Mémoire présenté le :

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

Par : BRUNEL Eva

Titre Mesures d'erreur du *Best Estimate* en assurance des emprunteurs

Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise :

Nom : BPCE Vie

Signature :

Directeur de mémoire en entreprise :

Nom : KERBOUL Laure

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Mesures d'erreur du *Best Estimate* en assurance des emprunteurs

Auteur :
Eva BRUNEL

Tuteur d'entreprise :
Adrien HELARY
Responsable :
Laure KERBOUL
Tuteur académique :
Stéphane LOISEL

25 Mars 2022

Résumé

Mots clés : *Best Estimate*, marge d'erreur, évaluation du risque, Solvabilité 2, Assurance des emprunteurs, algorithmes d'apprentissage, loi d'expérience, risque de modèle

Les sociétés d'assurance sont soumises à la norme européenne Solvabilité 2, exigeant notamment le calcul de provisions techniques composées du *Best Estimate* et de la marge pour risque. La complexité des modèles de projection permettant de calculer le *Best Estimate* entraîne des erreurs sur ce dernier. Ces modèles s'appuient sur les données du portefeuille et sur des hypothèses estimées en amont, ceci accentuant encore plus les erreurs dans le calcul du *Best Estimate*.

Afin de limiter d'éventuelles pertes financières, les sociétés d'assurance doivent réduire au maximum ces erreurs. Aussi, l'objectif de ce mémoire est de quantifier l'ensemble des erreurs possibles tout en les répartissant en trois catégories : les erreurs liées aux données en entrée du modèle, les erreurs liées aux hypothèses et les erreurs liées au modèle.

Pour ce faire, dans un premier temps, les anomalies présentes dans les données en entrée du modèle seront repérées à partir de différents algorithmes d'apprentissage. Dans un second temps, les erreurs liées aux hypothèses seront évaluées selon des méthodes statistiques telles que le *bootstrap*. Dans un troisième temps, les anomalies présentes dans le modèle seront détectées en effectuant du *backtesting* sur les flux projetés par le modèle. Plus précisément, le *backtesting* consiste à comparer les flux projetés et les flux réels. Enfin, l'impact de l'ensemble des erreurs sur la valeur du *Best Estimate* sera quantifié. Cette marge d'erreur sera ensuite comparée à la marge pour risque.

Abstract

Key words : Best Estimate, margin of error, risk assessment, Solvability 2, borrowers' insurance, Machine Learning, experience tables, backtesting

Insurance companies are subject to the European standard Solvability 2, which requires in particular the calculation of technical provisions composed of the Best Estimate and the risk margin. The complexity of forecasting models used for calculating the Best Estimate results in errors on the latter. These models are based on the data from the portfolio and on hypotheses estimated upstream, emphazing even more errors in the calculation of the Best Estimate.

In order to limit possible financial losses, insurance companies must reduce as much as possible these errors. Thus, the objective of this thesis is to quantify the whole possible errors, while dividing them into three categories : errors related to the model input data, assumption errors, and errors related to the model.

For those purpose, firstly, the anomalies present in the model input data will be identified thanks to different learning algorithms. Secondly, the assumptions errors will be evaluated according to statistical methods such as the bootstrapping. Thirdly, the anomalies present in the model will be detected by performing backtesting on the flows projected by the model. More specifically, backtesting consists in comparing the projected flows with the real flows. Finally, the impact of all the errors on the value of the Best Estimate will be quantified. This margin of error will then be compared to the risk margin.

Remerciements

Un grand merci à mon tuteur Adrien HELARY pour m'avoir admirablement accompagné dans le développement de ce mémoire et pour m'avoir permis de découvrir certaines facettes passionnantes de l'actuariat.

Je tiens aussi à remercier chaleureusement l'ensemble de l'équipe de la fonction actuarielle de BPCE Vie et entre autres ma responsable Laure KERBOUL, Laura PEREZ et Jim LOTHAMMER pour le temps qu'ils m'ont accordé dans la réalisation de ce mémoire, leurs précieux conseils et leurs relectures.

Mes remerciements vont également à Nicolas GEORGES, actuaire à la direction Produits de BPCE Vie, pour ses conseils affinés et son aide à la production technique de ce mémoire.

Pour finir, je remercie ma famille pour leurs encouragements tout au long de ce mémoire et leurs relectures.

Table des matières

Résumé **1**

Abstract **2**

Remerciements **3**

Introduction **2**

1 Contexte général **4**

1.1 Périmètre de l'étude 4

1.2 Contexte réglementaire 7

1.2.1 Introduction à Solvabilité 2 7

1.2.2 Le *Best Estimate* 9

1.3 La fonction actuarielle 11

1.4 Notions introductives sur les erreurs 12

2 Erreurs liées aux données utilisées **13**

2.1 Classification des erreurs sur les données 13

2.1.1 Critère d'exhaustivité 15

2.1.2 Critère d'exactitude 15

2.1.3 Caractère approprié 16

2.1.4 Erreurs liées à l'agrégation des données 16

2.1.5 Seuils des anomalies 17

2.2 Réduction de dimension des données 19

2.2.1 Analyse en Composante Principale (ACP) 19

2.2.2 UMAP 20
2.3 Algorithmes de détection d'anomalies 21
2.3.1 Apprentissage non-supervisé 22
2.3.1.1 K-moyennes 22
2.3.1.2 *Isolation Forest* 24
2.3.1.3 HDBSCAN 26
2.3.2 Apprentissage semi-supervisé : cas du *PU learning* 28
2.4 Loi de Benford 29

3 Erreurs liées aux hypothèses 31
3.1 Présentation des hypothèses 31
3.2 Classification des erreurs liées aux hypothèses 33
3.2.1 Anomalies sur les données de construction 33
3.2.2 Anomalies liées aux erreurs opérationnelles 34
3.2.3 Biais d'estimation (ou biais statistique d'échantillonnage) . . . 34
3.2.4 Biais de modélisation 35
3.2.4.1 Biais d'homogénéité du portefeuille 35
3.2.4.2 Biais temporel 37
3.2.4.3 Biais lié à la rupture des contraintes 37
3.2.4.4 Biais lié aux interactions entre l'actif et le passif . . . 38
3.3 Estimation d'une loi d'expérience 39
3.3.1 Introduction aux modèles de durée 40
3.3.2 Censures et troncatures 41
3.3.2.1 La censure à droite 42
3.3.2.2 La censure à gauche 42
3.3.2.3 La troncature 43
3.3.3 Modélisation non-paramétrique des taux bruts 43
3.3.3.1 Estimateur de Kaplan-Meier 44
3.3.3.2 Estimateur de Hoem 44
3.3.4 Lissage des taux bruts 45
3.3.4.1 Méthode des moyennes mobiles 45
3.3.4.2 Méthode des noyaux discrets 45

4 Erreurs liées au modèle de projection 47

4.1 Classification des erreurs liées au modèle 47
4.2 *Backtesting* 49
4.2.1 Test de couverture non conditionnelle 50
4.2.2 Test de Markov de Christoffersen (1998) 51
4.2.3 Tests basés sur la régression d'Engle et Manganelli (2004) . . 51

5 Mise en application en assurance des emprunteurs 53
5.1 Détection d'anomalies dans les données en entrée du modèle 53
5.1.1 Présentation des données 54
5.1.2 Première approche : analyses univariées 55
5.1.3 Réduction de dimension de la base 57
5.1.4 Sélection des anomalies : algorithmes de *clustering* 59
5.1.4.1 Application de l'algorithme K-moyennes 60
5.1.4.2 Application de l'algorithme *Isolation Forest* 64
5.1.4.3 Application de l'algorithme HDBSCAN 67
5.1.4.4 Regroupement des résultats 71
5.1.5 Analyse des anomalies 73
5.1.5.1 Analyses univariées sur les anomalies 73
5.1.5.2 Application de la Loi de Benford au capital restant dû 74
5.1.5.3 Application du *PU learning* aux anomalies 75
5.2 Estimation des erreurs liées aux hypothèses 76
5.2.1 Étude des biais associés aux lois d'expérience 76
5.2.1.1 Évaluation du biais d'homogénéité 77
5.2.1.2 Évaluation du biais temporel 82
5.2.1.3 Évaluation du biais d'estimation 86
5.2.1.4 Évaluation du biais total 88
5.2.2 Détection d'anomalies dans les données de construction des hypothèses 91
5.3 Détection d'anomalies dans le modèle de projection 92
5.3.1 *Backtesting* en tenant compte de la volatilité réelle du portefeuille 93
5.3.2 *Backtesting* en tenant compte des biais liés aux hypothèses . . 100
5.3.3 Avis final sur la qualité du modèle 104
5.4 Impact des différentes erreurs sur le *Best Estimate* 106

5.4.1 Impact des anomalies présentes dans les données en entrée du modèle 106
5.4.2 Impact des anomalies présentes dans les données de construction des hypothèses 106
5.4.3 Impact des biais liés aux hypothèses 107
5.4.4 Impact total sur le *Best Estimate* 109

Conclusion **111**

Bibliographie **113**

A Notations des indicateurs de contrôle **114**

B Analyses univariées sur les données et les anomalies **115**

Liste des tableaux

2.1 Synthèse des indicateurs et seuils des erreurs sur les données 18
2.2 Avantages et inconvénients de K-moyennes 23
2.3 Avantages et inconvénients de la méthode *Isolation Forest* 25
2.4 Avantages et inconvénients de la méthode HDBSCAN 27

5.1 Synthèse des incohérences détectées à partir d'analyses univariées . . 56
5.2 Temps d'exécution des algorithmes de réduction de dimension 57
5.3 Paramètres des algorithmes d'apprentissage non-supervisés 59
5.4 Proportion d'anomalies détectées par les k-moyennes 64
5.5 Proportion d'anomalies détectées par IF 64
5.6 Synthèse des anomalies détectées par les algorithmes de *clustering* . . 71
5.7 Analyse des anomalies détectées par les algorithmes de *clustering* . . 73
5.8 Impact de la suppression des anomalies sur la valeur du BE 106
5.9 Impact des anomalies présentes dans les données de construction des hypothèses sur la valeur du BE 107
5.10 Synthèse des BE obtenus à partir des biais liés aux lois de mortalité et de rachat 107
5.11 Impact des biais liés aux hypothèses sur la valeur du BE 109
5.12 Impact des différentes erreurs sur la valeur du BE 109

Table des figures

1.1 Sources d'erreurs sur l'évaluation du *Best Estimate* 5
1.2 Types de prêts en assurance des emprunteurs 5
1.3 Bilan Solvabilité 2 8
1.4 Composition du *Best Estimate* 9

2.1 Exemple du choix des axes en dimension 2 pour l'ACP 19
2.2 Fonctionnement de UMAP en dimension 3 21
2.3 Fonctionnement de k-moyennes 23
2.4 Fonctionnement d'*Isolation Forest* (source : lovelyanalytics.com) . . . 24
2.5 Formation des clusters par HDBSCAN 26
2.6 Fonctionnement de la technique espion 29
2.7 Loi de Benford - Fréquence d'apparition du 1er chiffre significatif . . . 30

3.1 Exemple d'évolution dans le temps des proportions de CSP dans le portefeuille 36
3.2 Exemple de *bootstrap* avec quatre groupes de risque 37
3.3 Le modèle multi-états de l'assurance des emprunteurs 39

5.1 Base de données réduite en dimension 2 58
5.2 Sélection du paramètre k 61
5.3 Sélection de la taille minimum d'un cluster 62
5.4 Anomalies détectées par k-moyennes 63
5.5 Sélection du score maximum 65
5.6 Anomalies détectées par IF 66

5.7 Différentes possibilités de partitionnement 68
5.8 Sélection de la largeur de tranche 69
5.9 Étapes du partitionnement 70
5.10 Anomalies détectées par HDBSCAN 71
5.11 Anomalies sélectionnées 72
5.12 Loi de Benford appliquée au capital restant dû 75
5.13 Moyenne des écarts relatifs entre la borne supérieure du biais d'homogénéité et le flux réel 78
5.14 Bornes brutes du biais d'homogénéité 80
5.15 Bornes lissées du biais d'homogénéité 81
5.16 Moyenne des écarts relatifs entre la borne supérieure du biais temporel et le flux réel 82
5.17 Bornes brutes représentant le biais temporel 84
5.18 Bornes lissées représentant le biais temporel 85
5.19 Bornes représentant le biais d'estimation 87
5.20 Moyenne des écarts relatifs entre la borne supérieure du biais d'estimation et le flux réel 88
5.21 Moyenne des écarts relatifs entre la borne supérieure du biais total et le flux réel 89
5.22 Biais totaux 90
5.23 *Backtesting* avec l'intervalle de volatilité réelle du portefeuille 97
5.24 Moyenne des écarts relatifs entre la borne supérieure et les flux projetés par année de projection 98
5.25 Tests de *backtesting* 98
5.26 Écarts relatifs entre les bornes supérieures et les flux réels 99
5.27 *Backtesting* avec les intervalles provenant des biais liés aux hypothèses 101
5.28 Proportion du flux réel situé au-delà de la borne supérieure 103
5.29 *Backtesting* final 105
5.30 Analyse de l'impact des anomalies dans les données en entrée des hypothèses sur le *Best Estimate* 108
5.31 Analyse de l'impact des erreurs sur le *Best Estimate* 110

Introduction

La directive Solvabilité 2 occupe une place importante dans le monde assurantiel actuel. Cette directive a pour objectif d'assurer la solvabilité des entreprises d'assurances face à des évènements exceptionnels. Pour ce faire, elle prévoit notamment des exigences quantitatives telles que le calcul de provisions techniques regroupant le *Best Estimate* et la marge pour risque.

Le *Best Estimate* correspond à la moyenne des flux de trésorerie futurs pondérée par leur probabilité de survenance d'un portefeuille d'assurance, compte tenu de la valeur temporelle de l'argent et d'une courbe pertinente des taux d'actualisation. Globalement, le *Best Estimate* est une représentation des engagements de l'assureur vis-à-vis de ses assurés à un instant donné. Aussi, les assureurs se doivent de fiabiliser autant que possible le calcul du *Best Estimate* afin d'évaluer au mieux les provisions techniques.

Le *Best Estimate* est évalué à partir d'un modèle de projection permettant d'estimer les différents flux futurs de l'assureur tels que les primes, les prestations ou encore les frais. Or, l'utilisation de données et d'hypothèses erronées dans ce modèle de projection pourrait entraîner une mauvaise estimation des provisions techniques. Il convient donc à l'assureur de s'assurer de la bonne qualité de ses données et de ses hypothèses. Par ailleurs, l'assureur doit prêter attention au risque de modèle se manifestant par un écart entre le phénomène modélisé et la réalité, dans le but d'évaluer au mieux son ratio de solvabilité.

La présente étude a pour objectif d'évaluer dans un premier temps la qualité des trois composantes utiles au calcul du *Best Estimate* en assurance des emprunteurs, à savoir les données, les hypothèses et le modèle de projection. Elle vise dans un second temps à estimer l'influence de la qualité des données, de la qualité des hypothèses et de la qualité du modèle sur la valeur du *Best Estimate*. Selon l'impact de la qualité de ces trois composantes sur la valeur du *Best Estimate*, ces trois composantes devront faire l'objet d'un retraitement plus ou moins conséquent.

Dans un premier temps, nous présenterons le contexte général de l'étude, laquelle porte sur les prêts personnels en assurance des emprunteurs. Cette étude a été réalisée au sein de la fonction actuarielle de BPCE Vie, l'une des quatre fonctions clés

mises en place par Solvabilité 2 pour le contrôle des risques internes des compagnies d'assurance.

Par la suite, nous étudierons le cadre théorique de la détection d'anomalies. Nous analyserons pour cela plusieurs algorithmes d'apprentissage permettant de repérer des anomalies dans une base de données. Nous nous pencherons également sur d'autres méthodes de détection d'anomalies, en particulier sur la loi de Benford.

Ensuite, nous présenterons les différents biais et anomalies pouvant être présents dans les hypothèses et proposerons plusieurs indicateurs permettant de les contrôler. Puis nous nous pencherons sur le cadre théorique de l'estimation des lois d'expérience, ces dernières représentant les hypothèses qui impactent le plus le *Best Estimate* et qui, de ce fait, seront retenues pour l'application.

Un dernier chapitre théorique portera sur les différentes erreurs présentes dans le modèle de projection et plus précisément les erreurs de modélisation. Nous présenterons l'approche la plus efficace pour évaluer le risque de modèle, à savoir le *backtesting*.

Finalement, nous appliquerons les différents processus de détection d'erreur aux données, aux hypothèses et au modèle utilisés pour le calcul du *Best Estimate* des prêts personnels. Les résultats obtenus seront ensuite analysés, puis l'impact de la qualité des données, des hypothèses et du modèle sur la valeur du *Best Estimate* sera estimé.

Chapitre 1: Contexte général

1.1 Périmètre de l'étude

Ce document a été développé au sein de la fonction actuarielle de BPCE Vie, l'une des quatre fonctions clés de Solvabilité 2. Cette fonction est notamment en charge de la validation des provisions techniques de Solvabilité 2 et de l'appréciation de "la suffisance et la qualité des données utilisées dans le calcul des provisions techniques" comme énoncé dans l'article 48 de la Directive 2009/138 Solvabilité 2. Dans ce document, nous analyserons les potentielles erreurs dans le calcul du *Best Estimate* (BE), le but étant de définir une erreur finale sur la valeur de ce BE. Comme illustré en figure 1.1, quatre sources d'erreurs potentielles ont été répertoriées :

- les données en entrée du modèle ;
- les données utilisées pour la construction des hypothèses ;
- les hypothèses ;
- et enfin le modèle de projection lui-même.

Dans la suite du mémoire, chacune de ces sources d'erreur seront étudiées.

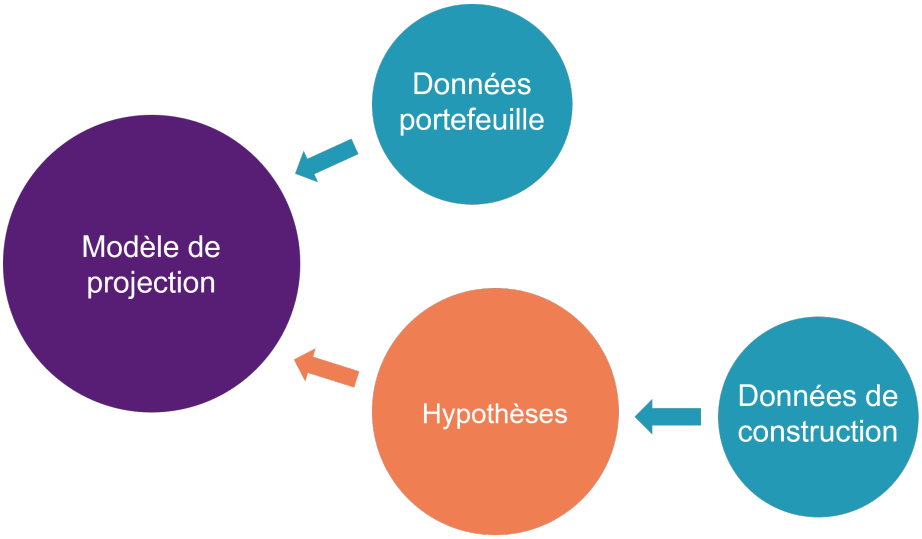


FIGURE 1.1 – Sources d'erreurs sur l'évaluation du *Best Estimate*

L'étude générale porte sur l'assurance des emprunteurs, laquelle peut être séparée en plusieurs types de prêts. Nous nous concentrerons sur les prêts personnels, appelés également crédits à la consommation (cf. figure 1.2).

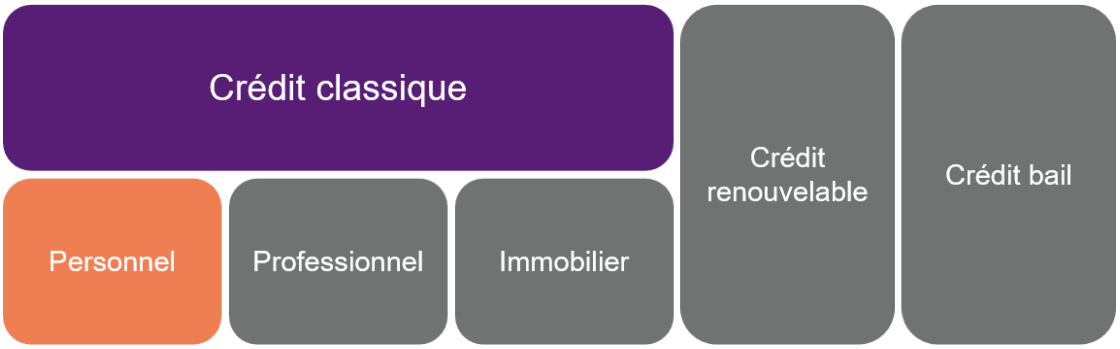


FIGURE 1.2 – Types de prêts en assurance des emprunteurs

Le prêt personnel peut être souscrit pour tout type d'usage et le client n'est pas tenu de justifier son achat. Il est réservé aux particuliers. Bien que la loi n'impose pas au client de souscrire une assurance, la banque qui prête exige systématiquement à l'assuré de souscrire un contrat d'assurance auprès de l'assureur de son choix. Cela permet à la banque d'être protégée en cas de décès, d'invalidité, d'arrêt de travail et éventuellement de perte d'emploi de son client. Le remboursement du prêt est différent selon les garanties, celles-ci sont présentées ci-après :

- **Garantie décès et PTIA (Perte Totale et Irréversible d'Autonomie)**
  Ces deux types de garanties sont regroupés dans une même catégorie. En cas de décès de l'assuré à un âge inférieur à l'âge limite, le prêteur est **entièrement remboursé**. Il en est de même en cas d'invalidité extrême de l'assuré (PTIA). La garantie PTIA est actionnée lorsqu'à la suite d'un grave accident ou d'une grave maladie, l'assuré a besoin de l'assistance d'une tierce personne dans

au moins trois des quatre actes du quotidien suivants : se laver, s'habiller, se nourrir et se déplacer. Dans ce cas, l'assuré est directement rattaché à la catégorie PTIA, sans passer par la phase d'observation de son état d'une durée maximale de trois ans requise dans le cas d'un arrêt de travail.

- **Garantie ITT**
  Lorsque l'assuré est en arrêt maladie ou victime d'un accident du travail, le prêteur est alors remboursé via des **mensualités**. Cette garantie ITT (Interruption Temporaire de Travail) n'est plus valable lorsque l'assuré dépasse la limite d'âge imposée par l'assureur ou en cas de départ en retraite.

  Une franchise de trois mois est également applicable, c'est-à-dire que le prêteur ne pourra commencer à être remboursé qu'une fois les trois premiers mois d'arrêt de travail écoulés.

  L'état de l'assuré est contrôlé pendant une période d'observation pouvant s'étaler sur trois ans maximum. On parle alors d'incapacité sur cette période. Lorsque son état est consolidé, c'est-à-dire que sa situation ne se dégrade ou ne se détériore plus, un taux d'invalidité est alors défini. Ce taux permet de déterminer le degré de prise en charge par la société d'assurance de l'assuré et le passage en invalidité.

- **Garanties IPT et IPP**
  L'invalidité est composée de deux états possibles. Si le taux est supérieur à 66%, l'assuré ayant souscrit l'option se trouve en IPT (Invalidité Permanente Totale). Si le taux est compris entre 33% et 66%, l'assuré se trouve en IPP (Invalidité Permanente Partielle). En deçà, on ne parle pas d'invalidité.

  Dans le cas où l'invalidité est prononcée, le capital restant dû est **entièrement remboursé**.

- **Garantie perte d'emploi**
  Lorsque l'assuré perd son emploi, le prêteur est remboursé sous forme de **mensualités**, selon certaines conditions définies par l'assureur dans le contrat.

A noter que dans notre cas, la garantie arrêt de travail (AT) est traitée au moyen d'une unique modélisation regroupant l'ensemble des garanties ITT, IPT et IPP, et la garantie perte d'emploi n'est pas modélisée.

Lors de la souscription d'un emprunt, un tableau d'amortissement est établi pour déterminer les mensualités de remboursement du client à l'établissement financier. Il inclut les intérêts et d'éventuels frais supplémentaires. Chaque période de remboursement fait état du capital restant à payer, appelé **capital restant dû**.

Par ailleurs, l'assuré a la possibilité de décaler le remboursement de certaines mensualités et/ou du taux d'intérêt ; on parle alors de **différé**. Il en existe deux types : le différé total et le différé partiel. Le différé total signifie qu'aucune partie du capital et aucun taux d'intérêt ne sont payés pendant une certaine période. Ils sont ainsi ajoutés au capital restant dû. Il est donc possible que celui-ci devienne plus élevé que le capital initial. Le différé est dit partiel lorsqu'il porte uniquement sur

le capital restant dû. Dans ce cas, seuls les intérêts sont payés par l'emprunteur à la date convenue initialement et le capital restant dû reste stable.

Concernant la tarification du contrat d'assurance, un **taux de prime** d'assurance est calculé en fonction du profil de l'assuré et varie selon les garanties comprises dans le contrat. Le contrat peut être tarifé sur le capital initial ou sur le capital restant dû. Dans le premier cas, la cotisation reste stable tout le long de la durée de vie du prêt. Dans le second cas, le capital restant dû fait évoluer les cotisations à chaque pas de temps ; on parle alors de cotisation variable ou dégressive. Notons que les taux de prime sont bien plus élevés pour les prêts personnels que pour les prêts immobiliers. En effet, du fait que le crédit à la consommation est plus accessible, la moyenne des niveaux de vie des emprunteurs est plus basse que pour les crédits immobiliers.

La prime d'un contrat d'assurance est définie en fonction du taux de prime, du capital assuré et de la quotité assurée. La quotité assurée correspond à la part du contrat qui est assurée sur une personne. Elle est comprise entre 0 et 1, et peut par exemple être strictement inférieure à 1 dans le cas où plusieurs personnes se sont jointes pour souscrire un emprunt. Par ailleurs, une commission est prélevée sur la prime de l'assuré et est reversée à l'établissement prêteur. Le **taux de commission** est déterminé par l'assureur et peut varier d'une année sur l'autre. On parle alors de **génération tarifaire du taux de commission**.

En résumé, la présente étude concernera les prêts personnels en assurance des emprunteurs avec les garanties suivantes : décès, PTIA, IPT, IPP, arrêt de travail et perte d'emploi. Cette étude s'inscrit dans le contexte réglementaire de Solvabilité 2 présenté par la suite.

1.2 Contexte réglementaire

La directive Solvabilité 2 a pour but principal de protéger les assurés contre la faillite de leur assureur. Elle assure la solvabilité des sociétés d'assurance dans le cas d'évènements catastrophiques survenant tous les 200 ans. Dans cette section, nous présenterons brièvement cette directive puis nous nous attarderons davantage sur le calcul du *Best Estimate* qui constitue la majeure partie des provisions de Solvabilité 2.

1.2.1 Introduction à Solvabilité 2

Solvabilité 2 est en vigueur depuis le 1er janvier 2016 et est à l'initiative de la Commission Européenne. Elle s'impose à l'ensemble des compagnies d'assurance et de réassurance de l'Union Européenne, à quelques exceptions près. Ce régime prudentiel vise également l'harmonisation du marché de l'assurance européen pour développer sa transparence et sa compétitivité. Il adopte une vision en valeur de marché pour le calcul des actifs et des passifs.

D'autres acteurs interviennent dans la mise en oeuvre de cette directive, dont principalement l'EIOPA et l'ACPR.

L'EIOPA, l'autorité européenne des assurances et des pensions professionnelles, a un rôle de superviseur auprès des compagnies. Elle est en charge de transmettre à la Commission Européenne des méthodologies dans le calcul des risques. Elle propose aussi des études d'impact, appelées QIS, aux compagnies et les questionne au travers de *consultation papers*. De plus, elle est responsable des évolutions de la Formule Standard pour le calcul des fonds propres.

Quant à l'ACPR, l'autorité de contrôle prudentiel et de résolution, elle surveille au quotidien les compagnies et est tenue sous Solvabilité 2 d'appliquer des contrôles réglementaires.

Solvabilité 2 se décompose en trois piliers :

- Le premier pilier concerne la mise en valeur économique des passifs d'assurance. Il comprend les calculs d'exigence de capital et l'évaluation des fonds propres éligibles, correspondant aux provisions techniques, au SCR (*Solvency Capital Requirement* en anglais) et au MCR (*Minimum Capital Requirement* en anglais) ;
- Le deuxième pilier regroupe l'évaluation interne des risques et de la solvabilité (ORSA), le système interne de gouvernance de la compagnie d'assurance, l'audit interne et le contrôle interne ;
- Le troisième pilier énonce les principes de transparence, d'information publique, d'information pour les superviseurs et d'informations publiées par les superviseurs.

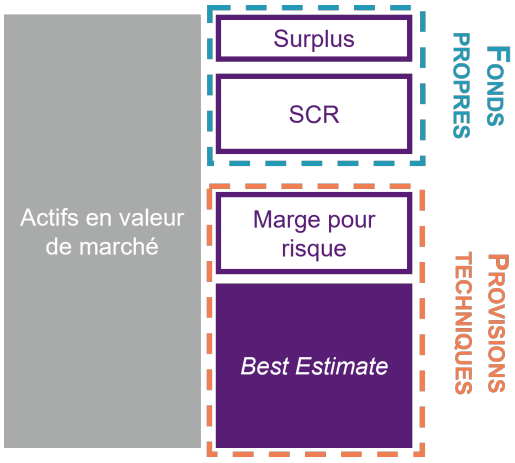


FIGURE 1.3 – Bilan Solvabilité 2

La fonction actuarielle est l'une des quatre fonctions clés énoncées par le pilier 2. Les missions de cette fonction seront présentées dans la section 1.3.

Rappelons que l'objectif de notre étude est de définir l'incertitude sur la valeur du BE. Les provisions techniques sous Solvabilité 2 sont constituées du BE et de la marge pour risque. L'un des intérêts de la marge pour risque est de compenser les erreurs d'estimation du BE. Cette marge pour risque s'exprime en fonction du SCR. La composition du BE peut à présent être expliquée.

1.2.2 Le *Best Estimate*

Le BE correspond à la moyenne pondérée par leur probabilité de survenance des flux de trésorerie futurs (*cash-flows* en anglais), compte tenu de la valeur temporelle de l'argent. Cette moyenne est estimée sur la base d'une courbe des taux pertinente. Ce BE est en fait la "meilleure estimation possible" des engagements de l'assureur vis-à-vis de ses assurés. Il se définit avec une approche *market consistent* qui signifie que l'évaluation du BE doit être représentative et cohérente avec le marché. Les valeurs des flux doivent être celles qui seraient utilisées si les flux étaient négociés sur le marché au moment de la projection.

Plus précisément, le BE est constitué du "BE de primes" et du "BE de sinistres", ayant chacun leurs propres flux comme illustré en figure 1.4. Notons que l'application présentée en section 5 sera faite sur le "BE de primes".

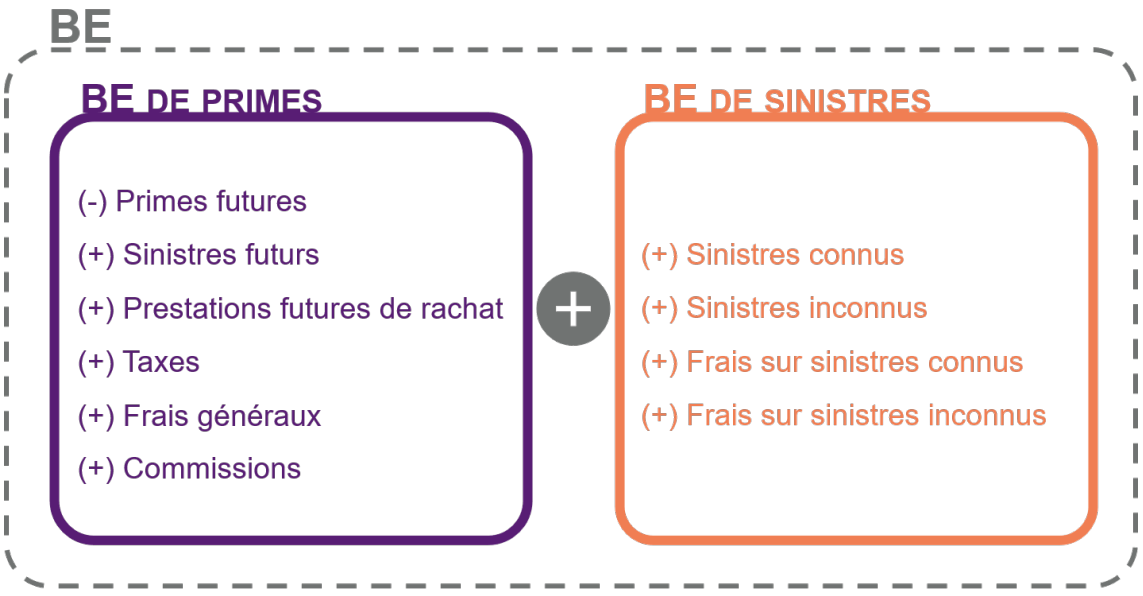


FIGURE 1.4 – Composition du *Best Estimate*

Un modèle de projection est utilisé pour l'estimation des flux futurs espérés. Le BE à l'instant 0 s'écrit alors :

$$BE_{0:T}(D_0, F, M, H) = \mathbb{E}^{\mathbb{P}\otimes\mathbb{Q}}[\Lambda_{0:T}(D_0, F, M, R)]$$

où

- $\Lambda_{0,T}(D_0, F, M)$ correspond à la valeur actuelle des flux futurs ;
- T correspond à la date de fin de projection ;
- D_0 désigne l'ensemble des données d'actifs et de passifs en entrée du modèle à la date 0 ;
- F désigne l'ensemble des règles de fonctionnement réglementaire et des règles contractuelles ;
- M désigne l'ensemble des décisions présentes et futures de gestion de l'assureur ;

- R désigne l'ensemble des risques couverts par l'assureur. Ces risques se séparent en deux parties, l'ensemble des risques du passif R_p et l'ensemble des risques de l'actif R_a ;
- H désigne l'ensemble des hypothèses permettant de couvrir l'ensemble des risques R.

$$R = R_p \cup R_a$$

Les flux $\Lambda_{0:T}(D_0, F, M, R)$ sont des variables aléatoires, pouvant dépendre de différents types de risques aléatoires stables ou variables dans le temps, et d'un cadre de fonctionnement.

$$\Lambda_{0:T}(D_0, F, M, R) = [\Lambda_0, ..., \Lambda_T]$$

À la date t, Λ_t est une variable aléatoire sous la probabilité historique $\mathbb{P}$ correspondant à la somme des flux découlant des primes, des sinistres, des frais et des commissions.

$$\Lambda_t(D_0, F, M, R_t|R_{t-1}) = \sum_{j \in L} CF_t^j(D_0, F, M, R_t|R_{t-1})$$

avec $L = \{-Primes, +Sinistres, +Frais, +Commissions, ...\}$. Elle s'écrit également de la façon suivante à l'instant t :

$$\Lambda_t(D_0, F, M,_t R|R_{t-1}) = f_t(D_0, F, M, R_t|R_{t-1})$$

avec f une fonction déterministe de $E \to \mathbb{R}^T$ ($E = \mathbb{R}$ ou $\mathbb{N}$).

Soit H l'ensemble des hypothèses utiles à la projection des provisions techniques, tel que $H = H_p|Y \cup H_a|A$, avec H_p les hypothèses du passif, H_a les hypothèses de l'actif, Y les données du passif et A les données de l'actif.

Si $R_p \sim \mathscr{L}^{R_p}(\theta^1, ..., \theta^i)$, avec $\mathscr{L}$ une loi quelconque de paramètres $\theta^1, ..., \theta^i$, alors $H_p|Y = \{\widehat{\theta^1}, ..., \widehat{\theta^i}\}$.

Si $R_a \sim Z_{0:T}$, où $Z_t \sim \mathscr{L}^{R_a}(\theta_t'^1, ..., \theta_t'^{i'})$ avec $\mathscr{L}$ une loi quelconque de paramètres $\theta_t'^1, ..., \theta_t'^{i'}$, alors $H_a|A = Z_{0:T}|\widehat{\theta_t'^1}, ..., \widehat{\theta_t'^{i'}}, \mathscr{F}_t$ où $\mathscr{F}_t$ représente la filtration.

Dans le cas d'un contrat emprunteur, la variable aléatoire des *cash flows* Λ_t dépend uniquement des risques du passif, son espérance P_t sous la probabilité historique s'écrit :

$$\begin{aligned} P_{0:T}(D_0, F, M, H_p|Y) &= \mathbb{E}^{\mathbb{P}}\left[\sum_{j \in L} CF_{O:T}^j(D_0, F, M, R)\right] \\ &= \mathbb{E}^{\mathbb{P}}[f_{0:T}(D_0, F, M, R, \theta^1, ..., \theta^i)] \\ &= f_{0:T}(D_0, F, M, R, \widehat{\theta^1}, ..., \widehat{\theta^i}) + \varepsilon_f \end{aligned}$$

avec $f_{0:T}$ représentant le modèle de projection et ε_f correspondant à l'ensemble des erreurs.

Or,

$$\mathbb{E}^{\mathbb{Q}}[P_{0:T}(D_0, F, M, H_p|Y)] = \sum_{k=0}^{T} \frac{P_k(D_0, F, M, H_p|Y)}{(1+r)^k}$$

où r est le taux sans risque. Ainsi,

$$BE_{0:T}(D_0, F, M, H) = \mathbb{E}^{\mathbb{Q}}[P_{0:T}(D_0, F, M, H_p|Y)] = \sum_{k=0}^{T} \frac{f_k(D_0, F, M, R, \widehat{\theta^1}, ..., \widehat{\theta^i}) + \varepsilon_f}{(1+r)^k}$$

En assurance des emprunteurs, il n'existe pas d'interaction entre l'actif et le passif, mis à part pour les rachats conjoncturels en crédit immobilier.

Deux types de rachats sont à différencier : les rachats structurels et les rachats conjoncturels. Les rachats structurels ont lieu chaque année en proportion relativement stable selon une loi de probabilité. En revanche, les rachats conjoncturels dépendent du comportement des assurés qui, lui, peut varier considérablement chaque année. La proportion de rachats conjoncturels peut être déterminée à partir de la différence entre le taux servi par les concurrents et celui servi par le contrat d'assurance. Si cette différence est positive et élevée, alors une partie des assurés a tendance à racheter leur contrat. Pour modéliser ces rachats, aussi appelés rachats dynamiques, un modèle stochastique est utilisé. Pour les prêts personnels, il existe très peu de rachats dynamiques. Ainsi, dans la présente étude, nous nous limiterons aux risques du passif. Le modèle de projection du BE sera donc dit déterministe.

Après avoir proposé une vision large de Solvabilité 2 et des provisions techniques, nous allons présenter les différentes missions de la fonction actuarielle.

