Mémoire présenté le :

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

Par : LANGLOIS Anaïs

Titre : Mesures d'erreur de la meilleure estimation des provisions techniques

Confidentialité : ☒ NON ☐ (Durée : ☐ 1 an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires — *Signature*

Y. MILOE

A. HASSLER

L. MARTINI

Membres présents du jury de l'ISFA

S. LOISEL

E. MASIELLO

Entreprise :
Nom : NATIXIS Assurances

Signature :

Directeur de mémoire en entreprise :
Nom : Mme Laure Kerboul

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

Résumé

Mots-clés : *Best Estimate*, mesure d'erreurs, prévoyance individuelle, détection d'anomalies, modèle de durée, tables d'expérience, *backtesting*, modèle de projection.

Le calcul des capitaux requis pour garantir la solvabilité de l'entreprise se base sur la projection des flux futurs de chaque portefeuille. Cela nécessite un contrôle du niveau des provisions par le régulateur, mais surtout en premier lieu par la compagnie elle-même. En effet, afin d'améliorer la fiabilité du calcul des provisions techniques, la Fonction Actuarielle doit garantir le caractère approprié des méthodologies, des modèles sous-jacents et des hypothèses utilisées. Pour évaluer la précision de la méthode de provisionnement utilisée, il est nécessaire de prendre en compte toutes les sources d'erreurs possibles associées au calcul des provisions techniques.

Partant de ce constat, l'objet du présent mémoire a pour but d'intégrer les erreurs générées par l'ensemble des éléments entrant en jeu dans le calcul de la meilleure estimation des provisions techniques : données, modèle et hypothèses. Ce mémoire propose de quantifier quel biais associer au calcul des engagements d'un portefeuille de contrats d'assurance. Il se base sur diverses approches. On peut citer la détection des anomalies sur chacune des bases de données utilisées, la quantification des différentes sources d'incertitudes intervenant dans la construction de loi d'expérience (fluctuations statistiques, hétérogénéité du portefeuille et dérive temporelle) ainsi que la réalisation d'un *backtesting* permettant de comparer les provisions techniques aux données empiriques correspondantes. Ce mémoire sera appliqué au risque "décès toutes causes" d'un portefeuille de prévoyance individuelle de Natixis Assurances.

Abstract

Keywords : Best Estimate, error measures, individual cover, anomaly detection, duration model, experience tables, backtesting, projection model.

The calculation of the required capital in order to the solvency of the company is based on the projection of future flows of each portfolio. This requires a control of the reserves' level by the regulator, but above all by the company itself. Indeed, in order to improve the reliability of the calculation of technical reserves, the Actuarial Function must guarantee the relevance of the methodologies, the underlying models and the assumptions used. To assess the accuracy of the reserving method used, it is necessary to take into account all possible sources of error associated with the calculation of technical reserves.

With this in mind, the purpose of this paper entitled "Error Measures for the Best Estimate of Technical Reserves", is to integrate the errors generated by all the elements involved in the calculation of the best estimate of technical reserves : data, model and assumptions. This paper proposes to quantify which bias can be associate with the calculation of the liabilities of an insurance. It's based on various approaches such as the detection of anomalies on each of the databases used, the quantification of the various sources of uncertainty involved in the construction of the law of experience (statistical fluctuations, portfolio heterogeneity and temporal drift) and the realization of a back-testing allowing the comparison of the technical provisions with the corresponding empirical data. It will be applied to the "death from all causes" risk of an individual insurance portfolio of Natixis Assurance group.

Avant-propos

Pour des raisons de confidentialité, les résultats chiffrés ont été modifiés sans perte de généralité et une partie des échelles supprimée. L'ensemble des calculs de ce mémoire a été réalisé à l'aide du langage R.

Remerciement

Je remercie tout particulièrement Adrien Helary, sans qui ce mémoire n'aurait pas été possible, pour son aide et sa disponibilité lors de mes interrogations.

Mes remerciements s'adressent aussi à toute l'équipe Fonction Actuarielle pour leur accueil chaleureux lors de mon alternance et leur soutien. Je tiens à remercier Muriel Combarnous pour m'avoir transmis son savoir, Jim Lothammer pour ses conseils avisés et Laure Kerboul pour son aide, sa relecture attentive, et pour m'avoir accueilli au sein de l'équipe.

De plus, je remercie mon tuteur académique, M. Frédéric Planchet, responsable du suivi de mon étude, pour ses conseils et sa disponibilité tout au long de la réalisation de ce mémoire.

Enfin, je remercie du fond du cœur mes amis, Olivier et Johanna, qui ont égayé la réalisation de ce mémoire au cours de nos nombreuses nuits blanches et qui, par leurs critiques constantes mais constructives, lui ont permis d'aboutir. Aussi, je remercie mes parents pour leur relecture assidue.

Table des matières

Introduction **1**

I Présentation générale du contexte de l'étude **5**
I.1 Contexte métier . . . 5
I.1.1 Périmètre de la prévoyance individuelle . . . 5
I.1.2 Risques de la prévoyance individuelle . . . 5
I.2 Contexte réglementaire . . . 7
I.2.1 Objectif de la réglementation . . . 7
I.2.2 Principes de gouvernance et fonctions clefs . . . 8
I.3 Provisionnement sous Solvabilité II . . . 9
I.3.1 Généralités . . . 9
I.3.2 Caractéristiques des provisions techniques . . . 11
I.4 Valorisation et description du *Best Estimate* . . . 12
I.4.1 Calcul de la Valeur Actuelle Nette et du *Best Estimate* . . . 12
I.4.2 Description élémentaire du *Best Estimate* . . . 13
I.4.3 Mise en œuvre du calcul de la meilleure estimation . . . 15
I.4.4 Eléments nécessaires au modèle décès et arrêt de travail . . . 16
I.4.5 Le modèle de projection multi-états . . . 17
I.4.6 Rôle de la fonction actuarielle dans le cadre du calcul des provisions techniques . 18
I.5 Objectifs de l'étude . . . 18

II Mesure d'erreurs relatives aux données **19**
II.1 Contextualisation et présentation de la démarche . . . 19
II.1.1 Généralités sur la qualité des données . . . 20
II.1.2 Exigence de qualité des données sous le référentiel de calcul Solvabilité II . . . 21
II.2 Présentation des méthodes de détection d'anomalies . . . 23
II.2.1 Typologie et revue des méthodes de détection d'anomalie . . . 24
II.2.2 KNN : les K plus proches voisins . . . 29
II.2.3 LOF : *Local Outlier Factor* . . . 30
II.2.4 $DBSCAN$: *density-based spatial clustering of applications with noise* . . . 32
II.2.5 IF : Isolation Forest . . . 35
II.3 Application . . . 37
II.3.1 Caractéristiques des bases de données utilisées lors de la détection d'anomalies . 37
II.3.2 Utilisation d'un algorithme de réduction de dimensions . . . 38
II.3.3 Mise en œuvre de la réduction de dimensions . . . 41
II.3.4 Application des algorithmes de détection d'anomalies . . . 43
II.3.5 Résultats des méthodes de détection . . . 46
II.4 Conclusion partielle . . . 47

III Mesure d'erreurs relatives à la construction des hypothèses actuarielles **49**
III.1 Contextualisation et présentation de la démarche . . . 49
III.1.1 Contexte d'utilisation des hypothèses *Best estimate* . . . 50
III.1.2 Caractéristiques des hypothèses actuarielles . . . 51
III.1.3 Estimation non-paramétrique des taux bruts . . . 52

III.1.4 Métrique de comparaison des taux bruts . . . 56
III.2 Sources d'incertitude relatives à la construction d'hypothèses . . . 57
III.2.1 Formalisation de l'erreur d'estimation . . . 57
III.2.2 Risques associés à l'estimation non-paramétrique . . . 58
III.2.3 Incertitudes possibles dans l'estimation d'un taux de survie . . . 59
III.2.4 Sources d'incertitude principales faisant l'objet d'une étude . . . 60
III.3 Application . . . 60
III.3.1 Erreur induite par la mauvaise qualité des données . . . 60
III.3.2 Formalisation de la procédure de ré-échantillonnage . . . 62
III.3.3 Incertitude causée par les fluctuations d'échantillonnage . . . 63
III.3.4 Incertitude causée par la dérive temporelle des portefeuilles de construction . . . 66
III.4 Conclusion partielle . . . 69

IV Quantification de l'erreur sur les *Best Estimates* 71
IV.1 Formalisation et présentation de la démarche . . . 71
IV.1.1 Formalisation des écarts . . . 71
IV.1.2 Obtention des hypothèses alternatives à partir des hypothèses existantes . . . 72
IV.1.3 Utilisation des différentes études . . . 72
IV.2 Incertitude générée par la qualité des données de construction . . . 74
IV.2.1 Sensibilité du *Best Estimate* à la mauvaise qualité des données . . . 74
IV.2.2 Différence d'impact selon l'hypothèse considérée . . . 75
IV.2.3 Ecart d'évolution entre les deux réseaux . . . 76
IV.2.4 Confortation des résultats . . . 76
IV.3 Incertitude générée par l'utilisation des hypothèses . . . 78
IV.3.1 Rappel des éléments utilisés . . . 78
IV.3.2 Impact des différentes études par hypothèses . . . 79
IV.3.3 Impact croisé des différentes études et sources . . . 80
IV.4 Comparaison des sources d'incertitude . . . 81

V Mesure d'erreurs relatives au modèle de projection 83
V.1 Contextualisation et présentation de la démarche . . . 83
V.1.1 Définition et objectif du *backtesting* . . . 83
V.1.2 Contexte règlementaire . . . 84
V.1.3 Choix de l'historique utilisé . . . 84
V.2 Passage d'un référentiel comptable à un référentiel prudentiel . . . 86
V.2.1 Principes de valorisation du *Best Estimate* . . . 86
V.2.2 Description des flux réels utilisés pour le backtesting . . . 87
V.3 Application . . . 88
V.3.1 Principaux retraitements effectués . . . 88
V.3.2 Analyse des résultats . . . 92
V.3.3 Axes d'amélioration de la méthode de backtesting . . . 94
V.4 Conclusion partielle . . . 94

Conclusion 94

A Anomalies détectées sur les autres bases 97

Introduction

Selon l'article 48 de la directive Solvabilité II, la fonction actuarielle valide le montant des provisions techniques et est garante de la justesse et de la pertinence des estimations. Elle est garante du caractère approprié des méthodologies, des modèles sous-jacents et des hypothèses utilisées pour le calcul des provisions techniques. Aussi, elle contribue au système de gestion des risques dans une démarche d'amélioration continue du calcul des provisions techniques.

En prévoyance individuelle, les passifs étant non-réplicables, les provisions techniques prudentielles sont égales à la somme du *Best Estimate* et de la marge de risque. Le *Best Estimate* correspond à l'actualisation de tous les flux probables futurs (cotisations, prestations, frais, fiscalité...) actualisés avec la courbe des taux sans risque de l'EIOPA. Les flux futurs probables sont estimés librement par les organismes d'assurance dans la mesure où la réglementation est respectée.

Ce *Best Estimate* représente la valorisation des engagements de l'assureur vis-à-vis des assurés et son calcul « fait appel à des méthodes actuarielles et statistiques adéquates, applicables et pertinentes » [article 77 de la Directive Solvabilité II]. Plus précisément, son calcul requiert des données représentant le portefeuille d'assuré, un modèle de projection lorsque cela est possible ainsi que des hypothèses comptables et actuarielles (hypothèses elle-même construites à l'aide de données et de modèles).

Or, il est important de remarquer qu'un modèle est une représentation imparfaite de la réalité, qu'une hypothèse n'est jamais parfaitement construite ni parfaitement adaptée à l'utilisation que l'on en fait et qu'une donnée ne peut être totalement fiable. La meilleure estimation possible des provisions techniques comporte donc inéluctablement une part d'erreur. Cette utilisation est acceptable tant que le risque est maîtrisé et que les erreurs sont limitées.

Partant de ce constat, l'objet du présent mémoire a pour but de mesurer ces erreurs en intégrant l'ensemble des éléments entrant en jeu dans le calcul du *Best Estimate* : données, modèle et hypothèses. Il sera appliqué au "risque décès toutes causes" d'un portefeuille de prévoyance individuelle. Par ailleurs, ce mémoire sera intitulé : "Mesures d'erreur de la meilleure estimation des provisions techniques".

Les différentes parties de cette étude se présentent de la façon suivante. Nous présenterons d'abord la prévoyance individuelle, les détails du contrat couvert par l'entreprise, le calcul de la meilleure estimation et son vocabulaire. Dans un second temps, nous présenterons les divers travaux menés dans ce mémoire pour parvenir à cette notion d'erreur associée au calcul du *Best Estimate*. Ceux-ci ont été articulés en trois grandes phases selon la source d'erreur considérée :

1. **Mesure d'erreurs relatives aux données :** la fiabilité des données utilisées sera analysée à l'aide d'algorithmes servant à détecter des anomalies.
2. **Mesure d'erreurs relatives aux hypothèses actuarielles :** nous détaillerons le processus de construction des hypothèses actuarielles (mortalité par sexe et résiliation) et les sources d'incertitude qui en résultent. Nous proposerons différentes méthodes pour estimer différentes sources d'erreurs (biais d'échantillonnage, stabilité de l'historique...).

3. **Mesure d'erreurs relatives à l'utilisation du modèle de projection :** nous étudierons l'écart entre la meilleure estimation des engagements de l'assureur et la réalité. Cette partie consiste en la réalisation d'un *backtesting* en prenant soin de justifier les méthodes et approximations servant à passer d'une norme comptable à prudentielle.

Enfin, dans une partie intermédiaire, nous analyserons les différentes sources d'erreur en termes d'impact sur le *Best Estimate*.

Chapitre I

Présentation générale du contexte de l'étude

Introduction : Dans ce mémoire, nous allons essayer de déterminer quel biais peut-on associer au calcul de la meilleure estimation des provisions techniques prudentielles. Cette problématique est appliquée à la garantie "décès toutes causes" des produits de prévoyance individuelle de l'entité Natixis Assurances.

Aussi, dans ce chapitre, nous allons commencer par présenter le portefeuille de contrat de prévoyance individuelle sur lequel ce mémoire a été construit. Dans un second temps, nous expliquerons davantage le contexte réglementaire pour ensuite préciser les principes de calculs relatifs à la meilleure estimation. Enfin, dans une troisième partie, nous expliquerons la façon dont cette dernière est calculée et pourquoi il est nécessaire d'en mesurer les différentes sources d'erreurs possibles.

I.1 Contexte métier

L'activité Prévoyance individuelle est couverte au sein du pôle Assurances de Personnes de Natixis Assurances. Les deux sociétés d'assurances françaises du pôle sont les suivantes :

- BPCE Vie, société d'assurance-vie comprenant le périmètre épargne ainsi que, en fonction de la nature de la garantie, une partie des périmètres de la prévoyance individuelle et de l'assurance des emprunteurs ;
- BPCE Prévoyance, société d'assurances non-vie mais d'assurances de personnes portant des risques de type prévoyance individuelle, emprunteur et certains produits IARD associés (assurance des moyens de paiements et des téléphones portables).

En premier lieu, les risques en prévoyance individuelle seront définis afin de mieux comprendre la mécanique des contrats et l'aléa sous-jacent. Cela va permettre de préciser le périmètre de l'étude.

I.1.1 Périmètre de la prévoyance individuelle

La prévoyance individuelle intervient en complément ou en remplacement de la prévoyance collective et vise à protéger les assurés et/ou leurs proches dans d'autres cas que la maladie proprement dite. Tandis que les contrats d'assurance collective de prévoyance sont généralement obligatoires, les contrats d'assurance individuels de prévoyance sont facultatifs.

I.1.2 Risques de la prévoyance individuelle

Les contrats d'assurance prévoyance individuelle proposent des garanties qui viendront s'ajouter aux prises en charge de base de la protection sociale obligatoire. L'objectif de la prévoyance individuelle est de couvrir des risques comme le décès, l'arrêt de travail, les pertes pécuniaires, l'obsèques, les accidents de la vie et la dépendance.

I.1.2.1 Principaux risques

Pour l'assuré, il s'agit de mettre ses proches et/ou soi-même à l'abri des aléas de la vie concernant les risques suivants :

Arrêt de travail : en prévoyance individuelle, l'arrêt de travail regroupe trois types de garanties, de la plus grave à la moins sévère : l'Incapacité Temporaire de Travail, l'Incapacité Permanente de Travail et l'Incapacité Permanente Partielle.

Dépendance : pouvant être totale ou partielle, la dépendance signifie une perte d'autonomie : la personne ne peut plus agir sans avoir recours à quelqu'un ou quelque chose pour effectuer des gestes quotidiens. La dépendance nécessite l'assistance constante d'une tierce personne pour effectuer au moins 2 des 4 actes de la vie quotidienne (s'alimenter, faire sa toilette, s'habiller/se déshabiller, effectuer des transferts). Cet état peut résulter d'un handicap physique ou du fait d'une maladie neurocognitive telle que la maladie d'Alzheimer.

Décès Dans le cadre de l'assurance-vie, l'assurance de personnes "décès toutes causes" a pour objet de garantir la ou les personne(s) assurée(s) contre le risque Décès, qu'il résulte d'une maladie ou qu'il soit la conséquence directe d'un accident. La garantie "décès toutes causes" nécessite l'agrément administratif de la branche vie 20 (vie-décès) dont dispose la société BPCE VIE.

Afin de limiter les risques de survenance de sinistres sur le portefeuille assuré, certaines clauses d'exclusions (conventionnelles et légales telles que le suicide et le meurtre par le bénéficiaire) sont prévues au contrat. En fonction des montants souscrits, une sélection médicale peut également être réalisée à l'adhésion.

En cas de décès de l'assuré, le versement est effectué sous forme de capital ou de rente aux bénéficiaires désignés dans le contrat. On trouve principalement les prestations suivantes :

- capital : un versement unique au bénéficiaire. Lors d'un versement en capital, l'assureur s'engage à payer le capital souscrit au bénéficiaire du contrat si l'assuré décède pendant la période de garantie.

  Remarque : la prestation décès est le plus souvent servie sous la forme d'un capital décès. Ce capital est exprimé en pourcentage de la base de garantie annuelle. Le capital peut être majoré par le nombre d'enfants, le décès simultané du conjoint ou la cause du décès : doublement du capital en cas d'accident.

- rente de conjoint : servie au bénéficiaire à compter du décès de l'assuré, jusqu'à un âge-limite ou de manière viagère.
- rente éducation : pour financer les études des enfants en cas de décès prématuré de l'un des parents. Elle prévoit le versement périodique d'une somme d'argent dont le montant est fixé au contrat. Le montant de la rente est souvent croissant par paliers jusqu'à l'âge terme de 26 ans.
- obsèques : la garantie obsèques propose, au moment du décès de l'assuré, une aide à l'organisation des obsèques et la couverture des frais funéraires au moyen du versement d'un capital aux bénéficiaires.

  Remarque : La garantie obsèques de BPCE Vie inclut une participation aux bénéfices discrétionnaires, c'est-à-dire qu'une part des bénéfices est reversée aux bénéficiaires selon les termes du contrat. Ce genre d'engagement, générant des interactions actif-passif, sa valorisation n'est pas intégrée dans le périmètre de l'étude qui porte sur le calcul déterministe des provisions techniques.

I.1.2.2 Garanties complémentaires au risque Décès

Le contrat peut également prévoir des garanties complémentaires en fonction de la cause ou des circonstances du décès (décès accidentel, orphelin ...). Ces garanties sont principalement utilisées pour assurer la sécurité financière des survivants après le décès. Au sein des produits considérés, on trouve :

- doublement accident ou triplement accident de la circulation : en complément de la garantie "décès toutes causes", cette garantie double le montant du capital si le décès survient par accident. Plus précisément, l'assuré est couvert en cas de décès suite à un accident si le sinistre survient dans un délai de 12 mois suivant la date d'accident. De plus, le montant du capital est triplé en cas d'accident de la circulation.
- maladie grave : il s'agit d'une option facultative garantissant le versement anticipé d'une fraction du capital décès en cas de diagnostic de l'une des maladies graves suivantes : infarctus du myocarde, coronaropathie grave, accident vasculaire cérébral ou médullaire, cancer, insuffisance rénale, greffe d'organes ou sclérose en plaques.

  Remarque : la garantie complémentaire "maladie grave" est proposée par les produits du portefeuille "décès toutes causes" sur l'un des deux réseaux de distribution et est incluse dans notre étude.
- Perte Totale et Irréversible d'Autonomie : également appelée Invalidité Absolue et Définitive (PTIA), cette garantie complémentaire représente l'impossibilité définitive pour l'assuré d'exercer une profession quelconque ou une activité pouvant lui procurer un gain ou un profit. Dans ce cas, l'assuré doit recourir, pour le reste de sa vie, à l'assistance d'une tierce personne pour effectuer les quatre actes ordinaires de la vie (s'habiller, se nourrir, se laver, se déplacer).

  Remarque : en cas de PTIA, l'assuré reçoit le paiement d'un capital décès. En effet, la PTIA est considérée comme une anticipation du décès et entre donc dans le suivi du risque "décès toutes causes". Aussi, dans le cadre de l'étude, la garantie "perte totale et irréversible d'autonomie" (PTIA) est incluse dans la garantie "décès toutes causes".

La prévoyance inclut également l'assurance des emprunteurs, mais ce périmètre ne sera pas abordé dans ce mémoire. Aussi, ce dernier va porter sur les contrats individuels du périmètre de la prévoyance concernant le décès.

I.2 Contexte réglementaire

Dans cette sous-partie, nous développerons davantage le contexte réglementaire dans lequel s'inscrit l'étude. Plus précisément, il s'agit de présenter brièvement la réglementation Solvabilité II, le rôle de la fonction actuarielle ainsi que la notion de provisions techniques.

I.2.1 Objectif de la réglementation

Solvabilité II est une réglementation européenne prenant effet depuis le 1er janvier 2016. Elle repose sur la Directive 2009/138/CE du Parlement européen et du conseil du 25 novembre 2009. Le but de la directive est de créer un cadre prudentiel davantage spécifique aux risques de chaque compagnie d'assurances. Elle insiste sur l'harmonisation des indicateurs de solvabilité et des pratiques au sein de l'Union européenne.

I.2.2 Principes de gouvernance et fonctions clefs

La réglementation Solvabilité II impose aux organismes d'assurance la mise en place d'un système de gouvernance dont le but est de garantir une gestion saine, prudente et efficace de l'activité. Plus précisément, la structure organisationnelle de l'entreprise doit être transparente et les responsabilités doivent être réparties et segmentées de façon claire et appropriée.

I.2.2.1 Présentation des fonctions clefs

Chaque fonction doit pouvoir exercer ses tâches et ses responsabilités de manière objective et indépendante des autres fonctions opérationnelles. Les fonctions clés sont placées sous l'autorité du directeur général, du directoire ou du dirigeant opérationnel.

Le système de gouvernance doit inclure la fonction de gestion des risques, la fonction actuarielle, la fonction de vérification de la conformité et la fonction d'audit interne. Leur rôle respectif est le suivant :

1. **La fonction gestion de risques**

   La fonction de gestion des risques est responsable de la coordination entre les activités de gestion des risques. Ses responsabilités comprennent également la détermination de l'appétence au risque et des limites globales de tolérance au risque, ainsi que l'approbation des principales stratégies et politiques de gestion des risques. Ses principales missions sont les suivantes :

   - veiller à la déclinaison opérationnelle d'un système de gestion globale des risques ;
   - identifier et évaluer les risques émergents ou pas encore pris en compte ;
   - assurer le suivi du profil du risque et des actions de maîtrise des risques engagées sur les risques prioritaires et/ou à forts impacts ;
   - réaliser le *reporting* des expositions aux risques et alerter l'organe d'administration sur les risques.

2. **La fonction vérification de la conformité**

   La fonction clé vérification de la conformité a pour rôle de conseiller l'organe d'administration et de gestion de contrôle, et de veiller au respect des dispositions législatives réglementaires et administratives propres à Solvabilité II.

   Cette fonction doit également prendre en compte un changement possible de l'environnement juridique et en évaluer les impacts sur l'activité de l'entreprise à travers, entre autres, un plan de conformité et une cartographie des risques de non-conformité. Le plan de conformité a pour objectif de détailler les activités de la fonction clé et les expositions au risque de conformité de tous les domaines pertinents de l'entreprise.

3. **La fonction audit interne**

   Le rôle de la fonction clé audit interne est d'évaluer la qualité et l'efficacité du contrôle interne, de la gestion des risques. L'une de ses principales missions est de mettre en place un plan d'audit qui tient compte de toutes les activités de l'entreprise et de son système de gouvernance.

4. **La fonction actuarielle**

   Régie par l'article R354-6 du code des assurances et l'article 272 des actes délégués, la fonction actuarielle fait partie intégrante de la mise en place du dispositif de gestion des risques. Son rôle concerne notamment l'analyse du pilotage technique de l'organisme, sa politique de souscription et la rédaction du rapport actuariel. Ce rapport est présenté au moins annuellement à l'organe de contrôle (conseil d'administration par exemple). Il synthétise les travaux effectués et présente la structure et les missions de la fonction actuarielle. Ces missions sont précisées par l'article 48 de la directive :

Ce rapport, mis à disposition de l'organisme de contrôle, synthétise les travaux effectués et présente la structure et les missions de la fonction actuarielle. Ces missions sont précisées par l'article 48 de la directive :

- Coordonner le calcul des provisions techniques : la fonction actuarielle valide le montant des provisions techniques. Elle est garante de la justesse et de la pertinence des estimations. Cette mission requiert un certain nombre d'études et analyses concernant l'incertitude des estimations, la qualité des données et la cohérence des hypothèses et méthodes utilisées.

  Remarque : Il s'agit de la principale mission de la fonction actuarielle. Aussi, les missions relatives au provisionnement prudentiel seront expliquées ci-après.

- Emettre un avis sur la politique globale de souscription : notamment en s'assurant que la politique de souscription prévoit les travaux et *reporting* nécessaires pour évaluer la viabilité et la rentabilité des portefeuilles. A l'aide d'indicateurs adaptés, cette fonction vieille à la suffisance des primes par rapport aux sinistres. En cela, elle analyse l'effet de l'inflation, du risque juridique, de l'évolution du portefeuille et des systèmes d'ajustement de primes.

- Emettre un avis sur l'adéquation des dispositions prises en matière de réassurance : notamment sur le caractère adéquat du profil de risque et de la politique de souscription, ainsi que sur la qualité des réassureurs. Plus précisément, afin de s'assurer du caractère pertinent des couvertures, elle analyse l'exposition aux risques et aux évènements, en tenant compte du profil de risque de l'entreprise.

- Contribuer à la mise en œuvre du système de gestion des risques : la fonction actuarielle, par ses travaux et les contrôles de second niveau qu'elle effectue, contribue au dispositif de gestion des risques. En surveillant tout risque ou manquement majeur qui serait susceptible de remettre en cause l'évaluation des engagements pour chacun des piliers, elle possède un rôle d'alerte vis-à-vis de l'organe de gouvernance.

La responsabilité de cette fonction est exercée par des personnes qui ont une connaissance et une pratique des statistiques, mathématiques actuarielles et financières et faisant preuve d'une expérience pertinente sur les normes professionnelles et autres normes applicables.

I.3 Provisionnement sous Solvabilité II

Dans cette partie, nous allons brièvement détailler les caractéristiques des provisions techniques prudentielles. Nous commencerons par préciser la nécessité d'un provisionnement et les principes sous-jacents aux provisions techniques afin d'expliquer la notion de meilleure estimation. Cela permettra de préciser la façon dont cette dernière est évaluée en prévoyance individuelle.

I.3.1 Généralités

Du fait de la commercialisation de ses produits, Natixis Assurances doit être en mesure de faire face à ses obligations contractuelles, et doit, le cas échéant, posséder suffisamment de capital pour rembourser ses assurés. C'est cette notion de provisionnement que nous allons maintenant développer.

I.3.1.1 Principe du provisionnement

Il est nécessaire, dans toute activité d'assurance, de pouvoir quantifier les risques afin de garantir la solvabilité face à l'aléa que constitue le paiement probable des prestations. En effet, puisque le produit d'assurance est vendu avant que l'on ne connaisse son coût définitif (principe appelé inversion

du cycle de production), l'assureur a besoin de constituer des réserves et de calculer ses engagements sur l'ensemble de son stock de contrats ainsi que sur de potentiels clients.

Les réserves font partie des passifs d'assurance et correspondent aux obligations contractuelles nettes d'un assureur pour chaque contrat d'assurance. La façon dont elles sont déterminées n'est pas figée, elle évolue au fil des directives et des réglementations. Le but du provisionnement est de quantifier les engagements de l'assureur afin de savoir quel montant mettre en réserve.

I.3.1.2 Bilan sous solvabilité II

Le bilan est un état représentatif du patrimoine de l'entreprise. Sous la Directive Solvabilité II, le bilan prudentiel se veut mieux adapté au profil de l'assureur. Il recense l'ensemble des actifs économiques et des emplois engagés par l'entreprise à un instant donné. Il se compose de deux parties égales et peut être représenté de la façon suivante :

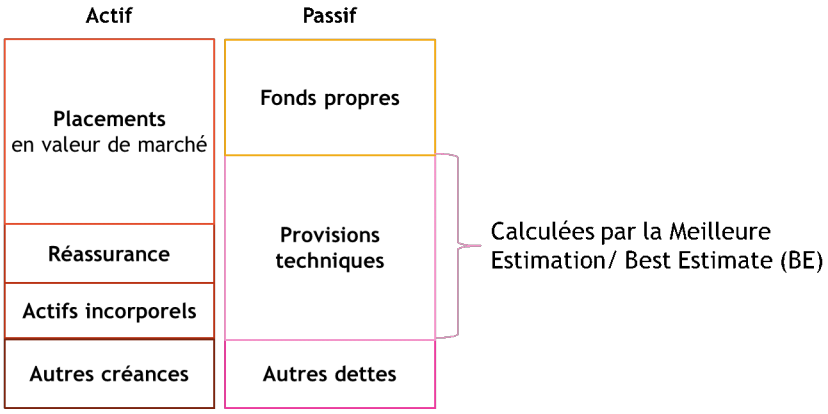


FIGURE I.1: Représentation d'un bilan économique sous la directive Solvabilité II

- L'actif d'une société est constitué de l'ensemble des biens qu'elle possède ainsi que l'ensemble des créances détenues sur des tiers. Dans les placements, on trouve principalement les produits financiers comme les obligations, les actions et les produits dérivés, et les biens immobiliers. En assurance, l'actif est essentiellement représenté par des placements, des créances et des disponibilités.

- Le passif est constitué de l'ensemble des dettes contractées par l'entreprise. Le passif est essentiellement composé des engagements contractés à l'égard des assurés, c'est-à-dire les provisions techniques. Les provisions techniques représentent montant des engagements de la société d'assurances à l'égard de l'ensemble des assurés.

Les fonds propres disponibles sont inscrits au passif du bilan et ont pour rôle de garantir la solvabilité future de l'entreprise : même si celle-ci fait une perte elle ne sera pas en faillite. Afin de déterminer ce montant, il faut sommer la totalité des actifs auxquels on ôte les provisions techniques et les autres dettes. Les provisions techniques sont obtenues via le calcul du *Best Estimate*. C'est pourquoi le *Best Estimate* est une notion importante sous Solvabilité II que nous allons ensuite détailler.

I.3.2 Caractéristiques des provisions techniques

Sous la précédente directive de solvabilité, les provisions techniques étaient déterminées en intégrant une marge de prudence dans chacun des paramètres utilisés lors du calcul des provisions. Pour la réforme Solvabilité II, la notion de cohérence avec le marché est appliquée au calcul des provisions techniques afin que les parties actif et passif du bilan correspondent. Aussi, les passifs sont calculés en valeur économique. Cela signifie qu'ils sont valorisés au montant pour lequel ils pourraient être transférés à une autre entité dans des conditions de concurrence normale.

I.3.2.1 Méthodes de valorisation des provisions techniques

Pour l'évaluation des provisions, la réglementation Solvabilité II distingue les passifs réplicables et les passifs non réplicables. Est considéré comme réplicable tout contrat d'assurance dont les flux peuvent être parfaitement représentés par des instruments financiers disponibles sur un marché suffisamment actif, liquide et transparent. Deux cas sont possibles :

- Dans le cas d'un passif réplicable, le montant de la provision technique est égal à la valeur de marché du portefeuille (*mark-to-market*).
- Si le passif d'assurance est non réplicable, la provision technique est égale à la somme du *Best Estimate* et de la marge pour risque.

I.3.2.2 Définition des provisions techniques

Comme expliqué, la valorisation économique repose sur l'hypothèse d'existence d'un marché liquide où l'on peut transférer des portefeuilles d'assurance ce qui n'est pas le cas. Aussi, en prévoyance individuelle, l'évaluation des provisions techniques se fait de façon simplifiée, en décomposant les provisions techniques comme la somme des éléments suivants :

- *Best Estimate* (BE) : aussi appelée "meilleure estimation des engagements", c'est la valeur actuelle probable des flux futurs des obligations de l'assureur. Plus précisément, d'après l'article 77 de la Directive Solvabilité II, « La meilleure estimation correspond à la moyenne pondérée par leur probabilité des flux de trésorerie futurs, compte tenu de la valeur temporelle de l'argent (valeur actuelle attendue des flux de trésorerie futurs), estimée sur la base de la courbe des taux sans risque pertinent. Le calcul de la meilleure estimation est fondé sur des informations actualisées et crédibles et des hypothèses réalistes et il fait appel à des méthodes actuarielles et statistiques adéquates, applicables et pertinentes ».

  Remarque : cette valeur actuelle probable correspond à une valeur moyenne. Ce concept écarte toute notion de marge de prudence dans les provisions techniques, même si le *Best Estimate* doit effectivement tenir compte de toutes les incertitudes liées aux flux futurs, ainsi que de la valeur temps de l'argent. Le calcul de cette estimation sera détaillé plus tard.

- Marge pour Risque (RM) : la marge pour risque permet de lisser les fluctuations des simulations, de sorte que, même si la meilleure estimation des engagements est inférieure à la valeur réelle des engagements, l'assureur conserve une certaine marge de solvabilité. La notion de marge de risque vient donc s'ajouter au «*Best Estimate*», le montant de provision attendu et évalué au plus juste, pour former le montant total de la provision à constituer. En un sens, cette marge intervient en complément, de façon à couvrir les différentes sources d'incertitudes relatives aux provisions techniques.

Remarque : la marge pour risque est calculée de manière à ce que le montant total des provisions techniques soit égal au montant qu'exigerait une entité de référence pour elle-même honorer ses engagements à la charge de l'assureur. Mathématiquement, la marge pour risque est évaluée en actualisant le coût du capital annuel généré par l'immobilisation du Capital de Solvabilité Requis (ou en anglais, *Solvency capital required SCR*). Ce dernier correspond au capital économique dont une entreprise d'assurance a besoin pour limiter sa probabilité de ruine à un an à $0,5\%$, c'est-à-dire à une seule occurrence tous les 200 ans. On estime généralement le coût de cette immobilisation à 6% par an sur la durée résiduelle des engagements utilisés pour le calcul du *Best Estimate*. Le calcul de la marge pour risque est réalisé sur l'activité entière, tenant compte de chacune des branches en fonction de leur contribution marginale au SCR.

L'évaluation de la marge pour risque n'est pas traitée dans ce mémoire.

Ainsi, en prévoyance individuelle, la directive Solvabilité II définit la valeur des provisions techniques comme étant égale à la somme de la meilleure estimation et de la marge de risque. C'est la notion de *Best Estimate* que nous allons maintenant détailler davantage tout au long de notre étude.

I.4 Valorisation et description du *Best Estimate*

La meilleure estimation des provisions techniques, c'est-à-dire le *Best Estimate*, se calcule en fonction de flux futurs estimés, actualisés selon la courbe des taux sans risque et repose sur le calcul de la valeur actuelle nette.

I.4.1 Calcul de la Valeur Actuelle Nette et du *Best Estimate*

Le *Best Estimate* est défini comme la valeur actuelle probable des flux de trésorerie futurs relatifs aux engagements d'assurance (cotisations, prestations, frais, commissions...). Plus précisément, ces flux sont estimés à l'aide d'hypothèses réalistes pour les différents facteurs de risques susceptibles de modifier la valeur des flux futurs. Ces derniers sont ensuite actualisés au moyen d'une courbe de taux sans risque à date d'évaluation $t = 0$ puis sommés pour obtenir le montant de *Best Estimate*.

On utilise la notation suivante afin de définir un vecteur de flux futurs (ou de *cash flows* (CF)) d'un portefeuille X : $CF_{0:T}(X) = \{CF_0(X), ..., CF_T(X)\}$.

Valeur Actuelle Nette (VAN) : indicateur utilisé en assurance, il constitue la base du calcul des engagements. La VAN consiste à faire la somme des *cash flows* futurs tout en les actualisant :

$$VAN_T(CF_{0:T}(X)) \;=\; \sum_{t=0}^{T} \frac{CF_t(X)}{(1+tx_t)^t}$$

où tx_t représente le taux d'actualisation de la période considérée et T l'horizon considéré.

Il est possible de calculer cette VAN sur de nombreux *cash flows* différents, tels que ceux obtenus par l'utilisation des modèles multi-états.

Expression mathématique : le *Best Estimate* représente le gain ou la perte espéré pour un portefeuille donné. Expliqué différemment, c'est le calcul de la VAN du résultat assureur espéré $Y_t(X)$ pour un portefeuille X :

$$\begin{aligned} BE_T(X) &= \mathbb{E}^{\mathbb{Q}}[Y_{0:T}(X)] \\ &= VAN_T(Y_{0:T}(X)) \\ &= \sum_{t=0}^{T} \frac{Y_t(X)}{(1+tx_t)^t} \end{aligned}$$

où tx_t représente le taux d'actualisation en univers risque neutre de la période considérée,
où X un portefeuille ou bien une tête et T l'horizon considéré.

I.4.2 Description élémentaire du *Best Estimate*

Il est possible de préciser un peu plus la formule du *Best Estimate* à l'aide du résultat technique de l'assureur. Pour pouvoir valoriser des provisions techniques, il est nécessaire de déterminer avec précision les engagements de l'entreprise envers les assurés et tout autre tier, et réciproquement de déterminer les engagements des assurés et tiers envers l'entreprise d'assurance ou de réassurance. Ces engagements correspondent aux flux futurs.

I.4.2.1 Résultat de l'assureur

Une fois le principe de calcul expliqué, il est nécessaire de détailler quels flux futurs seront estimés et comment ils sont assemblés pour reconstituer le résultat de l'assureur. Cette étape nous permettra de plus facilement analyser les résultats au cours de l'étude. Le résultat assureur espéré est obtenu à partir de la formule suivante :

$$Y_t(X) = \mathbb{E}^{\mathbb{P}}[\Lambda_t(X)]$$

où $\Lambda_t(X)$ représente le résultat assureur dans un état du monde historique (mesuré par la proba $\mathbb{P}$) à chaque pas de temps t du portefeuille X.

Tout d'abord, il existe deux catégories de flux futurs :

- les flux **entrants** comprennent les primes futures, ainsi que dans une moindre mesure, les montants recouvrables au titre de sauvetage et dérogations.
- les flux **sortants** sont à séparer en trois parties :
  - les prestations : couvrent l'ensemble des règlements des sinistres, les prestations en cas de décès, d'invalidité, de rachats, les rentes et les participations aux bénéfices ;
  - les frais : frais de gestion, d'administration et de gestion de sinistres. On y ajoute les taxes, c'est-à-dire les impôts à payer facturés aux souscripteurs comme les taxes sur les primes par exemple ;
  - les commissions : rémunération versée au réseau distributeur.

En utilisant les flux futurs, le résultat de l'assureur peut être détaillé de la manière suivante :

$$\boxed{\Lambda_t(X) = -Primes_t(X) + Prestations_t(X) + Frais_t(X) + Commissions_t(X)}$$

I.4.2.2 Description des flux futurs

Pour pouvoir analyser les résultats des futures études, il est nécessaire de détailler les flux constituant chacune des parties du *Best Estimate*. Nous allons commencer par quelques définitions comptables afin de mieux appréhender les différentes composantes.

