Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 19 septembre 2024

Par : Paul VAUJANY
Titre : Mesure de l'impact de la fiabilité des données sur un modèle de provisionnement en assurance santé collective

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Alexis MERX
Signature :

Entreprise :
AXA France
Signature :

Membres présents du jury de l'EURIA :
Catherine RAINER
Signature :

Directeur de mémoire en entreprise :
Brahim JAMALEDDINE
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

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

Remerciements

Je souhaite tout d'abord exprimer ma gratitude à AXA, qui m'a permis de réaliser ce mémoire dans un environnement professionnel enrichissant. Un grand merci à Brahim Jamaleddine, tuteur entreprise de mon mémoire et responsable de l'équipe Data Quality et Maîtrise Technique, pour ses conseils précieux et son soutien tout au long de cette expérience. Je tiens également à remercier Sandra Virlouvet, membre de l'équipe et encadrante de mon alternance, pour avoir facilité mon intégration et mon épanouissement professionnel pendant toute ma période d'alternance. Je n'oublie pas Meroine Ifoudine, dont l'accompagnement technique et l'expertise ont été d'une grande aide. Enfin, merci à toute l'équipe Data Quality et Maîtrise Technique pour leur accueil chaleureux et leur disponibilité.

Je tiens ensuite à remercier mon tuteur académique, Monsieur Dominique Abgrall, professeur à l'EURIA, pour ses conseils avisés, ses suggestions constructives et son soutien tout au long de ce travail.

Je souhaite également exprimer ma reconnaissance envers l'EURo Institut d'Actuariat qui m'a permis d'acquérir les bases théoriques essentielles pour la réalisation de ce mémoire. Je remercie l'ensemble du corps professoral pour la qualité de leur enseignement et leur engagement.

Je n'oublie pas l'ESILV, qui a largement contribué à mon parcours académique et professionnel en me permettant de suivre un double diplôme avec l'EURIA, renforçant ainsi mes compétences en actuariat.

Enfin, je souhaite adresser un remerciement tout particulier à mes amis et à ma famille, pour leur soutien constant et leurs encouragements tout au long de mes études.

Résumé

La qualité des données est fondamentale dans le domaine de l'actuariat, en particulier dans le contexte des contrats d'assurance au sein des grandes compagnies comme AXA. La capacité à collecter, stocker, et utiliser des données de haute qualité est essentielle pour évaluer et anticiper les risques liés à ces contrats, ainsi que pour établir des modèles de provisionnement précis et fiables.

Ce mémoire se propose d'explorer en profondeur l'impact de la *data* qualité sur la performance et la justesse des modèles de provisionnement appliqués aux contrats santé collectifs. L'objectif est de mettre en lumière les défis spécifiques rencontrés dans la gestion de la donnée, ainsi que les pistes d'amélioration qui pourraient en découler. Une attention particulière sera portée à la détection de données anormales, l'idée étant de développer un modèle automatisé avec des méthodes permettant de repérer ces anomalies. En outre, il sera opportun d'identifier les phénomènes conjoncturels, tels que les pandémies, qui ne devraient pas être pris en compte lors du provisionnement. Ces approches de détection, par calculs statistiques ou par *machine learning*, seront testées au préalable et la plus adaptée sera retenue. Un mécanisme de lissage mathématique corrigera alors les anomalies repérées. Enfin, une application concrète sur plusieurs segments d'étude sera présentée, mettant en évidence un modèle de provisionnement appliqué aux données initiales et aux données retraitées. Le but sera d'illustrer de manière pratique les enjeux et les bénéfices ou coûts associés à l'optimisation de la qualité des données dans ce contexte spécifique.

Les recherches contribueront à l'enrichissement des connaissances dans le domaine de l'actuariat en intégrant les dimensions pratiques et stratégiques liées à la fiabilisation des données. Ce travail ambitionne de démontrer comment une meilleure gestion des données peut affiner les prévisions et la robustesse des modèles actuariels, tout en offrant des perspectives concrètes d'amélioration continue dans le secteur de l'assurance santé collective.

Mots clefs: *Data* qualité, détection d'anomalies, *machine learning*, phénomènes conjoncturels, arbre de décision, lissage mathématique, retraitement des données, provisionnement, IBNR (*Incurred But Not Reported*)

Abstract

Data quality is fundamental in the field of actuarial science, particularly in the context of insurance contracts within large companies such as AXA. The ability to collect, store, and use high-quality data is essential for assessing and anticipating the risks associated with these contracts, as well as for establishing accurate and reliable reserving models.

This thesis aims to explore in depth the impact of data quality on the performance and accuracy of reserving models applied to group health contracts. The objective is to highlight the specific challenges encountered in data management and the potential areas of improvement that may arise. Particular attention will be given to the detection of anomalies in the data, with the aim of developing an automated model using methods to identify these anomalies. Furthermore, it will be crucial to identify cyclical phenomena, such as pandemics, which should not be considered in reserving. These detection approaches, using statistical calculations or machine learning, will be tested beforehand, and the most suitable one will be selected. A mathematical smoothing mechanism will then correct the identified anomalies. Finally, a practical application across several study segments will be presented, highlighting a reserving model applied to both the initial and reprocessed data. The goal is to provide a practical illustration of the challenges and benefits or costs associated with optimizing data quality in this specific context.

The research will contribute to the enhancement of knowledge in the field of actuarial science by integrating both practical and strategic dimensions related to data reliability enhancement. This work aims to demonstrate how better data management can refine forecasts and the robustness of actuarial models, while offering concrete prospects for continuous improvement in the group health insurance sector.

Keywords: Data quality, anomaly detection, machine learning, cyclical phenomena, decision tree, mathematical smoothing, data reprocessing, reserving, IBNR (Incurred But Not Reported)

Note de Synthèse

Contexte et problématique

La fiabilité des données désigne l'exactitude et l'intégrité des informations utilisées pour la prise de décisions. Dans le domaine de l'assurance, elle est essentielle pour évaluer les risques, calculer les primes et établir des prévisions justes, garantissant ainsi la stabilité financière de l'entreprise et la confiance des assurés.

Ce mémoire a pour ambition d'expérimenter diverses approches pour fiabiliser des données de règlement de sinistres en assurance santé collective. Cette branche de l'assurance est en constante évolution, notamment avec la réforme 100% Santé et le transfert de charge de la Sécurité sociale vers les compagnies d'assurance et les mutuelles. Elle est aussi soumise à des marges financières étroites en raison de la forte concurrence, du contrôle des coûts et de la nature imprévisible des dépenses de santé. Les calculs de provisions doivent alors être précis et suffisamment fidèles aux risques associés.

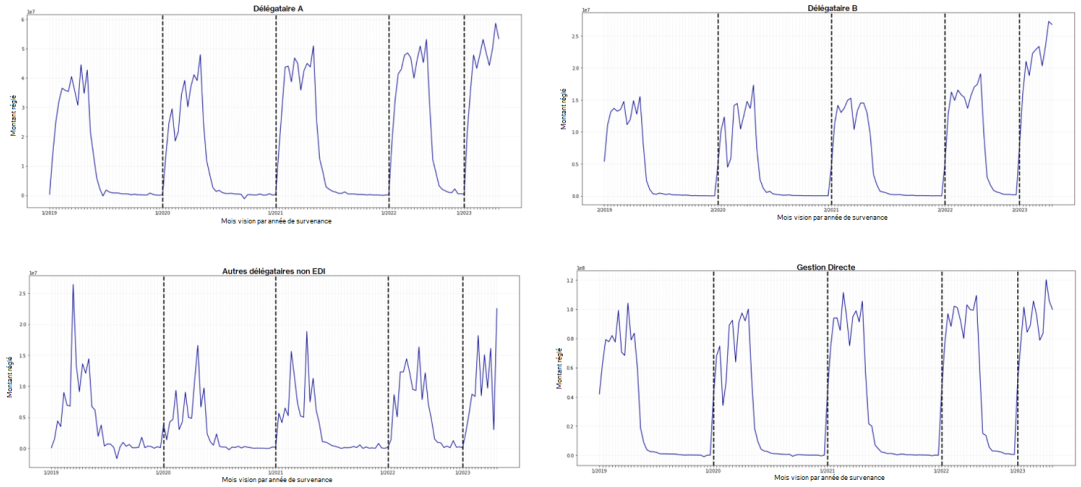
Pour assurer cette fidélité, les données utilisées dans les modèles de provisionnement sont tenues d'être représentatives des aléas en lien avec les sinistres en santé. Les processus humains et techniques, permettant la collecte, le stockage et la gestion des contrats, peuvent générer des données anormales et il est nécessaire de les corriger. Aussi, les compagnies d'assurance sont soumises à des exigences réglementaires strictes, ce qui renforce la nécessité d'utiliser des données cohérentes lors des estimations.

Différents algorithmes de détection d'anomalies ont été évalués, puis plusieurs types de correction par lissage ont été examinés afin de retraiter les valeurs anormales. L'application actuarielle du provisionnement sur plusieurs segments d'étude choisis au préalable a permis de conclure sur les conséquences de la fiabilisation des données.

Analyse des règlements de sinistre en santé

Les bases de données d'AXA Santé Collective relatives aux règlements des sinistres permettent d'analyser les flux concernant la garantie santé. L'étude technique porte sur quatre segments représentatifs du portefeuille d'assurance, en passant par la gestion directe ainsi que la gestion déléguée. Ces éléments couvrent environ 45% des règlements en santé et confirment la pertinence de cette étude.

