 6% par rapport au taux benchmark, 20 % des assurés rachètent leur contrat.

La mortalité est modélisée à partir de tables d'expérience ou réglementaires.

HYPOTHESES ASSUREUR

Futures décisions de gestion

Les futures décisions de gestion couvrent les décisions suivantes :

- La stratégie de PB et de taux cible (niveau de revalorisation cible) : elle détermine la partie discrétionnaire qui sera versée à l'assuré, au-delà de la revalorisation réglementaire. Pour servir le taux cible, l'assureur peut utiliser plusieurs leviers : reprise de PPB, réalisation de PVL ou abandon de marge en cas de déficit entre le disponible et la cible. En cas d'excès, réalisation de MVL, dotation à la PPB et reprise sur la réserve de capitalisation.

- La stratégie de gestion d'actifs : l'assureur projette chaque année l'actif avec réinvestissement des coupons, dividendes et nominaux, en reflétant sa politique d'investissement. L'allocation peut être fixe ou dépendre du temps ou du scénario.

Dépenses

Les frais d'administration, de gestion des sinistres et liés aux investissements doivent être pris en compte (article 31 du règlement délégué). Les hypothèses de coûts (par exemple unitaire) permettent d'estimer les dépenses à venir. L'assureur doit se placer en principe de continuité d'activité et les commissions doivent être modélisées selon les mécanismes prévus dans les accords de commissionnement.

Hypothèses économiques

Comme évoqué plus haut, les entreprises d'assurance et de réassurance établissent des provisions techniques prudentielles pour tous leurs engagements vis-à-vis des assurés, des bénéficiaires de contrats et des entreprises réassurées. Le calcul des provisions techniques prudentielles utilise les informations fournies par les marchés financiers et les données généralement disponibles sur les risques de souscription, en cohérence avec ces informations et données. Les provisions techniques prudentielles sont calculées d'une manière prudente, fiable et objective. Ce calcul peut comporter un ajustement égalisateur (non utilisé en France) ou une correction pour volatilité.

- Modèle de taux : élément central de l'ESG car il fournit l'évolution du numéraire pour les autres modèles (actions, immobilier) qui sont en « excess return ». Les modèles déplacés permettent la diffusion des taux négatifs. Les modèles sont calibrés sur des « nappes de volatilité ».
- Modèle action / immobilier : modèle plus ou moins complexe peut dépendre d'un ou plusieurs indices (CAC, private equity, infrastructure…). Un modèle courant est le modèle de Black and Scholes, avec recours à un mouvement brownien géométrique.

Ces modèles doivent faire l'objet de test pour s'assurer de leur cohérence. L'assureur doit vérifier la martingalité des prix, la convergence Monte-Carlo des scénarios et la cohérence avec les données de marché (market-consistency).

Courbe de taux

Le BE correspond à la somme actualisée des flux de trésorerie futurs :

$$BE = \sum_{t} \frac{Flux_t}{(1 + r_t)^t}$$

Avec r_t la courbe des taux sans risque pertinente soit le taux sans risque de base avec ou sans les mesures issues des ajustements du paquet branche longues (correction pour volatilité, ajustement égalisateur, mesures transitoires taux et PT) et $Flux_t$ l'ensemble des mouvements liés au versements aux primes, prestations et frais des contrats à la date t.

Distinction BEG / FDB

Lors du calcul du Best Estimate, l'organisme doit distinguer ce qui relève des Future Discretionary Benefits (FDB), revalorisations supplémentaires à l'exigence règlementaire de participation des assurés du reste du BE, dit Best Estimate Garanti (BEG). Lors du calcul du SCR, le montant de FDB est utilisé comme montant maximum de capacité d'absorption des pertes par les PT.

En conclusion, les fonds propres économiques proviennent de la valorisation économique du bilan des assureurs. Cette valorisation repose sur des données de marché, ou à défaut sur des modèles qui prennent en compte le caractère probabiliste des flux futurs et l'effet de l'actualisation sur la durée de vie des engagements. Les hypothèses et techniques de modélisations dépendent de la nature des engagements et des options retenues par chaque assureur. Il parait dès lors intéressant d'observer l'évolution des fonds propres économiques sur la base des données historiques communiquées à l'ACPR, à l'échelle du marché de l'assurance vie national, et de s'intéresser plus particulièrement aux principales composantes bilancielles de ces fonds propres économiques.

2. Chapitre 2 : constitution de la base de données

2.1. Présentation des données

2.1.1. Contexte

Les organismes d'assurance relevant du régime dit Solvabilité II sont tenus de remettre à l'ACPR[1], entre autres rapports, les états ARS/AES annuels et QRS/QES trimestriels qui fournissent des informations détaillées comme les fonds propres économiques, les provisions techniques, le bilan et les SCR et MCR de la société.

La fiabilité des données issues des remises réglementaires a fait l'objet d'un suivi spécifique par les services de l'ACPR et de l'EIOPA depuis la mise en œuvre de la nouvelle réglementation. Les autorités nationales et européennes contrôlent entre autres l'exhaustivité des remises des organismes soumis à la réglementation, en pratiquant le cas échéant des relances auprès les organismes qui n'auraient pas remis un état dans les délais impartis. Elle réalise également de nombreux contrôles sur la qualité des remises, au sein des différents états, afin de s'assurer de la cohérence des données renseignées par les organismes.

La nature et l'ampleur des contrôles réalisés par les parties prenantes, sans oublier les organismes eux-mêmes qui embarquent des contrôles automatisés au travers des outils de reporting qu'ils utilisent, tendent à donner une image de fiabilité des données issues des reporting utilisés dans la présente étude. Cependant il n'est pas exclu que ces remises contiennent certaines erreurs qui, à l'échelle de la présente étude pluriannuelle, peuvent être considérées comme acceptables pour la validité des résultats décrits.

2.1.2. Données de l'étude

L'étude se base sur les données incluses dans les remises quantitatives prudentielles dites ARS / AES pour les données annuelles et QRS / QES pour les données trimestrielles.

Ces données sont disponibles « ligne à ligne » pour chacun des organismes, ce qui représente un historique d'informations assez dense sur la période de référence : le périmètre recense 2 771 observations sur 197 organismes d'assurance vie et mixte sur la période du 31 décembre 2016 au 30 juin 2021 sur une fréquence de reporting trimestrielle, soit 22 trimestres d'observations de l'évolution des fonds propres économiques des sociétés d'assurance vie et mixte françaises. Les données sont issues de l'onglet S.02.01.01 relatif au bilan économique des assureurs. Toutes les variables sont utilisées, à l'exception des montants relatifs aux UC (à l'actif et au passif) car ils n'influent pas sur les fonds propres économiques. On exclut également les postes de bilan systématiquement non renseignés, à savoir les goodwill et les frais d'acquisition différés, qui ne s'appliquent pas dans le cadre du régime Solvabilité II. Comme indiqué précédemment, les données semblent exhaustives au regard des contrôles effectués par les parties prenantes avant la remise des bilans économiques. Nous n'avons pas identifié de données manquantes ou aberrantes suite à différents contrôles réalisés sur les principales variables (provisions techniques vie, placements et fonds propres).

Afin de définir quelles variables explicatives vont permettre de construire l'algorithme de prédiction des fonds propres de base, on rappelle la définition des fonds propres de base : les fonds propres de base sont constitués d'une part de l'excédent des actifs sur les passifs (valorisés selon l'article L.351-1 du Code des assurances), et

[1] Conformément à l'instruction n°2016-I-16 et aux articles L. 355-1 et L. 356-21 du code des assurances

d'autre part des passifs subordonnés. Dans le cadre de notre étude, nous prenons donc en compte l'ensemble des variables constitutives du bilan prudentiel des sociétés d'assurance dans un premier temps.

Les variables, avant sélection, sont présentées dans le tableau 1 ci-dessous :

TABLEAU 1 : VARIABLES POUR LE CALCUL DES FONDS PROPRES ECONOMIQUES

Numéro	Variable
1	Immobilisations incorporelles
2	Actifs d'impôts différés
3	Excédent du régime de retraite
4	Immobilisations corporelles détenues pour usage propre
5	Biens immobiliers (autres que détenus pour usage propre)
6	Détentions dans des entreprises liées, y compris participations
7	Actions – cotées
8	Actions – non cotées
9	Obligations d'État
10	Obligations d'entreprise
11	Titres structurés
12	Titres garantis
13	Organismes de placement collectif

14	Produits dérivés actif
15	Dépôts autres que les équivalents de trésorerie
16	Autres investissements
17	Prêts et prêts hypothécaires
18	Montants recouvrables au titre des contrats de réassurance
19	Dépôts auprès des cédantes
20	Créances nées d'opérations d'assurance et montants à recevoir d'intermédiaires
21	Créances nées d'opérations de réassurance
22	Autres créances (hors assurance)
23	Actions propres auto-détenues (directement)
24	Éléments de fonds propres ou fonds initial appelé(s), mais non encore payé(s)
25	Trésorerie et équivalents de trésorerie
26	Autres actifs non mentionnés dans les postes ci-dessus
27	Provisions techniques non-vie (hors santé)
28	Provisions techniques santé (similaire à la non-vie)
29	Provisions techniques santé (similaire à la vie)

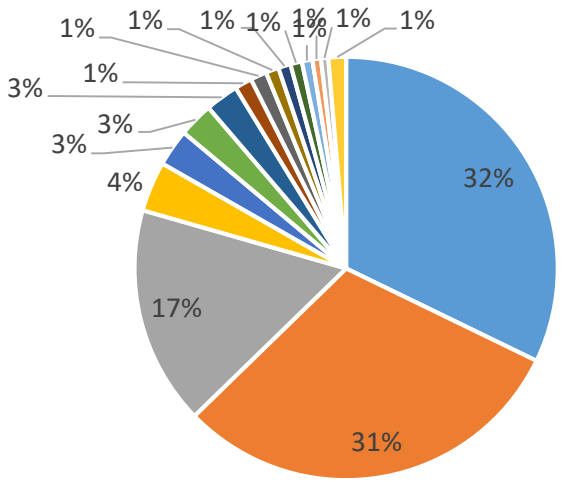
30	Provisions techniques vie (hors santé, UC et indexés)
31	Autres provisions techniques
32	Passifs éventuels
33	Provisions autres que les provisions techniques
34	Provisions pour retraite
35	Dépôts des réassureurs
36	Passifs d'impôts différés
37	Produits dérivés passif
38	Dettes envers des établissements de crédit
39	Dettes financières autres que celles envers les établissements de crédit
40	Dettes nées d'opérations d'assurance et montants dus aux intermédiaires
41	Dettes nées d'opérations de réassurance
42	Autres dettes (hors assurance)
43	Passifs subordonnés
44	Autres dettes non mentionnées dans les postes ci-dessus

Les figures 6 et 7 présentent la décomposition des principales variables afin de déterminer, en 1ère approche, les principaux contributeurs à la formation des fonds propres de base. Il s'agit d'une répartition moyenne basée sur

l'ensemble des observations des 22 trimestres afin de refléter au mieux la composition moyenne sur l'ensemble de la période. On observe :

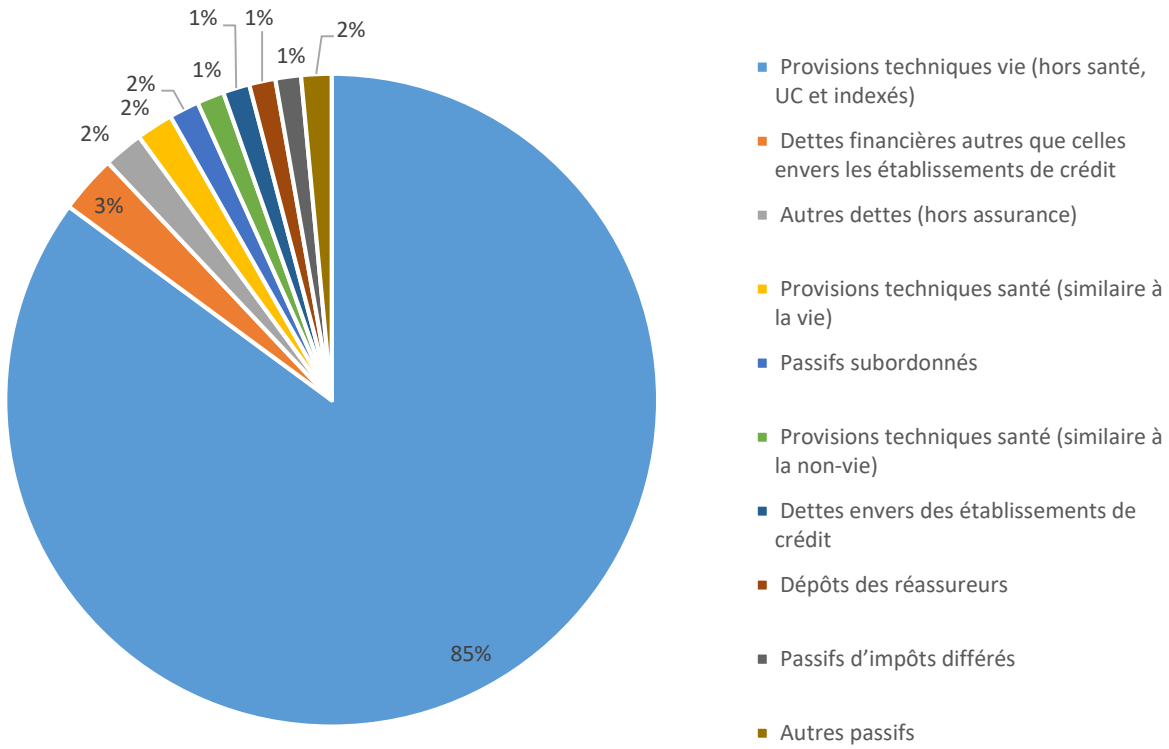
- À l'actif, comme illustré en figure 6, une forte contribution des obligations souveraines (32%), des obligations d'entreprise (31%) et des OPC (17%), et dans une moindre mesure de la réassurance (4%), des participations (3%) et des actions cotées (3%). Les autres actifs (10% en cumulé) sont individuellement faiblement contributeurs à la constitution de l'actif sur la période.

Figure 6 : Composition moyenne de l'actif (hors UC) sur la période d'observation



- Au passif, comme illustré en figure 7, on note une très forte concentration sur les PT vie (85%), alors que les dettes financières (3%), les autres dettes non assurantielles (2%), les PT santé SLT (2%) et les passifs subordonnés (2%) sont plus marginaux. Les autres passifs (6%) représentent le reste de la composition du passif en moyenne sur la période d'observation.

Figure 7 : Composition moyenne du passif (hors UC) sur la période d'observation



À ce stade, cette analyse n'est pas suffisante pour expliquer la composition des fonds propres de base pour plusieurs raisons :

- La part prépondérante de certaines variables dans l'actif ou le passif des sociétés de l'échantillon ne signifie pas qu'il existe une corrélation entre ces variables et le montant des fonds propres de base ;
- La volatilité des variables explicatives n'est pas prise en compte afin de déterminer la part de l'explication de chacune des variables dans la volatilité des fonds propres de base.

2.2. Analyse descriptive

L'analyse de la distribution des fonds propres de base dans les 4 scénarii que nous avons défini au 1er chapitre fournit une première analyse du comportement de la variable dans les différentes situations.

Pour rappel, nous avons établi 4 échantillons différents afin de tenter d'améliorer la prévision du modèle en faisant converger le comportement des échantillons vers le groupe le plus homogène possible, comme indiqué dans le tableau 2 ci-dessous :

TABLEAU 2 : ÉCHANTILLONS POUR LE CALCUL DES FONDS PROPRES ECONOMIQUES

Échantillon	Périmètre	Nombre d'observations
Échantillon 1	Toutes les données relatives aux sociétés vie et mixtes	2 771 observations sur 22 trimestres
Échantillon 2	Toutes les données relatives uniquement aux sociétés vie	734 observations sur 22 trimestres
Échantillon 3	Toutes les données relatives uniquement aux sociétés mixte	2 037 observations sur 22 trimestres
Échantillon 4	Toutes les données relatives aux sociétés vie et mixte dont plus de 70% des engagements sont constitués de contrats avec PB	807 observations sur 22 trimestres

2.2.1. Nuage des observations

En première analyse, nous nous intéressons à la répartition des observations des fonds propres économiques, afin de déceler d'éventuelles divergences de distribution entre les échantillons. On observe dans les figure 8.a à 8.d ci-dessous que les fonds propres économiques des sociétés vie présentent une moyenne et un maximum bien en dessous que les sociétés mixtes, et une variance plus faible. Les observations de l'échantillon 4 des sociétés portant majoritairement des contrats avec PB présentent des caractéristiques hybrides avec une dispersion qui s'approche des sociétés mixtes et une moyenne et un maximum supérieurs aux sociétés vie. Les caractéristiques du nuage se rapprochent de l'échantillon 1 (toutes sociétés) avec cependant un nombre d'observations en baisse notable (près de 2 000 observations en moins sur la période).

Figure 8.a : nuage des observations des fonds propres de base de l'échantillon 1 (sociétés vie et mixte)

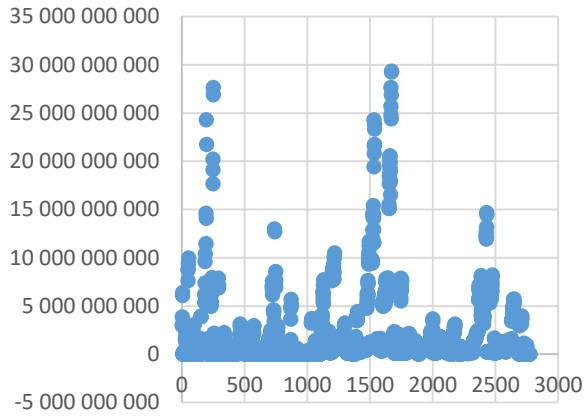


Figure 8.b : nuage des observations des fonds propres de base de l'échantillon 2 (sociétés vie)

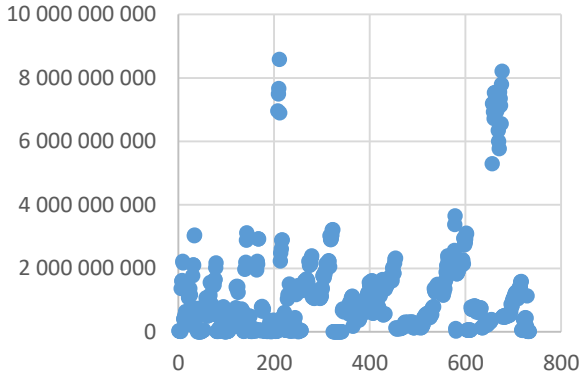


Figure 8.c : nuage des observations des fonds propres de base de l'échantillon 3 (sociétés mixte)

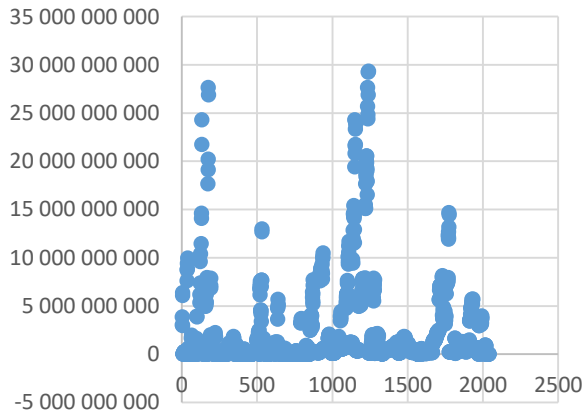
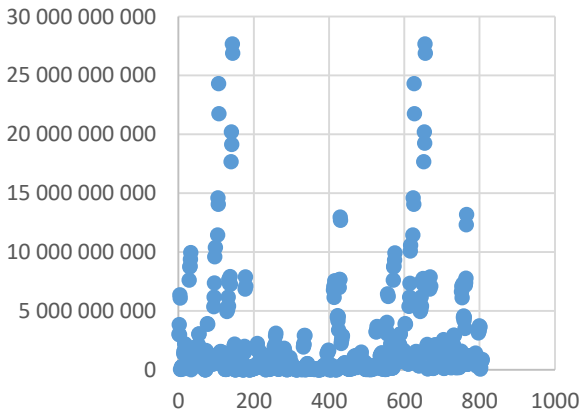


Figure 8.d : nuage des observations des fonds propres de base de l'échantillon 4 (sociétés majorité de contrats avec PB)



2.2.2. Principales statistiques

Dans l'analyse présentée dans le tableau 3 ci-dessous, on confirme cette analyse en observant notamment que la médiane des fonds propres des sociétés vie se rapproche beaucoup plus de la moyenne que les sociétés mixtes en raison de la dispersion plus contenue des observations des sociétés vie. A contrario l'échantillon 4 présente un écart entre médiane et moyenne qui s'approche de celui du 1er échantillon contenant toutes les sociétés.