Primes futures : les primes futures correspondent au montant des cotisations, diminué des taxes et commissions. Elles représentent ce que va toucher l'assureur durant la période de projection et sont calculées en fonction de la date de renouvellement et du délai de reconduction contractuel.

Les primes futures retenues dans le calcul du *Best Estimate* correspondent aux primes commerciales hors taxe restant à émettre sur les adhésions en cours à la date d'extraction des données. Ces primes sont probabilisées : par exemple, on y applique les probabilités mensuelles de décès et de résiliation correspondant aux caractéristiques de l'adhésion.

Prestations : les prestations incluses dans le calcul de la meilleure estimation correspondent aux engagements générés par le paiement des primes futures. Ce sont donc les sinistres couverts par les primes en question.

Les prestations futures des risques décès correspondent à la probabilité de décès (accidentel ou toutes causes) appliquée au capital garanti.

Frais généraux : les frais généraux observés sont théoriquement couverts par les chargements de gestions contenus dans les cotisations payées par les assurés. Ils représentent le coût additionnel engendré par la commercialisation et la gestion des contrats durant la période de projection et sont calculés en fonction de la date de renouvellement et du délai de reconduction contractuel.

Commission sur résultat : la rémunération des réseaux de distribution se compose d'une part fixe exprimée en pourcentage des primes perçues, distribuée à la souscription, ainsi que d'une rémunération variable calculée sur le résultat technique de l'assureur. On parle alors de commissions sur résultats. Elle correspond dans notre cas à 50% du résultat brut.

I.4.2.3 *Best Estimate* de primes et *Best Estimate* de sinsitres

La provision *Best Estimate* peut ensuite être décomposée entre deux parties distinctes. Selon la provenance des flux futurs, on distingue les flux associés aux survenances passées et les flux relatifs aux possibles survenances futures :

- **le *Best Estimate* de primes** : correspond à la période de garantie non encore écoulée. Plus précisément, cette partie couvre le risque que les primes non-acquises soient insuffisantes par rapport au montant des prestations pour sinistres futurs.

  Cela permet de décomposer les flux futurs de trésorerie relatifs au *Best Estimate* de primes de la façon suivante :

$$BE_T^{Primes}(X) = \sum_{t=0}^{T} \frac{\Lambda_t^{Primes}(X)}{(1+\ tx_t)^t}$$

  Où, les flux futurs du *Best Estimate* de primes sont les suivants :

$$\begin{aligned}
\Lambda_t^{Primes}(X) \quad = \quad & - \quad primes\, futures_t(X) \\
& + \quad commissions\, sur\, primes_t(X) \\
& + \quad prestations\, futures_t(X) \\
& + \quad frais\, \text{généraux}\, sur\, primes_t(X) \\
& + \quad frais\, \text{généraux}\, {}^{futurs}_{sur\, sinistres}{}_t(X) \\
& + \quad commissions\, sur\, \text{résultat}\, {}_{primes}{}_t(X)
\end{aligned}$$

- **le *Best Estimate* de sinistres** a pour objectif de couvrir les risques déjà survenus et les éventuels sinistres en cours à date d'évaluation. De même, on peut reconstituer le *Best Estimate* de sinistres de la façon suivante :

$$BE_T^{Sinistres}(X) = \sum_{t=0}^{T} \frac{\Lambda_t^{Sinistres}(X)}{(1+\ tx_t)^t}$$

  Où les flux futurs du *Best Estimate* de sinistres sont les suivants :

$$\begin{aligned}\Lambda_t^{Sinistres}(X) \quad = \quad & + \quad prestations\, connues_t(X) \\ & + \quad prestations\, inconnues_t(X) \\ & + \quad frais\, \text{généraux}_{sur\ sinistres}^{connus\ t}(X) \\ & + \quad frais\, \text{généraux}_{sur\ sinistres}^{inconnus\ t}(X) \\ & + \quad commissions\, sur\, \text{résultat}_t^{sinistres}(X)\end{aligned}$$

I.4.3 Mise en œuvre du calcul de la meilleure estimation

Après avoir expliqué les principes et les différentes composantes du *Best Estimate*, il convient d'expliquer la manière dont ce dernier est calculé. L'assureur ne pouvant connaître avec certitude le montant de primes futures, des sinistres et des frais occasionnés, il est obligé d'estimer les différents flux futurs au moyen de modèles.

Un modèle constitue une représentation de certains aspects du monde à partir d'hypothèses simplificatrices. Il permet d'obtenir des résultats à partir des données et d'hypothèses, et implique en cela un certain nombre de méthodes. L'une d'entre elles est la projection des états possibles de l'assuré.

Modèle Central : le modèle central de Natixis Assurances est un outil en charge de la diffusion du bilan et du compte de résultats sur les différents périmètres modélisés. Il permet de projeter les flux futurs et provisions que l'on peut associer aux portefeuilles de contrats, aux actifs des différents fonds comptables, aux chargements, frais généraux et commissions ainsi qu'aux traités de réassurance.

Le modèle central a pour vocation de répondre à l'enjeu de rationalisation des modèles actuariels utilisés dans l'entreprise et est aujourd'hui utilisé dans un certain nombre de processus, tels que l'évaluation des provisions et des bilans en normes comptables, IFRS et Solvabilité II. L'utilisation du modèle central permet de garantir la cohérence des indicateurs en sortie des modèles actuariels utilisés dans le pilotage de l'entreprise, exigence définie par l'article 120 de la directive Solvabilité II.

Selon le périmètre modélisé, les projections réalisées sont différentes :

- Projections stochastiques, avec interactions actif-passif : épargne et produits avec participation aux bénéfices discrétionnaire tels que l'assurance obsèques.
- Projections déterministes : seuls les flux de passif sont projetés. Il s'agit de la prévoyance individuelle et de l'assurance des emprunteurs.

  Remarque : compte tenu de l'absence d'interaction actif-passif sur ce périmètre, les portefeuilles en prévoyance individuelle peuvent être projetés indépendamment de l'actif, et même chaque contrat peut être projeté indépendamment au sein du portefeuille.

Aussi, en prévoyance individuelle, le modèle central décès et arrêt de travail (DCAT) repose sur un modèle de projection calibré à partir d'hypothèses déterministes.

I.4.4 Eléments nécessaires au modèle décès et arrêt de travail

Le modèle central couvre les garanties "décès toutes causes", "décès accidentel", "arrêt de travail" et "invalidité". Il consiste à projeter chaque assuré existant à date de projection via un modèle multi-états et nécessite les éléments suivants :

I.4.4.1 Paramètres du modèle central

Les « paramètres » correspondent aux informations de réglage utilisées dans le modèle. Le choix de mode de projection du modèle, la date de début de projection ou encore la graine nécessaire à la génération des scénarios économiques font partie des paramètres les plus importants.

I.4.4.2 Données utilisées

Pour les contrats relatifs à la prévoyance individuelle, plusieurs types d'informations sont utilisés par le modèle :

- données comptables bilancielles : l'objectif du modèle central étant la diffusion du bilan des filiales de Natixis Assurances, les bilans initiaux sont un *input* majeur du modèle.
- portefeuilles de contrats : la réalisation de projection se fait sur la base de contrats ligne par ligne. Chaque ligne représente un assuré présent à la date de projection et contient les garanties souscrites. D'autres éléments d'informations sont pris en compte : le genre de l'assuré, sa date de naissance, la date de souscription du contrat et, pour les contrats sinistrés, la date de déclaration du sinistre.
- frais généraux : les frais généraux projetés, représentant le coût additionnel engendré par la commercialisation et la gestion des contrats, ont une incidence majeure sur les indicateurs en sortie du modèle.

I.4.4.3 Hypothèses utilisées

Les « hypothèses » correspondent aux informations visant à modéliser de manière vraisemblable un ensemble de faits. Celles-ci doivent par conséquent être soumises au contrôle de l'expérience ou vérifiées concernant l'impact de leur utilisation. Ces hypothèses doivent dès lors faire l'objet d'une spécification de la méthodologie utilisée pour les établir, d'une validation de celles-ci et d'études de *backtesting*.

Plusieurs types d'hypothèses sont utilisés par le modèle central afin de mesurer les engagements générés par les portefeuilles de prévoyance individuelle :

- Lois biométriques : les lois biométriques incluent les tables de mortalité ainsi que les lois de passage d'état pour les modèles multi-états. Ces lois biométriques incluent les tables de mortalité ainsi que les lois de passage d'état pour les modèles multi états.
- Lois comportementales : les assurés ayant la capacité d'agir sur leurs contrats d'assurance de différentes manières (rachat partiel, rachat total, résiliation), il est possible de modéliser les probabilités d'occurrence de ces actions de manière statistique.
- Hypothèses de diffusion des frais généraux : l'objectif étant de diffuser les frais généraux des différentes filiales, il est nécessaire de faire des hypothèses quant à la diffusion de ceux-ci.
- Taxes et impôts : les taxes et impôts sont modélisés. En effet, il faut prendre en compte les spécificités aux différents pays dans lesquels Natixis Assurances a des filiales.

Remarque : les lois biométriques et comportementales font partie des hypothèses utilisées par le modèle de projection multi-états. Ce sont les lois qui feront l'objet d'une étude plus approfondie.

Une fois les données et les hypothèses utilisées explicitées, il convient d'en expliquer l'utilisation à travers le fonctionnement d'un modèle multi-état.

I.4.5 Le modèle de projection multi-états

Le modèle de projection du modèle central est un modèle multi-état. Du fait de la disponibilité de données individuelles de portefeuilles et de la suffisance de données nécessaires à la construction de lois de passage d'état, le choix d'un modèle multi-état a été conduit.

I.4.5.1 Fonctionnement d'un modèle multi-états

Ce type de modèle consiste à diffuser de manière prospective l'effectif des assurés. Le but est de faire vivre le portefeuille de contrats existants à date d'évaluation t, jusqu'à la date de fin des contrats. Le lecteur intéressé pourra se référer à la thèse de Q. Guibert (2015) [22]. Les états possibles d'un assuré sont étroitement liés aux garanties incluses dans les contrats d'assurance.

Etats possibles : au cours de la période de projection, l'assuré peut se situer dans les états suivants :

- *valide* : l'assuré n'est ni décédé, ni en arrêt de travail et n'a pas résilié ;
- *incapable* : l'assuré devient incapable. Il était autrefois valide et peut le redevenir ;
- *invalide* : généralement après avoir été incapable, l'assuré devient invalide. On considère qu'il ne peut plus redevenir valide ou incapable ;
- *résilié* : l'assuré choisissant de mettre fin au contrat ou bien de l'assureur suite au non-paiement des cotisations ;
- *décédé* ou en PTIA. En effet, l'état de PTIA est traité exactement de la même façon que l'état *décédé* ce qui signifie un assuré peut donc passer à l'état de PTIA n'importe quand.

Transitions possibles : Au départ, en $t = 0$, tout l'effectif se situe dans l'état *valide*, au pas suivant il est diffusé vers les autres états en fonction des probabilités associées aux différents changements d'état. Ensuite, le modèle calcule à chaque pas de temps les effectifs probables moyens dans chaque état en prenant en compte leurs interactions de façon que, à une date t, l'effectif est vu comme un effectif probabilisé.

Cette approche consiste à multiplier des probabilités conditionnelles aux effectifs probabilisés de la période précédente. Les assurés sont susceptibles de changer d'état selon le schéma suivant :

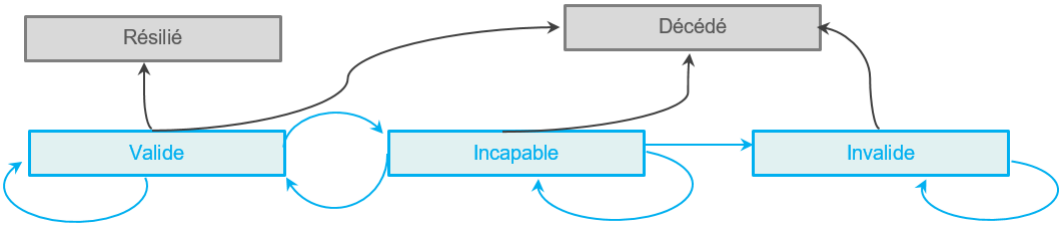


Figure I.2: Modèle multi-états

En observant ce schéma, on remarque que le décès et la résiliation sont des états absorbants : une fois les états *décédé* ou *résilié* atteint, l'assuré n'est plus considéré comme porteur de risque. A l'inverse, l'arrêt de travail et la validité sont des états transitoires. Pour le cas de l'incapacité et en fonction des lois de passage utilisées, l'ancienneté dans l'état peut être à prendre en compte dans la diffusion des effectifs. En effet, les chances de rétablissement sont généralement plus importantes pour un assuré en arrêt depuis trois mois plutôt que pour un assuré en arrêt depuis deux ans.

Concernant l'horizon de projection, les contrats de prévoyance individuelle couvrent le décès et l'arrêt de travail sur la durée d'effet des garanties. En prévoyance individuelle, mise à part pour l'assurance obsèques qui ne fait pas l'objet de notre étude, cette durée est limitée à un an.

I.4.6 Rôle de la fonction actuarielle dans le cadre du calcul des provisions techniques

Le calcul du *Best Estimate* se base sur des données dont le caractère est approprié, exhaustif et exact. Le Règlement Délégué insiste sur cette dimension du calcul des provisions, qui doit faire appel aux informations les plus actuelles et crédibles, ainsi qu'à des hypothèses réalistes.

Les divers travaux réalisés s'inscrivent dans une démarche d'amélioration constante des méthodologies, données, modèles et hypothèses sous-tendant le calcul des provisions techniques. Plus précisément, la fonction actuarielle est chargée de coordonner le calcul des provisions techniques en respectant les missions suivantes :

- garantir le caractère approprié des méthodologies, des modèles sous-jacents et des hypothèses utilisées pour le calcul des provisions techniques ;
- apprécier la suffisance et la qualité des données utilisées dans le calcul des provisions techniques ;
- comparer les meilleures estimations aux observations empiriques ;
- informer l'organe d'administration, de gestion ou de contrôle de la fiabilité et du caractère adéquat du calcul des provisions techniques.

Aussi, la fonction actuarielle présente tout au long de l'année les travaux qu'elle a effectués afin d'améliorer la fiabilité du calcul des provisions techniques. Elle produit diverses études visant à conforter la suffisance des provisions permettant également de déceler des défaillances potentielles liées aux données, hypothèses, ou méthodologies liées au modèle central. Des axes d'amélioration sont identifiés en vue, d'une part, de répondre plus pleinement aux exigences de la réglementation et, d'autre part, de fiabiliser procédures et résultats.

I.5 Objectifs de l'étude

Le calcul des capitaux requis pour garantir la solvabilité de l'entreprise se base sur la projection des flux futurs de chaque portefeuille. Cela nécessite un contrôle du niveau des provisions par l'autorité de contrôle, mais surtout en premier lieu par la compagnie elle-même. En effet, la réglementation Solvabilité II pousse les assureurs à s'interroger sur leur culture du risque ainsi que leur stratégie. Comme expliqué précédemment, la revue des données, modèles, hypothèses et résultats relatifs aux provisions techniques conduit à une appréciation de la part de la fonction actuarielle. En effet, la fonction actuarielle doit évaluer l'incertitude des provisions techniques.

Néanmoins, un modèle de provisionnement demeure une représentation imparfaite des engagements. C'est pourquoi, la meilleure estimation possible des provisions techniques comporte donc inéluctablement une part d'erreurs. Aussi, il est nécessaire de prendre en compte les sources d'erreurs possibles associées au calcul des provisions techniques. Plus précisément, le calcul du *Best Estimate* requiert des données représentant le portefeuille d'assuré, un modèle de projection ainsi que des hypothèses comptables et actuarielles (hypothèses elles-mêmes construites à l'aide de données et de modèles). Cela explique la nécessité de prendre en compte toutes les sources d'erreurs possibles associées au calcul des provisions techniques.

Partant de ce constat, l'objet du présent mémoire a pour but d'intégrer les erreurs générées par l'ensemble des éléments entrant en jeu dans le calcul de la meilleure estimation des provisions techniques : données, modèle et hypothèses. Il sera appliqué au risque "décès toutes causes" du portefeuille de prévoyance individuelle de Natixis Assurances. Ce portefeuille sera segmenté selon le réseau de distribution (réseau 1 et réseau 2) [1].

1. Pour des raisons de confidentialité, les réseaux de distribution ne sont pas spécifiés. Néanmoins, comme les réseaux se caractérisent par des cibles de clientèles, un système d'information, et une profondeur d'historique différents, nous allons quantifier les engagements séparément.

Chapitre II

Mesure d'erreurs relatives aux données

Introduction : La détection d'anomalies permet d'améliorer la qualité des données par suppression ou modification des données anormales. Dans d'autres cas, les anomalies traduisent un événement et apportent de nouvelles connaissances utiles. Par exemple, la détection d'anomalies peut prévenir un dommage matériel et donc inciter à la maintenance prédictive dans le secteur de l'industrie. Cette branche de la *data science* trouve son application dans plusieurs domaines comme la médecine, la cybersécurité, la finance, la prédiction des catastrophes naturelles, et bien d'autres domaines.

Dans les statistiques, une anomalie (également appelée aberration) est une observation qui s'écarte de la norme, au point d'éveiller la suspicion. Au sein d'un jeu de données, une donnée aberrante peut être due au hasard ou bien indiquer la présence d'erreurs, d'inexactitudes. Ces données anormales ne représentent pas fidèlement la réalité à laquelle elles se réfèrent. Aussi, elles doivent faire l'objet de recherches approfondies.

Dans le cadre d'une estimation *best estimate*, utiliser des données de bonne qualité est essentiel. En effet, les données étant la matière première de toute étude, utiliser des données de mauvaise qualité introduit un biais supplémentaire dans l'exactitude des résultats. Nous proposons donc une méthodologie pour mesurer l'erreur associée à la qualité des données dans le cadre du calcul du *Best Estimate*.

Pour ce faire, nous préciserons d'abord les méthodes actuelles permettant de garantir la fiabilité des données dans le cadre de la réglementation Solvabilité II. Nous décrirons ensuite les techniques que nous allons utiliser. Enfin, nous détaillerons la démarche adoptée, c'est-à-dire comment ces diverses méthodes de détection sont utilisées afin de répondre à notre objectif.

II.1 Contextualisation et présentation de la démarche

L'amélioration de la qualité des données constitue un enjeu majeur pour les organismes assureurs. Ces derniers font face au grand débit, toujours croissant, de flux de données associées à leurs clients, assurés, ayants droits, bénéficiaires et tiers... L'utilisation des données de qualité est essentielle pour toutes les activités de l'organisme : gestion, marketing, offre, relation client, conformité, actuariat, finance...

Elle permet notamment une meilleure une fiabilisation et accélération de la prise de décision ainsi qu'une meilleure maîtrise des risques. Dans le cade de l'optimisation du parcours client, la qualité des données utilisées est déterminante dans l'établissement d'un contact, la proposition d'un tarif adapté, puis par la suite, dans la fidélisation (service client personnalisé, mesure du niveau de satisfaction, facilitation des procédures).

II.1.1 Généralités sur la qualité des données

Les données aberrantes peuvent avoir de multiples causes : un appareil de mesure peut présenter un dysfonctionnement, il peut y avoir une erreur dans la transmission ou dans la retranscription des données, ou bien un changement de procédure, un comportement frauduleux ou une erreur humaine. De plus, il arrive que la donnée n'appartienne pas au domaine d'observation : l'échantillon a été « contaminé » par des individus différents de la population étudiée. C'est le cas lorsque des anomalies sont le résultat d'actions malveillantes telles que la fraude à l'assurance.

II.1.1.1 Données utilisées

Les données peuvent être définies comme l'ensemble des éléments connus, faits ou informations collectés à partir de recensements ou d'observations. Ces dernières peuvent être de nature variée (numérique, catégorielle ou ordinale) et nous verrons que cela conditionnera le choix des méthodes de détection à utiliser.

Les données peuvent être de deux types. Il peut s'agir de données brutes comme les données liées à la définition des contrats, des sinistres, de la réassurance. Si les données proviennent d'un calcul, ou plus largement de transformations effectuées sur les données brutes, ce sont des données calculées. Dans le cadre de notre étude, nous utilisons principalement de données relatives aux assurés, de données de sinistralité ou données liées aux contrats tels que des conditions particulières. Certaines colonnes, telles que des variables de censures, d'âge ou des variables catégorielles, ont été obtenues par retraitements, c'est-à-dire une combinaison d'opérations de recherches, de choix, de tri, de regroupement, et de concaténation.

Il est aussi important de prendre en compte la maille de la base de données, c'est-à-dire le degré d'agrégation ainsi que la finesse de l'information retranscrite. Une base de données correspond à une certaine façon d'organiser et de découper l'information. Ce découpage est basé sur des critères de segmentation et de regroupement. C'est l'utilisation que l'on en fait qui détermine sa structure.

Dans notre cas, le terme « donnée » désigne toute information utilisée directement ou indirectement pour procéder à l'évaluation de provisions techniques, en particulier lors de l'utilisation du modèle de projection ainsi que de méthodes actuarielles et statistiques. Cette utilisation doit être en adéquation avec les engagements d'assurance et les spécificités du portefeuille concerné. Pour les différentes bases considérées, chaque ligne des bases de données représente un unique assuré du portefeuille.

II.1.1.2 Anomalies que l'on souhaite détecter

Etant parfois saisies manuellement, il arrive que les données utilisées lors du calcul des provisions techniques soient mal renseignées et constituent des anomalies. Or, la présence d'anomalies génère des résultats erronés que ce soit dans la construction des hypothèses et dans la projection du portefeuille d'assuré. Ces erreurs affectent la fiabilité du calcul des provisions techniques. Aussi, il est nécessaire de produire des données de qualité garantissant un reflet fidèle des engagements des assureurs. On suppose ici que les erreurs contenues dans nos bases de données peuvent être les suivantes :

- Mauvaise insertion des données dans la base ou représentation d'un phénomène anormal (comportement frauduleux, échantillon contaminé.) ;
- Mauvaise récupération dans le système de gestion ;
- Mauvais retraitements effectués préalablement à l'utilisation de techniques statistiques. Par exemple, dans le cas de données manquantes, la base est parfois complétée selon des choix arbitraires.

II.1.2 Exigence de qualité des données sous le référentiel de calcul Solvabilité II

La qualité des données dépend d'abord du contexte dans lequel les données ont été saisies à l'origine, mais d'un point de vue décisionnel, leur intérêt dépend de la finalité dans laquelle l'utilisateur peut les utiliser. Par conséquent, la qualité des données ne doit pas être évaluée de manière absolue, mais doit être évaluée d'une manière adaptée à l'utilisation qui en est faite.

Disposer de données de qualité est un sujet abordé dans les trois piliers de la réglementation Solvabilité II. C'est un enjeu transversal à l'activité de la compagnie d'assurances. En effet, il est nécessaire d'utiliser des données fiables pour appréhender correctement les différents risques, mener les diverses études actuarielles, réaliser les tarifications et évaluer les provisions. Ainsi, les données sont la matière première des travaux, modèles et études. Elles doivent être adaptées aux besoins de l'assureur et être au centre des préoccupations.

C'est pourquoi, l'article 82 de la Directive établit que « les Etats veillent à ce que les entreprises d'assurances mettent en place des processus et procédures internes de nature à garantir le caractère approprié, l'exhaustivité et l'exactitude des données utilisées dans le calcul de leurs provisions techniques ».

Remarque : un défaut de qualité des données peut entraîner les difficultés suivantes :

- conduire à une dégradation ou à l'allongement des travaux et des analyses. Cela peut nuire au choix de l'assureur (mauvaise interprétation de la situation par exemple).
- générer un coût conséquent en fonds propres. En effet, une granularité trop faible et une tendance à corriger les anomalies de manière prudente conduisant généralement à des calculs pessimistes et donc à un surcoût en capital.

II.1.2.1 Missions de la fonction actuarielles relatives à la qualité des données

La qualité des données dans le cadre du calcul des provisions techniques est une des obligations de la fonction clé actuarielle. Cette dernière est tenue « d'apprécier la suffisance et la qualité des données utilisées dans le calcul des provisions techniques » et de mettre en place des « processus et procédures internes de nature à garantir le caractère approprié, l'exhaustivité et l'exactitude des données utilisées dans le calcul de leurs provisions techniques ».

De plus, en tant que fonction de contrôle de second niveau, elle s'inscrit également dans le processus de collecte et d'analyse de données en livrant post-production une appréciation de la suffisance et la qualité des données utilisées dans le calcul des provisions techniques. En effet, le rôle de la fonction actuarielle dans le cadre de la qualité des données est analogue à son rôle dans le calcul des provisions, la politique de réassurance ou la politique de souscription.

Ainsi, la fonction actuarielle doit être un garant permanent de la bonne maîtrise de la qualité des données. Cette appréciation fait partie intégrante du rapport actuariel. Pour cela, la fonction actuarielle s'appuie sur les notes de qualification visant à valider la qualité de la donnée conformément aux critères définis dans l'article 19 du règlement délégué.

II.1.2.2 Critère de qualité des données selon la réglementation Solvabilité II

La directive Solvabilité II comporte dans ses articles 82, 76 et 84, des exigences sur la qualité des données utilisées pour le calcul des provisions techniques. Elle demande ainsi aux entreprises d'assurances de mettre en place « des processus et procédures internes de nature à garantir le caractère approprié, l'exhaustivité et l'exactitude des données utilisées dans le calcul de leurs provisions techniques » (article 82, 1er §). Ces critères interviennent à des moments différents :

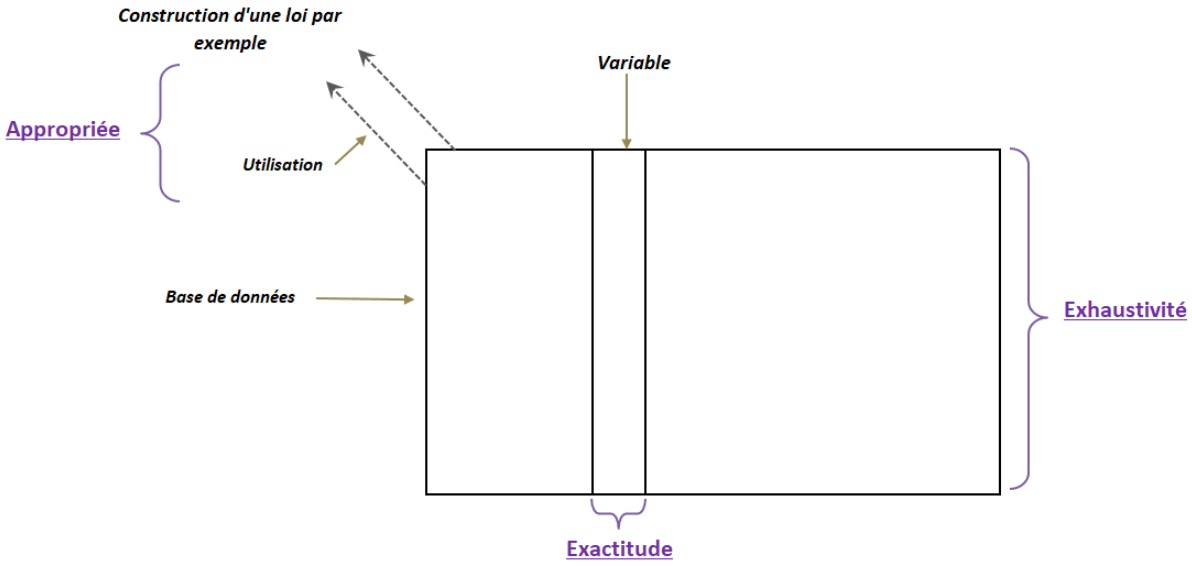


FIGURE II.1: Représentation des trois critères de qualité des données

Plus précisément, on peut expliciter les critères définissant la qualité des données dans le tableau suivant :

Exhaustivité	Complétude nécessaire au calcul des provisions techniques ▫ aucune donnée pertinente n'est exclue du calcul des provisions techniques sans justification. ▫ les données sont de granularité suffisante et incluent suffisamment d'historique pour évaluer au mieux les caractéristiques des risques sous-jacents ; ▫ les données remplissant la condition précédente sont disponibles pour chacun des groupes de risques utilisés dans le calcul des provisions techniques. Cela revient à s'assurer de la présence de toutes les variables sensibles à la construction des données d'entrée du modèle qui sont nécessaires aux calculs (i.e. contrôle sur le nombre de colonnes) ; ▫ le portefeuille assuré est reflété de façon exhaustive par l'ensemble des données (i.e. contrôle sur le nombre de lignes). Il n'y a pas de valeurs manquantes et des statistiques en montant et en nombre (par exemple, le chiffre d'affaires ou encore les montants garantis) sont définies par secteur d'activité et comparées à des statistiques de portefeuille pour répondre à ce besoin de contrôle.

Approprié	Reflet fidèle des engagements vis-à-vis des assurés ▫ les données sont cohérentes avec les finalités pour lesquelles elles seront utilisées ; ▫ le volume et la nature des données sont suffisants pour limiter l'erreur d'estimation dans le calcul des provisions techniques ; ▫ les données sont cohérentes avec les hypothèses sous-jacentes aux techniques actuarielles et statistiques utilisées dans les modèles ; ▫ les données reflètent les risques auxquels l'assureur est exposé à l'égard de ses engagements d'assurance ; ▫ les données ont été collectées, traitées et appliquées de manière transparente et structurée, et sont utilisées de manière cohérente dans la durée.
Exactitude	Degré élevé de confiance élevé porté par tous les intervenants ▫ l'origine et les étapes suivies par une donnée ainsi que la manière et la raison de l'évolution de cette dernière au fil du temps sont traçables ; ▫ les données liées à des périodes de temps différentes, utilisées pour la même estimation sont cohérentes. Cela signifie que les données anciennes ont la même qualité que les données récentes. Par exemple, les données sont enregistrées sans retard et de manière cohérente dans le temps ; ▫ les données ne contiennent pas d'erreurs matérielles (valeurs aberrantes et valeurs manquantes), par rapport aux seuils de tolérance définis par l'entreprise.

Cependant, il convient de remarquer que les moyens et les indicateurs mis en œuvre pour répondre à ces trois critères sont parfois insuffisants : ils sont la plupart du temps qualitatifs ou dépendent d'un seuil fixé par l'entreprise, et demeurent univariés. Aussi, ils ne prennent pas en compte la cohérence interne des bases de données utilisées.

C'est pourquoi, malgré les divers contrôles relatifs à la qualité des données, l'impact de l'utilisation de bases de données comportant des anomalies, dans la construction des hypothèses et dans l'utilisation du modèle de projection, sur le niveau d'erreur global du résultat est très peu mesuré.

II.2 Présentation des méthodes de détection d'anomalies

Nous allons utiliser une grande variété de méthodes de détection d'anomalies. Aussi il est nécessaire de présenter les solutions existantes, en particulier quelles méthodes de détection d'anomalies peuvent s'appliquer à nos jeux de données et à notre objectif.

II.2.0.1 Principales difficultés

Avant de chercher à identifier les individus mal saisis dans nos jeux de données, réfléchissons à ce que nous considérons comme une anomalie. Il n'y a pas de définition mathématique fixe de ce qu'est une anomalie.

Aussi, on peut définir, de façon abstraite, une anomalie comme étant une observation éloignée du reste des observations. Il s'agit principalement de données qui sortent de l'ordinaire, qui ne sont pas conforme à ce qui est attendu. Néanmoins, les cas d'usages sont très divers (par exemple des fluctuations de cours de bourse sont à interpréter différemment des fluctuations de température du corps) et il ne serait pas très pertinent de définir une métrique globale.

Une approche simple de détection d'anomalies consisterait à délimiter un espace représentant les données normales afin que toutes les observations n'appartenant pas à cet espace soient considérées comme des anomalies. Néanmoins, cette approche est difficile à mettre en œuvre pour les raisons suivantes :

- La frontière entre une donnée normale et une donnée anormale est difficile à définir : l'espace des données normales doit exclure toute la diversité des observations qui ne sont pas des anomalies ;
- Cette délimitation est difficilement fixe, certaines anomalies, situées à la frontière, peuvent se révéler normales et vice-versa. En effet, les données contiennent souvent du « bruit » qu'il est difficile à retirer et estompant la distinction entre les données fiables et les données aberrantes ;
- Les données aberrantes sont présentes en petites quantités au sein d'un jeu de données et ne couvrent généralement pas tous les cas de figure possible ce qui rend leur prédiction très difficile ;
- Les caractéristiques d'une donnée normale ne sont pas nécessairement stables dans le temps. Aussi, il arrive qu'une donnée correcte ne soit pas assez représentée dans les futurs jeux de données.
- Lorsque des anomalies sont le résultat d'actions malveillantes, les responsables s'adaptent souvent pour faire apparaître les observations anormales comme normales, rendant ainsi plus difficile leur détection. C'est le cas du blanchiment d'argent, des cyberattaques et des activités terroristes.

En raison des défis cités ci-dessus, le problème de détection des anomalies, dans sa forme la plus générale, n'est pas facile à résoudre.

II.2.1 Typologie et revue des méthodes de détection d'anomalie

La plupart des techniques de détection d'anomalies ont été introduites pour répondre à un problème précis de détection d'anomalies. Ces techniques reposent sur une grande variété de disciplines telles que les statistiques, l'apprentissage automatique, l'exploration de données, la théorie de l'information et la théorie spectrale. Pour cette étude nous proposons une méthode permettant rapidement d'isoler les corrections à apporter en détectant des individus d'un portefeuille que l'on peut considérer comme anormal. Nous allons donc, dans cette partie, tenter de répondre à la question suivante : Quelles méthodes utiliser pour déterminer si des points de données sont normaux ou anormaux ?

Cette partie théorique sur la détection d'anomalie s'inspire principalement de l'article CHANDOLA et *al.* [2009] [23] dans lequel les auteurs proposent une classification multicritère des méthodes de détection existantes dans la littérature et précisent que les méthodes à utiliser dépendent de la nature des données fournies, de la dimension des données, du type d'anomalies que l'on cherche à identifier, de la disponibilité de données identifiées comme « fiables » ou anormales, ainsi que du domaine d'application.

II.2.1.1 Nature des données

Un aspect clé de toute technique de détection d'anomalie est la nature des données d'entrée. Les données peuvent être de diverses natures (signaux, vidéos, séquences...) et plus ou moins structurées. Dans le cas multivarié, chaque point de données dispose de plusieurs attributs. Chaque observation peut ainsi être décrite à l'aide d'un ensemble d'attributs pouvant être de différents types : binaire, catégoriel ou continu. Tous les attributs peuvent être du même type ou peut être un mélange de différents types de données.

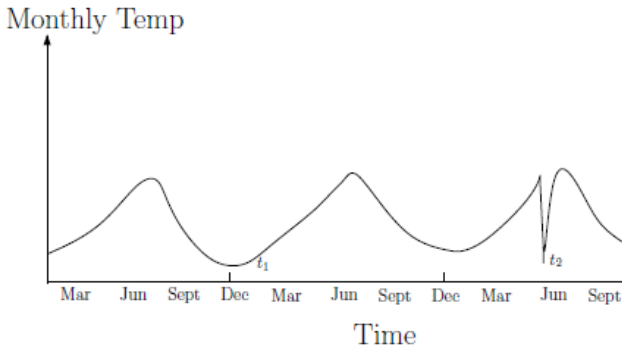
La nature des attributs détermine l'applicabilité des techniques de détection des anomalies. Par exemple, pour les techniques statistiques, des modèles différents sont utilisés pour les données continues et catégorielles. De même, pour les techniques basées sur le plus proche voisin, la nature des attributs détermine quelle mesure de distance utiliser.

Les données peuvent également être caractérisées au moyen de la relation qui lie les lignes entre elles. C'est le cas des données séquentielles, des données spatiales et des données graphiques. Néanmoins, la plupart des techniques de détection d'anomalies existantes traitent de données ponctuelles, dans lesquelles aucune relation n'est présumée entre les observations.

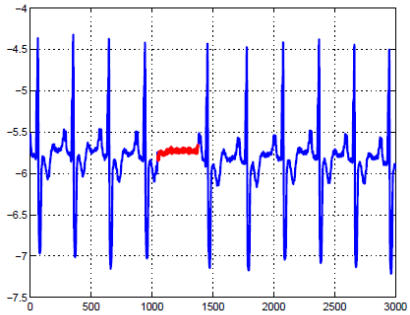
II.2.1.2 Description du type d'anomalie

Chandola, Banerjee et Kumar [23] précisent qu'un aspect important d'une technique de détection d'anomalie est la nature des anomalies que l'on espère détecter. En effet, Il est important de remarquer qu'une anomalie est toujours définie, implicitement ou explicitement, relativement à un modèle sous-jacent (c'est-à-dire une référence que l'on se donne). Cette référence peut dépendre des autres données observables ainsi que d'informations additionnelles comme le contexte par exemple. Aussi, les anomalies peuvent être classées en trois catégories :

- Anomalies ponctuelles : si un point de données est trop éloigné du reste, il tombe dans la catégorie des anomalies ponctuelles. Ce cas de figure concerne les données ponctuelles et fait l'objet de la majorité des recherches.
- Anomalies contextuelles : si l'événement est anormal pour des circonstances particulières, il s'agit anomalies contextuelles. À mesure que les données deviennent de plus en plus complexes, il est essentiel d'utiliser des méthodes de détection des anomalies propre à un contexte. Ce type d'anomalie est courant dans les données de séries chronologiques. Dans la figure ci-dessous, les températures t_1 et t_2 ont la même valeur mais se mesurent dans un contexte temporel différent. L'une est vraisemblablement une anomalie.
- Anomalies collectives : Si un ensemble de données est anormal par rapport à la totalité de données, cet ensemble est une anomalie collective. Ici, on ne cherche plus à identifier des anomalies mais des collections d'anomalies en comparant des ensembles de données au lieu de points individuels. Ce changement correspond à la détection d'anomalies et s'illustre dans la figure ci-dessous : la collection d'anomalies correspond ici à une rupture de rythme dans un électrocardiogramme. La région mise en surbrillance dénote une anomalie car la même valeur basse existe pour une durée anormalement longue (correspondant à une contraction auriculaire prématurée). Néanmoins les points appartenant à l'ensemble anormal ne sont pas pour autant des anomalies.



Série temporelle de températures mensuelles



Electrocardiogramme humain

FIGURE II.4: Illustration anomalies contextuelles et collectives

Le but de cette partie est de mesurer la part du *Best Estimate* que l'on peut associer aux données fiables. Plus précisément, il s'agira d'identifier les individus anormaux par rapport à l'ensemble des têtes d'un portefeuille. Ainsi, dans le cadre de cette étude, nous détectons uniquement les anomalies ponctuelles, c'est-à-dire une observation anormale au sens de différente des autres informations du jeu de données.

II.2.1.3 Disponibilité de données étiquetées

Déterminer si une observation est ou n'est pas une donnée aberrante est un exercice très subjectif. Il existe cependant des méthodes variées pour la détection des données aberrantes. Ces méthodes dépendent avant tout d'une information cruciale : la présence de données « étiquetées », c'est-à-dire déjà identifiées comme normale ou anormale. Selon de degré d'informations disponibles, les techniques de détection d'anomalies peuvent fonctionner d'après l'un des trois types d'apprentissage suivants :

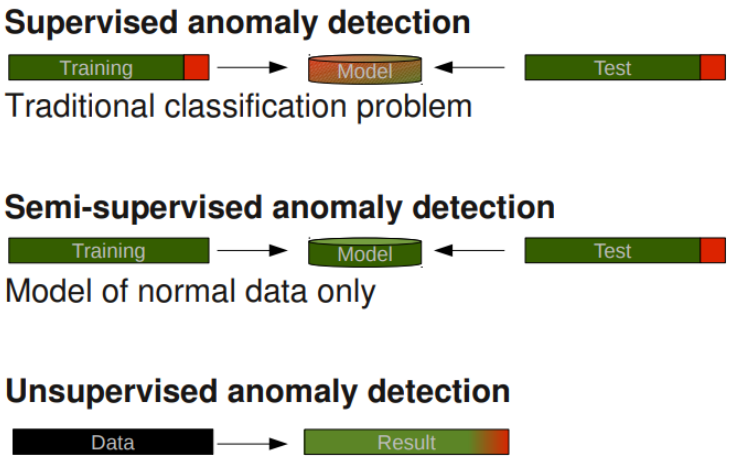


FIGURE II.5: Distinction entre les types d'apprentissage

- **Supervisé :** une des variables précise ce qui est une anomalie et de ce qui n'en est pas. Dans ce cas, toutes les données utilisées sont étiquetées. On suppose alors que les anomalies sont rares et qu'une nouvelle anomalie est similaire à celles déjà identifiées. La méthode de détection de données aberrantes est ici semblable à la construction de modèles prédictifs reposant sur des méthodes de classification ou de régression. Par exemple, pour la détection de *spams*, à partir de larges bases de données d'e-mails, le modèle apprend alors ce qui est un *spam* et pourra le détecter à la réception. Néanmoins, il est généralement difficile d'obtenir des étiquettes précises et représentatives de tous les cas anormaux.