Pour ces quatre exemples, il est possible de modéliser la trajectoire des règlements. L'historique des données concerne les cinq dernières années de survenance (2019 à 2023) et pour chacune de ces années, 36 mois de vision sont perçus, à l'exception des survenances 2022 et 2023. Ainsi, cette représentation illustre l'évolution des règlements en assurance santé collective, en tenant compte de la temporalité des sinistres et de leur règlement :



Des disparités entre ces courbes sont notables, bien que des valeurs plus contrastées soient identifiables au cours de la survenance 2020 pour l'ensemble de ces exemples. Désormais, le but est de repérer de manière automatisée les potentielles données incohérentes dans ces règlements.

Détection des anomalies

Plusieurs approches ont été examinées pour évaluer au mieux les anomalies dans les règlements de santé des quatre segments d'étude. Premièrement, un modèle statistique a été mis en place, fondé sur l'étude de la variance des données à l'aide d'un intervalle de confiance. Cet outil de référence rend possible l'estimation de la précision des modèles de détection automatisés.

Ensuite, des méthodes de *machine learning* non-supervisées, dont les facteurs de détermination d'anomalies diffèrent, ont été appliquées sur les exemples précédents : Local Outlier Factor (calcul de densité locale), One-Class SVM (frontière autour des données normales), DBSCAN (méthode de *clustering*) et Isolation Forest (choix par arbres de décision). Une approche supervisée avec XGBoost a été employée afin d'exercer un apprentissage automatisé de détection des données anormales proche du modèle statistique.

Enfin, la sélection de ces techniques a été réalisée sur la base du F1-score, métrique souvent utilisée dans le cadre d'une classification. En effet, pour chacun des modèles

testés, une variable binaire a été définie de la manière suivante : 1 si le montant réglé est considéré comme une anomalie, 0 sinon. Le tableau des F1-scores obtenus sur le délégataire A est le suivant :

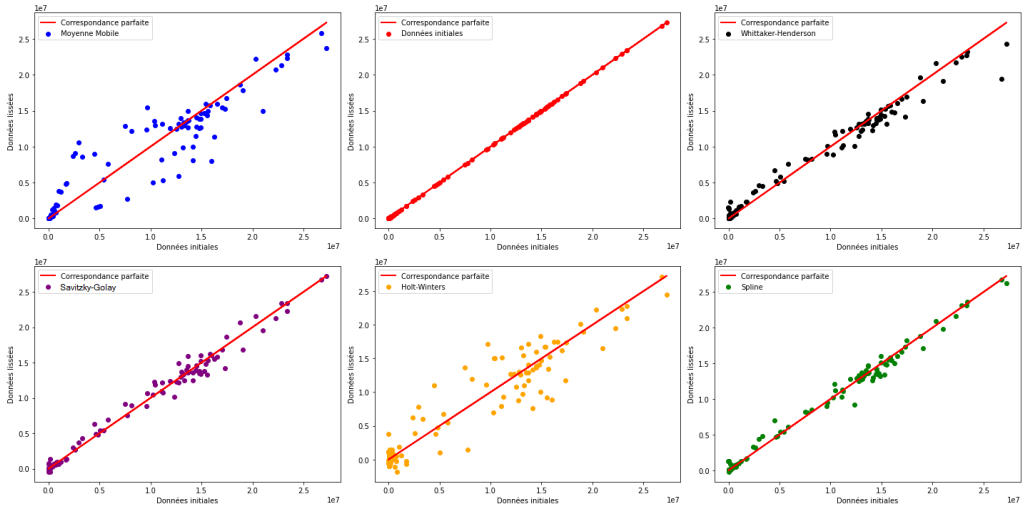
	Local Outlier Factor	One-Class SVM	DBSCAN	Isolation Forest	XGBoost
Délégataire A	10,53%	21,28%	31,25%	68,18%	84,62%
Délégataire B	28,07%	14,81%	51,69%	70,59%	66,67%
Autres délégataires non EDI	23,19%	22,22%	77,85%	36,78%	93,44%
Gestion Directe	14,29%	13,64%	36,36%	58,54%	66,67%

Isolation Forest et XGBoost sont les deux méthodes qui ont repéré au mieux les anomalies détectées par le modèle statistique. Elles semblent optimales pour les règlements en santé. Il est alors nécessaire de corriger ces données anormales.

Correction par lissage

L'objectif de l'étape de correction est d'annuler les effets potentiellement néfastes des anomalies sur les calculs des provisions. Pour cela, un moyen simple et adapté est le lissage mathématique. De nombreux outils ont été examinés afin de gommer les données anormales tout en conservant les tendances des règlements.

Le lissage peut être effectué à partir de plusieurs techniques comme des approches mathématiques (moyenne mobile, les splines), du traitement du signal avec des filtres ou du lissage exponentiel avec les séries temporelles. Le graphique de dispersion suivant présente l'explication de la variance de chacune des méthodes utilisées sur les données initiales du délégataire B :



Ces distributions ainsi que des mesures d'écart permettent de valider l'intérêt du filtre de Savitzky-Golay pour limiter les bruits dans les données.

Modélisation de la détection-correction

Les deux précédentes étapes interviennent dans le cadre du retraitement des données. Le schéma ci-dessous synthétise le modèle employé dans ce but :

Les données anormales repérées correspondent à des erreurs de gestion ou à des phénomènes conjonturels comme la pandémie du COVID-19.

A partir des données corrigées obtenues grâce au lissage des anomalies détectées, l'application d'une méthode de provisionnement est possible.

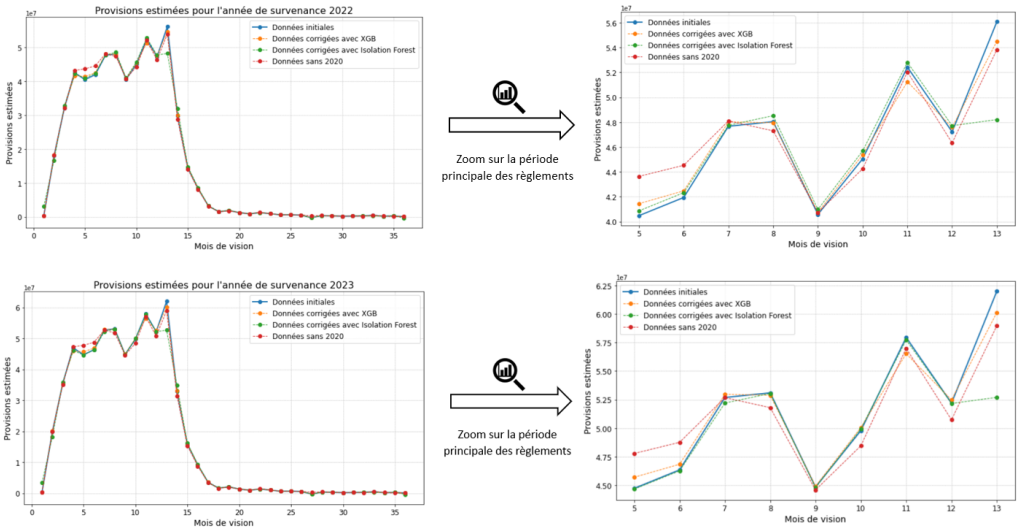
Estimation des provisions

Le modèle Chain Ladder a été retenu car il est très largement utilisé en actuariat. Dans le cadre de cette analyse, il projette les règlements des survenances 2022 et 2023 jusqu'au 36ème mois de vision. Une attention toute particulière est portée sur la comparaison des prévisions estimées à l'aide des données initiales et des données corrigées. Un provisionnement sans prendre en compte l'année de survenance 2020 est ajouté à l'étude pour mettre en parallèle l'impact de la pandémie sur les règlements de santé.

Le tableau ci-après détaille les IBNR calculés pour le délégataire A suivant quatre visions :

	Données initiales	Données corrigées avec XGBoost	Données corrigées avec Isolation Forest	Données initiales sans 2020
2022	3 003 001	2 945 301	2 481 897	3 409 597
delta		*-1,92%*	*-17,35%*	*13,54%*
2023	135 909 581	134 506 715	128 353 867	130 495 855
delta		*-1,03%*	*-5,56%*	*-3,98%*
TOTAL	**138 912 582**	**137 452 016**	**130 835 764**	**133 905 452**
delta		*-1,05%*	*-5,81%*	*-3,60%*

En confrontant les résultats obtenus, la correction à partir des deux approches de *machine learning* a estimé une revue à la baisse du provisionnement par rapport au modèle utilisant les données initiales. Cela peut s'expliquer avec la mise en parallèle des courbes de provisions sur les survenances 2022 et 2023.



Le modèle de détection-correction a créé des divergences dans les montants réglés avec les données initiales, en ajustant les valeurs associées à la pandémie et aux erreurs de gestion. L'analyse du délégataire B confirme cette conjecture alors que celle de la gestion directe la réfute. Cela peut être justifié par l'absence d'intermédiaires impliquant alors une diminution globale des anomalies et une réévaluation à la hausse des provisions.

Bilan de l'étude

Les résultats de ces recherches ont démontré l'intérêt et les enjeux autour de l'intégrité de l'information et de ses conséquences sur le provisionnement. Les modèles d'apprentissage automatisé sont capables de révéler assez précisément les valeurs incohérentes dans les données et l'ajustement par lissage limite leur influence dans les projections. Les montants de provisions alors estimés sont plus précis et fiables, en étant plus fidèles aux évènements passés et par conséquent plus représentatifs des risques actuels.

Face aux enjeux croissants liés aux données, d'autres pistes peuvent être explorées pour renforcer davantage leur fiabilité. L'application de techniques de *deep learning* pourrait améliorer la détection des anomalies en capturant des schémas complexes. Ces modèles pourraient également être étendus à des tâches telles que l'optimisation des processus de souscription ou la tarification afin de conclure sur la robustesse des calculs et des analyses actuarielles.

Executive Summary

Context and issues

Data reliability refers to the accuracy and integrity of information used for decision-making. In the insurance sector, it is crucial for assessing risks, calculating premiums, and making fair forecasts, thereby ensuring the financial stability of the company and the trust of policyholders.