TABLEAU 3 : PRINCIPALES STATISTIQUES DES 4 ECHANTILLONS DE L'ETUDE

	Echantillon 1 (toutes sociétés)	Echantillon 2 (sociétés vie)	Echantillon 3 (sociétés mixte)	Echantillon 4 (Majorité PB)
Observations	2 771	734	2 037	807
Moyenne	1 583 254 904 €	1 069 573 144 €	1 768 351 817 €	1 856 865 550 €
Écart type	3 235 974 992 €	1 419 632 792 €	3 659 461 962 €	3 676 464 595 €
1er quartile	125 288 051 €	161 936 506 €	111 275 790 €	154 747 795 €
Médiane	470 088 566 €	637 711 566 €	398 952 662 €	607 486 926 €
3ème quartile	1 343 720 094 €	1 425 462 267 €	1 198 041 182 €	1 543 328 700 €
Min	2 617 757 €	4 051 559 €	2 617 757 €	4 208 027 €
Max	29 320 663 263 €	8 575 650 222 €	29 320 663 263 €	27 657 686 668 €

2.3. Analyse univariée

Afin de compléter la première analyse statistique, l'objectif est d'identifier le lien de chacune des variables d'entrée X avec les fonds propres de base Y. Pour cela on utilise le coefficient de corrélation de Pearson$Correl(X,Y)$ qui est défini par la relation suivante :

$$Correl(X,Y) = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sqrt{\sum(x-\bar{x})^2 * \sum(y-\bar{y})^2}}$$

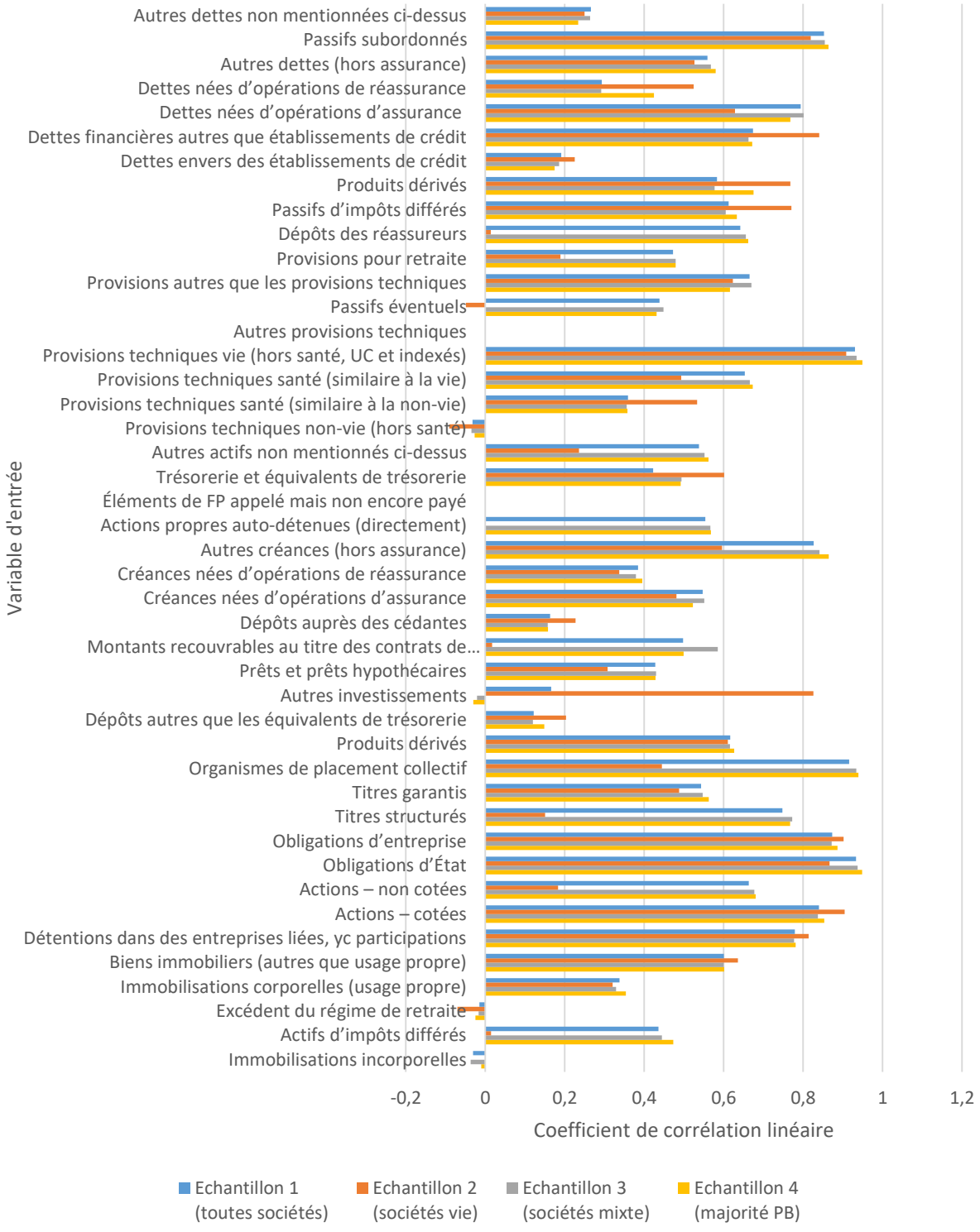
où $\bar{x}$ et $\bar{y}$ sont les moyennes des échantillons

Cette relation permet d'établir s'il existe ou non une corrélation linéaire entre X et Y. Un coefficient qui s'approche de 1 signale une corrélation linéaire qui tend vers la corrélation parfaite. A l'inverse, un coefficient de corrélation qui s'approche de 0 signifie qu'il n'existe pas de corrélation linéaire entre les variables.

Dans la figure 9, nous avons représenté l'ensemble de coefficients de corrélation de Pearson de chacune des variables avec les fonds propres économiques afin de déterminer quelles sont les variables les plus

linéairement corrélées aux fonds propres économiques, et ce pour chacun des 4 échantillons de l'étude. Il s'agit des corrélations entre chacune des variables et les fonds propres économiques.

Figure 9 : valeur des corrélations linéaire entre les variables d'entrée et les fonds propres prudentiels



Les corrélations linéaires les plus élevées sont, conformément à l'intuition, liées aux variables d'entrée les plus importantes du bilan prudentiel. Cependant, on observe quelques différences notables entre les corrélations et le poids des variables au bilan économique illustrées dans la figure 9 :

- La plus forte corrélation n'est pas celle de la variable la plus importante du bilan prudentiel, à savoir les provisions techniques vie pour les échantillons 1, 3 et 4.
- Les obligations d'entreprise sont moins corrélées qu'attendu, et notamment que les OPC alors qu'elles pèsent près du double des OPC au bilan prudentiel.
- Les passifs subordonnés, qui représentent une faible part du bilan prudentiel, sont très corrélés aux fonds propres économiques.
- Les montants recouvrables au titre de la réassurance et les dettes financières autres que les établissements de crédit sont faiblement corrélées aux fonds propres alors qu'ils représentent une part significative du bilan prudentiel.

Enfin, on note que les corrélations sont très homogènes sur les échantillons à l'exception notable de l'échantillon 2 qui présente une table de corrélation particulière, comme illustré dans le tableau 4 ci-dessous. On observe notamment la forte dépendance des fonds propres aux provisions techniques vie et aux actions cotées, au détriment des obligations d'État et d'entreprise comme pour les autres échantillons. La corrélation avec les OPC est remplacée par celle des autres investissements.

TABLEAU 4 : CORRELATION LINEAIRE DES PRINCIPALES VARIABLES AVEC LES FONDS PROPRES ECONOMIQUES

Numéro	Variable	Echantillon 1 (toutes sociétés)	Echantillon 2 (sociétés vie)	Echantillon 3 (sociétés mixte)	Echantillon 4 (majorité PB)	Montant moyen sur le périmètre sociétés vie et mixte au 31/12/2020
9	Obligations d'État	0,933	0,866	0,937	0,948	6 499 803 077
30	Provisions techniques vie (hors santé, UC et indexés)	0,931	0,908	0,934	0,949	15 923 390 854
13	Organismes de placement collectif	0,916	0,444	0,934	0,939	3 374 694 101
10	Obligations d'entreprise	0,873	0,902	0,872	0,887	6 164 974 336
43	Passifs subordonnés	0,853	0,819	0,854	0,864	277 865 750
7	Actions – cotées	0,840	0,905	0,837	0,853	512 103 789
22	Autres créances (hors assurance)	0,826	0,596	0,841	0,865	162 009 563
40	Dettes nées d'opérations d'assurance	0,794	0,629	0,801	0,768	58 742 417
6	Détentions dans des entreprises liées, yc participations	0,779	0,814	0,777	0,781	570 851 291
11	Titres structurés	0,748	0,150	0,773	0,767	534 153 175

En conclusion de l'analyse univariée, il est établi que les variables les plus fortement corrélées aux fonds propres de base sont en particulier celles qui représentent une part prépondérante au bilan prudentiel. Cependant on observe des divergences notables sur la nature des variables les plus corrélées ou sur le niveau de corrélation, qui ne suit pas toujours le poids relatif de la variable au bilan prudentiel. Enfin l'échantillon 2 des sociétés vie présente des corrélations qui lui sont propres, les autres échantillons étant plus homogènes.

2.4. Pré-sélection des variables d'intérêt

Afin de sélectionner les variables d'intérêt de notre modèle, nous nous basons sur l'analyse univariée des corrélations des fonds propres économiques avec les variables telles que présentées dans la partie précédente sur les corrélations de Pearson. Pour ce faire, nous retenons toutes les variables qui présentent un niveau de corrélation d'au moins 60% avec les fonds propres de base, et ce pour l'ensemble des échantillons. Cette sélection doit permettre d'améliorer la prédiction de notre modèle en limitant le bruit induit par les variables faiblement corrélées, et éviter par la même occasion les effets du surapprentissage.

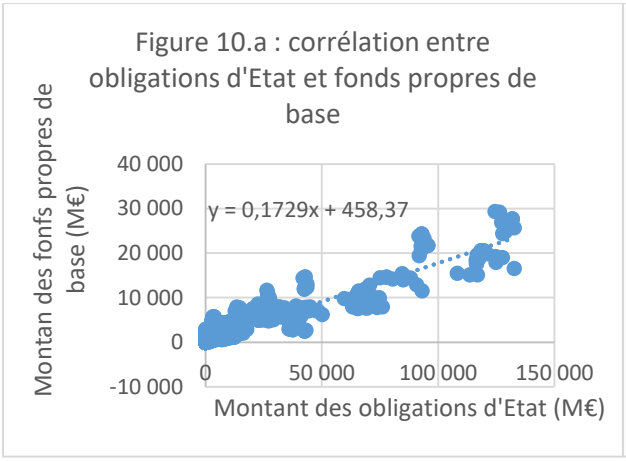
2.4.1. Sélection des variables pour les échantillons 1, 3 et 4

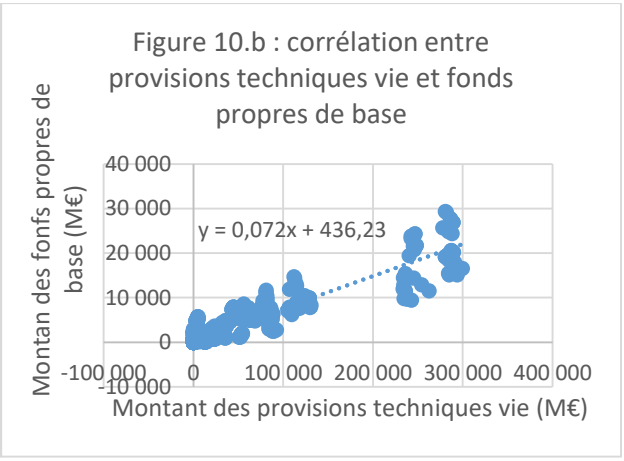
Concernant les échantillons 1, 3 et 4 qui sont homogènes en terme de corrélations, nous avons représenté graphiquement les corrélations linéaires des 6 principales variables dans les figures 10.a à 10.f suivantes.

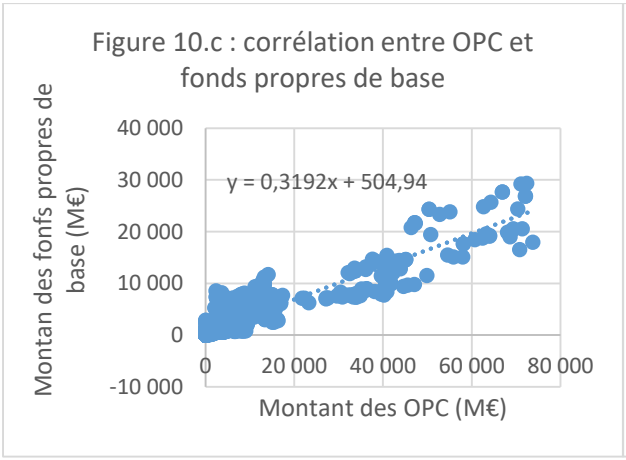
L'équation de la droite de régression linéaire, obtenue par la méthode des moindres carrés, est de la forme :

$$Y = bX + a \text{ avec } b = \frac{(n\sum x_i y_i - \sum x_i \sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} \text{ et } a = \bar{y} - b\bar{x}$$

où $\bar{x}$ et $\bar{y}$ sont les moyennes respectives des x_i et y_i.







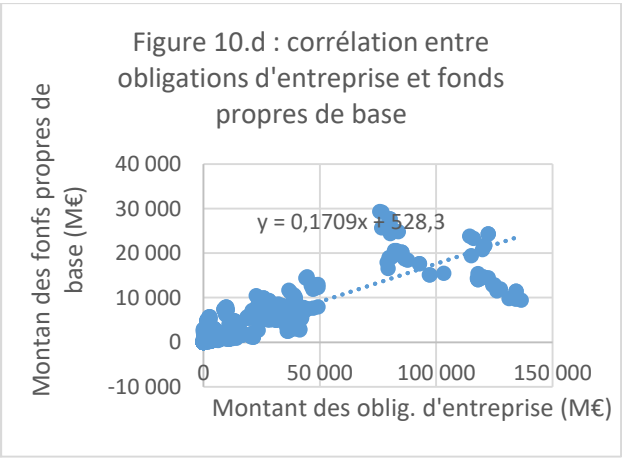


Figure 10.e : corrélation entre passifs subordonnés et fonds propres de base

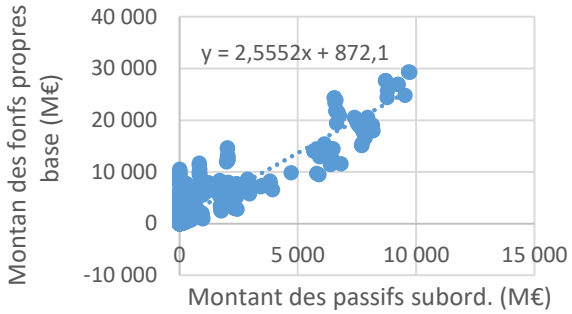
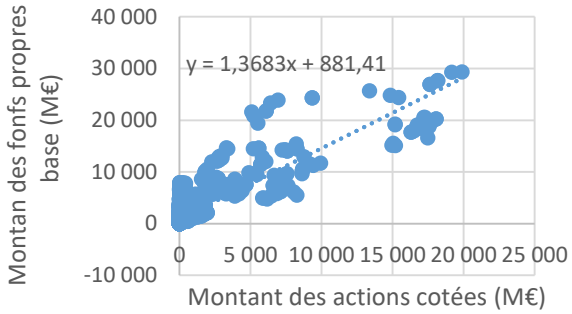


Figure 10.f : corrélation entre actions cotées et fonds propres de base



Au total nous retenons 18 variables explicatives sur ces échantillons 1, 3 et 4 qui présentent un niveau de corrélation d'au moins 60%. Ces variables sont listées dans le tableau 5 ci-dessous :

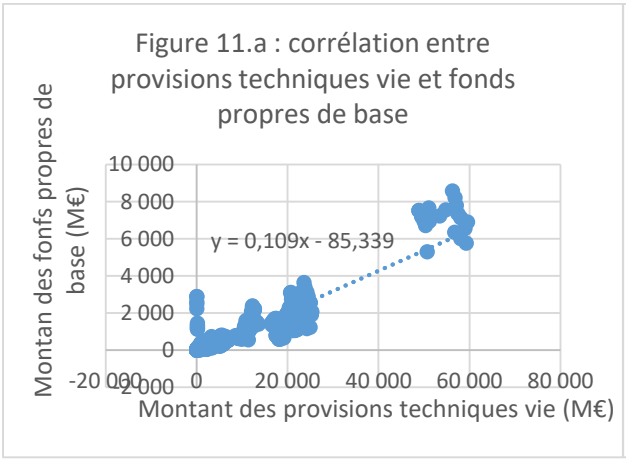
TABLEAU 5 : VARIABLES D'ENTREE DU MODELE POUR LES ECHANTILLONS 1, 3 ET 4

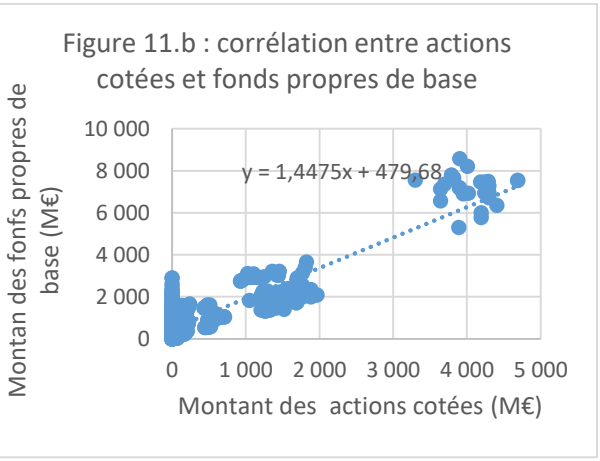
Numéro	Variables
9	Obligations d'État
30	Provisions techniques vie (hors santé, UC et indexés)
13	Organismes de placement collectif
10	Obligations d'entreprise
43	Passifs subordonnés
7	Actions – cotées
22	Autres créances (hors assurance)
40	Dettes nées d'opérations d'assurance

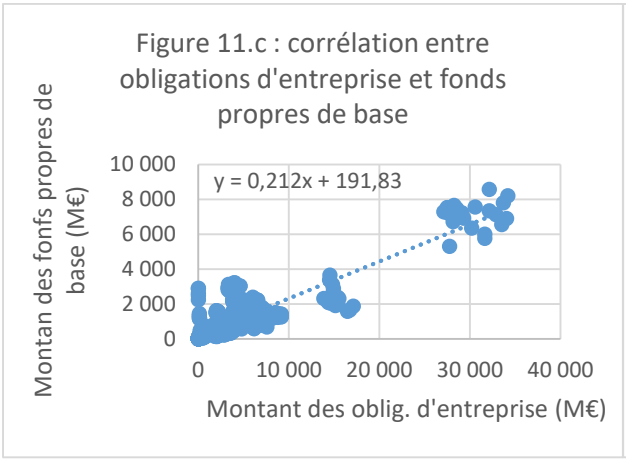
6	Détentions dans des entreprises liées, yc participations
11	Titres structurés
39	Dettes financières autres que établissements de crédit
33	Provisions autres que les provisions techniques
8	Actions – non cotées
29	Provisions techniques santé (similaire à la vie)
35	Dépôts des réassureurs
14	Produits dérivés passif
36	Passifs d'impôts différés
5	Biens immobiliers (autres que usage propre)

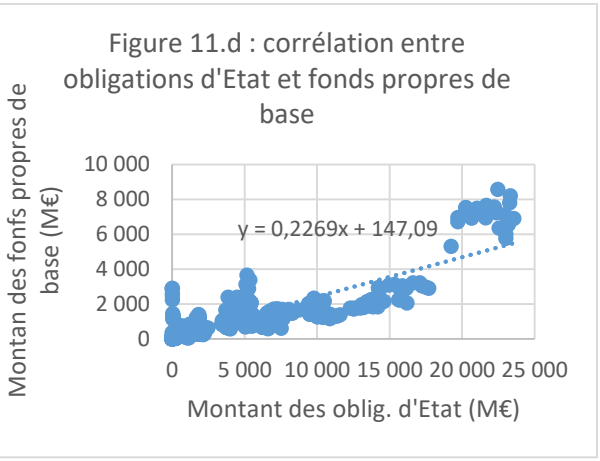
2.4.2. Sélection des variables pour l'échantillon 2

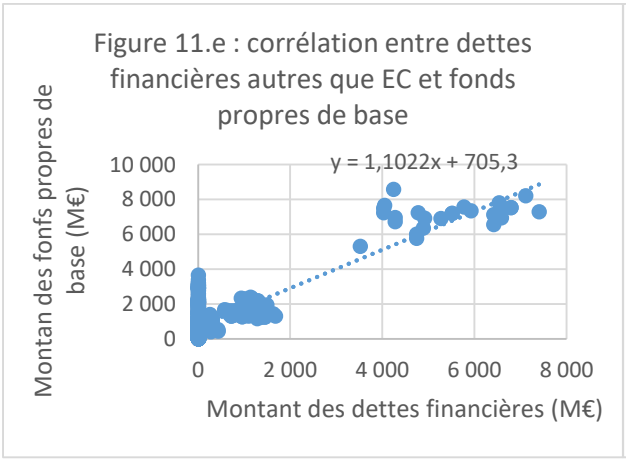
De la même manière, nous sélectionnons les variables dont la corrélation linéaire de Pearson est supérieure à 60% relativement à l'échantillon 2 des sociétés vie, qui présente une table de corrélation non homogène avec les autres échantillons, et qui est donc traitée individuellement. Les figures 11.a à 11.f ci-dessous représentent les corrélations linéaires des 6 principales variables avec les fonds propres de base :

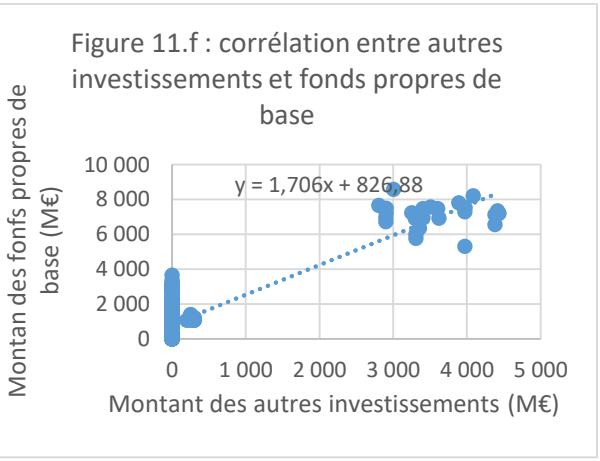












Nous retenons au final 15 variables dont le coefficient de corrélation avec les fonds propres de base est supérieur à 60% :

TABLEAU 6 : VARIABLES D'ENTREE DU MODELE POUR L'ECHANTILLON 2

Numéro	Variables
30	Provisions techniques vie (hors santé, UC et indexés)
7	Actions – cotées
10	Obligations d'entreprise
9	Obligations d'État
39	Dettes financières autres que établissements de crédit
16	Autres investissements
43	Passifs subordonnés
6	Détentions dans des entreprises liées, yc participations
36	Passifs d'impôts différés
37	Produits dérivés actif
5	Biens immobiliers (autres que usage propre)
40	Dettes nées d'opérations d'assurance
33	Provisions autres que les provisions techniques
14	Produits dérivés passif
25	Trésorerie et équivalents de trésorerie

En conclusion, l'analyse des données remises dans les états réglementaires nous a permis d'identifier les principales composantes du bilan prudentiel. Dans un second temps, nous avons analysé les corrélations entre les variations de ces composantes et celles des fonds propres de base. Cette analyse montre un écart notable entre les principales composantes et les principales corrélations. En outre, les corrélations ne sont pas apparues homogènes entre les différents échantillons, ce que nous avions pressenti dans l'analyse préalable des données, étant donné les différences entre les principales statistiques. Ces études préalables nous ont conduit à présélectionner nos variables d'intérêt différenciées par échantillon afin de maximiser la pertinence de nos algorithmes, dont nous allons détailler la construction dans le prochain chapitre.