- **Semi supervisé :** les données utilisées pour l'entrainement des algorithmes ne contiennent pas d'anomalies. Ce type de techniques, ne nécessitant pas d'étiquettes pour la classe d'anomalies est plus largement applicable que les techniques supervisées. L'approche typique utilisée dans ces techniques est de construire un modèle pour la classe correspondant au comportement normal pour ensuite identifier les anomalies dans les données de test. Les anomalies sont donc détectées parmi les nouvelles observations. Il existe un certain nombre d'approches appelées, ''One Class Classification" (ou encore ''Novelty détection").

- **non supervisé :** les données ne sont pas étiquetées. Les différents algorithmes tentent de distinguer les observations aberrantes en apprenant sur l'ensemble des données disponibles. Cela consiste à isoler une donnée aberrante sans avoir d'apriori sur ce qu'est une anomalie. Les techniques de cette catégorie de méthodes sont les plus largement applicables. Elles supposent seulement que les cas normaux sont beaucoup plus fréquents que les anomalies et que ces dernières sont déterminables par rapport aux informations contenues dans l'ensemble du jeu de données.

Si cette hypothèse n'est pas vraie, de telles techniques souffrent d'un taux élevé de fausses alertes et il y a beaucoup moins de certitudes sur les performances des algorithmes.

Dans notre cas, nous souhaitons contrôler la qualité des données. Nous ne disposons pas d'échantillons d'anomalies ni de données d'entrainement totalement fiables et utiliserons des techniques qui fonctionnent en mode non supervisé. Aussi, on considère la détection des anomalies comme l'identification d'éléments, d'événements ou d'observations rares qui soulèvent des soupçons en s'écartant significativement de l'ensemble du jeu de données.

II.2.1.4 Typologie des méthodes de détection d'anomalies non supervisées

La sélection de l'algorithme de détection des anomalies à utiliser est une activité complexe comportant de multiples considérations : type d'anomalie, données disponibles, performances, consommation de mémoire, évolutivité et robustesse (M. Goldstein et S. Uchida 2016) [14].

Nous avons vu précédemment que ce choix dépend principalement des données disponibles et de leurs propriétés. Il faut par ailleurs tenir compte également de la volumétrie de données, nos cinq portefeuilles étant composés de plus de 200 000 têtes chacun.

Nous allons enfin présenter les méthodes possibles en fonction de l'approche sous-jacente adoptée par chaque technique. Cette classification s'inspire de l'article TOGBE et *al.* [2020] [10]. Pour la majorité des catégories, nous appliquerons une technique de détection d'anomalies que l'on détaillera.

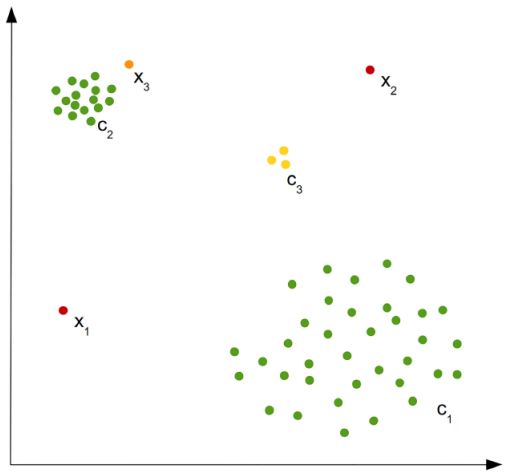
Les algorithmes non supervisés pouvant être utilisés sont basés sur les approches suivantes :

- **Méthodes statistiques** Les techniques statistiques peuvent être paramétriques ou non paramétriques. A la différence de l'approche non paramétrique, l'approche paramétrique suppose une connaissance a priori de la distribution des données. Dans le cadre des données multivariées, ces techniques ne permettent pas d'utiliser toute l'information.

- **Méthodes basées sur les plus proches voisins** Cette technique, basée sur la proximité entre les données, repose sur le calcul du voisinage de chaque observation. L'hypothèse clé de cette technique est que les points anormaux sont isolés des autres points. Ce type de technique détermine, pour une observation o, ses k plus proches voisins à travers le calcul de la distance entre toutes les observations du jeu de données.

- **Méthodes basées sur le *clustering*** Les techniques de *clustering* ont pour objectif principal de diviser le jeu de données en *clusters* contenant les données qui ont des comportements similaires. Les données qui appartiennent à de petits *clusters* ou à aucun *cluster* sont considérées comme des anomalies.

Remarque : distinction entre les méthodes basées sur les distances et celles basées sur les densités. Parmi les techniques basées sur la proximité, on distingue les techniques basées sur les plus proches voisins de celles basées sur le *clustering*. Cependant, une autre distinction est pertinente. En effet, alors que les méthodes basées sur la distance identifient les points éloignés de la majorité des autres points, les méthodes basées sur les densités partent du principe que la densité d'observations situées à proximité d'une anomalie est sensiblement inférieure à la densité d'observations à proximité d'un point normal.
Ces techniques sont parfois préférées aux méthodes basées sur la distance lorsque le jeu de données est composé de sous-ensembles de densités variables. Ce cas de figure correspond à l'illustration suivante :



On observe que x_1 et x_2 sont très différentes des autres *clusters* par rapport à leurs attributs et sont donc des anomalies **globales**. En regardant globalement l'ensemble de données, x_3 peut être considéré comme un point normal car il n'est pas trop loin du *cluster* c_2.

Cependant x_3 peut aussi être une anomalie locale : quand on se concentre uniquement sur le *cluster* c_2 et qu'on le compare à x_3 en omettant les autres points. Ce point est seulement anormal par rapport à son voisinage proche.

FIGURE II.7: Jeu de données bidimensionnelles

- **Les techniques basées sur le *Deep Learning*** Elles représentent une classe d'algorithmes d'apprentissage automatique supervisés ou non supervisés basés sur l'utilisation de plusieurs couches d'unité de traitement non-linéaire. Parmi ces méthodes on cite les auto-encoders (AE) et One-Class Neural Networks (OCNN).

- **Autres** D'autres techniques existent utilisant des approches variées comme celles basées sur les arbres de décision, les réseaux de neurones, les méthodes adaptées aux grandes dimensions par construction de sous-espaces ou par réduction de dimensions.

Dans la figure suivante, nous présentons une synthèse de cette classification basée sur l'approche sous-jacente ainsi que les algorithmes choisis, entourés en rouge :

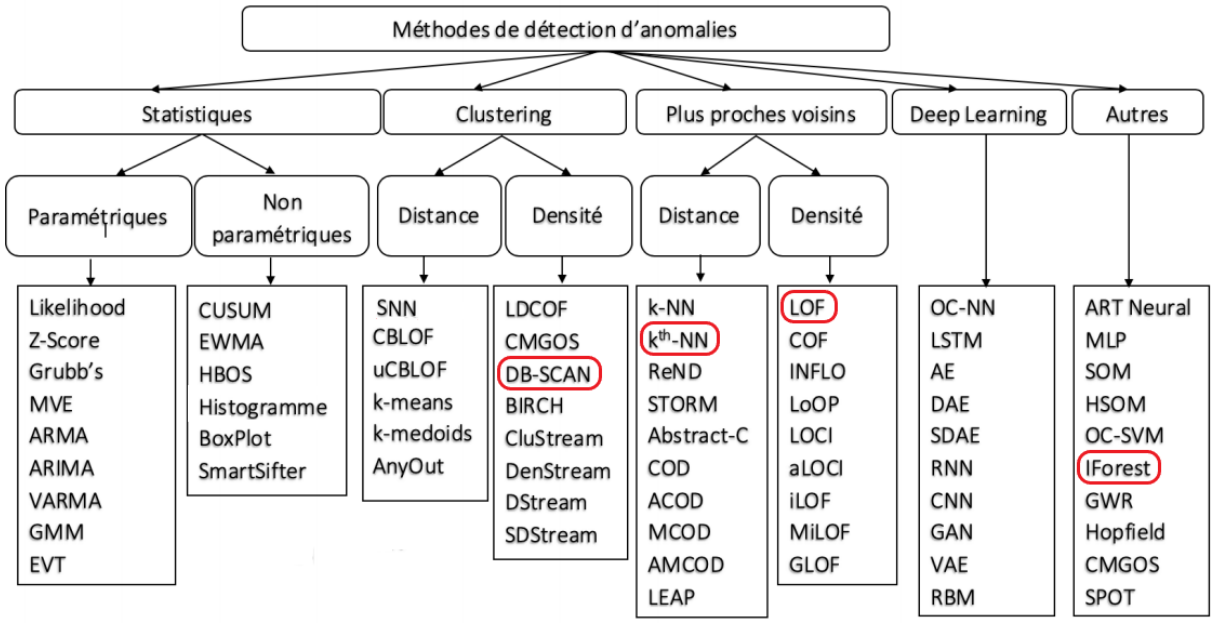


FIGURE II.8: Classification des différentes techniques de détection d'anomalies et techniques utilisées

II.2.1.5 Algorithmes de détection utilisés

Ainsi, les méthodes possibles se basent sur les différentes approches comme la classification, le *clustering*, la régression, les plus proches voisins et les outils statistiques.

Nous nous focalisons donc sur les algorithmes suivants, appartenant à la plupart des différentes catégories : KNN, LOF, $DBSCAN$ et IF. Ces algorithmes seront détaillés dans la section suivante. Ils seront ensuite testés sur cinq jeux de données différents dans le cadre de notre étude (partie application).

Nous avons fait le choix de comparer expérimentalement cinq méthodes de détection d'anomalies très utilisées dans la littérature : KNN, LOF, DBSCAN et Isolation Forest. Il s'agit de méthodes performantes compte tenu de la forte volumétrie des données. Nous commençons par décrire chacune de ces méthodes puis nous présentons les jeux de données utilisés ainsi que les résultats obtenus.

Remarque : l'objectif de cette étude est de détecter les individus aberrants sans aller trop loin dans le contexte théorique. Le lecteur intéressé par des descriptions détaillées ainsi que par des applications pratiques de l'apprentissage automatique pourra se référer aux références citées en bibliographie.

II.2.2 KNN : les K plus proches voisins

L'algorithme des k-voisins les plus proches [16] («k-nearest neighbors» ou KNN) est basé sur la distance entre les points. Initialement utilisé pour la classification, il s'utilise pour détection de valeurs aberrantes. Pour cette approche, on suppose que les anomalies ont tendance à s'écarter de la majorité des autres points. En ce sens, il s'agit d'une approche globale.

Algorithme : cette technique consiste à directement calculer un score pour chaque point. Ce score correspond à la distance de chaque observation par rapport au reste des observations. En effet, pour chaque point, on mesure la distance au k-ème plus proche voisin. Il est aussi possible de choisir de faire la moyenne des k distances pour calculer le score.

Ensuite, il suffit d'ordonner les observations et de sélectionner les observations ayant une $k - distance$ grande par rapport autres. Les observations qui ont les $n\%$ plus grandes $k - distances$ sont des anomalies, n étant un paramètre à fixer. Pour fixer n, on peut sélectionner un quantile correspondant à une quantité de points anormaux attendus :

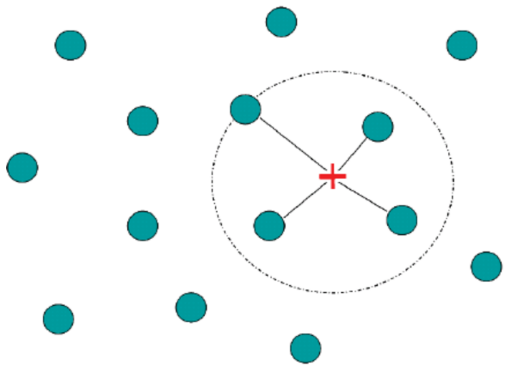


FIGURE II.9: **Illustration des k plus proches voisins**
exemple de calcul de k-distance du point en croix avec un k=4
source : "www.datasciencecentral.com"

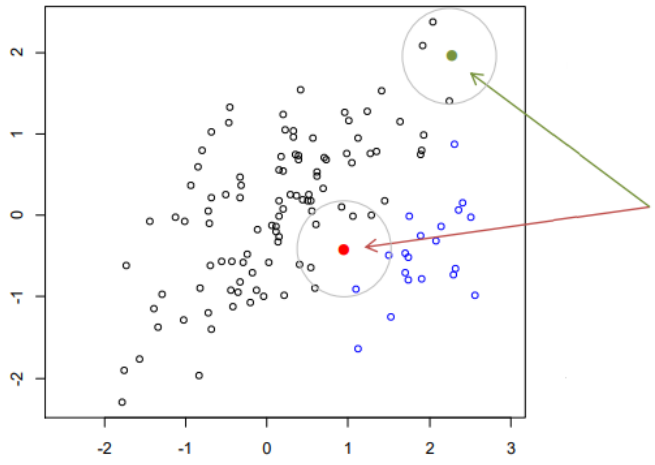
Avantages/Inconvénients :

- L'algorithme ne requiert que le paramètre k ce qui le rend simple et facile à mettre en œuvre. Il est polyvalent, pouvant être utilisé pour la classification, la régression et la détection d'anomalies.
- L'algorithme ralentit considérablement à mesure que le nombre d'observations et de variables augmente. En effet, l'algorithme parcourt l'ensemble des observations pour calculer chaque distance. De plus, cette approche est sensible à la valeur de k. Par exemple, si k est trop faible et s'il y a un *cluster* de plus de k points dont les points ont des petites distances par rapport à leurs voisins les plus proches, ces points seront tout de même considérés comme anormaux. De plus, la valeur du score est propre au jeu de données, aussi, il est parfois difficile de fixer un seuil à partir duquel les points sont des anomalies.

II.2.3 *LOF* : *Local Outlier Factor*

Le score de valeurs aberrantes locales (en anglais *Local Outlier Factor*) est un algorithme proposé par Markus et al. [2000] [6] permettant de trouver des points aberrants en mesurant l'écart local d'un point par rapport à ses voisins.

Principe : Cette méthode est également basée sur le plus proche voisin mais utilise la notion de densité. Elle provient initialement de la méthode *DBSCAN* qui sera présentée ci-après. L'hypothèse clé de cette approche est que les anomalies sont localisées dans les régions de faible densité. Plus précisément, on suppose que la densité autour d'un objet aberrant sera significativement différente de la densité autour de ses voisins. L'objectif de cette méthode est l'identification locale des points atypiques :



Dans une zone à forte densité, le point rouge qui s'écarte des autres (de ses voisins immédiats) devrait davantage interroger que le point vert situé dans une zone moins dense.

FIGURE II.11: Jeu de données bidimensionnelles

Formalisation : Notons $k-distance(A)$, la distance de l'objet A avec son k-ème plus proche voisin et l'ensemble des k voisins les plus proches $Nk(A)$. Cela nous permet de définir la "distance d'accessibilité" (ou reachability distance en anglais) :

$$reach - dist_k(A, B) = max\{k - distance(B), d(A, B)\}$$

En d'autres termes, la distance d'accessibilité d'un point A par rapport à B n'est pas uniquement la distance réelle entre ces deux points, mais au moins la distance k de B. Les objets qui appartiennent aux k voisins les plus proches de B sont considérés comme étant également distants. Ainsi, les points inclus dans le « rayon d'influence » d'ordre k de B sont considérés équivalents. L'utilisation de cette distance permet d'obtenir des résultats plus stables. Néanmoins, ce n'est pas une distance dans la définition mathématique, car elle n'est pas symétrique.

Remarque : la définition de la distance d'accessibilité justifie l'impact du paramètre k. Lorsque le point p est suffisamment loin du point central o (point p_2), alors la distance d'accessibilité entre les deux est leur distance réelle. A l'inverse, la distance d'accessibilité est alors égale la $k - distance$:

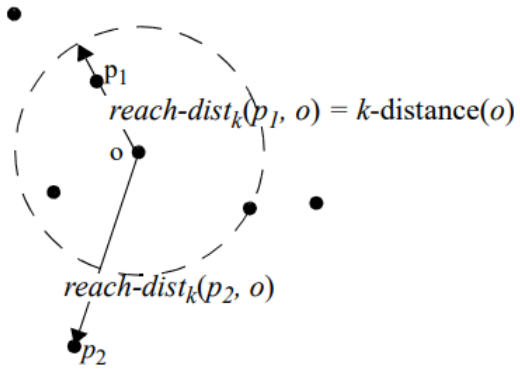


FIGURE II.12: distance d'accessibilité avec $k = 4$
avec pour le point o, p_2,le point suffisamment loin
et p_1 le point situé dans le voisinage (Breunig et *al.* [2000] [6])

Cette méthode réduit considérablement les fluctuations de statistiques de $reach - dist_k(p, o)$ pour tous les p proches de o. Cela correspond à un effet de lissage dont la force est contrôlée de la valeur de k. Plus la valeur de k est élevée, plus les distances d'accessibilité des objets dans la même zone sont similaires.

La distance d'accessibilité est ensuite utilisée pour définir la densité d'accessibilité locale. Pour un point A, elle est définie par l'équation suivante :

$$lrd_k(A) := 1/\left(\frac{\sum_{B \in N_k(A)} d_k(A, B)}{|N_k(A)|}\right)$$

où d_k est la distance d'accessibilité
où $|Nk(A)|$ est cardinal du $k - voisinage$ de A (vaut k s'il n'y a pas d'*ex aequo*)

C'est l'inverse de la distance moyenne d'accessibilité de l'objet A par rapport à ses voisins. Notons que ce n'est pas l'accessibilité moyenne des voisins de A (qui par définition serait la distance k(A)), mais la distance à laquelle A peut être "atteinte" de ses voisins [1].

Définition d'un point aberrant : les densités d'accessibilité locales sont ensuite comparées à celles des voisins en utilisant la densité d'accessibilité locale moyenne des voisins divisée par la densité d'accessibilité locale du point. Cela revient à calculer un score :

$$LOF_k(A) := \frac{\sum_{B \in N_k(A)} \frac{lrd_k(B)}{lrd_k(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} lrd_k(B)}{|N_k(A)|.lrd_k(A)}$$

1. Avec des points dupliqués, cette valeur peut devenir infinie.

On peut alors interpréter le score, correspondant à un ratio de densités locales, de la façon suivante :

- un score LOF inférieur à 1 indique que le point a une densité plus élevée que ses voisins ;
- un score LOF d'environ 1 indique que le point est comparable à ses voisins ;
- des valeurs significativement plus grandes que 1 indiquent des valeurs aberrantes.

Avantages/Inconvénients :

- A l'inverse de la méthode KNN, un point sera considéré comme une valeur aberrante s'il se trouve à une petite distance d'un sous-ensemble de très forte densité. En effet, la méthode LOF permet d'identifier les valeurs aberrantes locales. Par ailleurs, l'algorithme est simple, facile à mettre en œuvre et peut être appliqué dans n'importe quel contexte où une fonction de dissimilarité peut être définie.
- Au même titre que l'algorithme KNN, la technique est coûteuse en temps de calcul et en mémoire. Par ailleurs, les scores LOF sont des valeurs-quotient difficiles à interpréter dans l'absolu. En effet, une valeur de 1 correspond à une observation normale, mais une borne au-delà de laquelle une observation est atypique n'est pas explicite. Cela dépend du contexte et de la distance relative entre les points. Par exemple, dans un autre ensemble de données avec de fortes fluctuations locales, une valeur de 2 n'est pas forcément une anomalie. Le choix de la valeur de k est parfois difficile à fixer : si k est faible, le score est plus précis, mais les résultats sont instables ; si k est trop fort, il y a un risque de masquer les informations locales.

II.2.4 *DBSCAN* : *density-based spatial clustering of applications with noise*

L'algorithme DBSCAN (*density-based spatial clustering of applications with noise*) est une méthode de *clustering* proposée en 1996 (Martin et *al* [21]), utilisée pour former des *clusters* non convexes, comme des demi-lunes ou des cercles imbriqués les uns dans les autres. Pour cela, il se base sur la notion de densité entre les points.

L'hypothèse clé est que les observations normales appartiennent à des *clusters* de grande taille et denses, alors que les anomalies n'appartiennent à aucun *cluster* ou appartiennent à des petits *clusters* isolés.

Formalisation : l'algorithme DBSCAN utilise deux paramètres :

- Étant donné un nombre réel positif ϵ, on appelle l' ϵ-voisinage d'un point x l'ensemble des points du jeu de données dont la distance à x est inférieure à ϵ :

$$N_\epsilon(x) = \{u \in X \ |d(u,x) \leqslant \epsilon\}$$

- Étant donné un entier naturel n_{min}, on dit que x est un point cœur si son $\epsilon - voisinage$ contient au moins n_{min} points : $|N_\epsilon(x)| \geqslant n_{min}$

Pour des valeurs de ϵ et n_{min} fixées, nous sommes en mesure de catégoriser les points. Un point qui a au moins n_{min} points dans son voisinage de rayon ϵ est un point cœur. Si un point, dont le nombre de points dans son voisinage est inférieur à n_{min}, appartient au voisinage d'un point cœur, il s'agit d'un point frontière. Cela correspond au schéma suivant :

Si n_{min}=4,
le point orange n'est pas un point cœur car son $\epsilon - voisinage$ ne contient pas suffisamment de points.
Par contre, il appartient à l'ϵ -voisinage du point rouge, qui lui contient plus de n_{min} points.
Ainsi le point orange sera assigné au même *cluster* que le point rouge plutôt que d'être considéré comme du bruit.

FIGURE II.14: Type de points selon $DBSCAN$

La distance ϵ correspond au rayon minimal de voisinage et n_{min} représente le nombre minimum de points pour former un *cluster*. L'idée de base de l'algorithme est de déterminer pour un point donné son ϵ -voisinage et de vérifier qu'il contient bien n_{min} points ou plus. Ce point est alors considéré comme faisant partie définitivement d'un *cluster*. On parcourt ensuite l'$\epsilon - voisinage$ de proche en proche afin de trouver l'ensemble des points du *cluster*.

Cela nous amène à la définition de trois termes afin de mieux comprendre le fonctionnement de l'algorithme [2] :

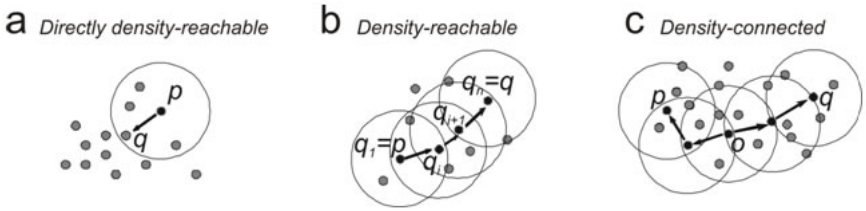


FIGURE II.15: Type de points selon DBSCAN

- **a** - *Direct density-reachable* : un point q est directement accessible par densité d'un autre point p si p est un point cœur et q est dans l'$\epsilon - voisinage$ de p ;
- **b** - *Density-reachable* : un point p est accessible par densité à q s'il existe une chaîne de points $q_1 \ldots q_n$, $q_1 = p$, $q_n = q$ tel que $q_i + 1$ est directement joignable par densité à partir de q_1 ;
- **c** - *Density-connected* : p et q sont connectés par densité s'il existait un point cœur O, pour lequel les deux points p et q sont accessibles par densité.

Algorithme : les *clusters* renvoyés par l'algorithme sont des "density-based clusters" : chaque *cluster* est un ensemble de points connectés. L'algorithme fonctionne de la manière suivante :

1. Pour chaque point x, calculer les distances aux autres points, puis trouver parmi ces points, les points voisins, situés à une distance inférieure à ϵ ;
2. Chaque point avec un nombre de voisins supérieur à n_{min} est marqué comme un point cœur ;
3. Pour chaque point cœur n'appartenant à aucun *cluster*, créer un nouveau *cluster* ;
4. Trouver tous les points connectés par densité au point cœur et les mettre dans le même *cluster* ;
5. Itérer sur le reste des points.

Un point initialement considéré comme du bruit par DBSCAN peut être ultérieurement rattaché à un autre *cluster* : il peut appartenir au voisinage d'un autre point, qui lui contient plus de n_{min} points.

2. source : http ://www.dmi.unict.it/ cassisi/DBStrata/help/methods.html

Définition d'un point aberrant : un point du jeu de données est dit aberrant si ce n'est pas un point cœur (son $\epsilon - voisinage$ ne contient pas n_{min} points ou plus) ni un point frontière (il n'est pas accessible par densité). Il s'agit d'un point qui n'appartient à aucun *cluster*.

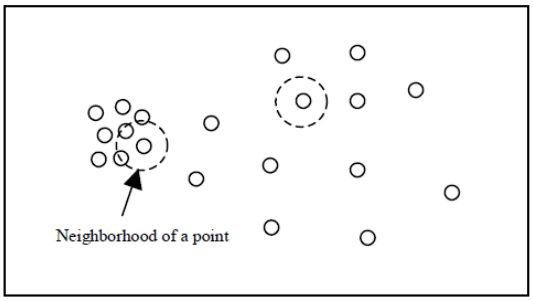
Avantages/Inconvénients :

- $DBSAN$ permet d'identifier des *clusters* de formes diverses. Cela revient à détecter au sein d'un jeu de données des associations et des structures qui sont difficiles à trouver manuellement, mais qui peuvent être pertinentes et utiles pour identifier des points anormaux. L'algorithme est assez simple à mettre en œuvre et ne requiert que deux paramètres. En opposition à la méthode des $k - means$, le nombre de *clusters* n'est pas défini à priori. De plus, l'algorithme peut être appliqué du moment qu'une fonction de dissimilarité est définissable.

- L'algorithme ne fournit pas de score, mais seulement une classification : il est possible que les anomalies soient situées dans des petits *clusters* ou des *clusters* de faibles densités. En grande dimension, cela perd sa significativité car les distances seront très similaires. Par ailleurs, la technique est coûteuse en temps de calcul et en mémoire.

  L'estimation des paramètres ϵ et n_{min} n'est pas un problème facile, car ces deux valeurs sont intrinsèquement liées à la topologie de l'espace à partitionner. Une trop faible valeur de ϵ et/ou une trop grande valeur de n_{min} peuvent empêcher l'algorithme de propager les *clusters*. A l'inverse, une trop grande valeur pour ϵ et/ou une trop faible valeur pour n_{min} peuvent conduire à renvoyer de fausses anomalies.

  Alors que $DBSCAN$ peut trouver des *clusters* de formes arbitraires, il ne peut pas gérer les données contenant des *clusters* de densité variable, car sa définition basée sur la densité ne peut pas identifier les points cœur des grappes de densité variable. En ajoutant un paramètre k prenant en compte les plus proches voisins partagés, l'algorithme SNN (Ertöz et al [2003] [8]) permet de trouver des *clusters* de différentes formes, tailles et densités. Il s'agit d'un algorithme de *clustering* basé sur la distance entre les points au lieu de la densité :



Pour n_{min}=4,
Si un utilisateur définit le voisinage d'un point en spécifiant un rayon particulier et cherche des points cœurs qui ont un nombre prédéfini de points dans ce rayon, alors soit le *cluster* dense sur la gauche sera identifié comme un *cluster*, et le reste sera marqué comme bruit, ou bien chaque point appartiendra à un seul *cluster*.

FIGURE II.17: Limite de $DBSCAN$, source : Ertöz et al [2003] [8]

II.2.5 *IF* : Isolation Forest

L'algorithme *Isolation Forest* (IF) (Liu et al. [2008] [13]) utilise une approche différente de la plupart des techniques de détection d'anomalies. Au lieu des mesures de distance et de densité couramment utilisées, IF isole explicitement les points anormaux par rapport à l'ensemble de données en se basant sur les arbres de décision et les forêts aléatoires. Cet algorithme repose sur l'hypothèse suivante : les anomalies, peu nombreuses et avec des caractéristiques différentes, sont plus sensibles à l'isolement que les données fiables.

Fonctionnement de l'algorithme : L'algorithme commence par créer des « arbres d'isolement » (*iTrees*). Il constitue un sous-échantillon de taille fixe et vérifie s'il peut être partitionné. Si ce n'est pas le cas, l'algorithme retourne la taille de l'échantillon et recommence. Ensuite, il sélectionne aléatoirement une caractéristique au sein du sous-échantillon, puis une valeur comprise entre les valeurs maximales et minimales de la caractéristique sélectionnée afin de séparer l'échantillon en deux. La construction de l'arbre est poursuivie en répétant l'opération précédente pour chaque variable jusqu'à l'obtention d'une observation par feuille :

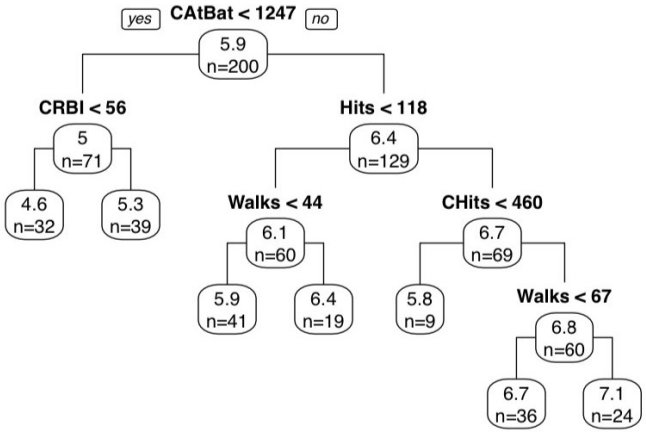


Figure II.18: Illustration de l'*isolation forest*

Enfin, afin d'éviter les problèmes dus au caractère aléatoire de l'algorithme, on répète la procédure en fonction du nombre d'arbres N spécifié [3].

Paramètres : L'algorithme consiste donc à construire N arbres d'isolation à partir d'un sous-échantillon aléatoire de taille n puis de mesurer la longueur du chemin pour chaque observation de l'arbre. Les principaux paramètres sont donc les suivants : le nombre d'arbres N, la taille des sous-échantillon $c(n)$ et le seuil de détection S à partir duquel x est une anomalie.

Définition d'un point aberrant : Considérons un jeu de données de n observations et une observation x, le score $S(x, n)$ d'aberrance de x est calculé par la formule suivante :

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

avec $h(x)$ la longueur du chemin entre la racine de l'arbre et x $\mathbb{E}(h(x))$ la moyenne des $h(x)$ de toute la forêt d'arbres et $c(n)$ la longueur du chemin moyen d'une observation depuis la racine dans le sous-échantillon de taille n du jeu de données.

3. En ce sens, IF n'est qu'une extension de *Random forest*, où l'objectif n'est plus la classification d'une observation mais l'identification des anomalies. De plus, alors que *Random Forest* travaille sur une méthode supervisée, *Isolation Forest* travaille sur une méthode non supervisée en raison de cette différence dans chacun de ses objectifs.

Le score d'une observation est donc la moyenne normalisée des longueurs des chemins depuis les racines de chacun des arbres. Il s'agit ensuite de fixex le seuil S à partir duquel l'observation x peut être considérée comme une anomalie. Plus il est facile de séparer rapidement l'observation de l'échantillon, plus cette dernière est isolée au plus près de la racine de l'arbre. Cela nous permet d'interpréter le score de la façon suivante :

- si le score $S(x, n)$ est proche de 1 alors x est très susceptible d'être une anomalie
- si $S(x, n)$ est inférieur à $0,5$, alors x est susceptible d'être une valeur normale

Ainsi, si, pour un échantillon donné, tous les points se voient attribuer un score d'anomalies d'environ 0.5, on peut supposer que l'échantillon ne comporte pas de données aberrantes. Le seuil S doit être supérieur à 0.5 et est choisi relativement au domaine d'application ainsi qu'aux différents scores mesurés.

Avantages/Inconvénients :

- En exploitant les techniques d'échantillonnage, IF est un modèle performant, nécessitant peu de mémoire, avec un petit nombre d'arbres. En effet, le mode de fonctionnement de l'*Isolation Forest* (pas de calcul de distance, pas de calcul de densité ni recherche de modèle) permet un traitement rapide même pour un jeu de données de grande taille ou de grande dimension. Comme il n'est pas nécessaire d'isoler tous les cas normaux, l'algorithme ignore la grande majorité de l'échantillon. La contribution de différents arbres indépendants dans la sélection des anomalies permet une très bonne précision.

  IF fait partie des méthodes de détection d'anomalies les plus récentes et les plus utilisées.

- L'un des principaux problèmes de l'application de l'*Isolation Forest* à la détection d'anomalies est la façon dont le « score d'anomalie » est calculé. Dans la figure ci-dessous, les zones avec un score faible en jaune ne correspondent pas aux données de formes sinusoïdales.

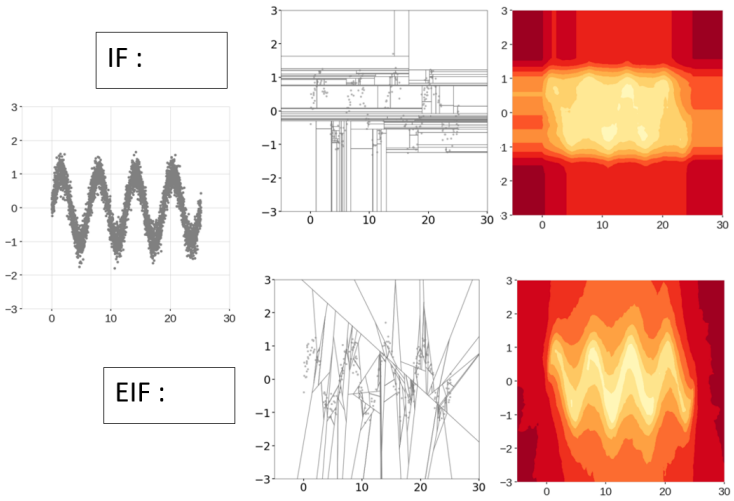


FIGURE II.19: Limites de l'IF [15]

Ce problème a été souligné par l'article de S. Hariri et al. [2018][15] dans lequel les auteurs proposent un modèle amélioré appelé *Extended Isolation Forest* (EIF) : au moment de la construction des arbres, au lieu de choisir une caractéristique aléatoire avec une valeur aléatoire, cette extension propose de sélectionner un vecteur normal aléatoire avec un point d'interception aléatoire.

Cette dernière remarque montre qu'il peut être intéressant de combiner ces algorithmes de détection afin d'identifier les anomalies locales et globales de façon exhaustive, au lieu de sélectionner une seule méthode de détection.

Nous utiliserons donc quatre algorithmes de détection d'anomalies. Par ailleurs, en raison des différentes limites expliquées, nous avons essayé d'utiliser d'autres méthodes telles que $GLOSH$ (*Global-Local Outlier Score from Hierarchies*), $HDBSCAN$ (*Hierachical density-based spatial clustering of applications with noise*), SNN (*Shared Nearest Neighbors*), EIF (*Extended Isolation Forest*). Néanmoins, les résultats n'ayant pas un apport suffisant par rapport aux quatre méthodes sélectionnées, celles-ci ne seront pas utilisées.

II.3 Application

Dans le cadre du calcul de la meilleure estimation, nous proposons de calculer un *Best Estimate* en utilisant exclusivement des données que l'on considère fiables, c'est-à-dire en enlevant les anomalies. Ces données fiables concernent aussi bien la projection du portefeuille d'assurés que la construction des hypothèses actuarielles.

Dans cette section, nous allons détailler comment nous avons appliqué les méthodes de détection sur les différents jeux de données. Pour cela, nous préciserons les retraitements mis en œuvre ainsi que le choix des paramètres de chacun des algorithmes. Nous allons ensuite présenter et interpréter les résultats afin de quantifier la part des provisions techniques que l'on peut attribuer à l'erreur générée par l'utilisation des données.

II.3.1 Caractéristiques des bases de données utilisées lors de la détection d'anomalies

Le but de cette partie est de quantifier l'erreur imputable à la qualité des données utilisées tant sur le portefeuille servant à la réalisation des projections que sur les portefeuilles servant à la construction d'hypothèses. Nous avons, au total, cinq bases de travail qui peuvent s'organiser de la façon suivante :

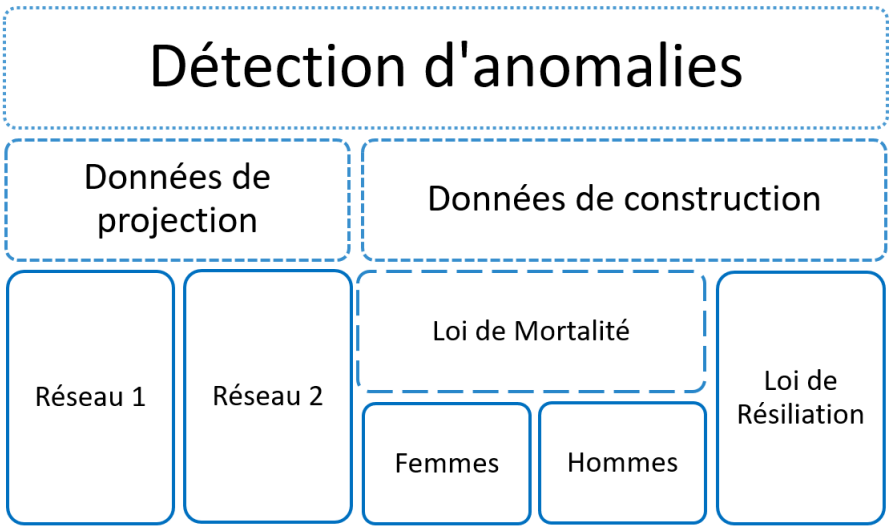


Figure II.20: Articulation entre les différentes bases de travail

En effet, les différentes bases de données se distinguent par leur utilisation, par le risque que l'on souhaite modéliser ainsi que le type de population représenté :

- Données de construction : utilisées dans la construction des hypothèses actuarielles servant au calcul du *Best Estimate* pour la garantie "décès toutes causes". Il s'agit d'une loi de mortalité par âge et par genre ainsi que d'une loi de résiliation par âge.
  - Données extraites le 30/09/2015 couvrant 5 années d'historique et servant à la construction d'une table de mortalité d'expérience. Cette loi est segmentée selon le genre :
    - Données de construction servant à la loi de mortalité pour les femmes ;
    - Données de construction servant à la loi de mortalité pour les hommes ;
  - Données extraites le 15/09/2015 et servant à la construction d'une table de résiliation.
- Données de projection : utilisées pour la réalisation des projections du portefeuille (MP, ou *Model Points*) représentant les contrats en cours pour la garantie "décès toutes causes" au 31/12/2018 dont il faut évaluer les engagements. Le *Best Estimate* étant segmenté par groupe homogène de risque, nous séparons les données de portefeuille selon le réseau de distribution :
  - Données de projection associées au réseau 1 ;
  - Données de projection associées au réseau 2.

Concernant les données de construction, nous nous servons des bases d'exposition utilisées juste avant la construction des taux bruts. En effet, les retraitements pré-construction ont déjà été réalisés : dédoublonnage par clients, suppression des lignes aberrantes et hors période d'observation, création de la variable de censure, des âges des assurés à l'entrée et à la sortie de la période d'exposition ...