This study aims to explore various approaches to enhance the reliability of claims data in collective health insurance. This branch of insurance is continuously evolving, particularly with the "100% Santé" reform and the shift of responsibilities from Social Security to insurance companies and mutual societies. It also faces narrow financial margins due to intense competition, cost control, and the unpredictable nature of healthcare expenses. Therefore, reserve calculations must be precise and closely aligned with the associated risks.

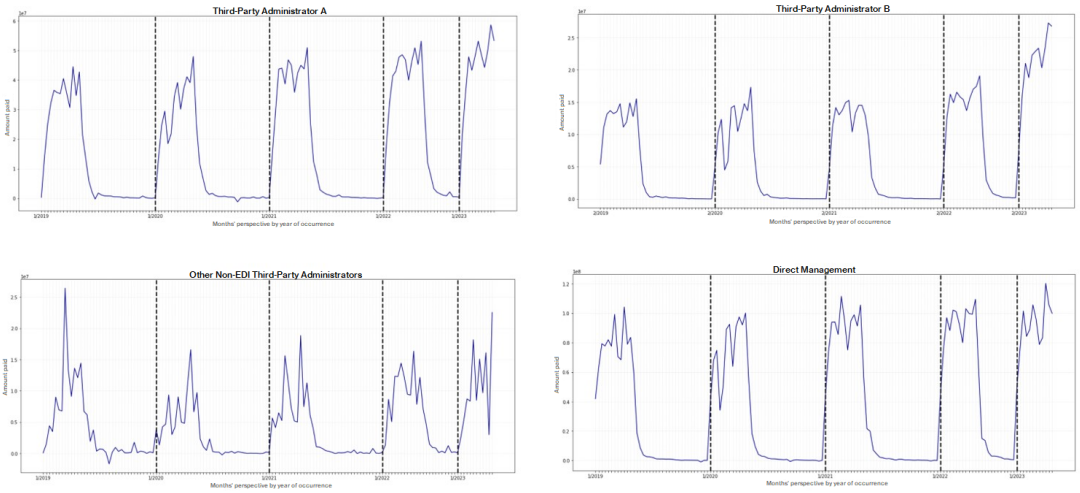
To ensure this accuracy, the data used in reserve models must accurately reflect the uncertainties related to health claims. Human and technical processes involved in the collection, storage, and management of contracts can produce anomalous data, which needs to be corrected. Furthermore, insurance companies are subject to stringent regulatory requirements, emphasizing the need for consistent data in estimations.

Various anomaly detection algorithms will be evaluated, and several smoothing correction techniques will be examined to reprocess anomalous values. Actuarial applications of reserving across several pre-selected study segments will help determine the impact of data reliability improvements.

Analysis of health claim payments

AXA Collective Health's databases related to claims payments enable the analysis of flows concerning health coverage. The technical study focuses on four representative segments of the insurance portfolio, including both direct management and delegated management. These elements cover approximately 45% of health claims payments and validate the relevance of this study.

For these four examples, it is possible to model the trajectory of claims payments. The data history covers the last five years of occurrence (2019 to 2023), with 36 months of data available for each of these years, except for the years 2022 and 2023. Thus, this representation illustrates the evolution of claims payments in collective health insurance, taking into account the timing of claims and their payments.



Differences between these curves are noticeable, although more contrasting values can be identified during the 2020 occurrence for all of these examples. The current goal is to automatically detect potential inconsistencies in these claims.

Anomaly detection

Several approaches have been examined to optimally assess anomalies in health claims across the four study segments. First, a statistical model was established, based on analyzing the variance of the data using a confidence interval. This reference tool enables the estimation of the accuracy of automated detection models.

Next, unsupervised machine learning methods, each with different anomaly detection factors, were applied to the previous examples : Local Outlier Factor (local density calculation), One-Class SVM (boundary around normal data), DBSCAN (clustering method), and Isolation Forest (decision tree-based selection). A supervised approach with XGBoost was employed to perform automated learning for detecting anomalies close to the statistical model.

Finally, the selection of these techniques was based on the F1-score, a metric commonly used in classification. For each tested model, a binary variable was defined as follows : 1 if the payment amount is considered an anomaly, 0 otherwise. The table of F1-scores obtained for Third-Party Administrator A is as follows :

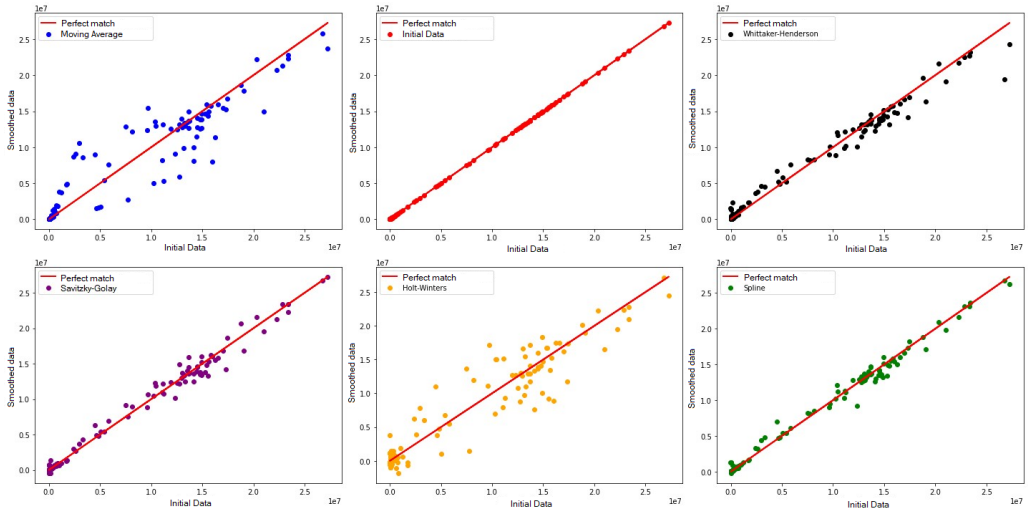
	Local Outlier Factor	One-Class SVM	DBSCAN	Isolation Forest	XGBoost
Third-Party Administrator A	10,53%	21,28%	31,25%	68,18%	84,62%
Third-Party Administrator B	28,07%	14,81%	51,69%	70,59%	66,67%
Other Non-EDI Third-Party Administrators	23,19%	22,22%	77,85%	36,78%	93,44%
Direct Management	14,29%	13,64%	36,36%	58,54%	66,67%

Isolation Forest and XGBoost are the two models that most effectively identified anomalies detected by the statistical model. They thus appear to be optimal for health claims. It is therefore necessary to correct these anomalous data.

Smoothing correction

The goal of the correction step is to eliminate the potentially detrimental effects of anomalies on the reserve calculations. To achieve this, a simple and appropriate method is mathematical smoothing. Various tools have been examined to smooth out anomalous data while preserving the trends in claims.

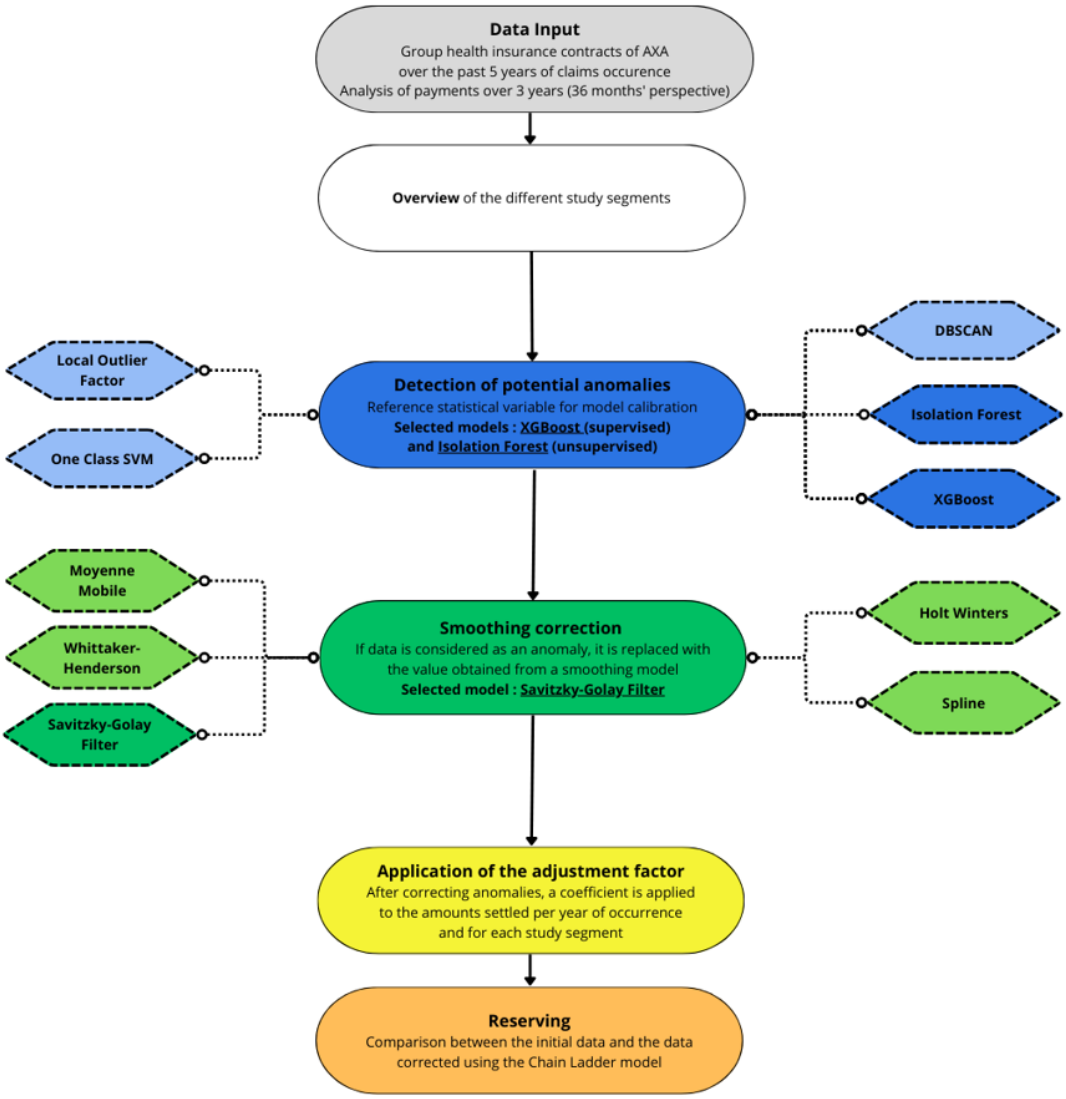
Smoothing can be performed using several techniques, such as mathematical approaches (moving average, splines), signal processing with filters, or exponential smoothing with time series. The following scatter plot illustrates the variance explained by each of the methods applied to the initial data from Third-Party Administrator B :



These distributions, along with deviation measures, validate the usefulness of the Savitzky-Golay filter for reducing noise in the data.