1.3 La fonction actuarielle

La fonction actuarielle produit une fois par an un rapport actuariel qui doit être validé par l'AMSB (en anglais, *Administration, management or supervisory body*), l'organe d'administration, de contrôle et de gestion. Ce document contient tous les travaux effectués par la fonction actuarielle, ses résultats et éventuellement des recommandations vis-à-vis des défaillances détectées.

Chez BPCE Vie, le direction de la fonction actuarielle est indépendante et divisée en deux pôles :

- Le pôle "Provisions techniques" qui :
  - s'occupe de la coordination du calcul des provisions techniques ;
  - vérifie le caractère approprié des hypothèses économiques (calibrage du générateur de scénarios économiques...), biométriques (tables de mortalité...), comportementales (taux de rachat, remboursements anticipés) et des futures décisions de gestion ;
  - analyse les méthodologies de calcul (du BE par exemple) ;
  - effectue des contrôles de second niveau sur la qualité des données ;
  - compare les meilleures estimations aux historiques (*backtesting*) ;
  - contribue à la mise en œuvre du système de gestion des risques (évaluation ORSA...).
- Le pôle "Politiques et fonction actuarielle groupe" qui émet des avis sur :
  - la politique globale de souscription ;
  - l'adéquation des dispositifs de réassurance.

La présente étude a été effectuée au sein du pôle "Provisions techniques".

À présent que le contexte général de l'étude a été présenté, nous pouvons nous pencher sur la théorie de la détection d'anomalies.

1.4 Notions introductives sur les erreurs

Dans la présente étude, deux types d'erreurs se distingueront : les anomalies et les biais.

Une **anomalie** est un écart fini et quantifiable entre le résultat d'un calcul d'estimation et la valeur réelle d'un phénomène certain. Elle peut être évaluée par un taux d'anomalie, un écart relatif ou encore un écart absolu. Dans le cas où les anomalies seraient difficilement quantifiables, des indicateurs peuvent être utilisés pour se rendre compte de leur impact.

Le **biais** est un écart presque certain entre le résultat d'un calcul probabiliste utilisé pour approcher un phénomène aléatoire et la valeur de ce phénomène. Généralement, un intervalle de confiance peut être obtenu afin de contrôler un biais. Dans la suite de cette étude, tous les biais évoqués seront liés aux hypothèses.

Chapitre 2: Erreurs liées aux données utilisées

La détection d'anomalies dans les données est aujourd'hui un réel enjeu pour les compagnies d'assurance. Les anomalies peuvent être à l'origine d'importantes pertes et nuire fortement aux résultats de l'entreprise. De nombreux chercheurs se sont penchés sur le sujet et ont proposé des façons diverses et variées de traiter les anomalies. Dans cette étude, nous avons testé plusieurs méthodes et avons retenu celles qui semblaient les plus adaptées à notre base de données, laquelle sera présentée dans le chapitre 5.

Ce chapitre est dédié au cadre théorique de la détection des anomalies. Dans un premier temps, nous classifierons les erreurs sur les données. Dans un second temps, nous décrirons deux méthodes de réduction de dimension qui nous seront indispensables par la suite pour utiliser les algorithmes de détection d'anomalies. Nous étudierons ensuite plusieurs de ces algorithmes. Enfin, nous présenterons la loi de Benford qui est aujourd'hui beaucoup utilisée pour la détection d'anomalies dans les séries aléatoires.

2.1 Classification des erreurs sur les données

Une anomalie est une donnée erronée. On peut tenter d'identifier une anomalie par les observations qui s'écartent considérablement de l'ensemble des observations. Il peut s'agir, par exemple, de données atypiques pour la distribution de probabilité observée. Toutefois, l'identification certaine d'une anomalie demande généralement une vérification humaine pour comprendre le comportement atypique de la donnée identifiée.

La détection d'anomalies permet d'améliorer la qualité des données et ainsi diminuer l'erreur sur la valeur finale du BE.

Les données utilisées pour le calcul des provisions techniques se décomposent en deux parties :

$$D = Y_{n,m}(t) \cup A_{n',m'}(t)$$

avec

- $Y_{n,m}(t)$: matrice des données du portefeuille de contrats utilisées à une date t ;
- n : nombre d'individus du portefeuille ;
- m : nombre de variables utiles à la projection.
- $A_{n',m'}(t)$: matrice des données du portefeuille d'actifs utilisées à une date t ;
- n' : nombre d'actifs du portefeuille ;
- m' : nombre de variables utiles à la projection.

Dans la mesure où ces deux matrices de données sont considérées indépendantes, l'erreur relative sur les données peut être séparée en deux composantes additives :

$$\varepsilon_D = \varepsilon_{Y_{n,m}(t)} + \varepsilon_{A_{n',m'}(t)}$$

Ces deux erreurs ne sont pas aléatoires, mais finies et quantifiables. Elles sont distribuées de manière identique, de la façon suivante :

$$\varepsilon_{Y_{n,m}(t)} = \varepsilon_{Exhaustivité,Y}(n|t) + \varepsilon_{Exactitude,Y}(m|t) + \varepsilon_{Ag,Y}(K|t)$$

$$\varepsilon_{A_{n',m'}(t)} = \varepsilon_{Exhaustivité,A}(n'|t) + \varepsilon_{Exactitude,A}(m'|t) + \varepsilon_{Ag,A}(K'|t)$$

où

- $\varepsilon_{Exhaustivite,Y}(n|t)$ correspond à l'erreur sur l'exhaustivité des données (validité du périmètre de projection par rapport au modèle, aux hypothèses et aux données nécessaires au calcul) ;
- $\varepsilon_{Exactitude,Y}(m|t)$ correspond à l'erreur sur l'exactitude des données (validité des valeurs de chaque variable pour chaque individu de la base) ;
- $\varepsilon_{Ag,Y}(K|t)$ correspond aux erreurs dues à l'agrégation des données en *model point* par rapport aux variables $K \subset [1, m]$.

Pour notre part, comme expliqué dans la section précédente, nous nous occuperons seulement des données relatives au passif :

$$D = Y_{n,m}(t)$$

Le règlement délégué complétant la directive Solvabilité 2 [1] impose dans son article 19 le respect de trois critères sur la qualité des données utilisées dans le calcul des provisions techniques : l'exhaustivité, l'exactitude et le caractère approprié. Ce dernier critère n'étant pas quantifiable, il n'a pas été ajouté dans les formules de calcul des erreurs liées aux données telles que décrites précédemment.

2.1.1 Critère d'exhaustivité

Les données sont dites exhaustives si :

- aucun contrat hors périmètre n'est présent ;
- aucune donnée nécessaire au calcul des provisions techniques n'est pas utilisée de manière volontaire et sans justification (e.g. aucune activité d'assurance ne doit être oubliée), ou de manière involontaire (e.g. aucune défaillance de processus concernant la sélection des données) ;
- l'historique est suffisamment profond afin *"d'apprécier les caractéristiques des risques sous-jacents et de dégager des tendances d'évolution des risques"* [1].

Plus généralement, le critère d'exhaustivité est validé si toutes les informations nécessaires au calcul des provisions sont utilisées et contiennent suffisamment de détails et d'historiques. Pour s'assurer du bon respect de ces conditions, plusieurs indicateurs sont utilisés :

- L'absence de doublons peut être vérifié par un taux du nombre de doublons sur la clé c. La maille choisie peut être fine mais elle doit tout de même être assez grande pour que l'indicateur ait un sens.
- Il est également intéressant d'étudier la cohérence dans la volumétrie des contrats. La différence du nombre de contrats entre deux dates successives doit être faible et intégralement justifiable. Il est également possible de procéder à un simple contrôle de l'écart relatif entre le volume des contrats en t et celui en $t-1$. Les volumes peuvent aussi être vérifiés à partir d'une source indépendante.

Une fois que ce critère d'exhaustivité est validé, le critère d'exactitude des données peut être étudié. Chez BPCE Vie, le contrôle du critère d'exhaustivité est considéré comme applicable au niveau "table", c'est-à-dire que la qualité globale de la base de données est évaluée. À contrario, le critère d'exactitude s'évalue localement dans la base de données, par exemple en étudiant chacun des *model points* ou chacun des contrats, ou encore chacune des variables.

2.1.2 Critère d'exactitude

Les données sont dites exactes si :

- elles ne comportent aucune erreur ;
- elles sont enregistrées sans retard dans le temps par rapport à la date de calcul.

Il est possible de vérifier l'exactitude des données à partir des mesures suivantes :

- L'importance des données manquantes, qui peut être observée par un taux du nombre de valeurs manquantes sur une variable $k \in [1, m]$;
- Le poids des données aberrantes, qui peut être étudié par un simple taux du nombre de données aberrantes sur une variable $k \in [1, m]$ (e.g. âge d'un assuré supérieur à la limite autorisée réglementairement) ;
- Le poids des données incohérentes, qui peut être analysé par un simple taux du nombre de données incohérentes sur les variables de $K \subset [1, m]$ (e.g. âge d'un assuré très supérieur à l'âge légal de retraite alors qu'il fait partie de la catégorie des professionnels).

Ces indicateurs peuvent être évalués sur une maille fine à la condition que cette maille soit suffisamment large pour que l'évaluation reste cohérente. Ensuite, le dernier critère défini par le règlement délégué, i.e. le caractère approprié des données, doit être vérifié.

2.1.3 Caractère approprié

Les données sont considérées comme appropriées si :

- elles sont adaptées au calcul des provisions techniques ;
- leur volume et leur nature ne conduisent pas à une estimation trop erronée ;
- elles sont en cohérence avec les hypothèses ;
- elles reflètent correctement les risques de l'entreprise ;
- elles sont collectées, traitées et appliquées de manière transparente et structurée ;
- elles sont utilisées de manière cohérente dans la durée.

Pour contrôler le respect du caractère approprié, des études de passage doivent être effectuées. Ces dernières ne seront pas détaillées ici.

Enfin, notamment pour optimiser les temps de calcul, les données sont généralement agrégées. En contrepartie, cela implique des erreurs sur l'estimation finale du BE qui doivent être maîtrisables et faibles.

2.1.4 Erreurs liées à l'agrégation des données

L'agrégation d'un portefeuille consiste à regrouper par classes homogènes les facteurs de risque. Concernant le portefeuille des contrats $Y_{n,m}(t)$, il s'agit de classifier les contrats par groupes selon des caractéristiques diverses (niveau des cotisations, niveau des prestations, garanties ...), puis d'attribuer à chacun de ces groupes un contrat représentatif.

L'article 35 du règlement délégué [1] précise que la projection des flux de trésorerie doit être faite séparément pour chacun des contrats. Cependant, le règlement délégué

indique également que l'agrégation de polices par *model point* est admise si ces polices et les risques sous-jacents sont similaires, et si cela n'introduit pas de biais significatif dans le profil de risque du portefeuille.

Les impacts sur le calcul du BE peuvent être observés à partir de l'écart relatif entre le BE obtenu avec agrégation des données et le BE obtenu sans agrégation. Cet écart est considéré comme un taux d'agrégation.

Le taux de compression est également intéressant à étudier. Il s'agit de l'écart relatif entre le volume des données initiales et le volume des données qui ont été agrégées dans une optique d'optimisation des performances de calcul.

Enfin, pour l'ensemble des anomalies (détectées à partir des critères d'exhaustivité et d'exactitude et les erreurs d'agrégation), il est nécessaire d'observer une cohérence dans :

- l'évolution temporelle des données : cela peut être fait par de simples écarts relatifs entre les statistiques des données à la date t et celles en $t-1$;
- dans les statistiques des données utilisées par les différentes équipes de la société d'assurance : des écarts relatifs peuvent être utilisés.

Tous ces indicateurs présentés précédemment sont ensuite comparés à des seuils qui permettront de valider ou non les données utilisées.

2.1.5 Seuils des anomalies

Pour les anomalies liées à l'exhaustivité et à l'exactitude des données, la littérature scientifique fait état de certains seuils :

- En dessous de 5%, l'impact sur un modèle est extrêmement marginal et ne nécessite pas de validation ;
- Entre 5% et 25%, l'impact sur le modèle est matériel et il convient d'analyser les données pour remédier à la situation ;
- Au-dessus de 25%, l'impact est jugé fort. Les données sont alors rejetées et le traitement est différé.

Lorsque les taux sont supérieurs au seuil de 25%, les données sont redressées. Par exemple, un âge moyen peut être utilisé pour une certaine catégorie d'assuré. Dans ce cas, il est nécessaire de documenter correctement ce redressement afin de pouvoir évaluer par la suite le biais qu'il entraîne lors de l'utilisation des données redressées.

Concernant les erreurs liées à l'agrégation des données, les seuils suivants pour l'indicateur paraissent appropriés :

- En dessous de $0,5\%$, l'agrégation des données est très satisfaisante et validée ;
- Entre $0,5\%$ et 1%, l'agrégation des données est quelque peu correcte et par conséquent une analyse sur les causes de cet écart est à réaliser ;

- Au dessus de 1%, l'agrégation des données est jugée mauvaise et doit faire l'objet d'un retraitement.

Le tableau 2.1 récapitule l'ensemble des anomalies sur les données avec leurs indicateurs de contrôles et leurs seuils associés. Les notations utilisées sont détaillées en annexe A.

Type d'erreur	Indicateur	Seuil de validation immédiate des données	Borne nécessitant une analyse des données	Seuil de rejet des données
$\varepsilon_{Exhaustivite}(n\|t)$	$\tau[doublons\ sur\ c]$ $\Delta_r[Stat_t, Stat_{t-1}]$ $\Delta_r[Volume_t, Volume_{t-1}]$ $\Delta_r[Stat_\circ, Stat_{metier}]$	$\leq 5\%$	entre 5% et 25%	$\geq 25\%$
$\varepsilon_{Exactitude}(m\|t)$	$\tau[k\ manquant]$ $\tau[k\ aberrant]$ $\tau[k\ incoherent]$ $\Delta_r[Stat_t, Stat_{t-1}]$			
$\varepsilon_{Ag}(K\|t)$	$\Delta_r[BE_{ag}, BE]$	$\leq 0,5\%$	entre $0,5\%$ et 1%	$\geq 1\%$
	$\Delta_r[Stat_\circ, Stat_{metier}]$	$\leq 5\%$	entre 5% et 25%	$\geq 25\%$

TABLE 2.1 – Synthèse des indicateurs et seuils des erreurs sur les données

Nous nous concentrerons dans ce document sur les anomalies liées à l'exhaustivité et à l'exactitude des données. Pour cela, nous utiliserons différentes techniques d'apprentissage des données qui seront présentées par la suite. Avant de les étudier, nous allons présenter quelques méthodes de réduction de dimension des données.

2.2 Réduction de dimension des données

La dimension d'une base de données correspond au nombre de variables contenues dans cette base. Considérons qu'une base de données est structurée sous forme de matrice. La réduction de dimension consiste alors à réduire le nombre de colonnes d'un jeu de données en gardant un maximum d'informations. Cela est très utile pour les jeux de données intégrant une grande quantité de variables. En effet, du fait que toutes ces variables représentent beaucoup d'informations, le coût de l'espace mémoire et du temps de calcul peuvent devenir très vite très importants. Par ailleurs, certaines variables ne sont pas forcément utiles à l'apprentissage des anomalies et peuvent également fausser les prédictions des algorithmes de détection d'anomalies. Ainsi, l'utilisation de ces algorithmes de réduction de dimension permet de supprimer ou de réduire l'impact négatif de certaines variables sur les résultats de l'apprentissage en se basant sur des critères objectifs et en prenant en compte les interactions entre les différentes variables.

2.2.1 Analyse en Composante Principale (ACP)

L'une des méthodes les plus utilisées pour réduire la dimension est l'Analyse en Composante Principale (ACP) (*Principal Components Analysis (PCA)* en anglais).

À partir du jeu de données initial, cette méthode produit un autre jeu de données de même dimension contenant plus d'informations dans la première colonne, puis un peu moins dans la deuxième colonne et encore moins dans la troisième colonne, etc.. Pour ce faire, une base orthonormée est construite de telle façon que la variance des données soit maximale sur chacun des axes. La figure 2.1 met en forme un exemple très simplifié avec un jeu de données initial à deux dimensions, dont les axes sont représentés en violet. Dans ce cas, les deux nouveaux axes formés par l'ACP sont représentés en orange ;

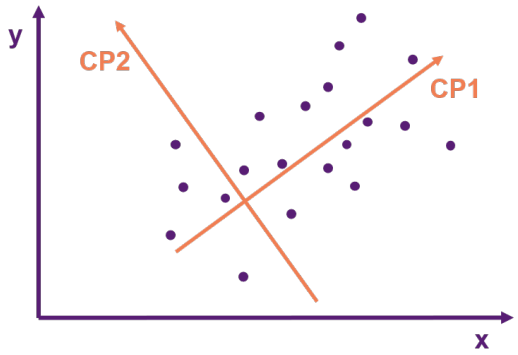


FIGURE 2.1 – Exemple du choix des axes en dimension 2 pour l'ACP

La démonstration mathématique de l'ACP est brièvement expliquée par la suite. Nous pourrons nous référer à l'article [2] de Jonathon Shlens pour plus d'explications.

Notons D la matrice des données, telle que $D \in \mathbb{R}^{n,m}$ avec n le nombre de contrats

et m le nombre de variables. Tout d'abord, la première direction v_1 est formée en maximisant la variance des données lorsqu'elles sont projetées sur cet axe. Les valeurs de la première colonne de la nouvelle base sont donc obtenues avec $c_1 = v_1^T D$.

Par simplification, supposons que la matrice D soit centrée. En réalité, il suffirait de soustraire à D son espérance si ce n'était pas le cas. La variance de c_1 devant être maximisée s'écrit alors :

$$V(c_1) = V(v_1^T D) = \mathbb{E}[v_1^T D - \mathbb{E}[v_1^T D]] = v_1^T \mathbb{E}[DD^T] v_1$$

Cela revient à maximiser $v_1^T \Sigma v_1$ avec Σ la matrice de covariance de D. Le vecteur v_1, appelé première composante principale de D, est le vecteur propre correspondant à la plus grande valeur propre de Σ. Une fois ce premier vecteur construit, le deuxième v_2 peut être formé en maximisant $v_2^T \Sigma v_2$ et en respectant une orthogonalité entre v_2 et v_1... et ainsi de suite pour tous les autres vecteurs v_i.

L'ACP est la méthode idéale lorsque les données sont linéaires. Cependant, elle supprime potentiellement des interactions complexes locales. Dans le cas où les données ne seraient pas homogènes, les transformations non linaires seraient plus adaptées. L'une de ces dernières est présentée ci-après.

2.2.2 UMAP

La deuxième méthode de réduction de dimension que nous allons étudier est appelée UMAP (*Uniform Manifold Approximation and Projection* en anglais). C'est une transformation non linéaire, de même que la méthode LLE (*Locally Linear Embedding* en anglais) proposée par Sam T. Roweis and Lawrence K. Saul en 2000, dans leur article *"Nonlinear Dimensionality Reduction by Locally Linear Embedding"* et la méthode t-SNE (*t-Stochastic Neighbour Embedding* en anglais). La méthode UMAP étant plus rapide que les deux méthodes précédentes, elle sera appliquée à nos données dans le chapitre 5.

Les transformations non linéaires se concentrent généralement sur la structure locale des données, contrairement aux transformations linéaires qui, elles, s'appuient sur leur structure globale. L'un des avantages de la réduction UMAP est qu'elle prend en compte les deux types de structure.

À partir d'un graphique initial à hautes dimensions, i.e. une base de données, UMAP recherche un graphique à plus faible dimension qui soit le plus ressemblant possible au premier. La forme et la typologie du graphique inital sont donc approximées. Par la suite, nous tenterons d'expliquer simplement la démarche qu'utilise UMAP. C'est une méthode qui a été prouvée mathématiquement et pour plus de théorie, le lecteur peut se référer à l'article [3] de Leland McInnes, John Healy et James Melville.

Globalement, le but est de retrouver la forme des données en connectant tous les points entre eux. Pour ce faire, des cercles de rayons identiques sont tracés autour de chaque point. Des points sont alors reliés si leurs cercles s'entrelacent. Ainsi dans les espaces de faible densité, les points auront très peu, voire pas de voisins, contrairement aux espaces de forte densité. Par conséquent, il n'est pas possible de

relier tous les points en gardant un rayon de taille fixe quelle que soit la densité de l'espace analysé. C'est pourquoi UMAP utilise une variable pour définir la taille des rayons. Comme le montre la figure 2.2, dans une zone de grande densité les rayons deviennent plus faibles et lorsque des points sont très éloignés des autres, leur rayon devient plus grand. Notons que sur cette figure, le diamètre de chaque point est relatif à son éloignement sur l'axe z.

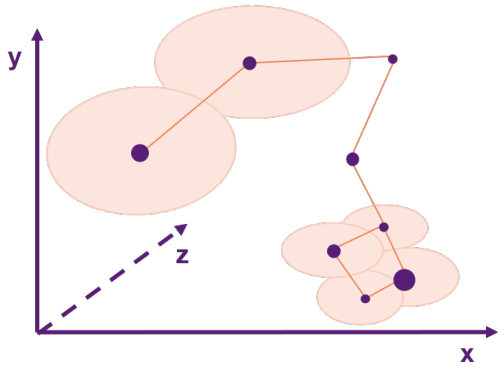


FIGURE 2.2 – Fonctionnement de UMAP en dimension 3

La densité de la variable du rayon est approximée à partir d'un paramètre k défini par l'utilisateur. Ce dernier correspond au nombre de voisins à trouver pour chaque point. Si k est grand, la structure globale est davantage conservée, à l'inverse si k est petit, la structure locale est mise en avant. Ce paramètre k représente donc la balance entre la structure globale et la structure locale. Il est généralement compris entre 2 et 100.

Selon UMAP, les connexions entre chaque point et leurs voisins se voient attribuer un poids, en d'autres termes une probabilité de connexion. Si les points sont éloignés les uns des autres, cette probabilité est plus faible et inversement. Pour obtenir une plus faible dimension, chacun de ces poids est étudié. Si les poids de points voisins sont élevés alors il est plus probable qu'ils restent ensemble dans une dimension plus faible.

Après avoir expliqué le fonctionnement des deux algorithmes de réduction de dimension de la base, l'ACP et l'UMAP, nous pouvons passer à l'étude des algorithmes de détection d'anomalies.

2.3 Algorithmes de détection d'anomalies

La littérature regroupe de nombreux travaux sur la détection d'anomalies, basés sur l'apprentissage, les statistiques ou encore les théories de probabilité. Dans ce document, nous nous intéresserons aux méthodes d'apprentissage automatique, *Machine Learning* en anglais, qui globalement font des prédictions sur la catégorie d'une observation : normale ou anormale. Ces méthodes d'apprentissage sont réparties en trois types : l'apprentissage supervisé, l'apprentissage non-supervisé et l'apprentissage semi-supervisé.

Apprentissage supervisé : pour évaluer l'anormalité d'une observation, les modèles supervisés s'appuient sur un ensemble de données qui sont déjà étiquetées, c'est-à-dire qu'elles font déjà partie d'une des deux catégories suivantes : soit la catégorie normale, soit la catégorie des anomalies. Elles sont également dites "labelisées" et portent le nom de données d'entraînement. Elles permettent de définir des profils de points normaux puis d'identifier les points anormaux à partir de ces profils. Des prédictions sont donc réalisées sur l'appartenance des données non étiquetées, dites données de test, à une catégorie. Ces prédictions sont présentées sous forme de probabilités.

Apprentissage non-supervisé : lorsqu'aucun jeu de données étiquetées n'est disponible, les méthodes non-supervisées sont alors utilisées. L'objectif est de rechercher des relations entre plusieurs données non étiquetées, et ainsi, former des groupes de données, appelés *clusters*. Pour ce faire, des distances entre des observations ou des distributions sont généralement calculées. On parle alors d'algorithmes de *clustering*.

Apprentissage semi-supervisé : les méthodes semi-supervisées reposent sur un ensemble de données dont une partie est étiquetée et l'autre non. Ces deux parties sont présentées en entrée de l'algorithme et permettent de classifier les données de test. Il s'agit là d'un mélange entre l'apprentissage supervisé et l'apprentissage non-supervisé.

Dans la prochaine partie, nous expliquerons les principes de quatre algorithmes de détection d'anomalies.

2.3.1 Apprentissage non-supervisé

Lorsqu'aucun jeu de données étiquetées n'est disponible, des algorithmes non-supervisés peuvent alors être utilisés. Nous allons en présenter trois d'entre eux.

2.3.1.1 K-moyennes

K-moyennes est l'un des plus simples algorithmes d'apprentissage non-supervisés. Cet algorithme nécessite un seul paramètre : k le nombre de *clusters* à construire. L'objectif est de définir k centroïdes, c'est-à-dire un centre pour chaque *cluster*, de telle sorte que les k *clusters* ne se chevauchent pas et soient les plus éloignés possible les uns des autres. Pour ce faire, k points sont sélectionnés au hasard parmi les données et sont initialisés en tant que centroïdes. Puis chaque point restant est attribué à l'un des centroïdes de façon à ce que la somme J des distances au carré entre les points du *cluster* et le centroïde soit minimale :

$$J(V) = \sum_{i=1}^{k} \sum_{j=1}^{k_i} \|x_{i,j} - v_i\|^2$$

avec

- $V = \{v_1, ..., v_k\}$ l'ensemble des centroïdes ;
- k_i le nombre de points dans le i^e *cluster* ;
- $(x_{i,j} : j \in \{1, ..., k_i\})$ les points appartenant au i^e *cluster* ;
- $\|.\|$ la distance euclidienne.

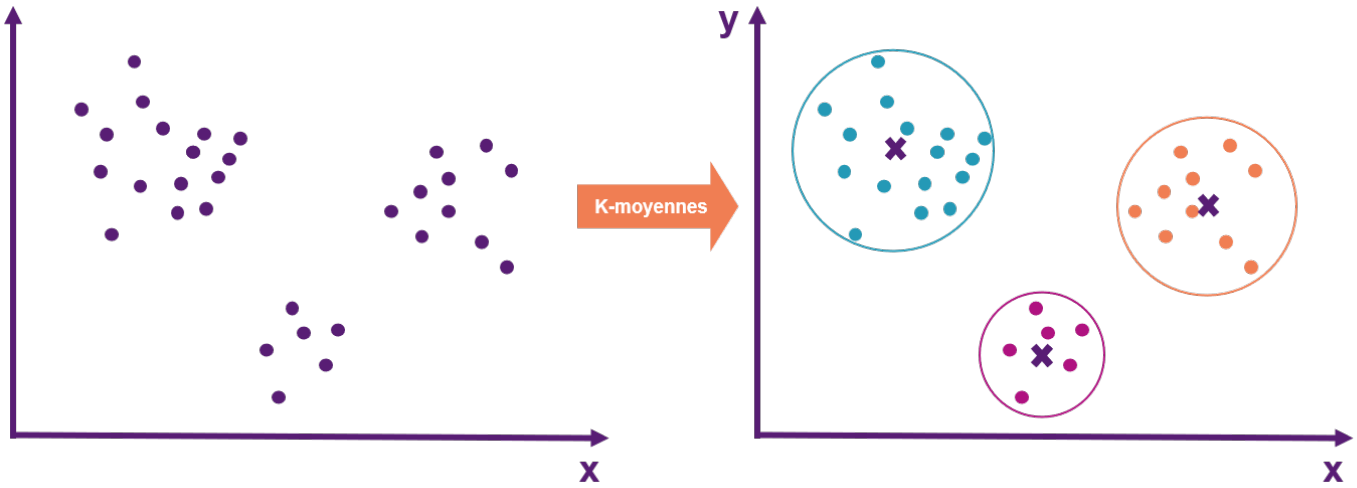


FIGURE 2.3 – Fonctionnement de k-moyennes

Il convient alors de calculer toutes les distances entre les points et les centroïdes. Ensuite les centroïdes sont remplacés par la moyenne de tous les points appartenant à leur *cluster*. Puis les autres points "non centroïdes" sont ré-attribués à l'un des nouveaux centroïdes. Ce processus est réitéré jusqu'à ce que l'affectation des points aux *clusters* ne change plus.

Ce procédé de *clustering* est simple et très efficace si des groupes de données sont facilement reconnaissables. Tous ses avantages et inconvénients sont résumés dans le tableau 2.2.

Avantages	Inconvénients
Rapide, robuste et simple	Difficulté à calibrer le nombre de clusters k
Complexité $O(kTn)$ avec T le nombre d'itérations et n le nombre de points	Choix aléatoire des premiers centroïdes
	Ne permet pas de créer des clusters avec des formes complexes

TABLE 2.2 – Avantages et inconvénients de K-moyennes

Nous n'avons jusqu'ici aucunement fait mention de la sélection des anomalies parmi tous les *clusters* formés, laquelle représente un processus délicat. La tendance serait de considérer ces anomalies comme correspondant aux groupes les moins denses puisqu'elles devraient très peu ressembler aux points normaux et donc être plus

écartées d'eux. Dans ce cas-là, la structure globale des données est prise en compte. Cependant, il est possible que plusieurs anomalies soient regroupées dans un groupe plus dense que des groupes ne contenant que des points normaux. Auquel cas, des groupes de points normaux pourraient être considérés comme anormaux.

Nous retiendrons que cette méthode peut nous permettre de détecter facilement des anomalies globales. Néanmoins, pour plus d'efficacité, il serait intéressant de l'associer à d'autres algorithmes de *clustering* utilisant des méthodes de *clustering* différentes, à commencer par *Isolation Forest*.

2.3.1.2 *Isolation Forest*

En 2008, KM. Ting et ZH. Zhou proposent dans leur article [4] une nouvelle approche non-supervisée appelée *Isolation Forest* (IF), basée sur l'isolement des anomalies. L'objectif d'IF est de définir un score pour chacun des points mesurant leur anormalité. Plus le score est élevé, plus le point semble atypique. Pour déterminer ces scores, IF partitionne aléatoirement les données. Pour cela, l'un des n attributs, i.e. variables, est choisi de manière aléatoire, puis une valeur y est tirée aléatoirement. À la suite de cette première découpe, une nouvelle variable est choisie, suivi d'un nouveau découpage... et ainsi de suite. Ces découpes peuvent être représentées par un arbre de décision où le nombre de coupures correspond au chemin parcouru de la racine à la feuille, comme illustré en figure 2.4. À l'inverse de la plupart des méthodes de *clustering*, dans la mesure où IF ne calcule aucune distance entre des points et n'utilise aucune mesure de densité, cela réduit considérablement le temps d'exécution de l'algorithme.

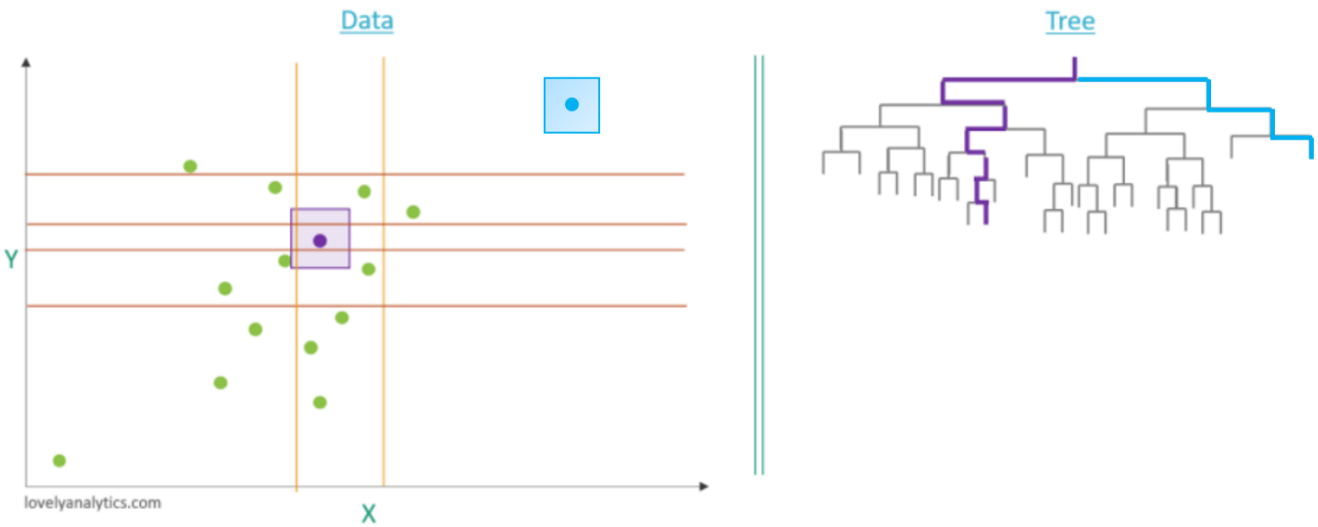


FIGURE 2.4 – Fonctionnement d'*Isolation Forest* (source : lovelyanalytics.com)

Le score d'anomalie d'un point x se base sur le chemin parcouru depuis le noeud racine jusqu'au noeud feuille, qui est noté $h(x)$. Plus le chemin est court, plus le

point est susceptible d'être une anomalie et donc plus son score est élevé. En effet, si un point est éloigné des autres points, il se retrouvera très vite isolé lors du partitionnement. Ainsi, IF interprète une anomalie comme étant un point écarté des autres points. La figure 2.4 illustre bien cela : le point bleu correspond à une anomalie et le point violet est considéré comme normal puisque son chemin est long.

Pour mieux estimer les anomalies, n arbres de cette sorte sont construits avec des partitions différentes. Le score final pour chaque point se calcule à partir de la moyenne des scores obtenus pour chacun des arbres. Plus précisément, il s'écrit de la façon suivante :

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

avec

- $E(h(x))$ la valeur moyenne de $h(x)$ sur tous les arbres ;
- $c(n)$ une valeur dépendant de n le nombre d'arbres.

La valeur de $c(n)$ se définit ainsi :

$$c(n) = \begin{cases} 2H(n-1) - \frac{2(n-1)}{n} & \text{si } n > 2 \\ 1 & \text{si } n = 2 \\ 0 & \text{sinon} \end{cases}$$

où $H(i) = ln(i) + \gamma$ avec γ la constante d'Euler.

Notons qu'un petit nombre d'arbres est suffisant pour assurer la convergence de la moyenne des scores i.e. pour que la moyenne des scores soit stable.

Avantages	Inconvénients
Très rapide	Ne permet pas de reconnaître les anomalies locales
Faible complexité, requiert peu de mémoire	Scores difficilement interprétables lorsque leurs minima et maxima sont proches
Facilité à calibrer le paramètre n	

TABLE 2.3 – Avantages et inconvénients de la méthode *Isolation Forest*

La grande qualité d'IF est sa rapidité d'exécution. De plus, elle fournit une mesure d'erreur pour les points, i.e. le score d'anomalie, lequel sera d'ailleurs utilisé dans l'application présentée au chapitre 5, dans la mise en oeuvre de l'algorithme de détection semi-supervisé *PU learning*. Les avantages et inconvénients de la méthode IF sont résumés dans le tableau 2.3.

Les algorithmes IF et k-moyennes analysent tous deux la structure globale des données, il serait donc intéressant d'étudier un algorithme qui adopte une approche plus locale. Nous avons choisi pour cela HDBSCAN.

2.3.1.3 HDBSCAN

L'algorithme HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise* en anglais) est une amélioration de DBSCAN. C'est une version hiérarchique qui n'a qu'un seul paramètre en entrée, $minPts$ le nombre de points minimum dans un *cluster*, contrairement à DBSCAN qui possède deux paramètres, $minPts$ et ϵ le rayon des *clusters*. HDBSCAN optimise ϵ localement sur chaque sous-branche du dendrogramme qu'il construit. Cette méthode est donc plus puissante que la méthode DBSCAN, tout en reprenant plusieurs de ses concepts.

Afin de bien comprendre le fonctionnement de cet algorithme, plusieurs termes doivent être définis. Pour plus d'informations, le lecteur peut se référer à la thèse de Courjault-Rade [5].

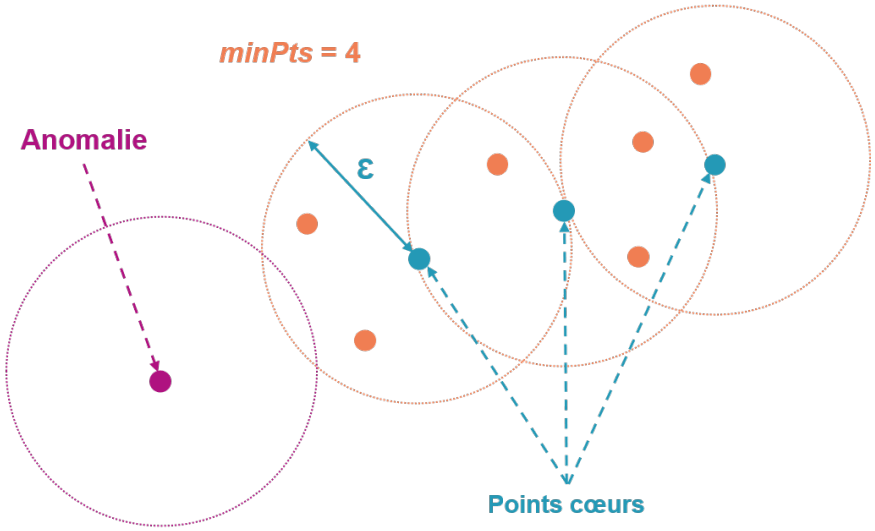


FIGURE 2.5 – Formation des clusters par HDBSCAN

Point coeur : comme illustré en figure 2.5, x_p est un point coeur si son nombre de voisins dans un rayon ϵ est supérieur ou égal à $minPts$, c'est à dire :

$$|N_\epsilon(x_p)| \geq minPts$$

avec

$$N_\epsilon(x_p) = \{x \in D | d(x, x_p) \leq \epsilon\}$$

avec d la distance euclidienne.

ϵ-atteignables : deux points x_p et x_q sont dits ϵ-atteignables si $x_p \in N_\epsilon(x_q)$ et $x_q \in N_\epsilon(x_p)$.

Densité-connectés : deux points coeurs sont dits densité-connectés s'ils sont directement ou de coeurs en coeurs ϵ-atteignables.

***Cluster* :** un *cluster* est un ensemble de points qui sont densité-connectés deux à deux.

Bruit : un point est un bruit s'il n'est pas un point coeur.

Distance-coeur : la distance-coeur d'un point x_p est la plus grande distance à ses voisins dans un rayon ϵ. Elle est notée $d_{coeur}(x_p)$.

ϵ-point coeur : un point x_p est un ϵ-point coeur si $d_{coeur}(x_p) < \epsilon$.

Distance d'accessibilité mutuelle : la distance d'accessibilité mutuelle (*mutual reachability distance* en anglais) entre deux points x_p et x_q se définit de la façon suivante :

$$d_{maccess}(x_p, x_q) = max(d_{coeur}(x_p), d_{coeur}(x_q), d(x_p, x_q))$$

Graphique d'accessibilité mutuelle : ce graphique G_{minPts} rassemble toutes les distances d'accessibilité mutuelle. Il a pour sommets tous les points de D, de sorte que chaque paire de sommets est connectée par une arrête ayant pour poids leur distance d'accessibilité mutuelle.