Remarque : différence entre base de construction et base de projection. Au lieu d'être construite à partir d'un historique de plusieurs années, les bases de projection sont anonymes et représentent l'ensemble des contrats en cours à une date fixe, c'est-à-dire des contrats non résiliés et dont les garanties sont encore effectives. Cependant, identifier les lignes anormales du portefeuille à projeter et les retirer n'est pas pertinent. En effet, dans le cadre de la politique de provisionnement, chaque tête du portefeuille doit être prise en compte dans le calcul du *Best Estimate*. Nous avons quand même mené la procédure de détection sur les bases de projections afin de vérifier la cohérence des portefeuilles projetés.

II.3.2 Utilisation d'un algorithme de réduction de dimensions

Afin d'illustrer les différents choix et méthodes utilisés, nous nous focaliserons sur une seule base de données : le portefeuille servant à la construction de la loi de mortalité de la population féminine concernant le risque "décès toutes causes".

Pour rappel, l'objectif est d'identifier, parmi les 175 220 individus représentés, les lignes que l'on peut considérer comme anormales. Dans notre cas, on suppose qu'il s'agit d'informations mal saisies dans des bases de données d'assurés et qui sont très différentes des autres données du portefeuille.

La première étape consiste à prendre des données dans un espace de grande dimension (une cinquantaine de colonnes dans notre cas), et à les remplacer par des données dans un espace de plus petite dimension. Dans la pratique, nous aurons donc deux vecteurs numériques représentant l'ensemble de la base de données. Cette étape nous permettra d'utiliser les divers algorithmes de détection d'anomalies de façon optimisée en matière de temps de calcul.

II.3.2.1 Nécessité et raison d'une réduction de dimensions

La visualisation des données peut souvent être un bon point de départ pour déceler les anomalies lorsqu'il s'agit d'une ou de deux variables. Cependant, lorsque l'on étend cette approche à des données de grande dimension, cela devient de plus en plus difficile. En effet, la base considérée contient 58 colonnes de type de données variées. Aussi, nous allons commencer par réduire notre base de données.

Par ailleurs, ce type de méthode apporte une solution au problème du fléau de la dimension (« Curse of dimensionality » en anglais, [4]). En effet, l'augmentation de la dimension a tendance à rendre les données plus éparses et donc à fausser les manières traditionnelles d'analyser les données.

En conservant un nombre élevé de dimensions, les points aberrants vont être de plus en plus difficiles à détecter. Dans ce cas, une valeur aberrante peut présenter un comportement normal du point de vue d'un attribut particulier, c'est-à-dire en ne la considérant que sous l'angle d'une seule dimension. En revanche, cette valeur peut s'avérer anormale lorsqu'on la considère en combinant plusieurs attributs, autrement dit en la considérant sous l'angle de plusieurs dimensions. La difficulté réside dans la diversité des configurations au sein d'un ensemble de données multivariées. En réduisant le nombre de dimensions, le nombre de combinaisons nécessaire pour détecter une anomalie diminue.

Ainsi, cette étape permet d'analyser les différentes bases de données de façon plus lisible mais surtout plus efficacement.

II.3.2.2 Retraitements préliminaires à l'application de l'algorithme de réduction de dimensions

L'objectif est donc de représenter les données du portefeuille sur un espace à faible dimension, tout en préservant au mieux la structure du jeu de données. Afin d'utiliser l'algorithme de réduction de dimensions, il est nécessaire d'utiliser des données auxquelles on peut associer une matrice de distance, ce qui correspond idéalement à des données de même nature. Or, les données utilisées sont de natures variées : dates, variables catégorielles, identifiants... Par exemple, la base d'étude pour la mortalité féminine contient les informations suivantes :

Type	*Exemple*
Quantités	Montant de primes annuelles Montant de prestations à régler Age limite de couverture
Catégories	Numéro de garantie élémentaire Numéro de catégorie socio-professionnelle Catégorie associée à l'état du dossier
Dates	Date d'effet du contrat Date de clôture du contrat si elle existe Date de déclaration du décès Date de naissance

Aussi, pour les diverses bases de données, les dates ont été converties en âge et en ancienneté (nombre de mois entre la date d'effet du contrat et la survenance du sinistre par exemple). De même, les variables catégorielles ont été encodées en *dummy variables* : les modalités des variables quantitatives sont éclatées en variables binaires, c'est-à-dire qu'elle ne prend que la valeur 0 ou 1 pour indiquer l'absence ou la présence d'une modalité de référence. Pour les variables catégorielles prenant un trop grand nombre de valeurs différentes, telles que la catégorie socio-professionnelle, nous les avons regroupés dans des catégories plus larges en fonction des informations disponibles.

De plus, puisque la détection d'anomalies ponctuelles consiste à isoler des individus par rapport à la totalité du portefeuille, nous avons supprimé les informations redondantes telles que certaines dates obtenues à partir d'autres variables. De même, nous avons enlevé les colonnes avec une unique valeur ou des valeurs uniques à chaque ligne, car celles-ci n'apportent aucune information supplémentaire permettant de distinguer les têtes entre elles.

II.3.2.3 Insuffisance d'une réduction linéaire

Lorsque les variables sont quantitatives, l'algorithme de réduction de dimensions linéaire le plus utilisé est l'"analyse en composantes principales" (APC). Cette technique consiste à effectuer une cartographie linéaire des données au sein d'un espace de dimension inférieure, de sorte que la variance des données réduites soit maximisée.

En pratique, la matrice de covariance des données est construite et les vecteurs propres de cette matrice sont calculés. Les vecteurs propres sont ensuite utilisés pour reconstruire une grande fraction de la variance des données originales ce qui nous donne la représentation suivante en conservant les deux premiers axes :

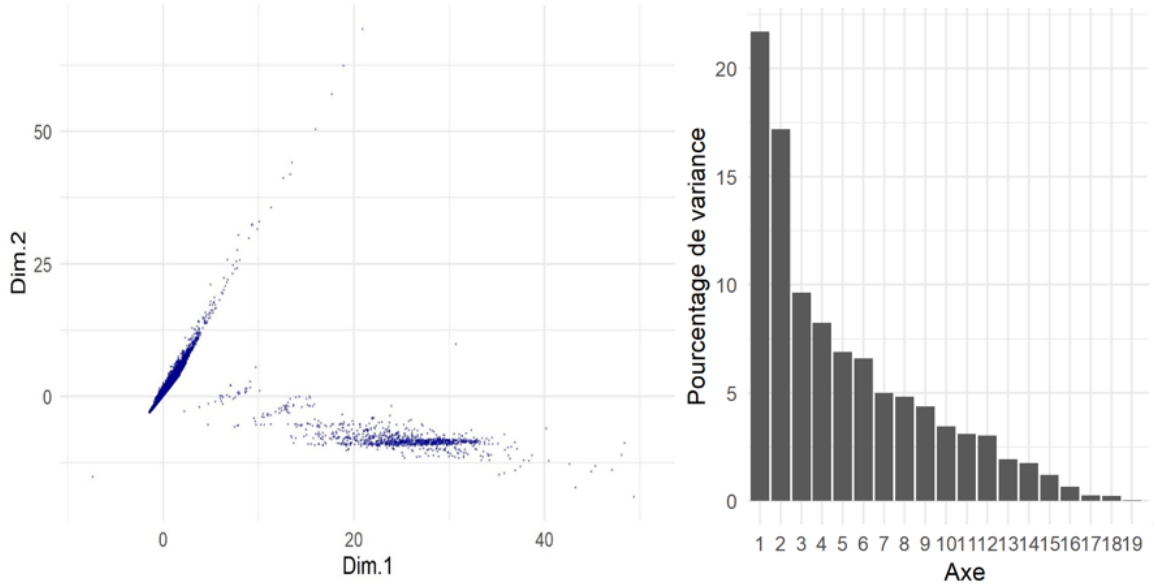


Figure II.21: Illustration de la réduction de dimensions par ACP

D'un point de vue théorique, l'ACP semble moins pertinente que la méthode expliquée ci-après puisqu'elle a pour but de réaliser une base orthogonale dans laquelle on projette nos données. Cette orthogonalité est mal adaptée quand les données à analyser vivent dans un sous-espace présentant une géométrie non-linéaire. C'est dans ce contexte que nous allons présenter une méthode de réduction de dimensions non-linéaire : **la méthode UMAP**.

II.3.2.4 Présentation de la méthode de réduction de dimensions utilisée

La méthode UMAP (*Uniform Manifold Approximation and Projection for Dimension Reduction*, McInnes et al. [2018] [9]) est une méthode de réduction de dimensions plus récente que l'ACP qui vise à préserver les distances faibles dans la projection. La méthode UMAP présente certains avantages par rapport à cette dernière notamment en termes de rapidité et de préservation de la position globale des groupes dans le résultat final.

Remarque : la méthode UMAP mobilise des notions mathématiques assez complexes, notamment d'analyse topologique. Dans le cadre de ce mémoire, nous nous contentons d'expliquer les principales étapes de l'algorithme.

Description simplifiée de l'algorithme :

- pour chaque point, on considère les n plus proches voisins. Le rayon de la sphère contenant ces n voisins est utilisé pour définir la base d'une "distance locale".
- cette "distance locale" est ensuite utilisée pour attribuer un "poids" (une probabilité) à la liaison entre le point considéré et chacun de ses n voisins les plus proches. L'algorithme part du principe qu'un point ne peut être totalement isolé, on a donc une probabilité certaine de lien entre un point donné et son plus proche voisin.
- ces ensembles de poids sont utilisés pour calculer un graphe pondéré : chaque noeud est un point, chaque arête est une liaison dont le poids est une "probabilité" de lien.
- dans le but de trouver la meilleure représentation possible en deux dimensions, on applique enfin à ce graphe pondéré un algorithme de *force-based layout*. Cet algorithme permet de positionner les nœuds d'un graphe, en se basant sur un système de forces appliqué entre les nœuds et les arcs, afin d'en faciliter sa visualisation.

Remarque : de nombreuses subtilités reposent sur la manière dont chaque arête est pondérée en fonction de la distance entre les points, ainsi que sur la façon dont les différentes distances sont "réconciliées" pour obtenir la pondération finale.

II.3.3 Mise en œuvre de la réduction de dimensions

Le temps de calcul n'est que d'une quinzaine de minutes pour une base d'un million de lignes et d'une cinquantaine de colonnes. On représente le portefeuille assuré servant à la détermination des taux bruts de mortalité pour les femmes pour le risque "décès toutes causes" dans un espace à deux dimensions.

Principaux paramètres La fonction *umap* du package *uwot* du logiciel R propose plus d'une trentaine de paramètres, dont certains influençant grandement les résultats :

- $n_{components}$: le nombre de dimensions du résultat final. Celui-ci sera fixé à 2 pour être visualisé ;
- $n_{neighbors}$: le nombre de plus proches voisins considérés dans le calcul du graphe des proximités. C'est le paramètre le plus important : avec une valeur élevée, le résultat donnera une vue plus globale de l'ensemble de points ; avec une valeur plus faible, les particularités locales des données seront mises en valeur. Cette valeur, généralement comprise entre 2 et 150, vaut 15 par défaut ;
- min_{dist} : distance minimale entre deux points. Visuellement, une valeur faible donnera des points beaucoup plus regroupés, une valeur élevée des points plus répartis (0.01 par défaut) ;
- *metric* : le type de distance utilisée.

Il est intéressant de remarquer que l'algorithme donne des résultats différents selon le type de distance utilisé :

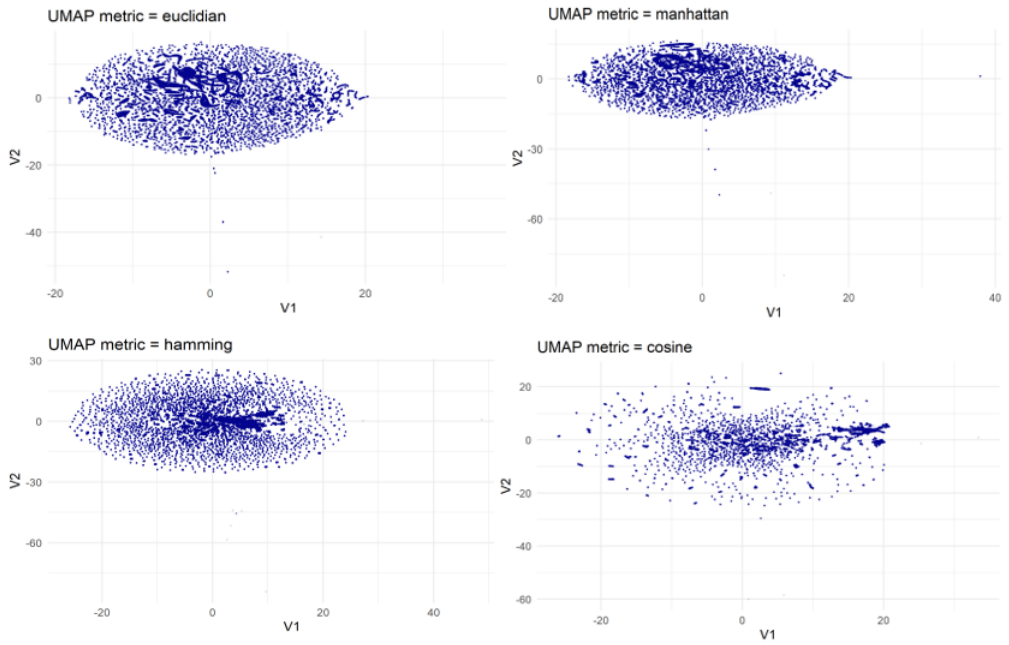


FIGURE II.22: Mesures de distances possibles pour UMAP

Quelle que soit la mesure de distance utilisée, on observe que le résultat, de forme elliptique, est visuellement très différent de celui de l'ACP : un grand groupe se distingue très nettement. A l'intérieur on peut repérer des "amas" plus petits et assez nombreux. Nous allons choisir la distance euclidienne qui est la distance la plus communément choisie dans le cadre de données numériques. Par ailleurs, les résultats sont assez différents selon le choix des paramètres $n_{neighbors}$ et min_{dist} :

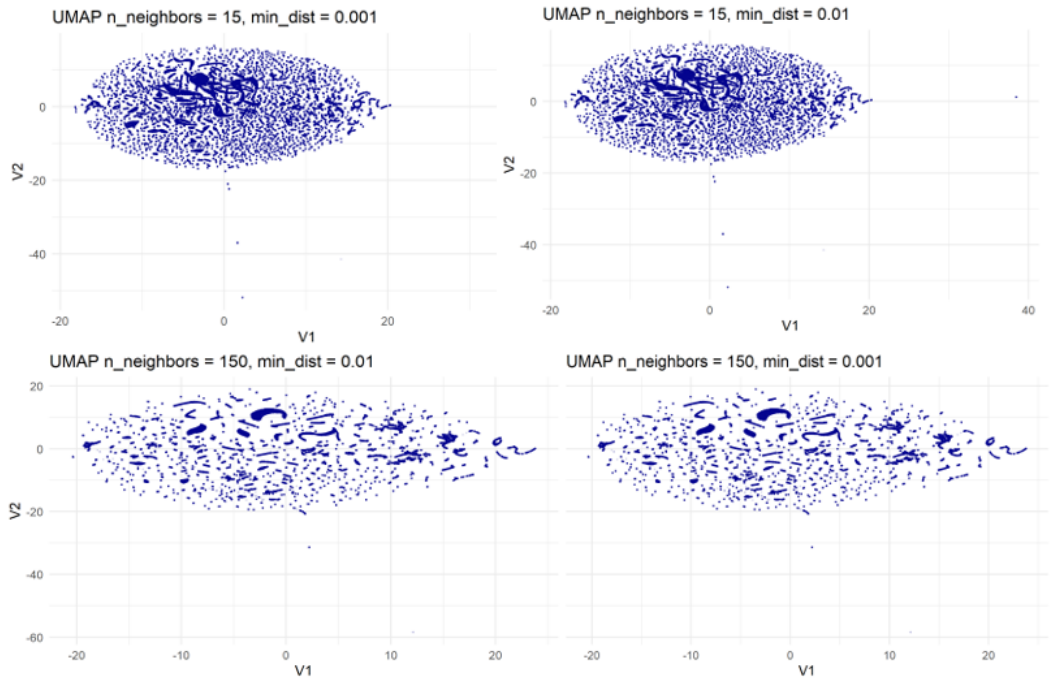


FIGURE II.23: Choix des paramètres de la réduction de dimensions

Quand on augmente la valeur de min_{dist} (graphiques de droite), les points sont moins regroupés mais certaines structures sont moins repérables. Quand on augmente le paramètre $n_{neighbors}$ (graphiques du bas), le temps de calcul augmente considérablement et on voit moins bien la structure globale au profit des structures locales. Nous retenons la représentation située en haut à droite. Aussi, pour toutes les bases de données à réduire, nous allons fixer $n_{neighbors} = 0.01$ et $min_{dist} = 15$.

II.3.4 Application des algorithmes de détection d'anomalies

Nous observons donc que certains des points de données sont des valeurs aberrantes évidentes, comme des valeurs extrêmement basses et extrêmement élevées. Il s'agit d'appliquer les méthodes de détection d'anomalies sur les données réduites. Cela permet de détecter plus efficacement les données anormales en combinant plusieurs attributs et en réduisant les temps de calcul des méthodes de détection.

II.3.4.1 Difficulté du paramétrage en détection non supervisé

Pour rappel, nous ne disposons pas d'échantillons d'anomalies ni de données d'entraînement totalement fiables et utilisons des techniques qui fonctionnent en mode non supervisé. Cela consiste à isoler une donnée aberrante sans avoir d'a priori sur ce qu'est une anomalie. Or, la littérature précise généralement uniquement que le choix des paramètres et/ou seuils de détections sont spécifiques au domaine d'application et aux données utilisées.

Aussi, le paramétrage et le choix du seuil de détection peuvent s'avérer délicats en l'absence d'avis d'expert. Dans notre cas, nous essayons de prendre en compte la volumétrie des données et les caractéristiques des points les plus isolés afin de fixer un seuil.

II.3.4.2 Sorties possibles des algorithmes de détection d'anomalies

La façon dont les anomalies sont ensuite identifiables dépend de l'algorithme de détection utilisé :

- L'algorithme peut nous donner un score, c'est-à-dire une grandeur numérique assignée à chaque point. Ce score représente une intensité d'aberrance, c'est-à-dire à quel point une donnée peut être considérée comme une anomalie. La sortie de ce genre de technique est donc une liste ordonnée et l'analyste peut choisir un seuil à partir duquel les points sont considérés comme des anomalies ou bien s'intéresser aux n points les plus anormaux.

  Ce type de sortie a l'avantage de permettre de sélectionner les anomalies les plus pertinentes tout en gardant en mémoire les caractéristiques de la base de données (distributions univariées, grandeurs numériques, nombre de lignes...). Cela laisse la liberté de choisir les anomalies tout en prenant le domaine d'application.

- Pour les autres techniques, dont le *clustering* fait partie, la sortie est une variable ordinale. Pour chaque point, une valeur non nulle désigne l'appartenance à un *cluster* et une valeur nulle désigne un point isolé.

  Néanmoins, ces techniques ne sont pas optimisées pour la détection d'anomalies puisque le but principal de l'algorithme est d'effectuer des regroupements. Ici, l'analyste ne peut pas directement sélectionner les anomalies les plus pertinentes. En effet, ce choix est fait indirectement par le choix des paramètres spécifiques à chaque algorithme. Ce choix est d'autant plus difficile en apprentissage non supervisé. Dans ce cas, on se contente de considérer les points isolés et peut choisir d'inclure les *clusters* de très petite taille, pouvant représenter un ensemble d'anomalies en comparaison des *clusters* les plus denses. L'enjeu est de savoir quel paramétrage et quel seuil choisir afin de détecter les anomalies. Il convient de prendre en compte le nombre de lignes de

la base en partant du principe que plus il y a de lignes, plus on est susceptible de détecter des anomalies.

II.3.4.3 Choix des paramètres de détection et des seuils de détections associés

Nous allons expliquer le choix des paramètres ainsi que les résultats obtenus avec la base de données servant à la construction de la loi de mortalité féminine. Il faut bien comprendre que cette étape est délicate, disposant de peu d'indications. Nous rappelons brièvement les méthodes de détection d'anomalies présentées dans la section précédente (II.2.1.5) :

Méthode de détection	Principe	Paramètres à fixer	Choix d'un seuil
KNN	Mesure la distance au $k-$ème plus proche voisin	k	oui
LOF	Compare la densité locale d'un point à ses voisins	k	oui
DBSCAN	Construit des *clusters* "connectés par densité"	k et n_{min}	non
IF	Construit N arbres d'isolation à partir d'un sous-échantillon aléatoire de taille n puis mesure la longueur moyenne du chemin de partitionnement pour chaque observation.	N et n	oui

Table II.1: Synthèse des méthodes de détection choisies

Pour chacun des algorithmes, nous avons déterminé les paramètres et/ou le score d'anomalies qui en résulte de la façon suivante :

LOF et KNN : Le paramètre commun à ces deux méthodes est le nombre de plus proches voisins partagés k. On choisit un seuil à partir duquel un point ayant un score supérieur est une anomalie. Pour en déterminer la valeur ainsi que le seuil associé, nous avons observé les différentes distributions des scores selon la valeur de k. Par exemple, observons la distribution extrême du score pour une valeur de $k = 15, 30, 90$ et 150 :

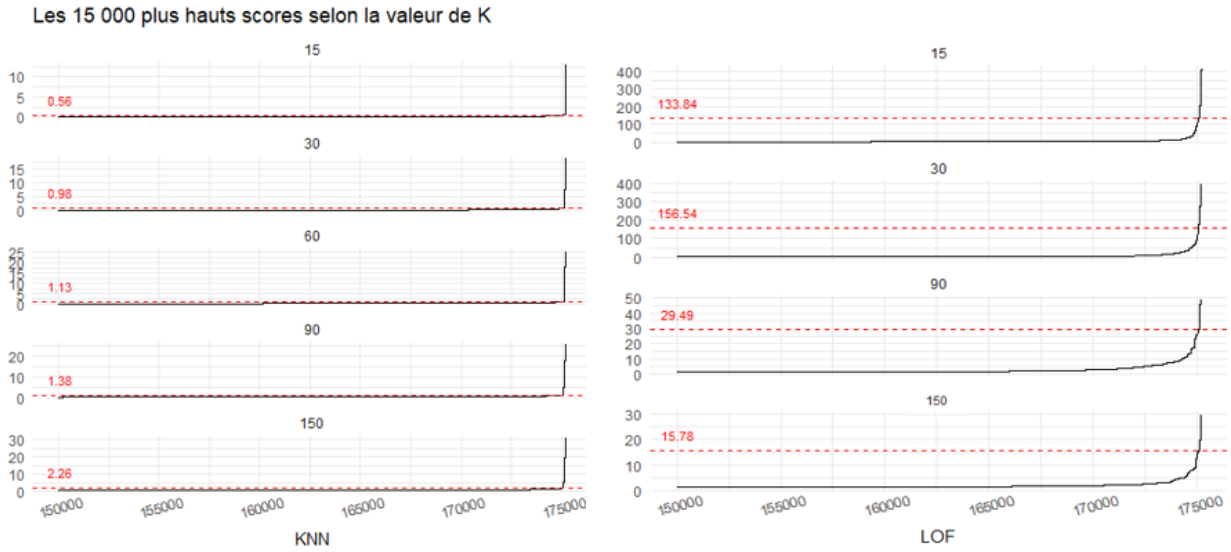


Figure II.24: Choix du seuil de détection d'anomalies pour les méthodes à un paramètre

Nous avons ajouté en rouge le score correspondant à une part de $0,05\%$ d'anomalies (soit 785 points) afin d'avoir une indication sur le nombre de valeurs détectées. Il s'agit de choisir un seuil en cohérence avec la distribution des scores pour une valeur de k fixée. Le choix de ce seuil consiste à déterminer un point "coude", c'est-à-dire une valeur du score correspondant à un fort redressement de la courbe (II.2.4). Aussi, pour une valeur de $k = 30$, nous choisissons un seuil de 156 pour la méthode LOF et $0,9$ pour la méthode KNN concernant la base de construction de la mortalité des femmes.

DBSCAN : Le choix des paramètres ϵ et n_{min} de cet algorithme est plus difficile à mettre en œuvre. Pour rappel, le voisinage de rayon ϵ est la zone dans laquelle les points sont considérés comme des voisins et n_{min} correspond au nombre de voisins à partir duquel un point peut être considéré comme un point cœur. De plus, un point du jeu de données est dit aberrant si ce n'est pas un "point cœur" (son $\epsilon - voisinage$ ne contient pas n_{min} points ou plus) ni un point frontière (il n'est pas accessible par densité).

En dimension deux, la littérature propose souvent de tracer les $k_{distances}$ de toutes les observations pour $k = n_{min} = 4$ [4]. Cela permet de déterminer ϵ grâce à la localisation d'un point frontière mais cela revenait à fixer une valeur trop forte pour ϵ si bien que l'algorithme classait tous les points dans aucun *cluster*. De même, quelle que soit la valeur de ϵ pour $n_{min} = 4$, l'algorithme ne détectait aucune anomalie. Aussi, compte tenu du nombre de lignes élevé, nous avons testé plusieurs combinaisons pour finalement prendre une valeur de n_{min} égale à 100 et sélectionner un ϵ égal à 1.

IF : Pour cette dernière méthode, le choix des paramètres N et n impacte uniquement la précision des résultats. Compte tenu de la taille de notre base, nous choisissons de fixer la taille du sous-échantillon à $N = 70079$ (soit 40% des lignes) et le nombre d'arbres à 200. La seule difficulté est donc de fixer un seuil. Le score représente la moyenne normalisée des longueurs des chemins depuis les racines de chacun des arbres : plus il est facile de séparer rapidement l'observation de l'échantillon, plus cette dernière est isolée au plus près de la racine de l'arbre. Après avoir observé la distribution des différents scores, nous fixons un seuil S de 0,75.

4. Ce que propose par défaut la commande *kNNdistplot()* du package R *dbscan*.

II.3.5 Résultats des méthodes de détection

Nous allons d'abord analyser les différentes anomalies détectées par les différentes méthodes. Afin de mieux visualiser le comportement de chacun des algorithmes, nous nous focaliserons sur la partie inférieure gauche de l'ellipse :

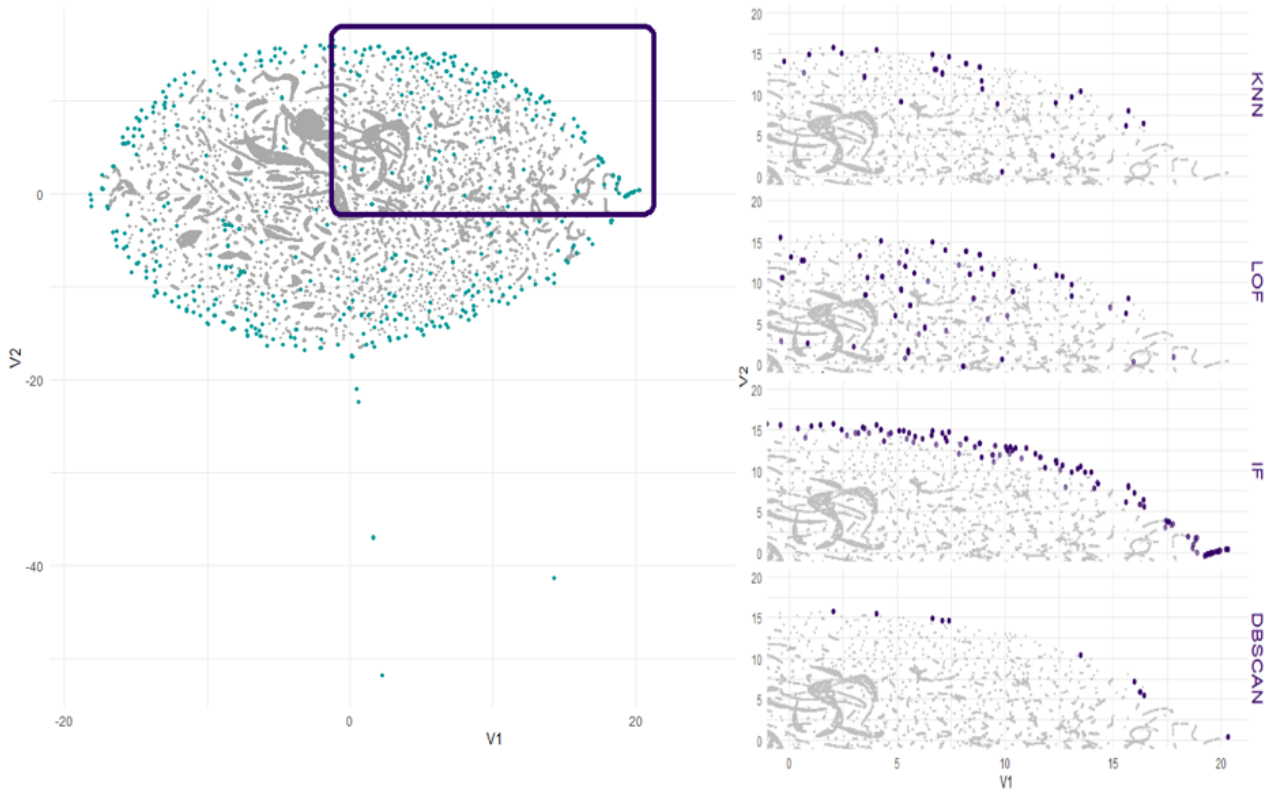


Figure II.25: Anomalies détectées en fonction de l'algorithme choisi

Les algorithmes renvoient des résultats différents :

- La méthode $DBSCAN$ est celle qui a détecté le moins d'anomalies et les points identifiés le sont déjà par d'autres méthodes. Cela est peut-être dû au choix des paramètres qui a été difficile à mettre en œuvre.
- La méthode IF une méthode globale au même titre que la méthode KNN, elle détecte surtout des points situés à l'extrémité de l'ellipse représentant l'ensemble des données.
- La méthode LOF reposant sur la notion de densité locale détecte davantage de points situés à l'intérieur de l'ellipse centrale.

Remarque : pour les quatre autres bases de données (les bases de construction des lois résiliation et de mortalité des hommes ainsi que les deux bases de projection), les illustrations des réductions de dimensions représentant les différentes anomalies détectées sont présentées en annexe.

II.3.5.1 Nombre d'anomalies détectées

Nous avons donc utilisé quatre méthodes de détection basées principalement sur du *clustering* ainsi que sur l'approche des plus proches voisins. Pour les cinq bases de données décrites au paragraphe II.3.1, le nombre d'anomalies détectées par méthode et pour chacune des bases est synthétisé dans le tableau ci-dessous :

Base de données	Décès Femmes	Décès Hommes	Rési -liation	Réseau 1	Réseau 2
Nombre de lignes	175 220	231 099	827 613	1 147 239	803 956
KNN	1358	979	2196	1805	1443
LOF	2097	2443	1023	193	27
DBSCAN	1011	557	3232	2858	1957
IF	1342	849	1565	548	215
Anomalies détectées par au moins une méthode	4230	3772	4815	3129	2071
Anomalies détectées par au moins deux méthodes	1200	720	2431	1810	1343
Anomalies détectées par au moins trois méthodes	410	288	684	359	210
Anomalies détectées par toutes les méthodes	79	48	86	106	18

On observe que très peu d'anomalies sont détectées par chacun des algorithmes ce qui est plutôt rassurant quant à la qualité des données utilisées. Par ailleurs, les quatre algorithmes de détection utilisés reposant sur des approches différentes, nous avons choisi de les combiner. Pour qu'un point soit détecté, il suffit qu'il soit identifié par une seule méthode de détection.

II.4 Conclusion partielle

Ainsi, dans le but de mesurer le niveau d'erreur associé au calcul du *Best Estimate*, nous avons commencé par proposer une méthode permettant d'identifier les lignes anormales dans chacune des bases utilisées. Cette méthode repose sur trois étapes successives :

- d'abord sur une standardisation des données pour avoir des données de type numérique ;
- ensuite sur une réduction de dimensions non-linéaire afin d'obtenir un vecteur de coordonnées ;
- enfin sur l'utilisation combinée d'algorithmes de détections.

De plus, cette procédure se base sur un apprentissage non supervisé et s'adapte bien aux données de grandes dimensions. Aussi, elle a l'avantage d'être générique et peut être utilisée dans bien d'autres contextes.

Pour la suite, nous avons pris la décision d'éliminer les observations anormales des bases de construction afin d'obtenir des bases de données les plus fidèles possible à la réalité. En effet, nous ne disposons pas assez d'avis d'expert pour pouvoir les corriger de façon pertinente. Aussi, nous allons utiliser les bases de données sans anomalies dans la suite de l'étude. Par exemple, nous allons visualiser les hypothèses de mortalité et de résiliation reconstruites à partir de bases sans anomalies et en quantifier l'impact sur le *Best Estimate*.

Pour les bases de projection, contrairement aux bases de construction des hypothèses, l'identification des lignes anormales ne sera pas utilisée par la suite. En effet, identifier les lignes anormales du portefeuille à projeter et les retirer n'est pas pertinent : dans le cadre de la politique de provisionnement, chaque tête du portefeuille doit être prise en compte dans le calcul du *Best Estimate*.

Dans ce cas, la détection d'anomalies nous a permis de vérifier la cohérence des portefeuilles projetés. Pour les deux bases de projection, on observe les lignes les plus anormales, c'est-à-dire représentées après réduction le plus loin de l'ellipse principale. Ces lignes correspondent aux individus les plus risqués. Il s'agit des individus ayant aussi souscrit la garantie "arrêt de travail" ainsi que les individus dont les montants de capitaux garantis en cas de décès sont les plus importants, et où les primes annuelles et les commissions sur primes sont les plus élevées.

Concernant les prolongements possibles de ce chapitre, il faut avoir en mémoire les limites des différents types d'apprentissage : pour l'apprentissage non supervisé, les méthodes sont difficiles à paramétrer en raison du manque de données d'entraînement tandis que pour l'apprentissage supervisé, il est difficile d'obtenir des données d'entraînement représentant les anomalies dans toute leur diversité. Or, il aurait été possible de mettre en place des méthodes de détection plus précises selon deux approches :

- Premièrement, concernant les données de projection, on pourrait implémenter une procédure de détection d'anomalies semi-supervisée. Il s'agit d'utiliser les différences entre les réseaux de distribution. En effet, le portefeuille du réseau 2 est plus récent et provient d'un système d'information moins bien contrôlé. Or, les hypothèses utilisées par le modèle de projection (dont les lois de mortalité et de résiliation) ont été construites à partir des données issues du réseau 1, correspondant au seul historique suffisant à disposition. Une méthode de détection possible serait de prendre comme données d'entraînement le portefeuille du réseau 1, auquel les anomalies ont été préalablement retirées, et de supposer que les deux portefeuilles sont comparables en termes de caractéristiques. Les anomalies du réseau 2 seront d'autant mieux détectées car considérées comme mal adaptées au calcul du *Best Estimate* qui est calibré à partir des données du portefeuille associé au réseau 1.

- Deuxièmement, il serait possible, en disposant d'un avis d'expert permettant d'identifier définitivement certaines anomalies, de mettre en place une méthode de détection plus précise. En effet, il est possible d'utiliser des méthodes de détection basées sur un apprentissage dit *Positive-unlabeled learning* (apprentissage positif-non étiqueté [5]) [Liu B, et al. (2003) [18]]. Ce type d'apprentissage consiste à supposer que l'on dispose déjà de têtes identifiées comme anormales, mais qu'il reste des têtes à identifier. Cette technique a l'avantage de mieux correspondre à la réalité. En effet, la fiabilisation des données normales est davantage chronophage que l'identification des données anormales, car ces dernières sont en nombre réduit.

5. Dans le sens où les données d'entraînement se composent d'un ensemble d'anomalies (étiquetage positif) et d'un ensemble non-étiqueté, contenant des anomalies et des données saines dans des proportions inconnues.

Chapitre III

Mesure d'erreurs relatives à la construction des hypothèses actuarielles

Introduction L'évaluation des engagements associés à un portefeuille d'assurance couvrant le risque "décès toutes causes" nécessite la mise en place d'une loi de mortalité et d'une loi de résiliation. Plus précisément, ces lois biométriques et comportementales sont des hypothèses utilisées par le modèle de projection multi-états servant au calcul des engagements. Ces deux hypothèses, l'une biométrique et l'autre comportementale, sont construites selon une approche non-paramétrique : les taux bruts sont obtenus à partir des informations provenant des données disponibles.

Dans le cadre de la réglementation Solvabilité II, ces hypothèses doivent être *Best Estimate*. Cela consiste à pouvoir déterminer les provisions techniques sur la base d'hypothèses et de paramètres pertinents. Par exemple, il s'agit de construire des tables d'expérience. Ces tables sont généralement construites à partir des données issues du portefeuille de l'assureur. Elles permettent de refléter au mieux les caractéristiques du portefeuille auquel on souhaite évaluer les engagements et ainsi d'appréhender le risque supporté de façon optimale.

L'objectif de cette partie est de mesurer quel niveau d'erreur de la meilleure estimation est générée par la construction et par l'utilisation des hypothèses actuarielles. Nous allons commencer par décrire les hypothèses utilisées ainsi que les sources d'incertitude que l'on peut associer à l'estimation d'une loi non-paramétrique par Kaplan-Meier. Nous allons ensuite proposer une méthode basée sur le ré-échantillonnage afin d'estimer ces différentes sources d'incertitude. Nous allons enfin détailler comment ces méthodes sont appliquées.

III.1 Contextualisation et présentation de la démarche

En prévoyance individuelle, les hypothèses utilisées lors de la valorisation des provisions techniques varient en fonction du type de modélisation et de la garantie considérée. Par exemple, la mortalité nécessite une loi de mortalité d'expérience ou à défaut règlementaire. Le comportement des assurés doit être pris en compte pour le calcul des provisions techniques. Aussi, la résiliation du contrat d'assurance nécessite une loi de résiliation afin d'estimer les primes perçues.

Nous allons commencer par décrire la notion de "meilleure estimation des hypothèses", le contexte de construction des deux hypothèses utilisées et ce que l'on entend par erreur d'estimation.

III.1.1 Contexte d'utilisation des hypothèses *Best estimate*

III.1.1.1 Définition d'une loi d'expérience

Afin de mieux prendre en compte le risque du portefeuille d'engagement, la réglementation autorise que les différentes hypothèses soient remplacées par des tables spécifiques dites « d'expérience ». Une loi d'expérience est construite à partir de données réelles. Les compagnies d'assurances construisent des tables d'expérience basées sur leur propre portefeuille afin de mieux appréhender le risque qu'elles couvrent.

La construction de table d'expérience consiste à comptabiliser l'occurrence d'un événement (décès, résiliation, survenance d'un arrêt de travail, reprise d'activité, etc.) dans une population donnée, en fonction des caractéristiques du portefeuille : âge, niveau de la couverture, genre... En pratique, l'obtention d'une table fiable est un exercice délicat, qui met en jeu plusieurs contraintes [11] :

- des aspects « biométriques » liés aux caractéristiques de la population sous risque ;
- les limitations et contraintes d'observation (franchises, résiliations, limitations dans la qualité des données disponibles) ;
- l'impact des couvertures sur le risque (sélection médicale, niveau et étendue des garanties).

Pour un risque tel que le décès, les lois d'expérience sont certifiées par un actuaire indépendant et agréé par l'une des associations d'actuaires reconnues par l'Autorité de contrôle des assurances et des mutuelles. Leur validité est limitée à 5 ans pour les lois de mortalité et 4 ans pour les lois de maintien. En l'absence de suivi, ces lois sont utilisables 2 ans seulement après leur certification.

III.1.1.2 Utilisation possible des hypothèses biométriques et comportementales

Il est nécessaire de construire les hypothèses actuarielles en cohérence avec la manière dont elles sont utilisées. En effet, il faut prendre en compte la finalité de l'utilisation de l'hypothèse. On distingue trois objectifs :

- tarification : les hypothèses sont utilisées pour évaluer la prime pure. Par exemple, la loi de mortalité permet de pondérer la fréquence de survenance des sinistres.
- provisionnement en norme comptable : les hypothèses permettent de quantifier les différentes provisions comptables (sinistres à payer, primes acquises et non émises...) et intègrent une marge de prudence ;
- provisionnement prudentiel : des hypothèses *Best Estimate* sont les hypothèses utilisées pour calculer les provisions techniques sur la base d'hypothèses et de paramètres pertinents. Le montant des provisions doit refléter les caractéristiques du portefeuille sous-jacent.