À titre de vérification, nous avons calculé les corrélations sur un échantillon tronqué de 1 400 observations de sociétés vie et mixte. Les résultats sont en ligne avec l'analyse supra, à savoir que les principales corrélations avec les fonds propres économiques sont identiques, à savoir : obligations d'Etat (91,9% contre 93,3% initialement), provisions techniques vie (93,1% contre 93,4%) et OPC (90,8% contre 91,6%).

3. Chapitre 3 : modélisation par apprentissage automatique

L'approche retenue consiste à déduire la valeur de la variable de sortie sur la base des valeurs des variables d'entrée. La variable de sortie étant réelle (positive sauf à prédire un risque de ruine), nous retenons une approche par régression (par opposition à une méthode discrète ou binaire) afin d'estimer notre variable de sortie.

Concernant le choix des méthodes d'apprentissage automatisé, comme nous disposons d'informations sur la variable de sortie, nous retenons un modèle d'apprentissage supervisé, capable d'apprendre de la relation entre les variables d'entrée et la variable de sortie afin de construire un algorithme le plus optimal possible.

Enfin, dans le cadre d'apprentissage ainsi défini, nous allons explorer plusieurs modèles, paramétriques (modèle de régression linéaire multiple non pénalisé) ou non paramétriques (arbres de régression, forêts aléatoires et gradient boosting), afin de comparer leur performance et retenir finalement celui dont l'erreur de prédiction sera la plus réduite. Un modèle est dit « paramétrique » lorsqu'il contient un nombre fini de paramètres. Lorsque de modèle dispose de paramètre infinis, celui-ci est dit « non-paramétrique ».

Afin de rester dans des modèles facilement interprétables, le choix retenu dans le présent mémoire consiste à ne pas explorer les résultats obtenus par le biais de méthodes de « deep learning » (apprentissage automatique profond) du type réseau de neurone par exemple qui peuvent s'avérer efficaces, mais dont le fonctionnement est souvent complexe à expliciter.

3.1. Cadre général de l'apprentissage machine

Dans le cadre de notre étude, nous disposons d'observations $d_n = \{(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)\}$ où chaque couple (x_i, y_i) est une réalisation indépendante et identiquement distribuée dans l'espace des observations $(\mathcal{X}, \mathcal{Y})$. L'apprentissage supervisé doit permettre de trouver une fonction $f: \mathcal{X} \to \mathcal{Y}$ telle que $f(x) \approx y$ pour toutes les paires $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ayant la même relation que les paires observées.

Pour résoudre le problème d'apprentissage supervisé, on doit trouver une fonction $f \in \mathcal{F}$ dont les prédictions soient les plus proches possibles des véritables étiquettes, sur tout l'espace X. Pour cela nous définissons la fonction de coût $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ tel que :

$$\ell(y, y') = 0 \text{ si } y = y' \text{ et}$$

$$> 0 \text{ si } y \neq y'$$

Cette fonction de perte va permettre de mesurer le coût ou erreur entre la prédiction y′ l'observation y. La performance globale du modèle $f: \mathcal{X} \to \mathcal{Y}$ est alors donnée par la fonction de risque $\mathcal{R}$ définie comme suit :

$$\mathcal{R} \coloneqq \mathbb{E}[\ell\big(Y, f(X)\big)]$$

Trouver le meilleur modèle revient alors à minimiser le risque $\mathcal{R}$ pour une fonction de perte ℓ, soit :

$$f^* \in argmin_f \mathcal{R}(f)$$

Comme nous ne connaissons pas les étiquettes de tous les points de $\mathcal{X}$, on approche le risque par son estimation sur les données observées. Le risque empirique est alors l'estimateur :

$$\mathcal{R}_n(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, f(x_i)\big)$$

Ce qui permet d'obtenir par minimisation du risque empirique le prédicteur :

$$f = argmin_f \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, f(x_i)\big)$$

Le choix de la fonction de coût est crucial puisque c'est sur celle-ci que repose le critère de performance du modèle. Ainsi un modèle performant pour une fonction de coût ℓ_1 peut ne pas l'être si on fait le choix d'une fonction de coût ℓ_2.

Dans le cadre d'une régression, nous considérons une fonction de coût quadratique telle que :

$$\mathcal{L}_{\mathcal{SE}} \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

$$y, f(x) \to (y - f(x))^2$$

3.1.1. Échantillonnage

Pour évaluer la qualité de prédiction de notre modèle, il est indispensable d'utiliser des données étiquetées qui n'ont pas servi à le construire. Nous définissons donc un jeu d'apprentissage et un jeu de test sur la base d'une partition en 2 jeux d_{ap} et d_{te} du jeu de données $d_n = \{(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)\}$, le premier servant à l'apprentissage du modèle, et le second à son évaluation. Dans notre étude, le jeu d'apprentissage intégrera 80% des données disponibles et le jeu de test 20% des données, soit la totalité des données restantes.

3.1.2. Indicateurs de performance

Afin de comparer la performance des modèles, nous devons utiliser des indicateurs de performance homogènes et comparables entre modèles. Ces indicateurs vont permettre de quantifier l'écart entre les prédictions et les valeurs réelles du jeu de données d_{te} de test.

Nous retenons les indicateurs suivants qui vont s'appliquer à l'ensemble des algorithmes :

- Erreur quadratique moyenne : étant données les étiquettes réelles y_i les prédictions $f(x_i)$ on appelle erreur quadratique moyenne ou MSE pour mean squared error la valeur suivante :

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (f(x_i) - y_i)^2$$

- Erreur absolue moyenne : étant données les étiquettes réelles y_i les prédictions $f(x_i)$ on appelle erreur absolue moyenne ou MAE pour mean absolute error la valeur suivante :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f(x_i) - y_i|$$

- Coefficient de détermination : étant données les étiquettes réelles y_i les prédictions $f(x_i)$ on appelle coefficient de détermination R^2 la valeur suivante :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)^2}$$

Le coefficient de détermination R^2 indique à quel point les valeurs prédites sont corrélées aux valeurs réelles.

3.2. Modèles paramétriques et non paramétriques

3.2.1. Régression linéaire multiple

3.2.1.1. Fonctionnement

Le modèle de référence en régression lorsque les variables d'entrée appartiennent à l'univers des réels en d dimensions est le modèle linéaire multiple. La régression linéaire multiple appartient aux modèles de la statistique descriptive classique. Son utilisation permet de comprendre les liens entre les données d'entrée et la variable d'intérêt, ce qui constitue une bonne entrée en matière avant l'application de modèles d'apprentissage automatisés.

La fonction de régression de ce modèle de paramètre β_i est une fonction linéaire de la forme suivante :

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d + \varepsilon$$

$$\text{avec } \mathbb{E}[\varepsilon|X = x] = 0 \text{ et } \mathbb{V}[\varepsilon|X = x] = \sigma^2$$

Dans le cas de la régression, le choix de la fonction de coût liée à la perte quadratique est souvent fait afin d'estimer les paramètres du modèle. La fonction est alors définie comme suit :

$$\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^+$$

$$(y, y') \rightarrow (y - y')^2$$

Le risque quadratique pour un modèle $m : \mathcal{X} \rightarrow \mathbb{R}$ est alors tel que :

$$\mathcal{R}(m) \coloneqq \mathbb{E}[(Y - m(X))^2]$$

Le modèle optimal pour la fonction de régression basée sur le risque quadratique est alors l'espérance conditionnelle de Y sachant X définie comme suit :

$$m^*(x) \coloneqq \mathbb{E}[Y|X = x]$$

L'approche par la perte quadratique consiste à minimiser les paramètres β_i comme suit :

$$\min_{\beta_1,\dots,\beta_d} \sum_{i=1}^{n} (Y_i - \beta_1 X_{i1} - \cdots - \beta_i X_{id})^2$$

Qui dispose d'un minorant explicite :

$$\hat{\beta}_{(n)} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$$

La fonction de régression est alors de la forme :

$$m_n(x) = \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_d X_d$$

3.2.1.2. Erreur de prédiction

Dans notre étude, nous nous intéressons dans un premier temps à la capacité de prédiction des fonds propres d'un modèle de régression linéaire multiple dont les variables explicatives ont été sélectionnées en fonction de leur niveau de corrélation linéaire avec la variable à prédire.

Dans le cas des échantillons 1, 3 et 4, nous utilisons les 18 variables d'entrée et nous établissons la performance du modèle de régression linéaire multiple. Dans le cas de l'échantillon 2, nous utilisons les 15 variables d'entrée sélectionnées précédemment. Les résultats sont représentés dans les figures 12.a à 12.h.

On observe dans tous les cas une répartition en partie normale des erreurs au centre de la distribution. Cependant, on voit dans les valeurs extrêmes des certaines distributions (notamment des échantillons 1 et 3) que le modèle produit des erreurs importantes sur certaines valeurs, qui viennent biaiser fortement la qualité de la prédiction.

Si on observe bien globalement une qualité de prédiction appréciable, certaines valeurs sont très nettement différentes de l'attendu, ce qui rend compte des limites de prédiction du modèle de régression linéaire, quelque-soit l'échantillon observé. C'est notamment le cas des échantillons 1 et 3 qui présentent un niveau d'erreur conséquent qui s'observe graphiquement. A contrario, le modèle apparait meilleur pour prédire les échantillons 3 et 4 qui ne contiennent pas certaines valeurs atypiques que le modèle a du mal à prédire.

Figure 12.a : résultat de la prédiction par le modèle de régression linéaire multiple sur l'échantillon 1 des sociétés vie et mixte

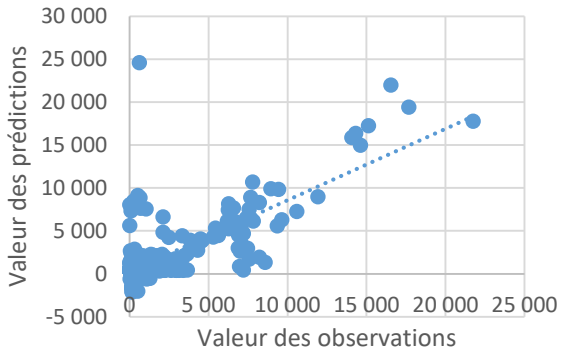


Figure 12.b : distribution des erreurs du modèle de régression linéaire multiple sur l'échantillon 1 des sociétés vie et mixte

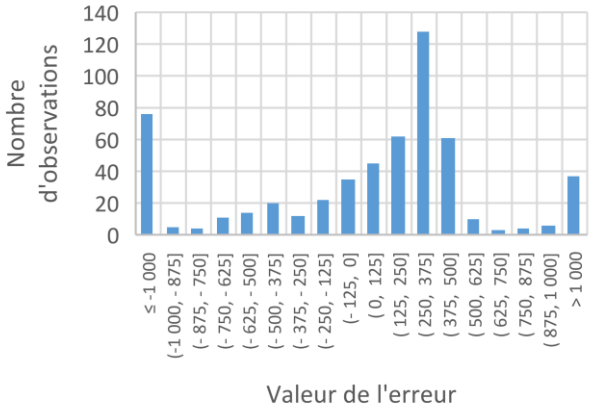


Figure 12.c : résultat de la prédiction par le modèle de régression linéaire multiple sur l'échantillon 2 des sociétés vie

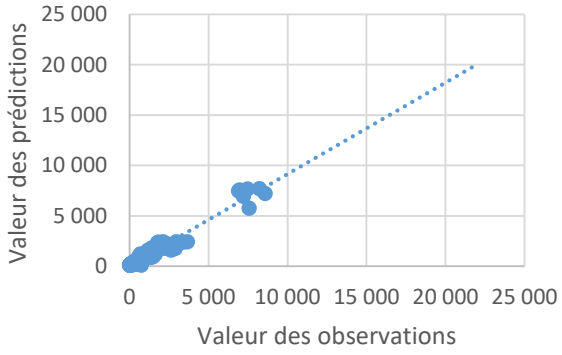


Figure 12.d : distribution des erreurs du modèle de régression linéaire multiple sur l'échantillon 2 des sociétés vie

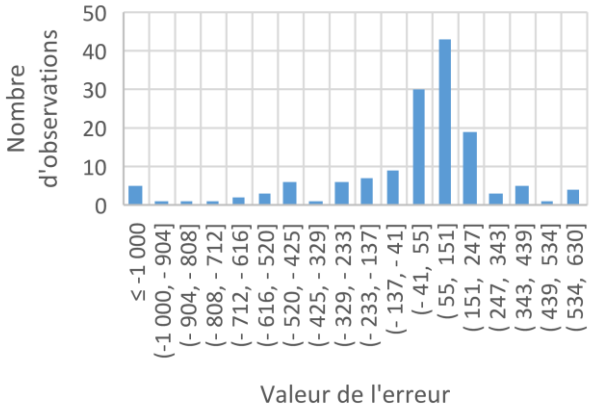


Figure 12.e : résultat de la prédiction par le modèle de régression linéaire multiple sur l'échantillon 3 des sociétés mixte

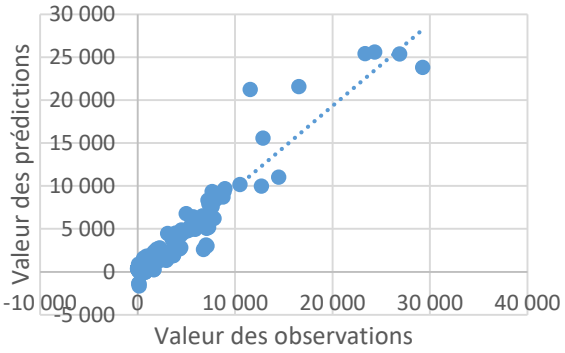


Figure 12.f : distribution des erreurs du modèle de régression linéaire multiple sur l'échantillon 3 des sociétés mixte

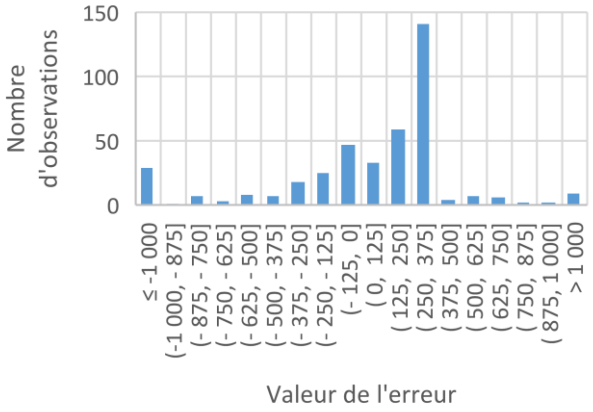


Figure 12.g : résultat de la prédiction par le modèle de régression linéaire multiple sur l'échantillon 4 des sociétés à majorité de PB

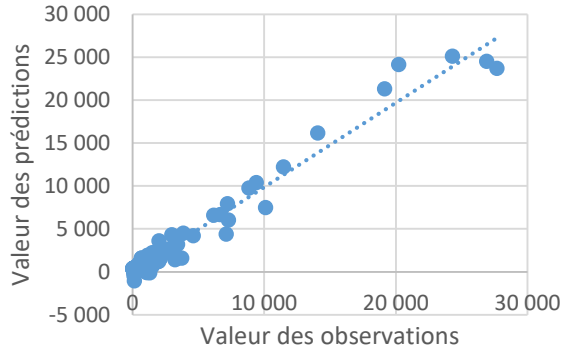
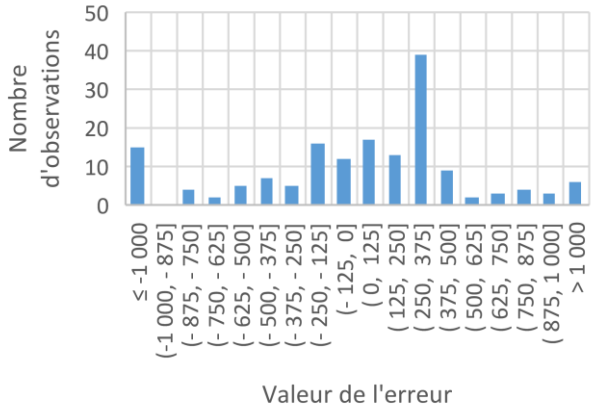


Figure 12.h : distribution des erreurs du modèle de régression linéaire multiple sur l'échantillon 4 des sociétés à majorité de PB



3.2.1.3. Performance du modèle

Afin d'établir la performance du modèle régression linéaire multiple, on s'intéresse aux indicateurs MSE (Mean Squared Error), MAE (Mean Absolute Error) et au score R^2 de la base de test. Les résultats sont récapitulés dans le tableau 7.

On observe que l'erreur quadratique et l'erreur absolue sont notablement moins élevées dans l'échantillon 2 des sociétés vie, qui présente un score R^2 appréciable (95,35%). On observe également que le modèle apparait relativement performant sur l'échantillon 4 des sociétés avec PB (score R^2 de 96,58%) avec cependant des erreurs quadratique et absolue relativement élevées soulignant la limite de prédiction du modèle sur certaines données. On retrouve le constat que nous avions fait précédemment lors de l'observation graphique des erreurs de prédiction que les échantillons 1 et 3 contiennent des valeurs atypiques mal prédites qui pénalisent grandement la qualité de la prédiction sur ces échantillons.

On observe que le score R^2 de la base d'apprentissage est quasiment toujours supérieur à celui de la base de test, ce qui tend à souligner un surapprentissage du modèle. Cependant, le résultat du score R^2 de la base d'apprentissage n'est pas non plus très élevé, ce qui confirme la capacité limitée du modèle à prévoir correctement les fonds propres de base.

TABLEAU 7 : INDICATEURS DE PERFORMANCE DU MODELE DE REGRESSION LINEAIRE MULTIPLE

Régression linéaire multiple	MSE (en M€²)	MAE (en M€)	R2 score test	R2 score apprent.
Échantillon 1 : sociétés vie et mixte	749 875	518	90,86%	93,83%
Échantillon 2 : sociétés vie	128 561	223	95,35%	96,33%
Échantillon 3 : sociétés mixte	819 601	457	92,97%	94,40%
Échantillon 4 : sociétés majorité PB	682 741	519	96,58%	95,26%

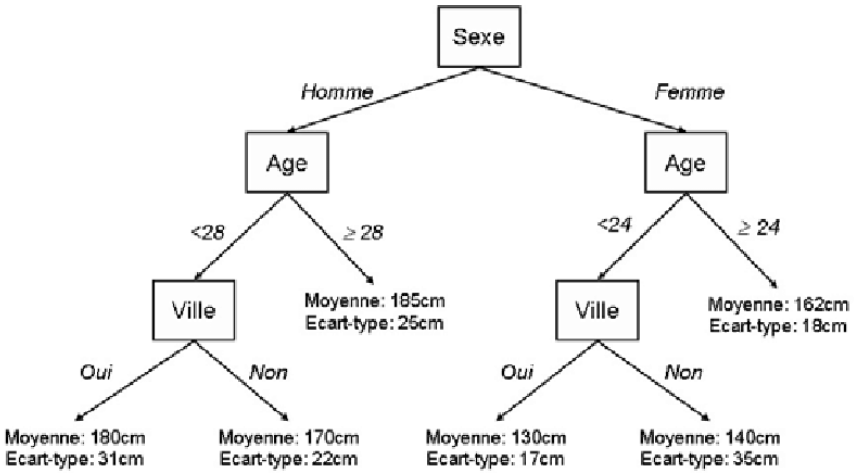
Dans le cadre de l'étude, il est à noter que nous n'avons pas retenu de modèle pénalisé (Ridge, Lasso ou Elastic Net) car la performance du modèle de régression linéaire multiple n'est pas apparue suffisante pour tenter de l'optimiser. Nous avons retenu à la place l'option de l'apprentissage supervisé qui est décrite dans les prochaines sections.