Detection-correction modeling

The two previous steps are part of the data reprocessing. The scheme below summarizes the model used for this purpose :



The identified outliers correspond to management errors or to circumstantial events such as the COVID-19 pandemic.

Based on the corrected data obtained through the smoothing of detected anomalies, the application of a reserving method is possible.

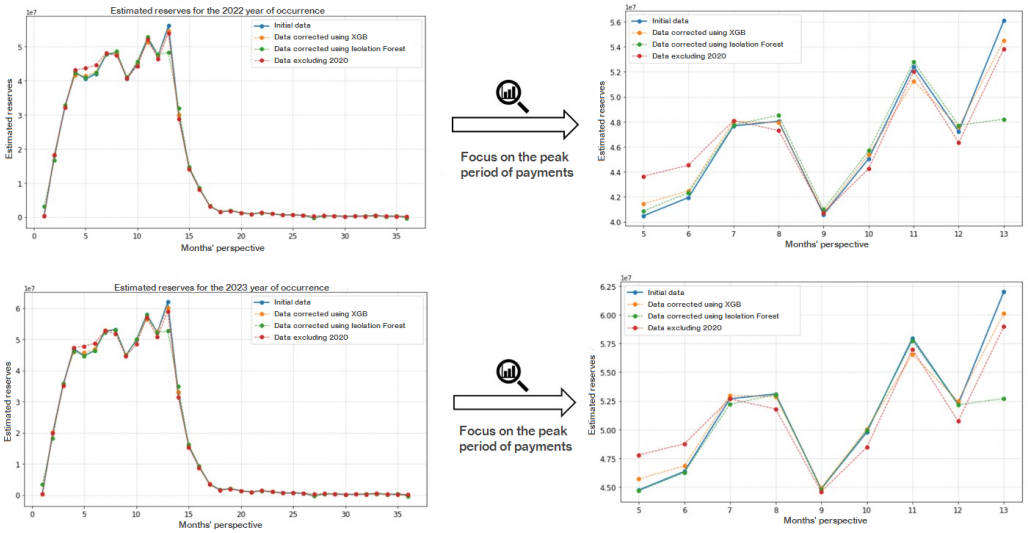
Reserves estimation

The Chain Ladder model was chosen because it is widely used in actuarial practice. In this analysis, it projects the claims for the years 2022 and 2023 up to the 36th month of projection. Particular attention is given to comparing the forecasts estimated using both the initial and corrected data. A reserving analysis excluding the year 2020 is also included to compare the impact of the pandemic on health claims.

The table below details the IBNR calculated for Third-Party Administrator A according to four projections :

	Initial data	Data corrected using XGBoost	Data corrected using Isolation Forest	Initial data excluding 2020
2022	3 003 001	2 945 301	2 481 897	3 409 597
delta		*-1,92%*	*-17,35%*	*13,54%*
2023	135 909 581	134 506 715	128 353 867	130 495 855
delta		*-1,03%*	*-5,56%*	*-3,98%*
TOTAL	**138 912 582**	**137 452 016**	**130 835 764**	**133 905 452**
delta		*-1,05%*	*-5,81%*	*-3,60%*

By comparing the obtained results, the adjustment based on the two machine learning approaches estimated a downward revision of the provisioning contrasted to the model using the initial data. This can be explained by comparing the reserve curves for the occurrences in 2022 and 2023.



The detection-correction model created discrepancies in the amounts settled with the initial data by adjusting values associated with the pandemic and management errors. The analysis of Third-Party Administrator B confirms this conjecture, while the analysis of direct management refutes it. This can be explained by the absence of intermediaries, leading to a overall reduction in anomalies and an upward reassessment of the provisions.

Study summary

The results of this research have highlighted the importance and challenges surrounding data integrity and its impact on reserving. Automated learning models have shown a strong ability to accurately detect inconsistencies in data, and smoothing adjustments have limited their influence on projections. The resulting estimated reserves are more accurate and reliable, reflecting past events more faithfully and thus providing a better representation of current risks.

Given the growing challenges related to data, further avenues can be explored to enhance its reliability. Applying deep learning techniques could improve anomaly detection by capturing complex patterns. These models could also be extended to tasks such as optimizing underwriting processes or pricing, ultimately ensuring the robustness of actuarial calculations and analyses.

Table des matières

Remerciements **1**

Résumé **3**

Abstract **4**

Note de Synthèse **5**

Executive Summary **11**

Introduction **22**

1 Contexte et actualités sur l'assurance santé **25**
1.1 Définitions et contexte 25
1.1.1 Les contrats santé 25
1.1.2 Les principes de l'assurance collective 27
1.1.3 Règlementation et évolution 27
1.2 Actualités et réformes 28
1.2.1 Télémédecine et digitalisation 28
1.2.2 Reste à charge zéro 29
1.2.3 Evolution des besoins des assurés 29
1.2.4 Transfert de charge 30
1.2.5 Impacts des nouvelles conventions 30
1.3 Objectifs du mémoire 31

2 Les enjeux autour de la donnée **33**
2.1 Les fondamentaux de la fiabilité des données 33
2.1.1 Ethique et intérêts 33
2.1.2 Facteurs détériorant les données 34
2.1.3 Mesures et évaluation de la qualité des données 35
2.2 Données disponibles en santé chez AXA 37
2.2.1 Périmètre d'étude 37
2.2.2 Les différents types de gestion des contrats 37
2.3 Construction de la base de données 38

2.3.1 Jointures et mise en place des élements constitutifs de la base . . . 38
2.3.1.1 Extractions et fusions . . . 38
2.3.1.2 Variable mois vision délégataire . . . 39
2.3.1.3 Segments d'étude . . . 39
2.3.2 Description des variables . . . 40
2.3.3 Statistiques sur la base . . . 41
2.3.3.1 Type de gestion . . . 41
2.3.3.2 Segment d'étude . . . 42
2.3.3.3 Montant, mois de vision et survenance . . . 43
2.3.4 Questionnement et enjeux . . . 46
2.3.4.1 Accélération de la gestion . . . 46
2.3.4.2 Différences de règlement par type de gestion . . . 49
2.3.4.3 Comparaison mois de vision comptable et délégataire . . . 50

3 Détection des anomalies et correction 53
3.1 Analyse détaillée des anomalies . . . 53
3.1.1 Contexte et détermination concrète des distorsions . . . 53
3.1.2 Démonstration de l'importance de la *data* qualité dans la prédiction 57
3.1.3 Zoom sur quatre segments d'étude . . . 58
3.1.4 Méthode statistique de détection des anomalies . . . 59
3.2 Automatisation par *machine learning* . . . 64
3.2.1 Local Outlier Factor (LOF) . . . 64
3.2.2 One-Class Support Vector Machine (OC-SVM) . . . 67
3.2.3 DBSCAN . . . 70
3.2.4 Isolation Forest . . . 73
3.2.5 Approche supervisée : XGBoost . . . 76
3.2.6 Comparatif et choix du modèle . . . 79
3.3 Ajustement des données par lissage . . . 81
3.3.1 Moyenne mobile . . . 81
3.3.2 Splines . . . 82
3.3.3 Whittaker-Henderson . . . 84
3.3.4 Triple lissage exponentiel de Holt-Winters . . . 85
3.3.5 Savitzky-Golay . . . 88
3.3.6 Comparatif des méthodes de lissage des données . . . 90

4 Effets du retraitement des données sur le provisionnement 94
4.1 Théories sur l'estimation des provisions . . . 94
4.2 Arbre final d'ajustement des données . . . 95
4.3 Modèle de provisionnement . . . 98
4.3.1 Modélisation théorique Chain Ladder . . . 98
4.3.2 Vérification des hypothèses du Chain Ladder . . . 99
4.3.3 Estimation des provisions . . . 101
4.3.4 Analyse sans 2020 . . . 102
4.4 Comparaison des estimations du provisionnement . . . 104

4.4.1 Délégataire A . . . 104
4.4.2 Délégataire B . . . 106
4.4.3 Gestion directe . . . 107
4.4.4 Bilan sur l'apport de la fiabilité des données . . . 108

Conclusion **111**

Annexes **114**

A Compléments sur les modèles de détection **115**
A.1 Arbre des méthodes *machine learning* de détection . . . 115
A.2 Tableau des méthodes *machine learning* non supervisées de détection . . . 116
A.3 Métriques d'évaluation pour la classification . . . 117
A.4 Détection des anomalies pour le délégataire A . . . 118
A.5 Détection des anomalies pour le délégataire B . . . 119
A.6 Détection des anomalies pour les autres délégataires non EDI . . . 120
A.7 Détection des anomalies pour la gestion directe . . . 121
A.8 Résultats des modèles de détection sur les quatre segments d'étude . . . 122

B Compléments sur les méthodes de lissage **123**
B.1 Mesures de performance du lissage . . . 123
B.2 Tableau récapitulatif des méthodes de lissage testées . . . 124
B.3 Comparaison des modèles de lissage sur le délégataire A . . . 125
B.4 Graphiques de dispersion du lissage des règlements du délégataire A . . . 125
B.5 Tableau des métriques du lissage des données du délégataire B . . . 126
B.6 Comparaison des modèles de lissage sur le délégataire B . . . 126
B.7 Tableau des métriques du lissage des données de la gestion directe . . . 127
B.8 Graphiques de dispersion du lissage des règlements de la gestion directe . 127

Table des figures **127**

Bibliographie **131**

Introduction

Dans le secteur de l'assurance santé collective, la précision et la fiabilité des modèles de provisionnement sont essentielles pour assurer la pertinence des résultats financiers des assureurs. Les provisions techniques calculées par les méthodes de provisionnement dépendent fortement de la qualité des données. Des anomalies dans ces données peuvent entraîner des estimations erronées, compromettant ainsi la capacité des assureurs à répondre à leurs engagements futurs. Dans ce contexte, la fiabilisation des données est devenue une priorité pour les actuaires et les professionnels de l'assurance.