Remarque : étant donné que l'on projette un portefeuille existant en *run off*, c'est-à-dire sans tenir compte des renouvellements et nouvelles souscriptions, la tarification ne change pas. Le calcul du *Best Estimate* est totalement indépendant de la tarification du portefeuille.

III.1.1.3 Meilleure estimation des hypothèses

Le calcul des provisions dans Solvabilité II est effectué dans le cadre de la « meilleure estimation possible ». L'idée est d'utiliser des hypothèses de projection qui sont en adéquation avec les propriétés des contrats du portefeuille afin que les hypothèses caractérisent le risque intrinsèque au portefeuille. Cela implique la prise en compte de l'expérience du portefeuille dans la détermination des hypothèses.

Par ailleurs, sous la précédente directive de solvabilité et dans chacun des paramètres utilisés lors du calcul des provisions, les provisions techniques étaient déterminées en intégrant une marge de prudence. Dans le cadre de la valorisation du *Best Estimate*, l'objectif est d'estimer les provisions de la façon la plus exacte possible. Il s'agit donc de construire des hypothèses *Best Estimate*, c'est-à-dire généralement moyennes plutôt que volontairement prudentes [3].

C'est pourquoi, la détermination de lois d'expérience s'avère indispensable, que celle-ci soit déterminée complètement à l'aide des données disponibles ou calibrée par positionnement par rapport à une référence externe.

III.1.2 Caractéristiques des hypothèses actuarielles

La construction des hypothèses et la mesure de l'erreur associée dans le cadre du calcul du *Best Estimate* sont basées sur une approche statistique non-paramétrique, c'est-à-dire qu'on ne fait pas d'hypothèse à priori sur la forme de la loi de survie. L'utilisation de cette approche est possible grâce à une volumétrie de données disponibles importante. Nous commencerons par une sous-partie portant sur les deux hypothèses ainsi qu'un exposé de la méthode de calcul des taux bruts via les approches classiques (méthode d'estimation de Hoem et Kaplan-Meier).

III.1.2.1 Présentation des risques faisant l'objet d'une hypothèse

Le calcul du *Best Estimate* pour la garantie "décès toutes causes" nécessite deux hypothèses actuarielles : une loi de mortalité par âge et par genre ainsi que d'une loi de résiliation par âge. Ce sont des tables déjà utilisées en hypothèse de projection du modèle central.

La résiliation correspond à un risque comportemental, les assurés ayant la capacité d'agir sur leurs contrats d'assurance de différentes manières. Elle se définit comme la cessation des garanties pour une cause autre que le décès de l'assuré ou une fin de couverture.

Remarque : dans ce portefeuille, les garanties des contrats étudiés sont accordées à compter de la date d'effet jusqu'au 31 décembre de l'année d'adhésion et sont renouvelables annuellement par tacite reconduction. Selon les conditions générales des contrats, l'assuré peut adresser volontairement une demande de résiliation effectuée par lettre recommandée avant le 1er décembre de chaque année avec un mois de préavis. Dans ce cas, nous parlons de la résiliation par le client. Dans le cas de la résiliation par l'assureur, ce dernier peut mettre fin aux garanties suite à un défaut du paiement des cotisations. [1]

III.1.2.2 Statistiques des bases de construction

Pour reconstruire les hypothèses de mortalité et de résiliation, nous avons utilisé les mêmes bases de données que dans le chapitre sur la détection d'anomalies. Il s'agit des bases d'exposition utilisées juste avant la construction des taux bruts et couvrant une période d'observation allant du 01/06/2012 au 01/06/2015 pour la résiliation et allant du 01/01/2010 au 31/12/2014 pour le décès. Ces bases d'exposition sont construites de la manière suivante : on utilise des bases dédoublonnées à la maille contrat-client. En effet, il n'y a qu'une seule ligne par individu et par contrat.

1. En effet, si une cotisation n'est pas payée dans les dix jours après son échéance, l'assureur adresse à l'adhérent une lettre recommandée de mise en demeure, et les garanties cessent 40 jours après l'envoi de la lettre à moins que la cotisation n'ait été versée durant cette période.

La base de construction pour la loi de mortalité est spécifique au risque "décès toutes causes" et possède 430 845 contrats dont 56,9% d'hommes et 43,1% de femmes. La loi de mortalité construite est segmentée par âge et par genre.

La base de construction de la loi de résiliation est commune au modèle central. En effet, la loi de résiliation est utilisée pour provisionner les garanties "décès toutes causes", "décès accidentel", "arrêt de travail" et "invalidité". Cette base, commune à plusieurs garanties, contient initialement 827 613 lignes. La loi de résiliation est uniquement segmentée par âge et considère sans distinguer la résiliation à l'initiative de l'assureur de celle par le client.

III.1.3 Estimation non-paramétrique des taux bruts

Pour les deux hypothèses, la méthode de construction est relativement la même. Il s'agit d'abord d'estimer les taux de sortie par l'estimateur de Kaplan-Meier pour ensuite réaliser un ajustement non-paramétrique. Cette étape n'est pas détaillée dans ce mémoire.

Nous rappelons ici la méthode d'estimation non-paramétrique des taux bruts de sortie ainsi que la façon dont le comportement des assurés a été observé pour construire les taux bruts.

III.1.3.1 Généralités sur les modèles de survie

Rappelons certaines notions basiques des modèles de survie, en reprenant les notations de F. Planchet et P. Therond dans l'ouvrage Modèles de durée - Applications actuarielles[12].

Dans notre cas, nous avons les notations suivantes : la variable aléatoire T représente la durée de survie, c'est-à-dire celle qu'un individu va mettre avant de décéder ou de résilier son contrat, et x l'âge de l'individu. Le taux de sortie pour un individu dont l'âge est compris entre x et $x+t$ est défini par :

$$ {}_tq_x = P(T \leqslant x+t | T > x) $$

Remarque : afin d'expliquer la démarche commune aux deux hypothèses construites, nous désignons, sans faire de distinction, une "sortie" comme un décès ou une résiliation selon la loi considérée.

Dans le cadre de la construction des lois d'expérience, on s'intéresse à l'estimation du taux par âge à horizon d'un an $q_x =_1 q_x$, c'est-à-dire la probabilité de résilier ou de décéder à l'âge x. Dans ce contexte, nous retenons l'expression de la fonction de survie, donnant la probabilité de sortie après une date t, qui caractérise entièrement la distribution de la durée de survie T : $S(t) = P(T > t))$

III.1.3.2 Prise en considération des censures et des troncatures

Dans le cadre de l'estimation des taux bruts, on fixe une fenêtre d'observation $[c, C]$ de plusieurs années. Par exemple, c vaut "01/01/2010" et C est égal à "31/12/2016" dans le cas de la mortalité. Nous allons définir la variable censure, à l'aide de laquelle les taux bruts de résiliation et de mortalité seront estimés.

La contrainte est la suivante : les contrats d'assurance du portefeuille observé ne sont pas tous entièrement couverts par l'intervalle choisi. Ceci est à prendre en compte dans le calcul de l'exposition du portefeuille. On distingue deux phénomènes qui peuvent affecter nos observations :

- On dit qu'il y a une **censure à droite** si la variable aléatoire T_x désignant la durée de survie à l'âge x n'est pas observable totalement au cours de la période C. La sortie est dite "non censurée" uniquement s'il a eu lieu avant C. La variable aléatoire $Z_x = \min(T_x, C)$ désigne la durée de vie future tenant compte de la fenêtre d'observations.

- On parle de **troncature** gauche si la variable aléatoire T_x n'est pas observable antérieurement à c. Contrairement à la censure, dans ce cas on est conscient de l'existence de l'information, mais cette dernière n'est pas mesurable.

Ainsi, les sorties sont réellement observées que lorsqu'elles ont lieu pendant la période d'observation et ne sont pas censurées. Cela nous permet de définir les sorties non censurées de façon adaptée selon la loi considérée :

- Pour la résiliation : il s'agit d'individus dont la date de résiliation est antérieure à la fin de la période d'observation ainsi qu'à un éventuel décès.
- Pour la mortalité : on considère les individus dont le décès survient avant la fin de la période d'observation et la cessation des garanties.

Cela correspond au schéma suivant :

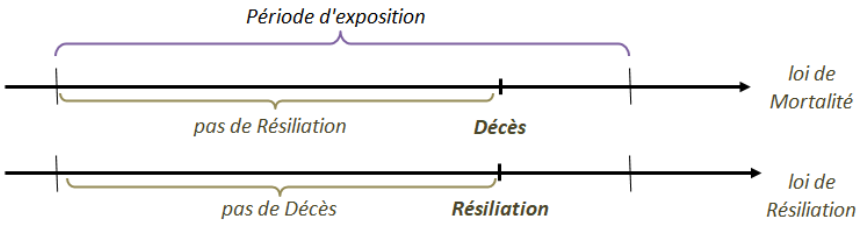


FIGURE III.1: Sorties non censurées selon la loi modélisée

Remarque : pour la loi de résiliation, comme la base historisée ne possède pas la date de résiliation exacte, nous utilisons comme date approximative la date de clôture du contrat.

III.1.3.3 Observation de l'exposition au risque et des sorties

On considère donc qu'une observation non censurée est une sortie, c'est-à-dire un décès ou une résiliation selon la loi estimée. De plus afin de comparer le nombre de censures par âge entier, on calcule l'exposition au risque totale par âge. Cette exposition correspond à la durée cumulée des individus passée dans chacun des âges : une personne pourra être présente à cinq âges consécutifs maximum avec une durée d'exposition à chaque fois comprise entre 0 et 1.

Pour chacune des bases de construction, les durées d'exposition au risque d'une part, et les sorties observées d'autre part, sont les suivantes :

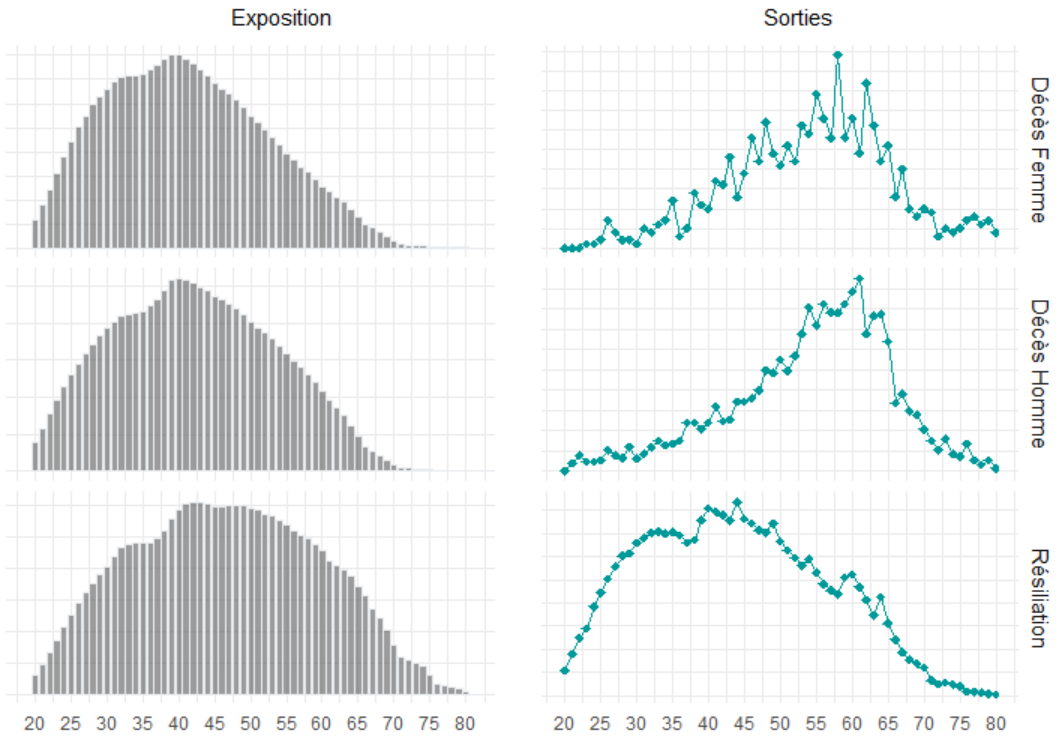


FIGURE III.2: Exposition (gauche) et nombre de sorties (droite) par âge selon la base de construction

On constate que les durées d'exposition sont semblables selon la base de construction. Concernant la mortalité, les sorties ne sont pas localisées aux âges les plus exposés (aux alentours de 45 ans) mais surviennent en majorité vers 60 ans. Cela s'explique simplement par une mortalité plus importante sur les âges élevés. Par ailleurs, on observe une légère augmentation des résiliations à 60 ans qui s'explique par le départ à la retraite.

III.1.3.4 Principaux estimateurs des taux bruts

Il s'agit de construire les hypothèses *Best Estimate* à partir d'une estimation non-paramétrique. Aussi, il peut être nécessaire d'expliquer les deux estimateurs utilisés : l'estimateur Kaplan-Meier de la fonction de survie et l'estimateur de Hoem des taux de sortie.

L'estimateur de Kaplan-Meier La méthode de Kaplan-Meier (E.Kaplan, P.Meyer, [1958] [19]) est une estimation non-paramétrique et unidimensionnelle de la fonction de survie. Sous réserve de disposer de suffisamment d'observations, on peut obtenir des estimations fiables pour la durée de maintien à l'âge x, notée T_x. Cet estimateur est une extension au cas censuré de la fonction de répartition empirique usuelle. De plus, sa simplicité de mise en œuvre, ainsi que le traitement des censures à droite en fait l'un des estimateurs les plus efficaces.

Afin de calculer les taux bruts, l'estimateur de Kaplan-Meier, de la fonction de survie $S(t)$ s'écrit de la manière suivante :

$$\widehat{S}(t) = \prod_{t_i<t} (1 - \frac{d_i}{n_i}) \tag{III.1}$$

Où t_i est un échantillon de durées censurées, n_i est le nombre d'individus à risque au temps t_i et d_i est le nombre de sorties au temps t_i.

A partir de la fonction de survie, nous en déduisons l'estimateur de la probabilité de résilier/décéder à l'âge x suivant :

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

La mesure la plus directe de l'incertitude sur les estimateurs passe par la construction d'intervalles de confiance. Ces intervalles sont généralement ponctuels (i.e. un âge fixé). En s'appuyant sur la propriété de la normalité asymptotique [2] de l'estimateur de Kaplan-Meier, on peut déterminer les intervalles de confiance de la fonction de survie de niveau α.

En prenant $\alpha \in [0,1]$, on a :

$$IC_\alpha = \left[1-(1-\hat{q}_x)\left(1-\phi^{-1}(1-\frac{\alpha}{2})\sqrt{\gamma(t)}\right), 1-(1-\hat{q}_x)\left(1+\phi^{-1}(1-\frac{\alpha}{2})\sqrt{\gamma(t)}\right)\right]$$

où $\gamma = \sum_{i=1|t_i<t}^{m} \frac{d_i}{n_i(n_i-d_i)}$

où$\phi^{-1}(1-\frac{\alpha}{2})$ correspond au quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite.

L'estimateur de Hoem est utilisé afin de calculer les taux bruts de façon plus rapide et de vérifier la cohérence des taux obtenus avec l'estimateur Kaplan-Meier [17].

L'estimateur de Hoem (1969) est un estimateur non-paramétrique qui permet de prendre en compte les censures et les troncatures dans la procédure de l'estimation des taux bruts. Il consiste à les estimer à l'aide d'un rapport entre le nombre de sorties observées et l'exposition.

Cet estimateur est défini en supposant que les taux de sortie sont constants au cours du temps (hypothèse très forte), nous pouvons déduire la relation pour un âge x fixé :

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

où $E_x(t) = \sum_i([\alpha_i, \beta_i] \cup]x, x+1])$ représente l'exposition au risque et d_x correspond au nombre de sorties à l'âge x d'un instant t. La deuxième partie de l'équation est vérifiée si le taux de sortie ne varie pas en fonction du temps.

Remarque : à la différence de l'exposition au risque précédemment calculée, le calcul de l'exposition par âge dans le cadre de l'utilisation de l'estimateur de Hoem associe une valeur de un à chaque sortie observée. Cela permet de garder de la cohérence entre le numérateur et le dénominateur de la fraction $\frac{d_x}{E_x}$.

Intervalle de confiance statistique : ayant estimé q_x par $\hat{q_x}$, il est nécessaire de calculer des intervalles de confiance statistiques sur les taux bruts. Pour cela, il est possible de s'appuyer sur l'estimateur des taux bruts de Hoem et sur l'exposition au risque E_x. En effet, d'après le théorème central-limite et en remplaçant la variance théorique par la variance estimée, on a la loi suivante :

$$\frac{d_x - \mathbb{E}(d_x)}{\sqrt{\mathbb{V}(d_x)}} = \frac{E_x\tilde{q}_x - E_x q_x}{\sqrt{E_x q_x(1-q_x)}} = \sqrt{\frac{E_x}{q_x(1-q_x)}}(\tilde{q}_x - q_x) \xrightarrow[x \longrightarrow +\infty]{\text{Loi}} \mathcal{N}(0,1)$$

2. En général, la normalité asymptotique de l'estimateur de Kaplan-Meier est vérifiée dans le cas où $n_i \geqslant 30$ (critère de Cochran).

Aussi l'intervalle de confiance asymptotique de niveau α pour q_x s'exprime comme :

$$\mathrm{IC}_\alpha = \left[\hat{q}_x + \phi^{-1}(\frac{\alpha}{2})\sqrt{\frac{\hat{q}_x(1-\hat{q}_x)}{E_x}}, \hat{q}_x + \phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{\hat{q}_x(1-\hat{q}_x)}{E_x}}\right]$$

Cet intervalle de confiance IC_α associé aux taux de sortie q_x servira de point de comparaison avec d'autres intervalles de confiances obtenus par ré-échantillonnage au cours des différentes études.

III.1.4 Métrique de comparaison des taux bruts

Afin de de quantifier l'écart entre les lois reconstruites et les lois utilisées en hypothèses, nous allons utiliser le ratio "observés sur attendus". L'objectif est de comparer les résultats des différentes études en utilisant le nombre de sorties observées.

Ratio "observés sur attendus" Le ratio "observés/attendus" est utilisé pour vérifier la validité des taux. Il consiste à comparer le nombre de sorties (décès ou résiliations) théoriques au nombre des décès observés. Pour une classe d'âge allant de m à M, la formule est la suivante :

$$Ratio_{OA} = \frac{Observ\acute{e}s}{Attendus} = \frac{\sum_{x=x_m}^{x_M} d_x}{\sum_{x=x_m}^{x_M} E_x \times \hat{q_x}}$$

où $E_x \times \hat{q_x}$ représente les sorties attendues pour un âge x.

Remarques :

- en prenant $x_m = 20$ et $x_M = 80$, nous avons un ratio "total". Il sera utilisé par la suite.
- selon le portefeuille de construction considéré, les grandeurs d_x et E_x restent inchangées.
- $\hat{q_x}$ peut être le $\hat{q}_x^{BRUT}$, c'est-à-dire les taux bruts ayant initialement servi à la construction des hypothèses, ou bien le $\hat{q}_x^{ALTER}$, obtenu via les différentes études.

C'est pourquoi, ce ratio va nous permettre d'observer à quel point les lois recalculées s'écartent des lois initiales en termes d'impact sur le portefeuille de construction.

III.1.4.1 Taux de sortie bruts par âge

Le graphique ci-dessous représente les taux de sortie bruts estimés selon la méthode de Hoem, les taux bruts déduits de l'estimation de la fonction de survie selon Kaplan-Meier, ainsi que les intervalles de confiance à 95% associés [3] :

3. Etant construit à partir de l'estimateur de Hoem, l'intervalle de confiance statistique est le même que l'intervalle de confiance associé à l'estimateur de Hoem.

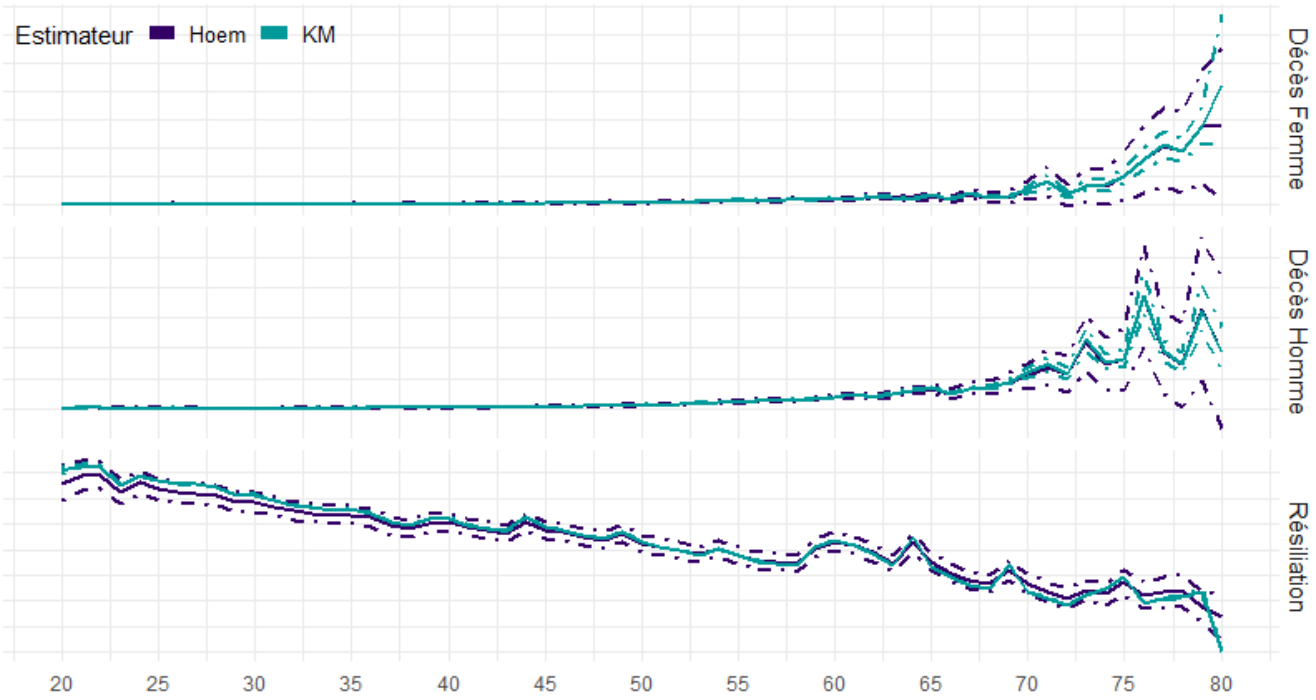


FIGURE III.3: Taux de sortie bruts selon les estimateurs de Hoem et Kaplan-Meier avec intervalles de confiance associés

Les taux estimés par la méthode de Kaplan-Meier sont très proches de ceux estimés par la méthode de Hoem et appartiennent bien aux intervalles de confiance. On note que les deux méthodes d'estimation des taux bruts sont comparables et que les taux bruts estimés selon l'approche Kaplan-Meier sont compris dans les bornes de l'intervalle de confiance asymptotique des taux bruts (sauf à 80 ans).

Concernant les taux de résiliation, nous pouvons observer des pics homogènes à 44, 49, 54, 59, 64 et 69 ans. Ce phénomène s'explique par l'augmentation tarifaire qui a lieu tous les 5 ans à partir du 35 ^{ième} anniversaire de l'assuré.

III.2 Sources d'incertitude relatives à la construction d'hypothèses

On peut considérer que les hypothèses construites ne sont qu'une approximation de la réalité et qu'il existe une loi de survie T définissable à partir de la population globale. Elle représente parfaitement les durées de survie que l'on souhaite estimer.

III.2.1 Formalisation de l'erreur d'estimation

Dans le cas général, une loi de durée est de la forme $T(\Theta) \cong T(\hat{\Theta})$ où T représente la durée de survie. Le paramètre Θ étant inconnu, il est remplacé par un estimateur $\hat{\Theta}$. Cependant, le processus d'estimation du phénomène statistique ainsi que les choix en termes de modèle introduisent des biais. Ce sont ces biais que nous allons essayer d'expliciter au travers de l'expression du taux de sortie.

Au moyen d'une méthode d'estimation non-paramétrique, nous obtenons le taux de sortie estimé brut $\widehat{q}_x(t)$ pour une âge x fixé. Nous considérons l'expression du taux de sortie représentant parfaitement la population globale $q_x(t)$. Il s'agit d'une variable définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ où I est un ensemble de dates discrètes équiréparties (généralement $\mathcal{I} = \mathbb{N}$ *ou* $\mathbb{Z}$) :

$$\begin{array}{rcl} \mathcal{I} \times \Omega & \longrightarrow & \mathbb{R} \\ (t, \omega) & \longmapsto & X_t(\omega) = q_x = \widehat{q}_x(t) + \varepsilon_x(t) \end{array}$$

La variable ε représente l'ensemble des écarts entre la loi estimée et la loi de survie caractérisant parfaitement le phénomène étudié et peut être interprétée comme l'imprécision dans la connaissance de la loi de durée sous-jacente. En effet, en provisionnant, l'assureur s'expose à des déviations adverses relatives à l'utilisation des hypothèses *Best Estimate*.

III.2.2 Risques associés à l'estimation non-paramétrique

L'objectif de la construction d'hypothèses *Best Estimate* est d'utiliser des hypothèses de projection qui ne s'écarte pas de façon significative du portefeuille de façon à prendre en compte le risque intrinsèque au portefeuille.

De manière analogue au mémoire de A. Jequier (2015) [1] sur la quantification de l'erreur d'estimation lors de la construction d'une loi de maintien, nous allons distinguer plusieurs sources d'incertitude intervenant dans la construction des hypothèses non-paramétriques :

- Fluctuations statistiques : du fait de sa taille réduite, l'échantillon ne permet pas une estimation suffisamment robuste de la loi de sortie. En effet, la normalité asymptotique est une hypothèse sous-jacente aux deux estimateurs précédemment cités et requiert suffisamment de données. Il est donc nécessaire d'avoir une exposition par âge suffisante pour chacune des lois considérées.

- Qualité des données : les données utilisées comportent parfois des erreurs de traitement et un manque de fiabilité. Les résultats relatifs à l'évaluation des différents risques se retrouvent dégradés du fait de la mauvaise qualité des données.

- Fluctuations d'échantillonnage : les estimations ne peuvent représenter totalement la diversité du phénomène étudié pour les raisons suivantes. Puisque chaque sous-portefeuille (homogène en termes de profil de risque) présente potentiellement une sinistralité différente, la loi globale n'est plus *Best Estimate* pour chaque sous-portefeuille. Aussi, n'étant pas nécessairement équirépartie, chaque sous-population du portefeuille n'est pas représentée de façon homogène. C'est d'ailleurs pour limiter cet effet qu'il faut utiliser une segmentation adaptée en termes de profil de risque, comme par exemple le genre pour la loi de mortalité.

  Plus précisément, au cours de la construction d'une loi d'expérience, se pose la question de la segmentation à retenir : avec une approche globale pour l'ensemble du portefeuille, on court le risque de n'être finalement *Best Estimate* sur aucune des sous-populations significatives composant le portefeuille. Cela revient à construire une table inadaptée dès que la répartition du portefeuille par profil de risque se déforme avec le temps. Cependant, cette segmentation ne peut être trop poussée afin de limiter les fluctuations statistiques induites par la taille réduite de l'échantillon. En effet, un échantillon restreint entraîne nécessairement des taux bruts beaucoup plus volatils. Aussi, ne pouvant pas segmenter le portefeuille suffisamment du fait d'une taille d'échantillon limitée, les fluctuations d'échantillonnages correspondent à la mauvaise prise en compte de l'hétérogénéité [4] du portefeuille.

4. On appelle "hétérogénéité" tout facteur pouvant traduire une différence de sinistralité entre deux populations.

- Dérive temporelle : les hypothèses sont obtenues à partir de données antérieures et sont appliquées sur des données récentes sans considération pour l'évolution du portefeuille en termes de risque. Or, le choix de la période d'observation engendre un biais dans la construction des taux pour deux raisons. Premièrement, l'estimation des taux bruts peut devenir inadaptée dans le cas d'un changement de conjoncture généré par un évènement exogène (tel qu'une épidémie ou une catastrophe naturelle). L'erreur temporelle induite par la conjoncture est très compliquée à estimer et est intégrée dans les différents calculs de SCR. Deuxièmement, la loi construite ne peut s'appliquer au portefeuille de projection à cause d'une raison structurelle. Par exemple, il est possible que la mortalité augmente au fil du temps si le portefeuille ne se renouvelle pas assez. Il s'agit alors d'une dérive temporelle que l'on cherche à quantifier.

Remarque : concernant les fluctuations statistiques, on peut considérer que cette source d'incertitude est prise en compte lors du lissage des taux bruts permettant d'obtenir les hypothèses finales. On a donc déjà une quantification possible de cette source d'erreur et il n'est pas nécessaire de reconstruire des taux bruts alternatifs.

III.2.3 Incertitudes possibles dans l'estimation d'un taux de survie

Il est donc possible d'associer plusieurs sources d'incertitude à la construction d'une loi d'expérience. Ces sources d'incertitude sont induites par la méthode de construction utilisée et se retrouvent dans l'ensemble des écarts entre les taux de sortie construits et ceux représentant parfaitement les durées de survie estimées :

$$\boxed{\begin{array}{rcl} \mathcal{I} \times \Omega & \longrightarrow & \mathbb{R} \\ (t, \omega) & \longmapsto & X_t(\omega) = q_x = \widehat{q}_x(\mathrm{t}) + \varepsilon_x(t) \\ & & \\ & & \text{Où } \varepsilon_x(t) = \varepsilon_{S_x} + \varepsilon_{D_x} + \varepsilon_{E_x} + \varepsilon_{T_x} + \varepsilon \end{array}}$$

où $\widehat{q}_x(t)$ est obtenu par estimation non-paramétrique et ε représente la source d'erreur résiduelle, c'est-à-dire "purement" aléatoire ou relative à d'autre sources d'erreurs éventuelles.

Les différentes sources d'incertitudes sont synthétisées dans le tableau suivant :

Source d'incertitude	Notation	Description
Fluctuations statistiques	ε_{S_x}	Non robustesse des taux bruts induite par la taille réduite de l'échantillon
Qualité des données	ε_{D_x}	Biais dans l'exactitude des résultats induit par la présence de données anormales
Fluctuations d'échantillonnage	ε_{E_x}	Mauvaise prise en compte de l'hétérogénéité du portefeuille
Dérive temporelle	ε_{T_x}	Evolution de la sinistralité entre la période d'observation et la période de projection

Table III.1: Principales sources d'erreurs dans le cadre d'une estimation non-paramétrique

Concernant le comportement de ces différentes sources d'incertitudes, on peut considérer que la dérive temporelle du portefeuille ε_{T_x} suit une tendance mais que ε_{S_x}, ε_{E_x} sont des variables centrées réduites. Concernant la qualité des données, on ne fait pas de conjecture sur le comportement de ε_{D_x} [5].

Par ailleurs, ces sources d'erreur ne sont pas totalement indépendantes :

- Par exemple, réduire les fluctuations statistiques en augmentant la taille de l'échantillon génère des fluctuations d'échantillonnage puisque le portefeuille est moins segmenté. Ces fluctuations sont induites par une mauvaise prise en compte des profils de risque hétérogènes du portefeuille.

- De même, concernant la dérive temporelle et les fluctuations d'échantillonnage, il est possible que la dérive du portefeuille s'explique en partie par l'évolution de la répartition des classes de risques. Par exemple, il se peut que les cadres supérieurs deviennent de plus en plus majoritaire alors que la base de construction prenait surtout en compte des agriculteurs.

III.2.4 Sources d'incertitude principales faisant l'objet d'une étude

Les taux de sortie étant calculés sur une période d'observation limitée, il est possible que les caractéristiques du portefeuille, en termes de profil de risque, aient évolué par rapport au moment de la construction des hypothèses. Il y a donc un risque de mauvaise estimation des taux de sortie. En estimation non-paramétrique, ce risque intervient principalement lors de l'estimation des taux bruts via l'effet lié aux fluctuations d'échantillonnage d'une part, ainsi que l'effet induit par la dérive temporelle d'autre part.

Ainsi, afin de quantifier les différentes sources d'incertitudes, deux études principales seront menées. L'une pour essayer de capter les fluctuations d'échantillonnage associées à l'hétérogénéité du portefeuille, l'autre permettant d'estimer les fluctuations temporelles induites par le choix de la période d'observation.

III.3 Application

Les différents types d'erreurs étant précisés, nous allons évaluer les risques associés à la modélisation des taux de sortie en étudiant l'impact de ces sources d'incertitude sur les taux bruts. L'objectif est d'anticiper une dérive du portefeuille en termes de répartition par catégories socio-professionnelles et par trimestre d'exposition. Nous commencerons par présenter les résultats pour la source d'erreur associée à l'utilisation des données et nous détaillerons la technique de ré-échantillonnage utilisée pour quantifier les deux autres sources.

III.3.1 Erreur induite par la mauvaise qualité des données

Afin de quantifier l'erreur imputable aux données des portefeuilles servant à la construction d'hypothèses, nous reprenons les résultats du premier chapitre portant sur les méthodes de détection d'anomalies. Pour rappel, les effectifs anormaux des bases de construction sont les suivants : 4230 individus pour la mortalité des femmes, 3772 pour la mortalité des hommes et 4815 pour la résiliation. Nous reconstruisons les taux bruts à partir des portefeuilles de construction auxquels nous avons retiré les individus détectés comme des anomalies :

5. En effet, ε_{D_x} correspond à l'écart entre la loi initiale et la loi recalculée sans anomalies, on ne sait pas encore l'effet de la suppression des lignes anormales sur les taux de sortie.

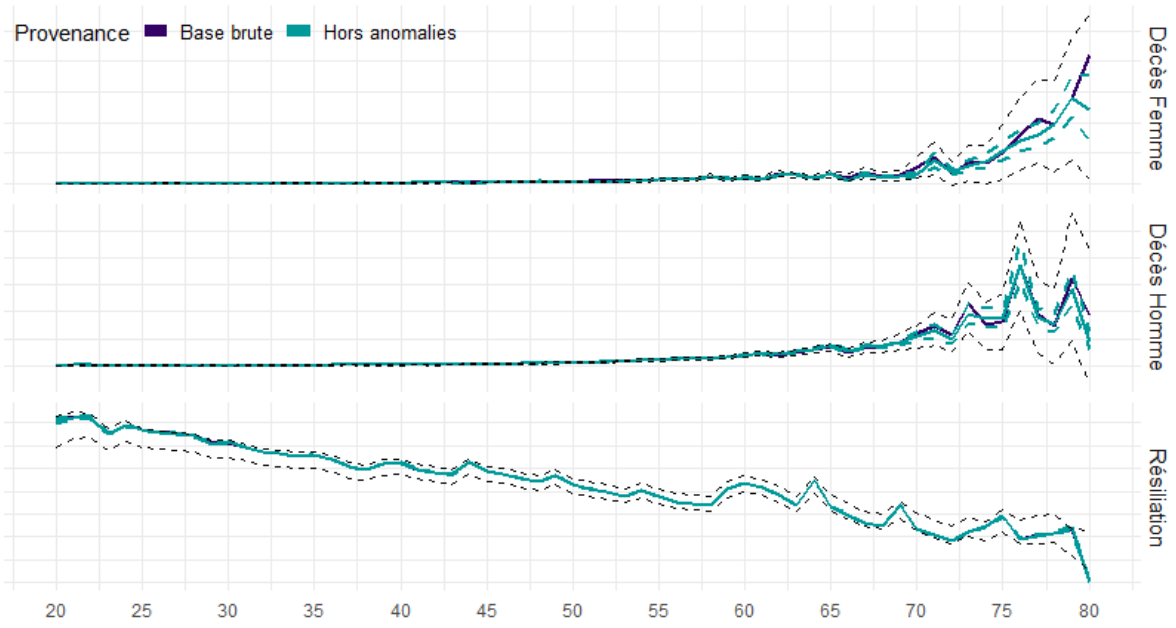


FIGURE III.4: Taux par âge recalculés à partir des bases de construction hors anomalies

Les taux recalculés après avoir supprimé les anomalies (en violet) sont globalement très proches des taux initiaux. En effet, les taux estimés et les intervalles de confiances associés sont situés à l'intérieur des bornes de confiance asymptotiques (en pointillé noir). Cela peut s'expliquer par la faible part d'anomalies détectées par rapport à la taille conséquente des bases de construction (plus de 170 000 lignes). Néanmoins, on constate des écarts sur les âges élevés pour la mortalité. Pour mieux se rendre compte, comparons le nombre de sorties observées au nombre de sorties estimées d'après nouvelles lois :

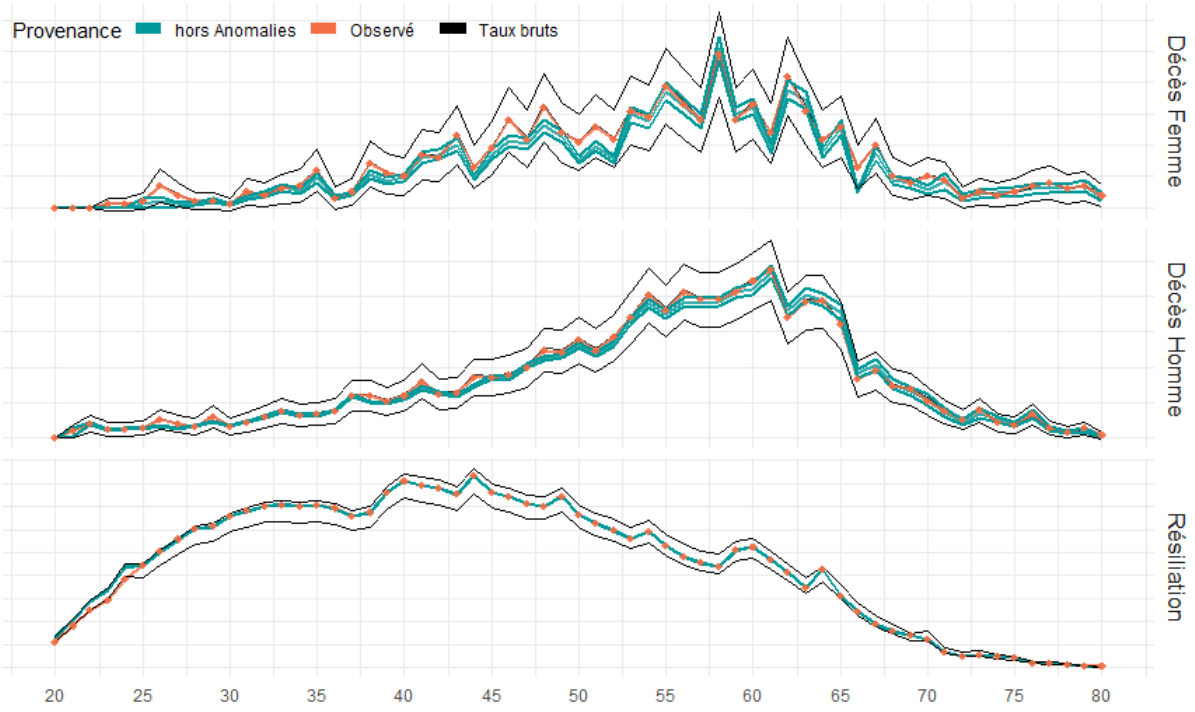


FIGURE III.5: Sorties provenant des taux recalculés par Kaplan-Meier sur les données sans anomalies

Afin d'isoler les différentes sources d'erreur, nous comparons les sorties observées des portefeuilles de construction initiaux (points orange) aux sorties estimées à l'aide des bases de construction sans anomalies (courbes bleues) [6]. On observe globalement que les sorties initialement observées sont supérieures aux sorties estimées avec les bases de construction sans anomalie. En effet, concernant la mortalité, le nombre de sorties par âge est globalement situé au-dessus des sorties associées à la borne supérieure des intervalles de confiance des taux estimés par Kaplan-Meier sur la base sans anomalie.