Rappelons que l'objectif d'HDBSCAN est de construire un dendrogramme. Tout d'abord, pour chaque paire de points (x_p, x_q), leur distance $d(x_p, x_q)$ est calculée. Puis, toutes les distances-coeurs $d_{coeur}(x_p)$ sont établies. Enfin, les distances d'accessibilité mutuelle $d_{maccess}(x_p, x_q)$ peuvent être déterminées pour chaque paire de points. À partir de ces dernières, le graphique d'accessibilité mutuelle est construit. Il contient $n(n-1)$ arêtes. Ce graphique est ensuite diminué en supprimant toutes les arêtes avec des poids supérieurs à ϵ. Ces dernières arêtes correspondent aux distances d'accessibilité mutuelle qui sont trop élevées et donc reflétant deux points trop éloignés l'un de l'autre. Au final, un arbre hiérarchique est obtenu.

Cet algorithme est très efficace et permet de détecter les anomalies locales. Ses avantages et inconvénients sont détaillés dans le tableau 2.4.

Avantages	Inconvénients
Un seul paramètre puissant et pertinent	Complexité $O(nlog(n))$, requiert beaucoup de mémoire vive
Très bonne gestion des clusters entremêlés et non-concentriques	
Idéal pour les données avec une structure hiérarchique	

TABLE 2.4 – Avantages et inconvénients de la méthode HDBSCAN

Les trois algorithmes de *clustering* présentés précédemment seront utilisés dans l'application présentée en chapitre 5. Nous allons maintenant étudier un dernier algorithme de détection d'anomalies.

2.3.2 Apprentissage semi-supervisé : cas du *PU learning*

Le *PU learning* est une méthode semi-supervisée qui s'appuie sur deux jeux de données :

- P (pour *Positive and labelled data* en anglais), correspondant à des données étiquetées normales, i.e. aucune anomalie n'y est présente ;
- U (pour *Unlabeled data* en anglais), correspondant aux données non étiquetées qui contiennent des données normales et des anomalies.

L'objectif est d'extraire parmi U les données qui sont négatives i.e. toutes les vraies anomalies. Pour cela, plusieurs techniques peuvent être utilisées : l'espion, la Cosine-Rocchio, la Rocchio, la Bayesian Naive, etc... Toutes ces techniques sont expliquées par A. Kaboutari, J. Bagherzadeh et F. Kheradmand dans leur article [6].

Prenons l'exemple de la méthode "espion". Elle se déroule de la façon suivante :

1. Des "espions" sont choisis aléatoirement dans la base P (15% de la base) ;
2. Les espions sont dès lors considérés comme non-étiquetés et sont injectés dans la base U. Deux nouvelles bases sont ainsi obtenues, P_s et U_s, telles que P_s correspond à la base P sans les espions et U_s correspond à la base U avec en plus les espions ;
3. L'algorithme d'apprentissage non-supervisé *randomForest* est exécuté sur $P_s + U_s$ (classification de 1^{er} degré) ;
4. Un seuil est déterminé pour sélectionner les négatifs fiables (NF) dans U à partir des prédictions de *randomForest* sur la base U_s contenant les espions. Par exemple, il peut correspondre au quantile à 15% de ces prédictions. A noter que le quantile doit être choisi de sorte qu'aucun espion ne se trouve dans les négatifs fiables. Tous les contrats dont la prédiction est supérieure au seuil déterminé sont considérés comme négatifs fiables ;
5. Les négatifs fiables sont extraits de U et forment une nouvelle base appelée NF ;
6. Un *randomForest* est à nouveau exécuté mais cette fois-ci sur $P_s + NF$ (classification de 2^{e} degré) afin d'obtenir des prédictions sur la nouvelle base non-étiquetée ;

Pour notre part, étant donné que nous n'avons pas de base de données étiquettée, nous utiliserons le *PU learning* pour vérifier que les anomalies que nous aurons trouvées à partir des algorithmes de *clustering* sont réellement des anomalies.

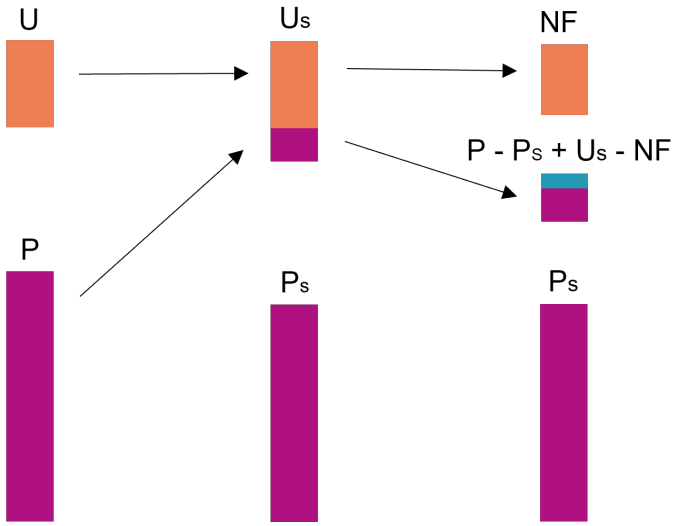


FIGURE 2.6 – Fonctionnement de la technique espion

2.4 Loi de Benford

La dernière technique de détection d'anomalies que nous allons étudier est la loi de Benford. En 1888, l'astronome canadien S. Newcomb découvrit que les premières pages des tables de logarithmes étaient plus abîmées que les autres. Cela signifiait que les nombres commençant par des petits chiffres tels que 1 ou 2, présents dans les premières pages des tables de logarithmes, étaient davantage regardés que les nombres commençant par des plus grands chiffres tels que 8 ou 9, présents dans les dernières pages des tables de logarithmes. L'article de Newcomb à ce sujet fut totalement ignoré jusqu'à ce qu'en 1938, F. Benford remarqua à son tour les traces d'usures sur les premières pages de tables numériques. En s'y intéressant davantage, il constata que pour certaines catégories de séries statistiques, la proportion de nombres avec 1 comme premier chiffre significatif était plus élevée que pour les nombres avec 2 comme premier chiffre significatif, et encore plus élevée que pour les nombres avec 3 comme premier chiffre significatif, etc...

Rappelons que le premier chiffre significatif du nombre 450 est 4 et le premier chiffre significatif de 0,0965 est 9.

F. Benford énonça la loi de Benford qui dit que la probabilité de trouver un nombre commençant par d, tel que $d \in \{1, 2, ..., 9\}$, est :

$$\mathbb{P}(X = d) = log_{10}(\frac{d+1}{d})$$

De même, il testa les probabilités d'apparition des deux premiers chiffres significatifs puis établit la loi suivante pour les deux premiers chiffres significatifs d_1d_2 d'un nombre, tels que $d_1d_2 \in \{10, 11, 12, ..., 99\}$:

$$\mathbb{P}(X = d_1d_2) = log_{10}(\frac{d_1d_2+1}{d_1d_2})$$

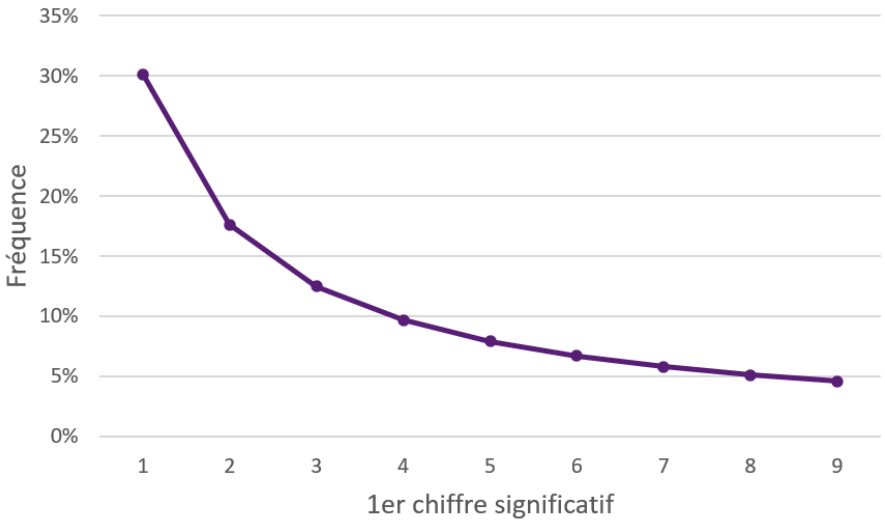


FIGURE 2.7 – Loi de Benford - Fréquence d'apparition du 1er chiffre significatif

Aujourd'hui cette loi est beaucoup utilisée pour détecter la fraude en assurance et peut également servir à repérer des anomalies dans un jeu de données. Néanmoins, elle ne s'applique qu'aux séries de données aléatoirement distribuées.

Les trois algorithmes de *clustering*, le *PU learning* et la loi de Benford, qui ont été détaillés dans cette section, seront par la suite appliqués aux données de BPCE Vie dans le chapitre 5.

Chapitre 3: Erreurs liées aux hypothèses

A présent que la détection des erreurs dans les données a été présentée, intéressons-nous au deuxième type d'erreur sur le BE, à savoir les erreurs liées à la construction des hypothèses utilisées dans le modèle de projection.

Dans un premier temps, une brève présentation de l'ensemble des hypothèses du modèle sera faite. Dans un second temps, chacune des erreurs potentielles sur l'estimation de ces hypothèses sera détaillée. Des indicateurs permettant de contrôler la qualité des hypothèses ainsi que des seuils de validation seront notamment présentés.

Parmi l'ensemble des hypothèses, nous avons retenu, pour l'application, les deux hypothèses qui impacteraient le plus la valeur du BE en cas de mauvaise estimation (cf. section 5). Il s'agit de la loi de mortalité et de la loi de rachat, aussi appelée loi de remboursement anticipé. Ainsi, la dernière partie de cette section vise à expliquer le fonctionnement et la construction de ces lois d'expérience, lesquelles sont estimées à partir de modèles de durée.

3.1 Présentation des hypothèses

Les hypothèses sont des modèles statistiques permettant d'estimer un phénomène ou un risque particulier R sur une période T, à partir d'informations disponibles sur le passé et sous certaines contraintes $\mathscr{C}_l$, appelées les l "contraintes".

Ces hypothèses sont injectées dans le modèle de projection et sont utiles au calcul du BE. Elles se séparent en deux catégories : les hypothèses non-économiques et les hypothèses économiques.

Les hypothèses non-économiques contiennent :

- les hypothèses biométriques telles que la loi de mortalité, la loi de rachat, et la loi de maintien en arrêt de travail ;
- les hypothèses comportementales qui prennent en compte le comportement de l'assuré, telles que la loi de rachat dynamique.

Les hypothèses économiques sont, quant à elles, liées à l'environnement économique, comme les taux d'inflation.

Afin que les hypothèses représentent au mieux le portefeuille, elles sont construites à partir des données présentes au passif et des données présentes à l'actif. Elles peuvent s'écrire de la façon suivante :

$$H = H_p|Y \cup H_a|A$$

où

- $H_p|Y$ représente les hypothèses du passif conditionnées à la base historisée des clients Y ;
- $H_a|A$ représente les hypothèses de l'actif conditionnées à la base historisée des actifs A.

Chacune de ces hypothèses se décompose ainsi :

$$H_p|Y = g_p(Y_{n,m}|T, \mathscr{F}_T, \mathscr{C}_l^p) + \varepsilon_{H_p|Y}$$

$$H_a|A = g_a(A_{n',m'}|T, \mathscr{F}_T, \mathscr{C}_l^a) + \varepsilon_{H_a|A}$$

où

- g_p est l'ensemble des méthodes utilisées pour calibrer les hypothèses du passif ;
- g_a est l'ensemble des méthodes utilisées pour calibrer les hypothèses de l'actif ;
- $Y_{n,m}|T, \mathscr{F}_T, \mathscr{C}_l^p$ est la matrice de données du portefeuille des clients sur la période T, soumise aux conditions de marché observées $\mathscr{F}_T$ sur cette période, sous les contraintes des différentes méthodes statistiques $\mathscr{C}_l^p$;
- n est le nombre d'individus dans le portefeuille du passif ;
- m est le nombre de variables utiles à la construction des hypothèses du passif ;
- $\varepsilon_{H_p|Y}$ correspond aux erreurs liées aux hypothèses du passif ;
- $A_{n,'m'}|T, \mathscr{F}_T, \mathscr{C}_l^a$ est la matrice de données du portefeuille des actifs sur la période T, soumise aux conditions de marché observées $\mathscr{F}_T$ sur cette période, sous les contraintes des différentes méthodes statistiques $\mathscr{C}_l^a$;
- n' est le nombre d'actifs dans le portefeuille d'actifs ;
- m' est le nombre de variables utiles à la construction des hypothèses de l'actif ;
- $\varepsilon_{H_a|A}$ correspond aux erreurs liées aux hypothèses de l'actif.

Les erreurs sur les hypothèses sont distribuées ainsi :

$$\varepsilon_H = \varepsilon_{H_a|A} + \varepsilon_{H_p|Y} \pm \varepsilon_{Corr[H_a|A, H_p|Y]}$$

Bien que les différentes hypothèses sont généralement indépendantes les unes des autres, il est nécessaire de s'assurer de cela sur les risques majeurs couverts. Si les hypothèses étaient corrélées, alors l'erreur $\varepsilon_{Corr[H_a|A, H_p|Y]}$ liée à cette corrélation devrait être ajoutée après compréhension des interactions et des liens de causalité.

3.2 Classification des erreurs liées aux hypothèses

Dans la présente étude, seul le passif sera traité, comme expliqué à la fin de la section 1.2.2. Ainsi, les erreurs liées aux hypothèses se décomposent de la manière suivante :

$$\varepsilon_H = \varepsilon_{H_p|Y} = \varepsilon_{Y_{n,m}} + \varepsilon^p_{op} + \varepsilon^p_s + \varepsilon^p_{\Delta^{hyp}_{mod}}$$

avec

- $\varepsilon_{Y_{n,m}}$ représentant les anomalies sur les données de construction ;
- ε^p_{op} représentant les anomalies liées aux erreurs opérationnelles ;
- ε^p_s correspondant au biais d'estimation de l'hypothèse (ou biais statistique d'échantillonnage dû à la calibration de l'hypothèse) ;
- $\varepsilon^p_{\Delta^{hyp}_{mod}}$ correspondant au biais de modélisation de l'hypothèse.

Chacune de ces erreurs est présentée dans la section suivante. En fonction de leur source, elles correspondent soit à des anomalies, soit à un biais. Les définitions de ces deux types d'erreur sont présentées en section 1.4.

Dans un premier temps, les différentes anomalies doivent être contrôlées et faire l'objet de corrections sur les hypothèses dans le cas où le contrôle s'avérait être négatif.

Dans un second temps, si aucune anomalie n'a été détectée ou si les anomalies détectées ne semblent pas être assez considérables pour faire l'objet d'une correction, les différents biais peuvent être évalués. L'objectif est d'obtenir un intervalle de confiance final sur le paramètre estimé pour chacune des hypothèses.

Pour ce faire, deux démarches différentes peuvent être adoptées. Ou bien, les biais sont considérés indépendants les uns des autres, comme illustré dans la formule ci-dessus. Dans ce cas, les biais sont évalués séparément, puis les intervalles de confiance associés à chacun de ces biais sont additionnés pour obtenir l'intervalle de confiance final. Ou bien, les interactions entre les biais sont prises en compte et dans ce cas, il est nécessaire d'effectuer un *bootstrap* sur l'ensemble des groupes de risque construits pour chacun des biais. Ainsi, la dimension des échantillons *bootstrappés* obtenus est multiple et est égale au nombre de biais. Dans cette démarche, l'erreur sur l'hypothèse est appréciée dans sa globalité. Ainsi, il n'est pas possible d'identifier l'impact de chaque biais sur le biais total.

Il convient dès lors de détailler à travers les sections suivantes les éventuelles anomalies et les différents biais sur les hypothèses.

3.2.1 Anomalies sur les données de construction

Source : Les causes racines de ces anomalies sont les mêmes que celles décrites dans la section 2.1.

Indicateurs de contrôle : Ce sont les mêmes que ceux définis dans la section 2.1.

Seuils : Ce sont les mêmes que ceux définis dans la section 2.1.

3.2.2 Anomalies liées aux erreurs opérationnelles

Source : Ce sont des erreurs résultant de procédures internes inadaptées, de membres du personnel de la société d'assurance, et/ou de l'inadéquation ou de la défaillance des systèmes utilisés pour les hypothèses. Par exemple, l'erreur pourrait être induite par l'utilisation d'une hypothèse $t-1$ à la place d'une nouvelle hypothèse en t, ou encore par la saisie d'une valeur erronée en hypothèse à la place de sa valeur réelle.

Indicateurs de contrôle : Ces erreurs de natures humaines ou informatiques doivent être évaluées en continu pour détecter des situations anormales dans l'évolution des risques. Des analyses de passage peuvent également être faites pour justifier l'utilisation et la construction de chacune des hypothèses. 100% des hypothèses doivent être justifiées, sinon elles ne peuvent être validées.

Ces anomalies ne seront pas étudiées dans la présente étude. Pour plus d'informations concernant l'évaluation de ces anomalies, veuillez vous référer au mémoire de Tondolo L. [7], utilisant des méthodes d'analyse de sensibilité.

3.2.3 Biais d'estimation (ou biais statistique d'échantillonnage)

Source : Cette erreur résulte de la mauvaise représentation du portefeuille de passifs par l'échantillon de données utilisé pour construire l'hypothèse. Elle baisse à mesure que le volume des données historiques utilisées augmente. En d'autres termes, ce biais statistique est dû à la mutualisation imparfaite d'un risque.

Indicateurs de contrôle : Le biais statistique peut être contrôlé à partir d'un intervalle de confiance.

Dans la présente étude, rappelons que seules les lois de mortalité et de rachat seront étudiées. Le biais statistique associé à l'estimation de ces lois peut être évalué par l'intervalle de confiance d'une loi binomiale :

$$IC_{Binom} = \left[max(0, \hat{q} - Q_{0.95}\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}); min(1, \hat{q} + Q_{0.95}\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}) \right]$$

avec $\hat{q}$ l'estimateur, n le volume de l'échantillon et $Q_{0.95}$ le quantile à 95% d'une loi normale centrée réduite.

Seuils : La valeur estimée $\hat{q}$ est à comparer avec les bornes inférieure et supérieure de l'intervalle de confiance. L'hypothèse est validée ou non en fonction de l'écart relatif entre la valeur estimée $\hat{q}$ et les bornes de l'intervalle, selon l'exemple suivant relatif à la borne supérieure :

- Si $\Delta_r[IC_{B,sup}, \hat{q}] < 25\%$, la précision de l'estimation est jugée satisfaisante ;
- Si $25\% \leq \Delta_r[IC_{B,sup}, \hat{q}] < 100\%$, la précision de l'estimation est tout juste suffisante et nécessite une analyse complémentaire pour améliorer la modélisation ;
- Si $100\% \leq \Delta_r[IC_{B,sup}, \hat{q}]$, la précision de l'estimation est insuffisante et l'hypothèse n'est pas valide.

3.2.4 Biais de modélisation

Ce biais est dû au choix de la méthode implémentée, laquelle ne permet pas toujours de reproduire exactement la réalité.

Cette erreur liée à la modélisation résulte des quatre facteurs détaillés ci-dessous :

$$\varepsilon^{p}_{\Delta^{hyp}_{mod}} = \varepsilon_{\mathscr{H}^{p}_{l}} + \varepsilon^{p}_{\Delta^{hyp}_{t}} + \varepsilon_{\Delta[Y,Z']} + \varepsilon_{\Delta A \times Y}$$

avec

- $\varepsilon^{p}_{\Delta^{hyp}_{t}}$ correspondant au biais d'homogénéité du portefeuille ;
- $\varepsilon_{\Delta[Y,Z']}$ correspondant au biais temporel ;
- $\varepsilon_{\mathscr{H}^{p}_{l}}$ correspondant au biais lié à la rupture des contraintes ;
- $\varepsilon_{\Delta A \times Y}$ correspondant au biais lié aux interactions entre l'actif et le passif.

Dans la formule ci-dessus, les quatre biais sont considérés comme indépendants entre eux. Si, d'après une analyse, la corrélation entre certains biais s'avère être importante, alors un correctif doit être appliqué au biais final de modélisation $\varepsilon^{p}_{\Delta^{hyp}_{mod}}$.

3.2.4.1 Biais d'homogénéité du portefeuille

Source : Ce biais est dû au fait que le portefeuille utilisé pour calibrer les hypothèses ne représente pas correctement le risque actuel. En d'autres termes, il résulte de la dérive des profils des assurés d'un portefeuille et de l'hétérogénéité du portefeuille vis-à-vis d'un risque. Par exemple, si les proportions des catégories socio-professionnelles (CSP) du portefeuille ne sont pas constantes au fil du temps comme illustré en figure 3.1 et si elles ne sont pas égales face au risque, de mortalité ou de rachat par exemple, alors un biais d'homogénéité apparaît dans la construction des hypothèses.

Indicateurs de contrôle : Pour contrôler ce biais, un *bootstrap* sur des classes homogènes peut être effectué. Pour ce faire, l'échantillon de valeurs observées, dit base de construction, est partitionné en plusieurs blocs de classes homogènes appelés "groupes de risque", sous réserve qu'il soit possible d'obtenir plusieurs classes homogènes. Généralement, les groupes de risque sont formés selon la catégorie socio-professionnelle de l'assuré, ou encore selon sa situation géographique. Ensuite, plusieurs échantillons "*bootstrappés*" sont construits aléatoirement à partir de ces blocs pour obtenir une nouvelle base et ainsi une distribution de l'hypothèse, comme illustré en figure 3.2. Un intervalle de confiance peut être obtenu à partir de l'ensemble de ces échantillons "*bootstrappés*". Le biais d'homogénéité du portefeuille correspond à l'écart entre le paramètre estimé et les bornes de l'intervalle.

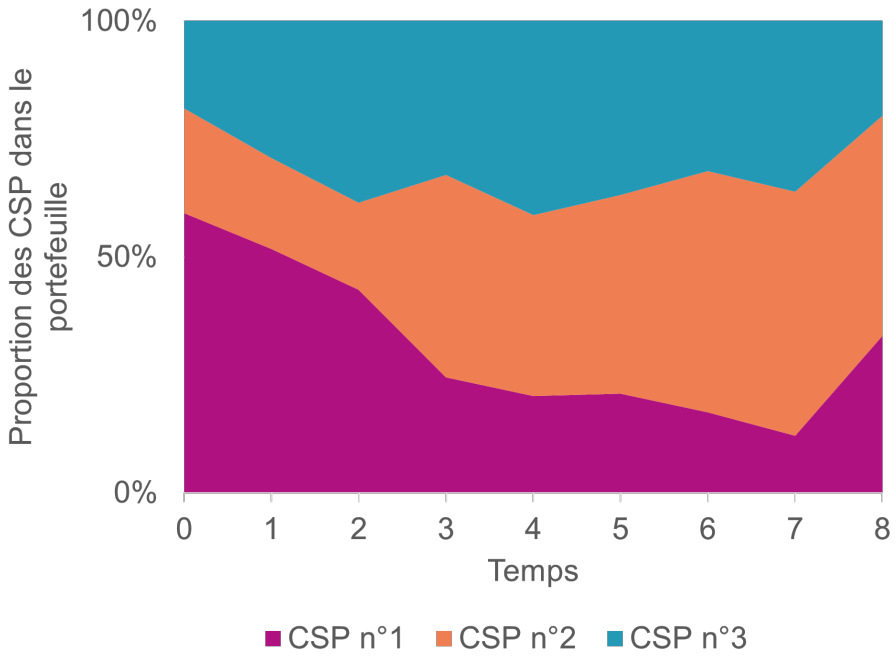


FIGURE 3.1 – Exemple d'évolution dans le temps des proportions de CSP dans le portefeuille

Deux méthodes peuvent être utilisées pour construire un intervalle de confiance à partir des B échantillons "*bootstrappés*". Notons q_B l'ensemble des paramètres obtenus à partir des B échantillons.

- Méthode des percentiles simples : Une fonction de répartition des paramètres q_B est construite. Les bornes inférieure et supérieure de l'intervalle de confiance à $1-\alpha\%$ correspondent aux quantiles à $1-\alpha\%$ de cette fonction de répartition.
- Méthode de l'erreur standard : La moyenne μ de q_B, l'erreur-type s de q_B et les quantiles $Q_{1-\alpha}$ de la loi Normale ou de la loi de Student sont utilisés pour construire l'intervalle de confiance à $1-\alpha\%$.

$$IC_B = [\mu - Q_{1-\alpha}s; \mu + Q_{1-\alpha}s]$$

Seuils : La valeur estimée $\hat{q}$ est à comparer avec les bornes inférieure et supérieure de l'intervalle de confiance. L'hypothèse est validée ou non en fonction de l'écart relatif entre la valeur estimée $\hat{q}$ et les bornes de l'intervalle, selon l'exemple suivant relatif à la borne supérieure :

- Si $\Delta_r[IC_{B,sup}, \hat{q}] < 25\%$, la précision de l'estimation est jugée satisfaisante ;
- Si $25\% \leq \Delta_r[IC_{B,sup}, \hat{q}] < 100\%$, la précision de l'estimation est tout juste suffisante et nécessite une analyse complémentaire pour améliorer la modélisation ;
- Si $100\% \leq \Delta_r[IC_{B,sup}, \hat{q}]$, la précision de l'estimation est insuffisante et l'hypothèse n'est pas valide.

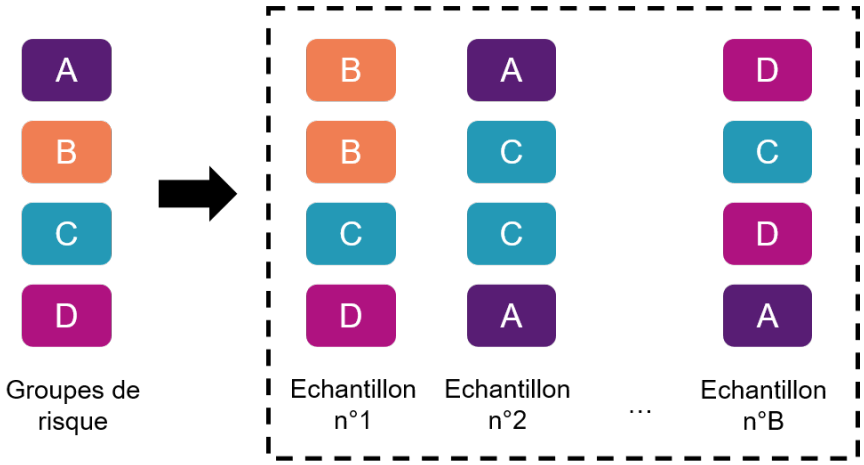


FIGURE 3.2 – Exemple de *bootstrap* avec quatre groupes de risque

3.2.4.2 Biais temporel

Source : Ce biais est lié aux évolutions du risque dans le temps. Par exemple, bien que le risque de décès évolue continuellement, la loi de mortalité ne peut être modifiée en temps réel.

Indicateurs de contrôle : De même que pour le biais d'homogénéité du portefeuille, ce biais peut être contrôlé en recourant au *bootstrap* temporel. Pour autant, aucune saisonnalité dans la série temporelle des données ne doit être constatée. Il s'agit de décomposer un échantillon de valeurs observées en plusieurs blocs indépendants constituant des segments temporels, appelés également "groupes de risque". Ensuite, B échantillons sont obtenus par tirage aléatoire avec remise dans l'ensemble de ces groupes de risque, comme illustré en figure 3.2. Enfin, un intervalle de confiance est obtenu à partir de ces échantillons "*bootstrappés*". Le biais temporel correspond à l'écart entre le paramètre estimé et les bornes de cet intervalle.

Seuils : La valeur estimée $\hat{q}$ est à comparer avec les bornes inférieure et supérieure de l'intervalle de confiance. L'hypothèse est validée ou non en fonction de l'écart relatif entre la valeur estimée $\hat{q}$ et les bornes de l'intervalle, selon l'exemple suivant relatif à la borne supérieure :

- Si $\Delta_r[IC_{B,sup}, \hat{q}] < 25\%$, la précision de l'estimation est jugée satisfaisante ;
- Si $25\% \leq \Delta_r[IC_{B,sup}, \hat{q}] < 100\%$, la précision de l'estimation est tout juste suffisante et nécessite une analyse complémentaire pour améliorer la modélisation ;
- Si $100\% \leq \Delta_r[IC_{B,sup}, \hat{q}]$, la précision de l'estimation est insuffisante et l'hypothèse n'est pas valide.

3.2.4.3 Biais lié à la rupture des contraintes

Source : Chaque contrainte doit être prise en compte dans la méthode statistique employée pour une hypothèse. Son absence a un impact plus ou moins élevé sur la valeur de l'hypothèse selon les risques. Une contrainte est une hypothèse faite pour

utiliser une méthode statistique permettant d'estimer une hypothèse en entrée du modèle de projection.

Indicateurs de contrôle : Il est nécessaire de vérifier si toutes les contraintes de la méthode choisie sont respectées, chaque contrainte non respectée pouvant entraîner des effets particuliers. Pour ce biais, il est plus difficile d'obtenir un intervalle de confiance. Toutefois, pour certaines contraintes, des indicateurs spécifiques peuvent être calculés. Par exemple, l'utilisation de l'estimateur de Kaplan-Meier nécessite une homogénéité dans le portefeuille et aucune évolution du risque dans le temps. Ces deux contraintes peuvent respectivement être évaluées à partir du biais d'homogénéité et du biais temporel précédemment présentés.

Seuils : Ce biais peut être évalué en fonction du pourcentage de contraintes respectées :

- Si toutes les contraintes sont vérifiées, l'hypothèse n'a pas de biais de contrainte et est validée pour ce biais ;
- Si 0% à 75% des contraintes ne sont pas vérifiées, une étude d'impact doit être réalisée ;
- Si plus de 75% des contraintes ne sont pas vérifiées, l'hypothèse doit être modifiée.

3.2.4.4 Biais lié aux interactions entre l'actif et le passif

Source : Ce biais est essentiellement présent dans les hypothèses du passif, et notamment les hypothèses comportementales. Par exemple, les rachats dynamiques sont liés aux évolutions des taux sur le marché, comme expliqué dans la section 1.2.2. Les interactions entre le marché et le portefeuille du passif sont délicats à estimer et sont sources d'erreurs.

Indicateurs de contrôle : Afin d'évaluer ce biais, tous les risques concernés doivent être identifiés et faire l'objet d'une analyse sur l'effet des interactions entre l'actif et le passif. Pour ce faire, des séries temporelles peuvent être utilisées. Cependant, ces risques ne sont généralement pas linéaires et demandent de la prudence dans leur traitement.

Seuils : Après une analyse temporelle entre les risques observés et les évolutions passées des marchés financiers, des coefficients de corrélation peuvent être obtenus et comparés aux seuils suivants :

- Si la valeur absolue de la corrélation est comprise entre 0% et 25%, aucune interaction actif/passif n'est identifiée ;
- Si la valeur absolue de la corrélation est comprise entre 25% et 90%, une analyse plus poussée doit être effectuée pour savoir si une modification est vraiment nécessaire dans la modélisation de l'hypothèse ;
- Si la valeur absolue de la corrélation est supérieure à 90%, la modélisation doit impérativement être modifiée pour tenir compte de ces interactions.

Dans la présente étude, les données historiques feront l'objet de diverses analyses afin de détecter d'éventuelles anomalies. De plus, le biais d'estimation, le biais d'homogénéité et le biais temporel seront évalués indépendamment pour chacune des hypothèses sélectionnées.

3.3 Estimation d'une loi d'expérience

L'ensemble des garanties de l'assurance des emprunteurs s'organise au travers d'un modèle multi-états, comme exposé en figure 3.3. Ce dernier se décompose en quatre états reliés entre eux par des lois de transition. Chaque état correspond à une situation possible de l'assuré. Ainsi, le modèle multi-états permet de suivre la situation de l'assuré et les changements associés tout au long de la durée de vie de son contrat d'assurance.

Les différents états possibles sont :

- l'assuré est valide ;
- l'assuré est en arrêt de travail ;
- l'assuré est décédé ;
- l'assuré rachète son contrat, i.e. il rembourse la totalité de son emprunt avant la fin de son contrat.

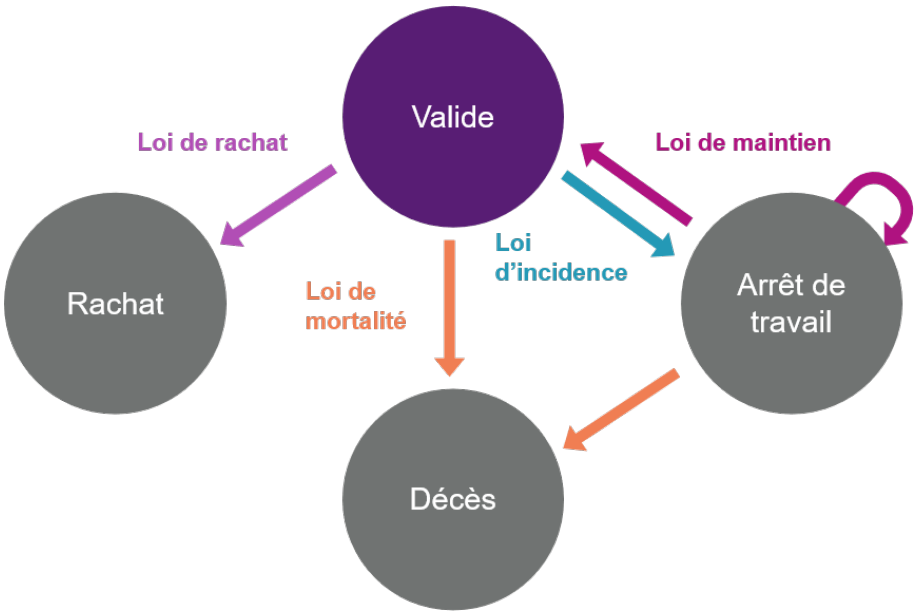


FIGURE 3.3 – Le modèle multi-états de l'assurance des emprunteurs

Dans la présente étude, les erreurs liées aux différentes lois de mortalité et à la loi de rachat feront l'objet d'une analyse approfondie. Au préalable, intéressons-nous à la façon dont elles sont estimées.

3.3.1 Introduction aux modèles de durée

Afin d'évaluer plus précisément les risques de décès, de rachat ou encore d'arrêt de travail, les organismes d'assurance peuvent utiliser leurs propres tables d'expérience. Ces tables sont rattachées à un portefeuille et sont construites sur les données historiques de ce portefeuille, i.e. sur un groupe d'individus présents dans le portefeuille précédent.

Ces lois d'expérience, utilisées en tant qu'hypothèses du modèle de projection, font appel à des modèles de durée. Le terme "durée" fait référence au temps écoulé jusqu'à la survenance d'un évènement. Cet évènement correspond au passage d'un état à un autre. Il peut s'agir du passage d'un individu de l'état "valide" à l'état "invalide", ou encore du passage de l'état "valide" à l'état "décès".

Pour résumer, les lois d'expérience permettent de mieux évaluer les engagements de l'assureur en estimant la survenance d'un évènement en fonction d'une ou plusieurs variables explicatives, telles que l'âge, le sexe et/ou l'ancienneté de l'assuré.

Lors de la construction de ces lois, plusieurs points doivent être correctement choisis et validés :

- L'unité de temps ;
- Les données historiques utilisées ;
- La méthode de construction des taux bruts ;
- La méthode de construction des taux lissés ;
- La fermeture de la table (estimation des taux aux bords).

À noter que les données utilisées doivent être exhaustives, exactes et appropriées, comme expliqué dans la section 2.

L'évaluation de cette durée consiste à estimer la distribution du temps de survie au travers d'une fonction de survie.

Supposons X une variable aléatoire positive ou nulle, représentant une durée de survie. La fonction de survie en t associée à cette variable s'écrit de la façon suivante :

$$S(t) = \mathbb{P}(X > t), t \geq 0$$

Elle correspond à la probabilité de survivre ou de rester dans le portefeuille jusqu'à l'instant t.

Usuellement, la fonction de répartition s'écrit :

$$F(t) = \mathbb{P}(X \leq t) = 1 - S(t)$$

Elle s'écrit également de la façon suivante :

$$F(t) = \int_0^t f(u)du$$

avec

$$f(u) = F'(t) = -S'(u)$$

Le risque instantané de décéder ou de racheter son contrat à un instant t se définit comme suit :

$$\lambda(t) = \frac{f(t)}{S(t)} = -ln(S(t))'$$

Il correspond à la probabilité qu'un individu meurt ou rachète son contrat juste après l'instant t sachant qu'il était respectivement en vie ou dans le portefeuille juste avant l'instant t.

Ainsi, le taux de hasard cumulé correspond à l'intégrale du risque instantané :

$$\Lambda(t) = \int_0^t \lambda(u)du = -ln(S(t))$$

A partir de cette équation, la fonction de survie peut être exprimée en fonction du taux de hasard cumulé :

$$S(t) = exp(-\Lambda(t))$$

Les modèles de durée présentés en section 3.3.3, permettent d'estimer cette fonction de survie et d'en déduire les taux de mortalité et de rachat.

3.3.2 Censures et troncatures

Les données historiques utilisées pour la construction des lois ne sont en réalité que partiellement recueillies dans le sens où elles appartiennent à une période d'observation déterminée et finie. Certaines informations n'appartenant pas à cette période d'observation sont donc manquantes. Ces données sont ainsi dites "censurées" ou "tronquées".

Le phénomène de censure est le phénomène le plus répandu dans la sélection des données. Une censure peut être dite "à droite" ou "à gauche".

Définissons trois variables pour un individu i :

- X_i son temps de survie ;
- C_i son temps de censure ;
- T_i la durée réellement observée.

Les variables X et C sont considérées indépendantes.

3.3.2.1 La censure à droite

On parle de censure à droite lorsque les durées de vie s'étendent au-delà de la période d'observation. Si l'évènement ne s'est pas produit avant la fin de cette période, alors il est noté que le temps de survie est supérieur à cette dernière valeur notée C_i.

Plus précisément, il existe trois types de censures à droite.

Censure de type I : Pour ce type de censure, le temps de censure C est fixé et ne dépend pas de i. Les temps de survie sont observés jusqu'à C. Ainsi, la réalisation de l'évènement est observée seulement si $X_i \leq C$. Sinon, il est retenu que l'évènement sera réalisé après C ($X_i > C$). Dans ce cas, la notation suivante est utilisée :

$$T_i = min(X_i, C) = X_i \wedge C$$

Censure de type II : Dans ce cas de censure, n individus sont observés jusqu'à ce que l'évènement ne soit réalisé que pour k individus parmi ces n. Ainsi, la durée observée pour les $n-k$ autres individus est retenue comme étant égale à la dernière durée observée $X_{(k)}$. Les durées réellement observées sont donc les suivantes :

$$T_{(1)} = X_{(1)}$$

$$T_{(2)} = X_{(2)}$$

$$...$$

$$T_{(k)} = X_{(k)}$$

$$T_{(k+1)} = X_{(k)}$$

$$...$$

$$T_{(n)} = X_{(k)}$$

Censure de type III (ou censure de type I aléatoire) : Dans ce dernier cas de censure à droite, nous considérons n variables aléatoires $C_1, ..., C_n$ i.i.d. représentant les temps de censure pour chacun des n individus. Un indicateur $\delta_i = \mathbf{1}_{X_i \leq C}$ renseignant la réalisation ou non de l'évènement pendant la durée de l'observation est utilisé.