Les enjeux liés à la qualité des données en assurance santé collective sont multiples. Des données inexactes peuvent entraîner des estimations biaisées des provisions, augmentant le risque financier pour les assureurs. De plus, les exigences réglementaires imposent des standards élevés en matière de précision des rapports financiers, renforçant ainsi la nécessité de données fiables.

Face à ces défis, la problématique centrale de ce mémoire est la suivante : comment la fiabilité des données affecte-t-elle les estimations des provisions en assurance santé collective, et quelles méthodes peuvent être mises en œuvre pour détecter et corriger les anomalies afin d'améliorer la précision des modèles de provisionnement ?

Ce mémoire vise à analyser l'impact des anomalies sur les modèles de provisionnement en assurance santé collective, en évaluant comment des données erronées peuvent influencer l'estimation des provisions. L'objectif est aussi de détecter de manière automatisée les phénomènes exceptionnels tels que les pandémies car ils ne doivent pas être pris en compte dans le cadre du provisionnement. Dans cette perspective, le mémoire sera développé en plusieurs parties.

Tout d'abord, la présentation du contexte autour de l'assurance santé collective sera rappelée, en mettant l'accent sur les dernières actualités et les réformes qui ont des conséquences en matière de gestion des contrats.

Puis sera développé et mis en avant les intérêts de la *data* qualité, la construction du jeu de données utilisé dans ce cadre et la visualisation des potentiels problèmes rencontrés dans la gestion des contrats d'assurance santé.

Ensuite, des techniques automatisées de détection des anomalies seront mises en place. Cela incluera l'utilisation de méthodes statistiques ainsi que l'application de modèles de *machine learning* non supervisés, tels que les SVM (*Support Vector Machine*), les forêts

aléatoires et les techniques de *clustering*, mais aussi supervisés avec le XGBoost. Des tests de précision des modèles permettront de confronter leur niveau de précision et d'adaptation à des données de règlements de sinistres en santé pour repérer les données anormales.

Une fois les anomalies détectées, diverses techniques de correction seront évaluées. Différentes méthodes de lissage, telles que la moyenne mobile, Whittaker-Henderson, les filtres, les splines ou approche avec les séries temporelles seront présentées et comparées en termes de réduction de variance et de maintien de la structure des données.

Enfin, un modèle de provisionnement sera appliqué pour calculer les IBNR. La méthode Chain Ladder sera mise en œuvre sur les données initiales et corrigées, et les résultats seront analysés pour évaluer les écarts et l'amélioration de la précision des estimations des provisions. Une discussion des résultats sera menée pour les interpréter, identifier les limites de l'étude et proposer des suggestions pour des recherches futures.

Le mémoire se concluera par une synthèse des principaux résultats, mettant en évidence l'importance de la fiabilisation des données sur les modèles de provisionnement en assurance santé collective. Il soulignera également les contributions de l'étude aux pratiques actuarielles, en fournissant des recommandations pratiques pour l'amélioration de la qualité des données et la solidité des estimations de provisions. Enfin, des pistes de recherche futures seront proposées pour approfondir l'analyse des impacts des anomalies de données dans un contexte assurantiel.

Chapitre 1

Contexte et actualités sur l'assurance santé

1.1 Définitions et contexte

1.1.1 Les contrats santé

L'assurance santé est un système conçu pour aider les personnes physiques à gérer les coûts liés aux soins médicaux. Elle fonctionne en répartissant le risque financier associé à la maladie ou aux problèmes de santé parmi un grand nombre de personnes assurées. Les assurés paient des primes régulières à l'assureur, qui utilise ces fonds pour couvrir les coûts des prises en charge médicales en cas de besoin.

Le contexte de l'assurance santé est étroitement lié à l'évolution des besoins en matière de soins et à la nécessité de protéger les individus contre les coûts élevés associés à la maladie et aux traitements médicaux.

L'assurance complémentaire santé, également connue sous le nom de mutuelle santé, fonctionne en complément de la Sécurité sociale. En France, la Sécurité sociale rembourse une partie des dépenses de santé, mais il reste souvent des frais à la charge de l'assuré. L'assurance complémentaire santé intervient pour couvrir ces frais supplémentaires. Les assurés paient des cotisations à leur mutuelle, qui utilise ces fonds pour compléter les remboursements de la Sécurité sociale et prendre en charge tout ou partie des dépenses restantes. Cela permet aux assurés de bénéficier d'une meilleure couverture pour leurs frais de santé, réduisant ainsi leur charge financière globale. Le schéma suivant récapitule les principes fondamentaux de l'assurance santé :

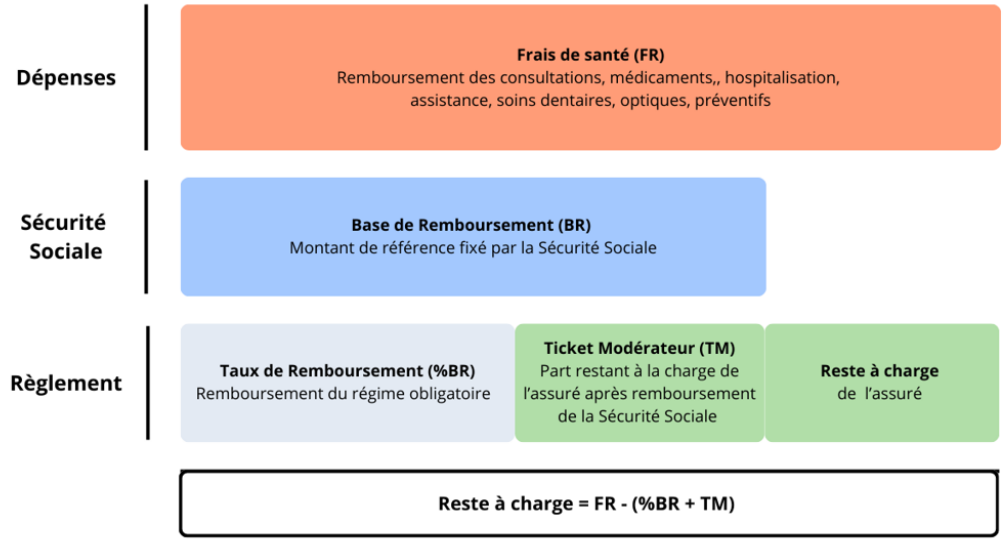


FIGURE 1.1 – Fonctionnement de l'assurance santé en France

Les différentes dépenses prises en charge par l'assurance complémentaire santé peuvent inclure les soins dentaires (comme les prothèses, les implants), l'optique (comme les lunettes, lentilles de contact), les frais d'hospitalisation (par exemple, le confort en chambre individuelle), les médecines douces (ostéopathie, acupuncture), les dépassements d'honoraires médicaux, les médicaments non remboursés par la Sécurité sociale, les frais liés à la maternité, la contraception, ou encore les prestations d'assistance et de prévention.

Ainsi, l'assurance complémentaire santé vient en appui à la Sécurité sociale en couvrant les dépenses de santé non remboursées et en offrant aux assurés une protection plus complète. Sa flexibilité permet aux assurés de choisir des garanties adaptées à leurs besoins spécifiques, en fonction de leur situation familiale, professionnelle et de santé. En effet, le fonctionnement de l'assurance santé repose sur le principe de la mutualisation des risques. Les assurés contribuent financièrement au système par le paiement de primes, et en retour, l'assureur s'engage à couvrir une partie ou la totalité des coûts des soins médicaux en cas de besoin.

Ses avantages sont donc multiples. Elle offre une protection financière aux assurés en cas de maladie ou d'accident corporel. De plus, l'assurance santé encourage la prévention et la prise en charge précoce des problèmes de santé, ce qui peut réduire les coûts à long terme pour l'ensemble du système de santé. Enfin, en répartissant le risque financier de manière équitable, l'assurance santé favorise la solidarité et l'inclusion sociale en garantissant que chacun ait accès à des soins de qualité, quel que soit son niveau de revenu.

1.1.2 Les principes de l'assurance collective

Le sujet abordé se concentrera plus particulièrement sur les contrats santé collectifs. Egalement connue sous le nom d'assurance groupe, il s'agit d'un type de couverture d'assurance offert par un employeur ou une autre entité à un groupe de personnes, telles que les employés d'une entreprise ou les membres d'une association. Ce type d'assurance présente plusieurs principes fondamentaux qui le distinguent des assurances individuelles.