L'estimation basée sur l'ensemble des données comporte ainsi une erreur de données, dans la mesure où elle conduit à une surestimation des sorties par rapport à ce qu'une estimation sans anomalies produirait. C'est pourquoi, conserver des anomalies dans les données entraîne une surestimation des engagements.

Remarque : les études qui suivent ont été dupliquées car effectuées à partir des bases de construction avec et sans anomalies. En effet, afin d'isoler l'erreur imputable à l'utilisation des hypothèses, il est nécessaire de repartir des bases de données sans anomalie. Les différents graphiques sont obtenus à partir des bases de construction brutes.

III.3.2 Formalisation de la procédure de ré-échantillonnage

Dans le cadre de la quantification de l'incertitude associée à l'erreur d'estimation, nous utilisons une procédure *Bootstrap*. Le *Bootstrap* est une méthode de ré-échantillonnage introduite par Efron en 1979 [5]. L'idée est de reconstruire des échantillons pour permettre une inférence statistique plus fine. Cette méthode est pragmatique et facile à mettre en œuvre, mais nécessite des outils informatiques performants. Elle repose sur deux hypothèses :

- l'indépendance des observations : ce qui permet un tirage avec remise ;
- l'unicité de la loi de distribution : chaque élément de l'échantillon initial est identiquement distribué.

Procédure : on construit N échantillons *bootstrap* en tirant au sort, avec remise, n observations au sein des n données de l'échantillon initial (i.e. le portefeuille de construction dans notre cas). Chaque observation (ou regroupement d'observations) est tirée de façon équiprobable afin de représenter au mieux le portefeuille considéré. Le résultat se présente sous la forme de $N = 1000$ portefeuilles alternatifs de la forme suivante :

$$\begin{array}{rcl} Portefeuille\ 1 & : & x_{11}, ..., x_{n,1} \\ Portefeuille\ 2 & : & x_{12}, ..., x_{n,2} \\ Portefeuille\ 3 & : & x_{13}, ..., x_{n,3} \\ ... & : & ... \\ Portefeuille\ N & : & x_{1,N}, ..., x_{n,N} \end{array}$$

6. Il aurait été plus simple de comparer les sorties observées de la base de construction sans anomalie avec les taux bruts initiaux, mais cela aurait perdu en cohérence avec les graphiques des études qui vont suivre.

Ensuite, nous allons reconstruire $N = 1000$ chroniques de taux bruts à l'aide de l'estimateur de Kaplan-Meier comme c'est le cas des taux bruts initiaux. Cela nous permet d'obtenir une loi "moyenne" (via une approximation de Monte-Carlo) ainsi qu'un intervalle de confiance (via des quantiles empiriques à 2,5% et à 97,5%). Ce qui nous donne pour un âge x fixé :

$$\hat{q_x}^{BOOT} = \frac{1}{N} \sum_{b \in \{1,\dots,N\}} \hat{q_x}_b^{BOOT} \quad ; \qquad IC_{N,\alpha}^{BOOT} = [X_{[\frac{N(1-\alpha)}{2}]+1}, X_{[\frac{N(1+\alpha)}{2}]+1}]$$

Où $N = 1000$ correspond ici au nombre d'échantillons *bootstrap* ;
Où α correspond à un niveau de confiance de 95%.

Ainsi, la méthode permet globalement de répondre à la question suivante : Si nous avions des données différentes mais en cohérence avec le portefeuille étudié, comment cela affectera-t-il notre estimation ?

Dans le cadre de cette étude, l'idée est d'appliquer des perturbations sur les données en construisant l'échantillon *bootstrap* et d'encadrer les différents taux bruts alternatifs pour mesurer les différentes sources d'incertitude sur les estimateurs.

III.3.3 Incertitude causée par les fluctuations d'échantillonnage

La première source d'erreurs d'estimation correspond au biais induit par l'échantillonnage, c'est-à-dire la segmentation des individus au sein de la population globale que l'on souhaite modéliser. En effet, il arrive souvent que les durées observées résultent de l'agrégation de sous-populations souvent inobservables ayant chacune un comportement spécifique. On parle alors d'hétérogénéité. [2]

Il aurait été idéal de construire des lois propres à chaque ensemble homogène. Cependant, cela n'est pas possible en pratique en raison la taille réduite des échantillons segmentés. En effet, un échantillon restreint entraîne nécessairement des taux bruts beaucoup plus volatils. Il est nécessaire de prendre en compte l'hétérogénéité de la population observée afin de limiter le bais associé.

Plus précisément, il s'agit de se demander en quoi la distribution des profils de risque est susceptible de modifier les taux bruts obtenus et par conséquent les provisions prudentielles qui en résultent. La population d'un portefeuille n'est jamais totalement homogène, l'objectif est de prendre en compte un éventuel changement dans la répartition des sous-populations au sein du portefeuille de construction considéré.

III.3.3.1 Catégorie socio-professionnelle comme déterminant du profil de risque

Outre l'âge et le genre, l'appartenance à la catégorie socio-professionnelle (CSP) est un élément déterminant dans l'évaluation de la mortalité et de la résiliation. En effet, la CSP utilise la profession comme critère d'appartenance à une catégorie sociale [7].

L'appartenance à la catégorie socio-professionnelle peut expliquer un certain niveau de richesse et de culture, donc un accès plus ou moins important à la santé ainsi qu'à la connaissance des conditions générales des contrats d'assurance et des possibilités de résiliation. De plus, la profession peut être un facteur explicatif d'un décès prématuré ou tardif : un ouvrier ayant un travail physique et pénible est susceptible de contracter plus de maladies chroniques qu'un cadre d'entreprise.

7. On aurait pu explorer d'autre facteur déterminant du profil de risque en fonction des données disponibles tel que la localisation géographique.

Nous observons cet effet sur nos portefeuilles de construction. Afin d'avoir une volumétrie suffisante pour chaque maille, nous avons défini des classes d'âge tous les cinq ans et nous avons regroupé les CSP en 4 catégories [8]. Cela consiste à observer le nombre de décès et de résiliations par rapport à l'exposition (la somme des durées passées) au sein de chaque classe d'âge pour chaque catégorie de CSP [9] :

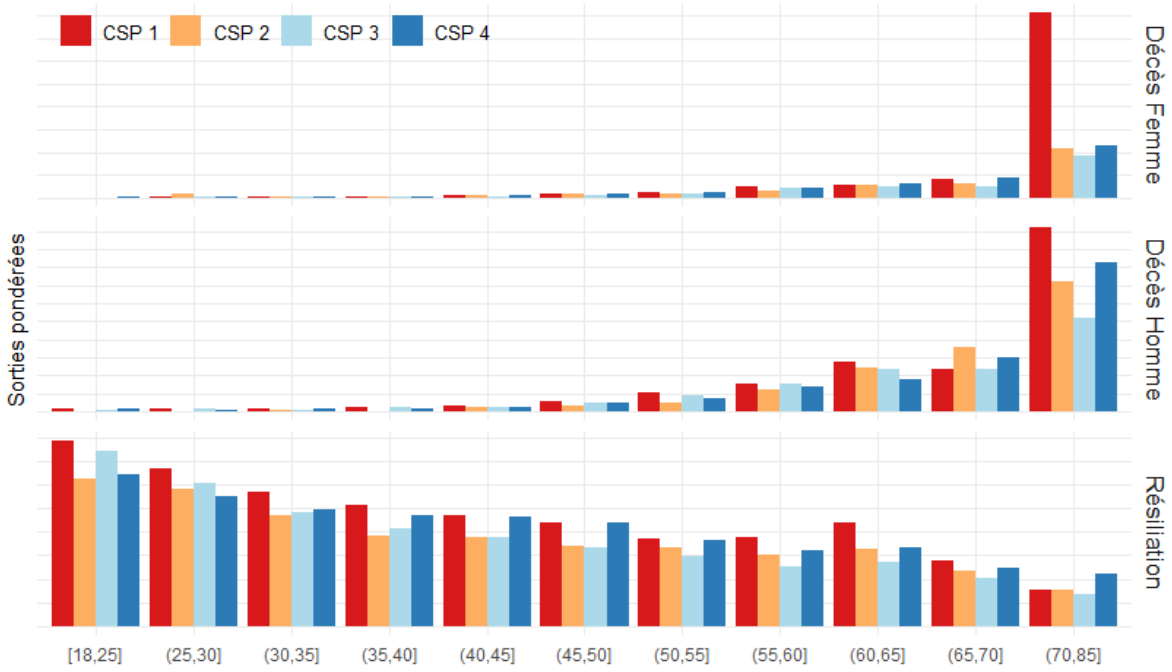


FIGURE III.6: Sorties pondérées par l'exposition au risque pour chaque classe d'âge selon la catégorie socio-professionnelle

Quelle que soit la loi considérée, on constate généralement des sorties pondérées supérieures pour le regroupement de CSP 1. A l'inverse, la mortalité de la CSP 3 est plus faible que pour les autres CSP.

On peut donc considérer que l'appartenance à la CSP est un facteur déterminant dans l'estimation des taux de sortie. Il s'agit ensuite de modifier l'échantillon de construction des taux bruts en perturbant l'appartenance à la catégorie socio-professionnelle.

III.3.3.2 Modalité des tirages selon l'appartenance à la CSP

Afin de quantifier l'erreur imputable aux fluctuations d'échantillonnage, nous allons prendre en compte une éventuelle évolution de la répartition par catégorie de CSP au sein du portefeuille. L'idée est de construire des lois équivalentes à partir d'autres échantillons possibles, avec un profil de risque sensiblement différent.

Pour chaque classe d'âge, un paquet de lignes est associé à un regroupement de CSP. Tout d'abord, pour chaque classe d'âge, nous effectuons un tirage avec remise pour obtenir une liste de regroupements de CSP. Afin de respecter les proportions d'appartenance à la classe d'âge, nous commençons par créer une table de pondération reposant sur l'exposition par regroupement de CSP et par classe d'âge. Par exemple, concernant la résiliation, cette table est la suivante :

8. Pour des raisons de confidentialité, les catégories de CSP et les modalités de regroupement ne sont pas précisées.
9. Il ne s'agit pas de taux bruts par âge au sens strict, l'exposition n'associant pas une valeur de un à chaque sortie.

Classe d'âge	CSP 1	CSP 2	CSP 3	CSP 4
[18, 25]	0,2%	87,2%	9,4%	3,3%
(25, 30]	1,2%	65,8%	20,5%	12,5%
(30, 35]	3,7%	51,0%	24,9%	20,4%
(35, 40]	6,4%	43,2%	26,5%	23,9%
(40, 45]	8,2%	38,8%	29,1%	23,9%
(45, 50]	9,3%	35,7%	30,5%	24,5%
(50, 55]	10,0%	34,6%	30,1%	25,2%
(55, 60]	11,9%	33,8%	28,5%	25,8%
(60, 65]	19,7%	32,0%	22,9%	25,4%
(65, 70]	41,2%	25,0%	10,4%	23,4%
(70, 80]	61,3%	18,2%	1,2%	19,3%

Table III.2: Illustration de la pondération par classe d'âge selon la CSP

Nous effectuons un tirage avec remise pour chaque classe d'âge tant que la somme des poids n'excède pas 1. Ensuite, on reconstitue les $N = 1000$ portefeuilles à partir des vecteurs de regroupements de CSP tirés. Chaque portefeuille est obtenu en faisant une combinaison de paquets lignes du portefeuille de construction découpés à la maille regroupement de CSP - classe d'âge.

III.3.3.3 Taux bruts obtenus par ré-échantillonnage selon l'appartenance à la CSP

Enfin, on poursuit la procédure *bootstrap* et on estime les $N = 1000$ taux bruts par Kaplan-Meier ainsi que les intervalles de confiance empiriques à 95%. On constate que les taux obtenus en moyenne sont très proches des taux initiaux ce qui conforte la procédure *bootstrap*. En effet, les deux lignes sont superposées :

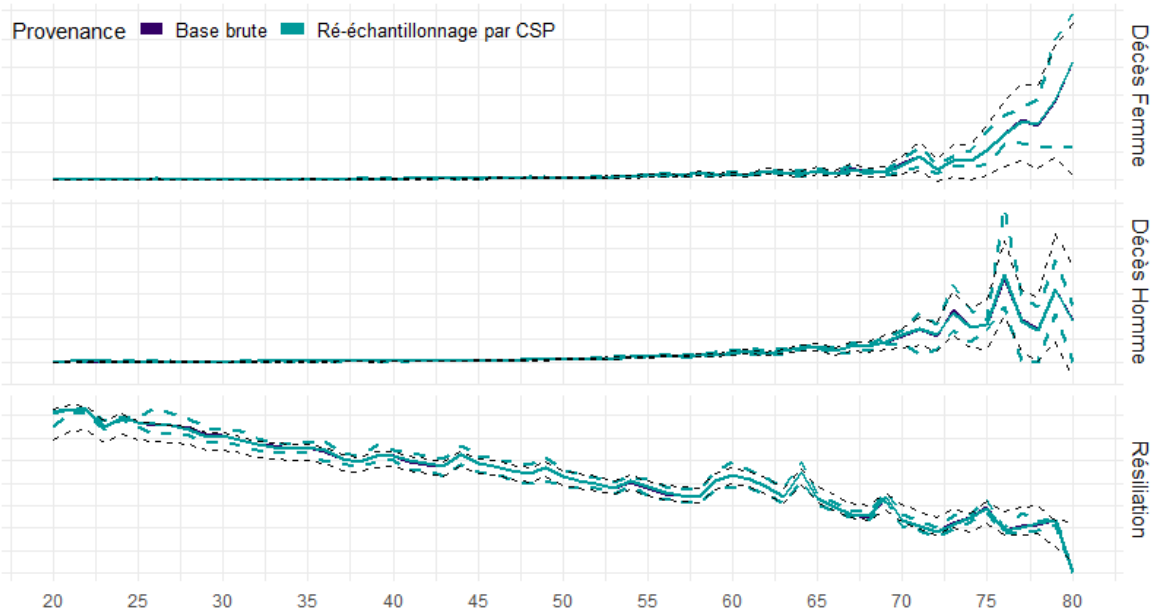


Figure III.7: Taux bruts obtenus par ré-échantillonnage selon la CSP

Néanmoins, le but de cette sous-partie est de prendre en compte la volatilité, c'est-à-dire à quel point les taux estimés sont susceptibles de varier du fait d'un changement dans la répartition des profils de risque. Pour cela, observons le nombre de sorties observées comparé au nombre de sorties estimées à l'aide des nouvelles lois :

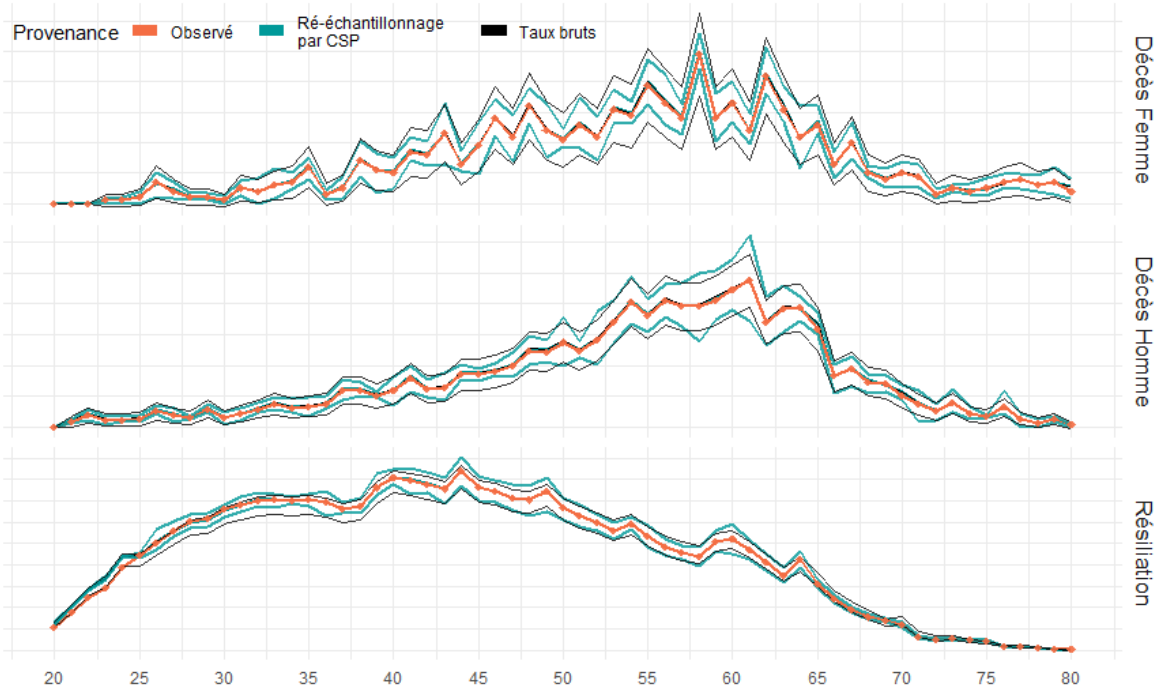


FIGURE III.8: Sorties estimées à partir des taux bruts recalculés par ré-échantillonnage *boostrap* selon l'appartenance à la CSP

On observe que les courbes bleues sont proches des intervalles de confiances asymptotiques (en noir). Ces courbes correspondent aux intervalles de confiances empiriques obtenus suite à la procédure *bootstrap*. Un point d'attention pourrait concerner la mortalité des femmes où les quantiles empiriques excèdent les bornes de l'intervalle de confiance.

III.3.4 Incertitude causée par la dérive temporelle des portefeuilles de construction

La plupart des techniques de modélisation se basent sur le lien entre les variables explicatives observées pendant une période antérieure mais sont utilisées pour prédire une valeur cible postérieure à l'étude à partir de la connaissance des variables observées sur la période $[c, C]$. Cette utilisation suppose implicitement une stabilité dans le temps de la population étudiée.

Or, le choix de la plage d'années à partir desquelles on construit la loi d'expérience est important dans la mesure où les estimations des paramètres dépendent sensiblement de ce choix et engendrent des prévisions différentes. La dernière source d'erreurs correspond au biais induit par le choix de la fenêtre d'observation. En effet, il est nécessaire de prendre en compte l'évolution du portefeuille en termes de risque. Plus précisément, il s'agit de se demander en quoi la temporalité est susceptible de modifier les taux bruts obtenus et par conséquent les provisions prudentielles qui en résultent.

III.3.4.1 Prise en compte du trimestre pour estimer la dérive temporelle

Il s'agit ici d'évaluer la stabilité des taux de sortie par rapport aux années couvertes par la période d'exposition. Pour cela, nous commençons par regarder le nombre de sorties mensuelles pour chacun des portefeuilles de construction :

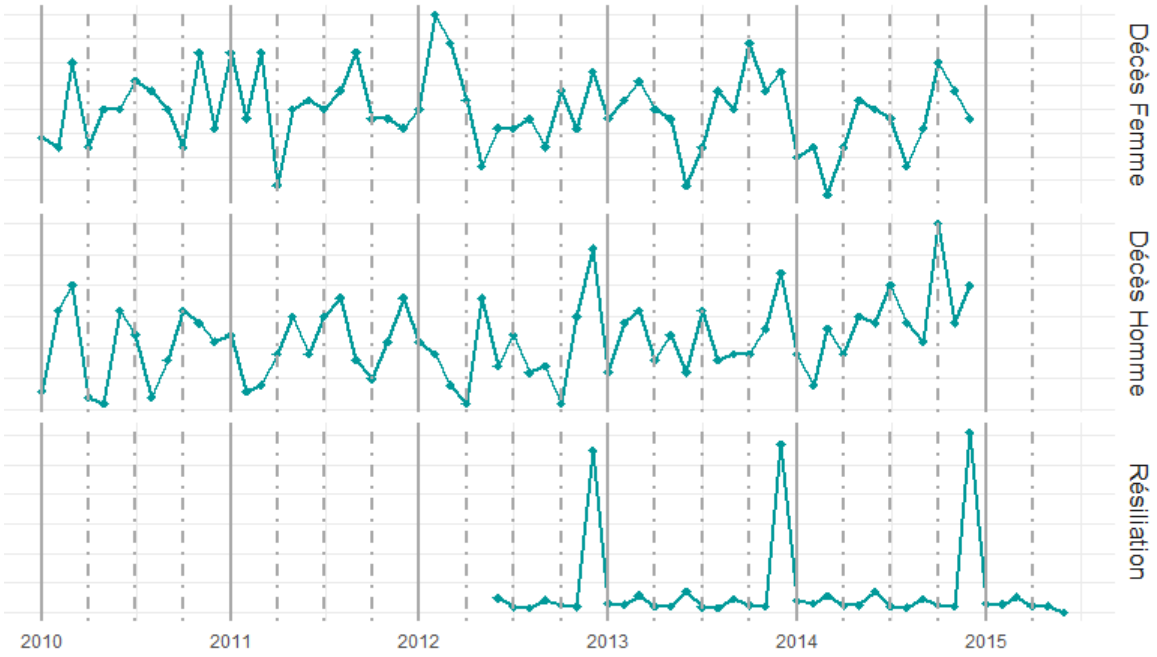


FIGURE III.9: Nombre de sorties mensuelles pour chaque loi

On observe généralement un nombre de sorties plus important en fin d'année. Pour la mortalité, cela peut s'expliquer par les grippes hivernales et les périodes de fête (abus d'alcool, suicides, accidents de la route...).

Concernant la résiliation, le pic au mois de décembre est dû à une clause contractuelle. En effet, selon les conditions générales des contrats, l'assuré peut adresser volontairement une demande de résiliation effectuée par lettre recommandée avant le 1er décembre de chaque année avec un mois de préavis ce qui fait que la résiliation à l'initiative du client pour les contrats annuels avec une tacite reconduction a pour échéance le 31 décembre. Par ailleurs, l'assuré pouvant choisir, lors de la souscription du contrat, le mode de versement des primes (annuel, semestriel, trimestriel, ou mensuel), on observe un effet saisonnier de la résiliation.

C'est pourquoi nous choisissons de redécouper les différentes bases d'exposition en fonction du trimestre d'observation.

III.3.4.2 Procédure de trimestrialisation des bases d'exposition

Pour faire cela, nous commençons par sélectionner les observations présentes dans chaque trimestre en fonction des dates d'effet, de décès et de clôture. Ensuite, nous avons redéfini les âges à l'entrée et à la sortie du trimestre d'exposition. Cela consiste à utiliser le minimum entre la date d'effet et le début du trimestre pour définir l'âge d'entrée et le maximum entre la date de sortie et la fin du trimestre pour calculer l'âge de sortie. Enfin, de façon adaptée à chaque risque, nous avons recalculé la variable représentant les sorties non censurées.

Par ailleurs, nous avons vérifié la ventilation du nombre de sorties, c'est-à-dire si le nombre de sorties de la base de construction initiale est bien égal au nombre total de sorties des bases trimestrielles. Nous avons donc 20 bases d'exposition trimestrielles pour la mortalité par genre et 16 bases pour la résiliation.

III.3.4.3 Taux bruts obtenus par ré-échantillonnage selon le trimestre d'observation

Une fois les bases trimestrielles crées, nous allons reconstruire les bases d'expositions en interchangeant les trimestres au cours de l'année. Plus précisément, il s'agit d'effectuer un ré-échantillonnage en prenant en compte les trimestres de l'année civile. Pour chacun des quatre trimestres, cela revient à tirer avec remise une année. Nous avons par exemple les tirages suivants :

$Portefeuille\ 1$	:	${T_1}^{A1};{T_2}^{A2};{T_3}^{A1};{T_4}^{A2};{T_1}^{A3};{T_2}^{A1};{T_3}^{A3};{T_4}^{A3};{T_1}^{A1};{T_2}^{A2};{T_3}^{A3};{T_4}^{A2}$
$Portefeuille\ 2$	:	${T_1}^{A2};{T_2}^{A3};{T_3}^{A1};{T_4}^{A2};{T_1}^{A3};{T_2}^{A1};{T_3}^{A3};{T_4}^{A3};{T_1}^{A1};{T_2}^{A1};{T_3}^{A3};{T_4}^{A3}$
$Portefeuille\ 3$	:	${T_1}^{A1};{T_2}^{A3};{T_3}^{A1};{T_4}^{A2};{T_1}^{A2};{T_2}^{A1};{T_3}^{A3};{T_4}^{A3};{T_1}^{A1};{T_2}^{A2};{T_3}^{A3};{T_4}^{A2}$
...	:	...
$Portefeuille\ N$	:	${T_1}^{A3};{T_2}^{A2};{T_3}^{A1};{T_4}^{A2};{T_1}^{A1};{T_2}^{A1};{T_3}^{A1};{T_4}^{A3};{T_1}^{A1};{T_2}^{A2};{T_3}^{A3};{T_4}^{A1}$

Table III.3: Exemple de tirages de trimestres
en prenant en compte l'année d'exposition i (pour la résiliation, i=3 années d'historique)

Ensuite, on reconstitue les $N = 1000$ portefeuilles à partir des différents trimestres tirés. Afin de gagner en temps de calcul et au lieu de simplement coller les bases entre elles, nous avons fusionné les trimestres en prenant en compte les individus présents sur plusieurs périodes consécutives. Cela revient à redéfinir les âges d'entrée et de sortie.

Les taux obtenus par Kaplan-Meier en faisant une approximation de Monte-Carlo ainsi que les intervalles de confiance empiriques sont les suivants :

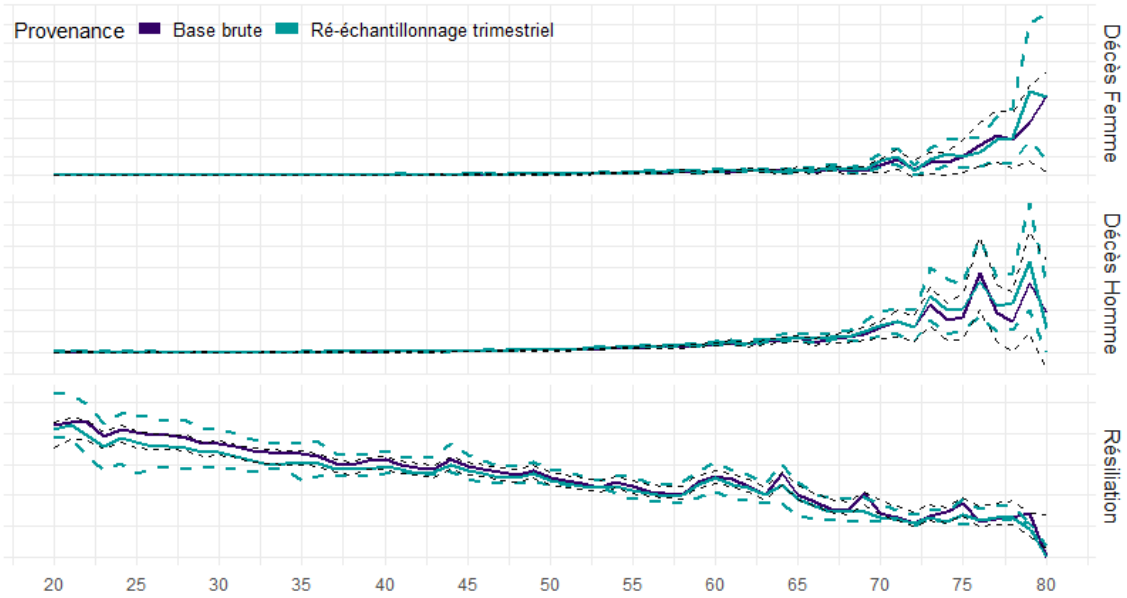


Figure III.10: Taux bruts recalculés après un ré-échantillonnage trimestriel

On constate que les taux obtenus sont plus éloignés des taux initiaux par rapport au ré-échantillonnage par CSP. De même, les intervalles de confiance empiriques semblent plus larges, notamment sur les âges faiblement exposés. Afin de vérifier, observons à nouveau le nombre de sorties observées comparé au nombre de sorties estimées à l'aide des nouvelles lois calculées par ré-échantillonnage trimestriel :

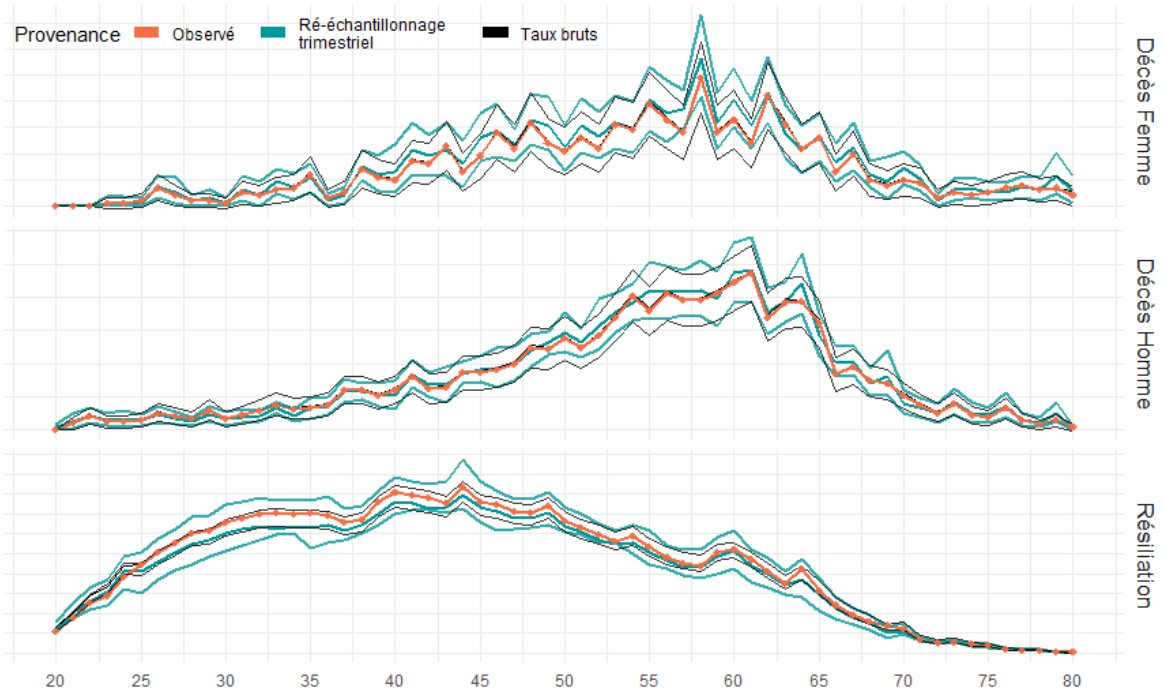


FIGURE III.11: Sorties estimées à partir des taux bruts recalculés par ré-échantillonnage *boostrap* selon le trimestre d'observation

On remarque que les écarts totaux mesurés sont plus importants que les écarts estimés lors de l'étude consacrée au ré-échantillonnage par CSP. En effet, le nombre de sorties associé aux quantiles empiriques dépasse les intervalles de confiance statistique, notamment dans le cas de la résiliation.

III.4 Conclusion partielle

Afin d'isoler l'erreur imputable à l'utilisation des hypothèses *Best Estimate*, nous sommes partis des bases servant à la construction des taux bruts des lois de résiliation et de mortalité. Outre l'incertitude générée par l'utilisation de données anormales, nous nous sommes concentrés sur deux sources principales d'incertitude : l'hétérogénéité du portefeuille et le choix de l'historique du portefeuille.

Pour faire cela, nous avons utilisé différentes procédures *bootstrap* pour mesurer la robustesse des lois face à la construction de portefeuilles alternatifs (avec une répartition par CSP ou des trimestres d'observation différents). L'idée était d'utiliser des données différentes mais en cohérence avec le profil de risque du portefeuille de construction utilisé.

Tout cela nous a permis d'obtenir des quantiles empiriques pour chaque âge ainsi que des chroniques de taux alternatifs. Afin d'avoir une vue d'ensemble sur les différents taux bruts recalculés, nous observons les ratios "observés sur attendus" totaux. Cela nous permet d'observer à quel point les lois recalculées s'écartent des lois initiales en terme d'impact sur le portefeuille de construction et de comparer les différents intervalles de confiance [10] :

Ratio O/A total	Décès Femme			Décès Homme			Résiliation		
Source d'incertitude	Sorties centrales	IC -	IC +	Sorties centrales	IC -	IC +	Sorties centrales	IC -	IC +
Fluctuations statistiques		68,9%	134,1%		78,7%	127,5%		94,1%	109,7%
Qualité des données	112,4%			102,6%			99,4%		
Fluctuations d'échantillonnage		76,4%	137,8%		80,8%	126,6%		93,0%	106,5%
Dérive temporelle		62,8%	144,3%		77,2%	128,0%		88,7%	119,2%

FIGURE III.12: Ratios "Observés/Attendus" obtenus à partir des taux bruts recalculés par ré-échantillonnage *boostrap* selon le trimestre d'observation

Finalement, plusieurs constats globaux peuvent être faits :

- Concernant l'incertitude associée à la qualité des données, on s'intéresse aux taux bruts alternatifs. On observe, pour la mortalité, une sous-estimation du nombre de sorties observées : le ratio O/A est supérieur à 1, le nombre de sorties attendues étant plus faible que l'observé. Cela signifie finalement que la mauvaise qualité des données génère une surestimation des hypothèses : les taux de sorties obtenus sans anomalies (auxquelles on a retiré le biais induit par la mauvaise qualité des données) sont inférieurs aux taux de sortie des bases de données initiales. Il s'agira ensuite d'en quantifier l'impact sur les *Best Estimates* modélisés.

- Concernant les bornes des intervalles de confiance, nous prenons comme point de comparaison l'intervalle de confiance asymptotique. Comme déjà observé, la prise en compte des trimestres d'observation génère le plus d'incertitude quelle que soit la loi construite. En effet, les ratios associés à la dérive temporelle sont plus éloignés par rapport aux autres ratios. A l'inverse, les ratios associés aux fluctuations d'échantillonnage sont moins dispersés. Les résultats, non présentés ici, sont semblables pour les bases obtenues sans anomalies.

Quant aux prolongements possibles de ce chapitre, outre l'appartenance à la CSP, il convient de se demander s'il n'y a pas d'autres sources d'hétérogénéité que l'on aurait pu prendre en compte tel que la localisation géographique ou les pratiques de vie (tabac, alcool). De plus, il aurait été pertinent d'utiliser d'autres années d'exposition plus récentes afin d'étayer l'étude sur la dérive temporelle. Par ailleurs, il ne faut pas oublier que les études mises en place reposent sur les données des bases de construction et restent dépendantes de données passées ce qui peut conduire à une sous-estimation des évolutions futures du portefeuille.

10. La colonne "Borne Inf." (resp. "Borne Sup.") correspond en réalité aux bornes supérieures des différents intervalles de confiance (resp. inférieures). En effet, il s'agit du ratio du nombre de sorties observés sur le nombre de sorties modélisées et seul le dénominateur varie.

Chapitre IV

Quantification de l'erreur sur les *Best Estimates*

Introduction : Après avoir présenté le fonctionnement du modèle central et quantifier l'impact des principales sources d'incertitudes (qualité des données, hétérogénéité du portefeuille et dérive temporelle) sur les taux bruts des différentes lois, il convient d'en estimer l'impact sur le niveau de provisionnement prudentiel.

Il s'agit d'une partie récapitulant les différents *Best Estimates* obtenus en utilisant les hypothèses recalculées au travers des différentes études. Plus précisément, on s'intéresse ici aux conséquences d'une erreur d'appréciation sur les taux de sortie en termes de niveau des provisions. Ces taux de sortie concernent la mortalité et la résiliation.

Nous allons voir que l'impact est différent selon l'hypothèse utilisée ainsi que le réseau de distribution considéré. Par ailleurs, nous allons essayer de comprendre comment ces différentes sources d'incertitudes interagissent entre elles.

IV.1 Formalisation et présentation de la démarche

Cette dernière étape consiste à utiliser les nouvelles hypothèses construites en données d'entrée du modèle central et à analyser les écarts entre les *Best Estimates* obtenus.

IV.1.1 Formalisation des écarts

Cette étape revient à calculer des sensibilités. La sensibilité, correspondant à la colonne "Evolution" dans les figures suivantes, est obtenue en modifiant les hypothèses de projection. Elle correspond à la formule :

$$Evolution \ = \ \frac{BE_{Total}^{ALTER} - BE_{Total}^{Central}}{BE_{Total}^{Central}}$$

où $BE_{Total}^{Central}$ représente le *Best Estimate* de référence ;

et BE_{Total}^{ALTER} représente un *Best Estimate* alternatif, c'est-à-dire recalculé.

IV.1.2 Obtention des hypothèses alternatives à partir des hypothèses existantes

La construction des lois d'expérience se fait en deux étapes : il s'agit d'abord d'estimer les taux de sortie par l'estimateur de Kaplan-Meier pour ensuite réaliser un ajustement non paramétrique. Cet ajustement consiste à lisser les taux bruts et à les prolonger sur les âges avec une exposition faible. Nous allons montrer comment les hypothèses alternatives ont été reconstruites à partir des taux bruts alternatifs et des hypothèses existantes.

Les différentes études ayant été menées en reculant les taux bruts, nous allons utiliser la relation suivante pour obtenir hypothèses alternatives. Pour un âge x fixé, on utilise la relation suivante :

$$\widehat{q}_x^{HYPO\ ALTER}(t) \ = \ \hat{q}_x^{HYPO}(t) \ \times \ \frac{\hat{q}_x^{BRUT\ ALTER}(t)}{\hat{q}_x^{BRUT}(t)}$$

où $\widehat{q}_x^{HYPO}(t)$ représente l'hypothèse initialement utilisée en entrée du modèle de projection ;
et où $\widehat{q}_x^{BRUT\ ALTER}(t)$ représente les loi calculées lors des différentes études.

Remarque : en raison d'un manque de temps, cette méthode repose sur une approximation que l'on peut facilement remettre en question [7]. En effet, les taux bruts présentent des irrégularités. Ces irrégularités résultent des fluctuations d'échantillonnage du fait d'une exposition variable selon les âges. Aussi, il est légitime de penser que la loi non-ajustée ne reflète pas la véritable loi de survie sous-jacente que l'on souhaite estimer. C'est d'autant plus le cas pour le risque de mortalité où les taux de sortie sont théoriquement croissants, augmentant avec le vieillissement.

Au lieu de raisonner en termes d'abattement sur les taux utilisés en hypothèse pour chacun des âges, il aurait été plus pertinent de raisonner en termes de taux d'abattement sur les taux au global ou par tranche d'âge. Cela aurait permis d'éviter des fluctuations d'échantillonnage et de représenter plus fidèlement la loi de survie que l'on veut estimer. Aussi, les différentes hypothèses alternatives conservent une variabilité non-souhaitable dans les taux de sortie estimés.

IV.1.3 Utilisation des différentes études

Voici un rappel des principales sources d'erreur intervenant dans la construction des taux bruts :

Source d'incertitude	Description	Etude mise en place
Fluctuations statistiques	Non robustesse des taux bruts induite par la taille réduite de l'échantillon	Utilisation des intervalles de confiances asymptotiques à 95% et des bornes de lissage associées
Qualité des données	Biais dans l'exactitude des résultats induit par la présence de données anormales	Procédure de détection d'anomalies reposant sur une réduction de dimension et l'utilisation combinée d'algorithmes de détections
Fluctuations d'échantillonnage	Mauvaise prise en compte de l'hétérogénéité du portefeuille	Procédure de ré-échantillonnage *bootstrap* selon l'appartenance à un regroupement de catégories socio-professionnelles
Dérive temporelle	Evolution de la sinistralité entre la période d'observation et la période de projection	Procédure de ré-échantillonnage *bootstrap* selon le trimestre d'observation

Table IV.1: Etude des principales sources d'erreurs dans le cadre d'une estimation non-paramétrique

Par ailleurs, dans le cadre de la quantification de l'impact des hypothèses alternatives sur le montant des *Best Estimates*, il est possible d'agir de deux façons : modifier les données de construction et/ou reconstruire les taux bruts à partir d'un ré-échantillonnage spécifique. Ces leviers correspondent aux études déjà menées :

- L'identification des lignes anormales des bases de projection n'étant pas pertinente, on cherche à quantifier l'incertitude de la meilleure estimation uniquement par rapport à l'utilisation des données de construction des hypothèses. Pour cela, nous sommes partis des bases servant à la construction des taux bruts des lois de résiliation et de mortalité. Les différents *Best Estimates* ont été calculés avec les hypothèses reconstruites avec des bases de construction alternatives, c'est–à-dire sans les anomalies identifiées au chapitre II.