- $\delta_i = 1$ si l'évènement est observé avant C_i. Dans ce cas, $T_i = X_i$
- $\delta_i = 0$ si l'évènement n'a pas été observé avant C_i. Dans ce cas, $T_i = C_i$. L'individu est donc dit censuré.

3.3.2.2 La censure à gauche

Les données sont censurées à gauche lorsque l'évènement a été réalisé avant le début de la période d'observation. Dans ce cas, il est seulement noté que l'évènement a été observé avant la date de la censure. Ainsi, pour chaque individu i, la durée réellement observée s'écrit de la manière suivante :

$$T_i = max(X_i, C_i) = X_i \vee C_i$$

3.3.2.3 La troncature

La troncature se rapporte à l'échantillon X.

Soit A un ensemble quelconque. En règle générale, une variable X est tronquée lorsque l'évènement est uniquement observable si $X_i \subset A$. Ainsi, l'échantillon "tronqué" correspond à l'ensemble des individus dont X_i de X appartient à l'ensemble A, les autres individus n'étant pas observables.

Soit Z un seuil prédéfini.

- **Troncature à droite :** Les données sont tronquées à droite lorsque X n'est observable que si $X < Z$.
- **Troncature à gauche :** À contrario, les données sont tronquées à gauche lorsque X n'est observable que si $X > Z$.

3.3.3 Modélisation non-paramétrique des taux bruts

Estimer une loi d'expérience revient à estimer un taux en fonction d'une ou plusieurs variables explicatives. Tout d'abord, les taux obtenus après modélisation sont dits **taux bruts** et forment une courbe qui comporte beaucoup de fluctuations. Ce phénomène est essentiellement dû à la taille de l'échantillon en entrée. Plus l'échantillon est petit, plus les fluctuations seront importantes. Afin de diminuer ces fluctuations, la courbe doit être lissée. Les taux finalement obtenus sont appelés **taux lissés**.

La littérature fait état de plusieurs méthodes de modélisation des lois d'expérience, lesquelles peuvent être regroupées dans trois catégories :

- **Les modèles non-paramétriques**, par exemple l'estimateur de Kaplan-Meier de la fonction de survie, l'estimateur d'Hoem du taux de survie et l'estimateur de Nelson-Aelen du risque cumulé ;
- **Les modèles semi-paramétriques**, par exemple les modèles à hasards proportionnels et le modèle de Cox ;
- **Les modèles paramétriques**, par exemple les modèles dont la distribution de la fonction de survie appartient à une famille de loi paramétrique connue telle que la loi de Weibull.

Les modèles non-paramétriques permettent de reproduire de façon fidèle les sorties d'un portefeuille historique. Quant aux modèles semi-paramétriques, ils sont en général légèrement plus complexes, mais sont tout autant utilisés en analyse de survie des données. Enfin, les modèles paramétriques sont efficaces lorsque le risque s'apparente à une loi paramétrique connue et que le volume de données n'est pas conséquent. Dans la présente étude, deux modèles non-paramétriques avec censure à droite de type III seront présentés.

3.3.3.1 Estimateur de Kaplan-Meier

L'un des estimateurs les plus fréquemment utilisés est l'estimateur de Kaplan-Meier. Afin d'expliquer son fonctionnement, nous prendrons le risque décès en exemple, le risque de rachat se modélisant exactement de la même façon.

Cet estimateur repose sur l'idée que l'énoncé *"un individu est en vie après l'instant t"* signifie la même chose que l'énoncé *"il était en vie à un instant t', tel que $t' < t$, et il n'est pas décédé en t"*. Cela revient à écrire la fonction de survie de la façon suivante :

$$S(t) = \mathbb{P}(X > t) = \mathbb{P}(X > t, X > t') = \mathbb{P}(X > t | X > t')\mathbb{P}(X > t')$$

De même, avec $t'' < t' < t$, la fonction de survie s'écrit :

$$S(t) = \mathbb{P}(X > t | X > t')\mathbb{P}(X > t' | X > t'')\mathbb{P}(X > t'')$$

Soit $T_{(k)}$, avec $k = 1, ..., n$, représentant plusieurs temps d'évènements. La probabilité qu'un individu soit toujours en vie après le temps $T_{(k)}$ s'écrit :

$$\mathbb{P}(X > T_{(k)}) = \prod_{i=1}^{k} \mathbb{P}(X > T_{(i)} | X > T_{(i-1)})$$

avec $T_{(0)} = 0$.

Ainsi, l'estimateur de Kaplan-Meier s'écrit :

$$\hat{S}(t) = \prod_{T_{(i)} < t, i=1,...,n} \left(1 - \frac{d_i}{e_i}\right)$$

avec

- d_k le nombre de décès entre $T_{(k-1)}$ et $T_{(k)}$;
- e_k le nombre d'individus exposés au risque de décéder entre $T_{(k-1)}$ et $T_{(k)}$ i.e. l'ensemble des individus en vie en $T_{(k-1)}$.

Notons qu'en cas de censure en $T_{(k)}$, $d_k = 0$.

Enfin, la probabilité qu'un individu meurt à l'âge x peut être estimée par :

$$\hat{q}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)}$$

3.3.3.2 Estimateur de Hoem

L'estimateur de Hoem (1969), aussi appelé estimateur du maximum de vraisemblance des taux bruts, est défini comme le rapport entre le nombre de sorties observées et l'exposition au risque sur une période d'observation.

A un instant t, la probabilité qu'un individu meurt à l'âge x s'écrit :

$$\hat{q}_x(t) = \frac{d_x(t)}{E_x(t)} = \frac{d_x}{E_x}$$

avec

- d_x le nombre de sorties à l'âge x sur l'ensemble de la période d'observation ;
- E_x le nombre d'individus d'âge x présents dans l'échantillon sur l'ensemble de période la d'observation.

Les estimateurs de Hoem et de Kaplan-Meier donnent à peu près les mêmes résultats. Ils permettent chacun d'obtenir une courbe de taux bruts. Comme expliqué précédemment, ces taux doivent être lissés pour être correctement utilisés dans le modèle de projection. Différentes méthodes de lissage sont présentées dans la section suivante.

3.3.4 Lissage des taux bruts

L'idée est de rechercher la méthode de lissage la plus adaptée aux données. Toutefois, plusieurs méthodes de lissage peuvent être associées pour améliorer le processus.

Les deux méthodes de lissage qui seront utilisées pour lisser les taux bruts des données BPCE Vie sont brièvement expliquées ci-après.

3.3.4.1 Méthode des moyennes mobiles

Le lissage par moyennes mobiles consiste à remplacer la valeur d'un point par une moyenne de l'ensemble des points autour de celui-ci.

Plus précisément, la moyenne mobile d'ordre $r+1$ d'un point $\hat{q}_x$ s'écrit :

$$\tilde{q}_x = \frac{1}{m} \sum_{i=-r}^{r} \hat{q}_{x+i}$$

avec $r = \frac{m-1}{2}$ et m un paramètre à fixer tel que $m \leq 2$.

Le lissage par moyennes mobiles a l'avantage d'être simple à mettre en oeuvre. En revanche, il n'est pas applicable aux bords. Pour remédier à cela, il peut éventuellement être associé à une autre méthode de lissage.

3.3.4.2 Méthode des noyaux discrets

La méthode des noyaux discrets permet d'estimer la densité de probabilité d'une variable aléatoire X, en se basant sur un échantillon statistique de n valeurs.

La fonction de survie s'écrit de la manière suivante :

$$\tilde{f}_{h,n}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

avec

- h un paramètre nommé fénêtre, réglant l'intensité du lissage ;
- K le noyau.

Plusieurs types de noyaux peuvent être utilisés. Le noyau gaussien $K(x) = \frac{1}{\sqrt{2\pi}} exp\left(\frac{-1}{x^2}\right)$ est le plus souvent choisi.

Chapitre 4: Erreurs liées au modèle de projection

Un modèle de projection permet de projeter les flux relatifs à chaque contrat en portefeuille sur une longue période, afin d'obtenir entre autres la valeur des provisions techniques. À noter qu'une mauvaise estimation de ces provisions techniques pourrait engendrer des conséquences néfastes sur la solvabilité d'une compagnie d'assurance. La qualité du modèle est donc un réel enjeu pour les compagnies d'assurance dans la maîtrise des risques.

Ce chapitre sera dédié au risque modèle et plus particulièrement aux méthodes permettant de détecter des erreurs dans un modèle de projection.

Les modèles de projection étant généralement très complexes, il n'est pas faisable de les étudier en détail pour repérer les éventuelles erreurs. L'une des méthodes les plus efficaces de détection d'anomalies est le *backtesting*.

Dans un premier temps, les différents anomalies qui peuvent être présentes dans un modèle seront détaillées. Dans un second temps, le processus de *backtesting* sera défini, puis plusieurs tests pour valider la qualité d'un modèle à partir des résultats d'un *backtesting* seront présentés.

4.1 Classification des erreurs liées au modèle

Les erreurs présentes dans un modèle de projection sont considérées comme des anomalies et non des biais, ces anomalies pouvant être décomposées en quatre catégories.

Anomalies liées à une mauvaise application des hypothèses

- **Source :** Les hypothèses du modèle, telles que les lois d'expérience, pourraient être utilisées de manière inappropriée dans le modèle. Par exemple, il pourrait s'agir d'une inadéquation entre les hypothèses sélectionnées et les règles de gestions et de fonctionnement de l'assureur. Dans ce cas là, le modèle pourrait par erreur appliquer un taux de mortalité à un assuré ayant un âge plus élevé que l'âge limite de couverture.

- **Indicateurs de contrôle :** Ces anomalies se détectent à partir d'un *backtest* (cf. section 4.2). L'approche consiste à remplacer les hypothèses dans le modèle par les bornes supérieure ou inférieure de leur estimateur. De ce fait, le modèle projette les bornes inférieure ou supérieure de l'intervalle de confiance des flux. Dans le cas d'une bonne application des hypothèses dans le modèle, les flux réels se situent à l'intérieur de l'intervalle de confiance des flux projetés.

  En revanche, lorsque les résultats du *backtest* s'avèrent mauvais, les anomalies peuvent être identifiées en effectuant un suivi plus approfondi des règles implémentées dans le modèle.
- **Seuils :** Afin de juger précisément de la bonne application des hypothèses dans le modèle, les flux présents dans le *backtest* doivent remplir les conditions présentées en section 5.3.2.

Anomalies liées aux évolutions dans le temps

- **Source :** Ces erreurs sont issues d'une part des évolutions dans le temps des facteurs de risques et des règles de gestion et de fonctionnement au cours de la période considérée, et d'autre part de la prise en compte de ces évolutions dans le modèle. Par exemple, un changement de loi, un nouveau produit ou le changement du gestionnaire de sinistre qui pourrait augmenter les délais de paiement, peuvent entraîner des modifications dans le modèle.
- **Indicateurs de contrôle :** Ce type d'erreurs peut également être détecté au travers d'un *backtest*.

  Par ailleurs, il est nécessaire de s'informer constamment des évolutions auprès des équipes de gestion et du juridique qui sont, dans les faits, au courant de tous les changements législatifs, contractuels et de règles de gestion qui pourraient impacter les risques.
- **Seuils :** Les conditions relatives à la validation d'un *backtest* seront présentées en section 4.2.

Anomalies liées à l'absence de prise en compte d'un risque ou aux interactions inappropriées entre un risque et les règles

- **Source :** Ces anomalies sont présentes dans le modèle pour les deux raisons suivantes. Ou bien certains risques, facteurs de ces risques, certaines règles de gestion ou règles de fonctionnement ne sont pas pris en compte dans le modèle. Il pourrait s'agir par exemple de l'oubli d'une règle de provisionnement, telle qu'une limite de règlement des sinistres. Ou bien les interactions entre les risques et les règles sont considérées de manière inappropriées.
- **Indicateurs de contrôle :** Ce type d'erreur peut également être détecté au travers d'un *backtest*.

  Lorsque les résultats du *backtest* s'avèrent mauvais, les anomalies peuvent être identifiées en effectuant un suivi plus approfondi des règles implémentées dans le modèle.
- **Seuils :** Les conditions relatives à la validation d'un *backtest* seront présentées en section 4.2.

Anomalies liées à la mauvaise implémentation de la méthode retenue

- **Source :** Les modèles étant généralement très complexes, ils nécessitent de bonnes connaissances techniques et théoriques pour être correctement mis en oeuvre, faute de quoi des erreurs seraient introduites dans le code du modèle. De plus, certains assureurs effectuent des approximations dans leurs modèles pour une raison de simplification. Les erreurs issues de ces approximations doivent être minimes et ne doivent pas impacter l'évaluation des risques.
- **Indicateurs de contrôle :** Ce type d'erreurs peut également être détecté au travers d'un *backtest*.
- **Seuils :** Les conditions relatives à la validation d'un *backtest* seront présentées en section 4.2.

Ces quatre types d'anomalies étant contrôlés chacun par du *backtesting*, il est généralement difficile de déterminer quels sont les types d'anomalies réellement présents dans le modèle lorsqu'un seul *backtest* est utilisé et dont le résultat s'avère négatif. Dans ce cas, des études complémentaires seraient nécessaires pour déterminer exactement le ou les types d'anomalies.

A présent, définissons concrètement le *backtesting*.

4.2 *Backtesting*

Le *backtesting* consiste à comparer la trajectoire des flux réels aux flux moyens projetés par le modèle. Dans le cas d'une bonne modélisation des flux, les valeurs réelles sont distribuées autour des valeurs estimées et correspondent en moyenne aux valeurs estimées.

Ainsi, les flux réels doivent se trouver à l'intérieur de l'intervalle de confiance des flux projetés qui ont été construits sur les données historiques du portefeuille. Néanmoins, il est possible que quelques valeurs soient situées à l'extérieur de l'intervalle de confiance sans pour autant remettre en cause le modèle de projection. L'objet de cette section est de définir des indicateurs permettant de juger de la qualité d'un modèle, à partir d'un *backtest*. Plus précisément, nous allons nous intéresser aux écarts entre les valeurs réelles et la borne supérieure de l'intervalle de confiance afin de déterminer si les flux ont été sous-estimés ou non. La démarche qui sera décrite peut également être utilisée pour la borne inférieure afin de déterminer si les flux ont été surestimés ou non.

Soit E une variable aléatoire précisant le nombre d'évènements où l'intervalle de confiance est franchi. Ces franchissements sont aussi appelés "violations". La variable E s'écrit de la façon suivante :

$$E = \sum_t \mathbf{1}_{\{x_t > IC^{\alpha}_{sup}\}}$$

où $\mathbf{1}_{\{x_t > IC^{\alpha}_{sup}\}}$ est une indicatrice qui vaut 1 si la valeur x_t à la date t a franchi l'intervalle de confiance de niveau de confiance $1-\alpha$.

Ainsi, $E \sim Binom(n, \alpha)$ où n est le nombre de points de comparaison.

Donc $\mathbb{E}[E] = n\alpha$.

La qualité du modèle est jugée satisfaisante si les deux hypothèses suivantes sont remplies. Campbell détaille ces deux hypothèses dans un article de la Réserve Fédérale des États-Unis [8].

- **Hypothèse de couverture non-conditionnelle :** Cette hypothèse concerne le nombre de violations de l'intervalle de confiance. Rappelons qu'un intervalle de confiance à $1-\alpha\%$ indique que dans $1-\alpha\%$ des cas, l'intervalle contient le flux estimé. Ainsi, une bonne estimation du flux signifie qu'au moins $1-\alpha\%$ des valeurs réelles se situe entre les bornes de l'intervalle. En d'autres termes, la proportion de violations de l'intervalle doit être inférieure ou égale à $\alpha\%$.

  $$\mathbb{E}[E] = n\alpha \ i.e. \forall t, \mathbb{P}(x_t > IC^{\alpha}_{sup}) = \alpha$$

  Si $\mathbb{E}[E] >>> n\alpha$, alors le flux projeté par le modèle est considéré comme sous-estimé.

- **Hypothèse d'indépendance :** Cette seconde hypothèse porte sur la manière dont les violations se produisent. Ces violations doivent être indépendantes entres elles. C'est-à-dire qu'un historique de violations ne doit pas permettre de déterminer une éventuelle prochaine violation de l'intervalle. Cette hypothèse d'indépendance s'écrit de la façon suivante : la variable $\mathbf{1}_{\{x_t > IC^{\alpha}_{sup}\}}$ est indépendante de la variable $\mathbf{1}_{\{x_{t+k} > IC^{\alpha}_{sup}\}}$ pour tout $k > 0$.

Dans la mesure où ces deux hypothèses sont indépendantes, elles doivent chacune être respectées pour valider la qualité d'un modèle.

En général, le test de Kupiec, détaillé par la suite, est utilisé pour valider l'hypothèse de couverture non conditionnelle. Concernant l'hypothèse d'indépendance, il existe plusieurs tests parmi lesquelles les plus utilisés sont le test de Markov proposé par Christoffersen (1998), les tests de durée de Christoffersen et Pelletier (2004) et les tests basés sur la régression d'Engle et Manganelli (2004). Deux d'entre eux seront présentés par la suite et appliqués dans la section 5.

4.2.1 Test de couverture non conditionnelle

Pour tester l'hypothèse de couverture non conditionnelle, il suffit de tester l'hypothèse nulle suivante :

$$H_0 : \mathbb{E}[I_t] = \alpha$$

$$H_1 : \mathbb{E}[I_t] \neq \alpha$$

avec $I_t = \mathbf{1}_{\{x_t > IC^{\alpha}_{sup}\}}$.

Kupiec propose en 1995 la statistique de ratio de vraisemblance associée à H_0 :

$$-2\ln\left[(1-\alpha)^{n-E}\alpha^{E}\right]+2\ln\left[\left(1-\frac{E}{n}\right)^{n-E}\frac{E}{n}^{E}\right]\xrightarrow[n\to\infty]{E}\chi^2(1)$$

avec $\chi^2(1)$ la loi du khi deux d'ordre 1.

Si cette première hypothèse n'est pas rejetée, alors un ou plusieurs tests d'indépendance peuvent être réalisés tels que le test de Christoffersen et le test d'Engle et Manganelli présentés dans les sections suivantes.

4.2.2 Test de Markov de Christoffersen (1998)

Christoffersen modélise en 1998 les franchissements $I_t(\alpha)$ par une chaîne de Markov avec une matrice de transition Π telle que

$$\Pi=\begin{pmatrix}\pi_{00} & \pi_{01}\\ \pi_{10} & \pi_{11}\end{pmatrix}$$

où $\pi_{ij}=P[I_t(\alpha)=j|I_{t-1}(\alpha)=i]$.

Il teste l'hypothèse nulle suivante :

$$H_0:\Pi_\alpha=\begin{pmatrix}1-\alpha & \alpha\\ 1-\alpha & \alpha\end{pmatrix}$$

La statistique de ratio de vraisemblance sous H_0 est définie comme :

$$-2\ln L(\Pi_\alpha,I_1(\alpha),...,I_n(\alpha))+2\ln L(\widehat{\Pi_\alpha},I_1(\alpha),...,I_n(\alpha))\xrightarrow[n\to\infty]{E}\chi^2(2)$$

avec $\chi^2(2)$ la loi du khi deux d'ordre 2, $\widehat{\Pi_\alpha}$ l'estimateur du maximum de vraisemblance et $L[.]$ la log-vraisemblance.

Dans ce cas-là, la log-vraisemblance est égale à :

$$L(\Pi_\alpha,I_1(\alpha),...,I_n(\alpha)=\pi_{00}^{n_{00}}\pi_{01}^{n_{01}}\pi_{10}^{n_{10}}\pi_{11}^{n_{11}}$$

où n_{ij} designe le nombre de fois où l'on observe $I_t(\alpha)=j|I_{t-1}(\alpha)=i$.

Ce test est facile à mettre en oeuvre mais il présente deux inconvénients non négligeables. Tout d'abord, l'indépendance est étudiée contre une forme spécifique de dépendance qui n'intègre pas de dépendance d'ordre supérieur à un. De plus, la chaîne de Markov ne permet pas d'étudier d'autres variables que celle des franchissements $I_t(\alpha)$ dans une possible dépendance des franchissements. Engel et Manganelli lèveront plus tard en 2004 ces deux défauts à travers leurs propres tests décrits dans la section suivante.

4.2.3 Tests basés sur la régression d'Engle et Manganelli (2004)

Les tests d'Engle et Manganelli sont fondés sur un modèle de régression linéaire. Plus précisément, la régression est appliquée au processus des violations centrées sur α et associées à $I_t(\alpha)$, appelés les *hits* :

$$Hit_t(\alpha)=I_t(\alpha)-\alpha$$

Le modèle de régression considéré est le suivant :

$$Hit_t(\alpha) = \delta + \sum_{k=1}^{K} \beta_k \; Hit_{t-k}(\alpha) + \sum_{k=1}^{K} \gamma_k \; g[Hit_{t-k}(\alpha), z_{t-k}] + \epsilon_t$$

où g est une fonction prenant compte des violations passées et des variables z_{t-k} appartenant à l'information disponible $\mathscr{F}_{t-1}$.

Le processus des innovations ϵ_t est un processus i.i.d. tel que :

$$\epsilon_t = \begin{cases} 1-\alpha & \text{avec une probabilité } \alpha \\ -\alpha & \text{avec une probabilité } 1-\alpha \end{cases}$$

L'hypothèse de couverture non-conditionnelle et l'hypothèse d'indépendance sont toutes les deux vérifiées si l'hypothèse jointe suivante n'est pas rejetée :

$$H_0 : \delta = \beta_k = \gamma_k = 0 \; \forall k = 1, ..., K$$

En effet, d'une part, l'indépendance des violations est validée si les β_k et les γ_k sont nuls. D'autre part, l'hypothèse de couverture non-conditionnelle est satisfaite si δ est nulle. En effet :

$$\mathbb{E}[Hit_t(\alpha)] = \mathbb{E}[\epsilon_t] = 0 \implies \mathbb{E}[I_t(\alpha)] = \alpha$$

Pour tester la nullité de tous les coefficients, le test de Wald peut par exemple être utilisé. Soit $\Psi = (\delta \; \beta_1 \; ... \; \beta_K \; \gamma_1 \; ... \; \gamma_K)'$ le vecteur des 2K+1 coefficients du modèle de régression et Z la matrice des variables explicatives. La statistique de Wald associée est la suivante :

$$\frac{\widehat{\Psi}' Z' Z \widehat{\Psi}}{\alpha(1-\alpha)} \xrightarrow[n\to\infty]{E} \chi^2(2K+1)$$

avec $\chi^2(2K+1)$ la loi du khi deux d'ordre $2K+1$.

Dans le cas où l'on voudrait seulement valider l'hypothèse d'indépendance, l'hypothèse nulle se restreint à la nullité des coefficients β_k et γ_k :

$$H_0 : \beta_k = \gamma_k = 0 \; \forall k = 1, ..., K$$

La statistique correspondante est :

$$\frac{\widehat{\Psi}' R' \left[R(Z'Z)^{-1}R'\right]^{-1} R\widehat{\Psi}}{\alpha(1-\alpha)} \xrightarrow[n\to\infty]{E} \chi^2(2K)$$

où $R = [0 : I_{2K}]$ est une matrice de dimension $2K \times 2K+1$, telle que $R\Psi = \Omega$ avec $\Omega = (\beta_1 \; ... \; \beta_K \; \gamma_1 \; ... \; \gamma_K)'$.

Chapitre 5: Mise en application en assurance des emprunteurs

Dans ce chapitre, nous détaillerons toute la démarche que nous avons suivie pour déterminer l'impact des différentes erreurs sur la valeur du BE.

Ainsi, ce chapitre sera décomposé en quatre phases. La première consistera à déterminer les anomalies présentes dans les données en entrée du modèle. La deuxième vise à évaluer les erreurs présentes dans les hypothèses. La troisième permettra de détecter d'éventuelles anomalies dans le modèle de projection. Enfin, la quatrième sera dédiée à la quantification de l'impact de toutes ces erreurs sur la valeur du BE.

Pour des raisons de confidentialité, tous les résultats qui seront présentés ont été multipliés par un facteur masqué.

Par ailleurs, l'ensemble de l'application a été implémenté dans le langage de programmation R.

5.1 Détection d'anomalies dans les données en entrée du modèle

Cette section est consacrée aux erreurs présentes dans les données en entrée du modèle de projection.

Dans un premier temps, nous présenterons les données sur lesquelles nous allons travailler et les analyserons une première fois pour repérer de potentielles sources d'anomalies.

Dans un second temps, nous appliquerons les deux techniques de réduction de dimension de données présentées dans le chapitre 2.

Cela nous permettra, dans un troisième temps, d'utiliser les trois algorithmes de *clustering*. Nous étudierons alors plus profondément les anomalies détectées par ces algorithmes. Nous tenterons de comprendre les raisons pour lesquelles elles ont été écartées et les confronterons aux données suspectées par la loi de Benford.

Enfin, pour s'assurer de l'efficacité des trois algorithmes de *clustering*, nous appliquerons un dernier algorithme de détection d'anomalie, le *PU learning*.

5.1.1 Présentation des données

La base de données que nous allons étudier a servi à la construction des *model points* pour le trimestre 4 de 2020. Elle a été utilisée et reste utilisable en entrée du modèle de projection du BE.

Tout d'abord, il est nécessaire de ne sélectionner qu'une partie des variables. En effet, certaines variables n'ont aucun intérêt pour évaluer la qualité des données, par exemple celles utilisées pour les jonctions avec d'autres tables. Les variables avec une seule modalité sont également mises à l'écart.

Ainsi, nous obtenons une base contenant 31 variables et environ un million de lignes. Chacune de ces lignes correspond à un contrat d'assurance des emprunteurs sur un crédit à la consommation.

Les variables étudiées sont les suivantes :

Identification du portefeuille
Identification de l'établissement bancaire
État du contrat (valide, décédé ou arrêt de travail)
Ancienneté de l'état pour un arrêt de travail, un décès ou une perte d'emploi
Date d'émission du contrat
Date de maturité du contrat
Taux d'intérêt
Taux de prime
Capital initial
Capital restant dû
Quotité assurée
Âge de l'assuré
Sexe de l'assuré
Sous-catégorie de l'assuré (standard, étudiant, jeune, senior...)
Âge limite de couverture pour la garantie décès
Type de différé
Durée du différé
Génération tarifaire du taux de commission

Maintenant que les données ont bien été sélectionnées, nous allons pouvoir les étudier. Commençons par repérer d'éventuelles données n'appartenant pas au périmètre défini (prêts personnels) et les données aberrantes pour répondre respectivement au critère d'exhaustivité et au critère d'exactitude définis dans la section 2.1.

5.1.2 Première approche : analyses univariées

Des analyses univariées peuvent être utilisées pour identifier certaines tendances et potentiellement repérer l'origine des erreurs qui seront détectées par les algorithmes de *clustering*. Le tableau 5.1 présente des incohérences détectées en analysant chacune des variables les unes après les autres.

Notons qu'il est possible que certains contrats correspondent à des prêts pour achat de biens privés onéreux (bateau, voiture, avion...), avec comme particularité d'avoir une durée longue et un capital initial élevé.

Par ailleurs, certains des contrats avec des taux de prime très faibles correspondent certainement à des prêts étudiants, assimilables en réalité à des prêts personnels.

Par la suite, il serait intéressant d'étudier comment les algorithmes de détection d'anomalies réagissent face à ces incohérences mises en évidences par des analyses univariées. Aussi, nous soumettrons à ces algorithmes la base de données intégrale.

Avant de lancer le processus de réduction de dimension des données, les variables catégorielles sont transformées en *dummy variables*, c'est-à-dire en plusieurs variables binaires.

Variable	Identification des incohérences	Explications	Nombre de contrats concernés
Durée du contrat	$>$ 10 ans	Les prêts personnels étant à court terme, la durée du contrat ne devrait pas excéder 10 ans.	0,6%
Maturité du contrat	$<$ 2017	Les contrats ayant une date de maturité inférieure à l'année 2017 concernent des décès et ont fait l'objet d'un retard de remboursement, alors que tout remboursement du sinistre se fait habituellement rapidement.	$< 0,1\%$
Taux de prime	Valeurs très faibles	Plusieurs têtes ont des taux de prime très faibles relevant plutôt des prêts immobiliers.	13%
Capital initial	Valeurs très faibles ($<$ 200€)	Le capital initial doit être assez élevé sinon aucun prêt ne serait consenti.	$< 0,1\%$
Capital initial	Valeurs très élevées ($>$ 75000€)	Le montant des prêts personnels ne sont généralement pas très élevés.	0,2%
Capital restant dû	$>$ Capital initial	Pour les contrats sans différé total et n'appartenant pas à une banque proche de la Suisse (prêts en devise), le capital restant dû ne devrait pas être supérieur au capital initial.	$< 0,1\%$

TABLE 5.1 – Synthèse des incohérences détectées à partir d'analyses univariées

5.1.3 Réduction de dimension de la base

Pour utiliser les algorithmes de détection d'anomalies, la dimension de notre base doit être réduite. Pour notre part, les données seront réduites en dimension 2 afin de faciliter leur visualisation.

Nous allons appliquer les deux algorithmes de réduction de dimension détaillés en section 2.2. Rappelons que l'algorithme ACP ne requiert aucun paramètre en entrée. En revanche, l'algorithme UMAP nécessite un paramètre k représentant le nombre de voisins de chaque point. Plus ce paramètre est élevé, plus la complexité est grande, ce qui nécessitera davantage de mémoire vive et augmentera les temps de calcul. Nous devons prendre en compte la taille de la mémoire vive de l'ordinateur pour dimensionner ce paramètre.

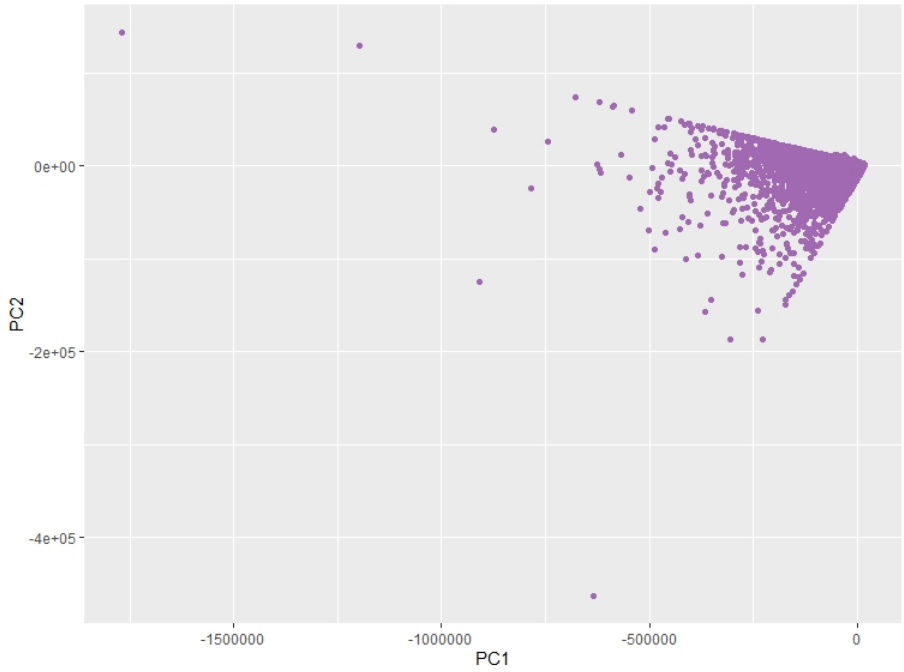
Alors que l'ACP s'appuie davantage sur la structure globale des données, il est intéressant de choisir une valeur de k relativement faible qui mettra plus en avant la structure locale avec la méthode UMAP. Ainsi, nous pouvons faire le choix de $k = 15$. Par la suite, les résultats obtenus à partir de ces deux algorithmes seront comparés.

Notons que le temps d'exécution de UMAP est bien plus long que celui de l'ACP, comme figuré dans le tableau 5.2. Cela s'explique par la complexité de l'algorithme UMAP.

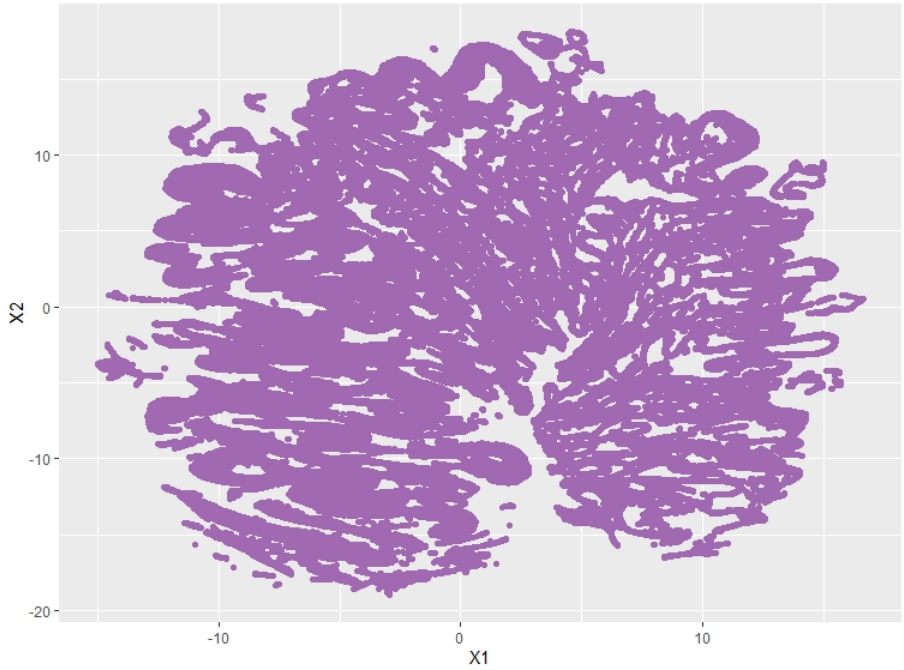
Algorithme	Temps d'exécution
ACP	16 secondes
UMAP	37 minutes

TABLE 5.2 – Temps d'exécution des algorithmes de réduction de dimension

Nous avons respectivement utilisé les fonctions *prcomp* du package *stats* et *umap* du package *uwot* pour l'ACP et l'UMAP. Comme illustré en figure 5.1, d'un point de vue graphique, les résultats semblent très différents d'une méthode à l'autre. Les données de l'ACP forment un cône et font ressortir des points très écartés que l'on peut facilement caractériser comme étant des anomalies. Au contraire, nous obtenons une forme plus arrondie pour les données de l'UMAP, avec moins de concentration, ce qui explique le caractère local de la méthode UMAP.



(a) ACP



(b) UMAP

FIGURE 5.1 – Base de données réduite en dimension 2

Rappelons que pour l'ACP, la variance doit être maximisée sur chacune des composantes principales. Comme la construction des composantes se fait une par une, la première composante *PC1* est celle qui a la variance la plus élevée. Pour notre part, elle détient 80,6% de la variance totale. La variance totale correspond à la somme des variances de chacune des composantes. La deuxième composante représente 18,6% de la variance totale. Comme nous souhaitons être en dimension 2, seules les deux premières composantes principales sont gardées. Dans la mesure où elles représentent à elles-seules presque 100% de la variance totale, il n'est donc pas démesuré de se limiter à ces deux composantes. Peu d'informations supplémentaires seraient conservées si nous avions choisi une dimension plus grande.

Une fois nos données ainsi réduites, nous pouvons leur appliquer des algorithmes de *clustering*.

5.1.4 Sélection des anomalies : algorithmes de *clustering*

Étant donné que nous ne disposons d'aucun jeu de données étiquetées, i.e. aucun jeu de données d'entraînement, sur lequel nous pourrions nous appuyer pour détecter par comparaison les anomalies de notre base, nous allons utiliser des méthodes non-supervisées.

Nous exécuterons les algorithmes d'apprentissage non-supervisés sur les deux jeux de données obtenus après réduction de dimension avec l'ACP et l'UMAP.

Parmi les anomalies que nous obtiendrons avec chacun de ces algorithmes, les points que nous retiendrons comme anomalies seront ceux qui auront été désignés "anomalie" par les trois méthodes de détection. L'idée est de mélanger plusieurs méthodes pour être certain du caractère anormal des points.

Algorithme	Paramètres
K-moyennes	k, le nombre de *clusters*
	$minT$, la taille minimum d'un *cluster* pour la sélection des anomalies
Isolation Forest	n, le nombre d'arbres
	s_{lim}, le score limite pour la sélection des anomalies
HDBSCAN	$minPts$, le nombre minimum de points dans un *cluster*

Table 5.3 – Paramètres des algorithmes d'apprentissage non-supervisés

Le sujet le plus délicat dans cette section est le calibrage des paramètres qui est plus ou moins évident selon les méthodes. Ils sont résumés dans le tableau 5.3.

5.1.4.1 Application de l'algorithme K-moyennes

Pour exécuter l'algorithme k-moyennes, nous avons choisi la fonction *h2o.kmeans* du logiciel H2O que nous avons appelé depuis le langage de programmation R. Cette fonction requiert un paramètre k correspondant au nombre de *clusters* formés. Il doit cependant être le plus adapté possible à nos données. Pour cela, il est possible de tracer la somme totale des carrés à l'intérieur des *clusters* (nommée *Tot.Withinss* ici et définie comme fonction J dans la section 2.3.1.1) en fonction de k (cf. figure 5.2). La meilleure valeur de k possible se situe au niveau du "coude" de la courbe". Au final, que ce soit pour les données provenant de l'ACP ou de UMAP, la meilleure valeur de k est identique, à savoir :

$$k_{acp} = k_{umap} = 600$$

L'algorithme de *clustering* peut maintenant être exécuté. Son temps de calcul est de deux minutes sur l'ensemble des données.