3.2.2. Arbre de décision

3.2.2.1. Fonctionnement

Les arbres de décision sont des algorithmes d'apprentissage supervisé non paramétriques capables d'adresser les situations de régression ou de classification. À titre d'illustration, on peut représenter l'arbre de régression suivant qui vise à déterminer la taille moyenne des individus en fonction de différentes variables explicatives que sont le sexe, l'âge et le lieu de résidence :

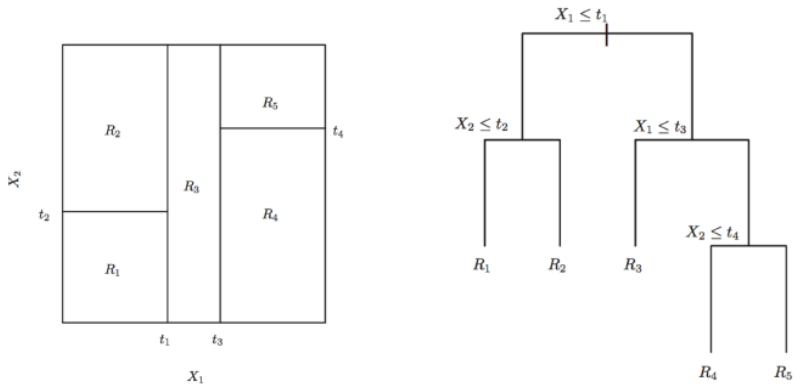
FIGURE N°13 : ILLUSTRATION DU FONCTIONNEMENT D'UN ARBRE DE REGRESSION



Ce modèle permet, comme dans le cadre général des modèles d'analyse statistique supervisé décrits plus haut, d'expliciter la valeur d'une variable de sortie Y sur la base de p variables d'entrée $X_1, \dots, X_p$. Dans notre cas, Y (représentant le montant des fonds propres de base) est une valeur continue, c'est donc un modèle d'arbre de régression qui sera utilisé.

Le principe de l'algorithme CART consiste à trouver une partition de l'espace des variables explicatives afin de séparer les observations en différents lots de dimension d. On obtient un « arbre » au travers du partitionnement récursif binaire de $\mathbb{R}^d$ en hyperrectangles. Chaque partition donne lieu à la création d'un « nœud » qui a exactement zéro ou deux développements. Le nœud terminal qui n'a plus de développement est dénommé « feuille ». La règle de partitionnement consiste à réaliser un vote de majorité qui permet de décider le classement dans l'une ou l'autre des parties répondant à l'interrogation « $X_j \geq \alpha$? » pour une coordonnée j et une partition α. Dans le cas d'une régression, c'est une simple moyenne dans chaque « feuille » qui va fournir la prédiction.

FIGURE N°14 : ILLUSTRATION DES ETAPES DE CONSTRUCTION D'UN ARBRE DECISIONNEL



La construction d'un arbre est heuristique puisqu'elle consiste à trouver deux régions qui sont le plus homogène possible, et n'intègre pas de regret sur la stratégie de choix de partition. La notion d'homogénéité en régression signifie que les observations sont concentrées autour de leur moyenne dans un nœud. Cette notion d'homogénéité est quantifiée par la variance au sein du nœud.

Sur le plan pratique : on souhaite partitionner un nœud Nentre un nœud gauche $N_L$et un nœud droit N_R. Le développement dépend d'un couple caractéristique/seuil noté (j, t) défini comme suit :

$$N_L(j,t) = \{x \in N : x_j < t\} \text{ et } N_R(j,t) = \{x \in N : x_j \geq t\}$$

Trouver la meilleure partition (j, t) consiste à comparer « l'impureté » de N avec celles de $N_L(j,t)$ et de $N_R(j,t)$ pour toutes les paires (j, t), et utiliser le gain d'information apporté par chaque paire (j, t).

L'impureté en régression est déterminée par la variance du nœud :

$$V(N) = \sum_{i:x_i \in N}(y_i - \bar{y}_N)^2 \quad \text{où} \quad \bar{y}_N = \frac{1}{|N|}\sum_{i:x_i \in N} y_i$$

$$\text{Avec } |N| = \#\{i: x_i \in N\}$$

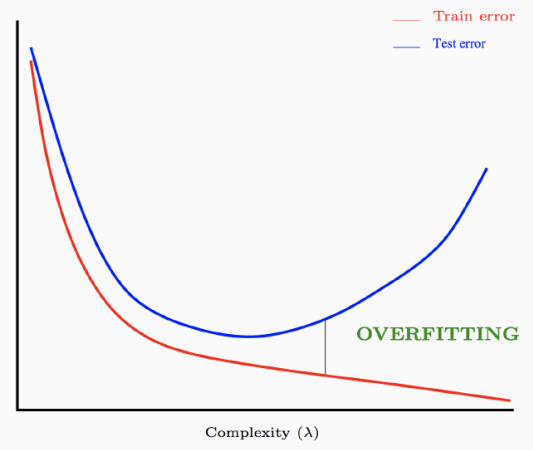
Le gain d'information est alors donné par :

$$IG(j,t) = V(N) - \frac{|N_L(j,t)|}{|N|}V\big(N_L(j,t)\big) - \frac{|N_R(j,t)|}{|N|}V\big(N_R(j,t)\big)$$

L'arbre est ainsi construit de manière récursive par la détermination du meilleur couple caractéristique/seuil (j, t) qui maximise $IG(j,t)$, ce qui conduit à la création de 2 feuilles jusqu'à ce que le critère d'arrêt soit atteint. Ce critère d'arrêt peut être la profondeur maximale d'un arbre ou l'augmentation de l'erreur de test par exemple.

Les algorithmes CART ont une tendance au « surapprentissage ». Le surapprentissage provient d'une complexité importante du modèle qui peut être très profond, et donc maximiser sa capacité d'apprentissage au détriment de sa capacité de prédiction sur un nouveau jeu de données.

FIGURE N°15 : ILLUSTRATION DU SURAPPRENTISSAGE D'UN ALGORITHME EN FONCTION DE SA COMPLEXITE



Les techniques de pénalisation permettent de répondre à la problématique de surapprentissage. Sans rentrer dans le détail, ces techniques permettent de limiter le nombre de paramètres d'un modèle afin d'en réduire la complexité, et donc la tendance au surapprentissage.

3.2.2.2. Erreur de prédiction

Nous appliquons l'algorithme d'arbre de décision aux données de nos 4 échantillons, en retenant à chaque fois les variables explicatives que nous avons défini suite à l'analyse des corrélations entre les variables d'entrée et les fonds propres de base.

Comme illustré dans les figures 16.a à 16.h, la capacité de prédiction du modèle apparait globalement meilleure que dans le modèle précédent de régression linéaire multiple. Comme précédemment, les échantillons 2 et 4 présentent un niveau d'erreur assez faible, et une distribution de ces erreurs très centrée autour de 0 pour l'échantillon 4, alors que certaines observations sont plus difficilement prédites de manière adéquate dans l'échantillon 2.

Concernant les échantillons 1 et 3, même si la précision s'améliore comparée à la régression linéaire multiple, l'algorithme présente quelques difficultés d'interprétation de certaines valeurs, qui se situent très loin de la valeur attendue. Ceci est d'autant plus marqué que l'échantillon est hétérogène. Ainsi la distribution de l'erreur de l'échantillon 1 présente une variance bien supérieure à celui de l'échantillon 3.

Figure 16.a : résultat de la prédiction par le modèle d'arbre décisionnel sur l'échantillon 1 des sociétés vie et mixte

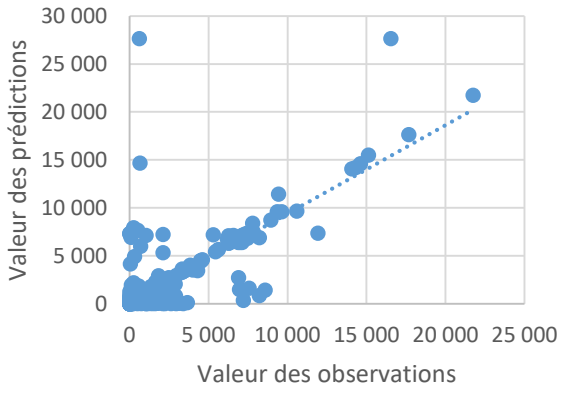


Figure 16.b : distribution des erreurs du modèle arbre décisionnel sur l'échantillon 1 des sociétés vie et mixte

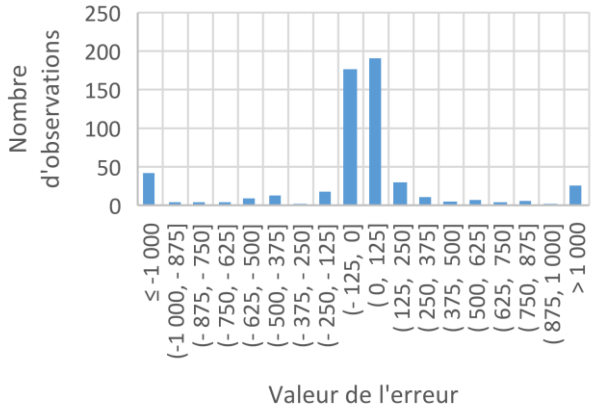


Figure 16.c : résultat de la prédiction par le modèle d'arbre décisionnel sur l'échantillon 2 des sociétés vie

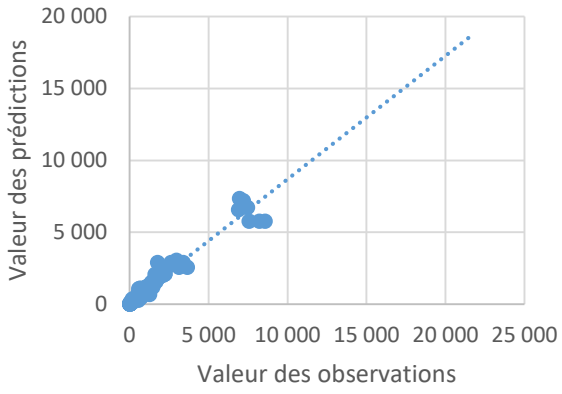


Figure 16.d : distribution des erreurs du modèle arbre décisionnel sur l'échantillon 2 des sociétés vie

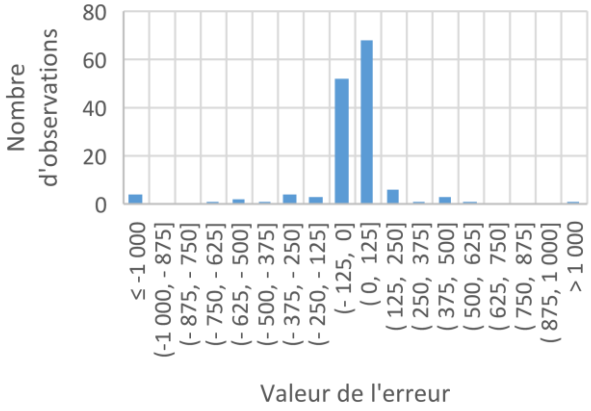


Figure 16.e : résultat de la prédiction par le modèle d'arbre décisionnel sur l'échantillon 3 des sociétés mixte

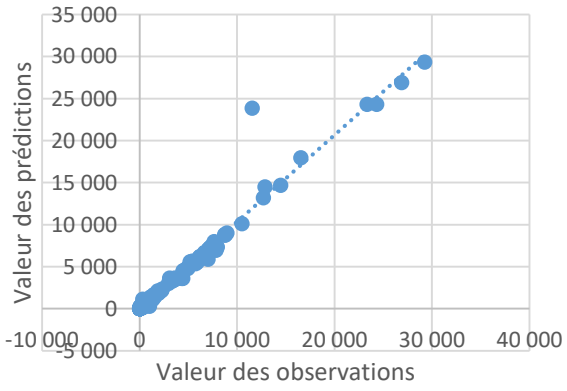


Figure 16.f : distribution des erreurs du modèle arbre décisionnel sur l'échantillon 3 des sociétés mixte

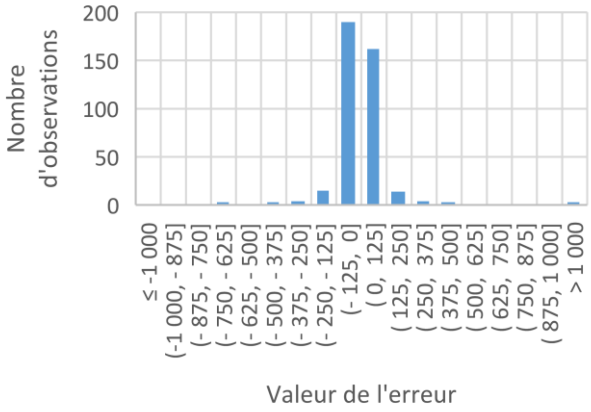


Figure 16.g : résultat de la prédiction par le modèle d'arbre décisionnel sur l'échantillon 4 des sociétés à majorité de PB

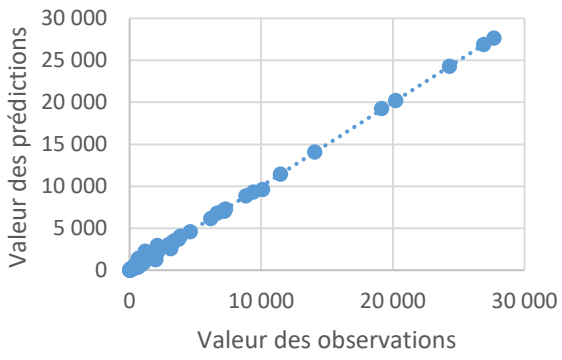
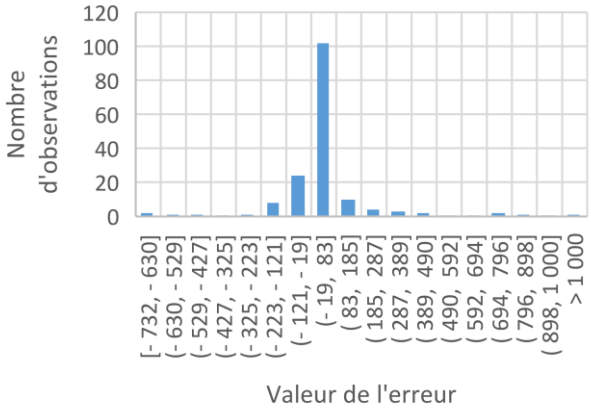


Figure 16.h : distribution des erreurs du modèle arbre décisionnel sur l'échantillon 4 des sociétés à majorité de PB



3.2.2.3. Performance du modèle

Comme illustré dans le tableau 8, nous observons que l'arbre de décision présente une capacité de prédiction assez forte. Le score R^2 de la base de test monte à 99,84% sur l'échantillon 4 des sociétés avec une majorité de PB, ce qui illustre la grande homogénéité de ce groupe qui permet une prévision bien meilleure que sur les autres échantillons. L'analyse du niveau de l'erreur quadratique et absolue confirme cette interprétation.

Sur les autres échantillons, la performance illustrée par le score R^2 de la base de test, apparait plus limitée, probablement en raison de la plus grande hétérogénéité de ces groupes. Enfin on observe le niveau très élevé du score R^2 de la base d'apprentissage (a minima 99,99%) qui signale avec force la tendance au surapprentissage du modèle d'arbre décisionnel. Ce constat nous oblige à rechercher d'autres modèles dont la capacité de prédiction peut-être mieux décorellée des données d'apprentissage, et qui vont probablement permettre une meilleure prédiction sur les échantillons 1, 2 et 3 plus hétérogènes.

TABLEAU 8 : INDICATEURS DE PERFORMANCE DU MODELE D'ARBRE DE DECISION

Arbre de décision	MSE (en M€²)	MAE (en M€)	R2 score test	R2 score apprent.
Échantillon 1 : sociétés vie et mixte	305 648	101	96,28%	99,99%
Échantillon 2 : sociétés vie	138 081	133	95,01%	100,00%
Échantillon 3 : sociétés mixte	248 087	93	97,87%	99,99%

Échantillon 4 : sociétés majorité PB	30 389	79	99,84%	100,00%

3.2.3. Forêt aléatoire

3.2.3.1. Fonctionnement

La forêt aléatoire est une méthode ensembliste. Elle utilise les arbres de régression et de classification (CART) dénommés « apprenants faibles » ainsi qu'une « combinaison » (bagging) de tels arbres afin de produire un « apprenant fort ». Chaque arbre est entrainé sur la base d'un échantillon avec remise de la base d'apprentissage, selon une méthode de validation croisée. Les apprenants faibles doivent être indépendants. Au lieu d'utiliser un unique arbre, le modèle permet d'utiliser une moyenne des prédictions fournies par plusieurs arbres. Cela permet de réduire le bruit et la variance d'un arbre unique. L'utilisation de « features bagging » permet en outre de réduire la corrélation entre les arbres en recourant à un sous échantillon aléatoire des variables explicatives pour déterminer les partitions. Ceci fonctionne efficacement lorsque certaines variables explicatives sont particulièrement importantes dans la qualité de la prédiction.

Les forêts aléatoires permettent un gain en précision très notable par rapport aux arbres décisionnels. Ils autorisent également l'utilisation d'échantillon de plus petites tailles et de variables explicatives plus nombreuses.

3.2.3.2. Erreur de prédiction

Nous avons appliqué l'algorithme de forêt aléatoire aux données de nos 4 échantillons. On observe dans les figures 17.a à 17.h que la qualité de la prédiction tend à s'améliorer par rapport à l'algorithme composé d'un unique arbre de décision, mais pas de manière uniforme sur les 4 échantillons.

En particulier, les prédictions de l'échantillon 1 se resserrent autour de la vraie valeur de manière assez notable, mais certaines valeurs restent difficiles à prédire pour le modèle, en témoigne les valeurs extrêmes toujours présentes dans la distribution des erreurs. L'échantillon 2 est encore mieux prédit par ce nouvel algorithme, qui réduit de manière significative les écarts sur certaines observations. Quant aux échantillons 3 et 4, le gain n'est pas forcément évident, et il faudra vérifier au moyen des indicateurs de qualité des modèles dans la prochaine section.

Au global, on observe toujours, comme pour l'algorithme précédent, une très forte capacité du modèle de forêt aléatoire à prédire efficacement les données des échantillons 2 et 4, qui présentent vraisemblablement un fonctionnement plus homogène que les échantillons 1 et 3.