- Outre l'erreur statistique, nous avons isolé deux sources principales d'incertitude intervenant au cours de l'estimation non-paramétrique des taux bruts : l'hétérogénéité du portefeuille et le choix de l'historique du portefeuille. Aussi, nous reprenons les taux bruts recalculés lors du chapitre III.

Enfin, les différentes études interviennent au cours du calcul des *Best Estimates* alternatifs selon le schéma suivant :

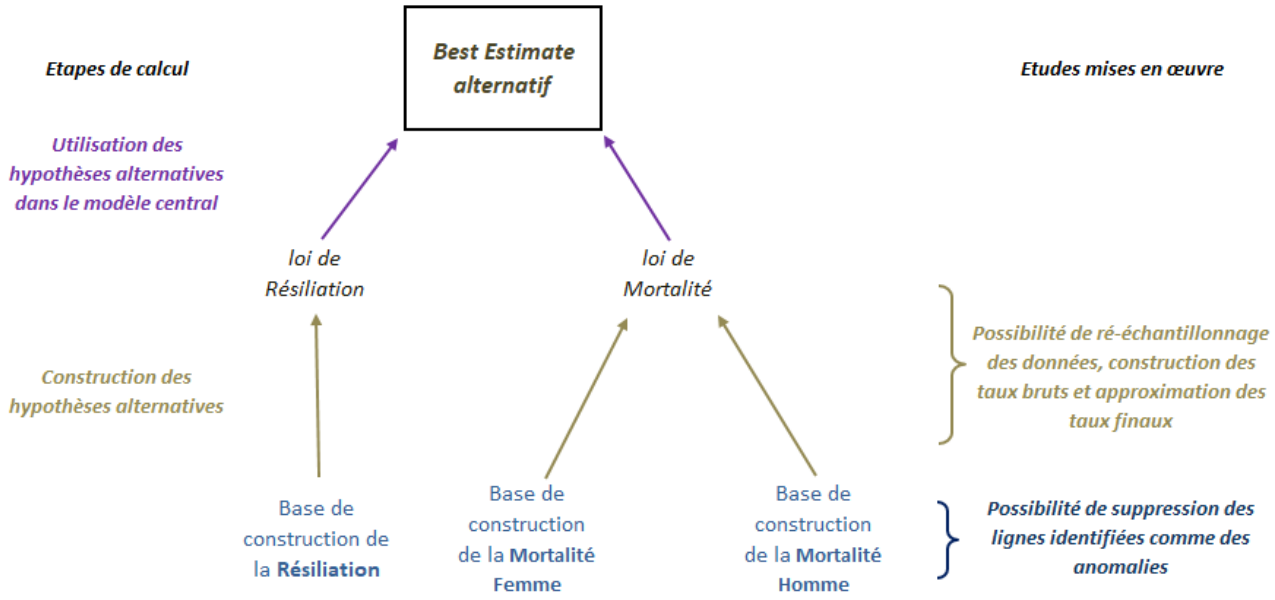


FIGURE IV.1: Procédure de calcul des BE alternatifs

IV.2 Incertitude générée par la qualité des données de construction

L'idée ici est d'isoler les erreurs générées par le modèle central de celles générées par les données et les hypothèses utilisées qui sont décrites dans les deux parties précédentes. Il s'agit d'utiliser les hypothèses de mortalité et/ou de résiliation reconstruites à partir de bases sans anomalies en entrée du modèle central.

IV.2.1 Sensibilité du *Best Estimate* à la mauvaise qualité des données

IV.2.1.1 Anomalies des bases de projection

Dans un premier temps, bien qu'identifier les lignes anormales du portefeuille à projeter et les retirer ne soit pas pertinent [1], il est tout de même utile de s'intéresser aux données anormales identifiées à partir des portefeuilles de projection. Cela consiste à projeter les individus du portefeuille duquel on souhaite quantifier les engagements sans les têtes anormales. Voici les résultats en termes d'impact sur les *Best Estimate* avec, pour rappel, le nombre d'anomalies détectées :

Réseau de distribution	Taille des bases	Anomalies Identifiées	% d'anomalies détectées	BE contenant des anomalies	BE sans anomalies	% du BE anormal
Réseau 1	1 147 239	3129	0,27%	248	246	1,3%
Réseau 2	803 956	2071	0,26%	19	17	10,3%

Table IV.2: Impact des données de projections anormales

Ce pourcentage du *Best Estimate* anormal représente la part des engagements qui est sujette à caution concernant l'utilisation des données de projection. Si l'on compare les deux réseaux, les deux portefeuilles contiennent une même proportion de têtes anormales. Néanmoins, l'impact sur le *Best Estimate* est 10 fois plus important pour le réseau 2. Cela s'explique par des montants initialement très différents selon le réseau de distribution considéré et par une tendance à l'identification des têtes les plus coûteuses, car atypiques. Par ailleurs, les données de ce réseau proviennent d'un système d'information moins bien contrôlé ce qui renforce la nécessité d'un contrôle approfondi pour ce sous-portefeuille.

IV.2.1.2 Anomalies des bases de construction

Ensuite, nous allons visualiser l'impact des hypothèses alternatives sur le *Best Estimate*. Nous nous intéressons dans un premier temps aux *Best Estimates* modélisés en utilisant conjointement les deux hypothèses recalculées à partir des bases d'exposition sans anomalies. Les résultats, selon chaque poste, sont les suivants [2] :

1. Dans le cadre de la politique de provisionnement, chaque tête du portefeuille doit être pris en compte dans le calcul du *Best Estimate*.
2. Pour des raisons de confidentialité, l'échelle relative aux montants a été modifiée.

Anomalies Hypothèses	Réseau 1				Réseau 2			
	Résultat	Central	Écarts	Évol.	Résultat	Central	Écarts	Évol.
BE PRIMES	-38	-37	-1,6	-4,1%	-33	-33	-0,4	-1,2%
Primes HT	-553	-553	0,0	0,0%	-201	-201	0,0	0,0%
Comm sur primes	72	72	0,0	0,0%	30	30	0,0	0,0%
Frais sur primes	49	49	0,0	0,0%	18	18	0,0	0,0%
Sinistres futurs	193	195	-2,1	-1,1%	51	51	-0,5	-1,1%
Frais sur sinistres futurs	36	36	-0,4	-1,1%	9	9	-0,1	-1,1%
Commissions sur resultat	165	164	0,9	0,5%	60	60	0,2	0,4%
BE SINISTRES	286	286	0,0	0,0%	52	52	0,0	0,0%
Sinistres connus	212	212	0,0	0,0%	38	38	0,0	0,0%
Sinistres inconnus	20	20	0,0	0,0%	5	5	0,0	0,0%
Frais sur sinistres connus	39	39	0,0	0,0%	7	7	0,0	0,0%
Frais sur sinistres inconnus	4	4	0,0	0,0%	1	1	0,0	0,0%
Commissions sur resultat	12	12	0,0	0,0%	1	1	0,0	0,0%
BE TOTAL	248	250	-1,6	-0,6%	19	19	-0,4	-2,2%

FIGURE IV.2: *Best Estimates* recalculés à partir des hypothèses construites sans les anomalies

Ce tableau nous permet de tirer plusieurs constats que nous allons approfondir dans ce qui suit.

IV.2.2 Différence d'impact selon l'hypothèse considérée

Selon la figure ci-dessus, on constate qu'une transformation des hypothèses se répercute uniquement sur le *Best Estimate* de primes. En effet, ce dernier correspond à la période de garantie non encore écoulée et vise à couvrir le risque que les primes futures soient insuffisantes par rapport au montant des prestations pour sinistres futurs. Il est le seul concerné par une modification des hypothèses de mortalité et de résiliation.

Plus précisément, seuls les flux dépendant des sinistres futurs sont modifiés. Dans le cadre de la garantie "décès toutes causes", ces sinistres représentent les prestations associées aux décès survenus au cours de l'année de projection. Aussi, on peut supposer que seule l'hypothèse de mortalité a un impact. Pour vérifier, distinguons l'impact des hypothèses reconstruites sans anomalies :

BE Total	Réseau 1		Réseau 2	
	Résultat	Évolution	Résultat	Évolution
Anomalies Résiliation	24 969 940	-0,028%	1 900 218	-0,089%
Anomalies Mortalité	24 824 924	-0,608%	1 862 398	-2,077%
Anomalies Hypothèses	24 817 237	-0,639%	1 860 577	-2,173%

FIGURE IV.3: BE sans anomalie selon l'hypothèse alternative utilisée

On observe le *Best Estimate* est très peu sensible à la loi de résiliation. Le *Best Estimate* a évolué de moins de 1% quel que soit le réseau considéré. Observons l'impact d'un choc homogène de -10% sur les taux de résiliations utilisés en hypothèse :

Résiliation -10%	Réseau 1				Réseau 2			
	Résultat	Central	Écarts	Évol.	Résultat	Central	Écarts	Évol.
BE PRIMES	**-37**	***-37***	**-0,4**	**-1,1%**	**-33**	***-33***	**-0,1**	**-0,4%**
Primes HT	-555	*-553*	-2,2	-0,4%	-202	*-201*	-0,8	-0,4%
Comm sur primes	72	*72*	0,3	0,4%	30	*30*	0,1	0,4%
Frais sur primes	49	*49*	0,2	0,4%	18	*18*	0,1	0,4%
Sinistres futurs	196	*195*	0,7	0,4%	52	*51*	0,2	0,4%
Frais sur sinistres futurs	36	*36*	0,1	0,4%	10	*9*	0,0	0,4%
Commissions sur resultat	164	*164*	0,5	0,3%	60	*60*	0,2	0,3%
BE SINISTRES	**286**	***286***	**0,0**	**0,0%**	**52**	***52***	**0,0**	**0,0%**
Sinistres connus	212	*212*	0,0	0,0%	38	*38*	0,0	0,0%
Sinistres inconnus	20	*20*	0,0	0,0%	5	*5*	0,0	0,0%
Frais sur sinistres connus	39	*39*	0,0	0,0%	7	*7*	0,0	0,0%
Frais sur sinistres inconnus	4	*4*	0,0	0,0%	1	*1*	0,0	0,0%
Commissions sur resultat	12	*12*	0,0	0,0%	1	*1*	0,0	0,0%
BE TOTAL	**249**	***250***	**-0,4**	**-0,2%**	**19**	***19***	**-0,1**	**-0,8%**

FIGURE IV.4: Sensibilité du *Best Estimate* pour une baisse homogène des taux de résiliation de 10%

On observe que l'hypothèse de résiliation modifie à la marge le *Best Estimate*. Seuls les flux dépendant du montant de primes futures (commissions sur primes, frais généraux sur primes, commission sur résultat du BE de primes) ainsi que sur les flux dépendant des sinistres futurs varient dans une moindre proportion.

A l'inverse, le décès faisant l'objet de la garantie "décès toutes causes", l'impact est beaucoup plus important sur le montant des sinistres futurs.

IV.2.3 Ecart d'évolution entre les deux réseaux

En s'intéressant à la différence d'échelle entre les montants, on constate que le *Best Estimate* du réseau 1 est 13 fois plus important. Concernant l'évolution globale, le *Best Estimate* du réseau 2 baisse de $2,2\%$. Il parait plus sensible à la détection des données de constructions anormales que le *Best Estimate* du réseau 1 qui lui baisse de $0,6\%$.

Néanmoins, pour comparer les réseaux de distribution, il est nécessaire de s'intéresser au *Best Estimate* de primes, car celui-ci est le seul à être modifié par un changement d'hypothèse de mortalité et de résiliation. On observe une évolution de $-4,1\%$ pour le réseau 1 et de $-1,2\%$ pour le réseau 2. C'est pourquoi, le réseau 1 est plus sensible aux résultats de la détection d'anomalies que le réseau 2.

IV.2.4 Confortation des résultats

Tout d'abord les évolutions étant négatives, on peut constater que la prise en compte de la mauvaise qualité des données révèle globalement une surestimation des engagements. Essayons de comprendre pourquoi.

Les algorithmes non supervisés de détection d'anomalies tentent de distinguer les observations aberrantes en apprenant sur l'ensemble des données du portefeuille considéré. Aussi, les données identifiées comme les plus anormales sont celles dont les caractéristiques sont les moins fréquentes. C'est le cas pour les lignes représentant des sorties non censurées (III.1.3.2).

Pour comprendre, observons la proportion d'anomalies détectées selon le type d'observation [3] :

Indicateur	**Décès Femmes**	**Décès Hommes**	**Rési -liation**
Taille	175 220	231 099	827 613
Nombre d'anomalies identifiées	4230	3772	4815
Nombre de sorties anormales	176	178	751
Nombre de censures anormales	4054	3594	4064
Proportion des sorties anormales par rapport au nombre total de sorties	19,7%	7,2%	0,5%
Proportion des censures anormales par rapport au nombre total de censures	2,3%	1,6%	0,6%

Pour la loi de mortalité, davantage de sorties anormales ont été détectées, ce qui a tendance à diminuer les taux bruts de sortie et par conséquent les taux ajutés. Observons ensuite les écarts par âge entre les hypothèses initiales et les hypothèses reconstruites sans anomalies :

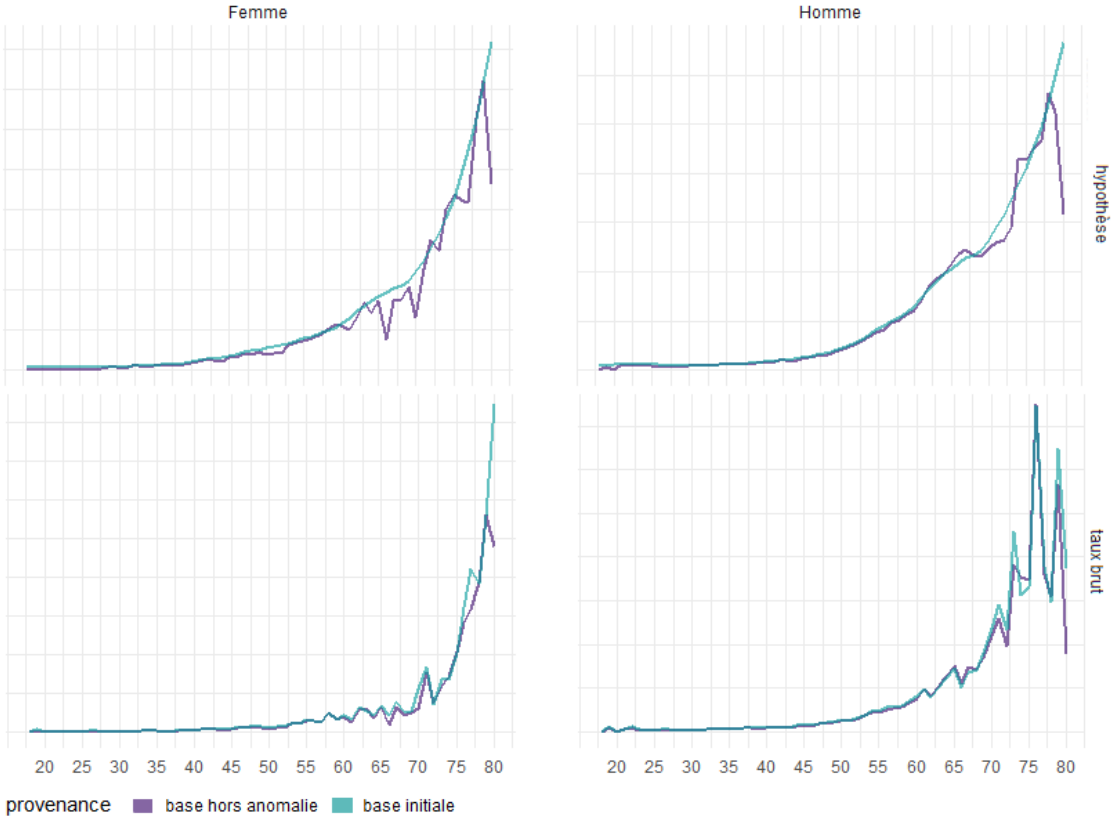


FIGURE IV.5: Taux de mortalité bruts et finaux selon la base de construction utilisée

3. Pour rappel, on appelle censure, l'opposé d'une sortie non censurée, c'est-à-dire un individu dont on n'a pas pu observer la sortie. Cette sortie est spécifique à la loi considérée (mortalité ou résiliation).

On observe globalement que l'hypothèse construite à partir des portefeuilles sans anomalies est plus faible que l'hypothèse initiale. Néanmoins, pour certains âges élevés, la loi alternative dépasse l'hypothèse initiale alors que ce n'est pas le cas pour les taux bruts. Cela est dû aux fortes fluctuations des taux de mortalité utilisés en tant qu'hypothèse alternative et s'explique par l'approximation utilisée dans la construction des taux finaux. Ce problème a été expliqué au paragraphe IV.1.2.

Ainsi, les résultats relatifs à l'évaluation des différents risques se retrouvent dégradés du fait de la mauvaise qualité des données. L'incertitude induite par les données de construction utilisées conduit à une surestimation des engagements.

Par ailleurs, l'utilisation de la loi de résiliation reconstruite sans anomalie a très peu d'impact par rapport à la loi de mortalité. Le réseau 1 est relativement plus sensible aux hypothèses utilisées. Cette sous-estimation est spécifique aux *Best Estimates* de primes et est surtout générée par la construction de la loi de mortalité. Ces constats restent vrais pour la suite de l'analyse.

IV.3 Incertitude générée par l'utilisation des hypothèses

L'objectif du chapitre III est de mesurer quel biais du *Best Estimate* peut-on attribuer aux hypothèses. On cherche à quantifier quelle part des engagements futurs vus à date de calcul est imputable à l'erreur de construction et d'utilisation des hypothèses.

IV.3.1 Rappel des éléments utilisés

En estimation non-paramétrique, le risque de mauvaise estimation des taux de sortie intervient lors de l'estimation des taux bruts. Les études du chapitre II nous ont permis d'obtenir des quantiles empiriques pour chaque âge ainsi que des chroniques de taux alternatifs.

Concernant les bornes des intervalles de confiance, nous prenons à nouveau comme point de comparaison l'intervalle de confiance asymptotique. Comme déjà observé, la prise en compte des trimestres d'observation génère le plus d'incertitude quelle que soit la loi construite. En effet, les ratios associés à la dérive temporelle étaient plus éloignés par rapport aux autres ratios. A l'inverse, les ratios associés aux fluctuations d'échantillonnage sont moins dispersés. Les résultats en termes d'impact sur les *Best Estimates*, seront aussi présentés pour les bases obtenues sans anomalies.

IV.3.2 Impact des différentes études par hypothèses

En réutilisant les taux bruts ainsi que les intervalles de confiances calculés au chapitre III, nous obtenons des hypothèses alternatives qui sont utilisées conjointement lors du calcul du *Best Estimates*.

A titre d'abréviation, nous utilisons le terme "CSP" pour faire référence aux études menées sur l'hétérogénéité. Le terme "trimestriel" correspond à l'étude sur la dérive temporelle. Le terme "Intervalles Statistiques" correspond aux bornes de l'intervalle de confiance statistique à 95%. Enfin, les extrémités des segments verticaux correspondent aux bornes des intervalles de confiance empirique précédemment calculés. Les résultats, en termes d'impact sur les différents postes du *Best Estimate* sont les suivants :

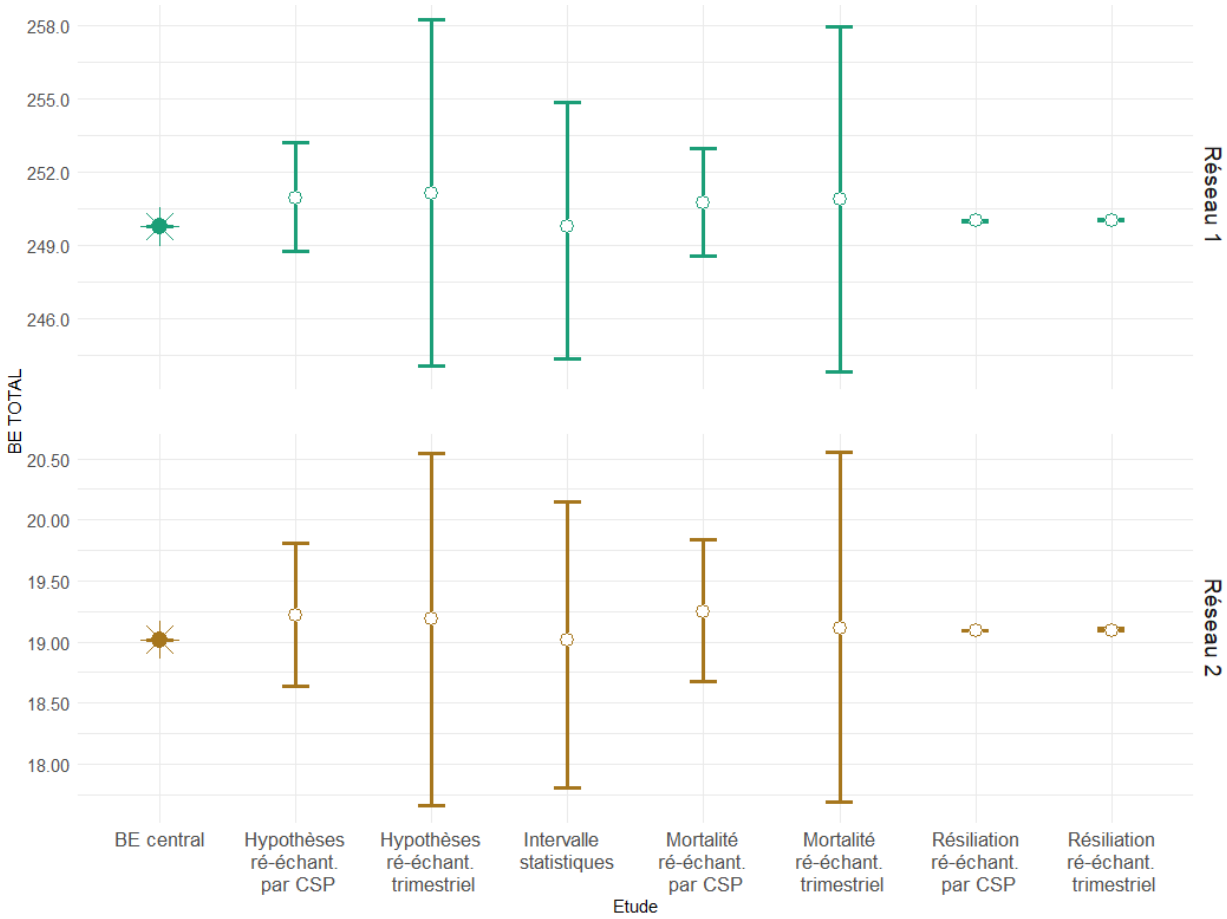


Figure IV.6: *Best Estimates* recalculés à partir des études sur les erreurs de construction d'hypothèses

A l'inverse des autres intervalles de confiances centrés sur les taux *bootstrap* obtenus par Monte-Carlo, les intervalles de confiance asymptotiques sont parfaitement centrés sur les taux initiaux. Néanmoins, pour chacun des réseaux, les *Best Estimates* centraux obtenus au travers des études basées sur du ré-échantillonnage ne sont pas trop éloignés du BE central.

Par ailleurs, on constate que l'impact sur les différents flux étant en cohérence avec les premiers résultats observés sur l'utilisation des hypothèses : la résiliation a très peu d'impact sur le *Best Estimate*. De plus, le biais induit par le choix de la période d'observation génère davantage de volatilité que l'hétérogénéité du portefeuille et la taille insuffisante des échantillons. Cela est cohérent avec les résultats observés au chapitre III.

Plus précisément, les résultats concernant l'utilisation des deux hypothèses reconstruites par ré-échantillonnage sont les suivants :

	BE Total			
	Réseau 1		Réseau 2	
	Résultat	Évolution	Résultat	Évolution
Hypothèses Stat sup	253,3	1,42%	19,7	3,84%
Hypothèses Stat inf	242,8	-2,78%	17,4	-8,47%
Hypothèses CSP	250,9	0,17%	19,3	0,90%
Hypothèses CSP sup	253,2	1,70%	19,9	4,70%
Hypothèses CSP inf	248,8	-0,40%	18,7	-1,46%
Hypothèses Trim	251,1	0,54%	19,2	0,90%
Hypothèses Trim sup	258,2	3,40%	20,6	8,54%
Hypothèses Trim inf	255,5	2,29%	17,8	-6,61%

FIGURE IV.7: *Best Estimate* recalculé à partir des études de ré-échantillonnage

IV.3.3 Impact croisé des différentes études et sources

Il s'agit de regarder l'impact des hypothèses alternatives en utilisant des données de construction sans anomalies. Cela revient à considérer conjointement plusieurs sources d'incertitude :

	BE Total Hors anomalies			
	Réseau 1		Réseau 2	
	Résultat	Évolution	Résultat	Évolution
Hypothèses Stat sup	254,8	2,03%	20,1	5,92%
Hypothèses Stat inf	244,3	-2,17%	17,8	-6,39%
Hypothèses CSP	249,4	-0,14%	18,9	-0,50%
Hypothèses CSP sup	251,7	0,77%	19,5	2,62%
Hypothèses CSP inf	247,2	-1,01%	18,3	-3,54%
Hypothèses Trim	249,6	-0,07%	18,8	-1,18%
Hypothèses Trim sup	256,7	2,79%	20,2	6,46%
Hypothèses Trim inf	242,5	-2,90%	17,4	-8,69%

FIGURE IV.8: *Best Estimate* recalculé à partir des études de ré-échantillonnage et de détection d'anomalie

Les taux négatifs obtenus s'expliquent en grande majorité par l'incertitude sur les données : l'utilisation des bases de construction sans anomalies diminue le nombre de sorties prises en compte dans l'estimation ce qui révèle globalement une surestimation des résultats en utilisant la loi initiale. On constate que la prise en compte des anomalies dans la construction des hypothèses correspond à une translation des écarts relatifs observés.

IV.4 Comparaison des sources d'incertitude

Ainsi, ce chapitre est concentré sur la quantification de l'impact des différentes hypothèses alternatives sur le montant des *Best Estimates*. Ces hypothèses alternatives sont obtenues à partir des différentes études et calculées avec des portefeuilles de construction avec et sans anomalies. Finalement, comparons les différents *Best Estimates* alternatifs afin d'identifier les sources d'erreurs les plus importantes :

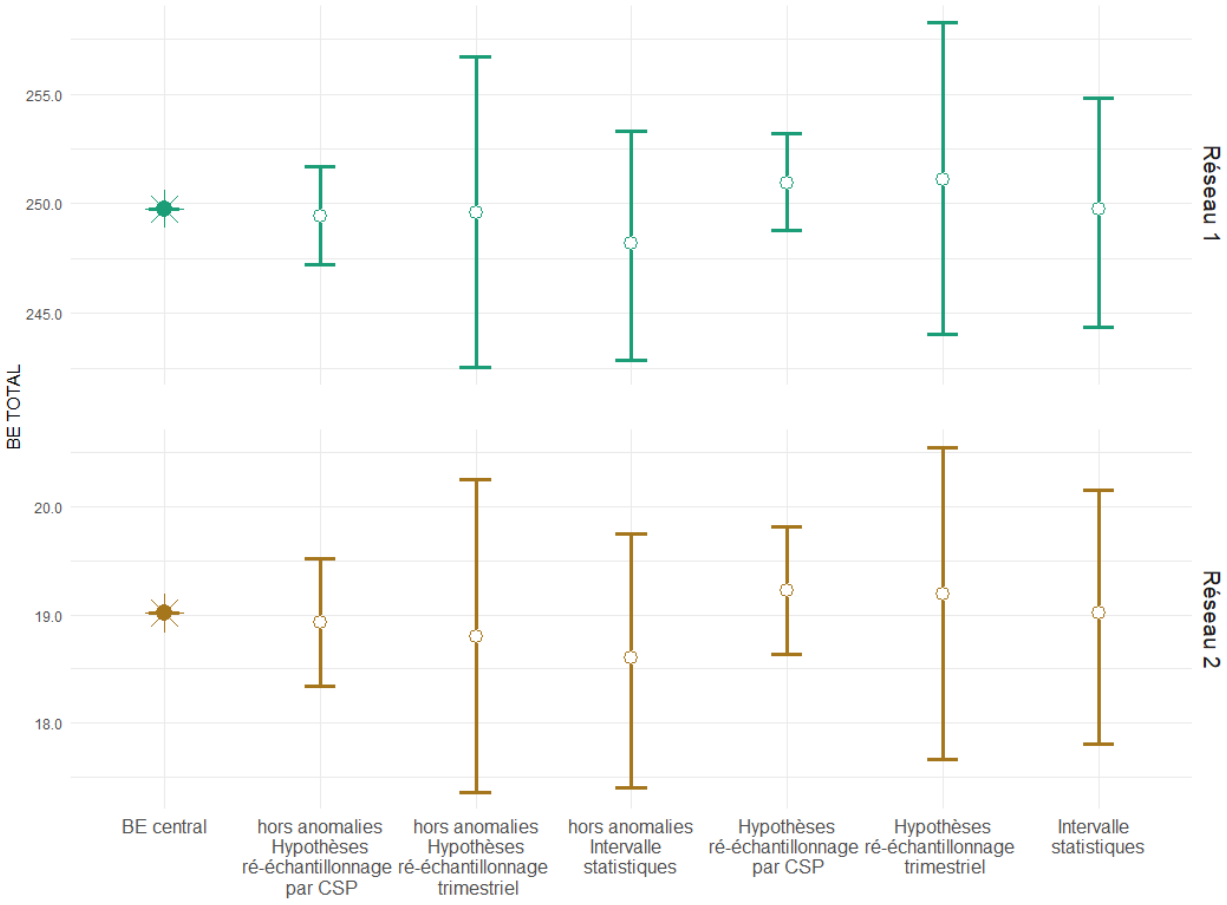


Figure IV.9: Synthèses des différents *Best Estimates* alternatifs selon chaque source d'incertitude

On observe que les *Best Estimates* calculés avec des données de construction sans anomalie correspondent à une translation des *Best Estimates* calculés à partir des données de construction initiales. L'intervalle de confiance statistique peut être considéré comme l'encadrement de référence du *Best Estimate*.

Ainsi, on retiendra l'encadrement le plus large du *Best Estimate* : celui obtenu via les hypothèses reconstruites sans anomalies et par ré-échantillonnage trimestriel. Il s'agit de l'étude ayant pour but d'évaluer la dérive temporelle conjointement à une mauvaise qualité des données.

Chapitre V

Mesure d'erreurs relatives au modèle de projection

Introduction : Le but d'un modèle de provisionnement est de prédire au plus juste l'espérance de valeur des engagements de la compagnie d'assurances. Après avoir présenté le fonctionnement du modèle central qui permet de projeter les flux futurs et provisions associés aux portefeuilles de contrats, il convient d'en évaluer la qualité.

L'objectif de cette troisième partie de l'étude est de mesurer quel biais du *Best Estimate* peut-on attribuer au modèle en tant qu'outil de calcul. On appelle "erreur de modèle", le risque dû à une mauvaise spécification du modèle par rapport à la réalité. On cherche à quantifier quelle part des engagements futurs vus à date de calcul est imputable à l'erreur du modèle. Plus précisément, on se demande quel est l'impact des erreurs du modèle sur le montant des provisions techniques prudentielles.

Au sein de la fonction actuarielle, l'évaluation de l'erreur de modèle peut être menée de différentes manières. Par exemple, on peut mettre en œuvre des analyses de sensibilité, une comparaison avec les résultats obtenus par d'autres méthodes dites "alternatives", des statistiques descriptives sur l'erreur d'estimation de paramètres de la méthode ou bien un *backtesting*.

Nous choisissons pour cette étude de réaliser un *backtesting*. Pour rappel, le portefeuille étudié concerne la garantie "décès toutes causes" et est distribué deux réseaux différents. L'idée ici est d'isoler les erreurs générées par le modèle central (permettant de calculer le *Best Estimate*) de celles générées par les données et les hypothèses utilisées qui sont décrites dans les deux parties précédentes.

Pour ce faire, nous allons essayer de quantifier l'écart entre les prévisions du modèle et la réalité.

V.1 Contextualisation et présentation de la démarche

Les provisions techniques exprimées en norme Solvabilité II doivent être déterminées sur la base d'hypothèses et de paramètres pertinents. Ceux-ci peuvent être estimés en fonction de l'expérience ou sur la base d'anticipations. Nous allons ici essayer de vérifier à quel point les provisions techniques obtenues par le modèle diffèrent de la réalité. Cette étape consiste en la réalisation d'un *backtesting*.

V.1.1 Définition et objectif du *backtesting*

Définition : le terme anglais *backtesting* signifie « test après coup » et consiste à évaluer la précision d'un modèle de prévision en comparant les résultats du modèle de projection aux données empiriques correspondantes. C'est une approche rétrospective qui se veut cohérente avec l'observé.

V.1.2 Contexte règlementaire

Cette analyse s'inscrit donc dans une démarche de suivi et de validation du calcul de la meilleure estimation et constitue une obligation règlementaire de la compagnie d'assurances.

La mise en œuvre de ces analyses est requise par les articles R354-6 ou R356-50 du Code des Assurances et par l'article 272 du Règlement Délégué 2015/35, point 4 : "Lorsqu'elle compare les meilleures estimations aux données tirées de l'expérience, la fonction actuarielle évalue la qualité des meilleures estimations établies dans le passé et exploite les enseignements tirés de cette évaluation pour améliorer la qualité des calculs en cours. La comparaison des meilleures estimations avec les données tirées de l'expérience inclut une comparaison des valeurs observées avec les estimations entrant dans le calcul de la meilleure estimation, afin que des conclusions puissent être tirées sur le caractère approprié, exact et complet des données et hypothèses utilisées ainsi que sur les méthodes employées pour les calculer".

Cela permet de détecter toute défaillance inhérente au provisionnement prudentiel et d'évaluer la précision de la méthode de provisionnement utilisée. En effet, en comparant les flux futurs espérés aux flux réalisés, le *backtesting* permet de conforter la pertinence des hypothèses. Les écarts significatifs qui ressortent de cet exercice sont ensuite analysés ce qui peut amener à remettre en question l'utilisation des hypothèses, voire des méthodes de projection.

Il s'agit donc de calculer un *Best Estimate* réel en se basant sur les flux observés.

V.1.3 Choix de l'historique utilisé

Pour rappel, le calcul du *Best Estimate* en prévoyance individuelle consiste à projeter chaque flux futur et provisions associés aux portefeuilles de contrats.

V.1.3.1 Flux futurs inclus dans la valorisation des provisions techniques

La notion de frontière des contrats permet d'identifier les engagements à considérer pour calculer les engagements et ceux à ne pas inclure. Dans le cadre de notre étude, elle délimite quels flux observés sont utilisés pour reconstituer des *Best Estimates* réels.

La frontière des contrats constitue une différence importante entre la norme comptable française actuelle et la norme Solvabilité II. En norme Solvabilité II, la notion de frontière des contrats précise que l'engagement doit être comptabilisé dès lors que l'assureur n'a plus la possibilité d'agir unilatéralement sur les termes du contrat.

En prévoyance individuelle, les contrats sont annuels à tacite reconduction, cette dernière étant réalisée en début d'année civile [20]. Cela signifie que les garanties sont couvertes durant un an maximum et que, chaque année, les assurés ont la possibilité de mettre un terme au contrat. On exclut donc les acquisitions postérieures à la date de projection. Concernant le renouvellement des contrats, on ne retient que le renouvellement certain. Cela correspond aux adhésions dont l'échéance annuelle coïncide avec la date de début de la période projetée.

V.1.3.2 Choix du *Best Estimate* de référence

Cette comparaison du réel par rapport à l'estimé nécessite d'avoir un recul suffisant. Plus précisément, parmi les données réelles utilisées, le montant des prestations constitue une part importante. Pour les prestations, l'objectif est de comparer l'estimation du coût définitif des sinistres avec les montants finaux réellement réglés. Aussi, il est nécessaire de positionner le calcul du *Best Estimate* suffisamment loin dans le passé afin que la quasi-totalité de la sinistralité soit clôturée, réglée et observée.

Cependant, utiliser des données associées à des règlements stabilisés présente une limite : en se positionnant trop loin dans le passé, les caractéristiques du portefeuille, en termes de risque, de pratiques de gestion et de sinistralité, peuvent être totalement différentes d'aujourd'hui. Les résultats ne seront alors pas généralisables aux années récentes alors que notre but est de mesurer le degré de maturité actuel du modèle central en termes de précision.

Aussi, pour éviter ce problème, nous choisissons de prendre les données les plus récentes possible : pour avoir une vision exhaustive de l'année N, la fonction actuarielle réalise ces analyses comparatives en N+1. Cela consiste à comparer la projection réalisée à la fin de l'année N-1, donc couvrant l'année N, avec les données réelles observées rétrospectivement. Cette idée correspond au schéma suivant :

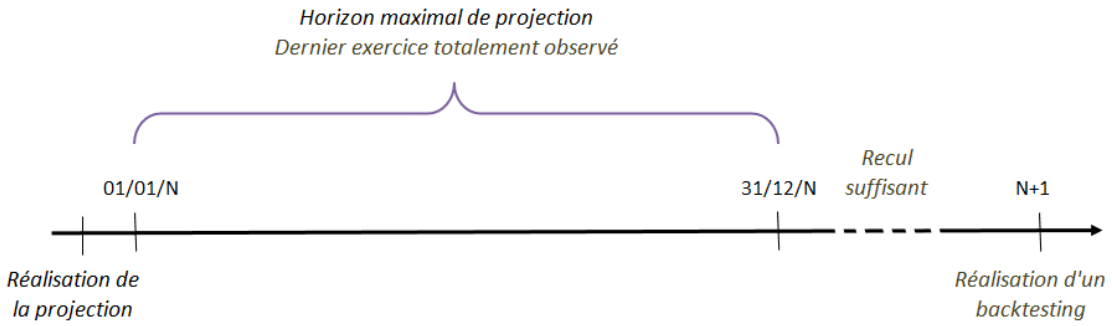


FIGURE V.1: Période d'observation

Dans notre cas, le dernier exercice totalement observé correspond à l'année $N = 2019$. Aussi, au moment où ce mémoire est rédigé (en 2020), nous disposons d'un recul temporel insuffisant pour valider le dernier exercice projeté au 31/12/2019 couvrant l'année comptable 2020. Le *Best Estimate* de référence est donc celui projeté au 31/12/2018.

V.1.3.3 Sources de données utilisées

Pour cette étude, nous choisissons de comparer la valeur estimée du *Best Estimate* projeté au 31/12/2018 et couvrant l'année comptable 2019 avec le BE dit « réel ». Le BE réel est obtenu à partir des données tirées de l'expérience. Nous utilisons donc principalement les bases de données suivantes :

- données comptables issues de l'arrêté des comptes au 31/12/2019 par risque, produit et par réseau de distribution. Elles incluent les provisions à constituer dans les comptes sociaux dont certaines seront définies plus tard ;
- données sinistres individuelles issues des fichiers de sinistralité bruts et cédés pour l'année comptable 2019 ;
- données de chiffre d'affaires par date d'échéance, c'est-à-dire la date à partir de laquelle les garanties du contrat ne sont plus actives. Ces données sont arrêtées au 31/12/2019 et permettent des estimations de chiffre d'affaires sans affaire nouvelle.
- montant des frais généraux, provenant du contrôle de gestion et données par entité, groupe homogène de risque ainsi que par ligne d'activité.

Ces bases de données sont délimitées afin de correspondre au périmètre de l'étude, c'est-à-dire le groupe homogène de risque « décès toutes causes » des réseaux de distributions considérés.

Toutefois, ces données d'expérience sont exprimées en normes comptables au 31/12/2019. Le provisionnement en normes sociales étant très prudent, utiliser ces données sans retraitement surestime *a priori* le *best estimate*. Aussi, évaluer la précision d'une méthode de prévision en utilisant simplement des données existantes n'est pas suffisant dans notre cas.