Tout d'abord, l'assurance collective repose sur le principe de la mutualisation des risques au sein du groupe assuré. En regroupant un grand nombre de personnes au sein d'une même police d'assurance, les risques individuels sont répartis de manière à atténuer l'impact financier de tout événement imprévu, tel qu'une maladie grave ou un accident. Cela permet de fournir une protection financière à l'ensemble du groupe, en partageant les coûts entre tous ses membres.

Un autre principe clé de l'assurance collective est celui de l'adhésion volontaire ou obligatoire. Dans de nombreux cas, l'adhésion à l'assurance collective est obligatoire pour les membres du groupe, comme c'est souvent le cas pour les employés d'une entreprise. Cela garantit une large participation au programme d'assurance, ce qui renforce la mutualisation des risques et permet de maintenir des tarifs stables. Cependant, dans certains cas, l'adhésion peut également être facultative, laissant aux individus le choix de participer ou non au programme. Cela peut dépendre des juridictions ou des législations spécifiques à la région ou à la situation professionelle de l'assuré.

De plus, l'assurance collective est souvent conçue pour offrir des conditions tarifaires avantageuses par rapport aux assurances individuelles. Grâce à la taille du groupe assuré et à la négociation de contrats collectifs, les primes d'assurance peuvent être plus compétitives, offrant ainsi une couverture plus étendue à un coût moindre pour les participants. Il est donc primordial pour l'assureur de maintenir un équilibre rigoureux entre le montant des sinistres et celui des primes perçues, afin de garantir la viabilité financière du contrat tout en assurant une protection optimale pour les assurés.

En outre, l'assurance collective peut inclure une gamme de garanties couvrant divers risques, tels que l'assurance maladie, l'assurance vie, l'assurance invalidité, l'assurance accident, l'assurance dentaire, et d'autres prestations complémentaires. Cela permet de répondre aux besoins variés des membres du groupe et de leur offrir une protection globale.

1.1.3 Règlementation et évolution

La réglementation en matière d'assurance santé collective est un élément essentiel à prendre en compte dans la conception et la gestion des programmes d'assurance. Voici quelques principes clés liés à la réglementation de l'assurance santé collective.

Les contrats d'assurance santé collective doivent être conçus et gérés en conformité avec les lois et réglementations en vigueur dans le pays ou la région où ils sont proposés. Cela inclut la conformité aux normes de protection des assurés, aux exigences en matière de solvabilité financière, et aux règles de gouvernance et de divulgation.

Ces réglementations peuvent comporter des dispositions visant à prévenir la discrimination à l'égard des bénéficiaires en raison de caractéristiques telles que l'âge, le sexe, l'état de santé ou d'autres facteurs protégés. En tant qu'actuaire, il est essentiel de s'assurer que les programmes respectent ces exigences de non-discrimination.

De nombreuses réglementations exigent une transparence accrue en ce qui concerne les coûts des primes, les prestations offertes et les processus de tarification. Il est important de s'assurer que les modèles de tarification et les prestations proposées respectent les normes de transparence de la réglementation.

Avec la montée en puissance des mégadonnées dans le domaine de l'assurance santé, la protection des données personnelles est également devenue une préoccupation majeure. Les réglementations telles que le RGPD (Règlement Général sur la Protection des Données) en Europe imposent des obligations strictes en matière de collecte, de stockage et d'utilisation de données personnelles des assurés. Il est capital de travailler en étroite collaboration avec les équipes juridiques et de conformité pour s'assurer que les programmes d'assurance santé collective respectent pleinement les réglementations en vigueur et les principes éthiques de la profession actuarielle.

Les récentes évolutions réglementaires en matière d'assurance santé, visant à renforcer l'accès aux soins pour tous, s'inscrivent directement dans la continuité de nouvelles initiatives.

1.2 Actualités et réformes

1.2.1 Télémédecine et digitalisation

Les progrès technologiques, tels que la télémédecine, les applications de suivi de la santé et les outils d'analyse des données, transforment la manière dont les programmes d'assurance santé collective sont conçus et gérés. Ces technologies offrent de nouvelles opportunités pour améliorer l'efficacité des soins de santé et la gestion des coûts, tout en offrant une meilleure expérience aux bénéficiaires.

La télémédecine, en permettant des consultations à distance, révolutionne la manière dont les soins de santé sont dispensés. Cette évolution technologique influence directement la gestion des contrats d'assurance en santé. Les assureurs doivent prendre en compte la montée en puissance de la télémédecine en intégrant des services de téléconsultation dans leurs offres. Cela peut se traduire par la couverture des consultations à distance et la prise en charge des frais y afférents au sein des contrats d'assurance santé.

Par ailleurs, la digitalisation des dossiers médicaux et des processus de remboursement offre de nombreux avantages. Les assureurs peuvent exploiter les données numériques pour améliorer la gestion des contrats, notamment en facilitant la collecte et l'analyse des informations médicales pertinentes. Cela peut conduire à des offres plus personnalisées et adaptées aux besoins spécifiques des assurés.

En ce qui concerne les prises en charge par les mutuelles et assureurs, la digitalisation permet une gestion plus efficace des remboursements. Les assurés peuvent soumettre leurs demandes de remboursement en ligne, ce qui accélère le processus et réduit la charge administrative pour les assureurs. De plus, la digitalisation favorise la transparence en offrant aux assurés un suivi en temps réel de leurs remboursements et des prestations couvertes par leur contrat, renforçant ainsi la confiance et la satisfaction des assurés.

1.2.2 Reste à charge zéro

Le reste à charge zéro, introduit dans le cadre de la réforme "100 % Santé", vise à réduire les coûts pour les assurés sur trois postes de dépenses : l'optique, le dentaire et l'audiologie. Ce dispositif prévoit un remboursement intégral par la Sécurité sociale et les mutuelles santé pour certaines lunettes, prothèses dentaires et aides auditives, éliminant ainsi tout frais à la charge de l'assuré pour ces postes de soins.

La réforme "100 % Santé" propose un panier de soins divers et identifiés, avec pour objectif de faciliter l'accès à des soins et équipements de qualité. Tous les équipements inclus dans ce panier seront pris en charge intégralement par la Sécurité sociale et les complémentaires santé.

Le reste à charge zéro est progressivement mis en place entre 2019 et 2021 pour les soins optiques, et est déjà effectif pour les prestations et équipements auditifs depuis janvier 2019. Tous les Français bénéficiant d'une complémentaire santé responsable ont accès au reste à charge zéro. Il n'y a pas de conditions d'âge ou d'état de santé pour en bénéficier.

1.2.3 Evolution des besoins des assurés

Les changements dans la législation, tels que les réformes de la santé ou les initiatives visant à promouvoir la transparence des coûts, ont un impact significatif sur la conception et la mise en œuvre des programmes d'assurance santé collective. Il est essentiel de suivre de près ces évolutions pour s'assurer que les programmes soient conformes aux exigences légales et répondent aux besoins changeants des employés.

Les attentes des employés en matière de couverture santé évoluent, et les entreprises doivent s'adapter pour rester compétitives sur le marché de l'emploi. Il apparaît alors une tendance vers une plus grande personnalisation des prestations, avec des offres flexibles qui permettent aux employés de choisir des options adaptées à leurs besoins individuels. De plus, la prise en compte croissante de la santé mentale dans les programmes d'assurance santé collective reflète une évolution importante dans la manière dont les soins de santé sont abordés en milieu professionnel.

En intégrant ces tendances dans la conception des programmes d'assurance santé collective, les entreprises peuvent mieux répondre à l'évolution des besoins de leurs employés tout en optimisant les coûts et en améliorant la santé et le bien-être de leur personnel.

1.2.4 Transfert de charge

L'impact réel de ces nouvelles réformes est constaté au travers d'un transfert significatif de la charge de la Sécurité sociale vers les compagnies d'assurance et les mutuelles. Ce transfert croissant a engendré des coûts supplémentaires substantiels pour ces acteurs. Les coûts liés à la prise en charge de prestations autrefois couvertes par la Sécurité sociale ont doublé au cours des dernières années, ce qui a nécessité une adaptation rapide des processus de tarification, de provisionnement et de gestion des sinistres pour assurer la soutenabilité financière des contrats d'assurance santé.

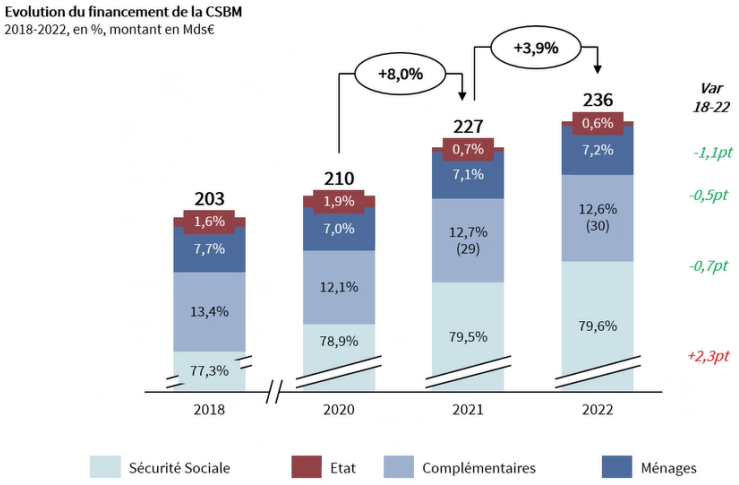


FIGURE 1.2 – Illustration du phénomène de transfert de charge

En parallèle, ce transfert de charge a suscité des réactions et des ajustements significatifs sur le marché de l'assurance santé. Les compagnies d'assurance ont dû repenser leurs politiques de tarification, développer de nouveaux mécanismes de gestion des risques, et renforcer leurs capacités de provisionnement pour faire face à cette évolution majeure. De plus, les mutuelles ont été confrontées à la nécessité de réévaluer leurs réserves et leurs stratégies d'investissement pour garantir leur solvabilité à long terme.