Figure 17.a : résultat de la prédiction par le modèle de forêt aléatoire sur l'échantillon 1 des sociétés vie et mixte

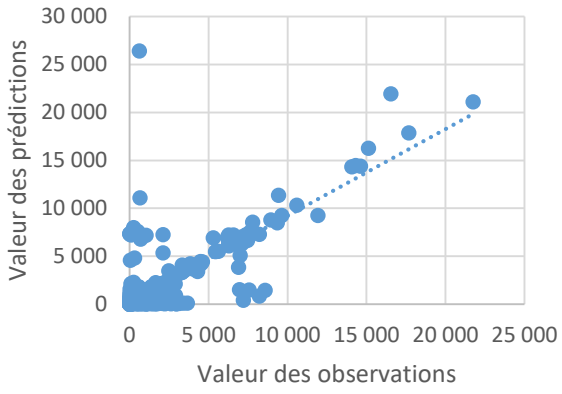


Figure 17.b : distribution des erreurs du modèle forêt aléatoire sur l'échantillon 1 des sociétés vie et mixte

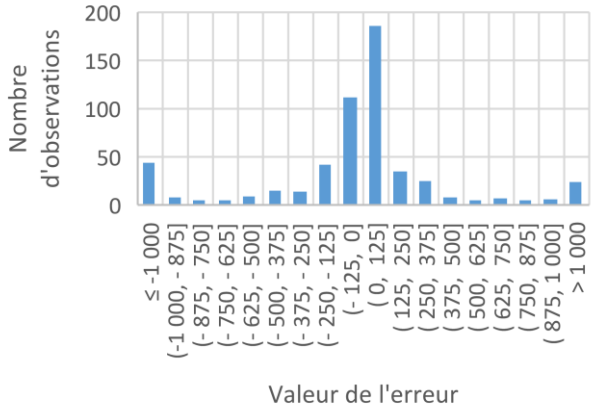


Figure 17.c : résultat de la prédiction par le modèle de forêt aléatoire sur l'échantillon 2 des sociétés vie

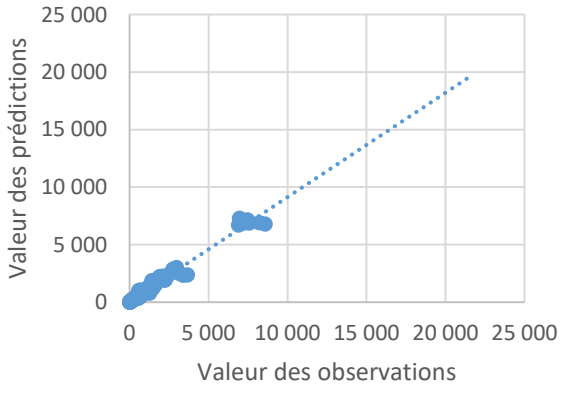


Figure 17.d : distribution des erreurs du modèle forêt aléatoire sur l'échantillon 2 des sociétés vie

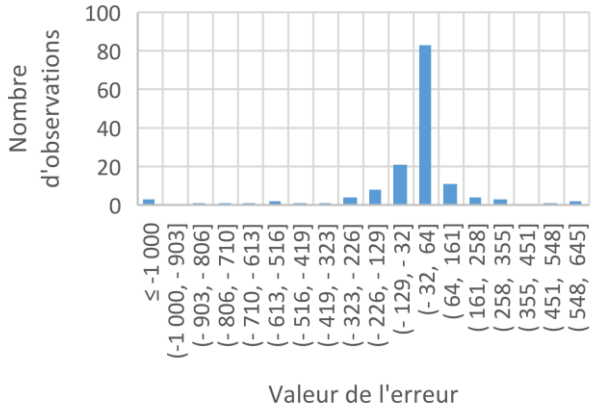


Figure 17.e : résultat de la prédiction par le modèle de forêt aléatoire sur l'échantillon 3 des sociétés mixte

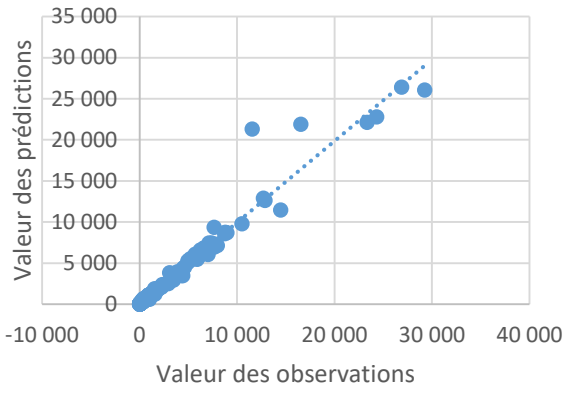


Figure 17.f : distribution des erreurs du modèle forêt aléatoire sur l'échantillon 3 des sociétés mixte

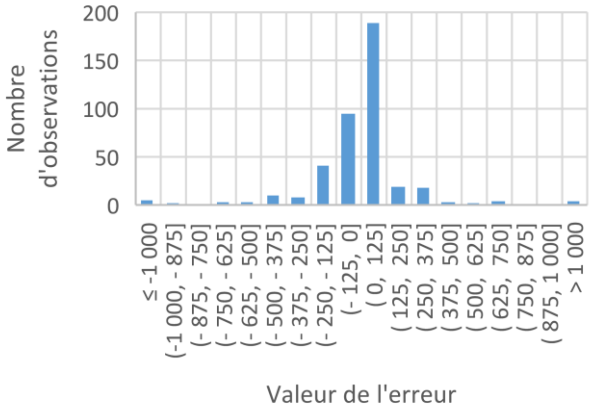


Figure 17.g : résultat de la prédiction par le modèle de forêt aléatoire sur l'échantillon 4 des sociétés à majorité de PB

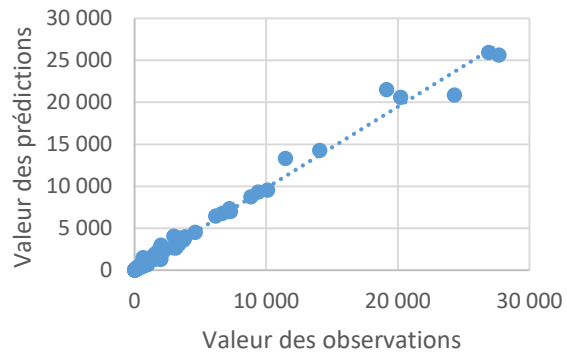
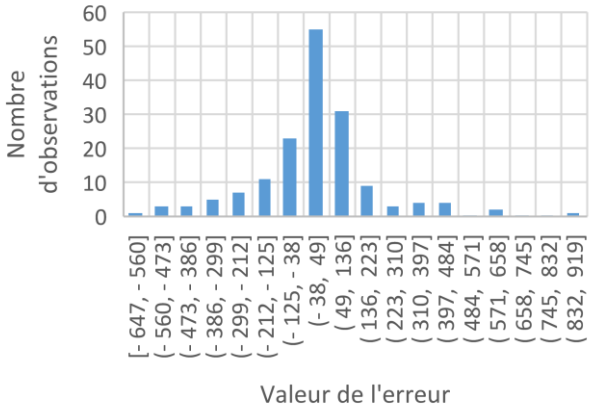


Figure 17.h : distribution des erreurs du modèle forêt aléatoire sur l'échantillon 4 des sociétés à majorité de PB



3.2.3.3. Performance du modèle

L'utilisation de forêts aléatoires nous permet d'améliorer notablement la qualité de la prévision des échantillons 1, 2. Le fonctionnement théorique de l'algorithme qui, en combinant des prévisions de « weak learners », est censé mieux apprendre, est vérifié pour ces 2 échantillons. Cependant la qualité de prédiction reste globalement très satisfaisante sur les 4 échantillons.

En particulier, si on s'intéresse au score R^2 sur la base de test, on observe que la prévision de l'échantillon 4, a priori le plus homogène, est toujours la meilleure (99,21%) parmi les 4 échantillons. On note que c'est l'échantillon le plus grand et le plus hétérogène qui bénéficie de l'apport de précision de ce nouvel algorithme, alors que les autres, dont la qualité de prévision était déjà bonne avec l'arbre de décision, en profite moins, comme illustré dans le tableau 9.

On observe enfin que le score R^2 sur la base d'apprentissage se rapproche plus de celui de la base de test. Ainsi l'algorithme présente une tendance au surapprentissage bien moindre que le modèle d'arbre décisionnel que nous avons étudié précédemment.

TABLEAU 9 : INDICATEURS DE PERFORMANCE DU MODELE DE FORET ALEATOIRE

Forêt aléatoire	MSE (en M€²)	MAE (en M€)	R2 score test	R2 score apprent.
Échantillon 1 : sociétés vie et mixte	115 104	108	98,59%	99,73%
Échantillon 2 : sociétés vie	77 376	112	97,20%	99,79%
Échantillon 3 : sociétés mixte	380 550	121	96,73%	99,77%
Échantillon 4 : sociétés majorité PB	156 575	169	99,21%	99,62%

3.2.4. Gradient boosting

3.2.4.1. Fonctionnement

Le boosting est une méthode ensembliste qui consiste à entrainer les algorithmes de manière séquentielle. À la différence des forêts aléatoires qui utilisent la technique du « bagging » consistant à tirer aléatoirement un échantillon avec remise et à entrainer à chaque fois un arbre sur ce tirage, puis de faire la moyenne pour obtenir la prédiction, le « boosting » consiste à réaliser un apprentissage séquentiel en modifiant les poids des échantillons d'apprentissage, ce qui permet au modèle de mieux tenir compte des observations difficiles à appréhender pour le « bagging ». Les algorithmes ne sont donc plus indépendants. Le modèle final correspond à une somme pondérée des sous-modèles qui composent l'algorithme.

Le gradient boosting est l'une des méthodes les plus répandue de boosting. Elle permet l'utilisation de différents types d'apprenants faibles et de différentes fonctions de perte. Les apprenants faibles sont entraînés de façon à corriger les erreurs des apprenants faibles précédents. A la fin, tous les apprenants faibles ont le même poids dans le système de vote.

La technique consiste à initier un premier apprenant faible simple qui consiste en la moyenne des observations. Ensuite on calcule l'écart entre cette moyenne et la réalité qui constitue le résidu. Le gradient boosting prédit à chaque étape les résidus. A partir des dernières prédictions, on calcule les nouveaux résidus, qui sont ensuite multipliés par un facteur inférieur à 1 de façon à avancer « pas à pas » pour écarter progressivement les prédictions de la moyenne. La prédiction de l'apprenant fort est alors la somme de celles des apprenants faibles.

3.2.4.2. Erreur de prédiction

A nouveau, malgré l'utilisation d'un algorithme plus complexe et robuste, la prédiction reste meilleure sur les échantillons les plus homogènes, à savoir l'échantillon 2 des sociétés vie et l'échantillon 4 des sociétés avec une majorité de PB.

L'apport du gradient boosting est très appréciable notamment sur l'échantillon 4 dont la distribution de l'erreur de prédiction se resserre très distinctement autour de zéro. C'est le cas aussi pour les échantillons 1 et 3 qui bénéficient de la mise en place de cet algorithme, comme on le voit dans les figures 18.a à 18.h.

Enfin la capacité de prédiction du gradient boosting sur les échantillons hétérogènes 1 et 3 reste modérée, même si, comme on le verra dans la prochaine section, la performance de ce modèle reste tout à fait satisfaisante.

Figure 18.a : résultat de la prédiction par le modèle de gradient boosting sur l'échantillon 1 des sociétés vie et mixte

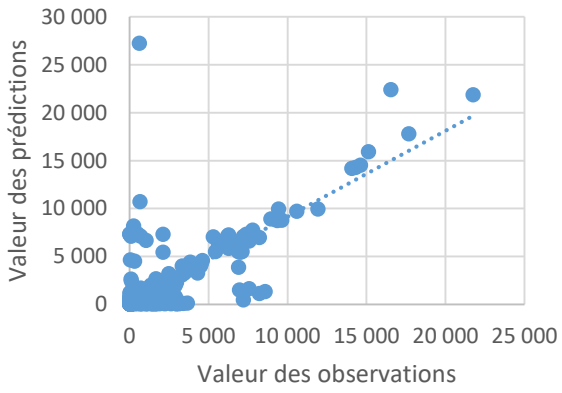


Figure 18.b : distribution des erreurs du modèle gradient boosting sur l'échantillon 1 des sociétés vie et mixte

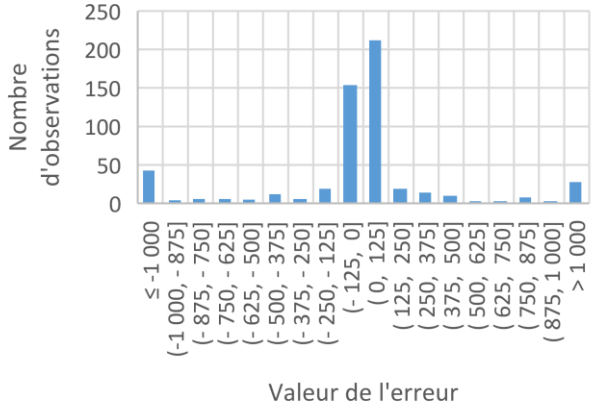


Figure 18.c : résultat de la prédiction par le modèle de gradient boosting sur l'échantillon 2 des sociétés vie

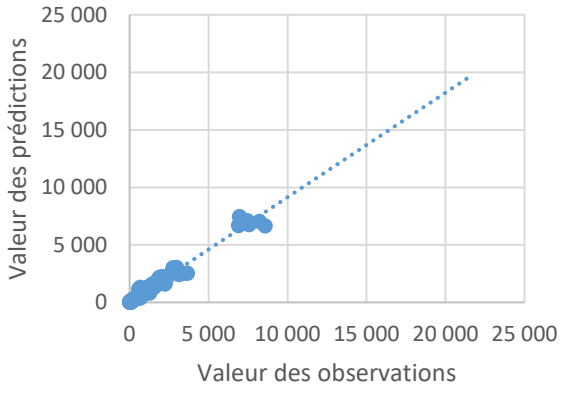


Figure 18.d : distribution des erreurs du modèle gradient boosting sur l'échantillon 2 des sociétés vie

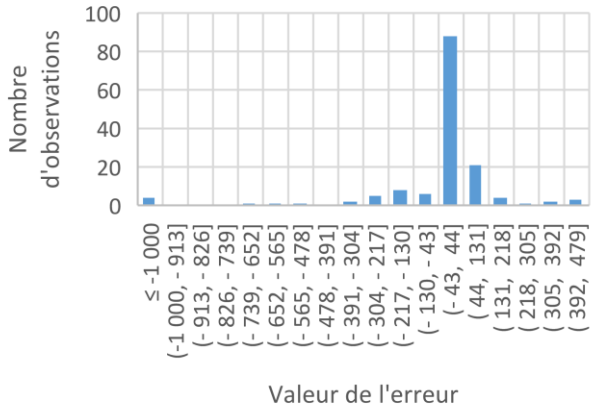


Figure 18.e : résultat de la prédiction par le modèle de gradient boosting sur l'échantillon 3 des sociétés mixte

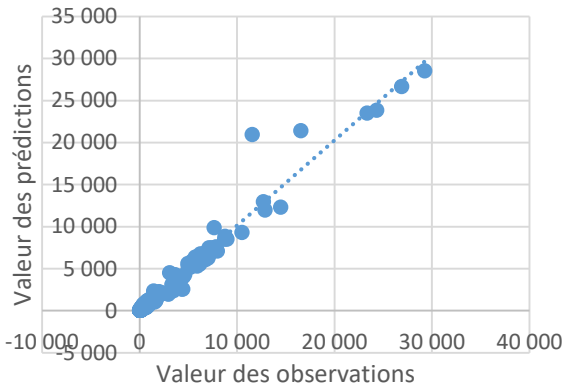


Figure 18.f : distribution des erreurs du modèle gradient boosting sur l'échantillon 3 des sociétés mixte

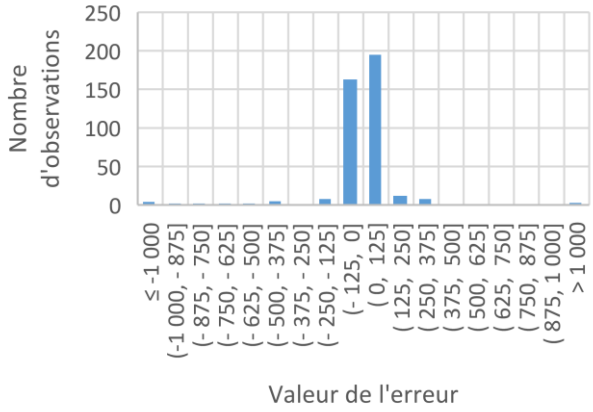


Figure 18.g : résultat de la prédiction par le modèle de gradient boosting sur l'échantillon 4 des sociétés à majorité de PB

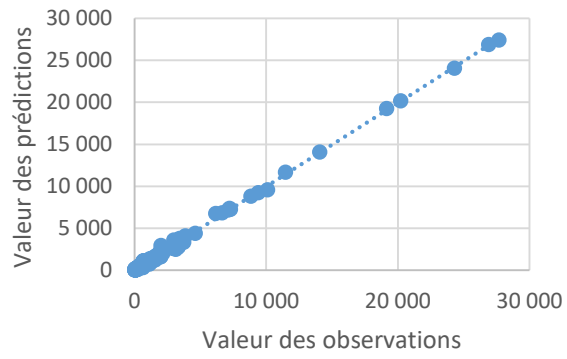
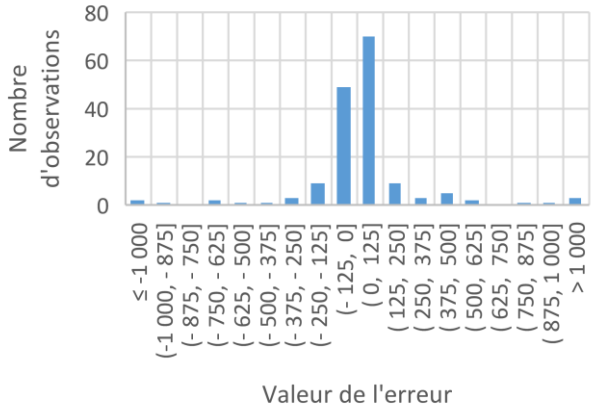


Figure 18.h : distribution des erreurs du modèle gradient boosting sur l'échantillon 4 des sociétés à majorité de PB



3.2.4.3. Performance du modèle

L'utilisation d'algorithme de gradient boosting doit, en théorie, permettre d'améliorer la prévision observée avec la forêt aléatoire, en permettant notamment de mieux tenir en compte des valeurs extrêmes et des apprenants faibles.

Comme illustré dans le tableau 10, on observe que, en effet, l'algorithme permet une amélioration de la qualité de la prévision dans certains cas. À nouveau, la prédiction sur l'échantillon 4 est la plus performante, quel que soit l'indicateur (score R^2 sur la base de test de 99,78%).

La qualité de la prédiction sur les échantillons 1 et 3 n'est pas réellement impacté par la mise en place de ce nouvel algorithme, dont les indicateurs restent relativement stables entre les 2 machines.

L'apport du gradient boosting concerne plus l'échantillon 4 qui voit son score R^2 amélioré. Étant donné la taille plus réduite de cet échantillon, ce dernier bénéficie davantage de l'apport de capacité d'apprentissage des apprenants faibles que les échantillons 1 et 3.

TABLEAU 10 : INDICATEURS DE PERFORMANCE DU MODELE DE GRADIENT BOOSTING

Gradient boosting	MSE (en M€²)	MAE (en M€)	R2 score test	R2 score apprent.
Échantillon 1 : sociétés vie et mixte	139 475	174	98,30%	99,55%
Échantillon 2 : sociétés vie	77 346	126	97,20%	99,78%
Échantillon 3 : sociétés mixte	371 580	182	96,81%	99,71%
Échantillon 4 : sociétés majorité PB	42 940	137	99,78%	99,85%

3.3. Comparaison de la performance des modèles

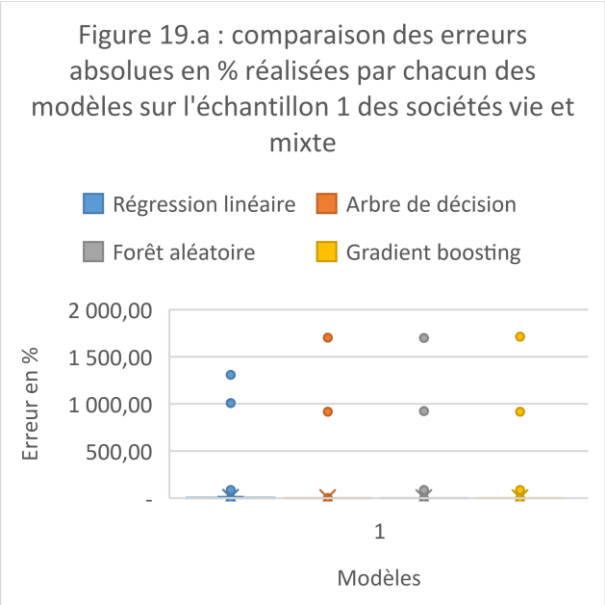
3.3.1. Comparaison des modèles

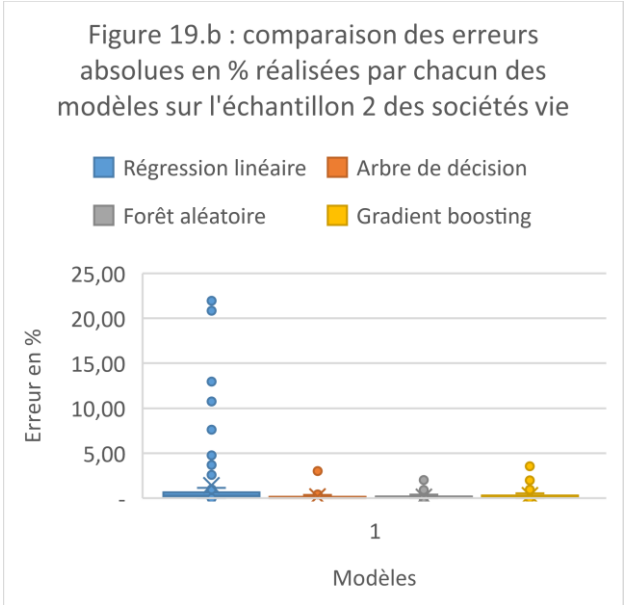
Afin de comparer la performance des algorithmes sur les différents échantillons, on réalise dans un premier temps une analyse comparée des erreurs absolues en pourcentage réalisées par les modèles. Les résultats sont donnés dans les figures 19.a à 19.d.

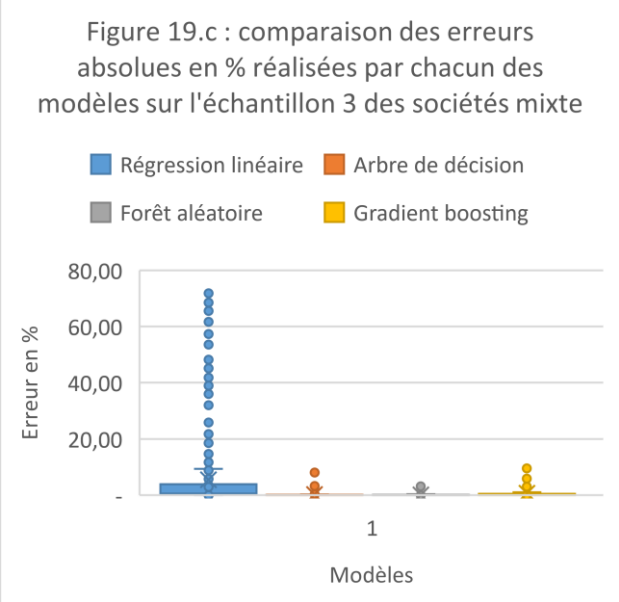
3.3.1.1. Distribution des erreurs absolues en pourcentage

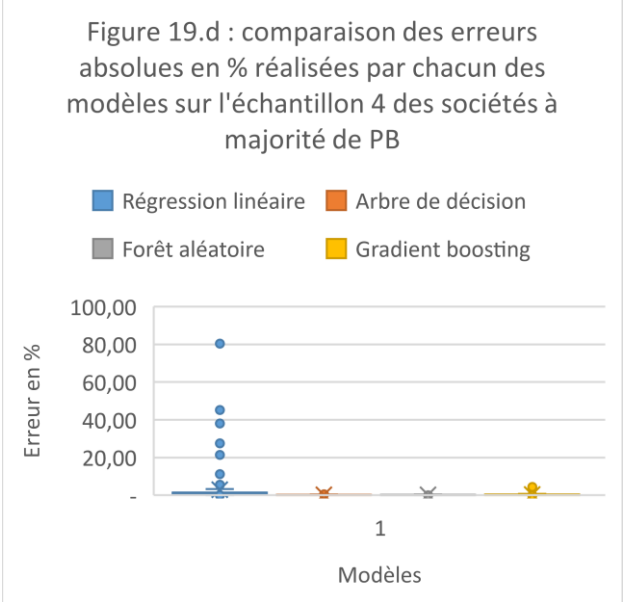
L'observation des résultats par échantillon montre que :

- Quel que soit l'algorithme utilisé, l'échantillon 1 des sociétés vie et mixte est celui pour lequel les prévisions sont les moins justes, notamment en raison d'écarts importants sur de nombreuses prévisions, qui se trompent parfois de plus de 1 000% ;
- L'échantillon 2, davantage homogène, présente des erreurs de prévision en pourcentage très limitées, moins de 5% pour les algorithmes non paramétriques ;
- L'échantillon 3 présente de bons résultats, la moyenne des écarts absolus étant très basse avec les algorithmes non paramétriques, mais certaines valeurs présentent des écarts de l'ordre de 10%, ce qui pénalise la qualité générale de la prédiction ;
- L'échantillon 4 présente davantage de stabilité entre les différents algorithmes non paramétriques, et les erreurs absolues en pourcentage sont très contenues. À nouveau, cet échantillon semble être celui le mieux appréhendé par les machines.