V.2 Passage d'un référentiel comptable à un référentiel prudentiel

Comme expliqué précédemment, pour mesurer la cohérence des provisions techniques avec le réel, il est nécessaire de retraiter les données tirées de l'expérience et exprimées en normes comptables afin de représenter au mieux les engagements de l'assureur *a posteriori*. Aussi, nous allons détailler, poste par poste du BE, comment cette traduction, de comptable à prudentielle est effectuée.

Aussi, les flux réels devant être retraités en vue de correspondre à une logique *Best Estimate*, nous allons préciser ce que nous entendons par là. Ensuite, après avoir explicité les différents flux tirés des données d'expérience, nous allons détailler leur utilisation à chaque retraitement.

V.2.1 Principes de valorisation du *Best Estimate*

Le référentiel Solvabilité II impose des exigences de calcul concernant le niveau de précision de la meilleure estimation des provisions techniques. Le calcul de l'espérance actualisée des flux futurs doit intégrer les points suivants :

Frontière des contrats : les flux projetés doivent appartenir au périmètre de contrats existant à date d'évaluation, c'est-à-dire à l'intérieur de la frontière du contrat

Projection en *run off* : jusqu'à extinction du portefeuille. Cela consiste à faire évoluer et vieillir au cours du temps, tous les contrats présents à date de projection. Aussi, on exclut des données utilisées les contrats relatifs aux nouvelles souscriptions et aux renouvellements postérieurs au 31/12/2018.

Valorisation économique : les engagements sont valorisés au montant pour lequel ils pourraient être transférés à une autre entité dans des conditions de concurrence normale.

Invariance du profil assuré : l'évaluation des flux futurs entrants et sortants est faite en considérant que les assurés du portefeuille ne vont pas changer au cours de la période de projection. Cela signifie que la proportion de décès, de résiliation et de frais reste stable. Ce principe légitime l'utilisation des hypothèses *best estimate*.

Remarque : Les provisions représentent une évaluation des engagements de l'assureur et sont en vision économique. Elles représentent une prévision à l'ultime du coût total des sinistres. À l'inverse, les provisions en normes comptables intègrent une vision prudente, c'est-à-dire qu'elles sont évaluées à l'aide d'hypothèses prudentes afin d'être suffisantes. Aussi, tenir compte de l'incertitude dans les flux de trésorerie futurs ne signifie pas qu'il faille inclure une marge de prudence dans les hypothèses de calcul. Au contraire, il convient d'effectuer la meilleure estimation, sans inclusion de marge de prudence.

Ainsi, pour correspondre aux exigences de calcul concernant les provisions techniques prudentielles et reconstituer les différents postes du *Best Estimate*, il est nécessaire d'effectuer des retraitements sur les flux réels.

V.2.2 Description des flux réels utilisés pour le backtesting

Afin de pouvoir expliquer les différents retraitements permettant d'obtenir une meilleure estimation des provisions techniques observées rétrospectivement, il est nécessaire de préciser quels les flux comptables sont utilisés. Par soucis de lisibilité, nous classons ces flux par type de retraitement.

V.2.2.1 Flux relatifs aux montants de primes

Primes hors-taxes : incluent les primes pures et les chargements de gestion ;

Commission sur prime : part de la prime permettant de rémunérer les réseaux de distribution ;

V.2.2.2 Flux et provisions relatifs aux montants de sinistres

Règlements : somme des prestations réglées au titre des sinistres survenus pendant l'exercice comptable concernant le risque "décès toutes causes". [1]

PSAP de clôture : provisions pour sinistres correspondant au paiement futur de tous les sinistres survenus et déclarés jusqu'à la fin de l'exercice. C'est la différence entre la charge finale estimée et les règlements effectués au 31/12/2019.

PSAP d'ouverture : provisions pour sinistres constituées au 01/01/2019 afin de couvrir les paiements à venir pour les sinistres déjà survenus, c'est-à-dire connus et inconnus.

PSNC : provisions pour sinistres non connus, c'est-à-dire pour les sinistres survenus en 2019 mais qui seront déclarés après le 31/12/2019. Cette provision est estimée au 31/12/2019 et correspond à l'estimation des sinistres qui sont déjà survenus, mais dont l'assureur n'a pas encore connaissance.

V.2.2.3 Flux relatifs aux montants de frais généraux

Les frais généraux observés sont théoriquement couverts par les chargements de gestions contenus dans les cotisations payées par les assurés. Ils représentent le coût additionnel engendré par la commercialisation et la gestion des contrats et sont décomposés comptablement selon leur destination :

Frais d'acquisition : commissions d'acquisition, frais liés aux réseaux de commercialisation, frais liés à l'établissement d'un contrat ;

Frais d'administration des contrats : encaissements des primes, comptabilisation, gestion de la relation client ;

Frais de prestation : de gestion des sinistres dont l'instruction des dossiers et la gestion des contentieux ;

Frais de placements : de gestion financière concernant l'analyse financière et les coûts de transaction.

Remarque : les frais d'acquisition découlent de l'émission d'un contrat. Dans un contexte de projection en *run off*, ces frais ne sont pas modélisés.

1. Les sinistres relatifs à des affaires nouvelles seront enlevés par la suite.

V.2.2.4 Composant de la commission sur résultat

Afin de déterminer la part du résultat technique à verser au réseau de distribution, il faut reconstituer l'assiette du résultat technique de l'assureur et donc définir la provision suivante :

Provisions d'égalisation d'ouverture (PE) : destiné à faire face aux fluctuations de sinistralité afférentes au risques décès et permettant d'égaliser les fluctuations des taux de sinistres ou de répondre à des dispositions contractuelles.

V.3 Application

Nous sommes en train de calculer le BE par approximation en utilisant des données d'expérience. L'objectif est d'observer la valeur que les différentes composantes du *Best Estimate* auraient prise si on avait observé ces dernières dans la réalité.

Aussi, après avoir détaillé les flux réels qui seront utilisés et expliqué la nécessité de correspondre aux hypothèses utilisées pour la valorisation des provisions techniques prudentielles, nous pouvons maintenant expliquer les différents retraitements mis en place pour obtenir le *Best Estimate* réel.

V.3.1 Principaux retraitements effectués

Le calcul du BE réel s'appuie sur les données issues de l'arrêté des comptes, retraitées pour simuler un portefeuille sans affaire nouvelle et une sinistralité à l'ultime. Ces retraitements sont communs à plusieurs postes du BE et la façon dont ils sont utilisés afin de correspondre au mieux à la valorisation *best estimate* est la suivante.

V.3.1.1 Retraitements relatifs aux montants de primes futures

Comme expliqué en première partie, la norme Solvabilité II requiert de prendre en compte les engagements induits par un contrat dès lors que la compagnie d'assurances n'a plus la possibilité d'agir unilatéralement sur les termes du contrat.

Pour estimer les primes sans affaire nouvelle, les primes perçues hors période de couverture sont retirées du chiffre d'affaires. Cela consiste à séparer les affaires nouvelles du stock dans les primes réelles. De plus, afin de ne pas prendre en compte les nouvelles acquisitions, on retire la part du chiffre d'affaires relative à l'année 2020, ce qui correspond aux $PPNA$ de clôture définis précédemment.

V.3.1.2 Retraitements relatifs aux montants de prestations

Nous allons expliquer, pas à pas, comment reconstituer les flux *best estimate* de sinistres futurs, connus et inconnus. Cela consiste à distinguer affecter les règlements et provisions selon le type de sinistre (connu, inconnu ou futur), à abaisser le montant des provisions en utilisant un taux de refus et à ajouter la provision pour sinistres inconnus.

1- Affectation des PSAP et règlements Tout d'abord, la charge ultime d'un sinistre est constituée des prestations déjà réglées et des prestations mises en provisions pour payer le reste de la charge. Les sinistres sont identifiés selon la logique suivante :

Futurs : la déclaration et la survenance des sinistres futurs étant observés après le 31/12/2018 pour notre étude, on obtient les flux $Reglements_{futurs}$ et $PSAP_{futurs}$ en agrégeant uniquement lorsque la date de survenance est supérieure au 31/12/2018.

Hors frontière : pour les sinistres futurs, la frontière des contrats est prise en compte afin d'inclure uniquement les engagements générés par le paiement des primes futures. On exclut les règlements relatifs aux sinistres survenus au moment où les garanties ne seraient plus actives si le contrat n'est pas renouvelé, soit après la date d'échéance du contrat.

Connus : la déclaration et la survenance des sinistres connus sont observées entre le 31/12/2018 et le 31/12/2019, aussi les $Reglements_{\ connus}$ et les $PSAP_{\ connus}$ s'obtiennent en considérant une date de survenance et de déclaration comprise entre le 31/12/2018 et le 31/12/2019.

Inconnus : pour les $Reglements_{\ inconnus}$ et les $PSAP_{\ inconnus}$, on considère uniquement les sinistres survenus entre le 01/01/2019 et le 31/12/2019 et déclarés après le 31/12/2019.

A titre illustratif, voici le nombre de sinistres survenus mensuellement pour chaque réseau de distribution :

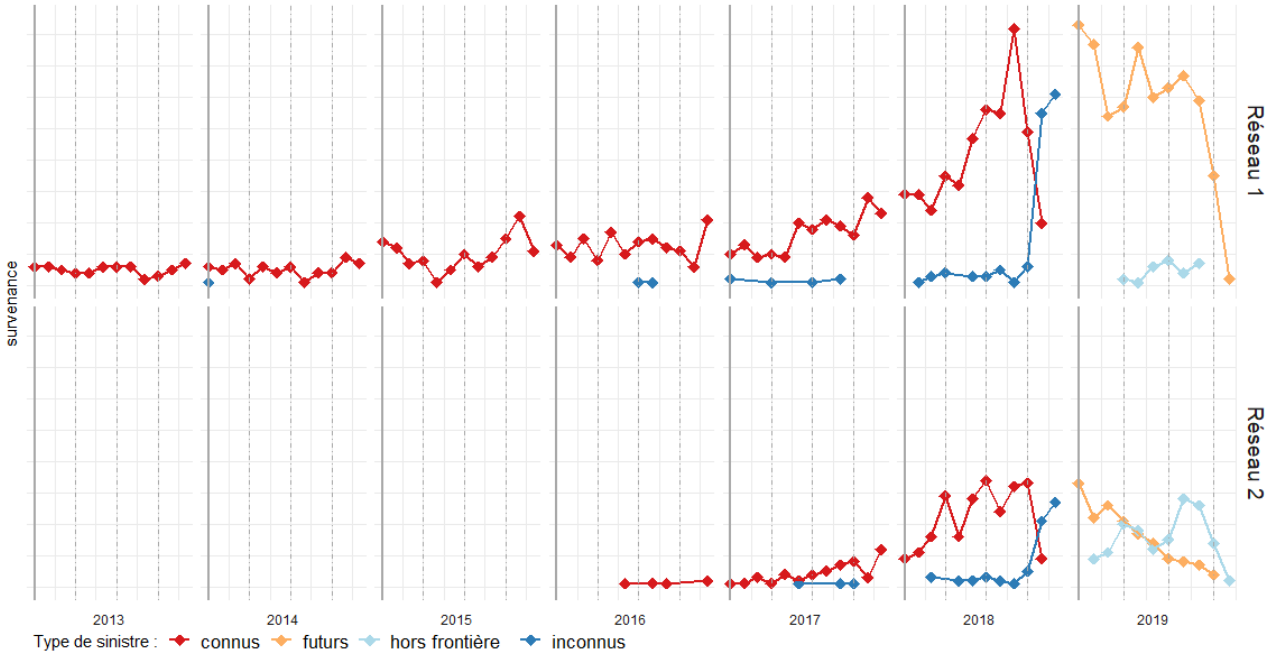


FIGURE V.2: Types de sinistres identifiés par rapport à la date de survenance

A partir des données individuelles de sinistralité, nous avons identifié le type de sinistre selon les critères définis plus haut. Les survenances des sinistres connus et inconnus appartiennent à la même période, la distinction se fait avec la date de déclaration. La distinction entre les sinistres futurs et hors frontière se fait en fonction de la date d'échéance du contrat.

Ainsi, afin de reconstituer les postes de sinistres, les *Reglements* ainsi que les $PSAP$ propres à chaque poste de sinistres du BE observés sont agrégés conditionnellement à la date de survenance, la date de déclaration des sinistres et les dates d'échéance.

2- Utilisation d'un taux de refus A la déclaration du sinistre, le montant provisionné par l'inventaire pour couvrir les règlements est souvent supérieur à la prestation réellement payée *in fine*. Voire, dans certains cas, le règlement de la prestation n'a pas lieu car, ne remplissant pas les termes du contrat, l'indemnisation est "refusée" par l'assureur.

Le montant de provisions, correspondant au paiement futur de tous les sinistres survenus et déclarés jusqu'à la fin de l'exercice, est abattu. Aussi, on utilise, une hypothèse de taux de refus. Ce $taux_{\ refus}$ représente la part de provision qui ne sera jamais consommée.

3- Ajout de la PSNC Enfin, pour approximer au mieux les sinistres, il faut ajouter les provisions constituées afin de couvrir les sinistres inconnus. Il s'agit de la PSNC constituée au 31/12/2019. Or, la PSNC porte sur l'évaluation de sinistres *a priori* déjà survenus mais dont la date de déclaration est ultérieure.

Aussi, on ventile cette provision, entre les sinistres futurs et les sinistres inconnus, selon le principe suivant : si le sinistre est déclaré avant le 31/12/2019 on considère que le sinistre est inconnu, s'il est déclaré après, il sera un sinistre futur. Pour cela, on utilise une hypothèse de liquidation par survenance, déterminée à partir de la cadence de règlement des prestations observées :

Total	N	$\leqslant N-1$
1	$Taux^{\ future}_{\ Liquidation}$	$Taux^{\ inconnus}_{\ Liquidation}$

Le taux associé à la première année de survenance permet de déterminer la part des inconnus estimés au 31/12/2019 à affecter au flux de sinistres futurs constitué au 31/12/2018. Cette partie de la provision pour sinistres non-connus représente finalement des sinistres futurs à date de projection car survenus après le 31/12/2018 et non déclarés. Ensuite, la proportion restante de la PSNC est affectée au flux de sinistres inconnus, c'est-à-dire survenus avant le 31/12/2018.

Après avoir expliqué les différentes étapes, les traitements relatifs aux flux de sinistres du BE peuvent être résumés de la façon suivante :

$$\begin{aligned}
Sinistres_{\ Futurs} &= Reglements_{\ futurs} && +PSAP_{\ futurs} \times taux_{\ refus} \\
& && +PSNC \times Taux^{\ inconnus}_{\ Liquidation} \\
Sinistres_{\ Connus} &= Reglements_{\ connus} && +PSAP_{connus} \times taux_{\ refus} \\
Sinistres_{\ Inconnus} &= Reglements_{\ inconnus} && +PSAP_{\ inconnus} \times taux_{\ refus} \\
& && +PSNC \times Taux^{\ inconnus}_{\ Liquidation}
\end{aligned}$$

V.3.1.3 Retraitements relatifs à la ventilation des frais généraux

En logique *Best Estimate*, les frais généraux d'acquisition ne sont pas inclus dans le calcul. Aussi, on regroupe les frais généraux en deux catégories : les frais généraux de prestations et les frais généraux de primes. Pour les frais généraux associés aux primes, on considère que les frais d'administration, de placements et les autres charges techniques :

$$Frais\ généraux\ de\ primes = administration + placement + autres\ charges\ techniques$$

Pour obtenir les frais généraux associés aux sinistres futurs, connus et inconnus, il suffit de ventiler les frais généraux de prestations à l'aide des postes du *Best Estimate* réel, c'est-à-dire en *prorata* des réglements estimés :

$$Fg_{\ prestations}^{\ futurs} = Fg_{\ prestations} \times \frac{Sinistres_{\ futurs}}{Sinistres_{\ futurs} + Sinistres_{\ connus} + Sinistres_{\ inconnus}}$$

$$Fg_{\ prestations}^{\ connus} = Fg_{\ prestations} \times \frac{Sinistres_{\ connus}}{Sinistres_{\ futurs} + Sinistres_{\ connus} + Sinistres_{\ inconnus}}$$

$$Fg_{\ prestations}^{\ inconnus} = Fg_{\ prestations} \times \frac{Sinistres_{\ inconnus}}{Sinistres_{\ futurs} + Sinistres_{\ connus} + Sinistres_{\ inconnus}}$$

V.3.1.4 Retraitements relatifs aux montants de commission sur résultats

Pour calculer les commissions sur résultats de primes et de sinistres, on applique le taux contractuel de commission sur résultats au résultat technique réel du portefeuille. il s'agit de reconstituer le résultat technique à partir de données comptables et de données retraitées :

$$\begin{aligned} Resultat_{\ prime} = Primes\ futures &+ PPNA_{\ ouverture} - Commissions\ sur\ primes \\ &- Chargements\ de\ gestions\ sans\ A.N. \\ &- Sinistres_{\ futurs} + PE_{\ ouverture} \end{aligned}$$

$$Resultat_{\ sinistres} = PSAP_{\ ouverture} - PSAP_{\ cloture} - Sinistres_{\ connus} - Sinistres_{\ inconnus}$$

Par ailleurs, pour obtenir les chargements de gestion sans affaire nouvelle on applique le taux de chargement de gestion aux primes futures :

$$chargements\ de\ gestion_{\ sans\ A.N.} = taux_{\ chargements\ de\ gestion} \times Primes\ futures$$

Remarque : la commission sur résultat du *Best Estimate* de sinistres est calculée quel que soit le signe du résultat. Néanmoins, si la somme des résultats des BE bruts de primes et de sinistres est négative, cette commission est nulle. Aussi, la commission sur résultat est présente dans les flux si la somme des deux résultats de BE est positive.

V.3.2 Analyse des résultats

Les résultats du modèle sur le périmètre de la prévoyance individuelle pour le risque "décès toutes causes" au 31/12/2018 ont ainsi été comparés aux données issues des comptes sociaux arrêtés au 31/12/2019, retraitées pour simuler une situation sans affaire nouvelle et une sinistralité à l'ultime.

Pour chacun des réseaux, les écarts entre la meilleure approximation des provisions techniques et la réalité sont les suivants. La colonne écart représente la variation entre le *Best Estimate* calculé via le modèle central au 31/12/2018 (noté Central) et le *Best Estimate* réel estimé à partir des flux comptables :

Backtest	Réseau 1				Réseau 2			
	Réel	Central	Écarts	Évol.	Réel	Central	Écarts	Évol.
BE PRIMES	**-85**	***-37***	**-48,4**	**-57%**	**-15**	***-33***	**17,9**	**121%**
Primes HT	-604	*-553*	-51,0	-8%	-109	*-201*	92,5	85%
Comm sur primes	78	*72*	6,1	8%	16	*30*	-13,9	-86%
Frais sur primes	56	*49*	7,2	13%	13	*18*	-4,9	-38%
Sinistres futurs	188	*195*	-7,2	-4%	27	*51*	-24,3	-90%
Frais sur sinistres futurs	9	*36*	-27,3	-309%	4	*9*	-5,8	-159%
Commissions sur resultat	188	*164*	23,8	13%	34	*60*	-25,6	-75%
BE SINISTRES	**273**	***286***	**-13,2**	**-5%**	**49**	***52***	**-2,8**	**-6%**
Sinistres connus	224	*212*	11,8	5%	34	*38*	-3,9	-11%
Sinistres inconnus	39	*20*	19,2	49%	8	*5*	3,6	43%
Frais sur sinistres connus	10	*39*	-28,5	-271%	5	*7*	-2,4	-51%
Frais sur sinistres inconnus	2	*4*	-1,9	-102%	1	*1*	0,2	21%
Commissions sur resultat	-2	*12*	-13,8	-726%	0	*1*	-0,4	-187%
BE TOTAL	**188**	***250***	**-61,6**	**-33%**	**34**	***19***	**15,1**	**44%**

Figure V.3: Résultats du *Backtesting*

On constate d'une part une surestimation des engagements pour le réseau 1 et une sous-estimation pour le réseau 2 d'autre part.

V.3.2.1 Explications pour le réseau 1

Pour le premier périmètre correspondant au réseau 1, cette vision prudente s'explique principalement par une sous-estimation des primes futures. L'écart sur les primes futures, qui se répercute sur les commissions sur primes ainsi que sur la commission sur résultat, peut s'expliquer par la raison suivante. La majeure partie du portefeuille se renouvelle en fin d'année et la réévaluation de la prime (retarification par âge) prenant effet au moment de la projection n'a pas été prise en compte. En effet, afin de tenir compte de l'éventuelle évolution annuelle de la prime, la prime annuelle retenue pour le calcul de la prime future doit être celle de l'échéance à venir au lieu de l'échéance passée. Le flux de primes futures, majoritairement composé de renouvellements ayant lieu en fin d'année, doit donc intégrer cette évolution tarifaire.

V.3.2.2 Explications pour le réseau 2

A propos du périmètre du réseau de distribution 2, l'écart de -33% entre la réalité et le *Best Estimate* modélisé est essentiellement dû au calcul des primes futures et à des hypothèses de projections mal adaptées. Concernant les primes futures, cette surestimation des primes réelle s'explique par une mauvaise prise en compte de la frontière des contrats. En effet, le renouvellement des contrats n'étant pas certain, seules les primes antérieures à la date d'échéance des contrats sont à comptabiliser. Aussi, le modèle de projection retenait à tort les primes futures perçues au-delà de la frontière des contrats.

Plus précisément, il s'agit de prendre en compte la spécificité du portefeuille concernant les dates d'effet et d'échéance ainsi que la périodicité de la prime lors de la projection.

Concernant les hypothèses de projection, il s'agit des taux servant au calcul des différents frais généraux, mais aussi des lois de mortalité et de résiliation. En effet, les hypothèses, dont la détermination s'appuie sur l'observation d'un historique, sont spécifiques au réseau 1. Ce choix s'explique par l'absence d'une profondeur d'historique suffisante pour le réseau 2.

Aussi, des différences éventuelles peuvent apparaître entre les portefeuilles. En effet, si l'on compare, pour chaque réseau de distribution, les survenances mensuelles de la base servant au *backtesting* avec celles de la base d'exposition servant à la construction de la loi de mortalité. A l'instar de la résiliation, il s'agit de l'hypothèse la plus importe dans le cadre de la garantie "décès toutes causes". Nous avons les résultats suivants :

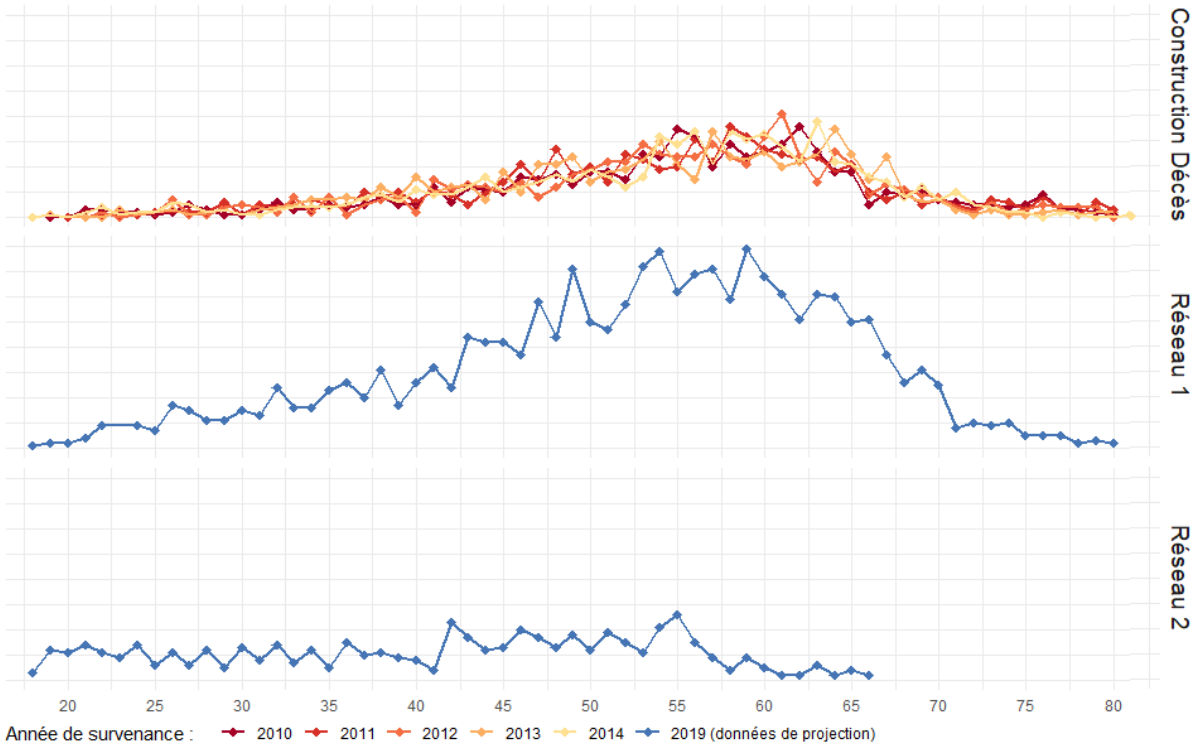


FIGURE V.4: Nombre de décès observés par âge selon l'année de survenance

Remarque pour la base de sinistralité utilisée lors du *backtesting*, nous avons exclu les sinistres n'appartenant pas à la frontière des contrats. De plus, ne disposant pas de la variable de genre dans la base de *backtesting*, nous avons agrégé les sinistres pour la base de construction seulement en fonction du mois d'exposition.

Tout cela nous permet de constater une augmentation du nombre de sinistres par rapport aux années de construction 2010-2014 pour le réseau 1, mais surtout de souligner l'inadéquation de l'hypothèse de mortalité avec le portefeuille du réseau 2. Les contrats étant commercialisés depuis le 01/01/2016, il s'agit d'un portefeuille avec une population plus jeune et en développement, auquel s'applique la sélection médicale. Ne disposant pas d'un historique suffisant, les hypothèses utilisées sont celles construites à partir des données associées au réseau 1.

V.3.3 Axes d'amélioration de la méthode de backtesting

Le calcul des *Best Estimates* réels s'appuie sur les données issues de l'arrêté des comptes, retraitées si nécessaire pour simuler un portefeuille sans affaire nouvelle et une sinistralité à l'ultime. Or, les retraitements effectués pour correspondre au mieux à une logique en valorisation économique introduisent forcément un biais. En effet, des approximations ont été faites pour estimer les montants de primes et de sinistres. Il s'agit de principalement de l'utilisation d'hypothèses calibrées sur des données antérieures telles qu'un taux de refus ainsi que des taux de ventilation de la $PSNC$.

De plus, ce *backtesting*, peut être amélioré avec un recul historique suffisant. Cela permettrait d'observer une sinistralité à l'ultime et de quantifier parfaitement les différents flux de prestations. Par ailleurs, ce recul suffisant permettrait de se passer des différentes hypothèses de retraitement utilisées. Au final, on obtiendrait un *backtesting* plus robuste car uniquement basé sur les flux comptables observés.

Ainsi, les différents écarts identifiés peuvent s'expliquer par l'imprécision du *Best Estimate* calculé, mais aussi par l'utilisation des hypothèses servant au retraitement des flux comptables. Cela explique la nécessité d'une prudence dans l'interprétation des résultats.

V.4 Conclusion partielle

Après avoir estimé l'erreur imputable à l'utilisation des données et à l'utilisation des hypothèses actuarielles, cette dernière partie nous permet de prendre du recul sur la notion de meilleure estimation. Cette partie repose sur un *backtesting* permettant de calculer des *Best Estimates* réels. Ce calcul s'appuie sur les données issues de l'arrêté des comptes, retraitées si nécessaire pour simuler un portefeuille sans affaire nouvelle et une sinistralité à l'ultime.

Au vu des résultats, on constate que, dans les deux portefeuilles, cette meilleure estimation n'est pas forcément proche de ce qui est observé dans la réalité. Néanmoins, il convient de prendre du recul sur le *backtesting*. En effet, les retraitements effectués pour correspondre au mieux à une logique en valorisation économique introduisent forcément un biais : l'utilisation des hypothèses et approximation afin de mieux coller à une valorisation prudentielle peut être aussi à l'origine des écarts observés. Par ailleurs, cette étude ne nous permet pas de quantifier des sources d'incertitude mais plutôt de souligner certaines limites quant à la modélisation mise en œuvre au moment du calcul des *Best Estimates*.

Aussi, ce dernier chapitre est proposé à titre exploratoire et devra faire l'objet d'améliorations dans le futur.

Conclusion

Dans le but d'évaluer les risques de primes et de réserves de la compagnie d'assurances, la fonction actuarielle garantit le caractère approprié des méthodologies, des modèles sous-jacents et des hypothèses utilisées pour le calcul des provisions techniques. Aussi, l'objectif de ce mémoire était d'expliquer, de la façon la plus complète, le niveau d'erreur que l'on peut associer au calcul de la meilleure estimation des provisions techniques. Dans ce cadre, ce mémoire s'attache à distinguer différentes sources d'erreurs en fonction des composants suivants :

- données d'entrée du modèle représentant le portefeuille projeté ;
- hypothèses dites *best estimate* construites et utilisées par le modèle central ;
- modèle de projection en tant qu'outil imparfait.

Cela revient à isoler les erreurs générées par le modèle central de celles générées par les données et les hypothèses utilisées par le modèle de projection. Cette démarche a été appliquée à un portefeuille de contrats auxquels s'applique la garantie "décès toutes causes". Ce portefeuille est segmenté selon le réseau de distribution.

Plus précisément, l'utilisation de données incorrectes engendre un biais intrinsèque dans l'exactitude des résultats. Aussi, nous avons commencé par proposer une méthodologie permettant de mesurer l'erreur associée à l'utilisation de données de mauvaise qualité. La procédure consiste à identifier les lignes anormales dans chacune des bases utilisées selon trois étapes successives : standardisation des données, réduction de dimension non-linéaire et utilisation combinée d'algorithmes de détection non supervisées.
Cette première étude nous a permis de visualiser les hypothèses de mortalité et de résiliation reconstruites à partir de bases sans anomalies et d'en quantifier l'impact sur le *Best Estimate*. On a vu que l'utilisation des bases de construction sans anomalies diminue le nombre de sorties prises en compte dans l'estimation. Cela signifie que la mauvaise qualité des données entraîne globalement une surestimation des engagements.

Dans un second temps, afin de quantifier l'erreur imputable à l'utilisation des lois d'expérience, nous avons proposé deux études basées sur du ré-échantillonnage *bootstrap*. En effet, outre l'incertitude générée par l'utilisation de données anormales, nous avons isolé trois sources principales d'incertitude lors de la construction d'hypothèses : les fluctuations statistiques, l'hétérogénéité de la population observée et le choix de la période d'observation du portefeuille. Les fluctuations statistiques sont quantifiées à l'aide des intervalles de confiance asymptotique et des bornes de lissage associées et ne font pas l'objet d'une procédure *bootstrap*. Le premier ré-échantillonnage a pour vocation de capter l'hétérogénéité des lois construites. Cette hétérogénéité a été représentée par l'appartenance à la catégorie socio-professionnelle. Le second ré-échantillonnage a pour but d'estimer les fluctuations temporelles induites par le choix des trimestres d'observation.
Ces études nous ont permis de reconstruire des hypothèses alternatives et d'en quantifier l'impact sur le *Best Estimate*. Nous avons observé que la dérive temporelle représentée par le choix de la période d'exposition est la source d'incertitude la plus importante. En raisonnant en termes d'impact sur le *Best Estimate* total, il s'agit de l'encadrement le plus large pour les deux réseaux de distribution.

L'objectif de cette troisième étape de l'étude est de mesurer quelle part d'erreur du *Best Estimate* peut-on attribuer au modèle en tant qu'outil de calcul. Nous avons essayé de quantifier quelle part des engagements futurs vus à date de calcul est imputable à l'erreur du modèle, c'est-à-dire le risque dû à une mauvaise spécification du modèle par rapport à la réalité. Nous avons choisi, pour cette dernière étude, de réaliser un *backtesting*. Cela revient à calculer des *Best Estimates* dits "réels" en s'appuyant sur les données issues de l'arrêté des comptes, retraitées si nécessaire pour simuler un portefeuille sans affaire nouvelle et une sinistralité à l'ultime.
Par ailleurs, ce chapitre nous a permis de prendre du recul sur la notion de meilleure estimation (*Best Estimate*). En effet, nous avons constaté, pour les deux réseaux de distribution, une sous-estimation ou une surestimation d'au moins 33% par rapport aux *Best Estimates* "réels". Ainsi, la meilleure estimation des engagements n'est pas forcément proche de ce qui est observé dans la réalité.

Par ailleurs, aux vues de la multiplicité des domaines d'études (modèles de durées, détection non supervisée d'anomalies, provisionnement prudentiel), ce fut un travail très enrichissant. Une des difficultés principales était de proposer une approche cohérente mêlant les différentes sources d'erreurs.

Néanmoins, la plupart des études proposées peuvent faire l'objet d'améliorations futures. C'est le cas concernant le chapitre sur les hypothèses actuarielles ainsi que le chapitre sur l'erreur de modèle. Outre l'appartenance à la CSP, il convient de se demander s'il n'y a pas d'autres sources d'hétérogénéité que l'on aurait pu prendre en compte en enrichissant les bases de données utilisées. On peut penser à la localisation géographique, à la sélection médicale, ou bien aux pratiques de vie (tabac, alcool). De plus, concernant la dérive temporelle, il aurait été pertinent d'utiliser d'autres années d'exposition plus récentes.
En ce qui concerne la mesure des écarts entre le *Best Estimate* et la réalité, les écarts observés ne sont pas assez satisfaisants pour servir de point de comparaison. En effet, ces écarts ne sont pas uniquement dus à une estimation inadaptée du *Best Estimate* mais aussi aux approximations mises en œuvre dans le cadre du *backtesting*. Enfin, dans le but d'isoler les erreurs générées par le modèle central de celles générées par les données et les hypothèses utilisées, il aurait été possible d'ajouter un chapitre sur l'agrégation de ces différentes erreurs et la possible corrélation entre ces dernières. Cela aurait permis d'avoir une mesure d'erreur de la meilleure estimation des provisions techniques globale.

Enfin, les travaux de ce mémoire ne constituent qu'un aspect très particulier de la mesure de l'erreur globale du *Best Estimate*. En effet, les différentes études sont spécifiques au portefeuille étudié, à la construction non-paramétrique des hypothèses et au modèle de provisionnement prudentiel permettant de valoriser les engagements associés à la garantie "décès toutes causes" de Natixis Assurances. Par ailleurs, ce mémoire constitue une première approche originale sur la mesure de l'erreur du *Best Estimate*. Cette approche sera, à terme, poursuivie en interne pour d'avoir une mesure cohérente et globale afin d'évaluer la qualité des estimations.

Annexe A

Anomalies détectées sur les autres bases

Hors la base de construction de la mortalité des femmes, nous présentons les résultats des réductions de dimensions. Les paramètres de l'algorithme de réduction sont les mêmes : nous avons choisi la distance euclidienne et fixé $n_{neighbors} = 0.01$ et $min_{dist} = 15$.

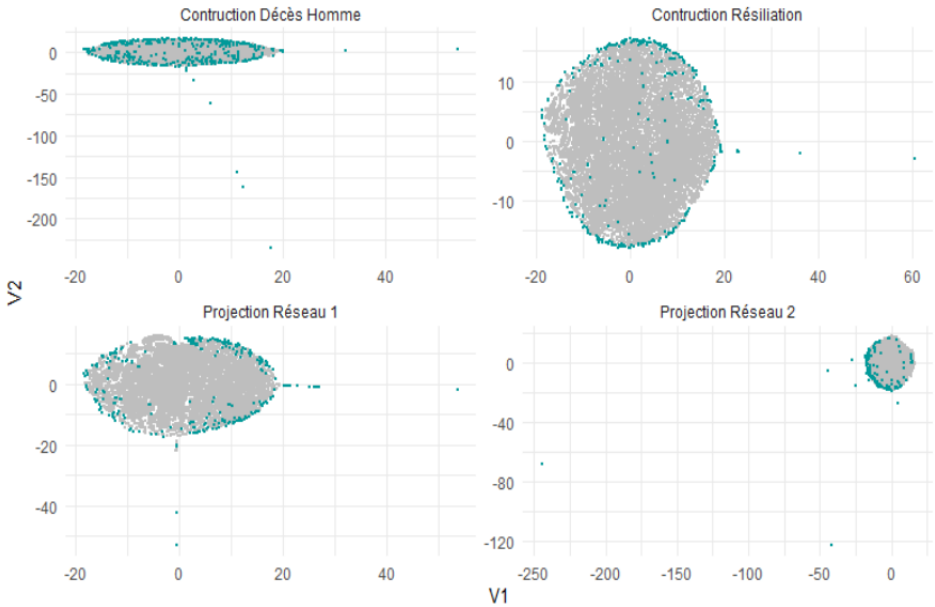


Figure A.1: Données réduites et points détectés au moins une fois

Les points bleus représentent les différentes anomalies détectées par au moins un des algorithmes KNN, LOF, $DBSCAN$ et IF.

Bibliographie

[1] Jéquier A. quantification erreur d'estimation construction d'une loi de maintien. Master's thesis, Diplôme Universitaire des Actuaires de Strasbourg, Institut des Actuaires, 2015.

[2] Kamega A. and Planchet F. Mesure du risque d'estimation associé à une table d'expérience. *Cahiers de recherche de l'ISFA*, (2136) :30, November 2010.

[3] ACPR. Analyse et synthèse : solvabilité 2, principaux enseignements de la cinquième étude quantitative d'impact (qis5). 2011.

[4] Charu C. Aggarwal. Outlier detection for high dimensional data. *In ACM Sigmod Record (Vol. 30, No. 2, pp. 37-46).*, 2001.

[5] Efron B. Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374) :312–319, 1981.

[6] Ng R. Breunig M., Kriegel H. and Sander. Lof : identifying density-based local outliers. pages 93–104, 2000.

[7] London D. *The revision of estimates.* Actex, 1995.

[8] Steinbach M. Kumar-V. Ertoz, L. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. *Proceedings of the 2003 SIAM international conference on data mining*, pages 47–58, 2003.

[9] McInnes L. et al. Umap, uniform manifold approximation and projection for dimension reduction. 2018.

[10] Togbe M. et al. Etude comparative des méthodes de détection d'anomalies. *Revue des Nouvelles Technologies de l'Information, Hermann, 2020, Extraction et Gestion des Connaissances EGC 2020.*, 2020.

[11] Planchet F. Mortalité, longévité, pandémie : Choix des hypothèses dans le cadre du pilier 1. 2012.

[12] Planchet F. Méthodes de lissage et d'ajustement. Technical Report Support de cours v4.1, Institut des Actuaires, 2018.

[13] Zhi-Hua Zhou Fei Tony Liu, Kaiming Ting. Isolation-based anomaly detection. 2012.

[14] Uchida S. Goldstein M. Comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. 2016.

[15] S. Hariri, M. Carrasco Kind, and R. J. Brunner. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2019.

[16] Karkkainen I. Hautamaki, V. and P. Franti. Outlier detection using k-nearest neighbour graph. 2004.

[17] Hoem J.M. Point estimation of forces of transition in demographic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 275–-289, 1971.

[18] Hyunjun Ju, Dongha Lee, Junyoung Hwang, Junghyun Namkung, and Hwanjo Yu. Pumad : Pu metric learning for anomaly detection. *Information Sciences*, 523 :167–183, 2020.

[19] E.L. Kaplan and P. Meier. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, (53) :457–481, 1958.

[20] Charrier M. Enjeux liés à la frontière des contrats sur les exigences quantitatives solvabilité ii en assurance emprunteur. Master's thesis, ISFA, mémoire d'actuariat, Institut des Actuaires, 2015.

[21] J. Sander M. Ester, H.-P. Kriegel and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, page 226–231, 1996.

[22] Guibert Q. *On the use of multi-state models to measure and manage the risks of an insurance contract*. Theses, Université Claude Bernard - Lyon I, 2015.

[23] V. Kumar V. Chandola, A. Banerjee. Anomaly detection : A survey. 2009.