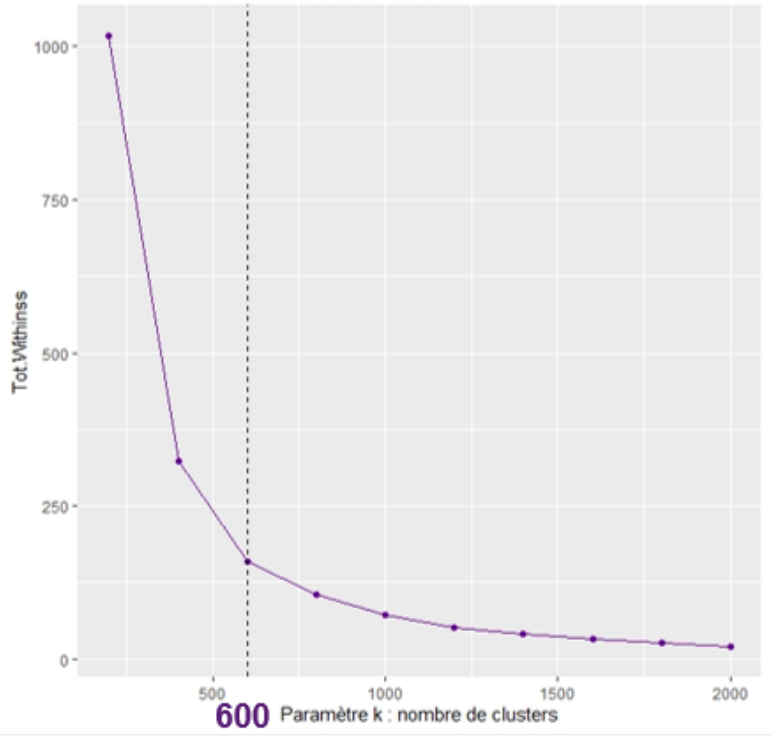
Une fois tous les *clusters* formés par l'algorithme, la sélection des anomalies peut commencer. Nous faisons le choix de désigner comme anomalies tous les points appartenant aux *clusters* de taille inférieure à $minT_{acp}$ et $minT_{umap}$ respectivement pour les données de l'ACP et de l'UMAP. Ces deux limites sont déterminées à partir d'un graphique représentant la proportion d'anomalies dans le jeu de données en fonction de la taille minimum des *clusters* (cf. figure 5.3).

Concernant les données de l'UMAP, $minT_{umap}$, égal à 720, est déterminé par le "coude" de la courbe.

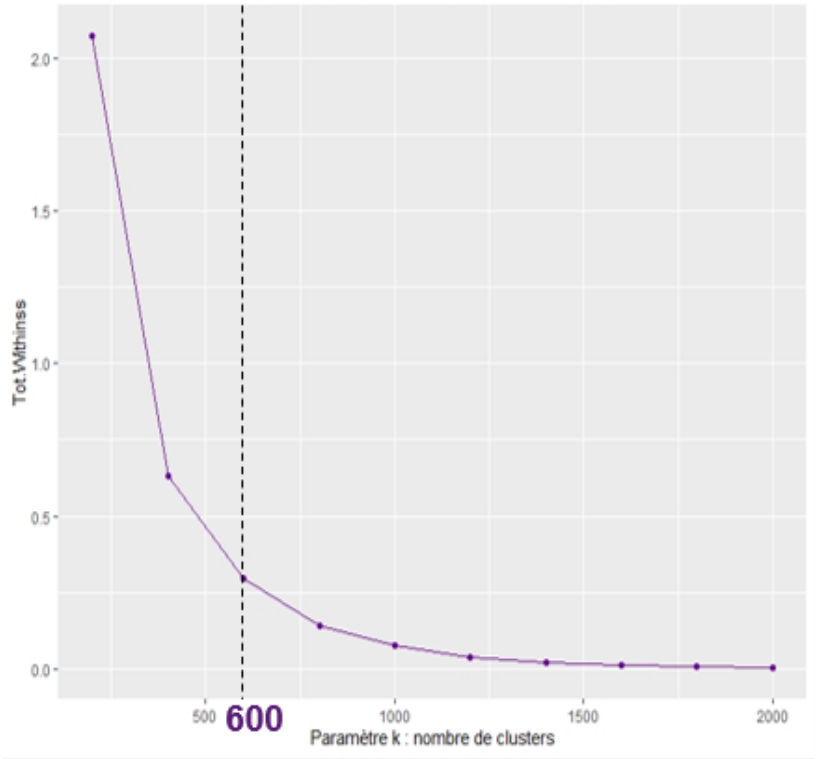
En revanche, pour les données de l'ACP, il est moins évident de définir cette limite. En effet, la courbe (a) de la figure 5.3 ne présente pas de "coude". Cela provient du fait que 60% des *clusters* ont une taille strictement inférieure à 6 points. Ainsi la très grande majorité des points est regroupée dans à peu près 5 *clusters* parmi les 600 existants. Nous avons fait le choix de retenir une taille limite de points par *cluster*, $minT_{acp}$, égale à 900, ce qui correspond au 1er palier de la courbe. Avec cette limite, les anomalies sont regroupées dans 92% des *clusters*. Cette sélection des anomalies reste assez imprécise. Aussi, nous pouvons donc nous demander pour la première fois si la méthode de réduction ACP est adaptée pour le traitement de notre base de données.

Les anomalies retenues sont représentées en violet sur la figure 5.4. Pour l'ACP, elles sont situées aux extrémités du cône que forme les données, ce qui paraît plutôt logique pour une approche globale. Elles représentent 2% des données et 40% d'entre elles sont également des anomalies selon la réduction UMAP. Pour les données de l'UMAP, les groupes d'anomalies, représentant 7% des données, sont assez dispersés. Une prépondérance est cependant notable sur la partie droite du graphique et en particulier en bas.

Nous retenons que davantage d'anomalies ont été détectées lorsque les données sont réduites par l'UMAP.

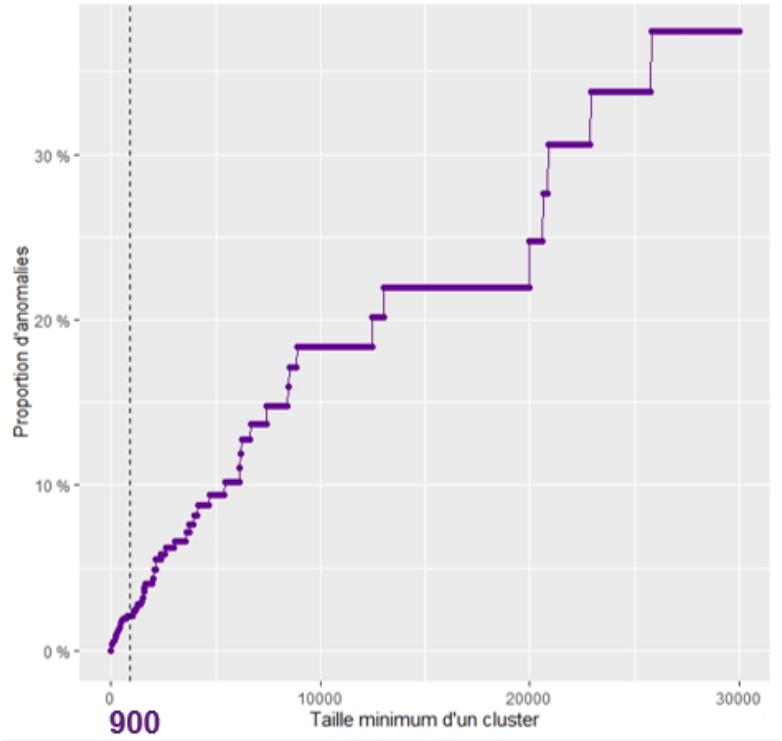


(a) ACP

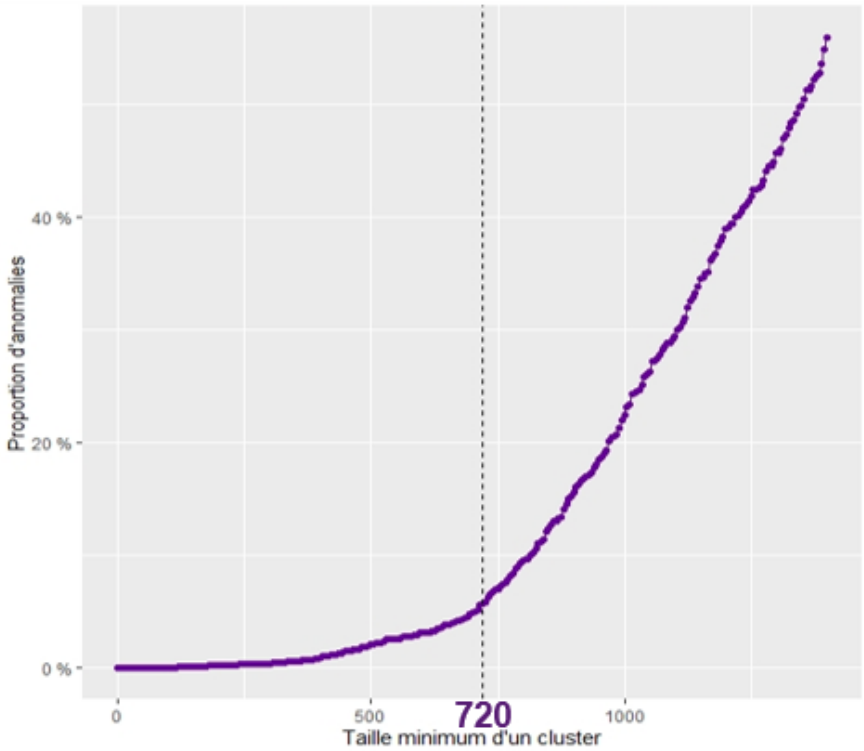


(b) UMAP

FIGURE 5.2 – Sélection du paramètre k

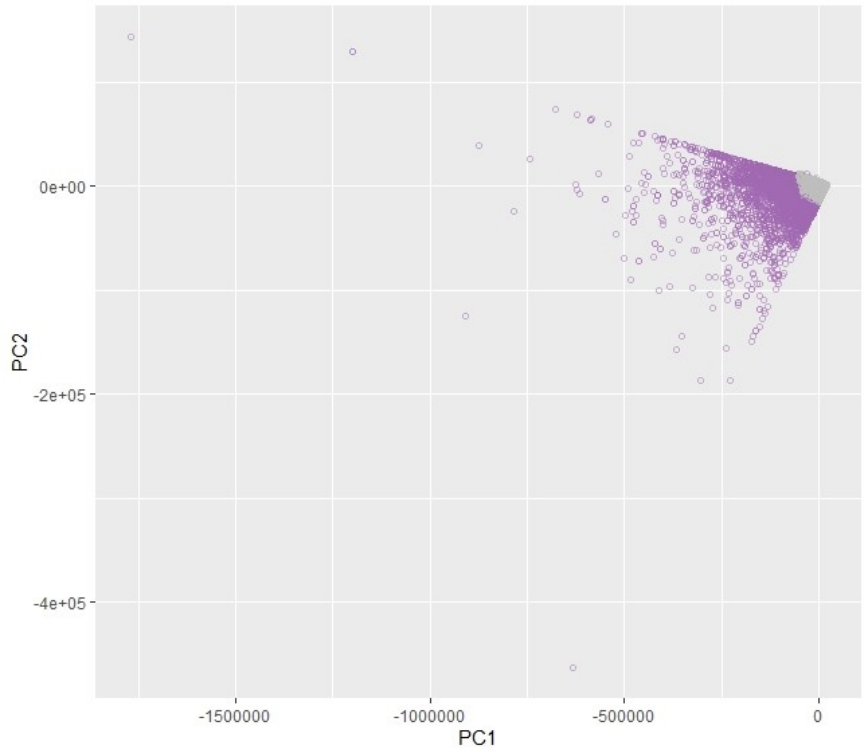


(a) ACP

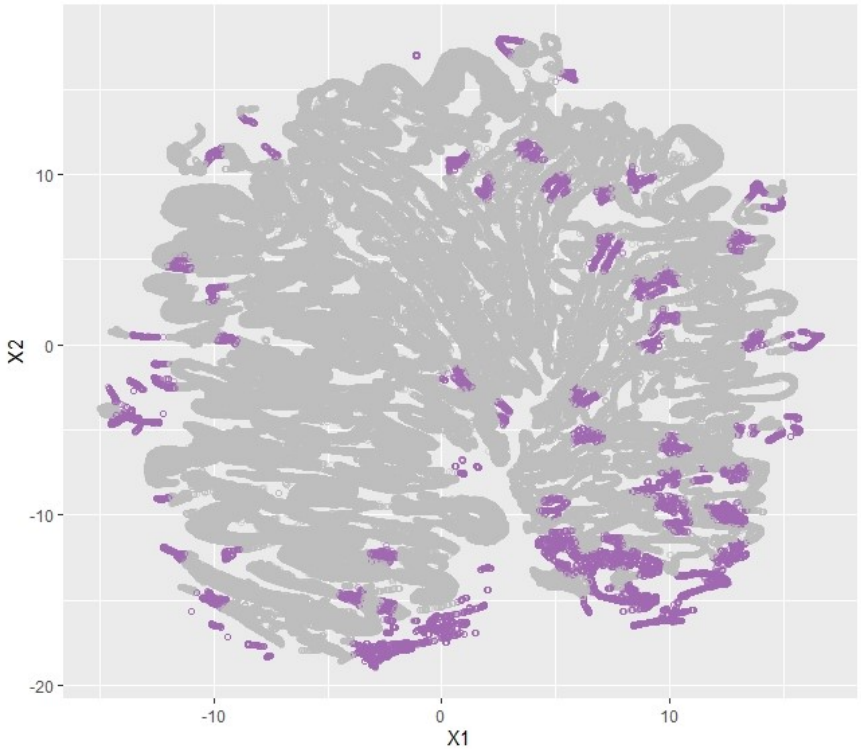


(b) UMAP

FIGURE 5.3 – Sélection de la taille minimum d'un cluster



(a) ACP



(b) UMAP

FIGURE 5.4 – Anomalies détectées par k-moyennes

Algorithme de réduction	Proportion d'anomalies
ACP	2%
UMAP	7%

Table 5.4 – Proportion d'anomalies détectées par les k-moyennes

5.1.4.2 Application de l'algorithme *Isolation Forest*

Pour cet algorithme de *clustering*, la fonction *isolationForest* du package *solitude* a été utilisée. Elle nécessite un seul paramètre : n le nombre d'arbres à réaliser. Nous avons observé la convergence de la moyenne des scores en fonction du nombre d'arbres et avons retenu $n = 100$, valeur à partir de laquelle la moyenne commence à être stable.

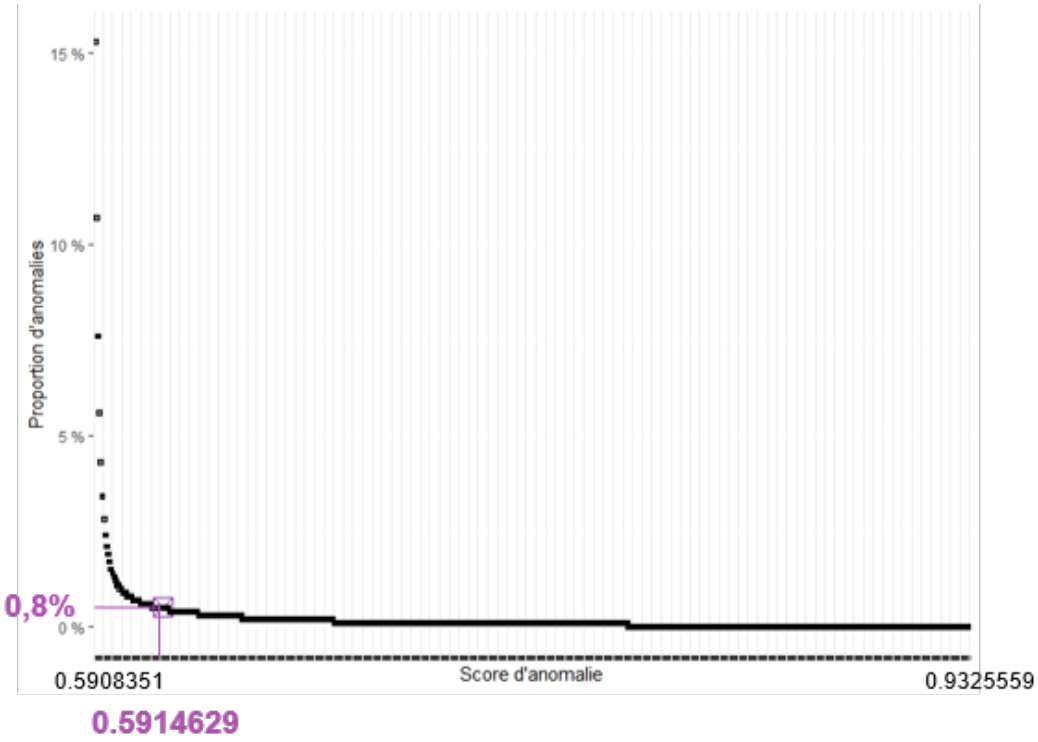
L'algorithme renvoie un score d'anomalie pour chacun des points. Rappelons que plus le score est élevé, plus le point est susceptible d'être une anomalie. La valeur des scores varie beaucoup d'un jeu de données quelconque à l'autre. Il n'existe pas de limite commune à tous les jeux de données au-delà de laquelle les données seraient anormales. Ainsi, nous avons décidé de tracer la proportion d'anomalies en fonction du seuil limite de score d'anomalie illustré en figure 5.5. Le score limite sélectionné est celui qui se trouve au niveau du "coude" de la courbe.

La figure 5.6 illustre les anomalies relevées sur les deux jeux de données réduits. Nous constatons que 93% des contrats sont étiquetés de la même façon par IF, qu'ils soient réduits en dimension par l'ACP ou par l'UMAP. Concernant les données de l'ACP, graphiquement aucune différence n'est visible avec la méthode des k-moyennes. Avec la réduction UMAP, les données atypiques se situent en très grande majorité aux bords de la figure. Cela est cohérent puisque IF se base sur la structure globale des données. En effet, Les points situés aux bords sont plus vite isolés selon le partitionnement d'IF.

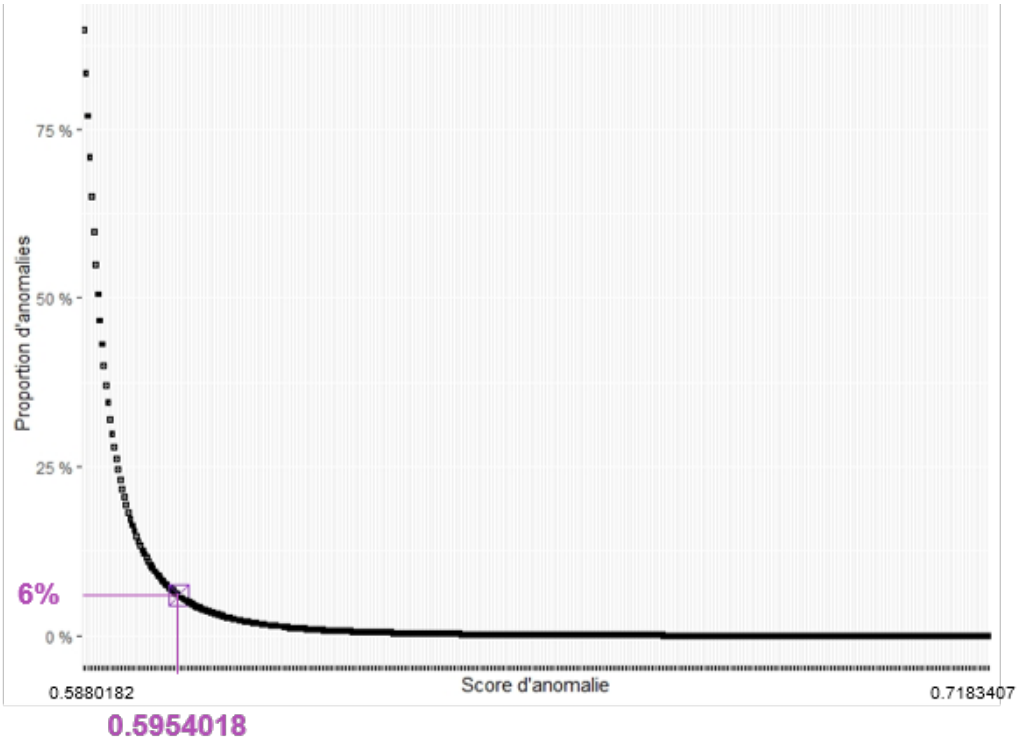
Algorithme de réduction	Proportion d'anomalies
ACP	0.8%
UMAP	6%

Table 5.5 – Proportion d'anomalies détectées par IF

Par ailleurs, nous avons appliqué IF à notre base de données initiale contenant 31 variables. IF nous le permet contrairement aux deux autres algorithmes d'apprentissage. Nous avons obtenu 2% d'anomalies parmi lesquels 11% sont également des anomalies constatées à partir de la base réduite. L'objectif étant de trouver un maximum d'anomalies, nous décidons de garder les anomalies détectées avec la base réduite dans la mesure où elles sont plus nombreuses.

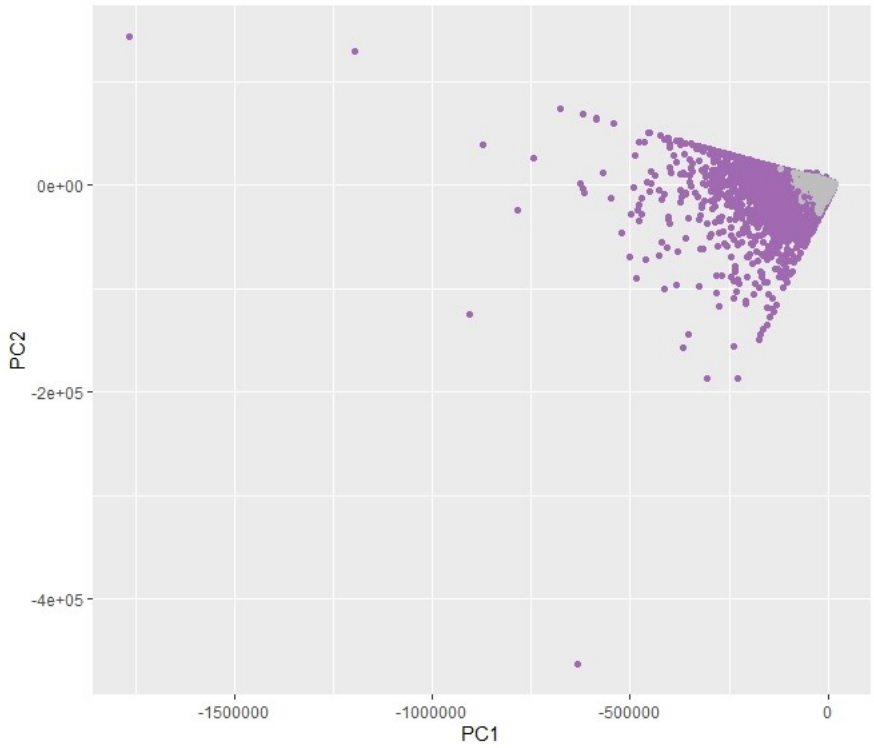


(a) ACP

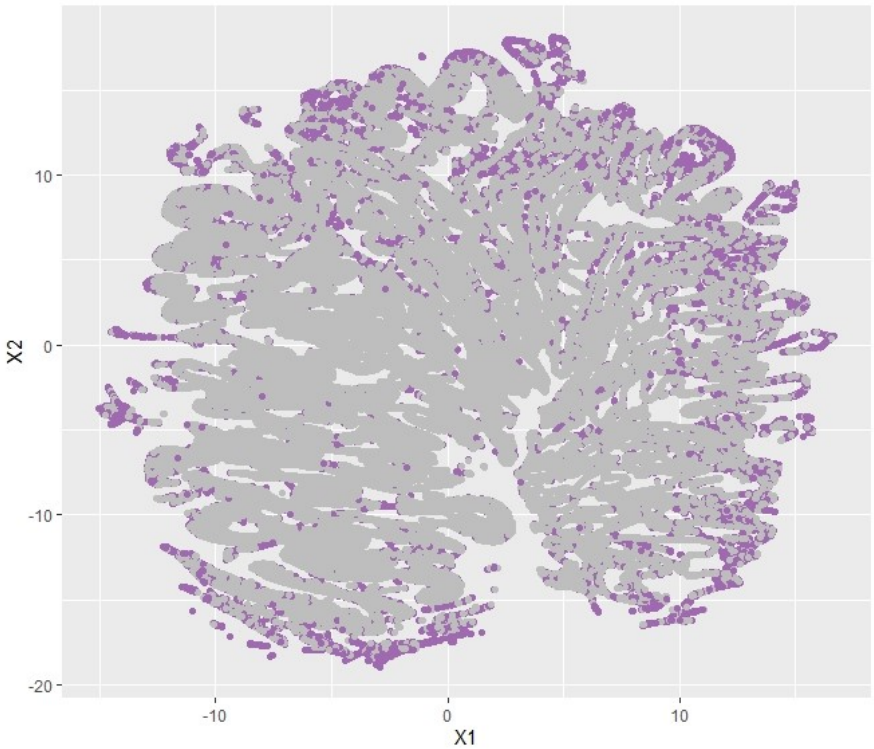


(b) UMAP

FIGURE 5.5 – Sélection du score maximum



(a) ACP



(b) UMAP

FIGURE 5.6 – Anomalies détectées par IF

À présent, il ne nous reste plus qu'à appliquer l'algorithme HDBSCAN à nos données. Cependant, du fait de la complexité élevée de cet algorithme, il est difficilement envisageable de l'appliquer directement à l'ensemble de nos données. Ainsi, nous avons contourné ce problème en partitionnant notre jeu de données.

5.1.4.3 Application de l'algorithme HDBSCAN

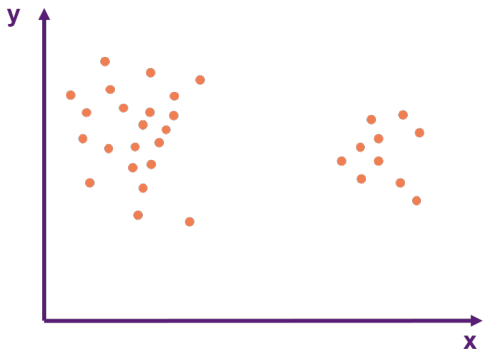
HDBSCAN fait partie de ces nombreux algorithmes d'apprentissage non-supervisés qu'il est difficile d'appliquer à l'ensemble de nos données pour cause de manque de mémoire vive. Là aussi, pour contourner ce problème, il est possible de partitionner nos données en blocs avant le lancement de l'algorithme HDBSCAN.

La question est alors la suivante : comment effectuer un partitionnement pertinent ? Pour y répondre, considérons tout d'abord un espace de données tel que représenté en (a), figure 5.7. Il serait alors envisageable de le décomposer en blocs en suivant un découpage sous forme de matrice, comme illustré en (b), 5.7. Dans ce cas, le découpage serait tellement fin, que certaines données pourraient passer pour des erreurs, car isolées dans un bloc (cellules sur-lignées), alors qu'elles appartiennent en réalité à un *cluster* de dimension plus importante que le bloc lui-même. Il apparaît donc clairement comme nécessaire d'avoir le moins possible de traits de découpage de l'espace de données. Voilà pourquoi nous recommandons dans un premier temps un découpage soit en tranches horizontales, soit en tranches verticales comme illustré en (c), figure 5.7. Dans un second temps, nous recommandons de choisir avec précaution la largeur des tranches de découpage. Dans ce sens, il apparaît évident que la largeur de chaque tranche de découpage doit être bien plus élevée que la distance moyenne entre deux points, sinon aucun regroupement de données ne pourra être effectué de manière pertinente par la suite par HDBSCAN. Un découpage idéal est illustré en (d), 5.7.

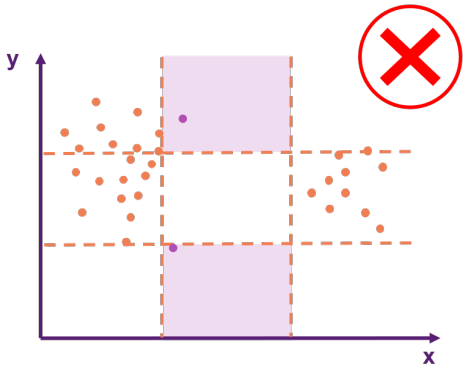
Dans cette sous-section, un partitionnement par découpage vertical sera appliqué uniquement aux données UMAP car leur répartition nous le permet. En effet, dans le cas de l'ACP, il est plus difficile de partitionner l'espace du fait que les données forment un cône. Puis l'algorithme HDBSCAN sera ensuite exécuté sur chacune des tranches obtenues.

Afin de choisir une première largeur de tranche d, il est possible de s'appuyer sur un graphique des k-distances. Nous proposons de reprendre le même graphique que celui qui serait utilisé pour déterminer le valeur optimale du paramètre ϵ de l'algorithme DBSCAN (cf. figure 5.8), avec k égal à $minPts$. Ce graphique permet d'obtenir la moyenne de la distance de chaque point par rapport à ses $minPts$ plus proches voisins, où $minPts$ est le paramètre commun à DBSCAN et HDBSCAN. Les distances en abscisse sont rangées par ordre croissant et la valeur optimale de ϵ se situe au niveau du "coude" de la courbe, soit : $\epsilon = 0,26$.

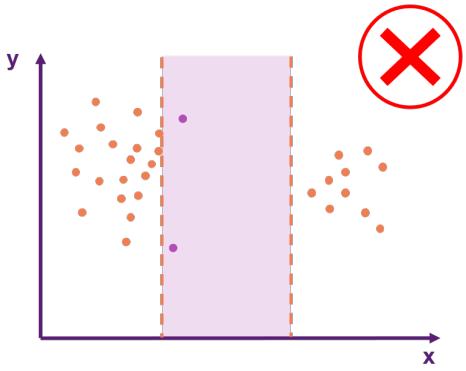
Rappelons que ϵ correspond au rayon des *clusters* et que par conséquent, notre largeur de tranche doit être bien plus élevée que celui-ci. Pour la figure 5.8, $minPts$ est égale à 100, valeur utilisée par la suite pour le calibrage d'HDBSCAN. Notons que ϵ varie très peu pour des $minPts$ entre 30 et 120.



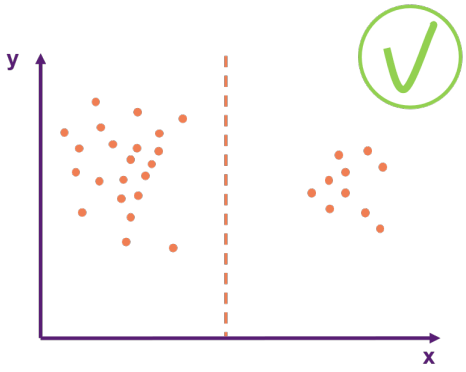
(a) Jeu de données quelconque



(b) 12 traits de découpage en matrice



(c) 2 traits de découpage vertical trop rapprochés



(d) 1 trait de découpage correctement réalisé

FIGURE 5.7 – Différentes possibilités de partitionnement

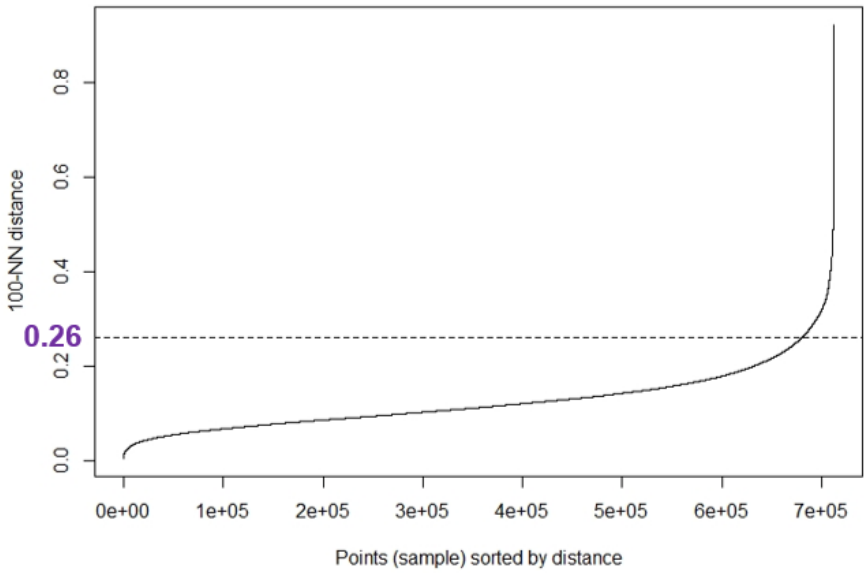


FIGURE 5.8 – Sélection de la largeur de tranche

Une première expérimentation a démontré qu'un ϵ de 0,26 conduit à des tranches ne contenant pas plus de 20 000 points. Or, HDBSCAN peut être lancé sur des groupes de données d'une taille maximale de 30 000 points. Ainsi, la largeur de bande peut être augmentée de telle façon que le nombre de points par tranches se rapproche davantage de 30 000. Finalement nous avons choisi de fixer la largeur de bande dans un premier temps à 0,55, valeur qui est bien au-delà du ϵ trouvé précédemment (cf. (a) de la figure 5.9 où $d = 0,55$).

Après une seconde expérimentation, nous constatons qu'une telle largeur de tranche conduit parfois à former des tranches ne contenant pas plus de 4 000 points, ce qui apparaît comme une trop faible densité. En effet, dans de telles conditions, soit ces points pourraient former différents *clusters* de petites tailles, soit certains points pourraient passer pour des anomalies. Le fait d'élargir la tranche créerait davantage de densité et regrouperait ces différents petits *clusters* en un seul et même *cluster* conséquent, minimisant ainsi le risque d'erreur d'isolement de points pertinents. Ainsi, certaines bandes ont été redimensionnées, voire regroupées, de sorte que finalement chaque tranche de l'espace ainsi découpé contienne entre 16 000 et 30 000 points (cf. (b) de la figure 5.9).

Dans un premier temps, HDBSCAN a été exécuté sur chacune de ces nouvelles tranches avec $minPts$ fixé à 100. Notons que plusieurs tests ont été effectués sur différentes tranches pour sélectionner la meilleur valeur de $minPts$. Ces tests consistent à tracer la dépendance entre la proportion d'anomalies et le paramètre $minPts$ pour plusieurs tranches. La meilleur valeur de $minPts$ pour chacune de ces tranches correspond au point situé au "coude" des courbes. Or, toutes les tranches ne proposent pas la même valeur de $minPts$. Ainsi, la valeur la plus optimale de $minPts$ peut être considérée comme celle ressortant le plus parmi toutes les tranches. Dans ce

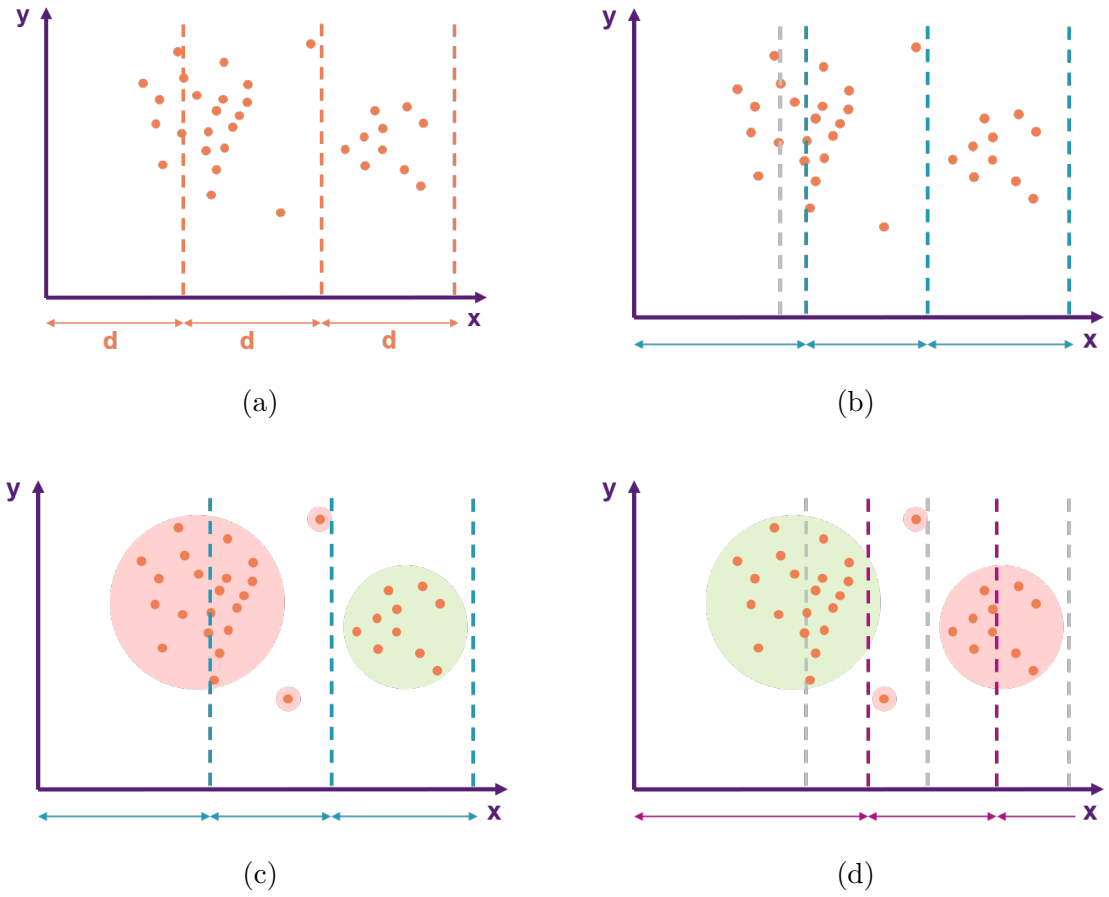


FIGURE 5.9 – Étapes du partitionnement

cas, la valeur finale de *minPts* est 100.

Il ne serait pas rigoureux de garder uniquement ce partitionnement dans le sens où des *clusters* peuvent être situés au niveau des frontières, i.e. à cheval sur un trait de découpage, auquel cas certains points pourraient être considérés comme des anomalies alors qu'ils ne le sont pas (cf. (c) de la figure 5.9). Ainsi, en complément, il est nécessaire d'effectuer un deuxième partitionnement avec des traits de découpage de tranches situés au milieu des tranches du premier partitionnement (cf. (d) de la figure 5.9). HDBSCAN est à nouveau lancé mais cette fois-ci sur chacune des tranches du deuxième partitionnement. Au final, nous considérons comme anomalies tous les points qui ont été détectés comme telles dans les deux partitionnements. Ces anomalies représentent in finé 8% de la base et sont représentées en figure 5.10, dans laquelle seuls les traits de découpage du premier partitionnement sont affichés.

Une fois les trois algorithmes de *clustering* effectués, tous les résultats peuvent être regroupés pour être certain de prendre seulement en compte les anomalies réelles.