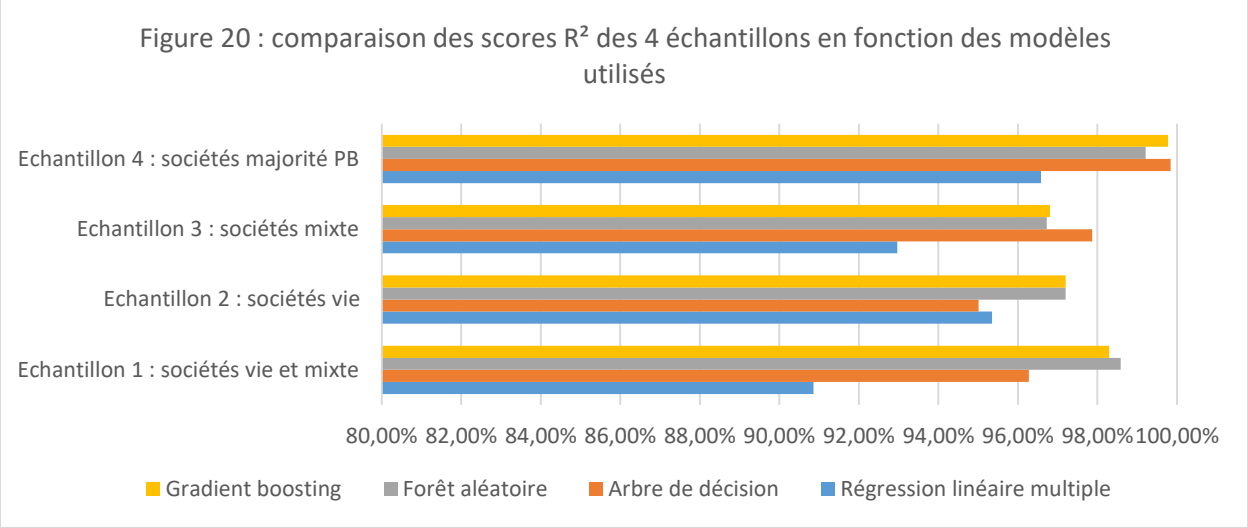




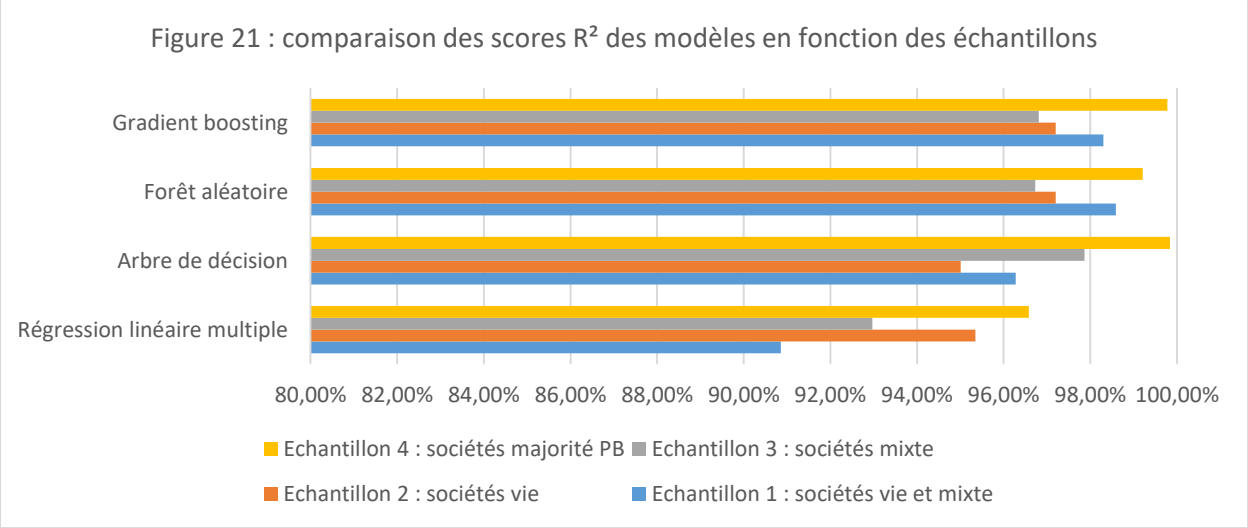


3.3.1.2. Performance basée sur le score R^2

L'analyse par échantillon présentée dans la figure 20 confirme que l'échantillon 4 des sociétés avec une majorité de PB est l'échantillon dont la qualité de prédiction est la meilleure, quelle que soit l'algorithme utilisé. Cependant, contrairement à ce que la distribution des erreurs absolues en pourcentage pouvait laisser penser, la qualité de la prédiction sur l'échantillon 1, notamment des algorithmes non paramétriques, est plutôt bonne (supérieure à 98% pour la forêt aléatoire et le gradient boosting). On perçoit ici l'apport de ces algorithmes sur les apprenants faibles et les observations erratiques qui ont tendance à être mieux interprétées par ces algorithmes.



L'analyse par modèle présentée dans la figure 21 confirme quant à elle la supériorité du modèle de gradient boosting sur la prédiction de la quasi-totalité des échantillons (hormis la forêt aléatoire qui donne un meilleur résultat sur l'échantillon 1, mais de très peu, et l'arbre de décision sur l'échantillon 4, mais à nouveau de très peu).



En conclusion, on retient comme modèle de prévision le modèle de gradient boosting qui présente en moyenne sur les 4 échantillons les meilleures qualités de prédiction, et une tendance limitée au surapprentissage. On note cependant la bonne performance générale des modèles non paramétriques, et la performance nettement plus réduite du modèle de régression linéaire multiple.

4. Chapitre 4 : étude de sensibilité

Après avoir sélectionné le modèle le plus efficace (ie de gradient boosting) pour prédire la valeur des fonds propres économiques de nos 4 échantillons, nous nous intéressons désormais à l'influence des variables constitutives des prévisions sur la prévision elle-même. Cette cartographie de l'importance des variables va nous permettre de comprendre quels sont les déterminants les plus significatifs qui entrent dans la composition des fonds propres prudentiels. Nous pourrons alors utiliser cette connaissance détaillée du fonctionnement des algorithmes pour évaluer l'impact d'un choc sur une ou plusieurs variables d'entrée du modèle sur la prédiction du niveau de fonds propres des sociétés du panel.

4.1. Analyse des variables d'importance

L'utilisation d'un modèle de type gradient boosting permet de déduire de manière simple et directe l'importance relative de chacune des variables dans la prédiction. Le principe consiste à établir un score qui correspond à l'utilisation de la variable dans la prédiction. Plus celle-ci est utilisée, plus le score est important. L'importance est calculée pour chaque variable de la base de données, ce qui permet de classer et de comparer la performance de chacune des variables. L'importance est alors estimée par la moyenne de tous les arbres de décision utilisés par le modèle.

La mesure du score $\widehat{I_j^2}$ correspond au nombre de fois où la variable est utilisée pour une décision, affecté d'un poids qui correspond au carré du gain d'information résultant de la décision à chaque nœud t précédent le nœud terminal J de l'arbre T, et moyenné sur l'ensemble des arbres, ce qui s'écrit :

$$\widehat{I_j^2}(T) = \sum_{t=1}^{J} \widehat{\imath_t^2} 1(v_t = j)$$

Avec v_t la variable de décision associée au nœud t, et $\widehat{\imath_t^2}$ correspondant au gain d'information résultant de l'amélioration empirique de l'erreur quadratique issue de la décision telle que $i^2(R_, ; R_r) = \frac{w_l w_r}{w_l + w_r}(\overline{y_l} - \overline{y_r})^2$ où $\overline{y_l}, \overline{y_r}$ sont les moyennes des décisions gauche et droite et w_l, w_r les poids correspondants.

4.1.1. Comparaison des variables d'importance des 4 échantillons

L'analyse des figures 22.a à 22.d montre que les variables d'importance varient d'un échantillon à l'autre. On observe cependant une constante majeure, c'est le rôle prépondérant des obligations d'État dans la prévision, quel que soit l'échantillon. Ensuite on retrouve, dans des proportions qui varient significativement entre les échantillons, le rôle important des passifs subordonnés, des actions cotées (pour les échantillons 1, 3 et 4), des OPC (également pour les échantillons 1, 3 et 4), et dans une moindre mesure des détentions dans les entreprises liées (pour tous les échantillons), des passifs d'impôts différés (également pour tous les organismes).

Les différences majeures entre les échantillons, au-delà de l'importance relative des variables précédentes qui varient dans la prédiction, est essentiellement liée à l'importance des provisions techniques vie (hors santé et UC) pour les échantillons 2 et 4. Dans ces échantillons, les provisions techniques vie sont le 2ème facteur d'importance de la prédiction, juste après les obligations d'État. De même, les obligations d'entreprise sont plus prépondérantes pour l'échantillon 2 que pour les autres échantillons.

Enfin on observe que les variables d'importance sont peu nombreuses : 5 à 6 variables maximum suivant les échantillons sont nécessaires pour réunir plus de 90% de l'importance dans la qualité de la prédiction.

Figure 22.a : score d'importance des variables du modèle de gradient boosting sur l'échantillon 1 des sociétés vie et mixte

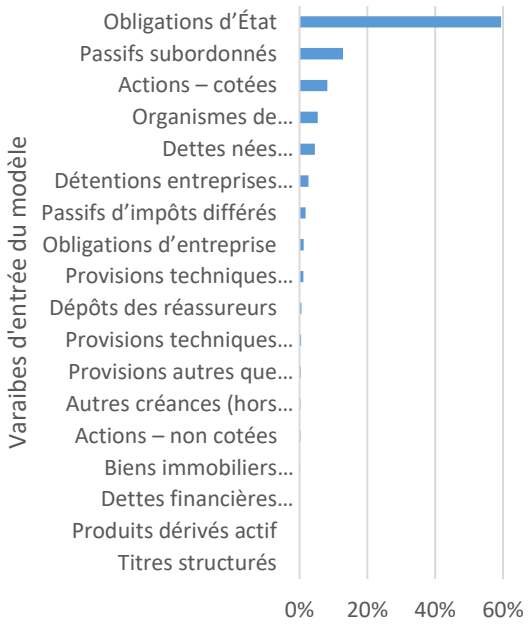


Figure 22.b : score d'importance des variables du modèle de gradient boosting sur l'échantillon 2 des sociétés vie

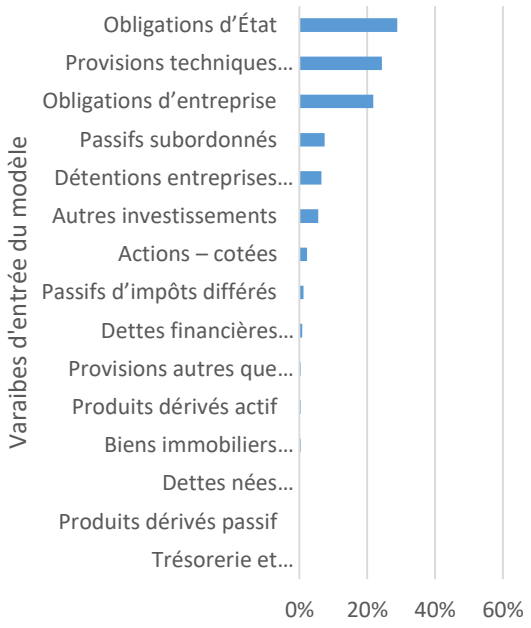


Figure 22.c : score d'importance des variables du modèle de gradient boosting sur l'échantillon 3 des sociétés mixte

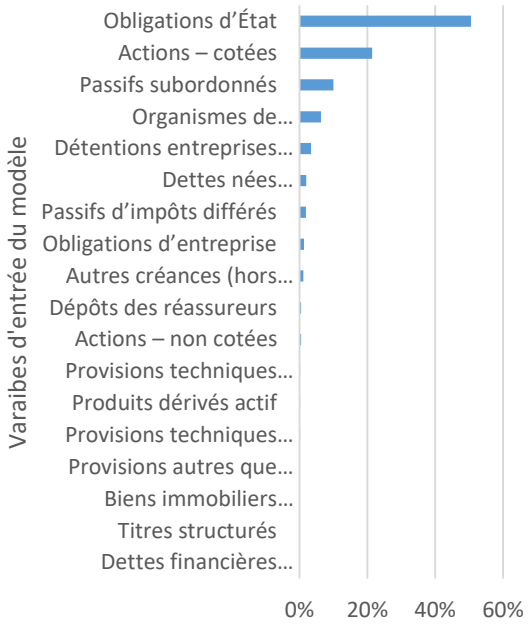
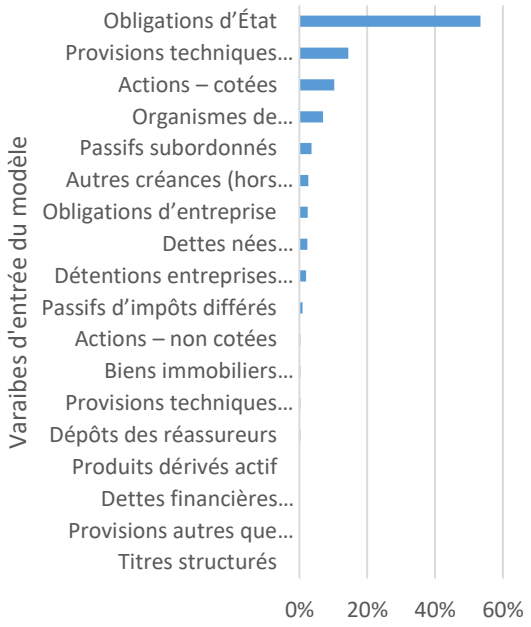


Figure 22.d : score d'importance des variables du modèle de gradient boosting sur l'échantillon 4 des sociétés avec majorité de PB



En synthèse, on observe que les variables suivantes sont les variables d'importance majeures des 4 échantillons de l'étude, avec leurs rangs associés dans le tableau 11 suivant :

TABLEAU 11 : CLASSEMENT DES VARIABLES D'IMPORTANCE PAR RANG ET PAR ECHANTILLON

Variable d'importance	Echantillon 1 (toutes sociétés)	Echantillon 2 (sociétés vie)	Echantillon 3 (sociétés mixte)	Echantillon 4 (majorité PB)
Obligations d'État	1	1	1	1
Passifs subordonnés	2	4	3	5
Provisions techniques vie	9	2	12	2
Actions – cotées	3	7	2	3

Organismes de placement collectif	4	N/A	4	4
Dettes nées d'opérations d'assu.	5	13	6	8
Détentions ent liées, yc particip.	6	5	5	9
Passifs d'impôts différés	7	8	7	10
Obligations d'entreprise	8	3	8	7
Autres investissements	N/A	6	N/A	N/A

4.1.2. Comparaison des variables d'importance avec les corrélations de Pearson et le poids des variables au bilan

Afin d'apprécier les apports de l'analyse par notre algorithme non paramétrique, on compare dans les figures 23 et 24 les variables d'importance de chaque échantillon avec leur poids moyen dans le bilan et avec la valeur du coefficient de corrélation de Pearson que nous avons déjà évoqué.

Le poids moyen au bilan de chacune des variables est donné par la somme des observations de la variable sur toutes les périodes, divisé par la somme de l'ensemble des postes du bilan également sur toutes les périodes. Ces poids moyens des variables varient donc d'un échantillon à l'autre puisque la composition du bilan diffère en fonction des typologies de sociétés (vie et mixte, vie, mixte ou à majorité de PB). On retrouve néanmoins, conformément à l'intuition, la part prépondérante des provisions techniques vie, des obligations d'entreprise, des obligations d'État pour la totalité des 4 échantillons dans des proportions très comparables.

Même si ces métriques (score d'importance, corrélations et poids au bilan) ne sont pas directement comparables, il parait intéressant d'observer comment l'importance relative de chacune des variables dans la composition des fonds propres prudentiels a évolué en fonction de la méthode retenue, de la plus simple (poids au bilan) à la plus complexe (algorithme gradient boosting).

4.1.2.1. Analyse comparée de l'importance des variables

Dans un 1er temps, on observe dans ces figures 23.a à 23.b plusieurs phénomènes intéressants concernant la comparaison au poids des variables au bilan prudentiel :

- L'importance des provisions techniques vie dans la prédiction est bien moins marquée que ne le laisse supposer leur poids au bilan prudentiel pour les échantillons 1 et 3.

- De la même façon, l'importance des obligations d'entreprise dans la prédiction est bien moins importante que ne le laisse supposer leur poids au bilan pour les échantillons 1 et 3.
- A contrario, les obligations d'État présentent une importance supérieure à leur poids au bilan pour l'ensemble des échantillons. Tout comme les actions cotées (sauf pour l'échantillon 2).
- L'importance des OPC est comparable à leur poids au bilan pour les échantillons 1, 3 et 4.

Figure 23.a : comparaison du score d'importance des variables du modèle de gradient boosting sur l'échantillon 1 des sociétés vie et mixte avec le poids moyen des variables au bilan prudentiel et le score de corrélation de Pearson

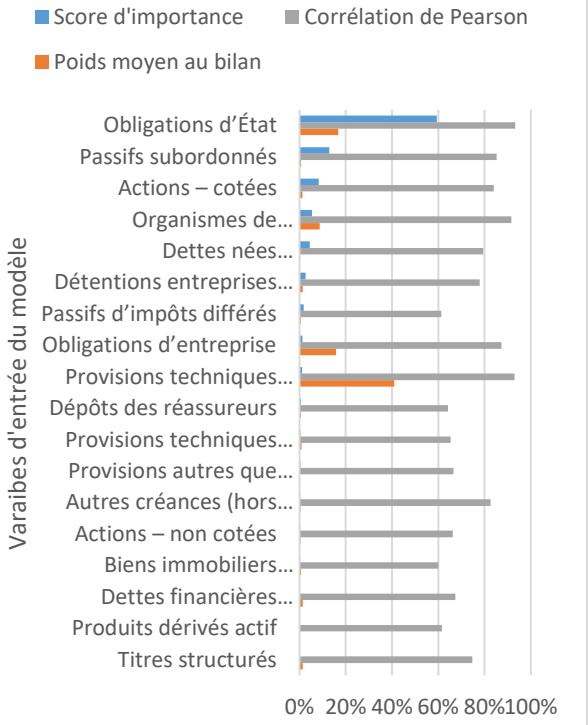
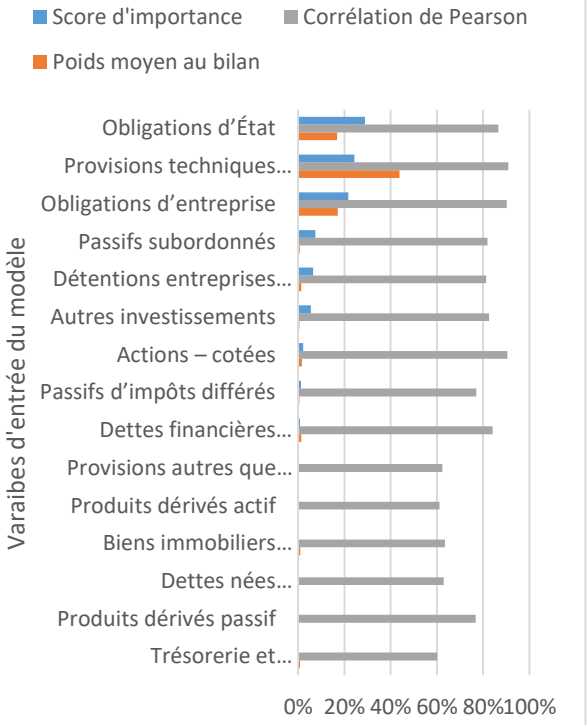


Figure 23.b : comparaison du score d'importance des variables du modèle de gradient boosting sur l'échantillon 2 des sociétés vie avec le poids moyen des variables au bilan prudentiel et le score de corrélation de Pearson



Dans un second temps, on s'intéresse à la comparaison du score d'importance avec celui des coefficients de corrélation de Pearson. On observe, toujours dans les figures 24.a à 24.b, les phénomènes notables suivants :

- L'importance des provisions techniques vie n'est pas aussi marquée que ne le laisse supposer le coefficient de corrélation pour les échantillons 1 et 3.
- De la même façon, l'importance des obligations d'entreprise est moins marquée que ne le laisse supposer leur coefficient de corrélation pour les échantillons 1 et 3.
- Les obligations d'État présentent une importance très comparable à leur corrélation pour tous les échantillons.
- Les actions cotées présentent une importance plus marquée que leur niveau de corrélation pour les échantillons 1, 3 et 4.