1.2.5 Impacts des nouvelles conventions

Les récentes évolutions législatives et les nouvelles conventions dans le domaine de l'assurance santé ont engendré des changements significatifs dans la gestion des contrats pour les compagnies d'assurance. Les modifications des obligations de couverture, les ajustements des modalités de remboursement des soins de santé, ainsi que les initiatives

visant à favoriser l'innovation dans les produits d'assurance ont profondément influencé le paysage de la couverture santé au cours des dernières années.

L'évaluation des risques associés aux nouvelles obligations de couverture a nécessité une réévaluation des modèles actuariels, tandis que la tarification des contrats a dû être ajustée pour prendre en compte les nouveaux critères imposés par les réformes législatives. Cela a suscité des réactions notables sur le marché de l'assurance santé. Les compagnies d'assurance ont répondu en cherchant à innover dans leurs offres de produits, à développer des partenariats stratégiques avec des acteurs de la santé, et à repenser leurs stratégies de distribution pour s'adapter aux nouvelles conventions. Ces initiatives ont non seulement façonné la compétitivité du marché de l'assurance santé, mais ont également eu un impact sur l'expérience des assurés, qui ont pu bénéficier de nouvelles options de couverture et de services, ainsi que sur la gestion des contrats.

1.3 Objectifs du mémoire

Bien que ce mémoire vise à répondre à une problématique autour de la fiabilité des données, l'étude des contrats santé collectifs et son contexte sont des éléments non négligeables à prendre en compte. En effet, pour identifier les enjeux derrière les anomalies, il faut chercher à connaître les raisons qui pourraient expliquer ces difficultés. L'ensemble des sujets et réformes évoqués dans la partie précédente peuvent avoir des conséquences dans la construction et l'analyse des bases de données.

C'est pourquoi il sera nécessaire dans un premier temps de comprendre les données disponibles au sein d'AXA concernant les règlements de sinistre pour les contrats santé collectifs, en s'intéressant aux informations et variables des bases de données. Puis, après l'observation, l'approfondissement et la compréhension des enjeux derrière les anomalies visibles dans les données, un modèle de détection devra être développé afin d'automatiser la recherche des problèmes de *data*. Ensuite, divers outils statistiques seront testés pour optimiser la correction de ces erreurs et enfin, un modèle de provisionnement sera appliqué, sur les données traitées et sur les données initiales, et l'objectif sera de comparer l'ensemble des résultats obtenus.

Chapitre 2

Les enjeux autour de la donnée

2.1 Les fondamentaux de la fiabilité des données

2.1.1 Ethique et intérêts

La qualité des données est un élément primordial dans de nombreux domaines, y compris l'assurance, et joue un rôle majeur dans la prise de décisions stratégiques, opérationnelles et analytiques. Pour décrire au mieux les fondamentaux de la fiabilité des données, il est nécessaire de rappeler son origine, puis d'aborder les enjeux, le fonctionnement et ses intérêts, en mettant l'accent sur son importance dans le domaine de l'assurance.

La préoccupation pour l'intégrité des informations remonte à plusieurs décennies, mais elle a pris une place croissante à l'ère de la transformation numérique et de l'explosion des volumes de données. Les premiers systèmes informatiques ont rapidement mis en lumière les défis liés à la gestion des données, conduisant à un intérêt croissant pour les normes de qualité et les bonnes pratiques. Au fil du temps, la complexité des systèmes d'information et l'évolution des réglementations ont renforcé l'importance de la cohérence des données pour garantir la fiabilité et la pertinence des informations utilisées dans la prise de décisions.

Dans le domaine de l'assurance, des données précises et fiables sont essentielles pour évaluer les risques, établir des tarifs appropriés, gérer les sinistres et offrir des produits et services adaptés aux besoins des clients. De plus, dans un contexte de réglementation stricte, la validité de l'information est capitale pour garantir la conformité aux exigences légales et réglementaires en matière de reporting, de protection des données et de lutte contre la fraude.

L'exactitude des données en assurance influence de manière significative plusieurs aspects clés de l'industrie. En effet, elle agit de manière significative sur l'analyse des risques et la prise de décision au sein des compagnies d'assurance. Des données de mau-

vaise qualité peuvent entraîner des évaluations erronées des aléas, compromettant ainsi la capacité des assureurs à gérer efficacement les risques assurés. Cela peut conduire à des décisions sub-optimales en matière de souscription, de tarification et de gestion des sinistres, impactant directement la rentabilité et la solidité financière des compagnies d'assurance.

De plus, la fiabilité des données a des implications directes sur les relations avec les assurés. Des données inexactes ou obsolètes peuvent entraîner des retards dans le traitement des demandes d'indemnisation, des erreurs dans les communications avec les assurés et une insatisfaction générale de la clientèle. En revanche, des données de haute qualité permettent aux assureurs de fournir un service plus efficace, précis et personnalisé, renforçant ainsi la confiance et la fidélité des assurés.

Le cadre réglementaire entourant la qualité des données en assurance est nécessaire pour garantir la protection des bénéficiaires et la fiabilité des informations utilisées dans le secteur de l'assurance. Les réglementations telles que le RGPD en Europe et d'autres lois sur la protection des données imposent des exigences strictes en matière de qualité et de sécurité des données. Cela signifie que les assureurs doivent vérifier que les données soient précises et sécurisées contre tout accès non autorisé mais ils doivent aussi aussi veiller à ce que l'utilisation de ces données soit justifiée, transparente et conforme aux attentes des individus concernés.

2.1.2 Facteurs détériorant les données

La précision des informations en assurance est influencée par une multitude de facteurs, allant de l'origine des données et des processus de collecte, à leur stockage et leur gestion, ainsi que l'intégration de données provenant de sources multiples. Comprendre ces facteurs est essentiel pour identifier les sources potentielles d'erreurs et d'incohérences dans les données, et pour élaborer des stratégies visant à améliorer leur qualité.

Tout d'abord, les assureurs font souvent appel à une variété de sources de données, telles que les formulaires de demande, les antécédents médicaux, les rapports d'expertise et les bases de données externes. Cependant, la diversité des sources et des formats de données peut entraîner des incohérences, des erreurs de saisie et des lacunes qui compromettent l'appréciation des risques et le calcul des tarifs des contrats.

Ensuite, le stockage et la gestion des données représentent un autre défi majeur pour la validité des informations. Les assureurs doivent veiller à ce que les données soient correctement stockées, sécurisées et accessibles, tout en préservant leur intégrité et leur précision. Des pratiques inadéquates de stockage des données, telles que des systèmes obsolètes ou des processus de sauvegarde inefficaces, peuvent entraîner des pertes de données, des altérations non intentionnelles et des difficultés à retrouver et à utiliser les informations nécessaires.

Enfin, L'intégration de données provenant de sources multiples constitue un défi supplémentaire pour garantir la fiabilité des informations en assurance. Les assureurs doivent souvent agréger des données provenant de différentes sources internes et externes, telles que les systèmes de gestion des sinistres, les partenaires commerciaux et les bases de données publiques, pour obtenir une vue d'ensemble complète des risques assurés. Cependant, cette intégration peut entraîner des incompatibilités entre les formats de données, des doublons et des incohérences, compromettant ainsi l'exactitude des informations utilisées pour prendre des décisions critiques en matière d'assurance. A titre d'exemple, il est possible qu'un gestionnaire comptable se trompe sur l'écriture d'un montant (un contrat présentant un règlement de 100k€ passant à 1M€). Ces erreurs opérationnelles peuvent alors modifier les résultats d'analyse, elles doivent donc être corrigées en amont.

De manière générale, que ce soit en lien avec un processus humain ou technique, les données ont le risque d'être à un moment altérées ce qui renforce la pertinence de les ajuster. Dans l'évaluation des risques sous Solvabilité 2, le risque opérationnel est bien pris en compte et peut représenter une part conséquente du coût en capital, démontrant le facteur déterminant que constitue la fiabilité de l'information.

2.1.3 Mesures et évaluation de la qualité des données

La gestion de la cohérence des données implique plusieurs étapes, telles que la collecte, le stockage, la validation, la normalisation, la déduplication, la correction, la documentation et la gouvernance des données. Des outils spécialisés et des processus rigoureux sont utilisés pour surveiller et améliorer en continu l'exactitude des données. Cela peut inclure l'automatisation des contrôles, l'utilisation de modèles de données normalisés, la mise en place de processus de validation croisée et la formation des utilisateurs finaux. DataValueConsulting, cabinet de conseil et d'intégration en *data*, résume dans le schéma suivant les principes de la fiabilité des données :

FIGURE 2.1 – Indicateurs clés et objectifs de la *data* qualité

Les enjeux liés à la qualité des données sont multiples. Il s'agit notamment de garantir l'exactitude, la cohérence, l'intégrité, la complétude, la validité et la fiabilité des

données. Des erreurs ou des lacunes dans les données peuvent entraîner des décisions erronées, des analyses biaisées, des coûts supplémentaires, une perte de confiance des clients, des problèmes de conformité réglementaire et une diminution de la compétitivité. Par conséquent, la précision de l'information est devenue un élément stratégique pour de nombreuses organisations, y compris les compagnies d'assurance.

Les métriques et indicateurs de la cohérence des données jouent un rôle premier dans l'évaluation de la fiabilité des informations utilisées dans le secteur de l'assurance. Parmi les métriques utilisées, la complétude, l'exactitude, la cohérence et la pertinence sont les plus essentielles. La complétude mesure si toutes les données nécessaires sont présentes, l'exactitude évalue la précision des données, la cohérence vérifie la concordance des informations à travers différents systèmes, et la pertinence évalue si les données sont appropriées pour l'usage auquel elles sont destinées.