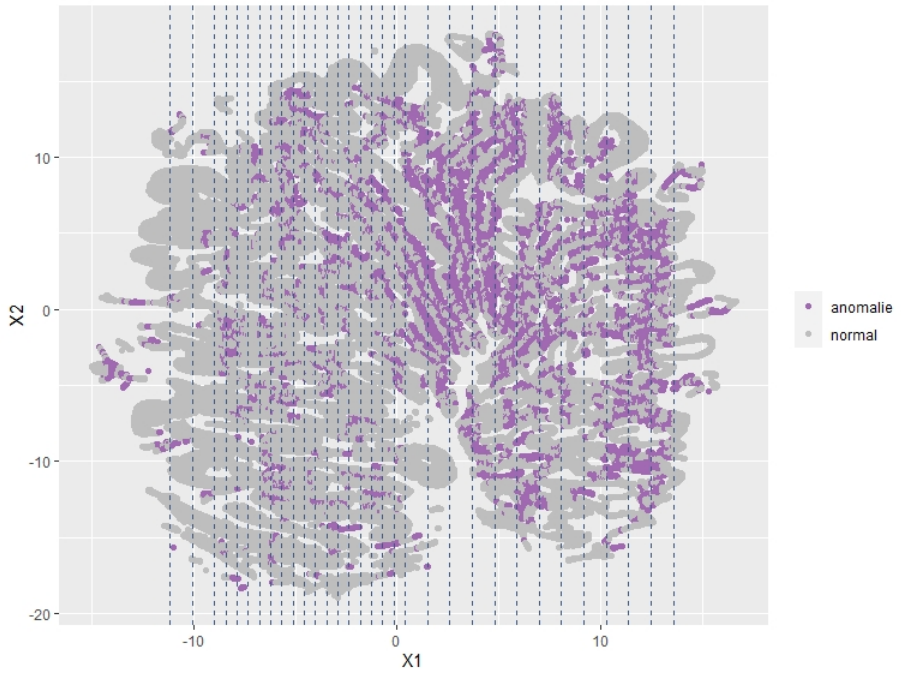


FIGURE 5.10 – Anomalies détectées par HDBSCAN

5.1.4.4 Regroupement des résultats

Les anomalies finales sont celles qui ont été labelisées anomalies par les trois algorithmes d'apprentissage et sont représentées en figure 5.11. L'ACP ayant été moins pertinente dans le cas des k-moyennes et inutilisable pour le partitionnement d'HDBSCAN, nous n'avons retenu que la réduction de dimension par UMAP.

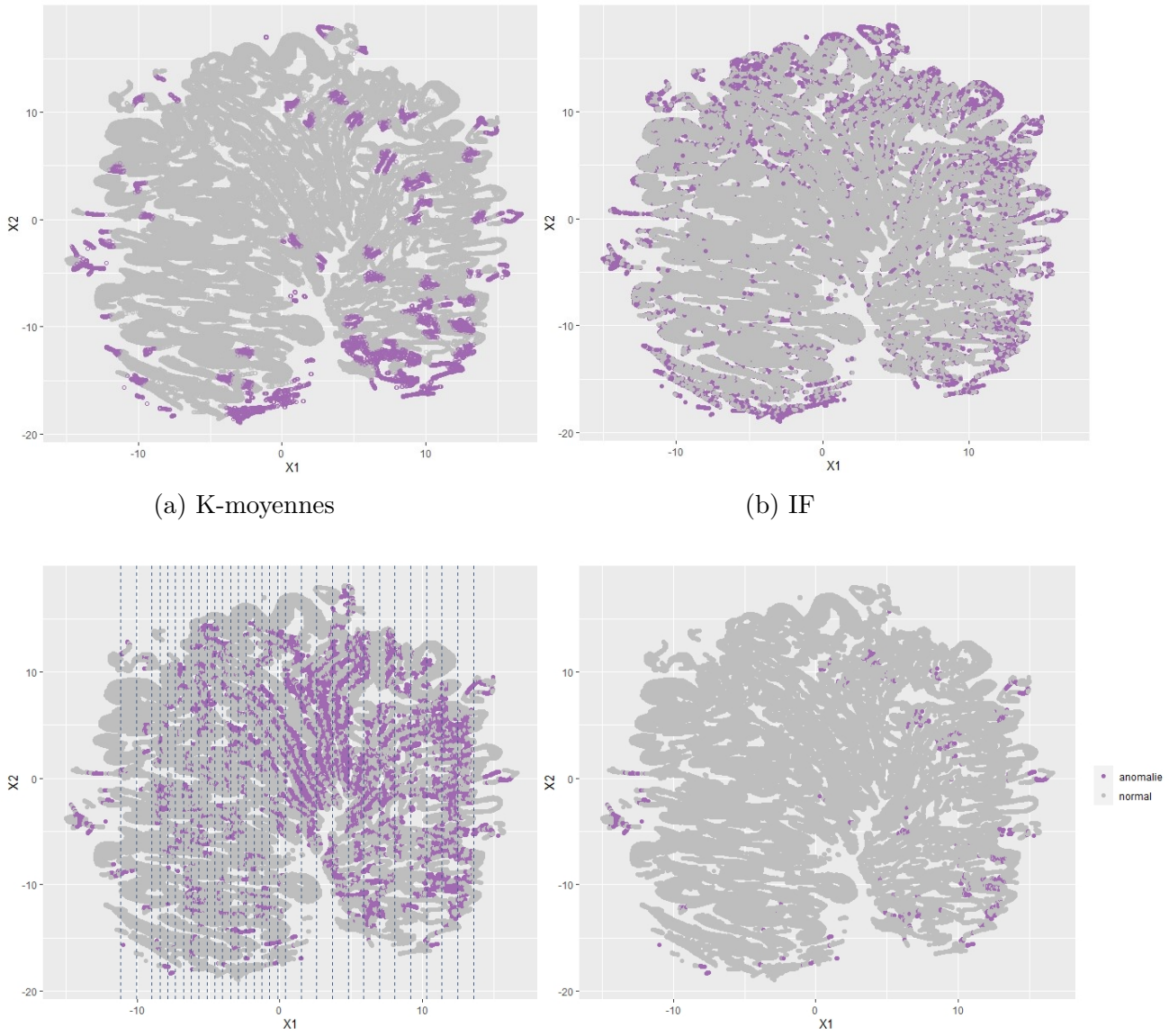
Sur la figure 5.11, les graphiques mettent bien en exergue les résultats obtenus à partir des trois algorithmes de *clustering*. Pour les k-moyennes, nous pouvons remarquer des groupes d'anomalies éparpillés. Pour IF, les anomalies sont plutôt situées en périphérie. Alors que pour HDBSCAN, les anomalies détectées sont mieux réparties, cela démontre bien son approche locale.

Algorithme	Proportion d'anomalies
K-moyennes	7%
IF	6%
HDBSCAN	8%
HDBSCAN - IF - k-moyennes	0,3%

TABLE 5.6 – Synthèse des anomalies détectées par les algorithmes de *clustering*

Comme illustrées dans le tableau 5.6, les proportions d'anomalies détectées par les algorithmes sont assez proches. Lorsque nous prenons l'intersection des trois groupes d'anomalies, le nombre d'anomalies est considérablement réduit : 0,3% des données.

À présent, il est intéressant d'analyser plus profondément ces anomalies.



(a) K-moyennes

(b) IF

(c) HDBSCAN

(d) HDBSCAN - IF - K-moyennes

FIGURE 5.11 – Anomalies sélectionnées

5.1.5 Analyse des anomalies

Dans cette section, nous effectuerons des analyses sur les anomalies obtenues pour tenter de comprendre la cause de leur présence dans la base et de vérifier si les données incohérentes détectées dans la partie 5.1.2 ont également été identifiées comme anormales par les algorithmes de *clustering*. Pour finir, nous tenterons de vérifier la cohérence des résultats en utilisant la loi de Benford, laquelle a été beaucoup étudiée ces dernières années dans la littérature. Enfin, nous appliquerons une méthode semi-supervisée à nos anomalies pour s'assurer de l'efficacité des algorithmes de *clustering*.

5.1.5.1 Analyses univariées sur les anomalies

Il est intéressant de comparer les anomalies obtenues par les algorithmes de *clustering* avec celles détectées à partir des analyses univariées en section 5.1.2. La plupart des données atypiques identifiées en analyse univariée n'ont pas été détectées anormales par les algorithmes d'apprentissage, comme le montre le tableau 5.7. Il est fort probable que cela soit dû à la réduction de dimension, opération au cours de laquelle des informations ont été perdues lors du croisement des différentes variables. Des *violin plots* sont disponibles en annexe B, représentant les densités de variables quantitatives pour la base de donnée initiale et les anomalies.

Variable	Identification des contrats atypiques	Proportion de contrats dans la base de données initiale	Proportion de contrats présents dans les anomalies détectées par les algorithmes
Durée du contrat	> 10 ans	0,6%	2%
Maturité du contrat	< 2017	$< 0,1\%$	0%
Taux de prime	Valeurs très faibles	13%	36%
Capital initial	Valeurs très faibles	$< 0,1\%$	0%
Capital initial	Valeurs très élevées	0,2%	0%
Capital restant dû	Contrats ayant un capital restant dû supérieur au capital initial, n'ayant pas de différé total et n'appartenant pas à une banque proche de la Suisse (prêts en devise)	$< 0,1\%$	0%

TABLE 5.7 – Analyse des anomalies détectées par les algorithmes de *clustering*

Il est important de relever que les algorithmes ont détecté 36% de contrats ayant un taux de prime faible, alors que l'analyse univariée n'en relevait que 13%. Rappelons

que les taux de prime pour les prêts immobiliers sont nettement plus faibles que pour les prêts personnels. Seuls les prêts étudiants dans les prêts personnels peuvent avoir des taux de prime faibles. Or, les contrats ayant des taux de prime très faibles dans la base ne concernent pas seulement des prêts étudiants. Aussi, nous en concluons une éventuelle mauvaise classification de contrats : certains prêts immobiliers sont potentiellement présents dans cette base, alors que cette dernière ne devrait contenir que des prêts personnels.

5.1.5.2 Application de la Loi de Benford au capital restant dû

Pour vérifier la cohérence de nos résultats, nous allons ici nous intéresser à la loi de Benford.

Dans notre cas, sans tenir compte des contrats qui ont été souscrits en 2020, seule la variable "capital restant dû" est aléatoirement distribuée. En effet, les contrats souscrits en 2020 ne doivent pas être pris en compte dans l'application de la loi de Benford puisqu'ils sont égaux aux capitaux initiaux. En tant que tel, leur valeur ne peut être aléatoirement distribuée dans la mesure où environ 50% des capitaux initiaux appartiennent à un ensemble de 10 nombres ronds (10 000, 15 000, 20 000, etc.).

Le package *benford.analysis* a été utilisé pour appliquer la loi de Benford à nos capitaux restant dus. Les résultats sont illustrés en figure 5.12. Par analyse du premier chiffre significatif, les nombres commençant par le chiffre 2 sont suspectés, c'est-à-dire qu'il est fortement probable que des anomalies se situent dans les contrats dont le capital restant dû commence par le chiffre 2 (par exemple 20 000€). Cela est cohérent avec les analyses des sections précédentes puisque 53% des anomalies trouvées par les algorithmes de *clustering* sont des contrats de ce type, alors qu'ils représentent 30% de la base initiale.

Par ailleurs, plusieurs des capitaux restant dus commençant par 10 ou 70 (par exemple 10 000€ ou 70 000€) doivent être mal renseignés. Ce sont les deux premiers chiffres pour lesquels les différences au carré entre la loi de Benford et les données sont les plus élevées. Cela est de nouveau cohérent avec les résultats des algorithmes d'apprentissage puisque 45% des anomalies sont des contrats avec des capitaux restant dus commençant par 10 ou 70, alors que ces contrats représentent 18% de la base initiale. En utilisant une différence absolue entre la loi de Benford et les données, ce sont les capitaux restant dus commençant par 10 et 15 qui sont suspectés. Ils représentent 46% des anomalies contre 20% dans la base initiale.

La loi de Benford confirme que la sélection des anomalies est cohérente si l'on regarde leurs capitaux restant dus.

À présent, nous allons utiliser une dernière approche pour être certains de nos anomalies.

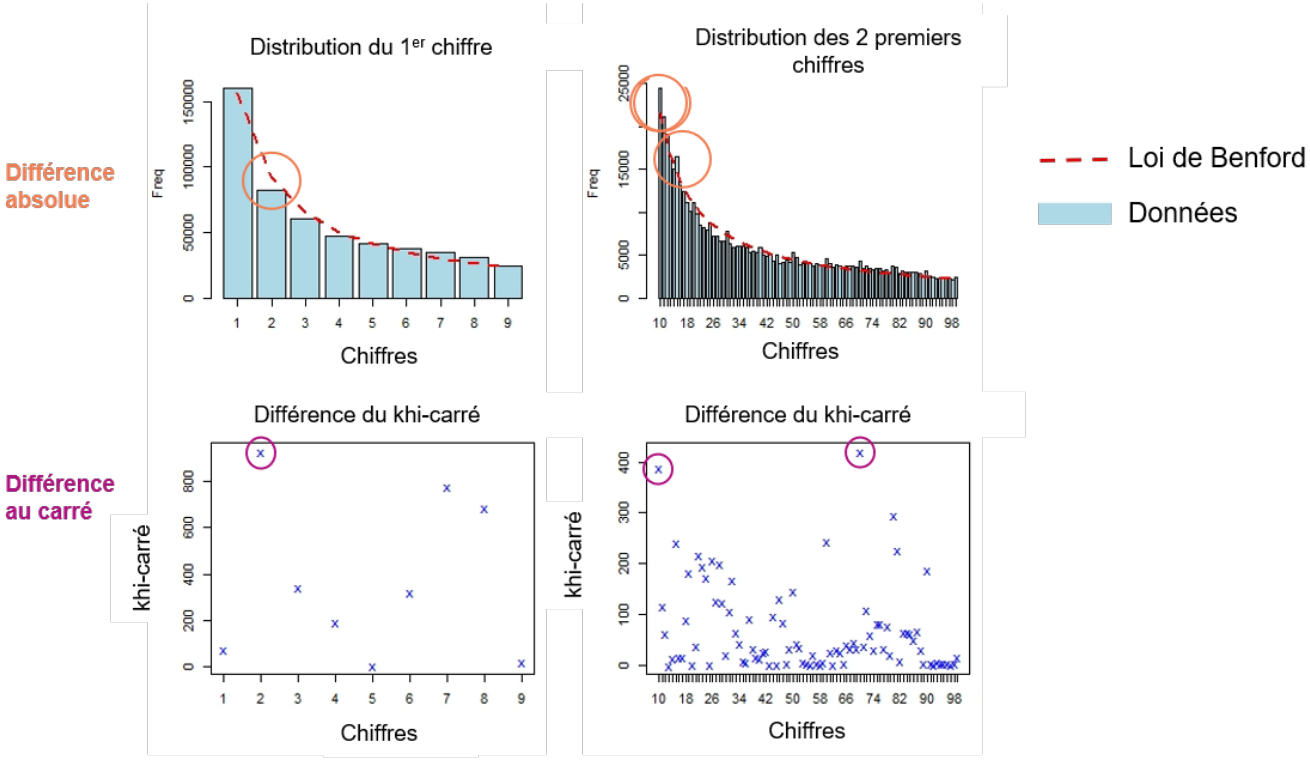


FIGURE 5.12 – Loi de Benford appliquée au capital restant dû

5.1.5.3 Application du *PU learning* aux anomalies

Dans la mesure où nous ne disposons d'aucun jeu de données étiquetées, nous allons choisir comme jeu de données pour le *PU learning* les bases P et U suivantes (cf. section 2.3.2).

- P : base contenant uniquement les points avec les plus petits scores d'anomalies donnés par IF (rappelons que plus le score d'anomalie est bas, plus le point analysé semble normal selon IF) ;
- U : base contenant uniquement les anomalies détectées par les algorithmes de *clustering*.

Pour la construction de l'algorithme, nous nous sommes inspirés des travaux de J. Long [6] et avons choisi la technique de l'espion pour confirmer les anomalies de la base U.

Après application, nous constatons que seulement 0,6% des contrats détectés comme anomalies, ne seraient pas des anomalies. Cela montre bien l'efficacité des trois algorithmes non-supervisés.

Par ailleurs, dix contrats anormaux selon les algorithmes de *clustering* et le *PU learning* ont été sélectionnés au hasard pour être analysés en détail. Ils semblent être anormaux pour l'une ou plusieurs des raisons suivantes :

- Taux d'intérêt élevé par rapport au profil de l'assuré ;
- Taux de prime élevé par rapport au profil de l'assuré ;
- Catégorie socio-professionnelle anormale par rapport au profil de l'assuré ;
- Amortissement de l'emprunt inexact.

Seul un contrat parmi les dix ne semble pas contenir d'anomalies. Il s'agit d'un cas extrême où l'âge de l'assuré est très élevé. Ainsi, il est fortement possible que les algorithmes aient isolé ce cas du fait de son caractère inhabituel. Cela montre bien que les algorithmes ne sont pas fiables à 100%. Il peuvent parfois retourner des contrats qui leur semblent erronés alors qu'ils contiennent en réalité aucune erreur. Cependant, ils permettent d'éliminer une très grande partie des contrats qui ne contiennent pas d'erreurs. Ainsi, il est beaucoup plus rapide d'analyser les contrats un par un pour s'assurer des anomalies. Les algorithmes de *clustering* constituent donc un réel gain de temps.

5.2 Estimation des erreurs liées aux hypothèses

Dans cette section, nous nous intéressons aux erreurs liées aux hypothèses. Pour ce faire, la loi de mortalité relative aux femmes, la loi de mortalité relative aux hommes et la loi de rachat seront étudiées. Les données utilisées pour la construction de ces lois ont été soumises à une censure à droite de type III, comme expliqué dans la section 3.3.2.

L'estimateur de Kaplan-Meier, présenté en section 3.3.3.1, a été utilisé pour déterminer les taux bruts de ces trois lois. Ensuite, ces taux bruts ont été lissés avec la méthode des noyaux discrets, laquelle est décrite dans la section 3.3.4.2.

5.2.1 Étude des biais associés aux lois d'expérience

Trois biais parmi les cinq présentés dans la section 3.2 seront évalués, à savoir le biais d'homogénéité, le biais temporel et le biais d'estimation. Pour rappel, l'objectif est de construire un intervalle de confiance global pour chacune des trois lois d'expérience. Pour ce faire, un intervalle de confiance sera déterminé pour chacun des trois biais, permettant ensuite d'obtenir un intervalle de confiance global.

Afin de valider l'estimation des hypothèses, chaque borne supérieure des intervalles de confiance sera comparée aux seuils décrits en section 3.2 et rappelés ci-dessous :

- Si $\Delta_r[IC_{B,sup}, \hat{q}] < 25\%$, la précision de l'estimation est jugée satisfaisante ;
- Si $25\% \leq \Delta_r[IC_{B,sup}, \hat{q}] < 100\%$, la précision de l'estimation est tout juste suffisante et nécessite une analyse complémentaire pour améliorer la modélisation ;
- Si $100\% \leq \Delta_r[IC_{B,sup}, \hat{q}]$, la précision de l'estimation est insuffisante et l'hypothèse n'est pas valide.

Dans notre cas, la période d'observation de la mortalité s'étale sur environ six ans, et plus précisément de janvier 2015 à septembre 2021. Quant aux rachats, ils seront observés sur plus de six ans, de janvier 2015 à mars 2021.

5.2.1.1 Évaluation du biais d'homogénéité

La démarche suivie pour déterminer le biais d'homogénéité est la même que celle présentée en section 3.2.4.1. Elle est davantage détaillée ci-dessous :

1. **Création des groupes de risque :** Des groupes de risque sont formés à partir des catégories socio-professionnelles pour chacune des lois utilisées.

   Pour ce faire, des ratios correspondant respectivement aux décès ou rachat observés sur les décès ou rachats attendus sont calculés sur l'ensemble de la période d'observation pour chacune des catégories socio-professionnelles. Plus précisément, ces ratios s'écrivent de la façon suivante :

   $$Ratio_{obs/att} = \frac{d_i}{\sum_i e_i * \hat{q}_i}$$

   avec

   - i représentant l'âge ou l'ancienneté de l'assuré ;
   - d_i le nombre de décès ou de rachat observés à l'âge i ou l'ancienneté i ;
   - e_i le nombre d'individus d'âge i ou d'ancienneté i ;
   - $\hat{q}_i$ les taux estimés de mortalité ou de rachat.

   Les catégories socio-professionnelles ayant des ratios semblables sont regroupées pour former des groupes de risque, de telle sorte que les volumes d'assurés dans chacun des groupes soient à peu près égaux. Au total, entre sept et neuf groupes de risque sont créés dans les portefeuilles utilisés pour la construction des différentes lois.

2. **Lancement du *bootstrap* :** 1000 échantillons "*bootstrappés*" sont construits. Puis pour chacun des ces échantillons, des taux bruts, appelés taux bruts "*bootstrappés*", sont calculés à partir de l'estimateur de Hoem, lequel étant décrit en section 3.3.3.2.

3. **Détermination des bornes brutes de l'intervalle de confiance :** Des intervalles de confiance sont calculés selon la méthode des percentiles simples en prenant le quantile à 95% des taux bruts "*bootstrappés*". La figure 5.14 illustre les intervalles ainsi obtenus et fait apparaître ce que l'on appelle les "taux lissés centraux", à savoir les taux lissés de mortalité ou de rachat provenant des lois initiales étudiées.

4. **Lissage des bornes de l'intervalle de confiance :** Les écarts entre les bornes brutes et les taux lissés centraux sont lissés par moyenne mobile d'ordre 2, une méthode de lissage présentée en section 3.3.4.1. Enfin, les bornes finales de l'intervalle de confiance correspondent aux taux lissés centraux auxquels sont ajoutés ces écarts lissés, comme illustrées en figure 5.15.

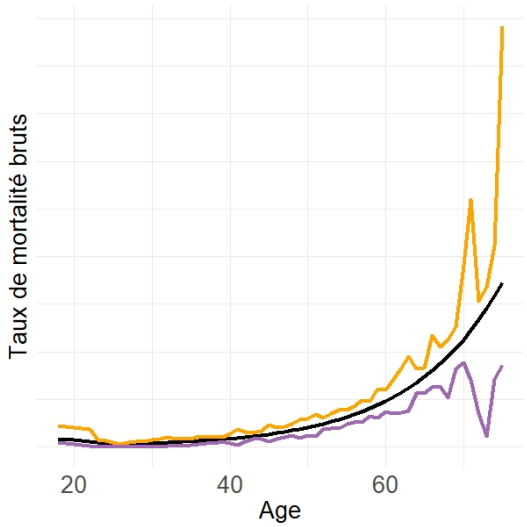
	Biais d'homogénéité	
Tranches d'âge	**Loi de mortalité Femmes**	**Loi de mortalité Hommes**
[18:24]	162%	191%
[25:31]	66%	47%
[32:38]	67%	53%
[39:45]	54%	34%
[46:52]	43%	21%
[53:75]	41%	27%
Tranches d'ancienneté	**Loi de rachat**	
[0:4]	9%	
[5:9]	13%	
[9:14]	44%	

FIGURE 5.13 – Moyenne des écarts relatifs entre la borne supérieure du biais d'homogénéité et le flux réel

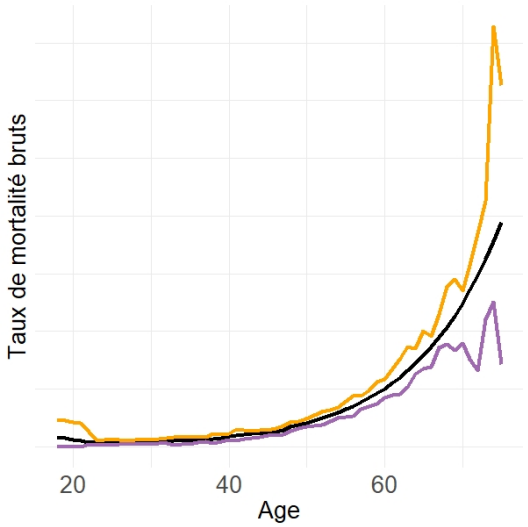
A partir des écarts relatifs détaillés dans le tableau 5.13, les avis suivants peuvent être émis sur l'estimation des lois d'expérience vis-à-vis de l'homogénéité du portefeuille :

- Les écarts relatifs sur la **loi de mortalité Femmes** et sur la **loi de mortalité Hommes** sont suffisants sur les tranches d'âge ayant le plus d'assurés. Aussi, la précision de l'estimation pour les deux lois est jugée tout juste suffisante et nécessite une étude d'impact complémentaire sur les écarts d'expérience, i.e. sur la différence entre les flux projetés et les flux réels. Cette analyse sera faite dans la section 5.3.2 ;
- Les écarts relatifs sur la **loi de rachat** sont satisfaisants sur les neuf premières années d'ancienneté. Or, la durée moyenne d'un prêt personnel étant située entre trois et quatre ans, l'impact constaté pour des anciennetés de neuf à quatorze ans est faible. Aussi, la précision de l'estimation est jugée satisfaisante.

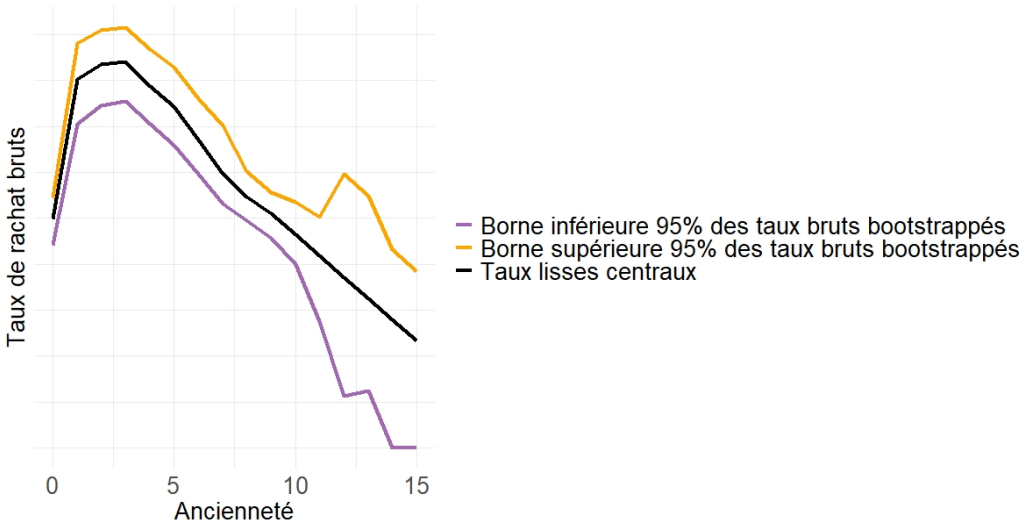
Notons que les bornes de ces intervalles de confiance représentent les pires scénarios possibles en terme de mortalité ou de rachat, du fait de l'hétérogénéité du portefeuille utilisé pour la construction.



(a) Loi de mortalité Femmes

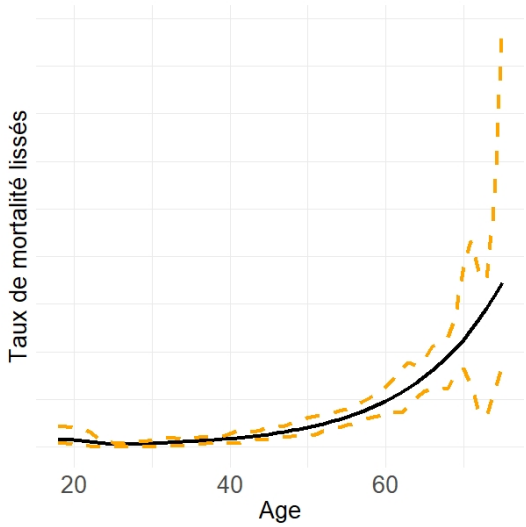


(b) Loi de mortalité Hommes

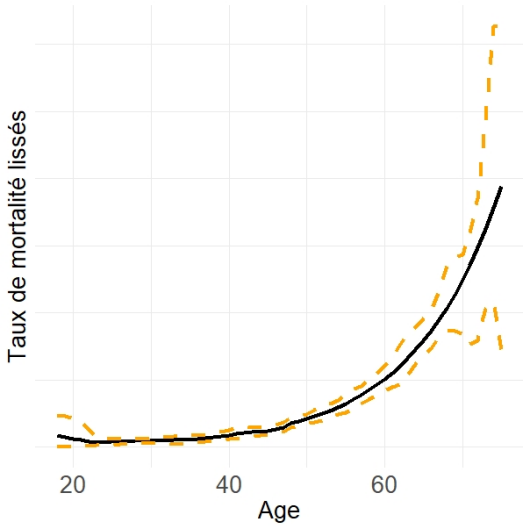


(c) Loi de rachat

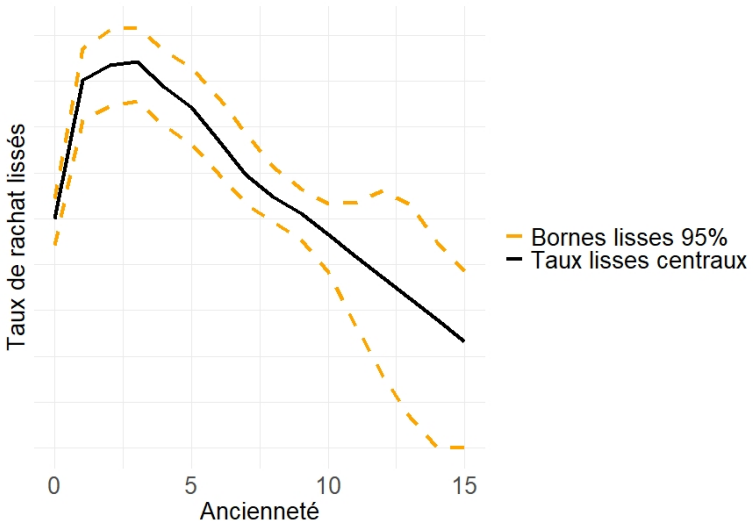
FIGURE 5.14 – Bornes brutes du biais d'homogénéité



(a) Loi de mortalité Femmes



(b) Loi de mortalité Hommes



(c) Loi de rachat

FIGURE 5.15 – Bornes lissées du biais d'homogénéité

A présent, passons à l'évaluation du deuxième biais de modélisation, à savoir le biais temporel.

5.2.1.2 Évaluation du biais temporel

Comme expliqué en section 3.2.4.2, le biais temporel est également évalué au travers d'un intervalle de confiance. Ce dernier est construit de la même manière que celui défini pour le biais d'homogénéité en section précédente. Bien entendu, seule la création des groupes de risque se fait d'une autre manière.

Pour le biais temporel, un groupe de risque correspond à un trimestre. En d'autres termes, un contrat d'assurance appartient à un groupe de risque s'il a été présent en portefeuille pendant le trimestre concerné.

Les résultats du biais temporel sont présentés en figure 5.17 et 5.18.

	Biais temporel	
Tranches d'âge	**Loi de mortalité Femmes**	**Loi de mortalité Hommes**
[18:24]	129%	146%
[25:31]	102%	62%
[32:38]	70%	49%
[39:45]	48%	36%
[46:52]	38%	22%
[53:75]	27%	16%
Tranches d'ancienneté	Loi de rachat	
[0:4]	2%	
[5:9]	6%	
[9:14]	61%	

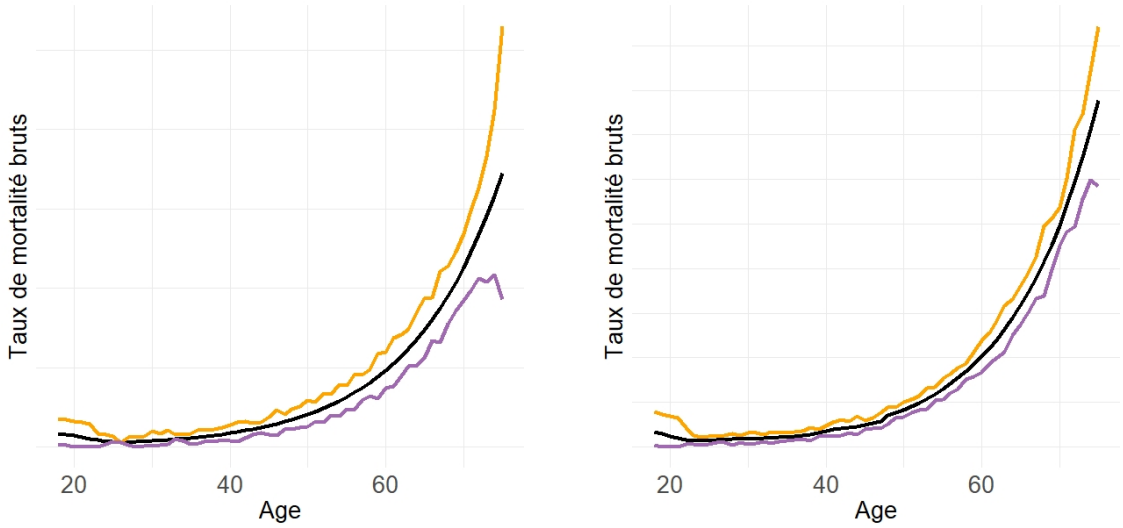
FIGURE 5.16 – Moyenne des écarts relatifs entre la borne supérieure du biais temporel et le flux réel

A partir des écarts relatifs détaillés dans le tableau 5.16, les avis suivants peuvent être émis sur l'estimation des lois d'expérience vis-à-vis de l'évolution temporelle du portefeuille :

- Les écarts relatifs sur la **loi de mortalité Femmes** et sur la **loi de mortalité Hommes** sont suffisants sur les tranches d'âge ayant le plus d'assurés. Aussi, la précision de l'estimation pour les deux lois est jugée tout juste suffisante et nécessite une étude d'impact complémentaire sur les écarts d'expérience, i.e. sur la différence entre les flux projetés et les flux réels. Cette analyse sera faite dans la section 5.3.2 ;
- Les écarts relatifs sur la **loi de rachat** sont satisfaisants sur les neuf premières années d'ancienneté. De même que pour le biais d'homogénéité, la durée moyenne d'un prêt personnel étant située entre trois et quatre ans, l'impact constaté pour des anciennetés de neuf à quatorze ans est faible. Aussi, la précision de l'estimation est jugée satisfaisante.

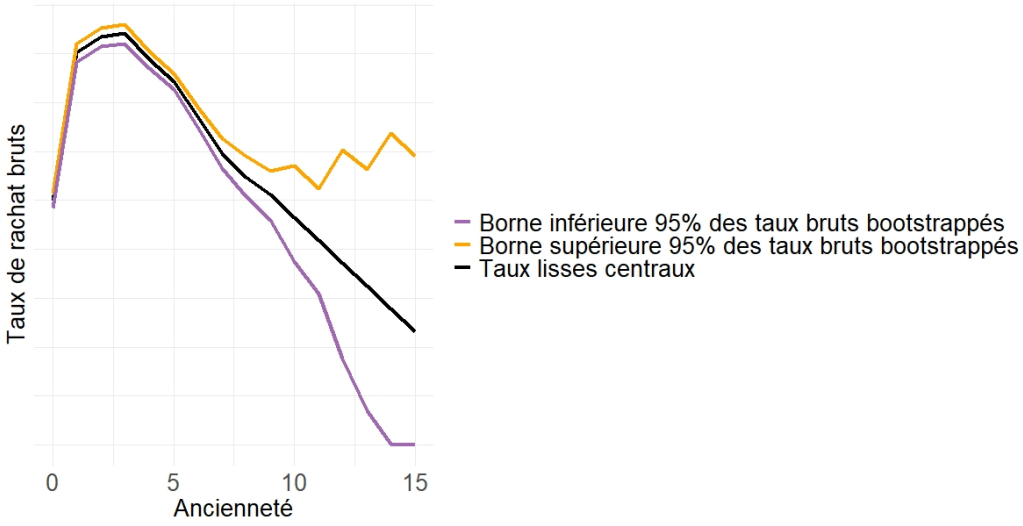
Tout comme pour le biais d'homogénéité, les bornes des intervalles de confiance représentent les pires scénarios possibles.

Évaluons maintenant le dernier biais sur les hypothèses, à savoir le biais d'estimation.



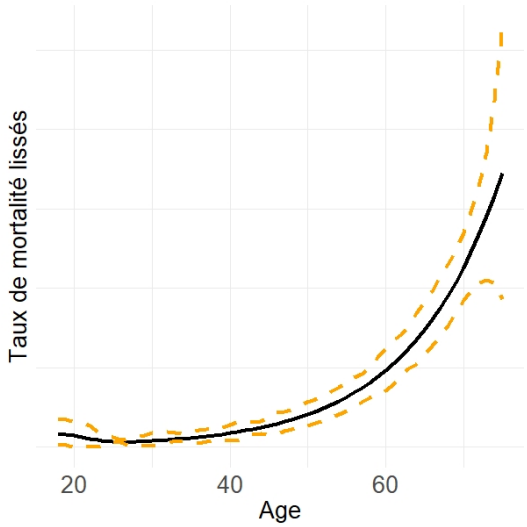
(a) Loi de mortalité Femmes

(b) Loi de mortalité Hommes

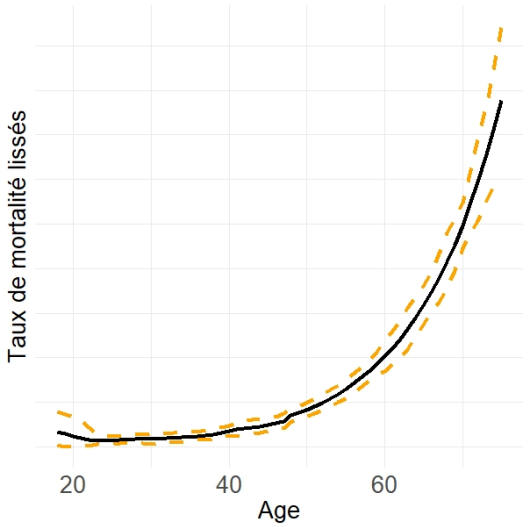


(c) Loi de rachat

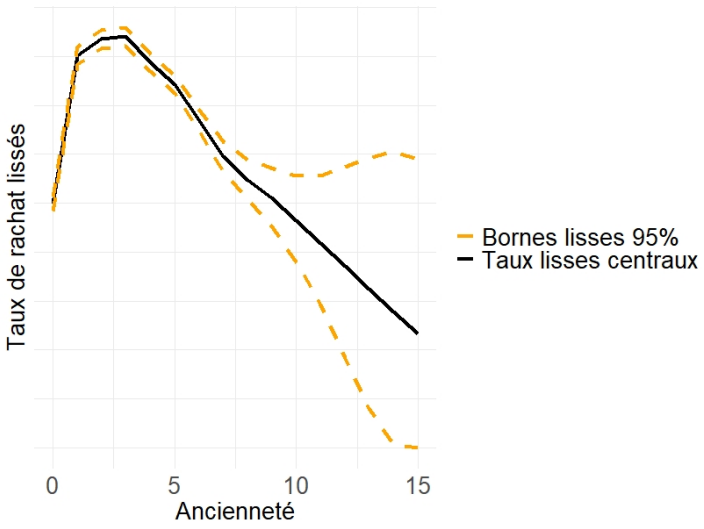
FIGURE 5.17 – Bornes brutes représentant le biais temporel



(a) Loi de mortalité Femmes



(b) Loi de mortalité Hommes



(c) Loi de rachat

FIGURE 5.18 – Bornes lissées représentant le biais temporel

5.2.1.3 Évaluation du biais d'estimation

Le biais d'estimation est également estimé à partir d'un intervalle de confiance, comme décrit en section 3.2.3.