- On observe une plus grande cohérence entre les corrélations et l'importance pour les échantillons 2 et 4.
- Certaines variables moins corrélées présentent une importance plus marquée qu'attendu spécifiquement dans tous les échantillons : les passifs subordonnées, les actions cotées, les dettes nées d'opération d'assurance, les détentions dans les entreprises liées et les passifs d'impôt différés.

Figure 24.a : comparaison du score d'importance des variables du modèle de gradient boosting sur l'échantillon 3 des sociétés mixte avec le poids moyen des variables au bilan prudentiel et le score de corrélation de Pearson

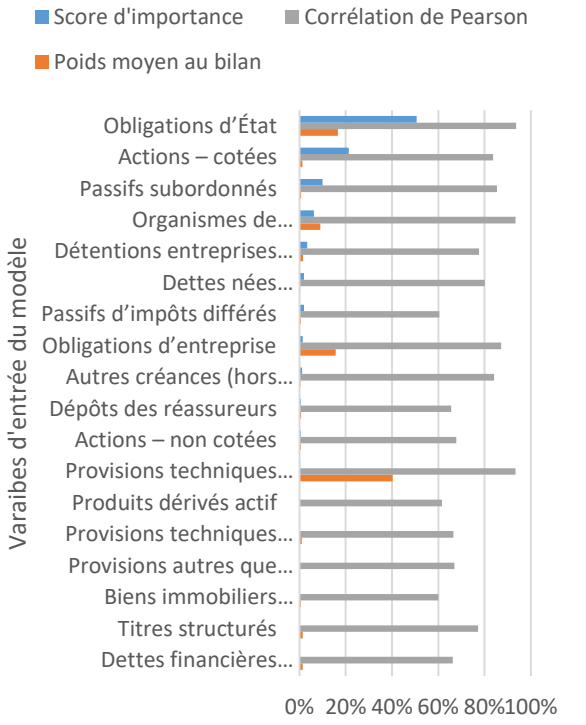
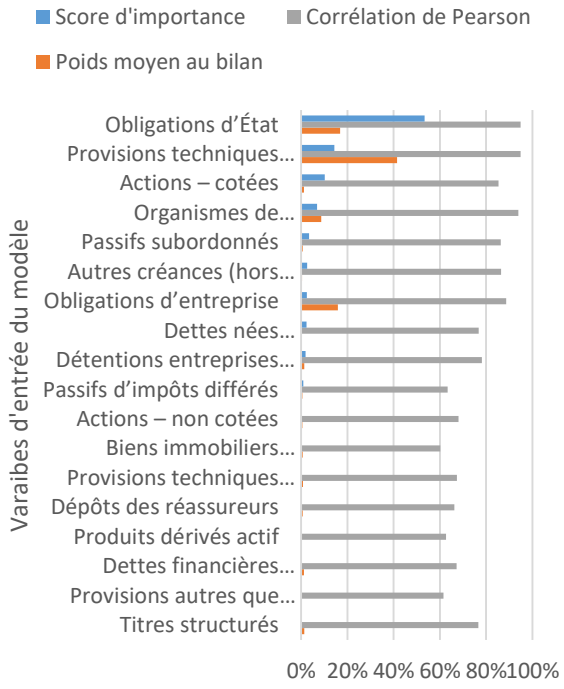


Figure 24.b : comparaison du score d'importance des variables du modèle de gradient boosting sur l'échantillon 4 des sociétés avec majorité de PB avec le poids moyen des variables au bilan prudentiel et le score de corrélation de Pearson



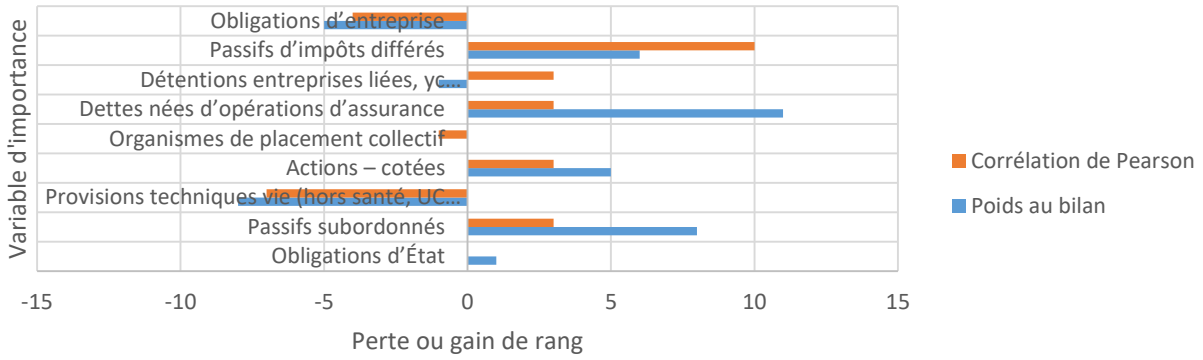
4.1.2.2. Synthèse de la significativité des variables

On observe dans les figures 25.a à 25.d les changements d'importance des variables entre les différentes méthodes au travers de la variation du rang d'importance des variables dans la prédiction avec l'une ou l'autre méthode. On retrouve les constats ci-dessus, à savoir que certaines variables qu'on considérait comme majeures pour expliquer la variation des fonds propres en utilisant une méthode simpliste comme le poids au bilan ou plus élaborée comme les corrélations linéaires, se trouvent remises en cause par l'analyse d'importance du modèle de gradient boosting.

- Concernant l'échantillon 1, dans la figure 25.a, on confirme notre 1ère analyse, à savoir que les obligations d'entreprise et les provisions techniques vie sont moins déterminantes qu'attendu dans la

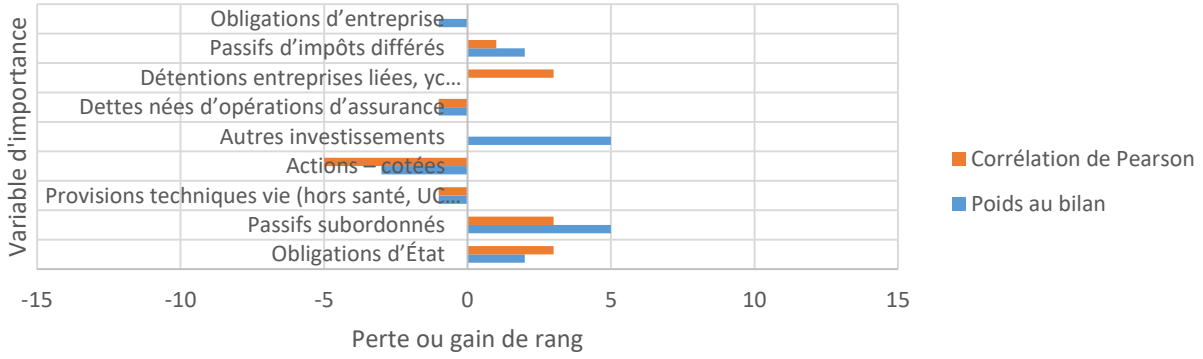
prévision avec l'algorithme de gradient boosting. A l'inverse, on note l'importance plus marquée des passifs d'impôt différés, des passifs subordonnés et des dettes nées d'opération d'assurance :

Figure 25.a : synthèse des variations de rang des variables d'importance du modèle gradient boosting comparé au rang des corrélations de Pearson et au rang des poids au bilan pour l'échantillon 1 des sociétés vie et mixte



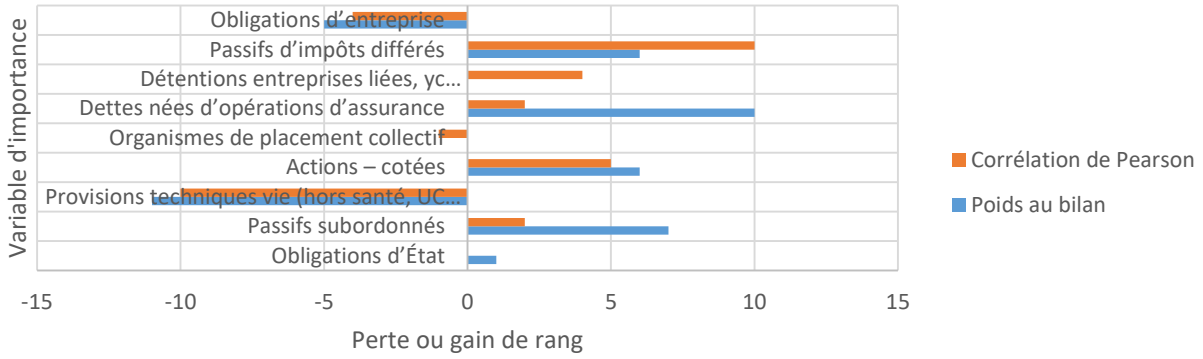
- Concernant l'échantillon 2, on observe dans la figure 25.b que ce sont les actions cotées qui perdent de leur importance au profit des passifs subordonnés, des autres investissements et des obligations d'État :

Figure 25.b : synthèse des variations de rang des variables d'importance du modèle gradient boosting comparé au rang des corrélations de Pearson et au rang des poids au bilan pour l'échantillon 2 des sociétés vie



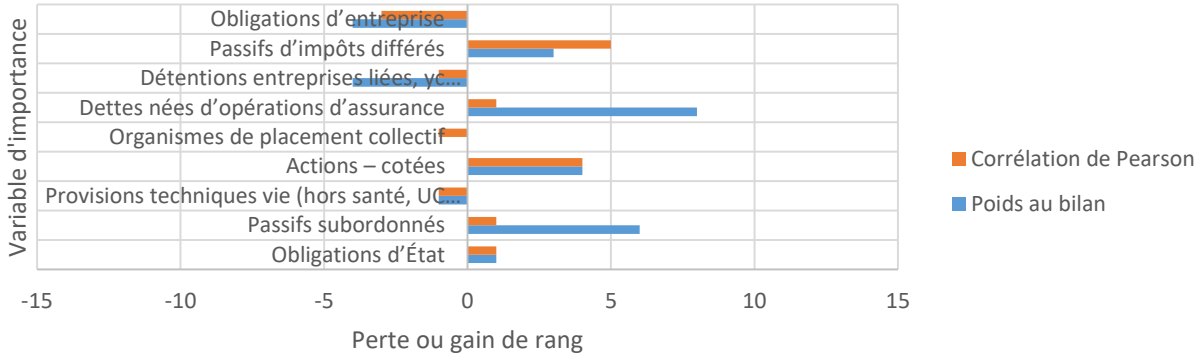
- Concernant l'échantillon 3, on observe dans la figure 25.c que, comme pour l'échantillon 1, ce sont les provisions techniques vie et les obligations d'entreprise qui perdent en importance, alors que le rang des passifs d'impôt différés, des passifs subordonnés et des dettes nées d'opération d'assurance progressent :

Figure 25.c : synthèse des variations de rang des variables d'importance du modèle gradient boosting comparé au rang des corrélations de Pearson et au rang des poids au bilan pour l'échantillon 3 des sociétés mixte



- Concernant l'échantillon 4, on observe enfin dans la figure 25.d que les passifs d'impôts différés, les dettes nées d'opérations d'assurance et les passifs subordonnés sont plus importants au détriment des obligations d'entreprise et des détentions dans les entreprises liées.

Figure 25.d : synthèse des variations de rang des variables d'importance du modèle gradient boosting comparé au rang des corrélations de Pearson et au rang des poids au bilan pour l'échantillon 4 des sociétés avec majorité de PB



En synthèse de ces constats, on observe que l'algorithme de gradient boosting accorde davantage d'importance à certaines variables, en particulier les passifs d'impôt différés, les passifs subordonnés et les dettes nées d'opérations d'assurance, pour la majorité des échantillons, que leurs poids respectifs au bilan ou leur corrélation linéaire avec les fonds propres de base. A l'inverse, la significativité des obligations d'entreprise et des provisions techniques vie est majoritairement en retrait comparé à ces métriques.

Ainsi la modification de la métrique de mesure de l'importance des variables apporte une véritable plus-value à la prédiction lorsqu'on compare le score R^2 du modèle de régression linéaire multiple et celui du modèle de gradient boosting par exemple (respectivement en moyenne 93,94% contre 98,02% sur tous les échantillons). Ce

gain de précision est obtenu par l'utilisation d'un algorithme plus fiable de sélection de la significativité des variables.

4.2. Définition des chocs bilanciels

Nous restreignons à présent l'étude à l'unique échantillon 4 des sociétés avec majorité de PB qui est celui le mieux décrit par notre algorithme de gradient boosting. Afin d'observer l'impact de différentes variations sur les fonds propres prudentiels des sociétés à majorité de PB, nous allons dans un premier temps calibrer les chocs à appliquer. Pour ce faire, nous partons de l'observation du panel afin d'observer la distribution des variations annuelles des principales variables.

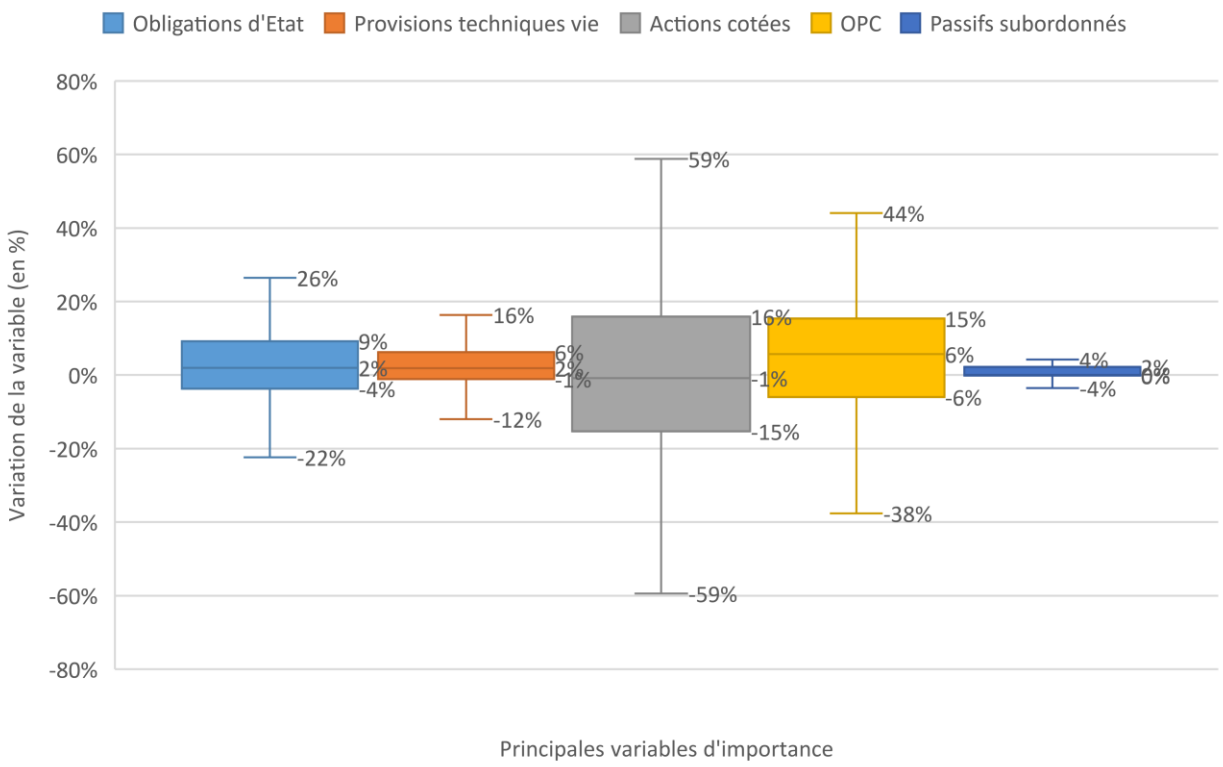
4.2.1. Calibration des chocs

L'approche retenue dans l'étude consiste à calibrer les chocs en fonction des variations historiques observées sur l'échantillon. Le calibrage correspond ainsi aux principales variations observées sur le panel à horizon un an (moyenne, quartiles, min, max notamment), ajustées le cas échéant pour prendre en compte le comportement par nature non homogène entre les variations des différentes variables. Le principe étant que les chocs retenus doivent être à nouveau observables dans des conditions de marché normales avec une grande probabilité.

On observe dans la figure 26 que les variations annuelles des principales variables d'importance sont comprises (en excluant les variations extrêmes) entre :

- Pour une approche par quartile : +16% (1er quartile de la variation des actions cotées) et -15% (dernier quartile de la variation annuelle des actions cotées) ;
- Pour une approche min / max : +59% (max de la variation des actions cotées) et -59% (min de la variation des actions cotées).

Figure 26 : variation annuelle des 5 principales variables d'importance des sociétés à majorité de PB



Afin de trouver une calibration qui reflète au mieux les variations observées des 5 variables, en compensant les variations plus volatiles de certaines (actions cotées notamment), et pour s'aligner sur la majorité des variations observées de la variable de plus grande importance (obligations d'État), on calibre comme suit les chocs à appliquer :

TABLEAU 12 : VALEUR DES CHOCS A APPLIQUER AUX VARIABLES D'IMPORTANCE

Choc	Nature
Choc 1	Hausse annuelle de +20% de la valeur de la variable
Choc 2	Hausse annuelle de +10% de la valeur de la variable
Choc 3	Baisse annuelle de -10% de la valeur de la variable

Choc 4	Baisse annuelle de -20% de la valeur de la variable
Choc 5	Baisse annuelle de -40% de la valeur de la variable

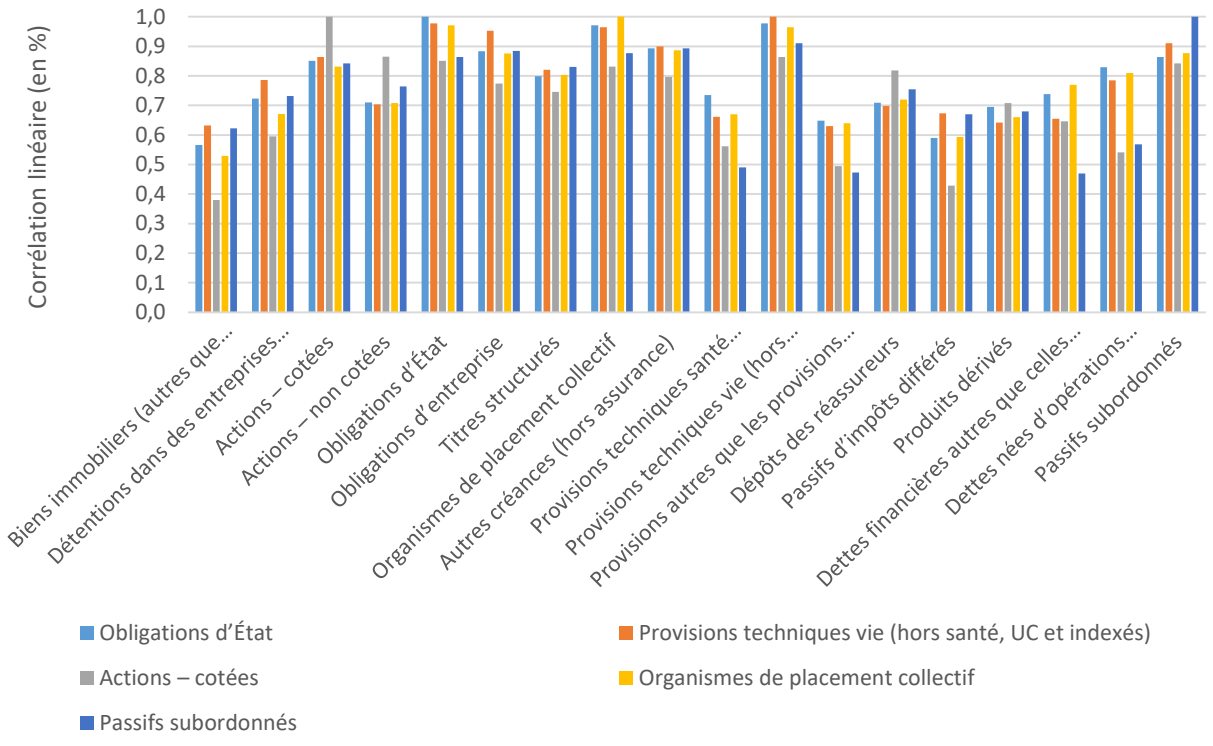
4.2.2. Matrice de corrélation

En première partie de notre étude les corrélations linéaires entre les variables d'entrée et les fonds propres prudentiels. Dans la même logique, nous nous intéressons désormais aux corrélations linéaires entre les 5 principales variables d'importance et les autres variables. Nous utiliserons les corrélations linéaires historiques ainsi calculées comme matrice de corrélation. Ainsi, tout choc appliqué à une variable d'importance modifiera la valeur des autres variables d'entrée du modèle à hauteur de la corrélation linéaire entre la variable d'importance et les autres variables.

Cette approche des corrélations sur la base de l'analyse historique permet d'assurer que tout choc appliqué à une variable d'importance sera appliqué aux autres variables sur la base des observations des corrélations réellement observées sur l'échantillon. En prenant en compte l'interdépendance des variables sur une durée suffisamment grande et avec un nombre d'observations suffisamment important, on permet au modèle de reproduire les covariances entre variables, et donc notamment les interactions entre les actifs et les passifs, qui sont fondamentaux pour les sociétés d'assurance vie et notamment celles disposant d'une majorité d'engagements avec participation aux bénéfices.