Un décès ou un rachat correspondent à des évènements et peuvent donc être représentés par une indicatrice renseignant la réalisation ou non de l'évènement. En ce sens, le biais d'estimation des lois de mortalité et de rachat peut être estimé par l'intervalle de confiance à 95% d'une loi binomiale.

Ainsi, il s'écrit de la façon suivante :

$$IC_{Binom} = \left[max(0, \hat{q} - Q_{0.95}\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}); min(1, \hat{q} + Q_{0.95}\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}) \right]$$

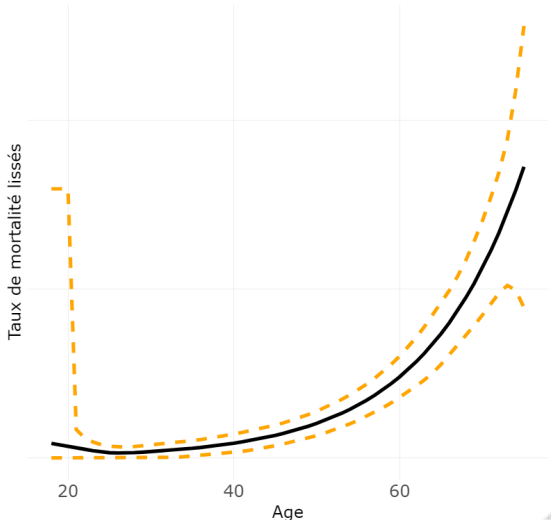
avec

- $\hat{q}$ les taux lissés centraux ;
- n l'exposition sur la période d'observation ;
- $Q_{0.95}$ le quantile à 95% d'une loi normale centrée réduite.

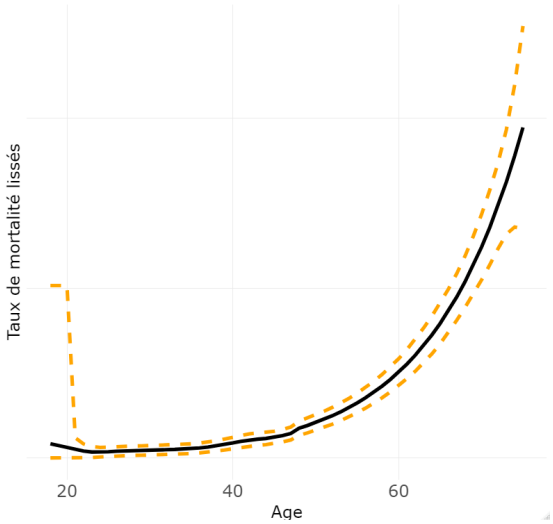
Les résultats du biais d'estimation sont présentés en figure 5.19.

A partir des écarts relatifs détaillés dans le tableau 5.20, les avis suivants peuvent être émis sur l'estimation des lois d'expérience vis-à-vis de l'échantillonnage du portefeuille :

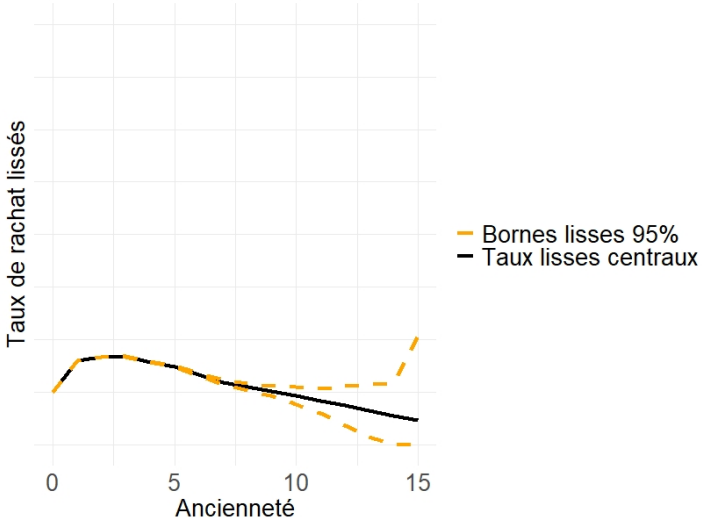
- Les écarts relatifs sur la **loi de mortalité Femmes** et sur la **loi de mortalité Hommes** sont suffisants sur les tranches d'âge ayant le plus d'assurés. Aussi, la précision de l'estimation pour les deux lois est jugée tout juste suffisante et nécessite une étude d'impact complémentaire sur les écarts d'expérience, i.e. sur la différence entre les flux projetés et les flux réels. Cette analyse sera faite dans la section 5.3.2 ;
- Les écarts relatifs sur la **loi de rachat** sont satisfaisants sur les neuf premières années d'ancienneté. De même que pour les deux biais précédents, la durée moyenne d'un prêt personnel étant située entre trois et quatre ans, l'impact constaté pour des anciennetés de neuf à quatorze ans est faible. Aussi, la précision de l'estimation est jugée satisfaisante.



(a) Loi de mortalité Femmes



(b) Loi de mortalité Hommes



(c) Loi de rachat

FIGURE 5.19 – Bornes représentant le biais d'estimation

	Biais d'estimation	
Tranches d'âge	**Loi de mortalité Femmes**	**Loi de mortalité Hommes**
[18:24]	>200%	>200%
[25:31]	114%	66%
[32:38]	82%	52%
[39:45]	56%	35%
[46:52]	37%	23%
[53:75]	28%	17%
Tranches d'ancienneté	Loi de rachat	
[0:4]	<1%	
[5:9]	4%	
[9:14]	56%	

FIGURE 5.20 – Moyenne des écarts relatifs entre la borne supérieure du biais d'estimation et le flux réel

A présent que les trois biais ont été estimés, le biais total sur chacune des hypothèses va pouvoir être déterminé dans la section suivante.

5.2.1.4 Évaluation du biais total

Afin d'obtenir les bornes supérieure et inférieure du biais total des hypothèses, nous allons regrouper le biais d'homogénéité, le biais temporel et le biais d'estimation. Les biais totaux sont représentés en figure 5.22 pour chacune des lois d'expérience et sont construits de la façon suivante :

- Le biais total de la **loi de rachat** est obtenu en additionnant le biais d'homogénéité, le biais temporel et le biais d'estimation ;
- En revanche, les biais totaux de la **loi de mortalité Femmes** et de la **loi de mortalité Hommes** sont obtenus en prenant le maximum du biais d'homogénéité, du biais temporel et du biais d'estimation. Pour rappel, les bais d'homogénéité et temporel s'évaluent notamment à partir des expositions sur chacun des âges. Certaines tranches d'âges étant très peu représentées dans le portefeuille de construction, un biais d'estimation apparaît au travers de l'évaluation des biais d'homogénéité et temporel. C'est pourquoi il est préférable de prendre le maximum des trois biais au lieu de les additionner.

Plus l'intervalle de confiance de ce biais total est étroit, plus l'erreur sur l'estimation de la loi est faible. Ainsi, un intervalle étroit signifie que la loi d'expérience est correctement estimée vis-à-vis du portefeuille actuel.

Concernant les lois de mortalité, l'erreur est beaucoup plus importante aux bords, c'est-à-dire pour les populations très jeunes et très âgées. Cela s'explique par un

manque d'exposition sur ces âges (effectifs faibles en portefeuille).

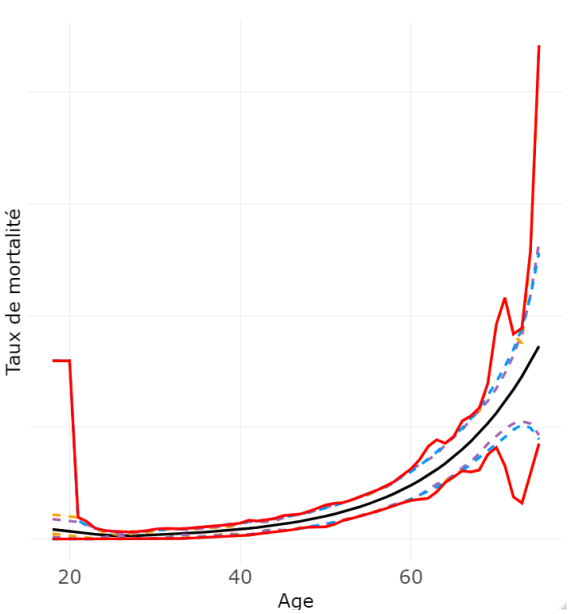
A partir des écarts relatifs détaillés dans le tableau 5.21, les avis suivants peuvent être émis sur l'estimation globale des lois d'expérience :

- Les écarts relatifs sur la **loi de mortalité Femmes** et sur la **loi de mortalité Hommes** sont insuffisants sur les tranches d'âge ayant le plus d'assurés. Aussi, la précision de l'estimation pour les deux lois est jugée insuffisante et nécessite d'autant plus une étude d'impact complémentaire sur les écarts d'expérience, i.e. sur la différence entre les flux projetés et les flux réels. Cette analyse sera faite dans la section 5.3.2 ;
- Les écarts relatifs sur la **loi de rachat** sont satisfaisants sur les neuf premières années d'ancienneté. De même que pour les trois biais précédents, la durée moyenne d'un prêt personnel étant située entre trois et quatre ans, l'impact constaté pour des anciennetés de neuf à quatorze ans est faible. Aussi, la précision de l'estimation est jugée satisfaisante.

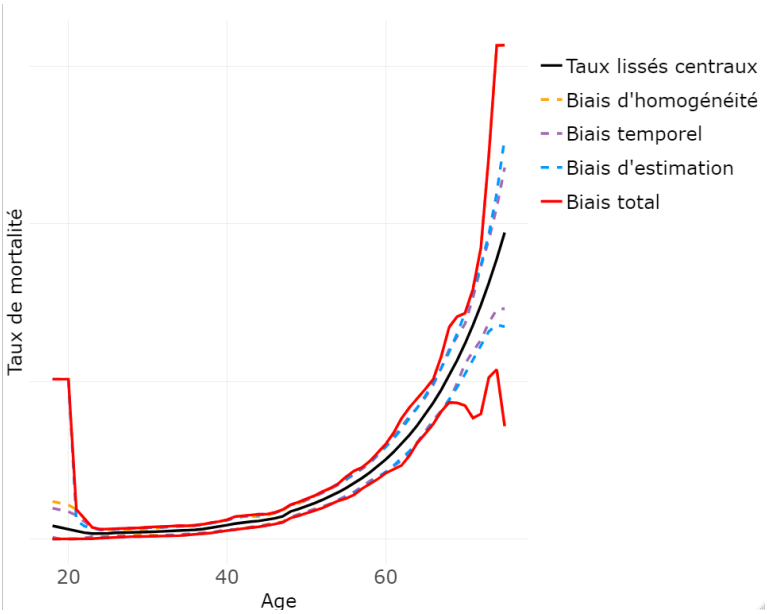
	Biais d'homogénéité	Biais temporel	Biais d'estimation	Biais total
Tranches d'âge		**Loi de mortalité Femmes**		
[18:24]	162%	129%	>200%	>200%
[25:31]	66%	102%	114%	114%
[32:38]	67%	70%	82%	82%
[39:45]	54%	48%	56%	56%
[46:52]	43%	38%	37%	43%
[53:75]	41%	27%	28%	41%
Tranches d'âge		**Loi de mortalité Hommes**		
[18:24]	191%	146%	>200%	>200%
[25:31]	47%	62%	66%	66%
[32:38]	53%	49%	52%	53%
[39:45]	34%	36%	35%	36%
[46:52]	21%	22%	23%	23%
[53:75]	27%	16%	17%	27%
Tranches d'ancienneté		**Loi de rachat**		
[0:4]	9%	2%	<1%	12%
[5:9]	13%	6%	4%	23%
[9:14]	44%	61%	56%	161%

FIGURE 5.21 – Moyenne des écarts relatifs entre la borne supérieure du biais total et le flux réel

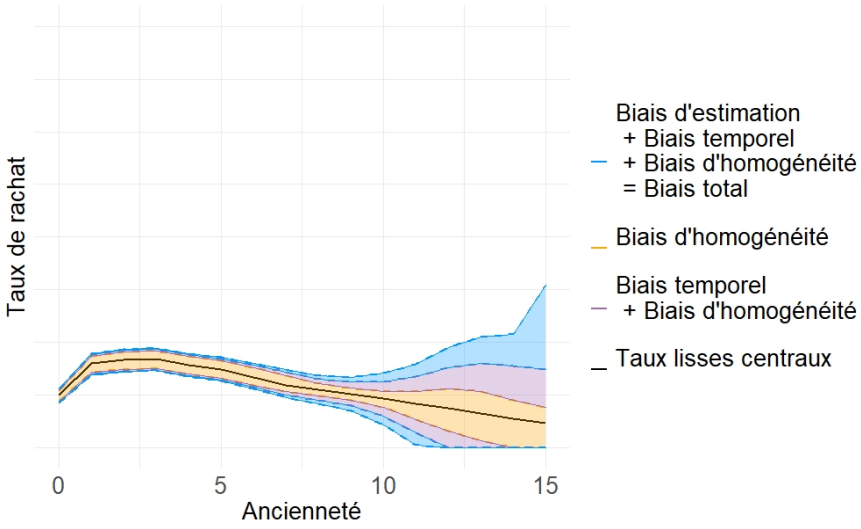
Une ultime erreur liée aux hypothèses sera étudiée dans la section suivante, il s'agit de l'erreur liée aux anomalies présentes dans les données utilisées pour la construction des lois.



(a) Loi de mortalité Femmes



(b) Loi de mortalité Hommes



(c) Loi de rachat

FIGURE 5.22 – Biais totaux

5.2.2 Détection d'anomalies dans les données de construction des hypothèses

La détection d'anomalies dans les données utilisées pour construire des lois de mortalité et de rachat peut être menée en suivant la même démarche que celle présentée dans la section 5.1. Toutefois, dans la mesure où les anomalies présentes dans les données des hypothèses ont un moindre impact sur la valeur du BE, le processus de détection d'anomalies sera réduit dans la présente section aux algorithmes les plus pertinents.

Pour rappel, les données étudiées correspondent aux données historiques du portefeuille sur la période d'observation.

En résumé, la démarche suivante est adoptée pour détecter les anomalies :

1. **Sélection des variables :** Seules les variables utiles à la construction des lois d'expérience sont sélectionnées (âge de l'assuré à la survenance du sinistre, âge d'entrée en portefeuille, âge de sortie du portefeuille, etc.).

2. **Réduction de dimension des données :** Les données sont réduites en dimension 2 à partir, d'une part, de l'algorithme ACP, et d'autre part, de l'algorithme UMAP. Ainsi, deux bases de données différentes sont obtenues.

3. **Lancement des algorithmes de *clustering* :** Les algorithmes K-moyennes et *Isolation Forest* sont chacun appliqués aux deux bases de données réduites. Les paramètres en entrée des algorithmes sont optimisés selon les méthodes présentées en section 5.1.

4. **Regroupement des anomalies :** Quatre groupes d'anomalies ont été obtenus à partir des différents algorithmes.

   — **Groupe 1 :** ACP et K-moyennes (2% de la base de données)
   — **Groupe 2 :** ACP et *Isolation Forest* (2% de la base de données)
   — **Groupe 3 :** UMAP et K-moyennes (4% de la base de données)
   — **Groupe 4 :** UMAP et *Isolation Forest* (3% de la base de données)

   L'idée est de prendre l'intersection de plusieurs de ces groupes pour définir les anomalies finalement retenues. Effectivement, ce procédé tend à confirmer le caractère anormal des données en question.

Notons d'une part qu'il est préférable de rassembler deux algorithmes de *clustering* utilisant des méthodes différentes de sélection. D'autre part, la méthode UMAP semble plus adaptée que la méthode ACP pour la sélection des anomalies dans ces données, dans la mesure où son paramétrage est optimisable dans notre cas d'étude, contrairement à la méthode ACP. Ainsi, deux cas possibles de regroupement d'anomalies se présentent :

— **Cas n°1 :** Groupe final = groupe 1 ∩ groupe 2 ∩ groupe 3 ∩ groupe 4 (1% de la base de données)

— **Cas n°2 :** Groupe final = groupe 3 ∩ groupe 4 (0,1% de la base de données)

Dans le cas n°1, les lois d'expérience reconstruites à partir des données sans ces anomalies sont davantage déformées que dans le cas n°2. Ainsi, le BE calculé à partir de ces nouvelles lois d'expérience diffère encore plus du BE central. Ces deux BE sont présentés dans la suite de l'étude en section 5.4. Au final, l'impact sur le BE étant plus important, les anomalies finales sélectionnées correspondent à tous les contrats détectés anormaux par les deux algorithmes de *clustering*, K-moyennes et *Isolation Forest*, que ce soit avec la réduction ACP ou la réduction UMAP.

5.3 Détection d'anomalies dans le modèle de projection

Les dernières erreurs à étudier sont celles provenant du modèle de projection du BE, le *backtesting* étant la meilleure approche pour détecter ce type d'erreurs.

Dans la présente étude, les *backtests* ont été réalisés à partir de l'année 2012. Les flux réels sont donc observables sur une large période de sept ans. Le portefeuille utilisé pour projeter les flux correspond à un portefeuille en stock au 31 décembre 2011. Il est projeté en *run-off*, comme l'exige Solvabilité 2. Cela signifie que les affaires nouvelles (i.e. les nouveaux contrats) arrivant en cours de projection ne sont pas pris en compte et que tous les engagements de l'assureur à l'égard des assurés présents au 31 décembre 2011 en portefeuille sur toute la période de leur contrat sont comptabilisés.

Les deux flux principaux du BE, à savoir les primes perçues et les prestations versées, feront chacun l'objet d'un *backtesting*. Notons que les flux seront présentés bruts de réassurance. De plus, ces flux seront agrégés à la maille "Garantie x Pas de temps". Plus la maille choisie est fine, plus les écarts entre les flux réels et les flux projetés sont précis.

Deux types d'intervalle de confiance autour des flux projetés seront utilisés afin d'évaluer la qualité du modèle. Le premier intervalle de confiance représentera la volatilité du portefeuille réel sur la période de projection. Cette volatilité sur l'effectif réel des contrats en portefeuille a un impact plus ou moins important sur le flux réel selon sa catégorie (primes ou charges de sinistre). Un flux peut être très sensible à une sortie de portefeuille, ce qui le rendrait très instable. Auquel cas son intervalle de volatilité serait très large. Par exemple, les prestations décès sont très

volatiles d'un mois à l'autre puisque comme le nombre de décès sur un mois donné est très faible, un seul décès en plus par rapport à un autre mois peut augmenter considérablement le montant des prestations. Cet intervalle est calculé à chaque pas de temps indépendamment des résultats obtenus sur les pas de temps précédents. Le portefeuille étant projeté en *run-off*, l'effectif réel du portefeuille diminue au fur et à mesure de la projection, ainsi l'intervalle de volatilité s'agrandit au fur et à mesure de la projection.

Le deuxième intervalle de confiance sera construit à partir des biais évalués précédemment en section 5.2.1. Plus précisément, les bornes des intervalles de confiance des hypothèses seront directement utilisées comme hypothèses dans le modèle de projection. Ainsi, ce deuxième intervalle permettra de prendre en compte les potentielles erreurs d'estimation des hypothèses dans le modèle de projection et l'effet cumulatif dans le temps de l'utilisation d'hypothèses erronées.

Par ailleurs, bien que le *backtesting* soit avant tout utilisé pour détecter les erreurs dans le modèle, il a également permis de repérer certaines erreurs dans les données réelles. Plus précisément, la somme des prestations versées paraissait démesurée concernant plusieurs mois de la période étudiée. En regardant de plus près chacune des prestations versées pour chacun des mois concernés avant agrégation, nous avons remarqué que la date d'effet du début de l'assurance était parfois mal renseignée. Ainsi, certains sinistres étaient comptabilisés alors que l'assuré n'était pas présent en portefeuille en décembre 2011. Ces erreurs ont fait l'objet d'une correction et ne sont pas présentes dans les figures des sections suivantes.

5.3.1 *Backtesting* en tenant compte de la volatilité réelle du portefeuille

Un flux réel présente parfois des fluctuations très importantes, en particulier les charges de sinistres relatives aux décès. Par exemple, la charge de sinistre peut, pour un mois donné, exploser à cause du décès accidentel d'un assuré ayant un capital restant dû très élevé au regard des autres prêts en portefeuille.

Cependant, le flux réel devrait toujours osciller autour du flux projeté, i.e. il devrait en moyenne correspondre au flux projeté.

Afin de déterminer l'écart acceptable entre le flux projeté et le flux réel, un intervalle de confiance, dit "de volatilité", peut être construit autour du flux projeté. Il correspond à volatilité certaine de l'effectif réel du portefeuille tout au long de la période de projection. Plus l'effectif du portefeuille est important, plus la volatilité du flux réel est faible, et ainsi l'intervalle de volatilité est plus étroit.

L'intervalle de volatilité s'établit alors de la façon suivante :

1. **Calcul de l'effectif du portefeuille réel à chaque mois de la projection**

2. **Calcul du flux de chaque assuré à chaque mois de la projection**

   D'une part, les primes d'assurance sont calculées pour chaque assuré. Pour rappel, elles correspondent à la multiplication du taux de prime, du capital initial et de la quotitée assurée. Pour les prêts personnels, ces primes restent inchangées tout au long de la projection.

   D'autre part, les capitaux restant dus sont calculés pour chaque assuré à chaque mois de projection en réalisant un simple amortissement du prêt à partir du capital initial, du taux d'intérêt et de l'éventuel différé.

3. **Calcul de l'intervalle de confiance des taux à chaque mois de projection**

   Des taux supérieurs et inférieurs de mortalité et de rachat sont calculés à chaque pas de temps en fonction de l'effectif du portefeuille. Ils correspondent aux pires cas possibles en terme de décès ou de rachat à un mois donné. Pour ce faire, l'intervalle de confiance de la loi binomiale est une nouvelle fois utilisé.

   À un mois t, l'intervalle de confiance à 95% des taux de rachat s'écrit de la façon suivante :

$$IC_{Binom,a,t} =$$

$$\left[\hat{q}_{inf,a,t} = max(0, \hat{q_a} - Q_{0.95}\sqrt{\frac{\hat{q_a}(1-\hat{q_a})}{N}}); hatq_{sup,a,t} = min(1, \hat{q_a} + Q_{0.95}\sqrt{\frac{\hat{q_a}(1-\hat{q_a})}{N}})\right]$$

   avec

   - $\hat{q_a}$ les taux de rachat lisses centraux en fonction de l'ancienneté a de l'assuré ;
   - N l'effectif du portefeuille réel au mois t ;
   - $Q_{0.95}$ le quantile à 95% d'une loi normale centrée réduite.

   À un mois t, l'intervalle de confiance à 95% des taux de mortalité s'écrit de la façon suivante :

$$IC_{Binom,x,t} =$$

$$\left[\hat{q}_{inf,x,s,t} = max(0, \hat{q_x} - Q_{0.95}\sqrt{\frac{\hat{q_x}(1-\hat{q_x})}{N}}); \hat{q}_{sup,x,s,t} = min(1, \hat{q_x} + Q_{0.95}\sqrt{\frac{\hat{q_x}(1-\hat{q_x})}{N}})\right]$$

   avec

   - $\hat{q_x}$ les taux de mortalité lisses centraux en fonction de l'âge x de l'assuré ;
   - N l'effectif du portefeuille réel au mois t ;
   - $Q_{0.95}$ le quantile à 95% d'une loi normale centrée réduite.

4. **Calcul de l'intervalle de confiance des flux à chaque mois de la projection**

   Enfin, les flux relevant des pires scénarios possibles de mortalité ou de rachat sont estimés à chaque pas de projection.

   Concernant les primes, le montant le plus élevé est obtenu à partir des taux de rachat inférieurs. À contrario, le montant le plus faible s'évalue à partir des taux de rachat supérieurs. Notons que les décès ont peu d'impact sur la valeur totale des primes, du fait que la proportion de décès est beaucoup plus faible que la proportion de rachat.

   Finalement, l'intervalle de confiance à 95% des primes s'écrit ainsi :

$$IC_{P,t} = [F_P(t, \hat{q}_{sup,a,t}, PRM); F_P(t, \hat{q}_{inf,a,t}, PRM)]$$

   avec

   - F_{PF} la fonction retournant le flux de primes ;
   - $\hat{q}_{sup,a,t}$ les taux de rachat supérieurs en fonction de l'ancienneté a de l'assuré et du temps de projection t ;
   - $\hat{q}_{inf,a,t}$ les taux de rachat inférieurs en fonction de l'ancienneté a de l'assuré et du temps de projection t ;
   - PRM les primes d'assurance des assurés.

   Concernant les prestations, le montant le plus élevé à un mois t est évalué à partir des taux supérieurs de mortalité. À contrario, le montant le plus faible à un mois t est évalué à partir des taux inférieurs de mortalité.

   Ainsi, l'intervalle de confiance à 95% des prestations futures est le suivant :

$$IC_{PF,t} = [F_{PF}(t, \hat{q}_{inf,x,s,t}, CRD_t); F_{PF}(t, \hat{q}_{sup,x,s,t}, CRD_t)]$$

   avec

   - F_{PF} la fonction retournant le flux de prestations futures ;
   - $\hat{q}_{sup,x,s,t}$ les taux de mortalité supérieurs en fonction de l'âge x de l'assuré, du sexe s de l'assuré et du temps de projection t ;
   - $\hat{q}_{inf,x,s,t}$ les taux de mortalité inférieurs en fonction de l'âge x de l'assuré, du sexe s de l'assuré et du temps de projection t ;
   - CRD_t les capitaux restant dus des assurés au temps de projection t.

Des *backtesting* ont été réalisés sur les primes et les prestations avec ces intervalles de volatilité (figure 5.23). Deux informations ressortent de ces *backtests* à partir des deux indicateurs présentés ci-après.

Indicateur n°1 :

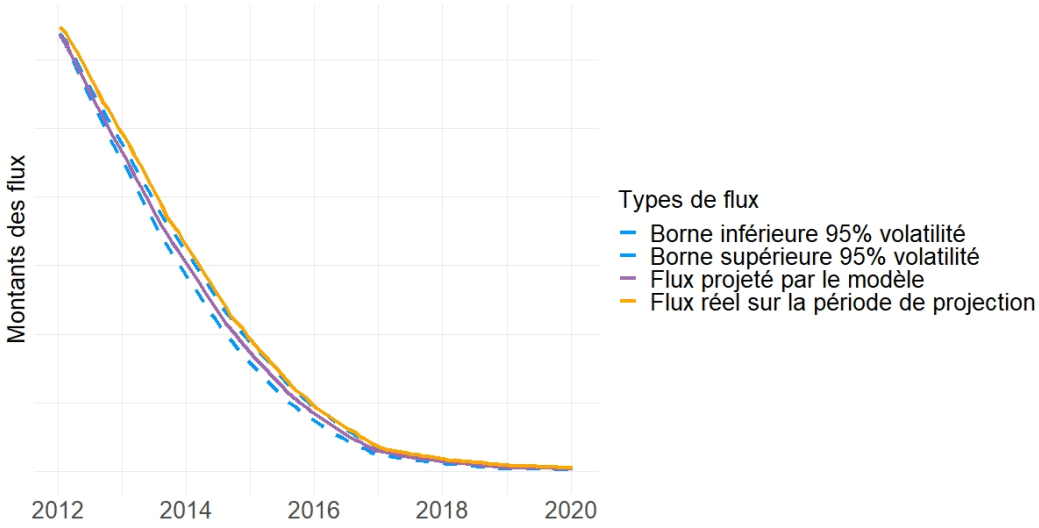
Le niveau de volatilité du portefeuille réel peut être mesuré à partir des écarts relatifs entre la borne supérieure et le flux réel. Tous ces écarts relatifs sont résumés dans le tableau 5.24.

- **Primes futures**

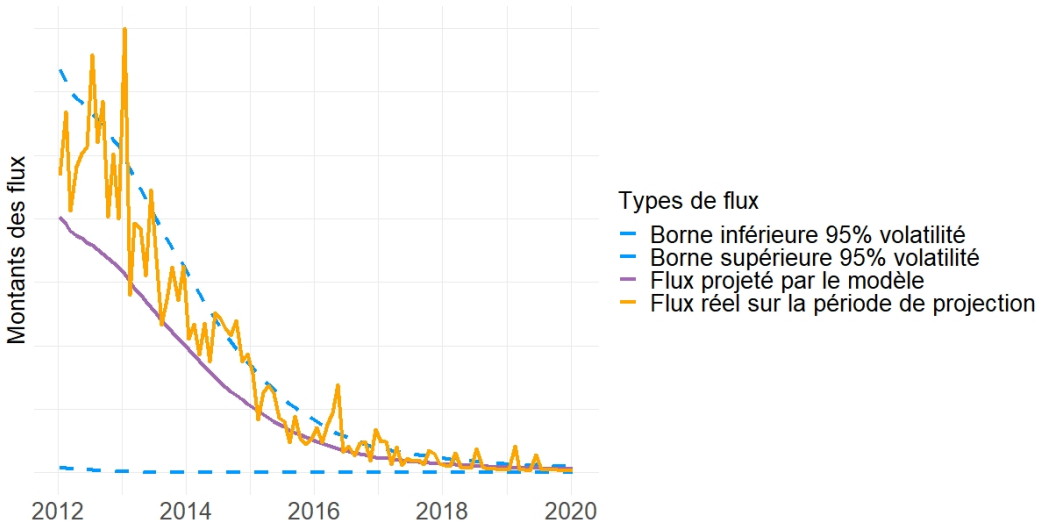
  L'intervalle de volatilité est assez étroit puisque l'écart relatif entre la borne supérieure de l'intervalle et les primes projetées reste inférieur à 25% tout au long de la projection. Cela est plutôt logique puisque les primes présentent peu d'aléa. Ainsi le flux réel est plutôt lisse.

- **Prestations futures**

  AU contraire, les prestations présentent beaucoup d'aléas et sont donc très volatiles. Cela se remarque par l'alternance de pics dans le flux réel qui est représenté en jaune sur la figure (b). Ainsi, l'intervalle de volatilité des prestations est plus large. En effet, les écarts relatifs entre la borne supérieure et le flux projeté se situent entre 58% et 65% tout au long de la projection.



(a) Primes futures



(b) Prestations futures

FIGURE 5.23 – *Backtesting* avec l'intervalle de volatilité réelle du portefeuille

	Intervalle de volatilité	Intervalle construit à partir des biais	Superposition des deux intervalles
Année projetée	**Primes**		
2012	1%	3%	4%
2013	4%	9%	13%
2014	7%	17%	24%
2015	11%	24%	35%
2016	14%	32%	47%
2017	19%	40%	59%
2018	21%	44%	65%
Année projetée	**Prestations futures**		
2012	58%	68%	126%
2013	60%	57%	117%
2014	61%	49%	110%
2015	62%	43%	106%
2016	64%	41%	104%
2017	65%	40%	104%
2018	65%	39%	104%

FIGURE 5.24 – Moyenne des écarts relatifs entre la borne supérieure et les flux projetés par année de projection

Indicateur n°2 :

Étant donné que ces intervalles de confiance à 95% ont été construits sur un échantillon du portefeuille, les tests de *backtesting* tels que détaillés en section 4.2 peuvent être appliqués. Ils permettent d'évaluer la qualité de l'estimation du flux en tenant compte de l'erreur de volatilité du portefeuille. Les valeurs des statistiques sont résumées dans le tableau 5.25. Si elles sont supérieures aux valeurs critiques, alors elles sont surlignées en rouge et l'hypothèse nulle est rejetée.

	Primes	Prestations futures	Valeur critique
Test de couverture non-conditionnelle	371,5	39,6	3,8
Test d'indépendance Christoffersen (1998)	5	3,9	6
Test d'indépendance Engel & Manganelli (2004)	502,6	15,3	18,3

FIGURE 5.25 – Tests de *backtesting*

Rappelons qu'un modèle est validé s'il remplit l'hypothèse de couverture non-conditionnelle et l'hypothèse d'indépendance des violations.

- **Primes futures**

  L'hypothèse de couverture non-conditionnelle n'est pas vérifiée. Cela est cohérent puisque 93% du flux réel se situe au-delà de la borne supérieure des intervalles.

  L'hypothèse d'indépendance est validée par le test de Christoffersen mais rejetée par le test d'Engel et Manganelli (avec $K = 5$).

  Cette dépendance entre les violations est illustrée par la figure 5.26 qui représente les écarts absolus entre la borne supérieure et le flux réel pour chaque type de flux. En effet, toutes les violations se succèdent au début de la projection. Notons que dans la mesure où la ligne rouge correspond à un écart nul, les points situés au-dessus de celle-ci correspondent à un écart positif.

- **Prestations futures**

  L'hypothèse de couverture non-conditionnelle n'est pas vérifiée. Cela est cohérent puisque 22% du flux réel se situe au-delà de la borne supérieure des intervalles. Toutefois, les violations de l'intervalle se trouvent essentiellement en fin de projection et correspondent à des montants de flux très faibles. Ainsi, la criticité du risque est considérée faible.

  L'hypothèse d'indépendance est validée par le test de Christoffersen et le test d'Engel et Manganelli (avec $K = 5$). De ce fait, les violations de l'intervalle des prestations peuvent être considérées comme indépendantes entre elles. La figure 5.26 conforte ces résultats. En effet, aucune dépendance n'est visible entre les points situés au-dessus de la ligne rouge.

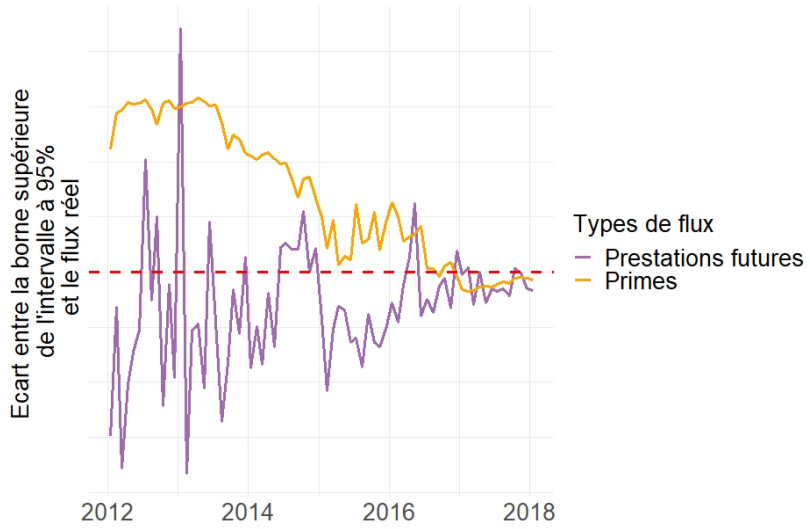


FIGURE 5.26 – Écarts relatifs entre les bornes supérieures et les flux réels

Pour résumer, les prestations futures paraissent bien estimées par le modèle. Elles semblent seulement être sous-estimées en fin de projection. Néanmoins, l'impact de cette sous-estimation sur la valeur du BE est non matérielle du fait des faibles montants concernés. Par ailleurs, bien que le flux réel des primes se situe dans

l'ensemble à l'extérieur de l'intervalle de volatilité, cet intervalle est relativement faible par rapport à la valeur des flux. Ainsi, la qualité du modèle ne peut pas être remise en cause. Nous retenons que la volatilité actuelle du portefeuille a un faible impact sur le flux de primes.

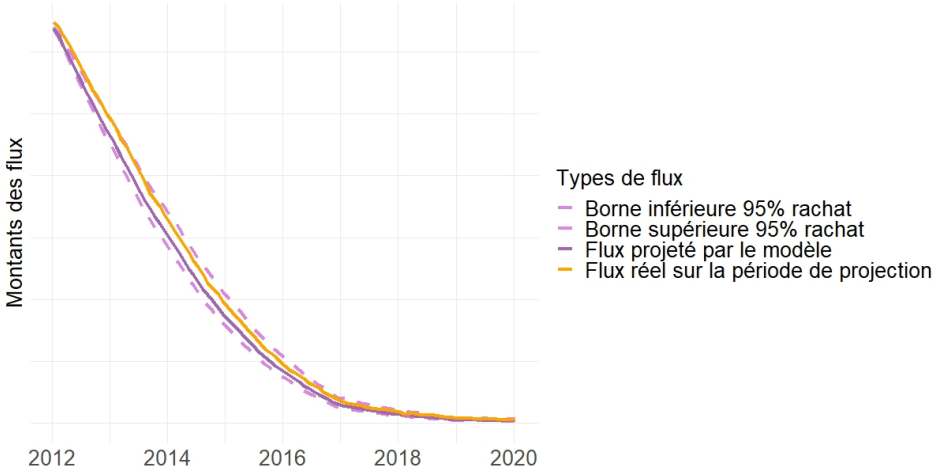
Afin d'étendre notre analyse sur la qualité du modèle, la pertinence de l'application des hypothèses par le modèle sera étudiée dans la section suivante.

5.3.2 *Backtesting* en tenant compte des biais liés aux hypothèses

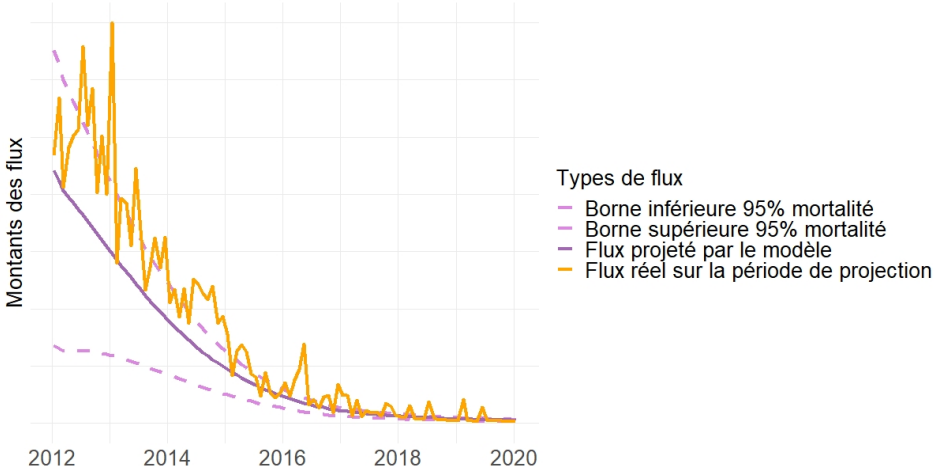
Rappelons que nous avons obtenu un ensemble de bornes supérieures et inférieures aux lois de mortalité et de rachat en section 5.2.1. Ces bornes correspondent aux pires situations possibles en terme de mortalité et de rachat vis-à-vis du non respect des contraintes de construction des hypothèses que nous avons identifiées. L'idée est donc d'incorporer ces bornes dans le modèle de projection afin d'obtenir un intervalle de confiance des flux projetés. De ce fait, le modèle de projection cumulera ces pires situations dans le temps i.e tout au long de la période de projection.

Si les flux réels se situent à l'intérieur des intervalles de confiance, alors l'application des lois d'expérience dans le modèle est satisfaisante. En revanche, si les flux réels sont en partie situés à l'extérieur de l'intervalle, alors les points concernés doivent faire l'objet d'une analyse plus approfondie, auquel cas nous proposons deux contrôles différents. Avant de se pencher sur les indicateurs de ces contrôles, présentons les trois *backtests* de la figure 5.27.

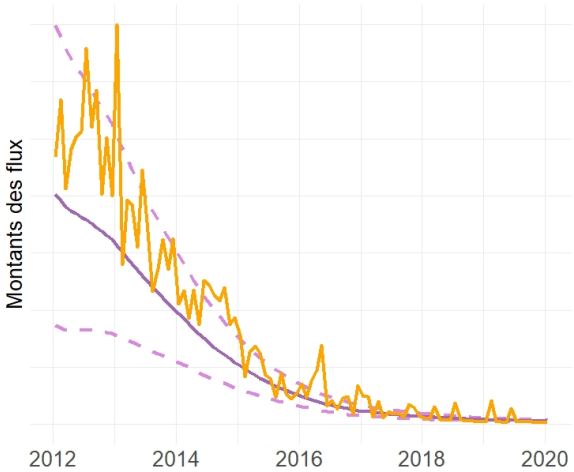
- Le *backtest* (a) présente les primes projetées et réellement perçues. L'intervalle de confiance a été obtenu en injectant dans le modèle les bornes de l'intervalle de confiance de la loi de rachat.
- Le *backtest* (b) présente les charges de sinistres projetées et réelles. L'intervalle de confiance a été réalisé en injectant dans le modèle les bornes des intervalles de confiance des anciennes lois de mortalité. Ces dernières ont été construites sur un échantillon historique avec des assurés de toutes anciennetés.
- Le *backtest* (c) présente les charges de sinistres projetées et réelles. L'intervalle de confiance a été réalisé en injectant dans le modèle les bornes des intervalles de confiance des nouvelles lois de mortalité. Ces dernières ont été construites en tenant compte de la sélection médicale à la souscription de l'assurance.



(a) Primes futures



(b) Prestations futures : lois de mortalité construites sur toutes les anciennetés



(c) Prestations futures : lois de mortalité construites sur les anciennetés supérieures ou égales à 2 ans

FIGURE 5.27 – *Backtesting* avec les intervalles provenant des biais liés aux hypothèses

La **sélection médicale** est un phénomène qui suscite beaucoup d'attention chez les assureurs. À la souscription d'un contrat d'assurance, l'assureur est en droit de demander à son client de remplir un questionnaire de santé. En conséquence, l'assureur peut estimer le risque porté vis-à-vis de son client et, de ce fait, accepter ou refuser de l'assurer. L'assureur peut refuser les clients qui ont de graves problèmes de santé avec un risque élevé de ne pas rembourser le prêt. Ainsi, la mortalité des assurés est plus faible les deux premières années d'ancienneté essentiellement. Ce phénomène doit être pris en compte dans la modélisation du risque décès. Pour ce faire, tout d'abord, les lois de mortalité peuvent être estimées à partir d'un échantillon comportant uniquement des assurés d'une ancienneté supérieure ou égale à deux ans (lois utilisées en figure (c)). Ensuite, un taux d'abattement peut être appliqué aux assurés ayant une ancienneté inférieure à deux ans.

Indicateur n°1 :

Les intervalles tracés en figure 5.27 délimitent le domaine auquel pourrait appartenir en réalité le flux projeté. En d'autres termes, ce domaine correspond à l'amplitude d'erreur du flux projeté (i.e. espéré) vis-à-vis de l'utilisation des hypothèses dans le modèle. Ainsi, le flux projeté pourrait potentiellement se situer au niveau de la borne supérieure et dans ce cas, le flux projeté serait sous-estimé. Or, ce flux projeté a été estimé en moyenne puisqu'il s'agit d'une espérance. Aussi, dans le cas précédent, il serait normal que le flux réel se situe à 50% en probabilité au-dessus de la borne et à 50% en dessous. De ce fait, nous retenons les marqueurs suivants :

- Si l'ensemble du flux réel se situe à l'intérieur de l'intervalle, alors la qualité du modèle vis-à-vis de l'application des hypothèses est jugée satisfaisante ;
- Si moins de 50% du flux réel se situe à l'extérieur de l'intervalle, alors la qualité du modèle vis-à-vis de l'application des hypothèses est jugée tout juste suffisante ;
- Si plus de 50% du flux réel se situe à l'extérieur de l'intervalle, alors la qualité du modèle vis-à-vis de l'application des hypothèses est jugée insatisfaisante et une étude plus poussée sur l'utilisation des hypothèses dans le modèle doit être menée.

Notons qu'une sous-estimation du flux projeté est beaucoup plus critique qu'une sur-estimation. Donc, s'il s'agit d'une sous-estimation du flux réel dans plus de 50% des cas, l'avis correctionnel sur la qualité du modèle est d'autant plus important.

À partir des proportions du flux réel situé au-delà des bornes supérieures répertoriées dans le tableau 5.28, les avis suivants sont émis sur les différents flux :

- **Primes futures**

  Moins de 50% du flux réel se situe au-delà de la borne supérieure. Ainsi, la qualité du modèle vis-à-vis de l'application de la loi de rachat est jugée tout juste suffisante.

  Notons que sur les premières années de projection, le flux réel se situe au-dessus de la borne supérieure. Cela s'explique par le fait que l'espérance de vie a augmenté depuis 2012 d'un point de vue général, ainsi la mortalité pour

un même portefeuille a diminué depuis 2012. Or, les lois de mortalité ont été construites sur les données de 2015 à 2020. Donc les lois de mortalité sous-estiment légèrement le risque de mortalité sur les années antérieures à 2015.

- **Prestations futures**

  Moins de 50% du flux réel du *backtest* (c) se situe au-delà de la borne supérieure. Ainsi, la qualité du modèle vis-à-vis de l'application des nouvelles lois de mortalité est jugée tout juste suffisante.

  Par ailleurs, le *backtesting* est un excellent outil pour évaluer les conséquences d'un changement de loi. Prenons l'exemple des backtests (b) et (c). Ils montrent que les nouvelles lois de mortalité construites sur les anciennetés supérieures ou égales à deux ans sont plus adaptées au portefeuille et reflètent mieux la réalité. En effet, 56% du flux réel se situe au-dessus de la borne supérieure de l'intervalle construit avec les anciennes lois, contre 31% pour l'intervalle construit avec les lois tenant compte de la sélection médicale.

	Intervalle de volatilité	Intervalle construit à partir des biais	Superposition des deux intervalles
Primes	93%	16%	8%
Prestations futures	22%	31%	3%

FIGURE 5.28 – Proportion du flux réel situé au-delà de la borne supérieure

Indicateur n°2 :

Les écarts relatifs entre les bornes de l'intervalle et le flux projeté sont de bons indicateurs de l'incertitude liée aux hypothèses sur le flux projeté. Ils sont résumés dans le tableau 5.24 pour les *backtests* (a) et (c). Les seuils des biais présentés en section 3.2 peuvent à nouveau être utilisés.

- **Primes futures**

  Les écarts relatifs avec la borne supérieure sont inférieurs à 25% les cinq premières années de projection. Ainsi, l'estimation des primes futures par la loi de rachat est jugée satisfaisante.

- **Prestations futures**

  Les écarts relatifs avec la borne supérieure se situent entre 39% et 69% les sept premières années de projection. Ainsi, l'estimation des prestations futures par les lois de mortalité tenant compte de la sélection médicale est jugée tout juste suffisante.

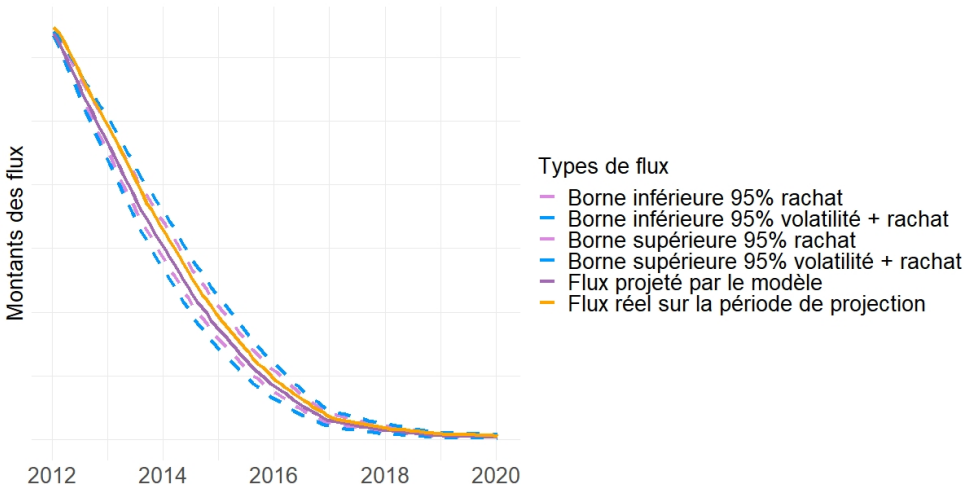
Pour résumer, l'intervalle de confiance basé sur les biais des hypothèses permet de valider la bonne modélisation des hypothèses. De plus, cet intervalle peut être utilisé comme un outil de pilotage par la société d'assurance. Il constitue une aide à la prise de décision dans le choix des hypothèses. Tout changement de loi pouvant engendrer un impact très important sur les résultats et la solvabilité de la société, il faut pouvoir identifier et quantifier cet impact avant la mise en place de la modification.

5.3.3 Avis final sur la qualité du modèle

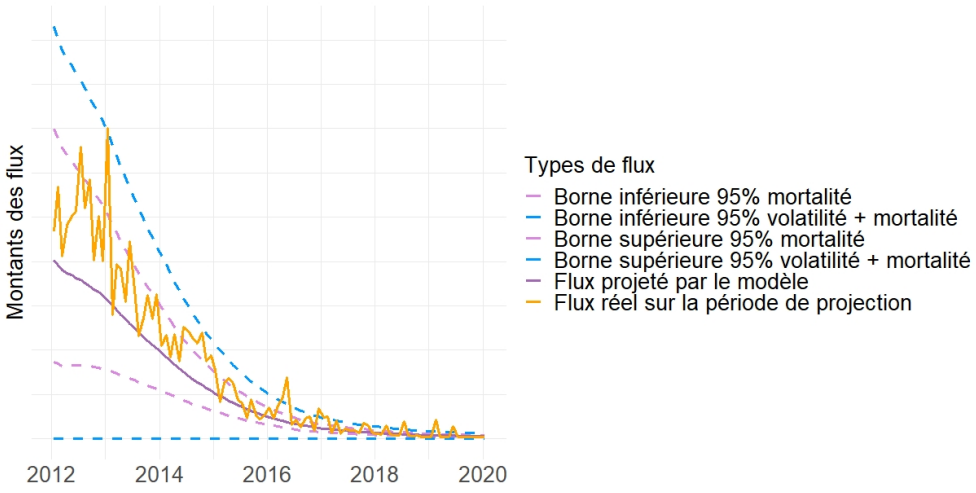
Afin d'avoir une idée globale de l'incertitude du flux projeté, nous pouvons associer l'intervalle de volatilité et l'intervalle obtenu à partir des biais. Ces deux intervalles étant indépendants, ils peuvent tout simplement être additionnés.

Les résultats sont illustrés en figure 5.29, les bornes finales étant représentées en pointillés bleus.

Pour conclure, au vu de la très faible proportion de flux réel au dessus de la borne supérieure (figure 5.28) et des écarts relatifs satisfaisants sur les premières années (figure 5.24), la projection des **primes futures** par le modèle est jugée satisfaisante. Concernant les prestations futures, les écarts relatifs sont plutôt élevés mais seulement 3% du flux réel se situe à l'extérieur de l'intervalle de confiance. Ainsi, la projection des **prestations futures** par le modèle est jugée tout juste suffisante et nécessite des analyses complémentaires sur les premières années de projection afin de repérer les prêts majoritairement touchés par cette légère sous-estimation du flux projeté.



(a) Primes



(b) Prestations futures

FIGURE 5.29 – *Backtesting* final

5.4 Impact des différentes erreurs sur le *Best Estimate*

Nous commençons à présent la phase finale de nos expérimentations, à savoir l'analyse de l'impact des erreurs sur la valeur du BE. Nous nous concentrerons sur le BE de primes des prêts personnels. Par la suite, il pourra être nommé "BE central", faisant référence au "scénario central".

5.4.1 Impact des anomalies présentes dans les données en entrée du modèle

Commençons par estimer l'impact des anomalies présentes dans les données en entrée du modèle qui ont été détectées en section 5.1. Pour ce faire, nous allons calculer la différence entre deux "BE prêt personnel", l'un à partir de la base initiale des prêts personnels et l'autre à partir de la même base à laquelle nous avons soustrait les anomalies.

Les anomalies soustraites correspondent aux données anormales qui ont été détectées par les trois algorithmes de *clustering* complétés par le *PU learning* (0,3% de la base initiale).

En définitive, la suppression des anomalies de notre base initiale augmente de 3% le "BE prêt personnel".

Impact sur le BE de primes prêts personnels	+ 3%

TABLE 5.8 – Impact de la suppression des anomalies sur la valeur du BE

Cet impact permet de quantifier l'erreur maximale que pourrait entraîner les éventuelles anomalies sur la valeur du BE.

Passons maintenant à l'impact des erreurs liées aux hypothèses sur le BE.

5.4.2 Impact des anomalies présentes dans les données de construction des hypothèses

Tout d'abord, nous pouvons évaluer l'impact des anomalies présentes dans les données de construction des hypothèses sur le BE. Pour ce faire, les trois lois d'expérience sont reconstruites avec les données auxquelles les anomalies ont été soustraites. Cette reconstruction doit suivre exactement la même méthode que celle suivie pour la construction des lois centrales, c'est-à-dire entre autres utiliser le même estimateur et la même méthode de lissage.

Ensuite, il suffit d'exécuter le modèle de projection avec ces trois nouvelles lois d'expérience en tant qu'hypothèses afin d'obtenir un nouveau BE, noté $BE_{HypSansAnomalies}$. L'impact des anomalies sur la valeur du BE correspond à l'écart relatif :

$$\Delta_r[BE_{HypSansAnomalies}, BE_{Central}]$$

Dans notre cas, cet écart est égal à 0,4%, ce qui est relativement faible.

Impact sur le BE de primes prêts personnels	+ 0,4%

TABLE 5.9 – Impact des anomalies présentes dans les données de construction des hypothèses sur la valeur du BE

L'impact sur le BE des biais liés aux hypothèses est quantifié dans la section suivante.

5.4.3 Impact des biais liés aux hypothèses

Rappelons qu'un intervalle de confiance, représentant l'ensemble des biais, a été construit pour chacune des trois lois d'expérience dans la section 5.2.1.

Tout d'abord, il est nécessaire d'exécuter le modèle de projection en renseignant en hypothèses les bornes supérieures et inférieures des intervalles de confiance pour les lois de mortalité et/ou de rachat. Ainsi, huit BE différents peuvent être obtenus, lesquels étant résumés dans le tableau 5.10.

Identification du BE	Loi de mortalité utilisée	Loi de rachat utilisée
$BE_{MortaSup}$	Supérieure	Centrale
$BE_{MortaInf}$	Inférieure	Centrale
$BE_{RachatSup}$	Centrale	Supérieure
$BE_{RachatInf}$	Centrale	Inférieure
$BE_{MortaSup,RachatSup}$	Supérieure	Supérieure
$BE_{MortaSup,RachatInf}$	Supérieure	Inférieure
$BE_{MortaInf,RachatSup}$	Inférieure	Supérieure
$BE_{MortaInf,RachatInf}$	Inférieure	Inférieure

TABLE 5.10 – Synthèse des BE obtenus à partir des biais liés aux lois de mortalité et de rachat

Les écarts relatifs $\Delta_r[BE_\circ, BE_{Central}]$ et $\Delta_r[BE_{\circ,\circ}, BE_{Central}]$, entre le BE central et ces différents BE sont illustrés en gras en figure 5.30. Le scénario central est représenté par le carré au milieu de la figure, dont la valeur est évidemment nulle. Les valeurs grises correspondent à la moyenne des carrés situés en coins supérieurs et inférieurs de chacune d'elles. Plus précisément, elles sont calculées de la façon suivante, avec $X_{i,j}$, $i, j \in [1, 9]^2$ la matrice représentée par le graphique :

$$x_{i,j} = \frac{x_{max(1,i-1),max(1,j-1)} + x_{max(1,i-1),min(9,j+1)} + x_{min(9,i+1),min(9,j+1)} + x_{min(9,i+1),max(1,j-1)}}{4}$$

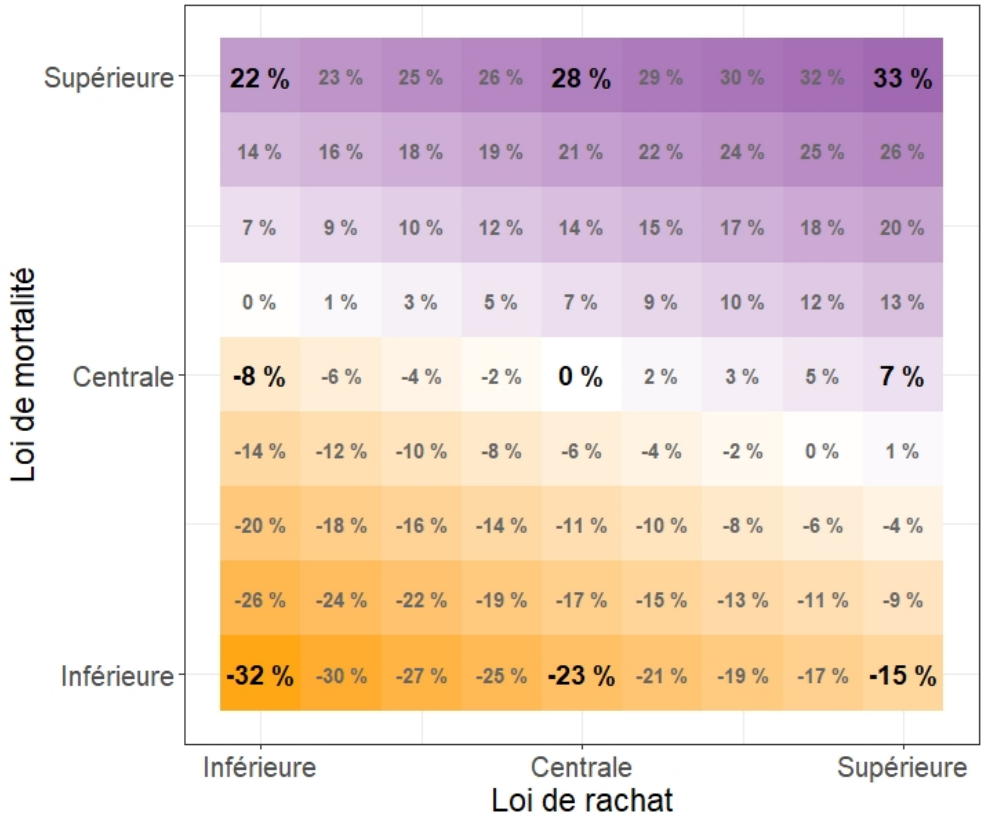


FIGURE 5.30 – Analyse de l'impact des anomalies dans les données en entrée des hypothèses sur le *Best Estimate*

Ce graphique montre que le pire scénario possible serait la combinaison d'une forte augmentation de la mortalité et d'une forte augmentation des rachats (coin supérieur droit : $x_{1,9} = 33\%$). En d'autres termes, une sous-estimation de la loi de mortalité combinée à une sous-estimation de la loi de rachat correspondrait au pire cas possible. Par ailleurs, l'augmentation de la mortalité a un impact beaucoup plus négatif que l'augmentation des rachats ($x_{1,5} > x_{5,9}$).

À partir de ce graphique, il est possible de déterminer la matrice de corrélation suivante entre les risques de mortalité et de rachat :

$$\begin{pmatrix} \frac{BE_{MortaSup,RachatInf} - BE_{Central}}{BE_{MortaSup} + BE_{RachatInf} - 2BE_{Central}} & \frac{BE_{MortaSup,RachatSup} - BE_{Central}}{BE_{MortaSup} + BE_{RachatSup} - 2BE_{Central}} \\ \frac{BE_{MortaInf,RachatInf} - BE_{Central}}{BE_{MortaInf} + BE_{RachatInf} - 2BE_{Central}} & \frac{BE_{MortaInf,RachatSup} - BE_{Central}}{BE_{MortaInf} + BE_{RachatSup} - 2BE_{Central}} \end{pmatrix}$$

$$= \begin{pmatrix} 1,09 & 0,96 \\ 1,05 & 0,92 \end{pmatrix}$$

On en déduit que les risques d'augmentation de la mortalité et d'augmentation des rachats sont sous-additifs. C'est-à-dire que l'augmentation de la mortalité compense une partie de l'augmentation des rachats. Il en est de même pour les risques de baisse de la mortalité et d'augmentation des rachats. À contrario, les risques d'aug-

mentation de la mortalité et de baisse des rachats sont sur-additifs (i.e. le premier phénomène amplifie le second), de même que les risques de baisse de la mortalité et de baisse des rachats.

À l'avenir, seuls les $BE_{MortaSup}, BE_{MortaInf}, BE_{RachatSup}$ et $BE_{RachatInf}$ pourront être évalués à partir du modèle de projection. Les autres BE pourront être calculés à partir de la matrice de corrélation ci-dessus. Cependant, il sera nécessaire de reconstruire cette matrice dans les prochaines années afin de s'assurer qu'elle ne varie que très peu. De même, elle devra être reconstruite en cas de changement de loi de mortalité ou de rachat.

Cette matrice de corrélation est un outil très intéressant pour l'évaluation des risques, *risk assessment* en anglais, et en particulier pour évaluer l'impact d'une augmentation future de la mortalité et/ou des rachats dans une compagnie d'assurance.

Finalement, l'impact des biais sur la valeur du BE peut être évalué en prenant le maximum des écarts relatifs présents en figure 5.30. Pour rappel, cet écart maximal serait valable dans le pire des scénarios possibles en terme de décès et de rachat.

Impact sur le BE de primes prêts personnels	+ 33%

TABLE 5.11 – Impact des biais liés aux hypothèses sur la valeur du BE

5.4.4 Impact total sur le *Best Estimate*

Pour conclure la présente étude, quantifions l'impact total sur le BE des biais liés au hypothèses, des anomalies dans les données de construction des hypothèses et des anomalies dans les données en entrée du modèle.

Ces erreurs étant totalement indépendantes entre elles, l'erreur totale sur le BE correspond à la somme des trois impacts déterminés précédemment.

Biais liés aux hypothèses	+ 33%
Anomalies dans les données de construction des hypothèses	+ 0,4%
Anomalies dans les données en entrée du modèle	+ 3%
Erreur totale sur le BE de primes prêts personnels	+ 36,4%

TABLE 5.12 – Impact des différentes erreurs sur la valeur du BE

Au final, sans prendre en compte les anomalies présentes dans le modèle de projection, l'incertitude sur le BE est de 36,4%. Cela signifie que le BE pourrait être sous-estimé de 36,4% dans le pire des cas possibles. Comme illustrée en figure 5.31, cette incertitude provient essentiellement des biais liés aux hypothèses.

Afin de valider ou non l'estimation du BE, il est possible de comparer l'erreur totale à la marge pour risque agrégée à la maille "prêts personnels". En effet, cette marge

exigée par Solvabilité 2 permet notamment de compenser l'erreur d'incertitude du BE, comme expliqué en section 1.2.1. Finalement, l'erreur totale reste inférieure à la marge pour risque. Cependant, bien que la majorité des erreurs aient été quantifiées dans cette présente étude, elle s'en rapproche fortement.

Ainsi, l'estimation du BE est jugée tout juste suffisante et pour poursuivre la présente étude, il serait intéressant de quantifier les erreurs opérationnelles et d'évaluer leur impact sur la valeur du BE.

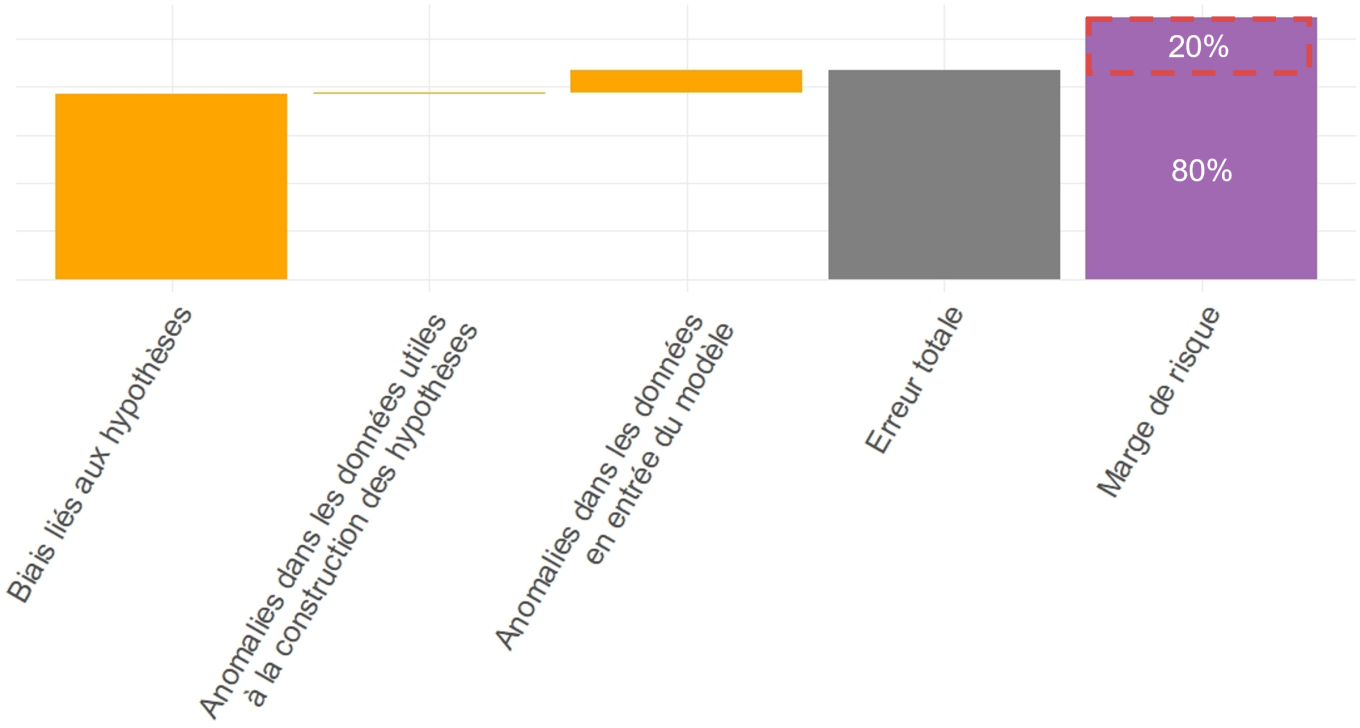


FIGURE 5.31 – Analyse de l'impact des erreurs sur le *Best Estimate*

Conclusion

Les différentes analyses conduites au cours de cette étude ont permis de mettre en évidence plusieurs erreurs dans les bases de données, les hypothèses et le modèle de projection utilisés pour le calcul du *Best Estimate* des prêts personnels.

En premier lieu, trois algorithmes d'apprentissage non-supervisé ont permis de détecter 0,3% de potentielles anomalies dans la base de données utilisée en entrée du modèle de projection. À noter qu'il a été choisi de considérer comme potentielles anomalies toutes les données qui ont été détectées anormales par l'ensemble des trois algorithmes. Un algorithme d'apprentissage semi-supervisé, le *PU learning*, a ensuite permis de s'assurer de l'efficacité des trois algorithmes. L'application de la loi de Benford nous a également amené à apprécier la cohérence des résultats des trois algorithmes.

En deuxième lieu, les lois de mortalité et de rachat ont fait l'objet d'une étude approfondie. Afin de quantifier l'incertitude liée à ces lois représentant des hypothèses, un intervalle de confiance regroupant le biais d'homogénéité du portefeuille, le biais temporel et le biais d'estimation a d'une part été estimé pour chacune des lois. Cet intervalle de confiance global dessine l'ensemble des scénarios possibles, les bornes de cet intervalle reflétant les pires scénarios. Notons que les biais dus aux évolutions temporelles et à l'hétérogénéité du portefeuille ont été évalués à partir d'un *bootstrap* sur des groupes de risque. Le biais d'estimation à quant à lui été évalué par l'intervalle de confiance d'une loi binomiale. D'autre part, les données utilisées pour la construction des lois ont été analysées : 1% d'entre elles ont été détectées anormales par des algorithmes d'apprentissage non-supervisé utilisés également dans le cas des données en entrée du modèle.

En troisième lieu, des *backtesting* ont été réalisés sur les primes perçues et les prestations versées afin d'évaluer le risque de modèle. Deux intervalles de confiance ont été utilisés pour encadrer les flux projetés : un intervalle représentant la volatilité réelle du portefeuille et un intervalle construit à partir des biais des hypothèses. Finalement, en comparant les flux réels à ces intervalles de confiance, le modèle semble estimer correctement les primes futures et assez bien les prestations futures.

En définitive, l'ensemble des anomalies détectées dans les données en entrée du

modèle et dans les données utilisées pour la construction des hypothèses et les biais liés aux hypothèses augmentent de 36% la valeur du *Best Estimate*. Cette marge d'erreur du *Best Estimate* prend en compte les pires scénarios possibles, notamment en termes de décès et de rachat. Elle reste toutefois inférieure à la marge pour risque exigée par Solvabilité 2 et allouée à l'activité des prêts personnels. Cette marge pour risque couvre donc bien les risques d'estimation potentiels sur le périmètre des prêts personnels.

Pour aller plus loin, cette première étude pourrait être complétée par des analyses sur les erreurs opérationnelles dans les hypothèses et le modèle, lesquelles erreurs augmenteraient l'incertitude du *Best Estimate*. Aussi, pour compléter les *backtests* de la présente étude, il serait intéressant de prendre en compte l'augmentation de l'espérance de vie des assurés au fil du temps. Pour ce faire, un coefficient pourrait être appliqué aux prestations futures projetées, celui-ci tiendrait compte de la baisse de la mortalité sur les années antérieures à la première année observée lors de la construction des lois de mortalité.

Bibliographie

[1] *Règlement Délégué (UE) 2015/35 de la Commission - du 10 octobre 2014 - complétant la directive 2009/138/CE du Parlement européen et du Conseil sur l'accès aux activités de l'assurance et de la réassurance et leur exercice (solvabilité 2)*, Journal officiel de l'Union Européenne, 2015

[2] *A tutorial on principal component analysis*, Shlens Jonathon, 2014

[3] *Umap : Uniform manifold approximation and projection for dimension reduction*, McInnes John, Leland Melville et Healy James, 2018

[4] *Isolation forest*, Fei Tony Liu, Kai Ming Ting, 2008

[5] *Ballstering : un algorithme de clustering dédié à de grands échantillons*, Université de Toulouse, Courjault-Rade Vincent, 2008

[6] *An evaluation of two-step techniques for positive-unlabeled learning in text classification*, *International Journal of Computer Applications Technology and Research, Volume 3*, Kaboutari Azam, Bagherzadeh Jamshid, Kheradmand Fatemeh, 2014

[6] *Flexible Procedure for Mixture Proportion Estimation in Positive-Unlabeled Learning*, *github.com*, Long James, 2020

[7] *Quantification de l'impact des erreurs opérationnelles sur les variables d'entrée d'un outil de calcul et de reporting Solvabilité 2 : approche par analyse de sensibilité*, Tondolo L., mémoire, Institut des actuaires, 2019

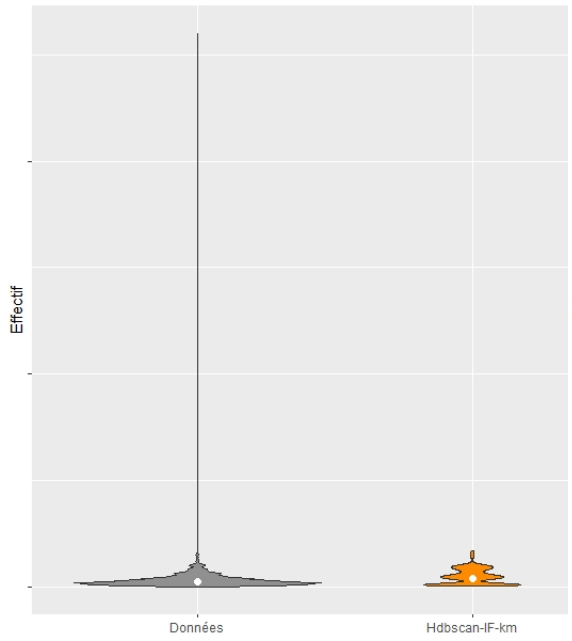
[8] *A review of backtesting and backtesting procedures*, Campbell S. D., *Journal of Risk*

Annexe A: Notations des indicateurs de contrôle

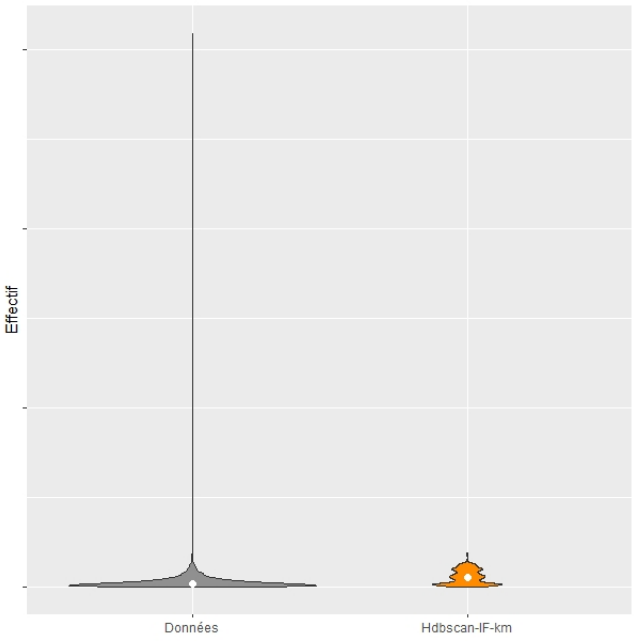
Les notations utilisées pour les indicateurs de contrôles des erreurs sont les suivantes :

- $\Delta_r[x, y] = \frac{x-y}{y}$: écart relatif entre deux valeurs x et y, exprimé en pourcentage ;
- $\Delta_a[x, y] = x - y$: écart absolu entre deux valeurs x et y ;
- $\tau[x] = \frac{Nb_x}{Nb_{total}}$: taux de l'anomalie x d'un portefeuille par rapport à un groupe, exprimé en pourcentage ;
- $Stat_X[k]$: une statistique (moyenne, écart type, variance, fréquence, coefficient de variation ...) sur la variable k de la matrice X.

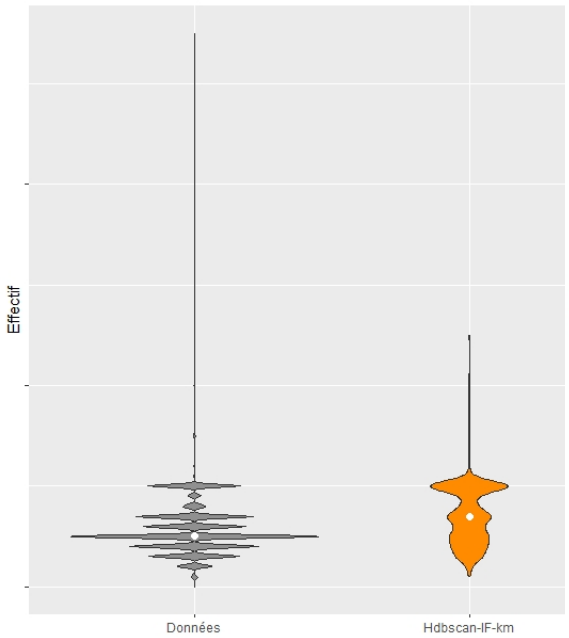
Annexe B: Analyses univariées sur les données et les anomalies



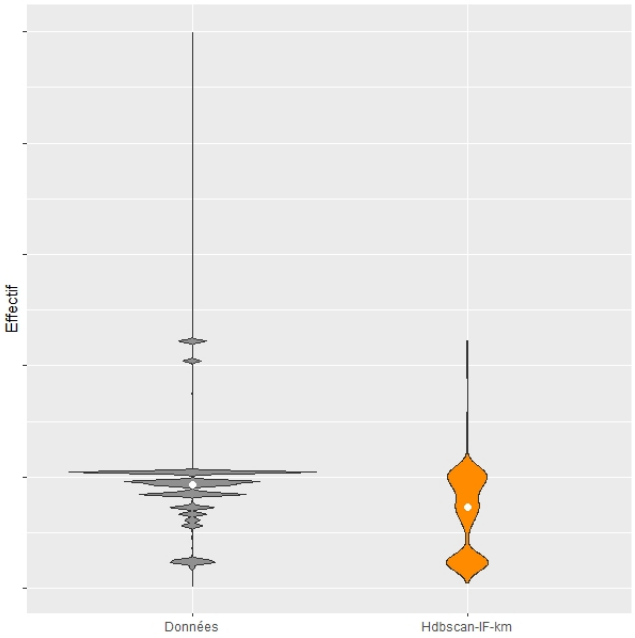
Capital initial



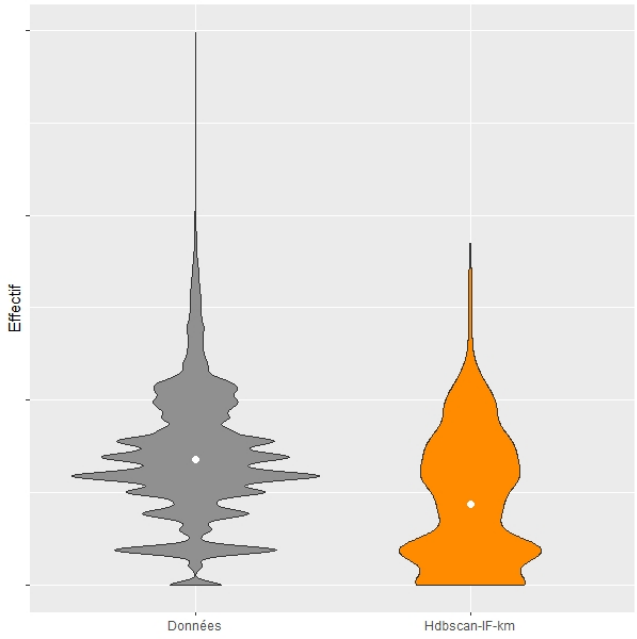
Capital restant dû



Durée du contrat



Taux de prime



Taux d'intérêt