On observe dans la figure 27 l'ensemble des corrélations linéaires entre les 5 variables d'importance du modèle de gradient boosting et les autres variables d'entrée du modèle, qui constitue donc la matrice de corrélation de nos chocs :

Figure 27 : matrice de corrélation provenant des corrélations linéaires entre les variables d'importance du modèle et les autres variables du modèle



4.3. Analyse de la sensibilité des fonds propres prudentiels

Après avoir calibré nos chocs et défini leur corrélation, nous nous intéressons désormais à l'impact d'un choc sur les principales composantes du bilan économique sur la valeur des fonds propres prudentiels.

Nous allons donc observer l'impact sur les fonds propres prudentiels d'une variation de valeur forfaitaire des 5 principales variables d'intérêt (celles dont le score d'importance issu de l'algorithme de gradient boosting est le plus important). Comme décrit précédemment, ces variations prennent la forme d'un choc forfaitaire de +20%, +10%, -10%, -20% et -40% appliqué à la variable d'intérêt, en tenant compte des corrélations avec les autres variables telles qu'observées dans l'échantillon.

4.3.1. Sensibilité aux obligations d'État

Afin d'évaluer la sensibilité des fonds propres prudentiels à une variation de valeur des obligation d'État, on applique les chocs 1 à 5 à la variable d'importance, en tenant compte des corrélations avec toutes les autres variables du modèle. Puis on applique le modèle de gradient boosting préalablement calibré sur les données d'apprentissage afin de prédire, pour chacune des observations, le montant de fonds propres après choc. On évalue enfin la variation de valeur en pourcentage entre les fonds propres avant choc et les fonds propres prédits après choc.

Comme le laisse présager le score d'importance, la sensibilité aux obligations d'État est la plus significative. Si on simule une variation de la valeur de cette variable, on observe que la prédiction de fonds propres prudentiels varie de manière significative, comme observé dans les figures 28.a et 28.b ci-après :

Figure 28.a : impact des différents scénarios de variation de valeur des obligations d'Etat sur la valeur des fonds propres prudentiels des sociétés à majorité de PB (hors valeurs extrêmes)

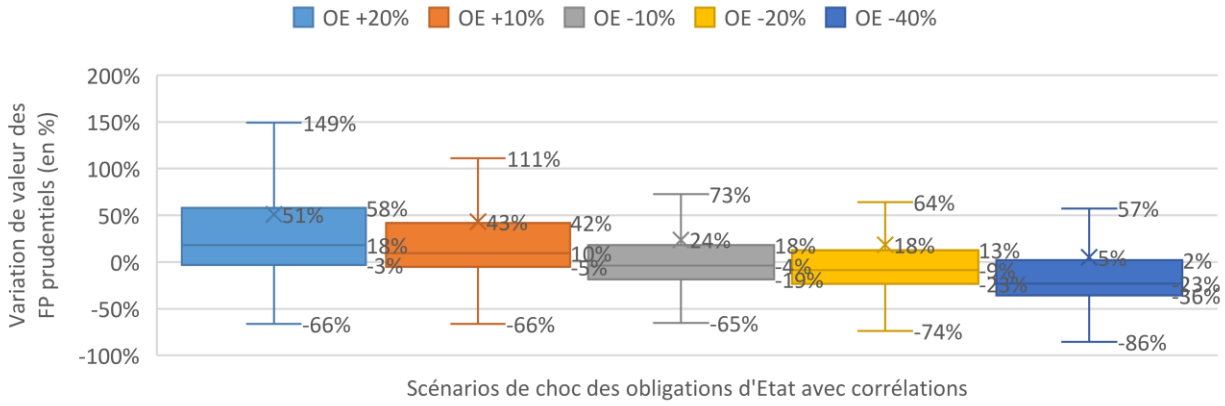
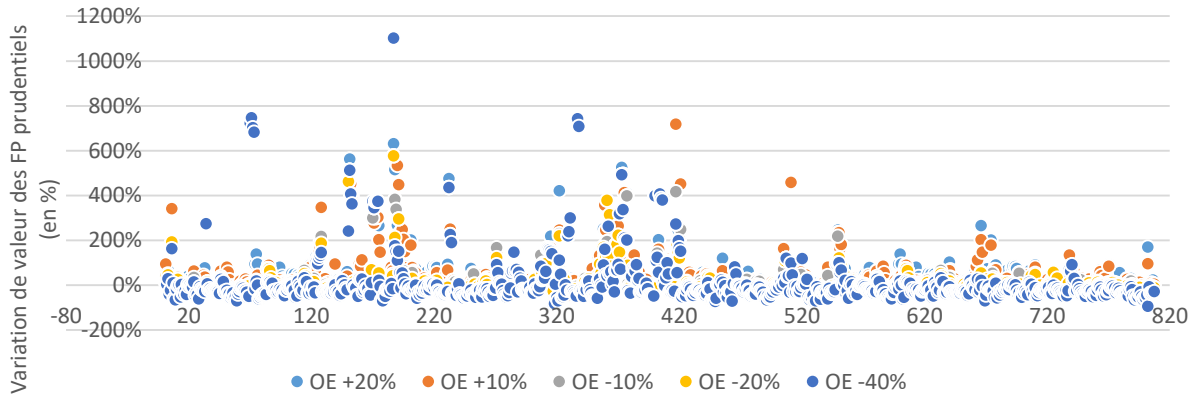


Figure 28.b : impact des différents scénarios de variation de valeur des obligations d'Etat sur la valeur des fonds propres prudentiels des sociétés à majorité de PB



Les observations ci-dessus sont riches de plusieurs enseignements :

- En médiane, l'impact d'un choc des obligations d'État est de même sens que la variation de valeur des fonds propres prudentiels ;
- En médiane, la variation de valeur des fonds propres prudentiels est inférieure ou égale au choc appliqué sur les obligations d'État ;
- L'écart interquartile dans les 5 scénarios de choc tend à diminuer avec les scénarios les plus défavorables de choc des obligations d'État, mais reste important dans tous les cas.

Il apparait donc que les fonds propres prudentiels sont très sensibles à la valeur des obligations d'État. Si on s'intéresse au comportement médian des observations, on conclut qu'il existe une élasticité positive entre les

variations des 2 variables, ce qui semble intuitif. Cependant, les prédictions obtenues à partir du modèle de gradient boosting calibré sur l'historique des observations, dont la qualité de prédiction a été précédemment démontré, montre une grande variabilité d'impact sur les fonds propres prudentiels, comme en témoigne l'importance de l'écart interquartile. Il n'est donc pas possible de conclure à un lien systématique entre une hausse (respectivement une baisse) de la valeur des obligations d'État et une hausse (respectivement une baisse) des fonds propres prudentiels.

A titre de vérification du modèle, nous avons testé sa qualité de prédiction sur une période de stress des obligations d'Etat à fin 2019. Les résultats confirment l'efficacité du modèle puisque le coefficient de détermination R^2 est alors de 99,87% avec 161 observations à fin 2019, contre 99,84% sur la totalité des observations de l'échantillon 4 soit 807 observations sur la totalité des 22 trimestres.

4.3.2. Sensibilité aux autres variables

Du fait de la corrélation, tout choc appliqué aux autres variables d'importance a un impact important sur la valeur des obligations d'État. Cette corrélation étant d'au moins 85% selon les valeurs de la figure 27 calculée précédemment, l'effet de contagion est important entre le choc sur les autres variables d'importance et les obligations d'État, qui présente pour mémoire un score d'importance très significatif dans la prédiction au moyen de l'algorithme de gradient boosting.

En conséquence, on observe dans les figures 29.a à 29.b que les constats que nous avons fait concernant le lien entre les fonds propres prudentiels et les obligations d'État sont toujours valables s'agissant d'appliquer les chocs corrélés aux autres variables d'intérêt. À savoir que, en médiane, un choc appliqué aux autres variables est de même sens que la variation des fonds propres économiques et que la variation de valeur des fonds propres est inférieure ou égale à la variation appliquée aux variables d'entrée.

Figure 29.a : impact des différents scénarios de variation de valeur des provisions techniques vie sur la valeur des fonds propres prudentiels des sociétés à majorité de PB (hors valeurs extrêmes)

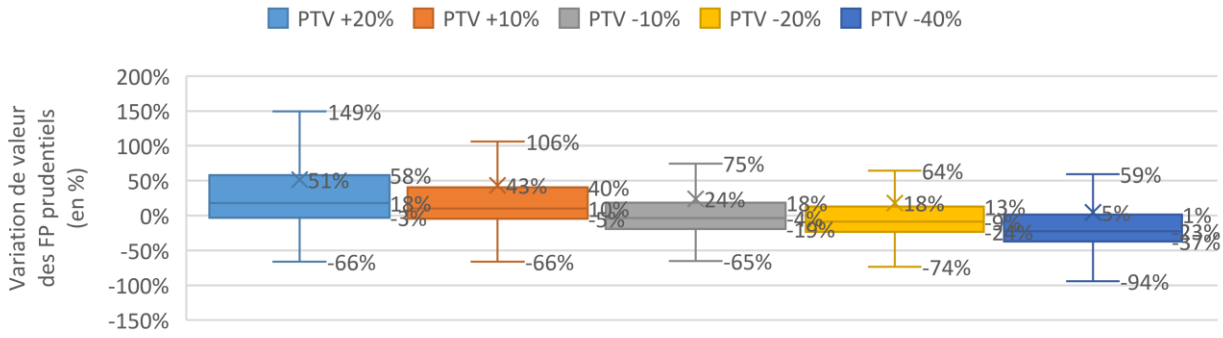


Figure 29.b : impact des différents scénarios de variation de valeur des actions cotées sur la valeur des fonds propres prudentiels des sociétés à majorité de PB (hors valeurs extrèmes)

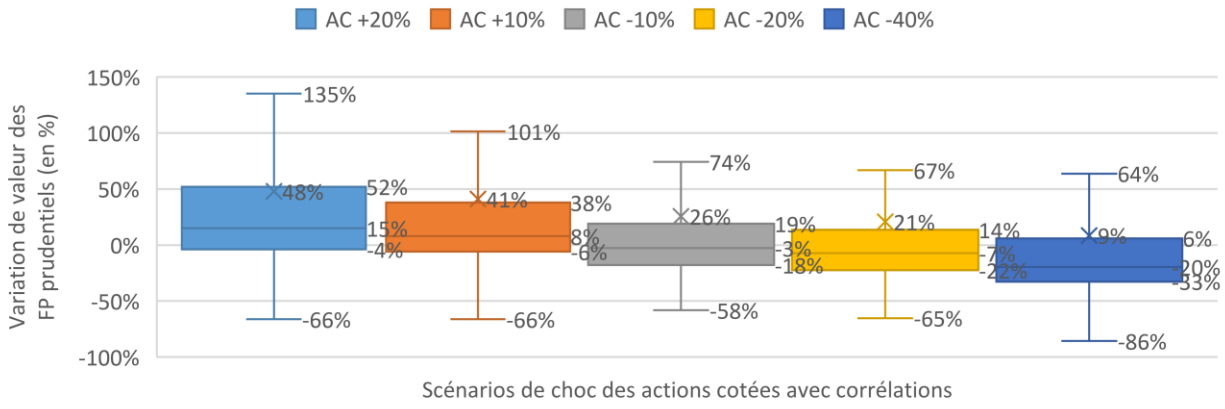


Figure 29.c : impact des différents scénarios de variation de valeur des OPC sur la valeur des fonds propres prudentiels des sociétés à majorité de PB (hors valeurs extrèmes)

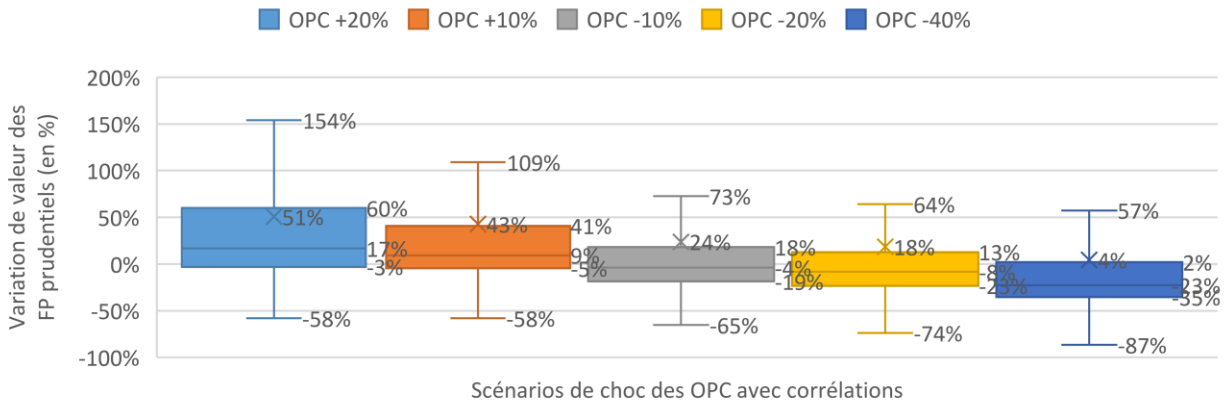
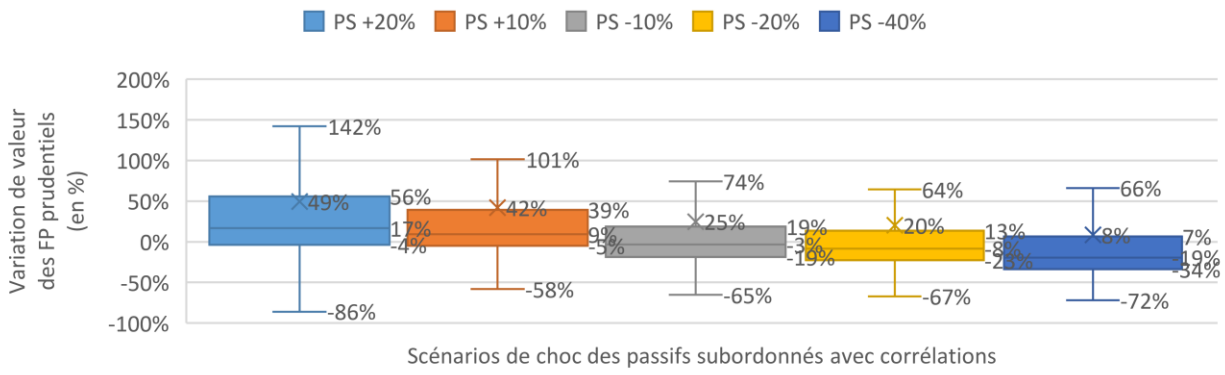


Figure 29.d : impact des différents scénarios de variation de valeur des passifs subordonnés sur la valeur des fonds propres prudentiels des sociétés à majorité de PB (hors valeurs extrêmes)



En conclusion, l'importance des obligations d'État dans la prédiction des fonds propres économiques par le modèle de gradient boosting est très significative, quel que soit l'échantillon (sociétés vie et mixte, vie, mixte ou majorité de PB). Celle des autres variables l'est beaucoup moins. Qui plus est, en appliquant différents chocs corrélés sur base historique, on retrouve le rôle primordial des obligations d'État dans les prédictions des fonds propres économiques des sociétés à majorité de PB. Enfin, l'algorithme, dont la qualité de prédiction a été démontrée dans le chapitre précédent, ne permet pas de conclure à une relation linéaire entre la variation de valeur des variables d'entrée et les fonds propres économiques. En revanche, l'algorithme est capable de restituer précisément la valeur des fonds propres économiques quel que soit la valeur des 5 variables significatives en entrée du modèle, grâce à une reproduction fidèle et dynamique des interactions entre les composantes du bilan des assureurs, et notamment les interactions entre les actifs et les passifs.

5. Conclusion

Les fonds propres économiques des sociétés spécialisées en assurance vie sont un élément déterminant de leur solvabilité. Dans un contexte de baisse des taux d'intérêt, il apparait nécessaire de déterminer les principales composantes de variation des fonds propres économiques, afin notamment d'évaluer l'impact du changement des conditions de marché sur les assureurs. Pour ce faire, la présente étude vise à construire un outil de prédiction des fonds propres économiques basé sur l'observation des évolutions du bilan économique des assureurs. La qualité de cet outil est appréciée au regard de sa capacité à effectuer des prédictions de fonds propres qui soient proches de ceux réellement observés dans l'échantillon. Cet outil, une fois construit et son efficacité confirmée, doit permettre d'identifier les principales variables qui influent sur les fonds propres économiques, et ensuite de déterminer l'impact d'un choc appliqué à ces variables sur la valeur des fonds propres économiques.

Dans le cadre de notre étude, nous avons ainsi observé l'évolution des fonds propres économiques de 197 sociétés d'assurance vie et mixte sur une période de 22 trimestres consécutifs (soit 2 771 observations) en lien avec les évolutions de leur bilan économique. Dans un premier temps, une analyse statistique descriptive a permis de déterminer les principales composantes et les principales corrélations entre les variables du bilan économique et les fonds propres. Ensuite, plusieurs modèles d'apprentissage automatisé ont été proposés afin de prédire l'évolution des fonds propres en fonction des observations des variables du bilan économique sélectionnées. Ces modèles ont été comparés sur plusieurs échantillons (vie et mixte, vie, mixte et majorité de PB) construits pour améliorer la prédiction. Il en résulte une qualité de prévision appréciable avec les algorithmes non paramétriques, en particulier sur l'échantillon des sociétés avec une majorité de contrat avec PB en utilisant l'algorithme ensembliste d'arbres de décision de gradient boosting. Enfin, après avoir listé les variables d'importance et calibré les chocs et les corrélations sur la base des données historiques, nous avons pu apprécier la sensibilité des fonds propres économiques aux différentes variables. Il en résulte une très grande dépendance des fonds propres à la valeur des obligations d'État. Et a contrario, on observe que d'autres variables comme les obligations d'entreprise, les actions ou les OPC sont moins déterminantes que ne le laisse présager leur poids au bilan économique ou leur corrélation linéaire avec les fonds propres.

Les résultats ainsi obtenus permettent d'estimer l'évolution des fonds propres économiques en fonction de chocs qui peuvent se produire sur ses principales composantes, sans nécessairement simuler des milliers de scénarios et leurs impacts sur le bilan prudentiel dans plusieurs états du monde, comme le font individuellement les assureurs pour calculer la valeur économique de leurs engagements, et ensuite de leurs fonds propres. De plus, ces résultats sont obtenus en utilisant des données historiques relativement longues qui ont permis d'observer l'impact de différentes situations de marché sur les fonds propres économiques, et notamment de la baisse importante des taux d'intérêt, de la hausse des spreads ou de la baisse importante des marchés actions, qui se sont produits durant les 22 trimestres de la période d'observation.

Toutefois, les conclusions de l'étude, qui ne constituent pas une position de l'ACPR, ne sont pas entièrement généralisables : certaines situations de marché ne sont pas incluses dans les données, et notamment une hausse brutale des taux d'intérêt, scénario très sensible pour la solvabilité des assureurs vie et mixte qui pourrait provoquer une hausse massive des rachats et une forte baisse des plus-values obligataires. Ensuite, l'étude ne prend pas en compte l'effet de l'arrêté PPB sur les fonds propres disponibles pour couvrir l'exigence de capital, ni les règles d'écrêtement de ces fonds propres qui diminuent les fonds propres économiques disponibles servant au calcul de la solvabilité. Enfin, les prédictions ne prennent pas en compte directement l'impact des variations économiques (taux d'intérêt, spread de crédit ou indice action par exemple) mais indirectement leur impact sur

les principales variables du bilan économique (provisions techniques, placements…) qui dépendent également d'autres facteurs, et notamment des décisions stratégiques futures comme la stratégie de placement, de participation aux bénéfices ou de réassurance.

Afin de compléter la présente étude, il serait intéressant d'envisager la construction d'un algorithme de prédiction du capital de solvabilité requis, basée également sur les états réglementaires. Ceci permettrait d'avoir une vision plus complète de l'évolution du ratio de solvabilité des assureurs spécialisés en vie. Enfin, il serait intéressant d'intégrer une analyse de l'impact de l'évolution des facteurs économiques (taux d'intérêt, spread de crédit ou indice action) et des futures décisions de gestion sur le calcul des fonds propres économiques, afin d'affiner les résultats de la présente étude.

Bibliographie

ACPR (2021), *La situation des assureurs soumis à Solvabilité II en France au premier semestre 2021*, Analyses et synthèses, n°129-2021.

Buzzi, A. (2017), *Approximation du bilan économique par apprentissage automatique et application à l'ORSA*, Mémoire IA.

Dubois, D., Féderié, A. & Ranaivozanany, V. (2021), *Appliquer la data science à l'assurance*, L'argus de l'assurance éditions.

Jakobowicz, E. (2018), *Python pour le data scientist*, Dunod.

Juilliard, MF. (2021), *L'assurance vie épargne en France enfin l'âge de raison ?*, Risques, n°127, 88-92.

Lemberger, P., Batty, M., Morel, M. & Raffaëlli, JL. (2019), *Big Data et Machine Learning*, Dunod.

Tassi, P. (2004), *Méthodes statistiques*, Economica.

Soix, E. (2018), *Estimation du ratio de solvabilité par apprentissage statistique supervisé*, Mémoire IA.

Vannieuwenhuyze, A. (2019), *Intelligence artificielle vulgarisée*, Eni.