En plus des métriques, divers outils et méthodes sont utilisés pour évaluer l'intégrité des données en assurance. Les outils de gestion, tels que les logiciels de nettoyage et de déduplication, peuvent être déployés pour améliorer la qualité des données. Parallèlement, des audits réguliers et des processus de validation, incluant des contrôles manuels et automatisés, sont mis en place pour garantir que les données répondent aux normes de qualité requises.

Il est également important de noter que l'évaluation de la fiabilité des données est un processus continu et évolutif. Ce constat s'intègre bien au contexte de l'assurance santé qui exige une actualisation constante des informations pour s'adapter aux évolutions réglementaires, aux besoins des assurés, et aux avancées technologiques, afin de garantir une gestion efficace des risques et des prestations.

Il conviendra alors d'analyser en détails la justesse des informations détenues par AXA Santé Collective afin d'identifier les enjeux clés et évaluer la pertinence de cette étude.

2.2 Données disponibles en santé chez AXA

2.2.1 Périmètre d'étude

De nombreuses bases de données sont disponibles au sein de la direction pilotage technique d'AXA Santé Collective en ce qui concerne la prévoyance santé. Dans le cadre de cette étude, ce sont les règlements des sinistres qui vont être analysés. La base disposant de ces informations s'appelle "Mouvements sinistres" et répertorie les flux de règlements pour les contrats collectifs en santé et prévoyance. Sur chaque ligne de la base sont renseignées différentes informations concernant le dossier associé (numéro du contrat, numéro de sinistre, type de couverture...) ainsi que le montant du règlement effectué. D'autres données essentielles dans cette étude vont être récupérées depuis deux autres bases : "Cartographie" qui regroupe de nombreuses variables qualitatives et quantitatives concernant les contrats d'assurance et "Facture" qui présente des informations supplémentaires, notamment à propos de la gestion de ces contrats.

Ces trois bases de données sont elles-mêmes extraites de l'entrepôt de données d'AXA Santé Collective et c'est à partir de la fusion de ces dernières que la base étudiée dans ce mémoire va pouvoir être générée. Pour cela, le logiciel WPS sera utilisé. Il est fondé sur le langage de programmation SAS et est conçu pour la gestion de bases de données conséquentes. Une fois la base créée, l'étude se poursuivra sur l'outil Databricks qui est adapté à la *Big Data* et à l'analyse technique poussée avec des modèles statistiques sous Python. Dans l'optique de gestion et de développement de modèles sur une base de données très volumineuse, la librairie PySpark devra être employée.

2.2.2 Les différents types de gestion des contrats

AXA étant un acteur majeur de l'assurance santé collective, elle adopte une approche mixte de gestion de ses contrats, c'est-à-dire qu'une partie d'entre eux est déléguée à des intermédiaires en fonction des besoins spécifiques des activités et des opportunités sur le marché de l'assurance. En effet, la gestion directe et la gestion déléguée sont deux approches clés, chacune avec ses propres principes fondamentaux.

La gestion directe implique que l'assureur assume directement la responsabilité de la souscription des polices d'assurance, de la tarification, de la gestion des sinistres et de la réassurance, le tout en interne. Cela signifie que toutes les fonctions essentielles de l'assurance sont gérées et exécutées par l'assureur lui-même, sans recourir à des tiers pour ces services.

Les fondements de la gestion directe incluent :
— Un contrôle complet : l'assureur exerce une maîtrise sur toutes les décisions et activités liées à l'assurance, ce qui lui permet de maintenir une vision globale et cohérente de ses opérations.
— Une expertise interne : l'assureur doit disposer des ressources nécessaires pour

gérer toutes les facettes de l'assurance, de la souscription des risques à la gestion des sinistres, en passant par la tarification et la réassurance.
— Une responsabilité directe : l'assureur joue un rôle clé dans la performance de ses produits d'assurance, ce qui le rend responsable des résultats financiers et de la satisfaction des assurés.

En revanche, la gestion déléguée implique qu'une partie des fonctions stratégiques de l'assurance soit confiée à des tiers spécialisés, tels que des courtiers, des réassureurs ou d'autres assureurs. Les éléments essentiels de la gestion déléguée englobent :
— Des partenariats stratégiques : l'assureur doit établir des liens solides avec des prestataires de services externes, en s'assurant qu'ils partagent la même vision en matière d'assurance.
— Une expertise spécialisée : en confiant certaines tâches à des tiers spécialisés, l'assureur peut bénéficier de l'expérience de ces prestataires pour renforcer sa propre offre de produits d'assurance.
— Une gestion des risques : l'assureur doit exercer une diligence raisonnable dans le choix de ses partenaires délégués, en vérifiant qu'ils respectent les normes de gouvernance, de conformité et de gestion des risques.

La combinaison de ces deux approches permet à AXA de gérer un grand nombre de contrats santé collectifs. Néanmoins, cela pourrait avoir des conséquences sur les informations disponibles dans les bases et sur l'exactitude des données.

2.3 Construction de la base de données

2.3.1 Jointures et mise en place des élements constitutifs de la base

2.3.1.1 Extractions et fusions

Comme présenté en introduction de cette partie, 3 bases vont être utilisées : "Mouvements sinistres", "Cartographie" et "Facture". Dans un premier temps, il faut choisir les éléments nécessaires à l'étude présents dans la base "Mouvements sinistres".

La première sélection à effectuer sur la base est sur le type de garantie, ici uniquement les garanties santé. Puis, l'analyse tout au long de ce mémoire doit porter sur des périodes de survenance récentes, en équilibrant d'un côté un nombre suffisamment important d'années tout en limitant la quantité de données pour la base finale. Ainsi, la décision de sélectionner les cinq dernières années de survenance (entre 2019 et 2023) semble être un choix optimal et adapté à l'objectif de la recherche. L'idée étant de détecter des anomalies, la survenance 2020 sera profitable pour l'évaluation des modèles compte tenu du contexte pandémique et son impact sur les dates de règlement de sinistre.

La première jointure entre la base "Mouvements sinistres" et "Cartographie" a pour objectif d'ajouter des variables dans la base de données finale. Il s'agit du type de gestion du contrat, des noms du délégataire et du client concernés, du chiffre d'affaire du délégataire ainsi que si le contrat est avec ou sans participation aux bénéfices. Les clés de cette jointure sont le numéro du contrat, le numéro de sinistre ainsi que l'année de survenance.

Une variable fondamentale disponible dans la base "Mouvements sinistres" est la variable "mois_vision". Elle correspond à la différence entre la période comptable et la date de survenance du sinistre :

$$mois\_vision = (ANNCPT - an\_surv) \times 12 + MOISCPT \tag{2.1}$$

Elle permettra d'identifier les cadences de règlement en santé. Cependant, pour répondre à certains enjeux de l'étude, il va falloir créer deux nouvelles variables.

2.3.1.2 Variable mois vision délégataire

Les montants de règlement dans la base "Mouvements sinistres" dépendent uniquement des dates de survenance et comptables. Or, il est possible que des anomalies dans les données apparaissent à cause de problèmes de gestion. Cela peut être dû aux échanges entre les équipes de comptabilité et les délégataires qui impliqueraient des écarts notamment au niveau des périodes de règlement. Entre le moment où le délégataire a complété le règlement et la date exacte de prise en compte de ce règlement par la comptabilité, du temps peut s'écouler et des erreurs de manipulation de *data* peuvent apparaître.

C'est alors que la base "Facture" va pouvoir être utile. Elle dispose d'informations provenant directement du délégataire, spécifiquement les dates de début et de fin de règlement de ce dernier. En réalisant la différence entre les périodes de fin de règlement et de survenance, la variable nommée "mois_vision_delegataire" peut être définie :

$$mois\_vision\_delegataire = (an\_fin\_regl - an\_surv) \times 12 + mois\_fin\_regl \tag{2.2}$$

Il sera alors possible de comparer les trajectoires des montants de règlement selon le mois vision comptable et délégataire et mettre en exergue quelques différenciations à prendre en compte dans l'analyse de la fiabilité des données.

2.3.1.3 Segments d'étude

Un facteur primordial du sujet est le type de gestion des contrats santé collectifs. Il faudra distinguer la cadence de règlements selon s'il s'agit d'une gestion directe ou déléguée. Mais, il faudra également réaliser un zoom sur certains délégataires dont les données pourraient être plus erronées que d'autres.

Pour cela, il faut créer une nouvelle variable nommée "Segment_Gestion" pour séparer clairement les différents axes d'analyse : d'un côté, la gestion directe et de l'autre la gestion délégataire. Pour cette dernière, il a fallu créer un classement selon le chiffre d'affaires afin de générer un top 20 de ces délégataires. Les autres délégataires avec moins d'enjeux sont regroupés dans deux autres catégories : "Autres EDI" qui sont concernés par le processus d'échange de données automatisées avec AXA Santé Collective et "Autres non EDI" qui correspondent au peu de délégataires qui n'ont pas optés pour cette procédure. Il sera donc intéressant de comparer ces différents segments notamment en ce qui concerne l'étude de données anormales car le type et la quantité d'anomalies pourraient différer.

Cette nouvelle variable qui correspond à la segmentation des périmètres d'étude en quatre socles se traduit alors de la manière suivante :

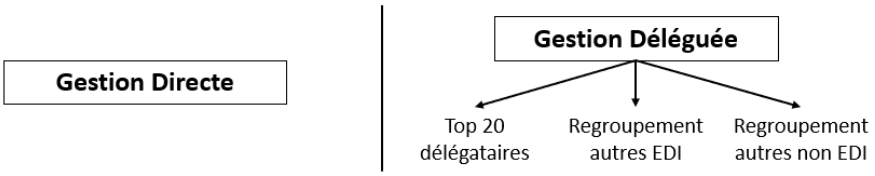


FIGURE 2.2 – Classification des segments d'étude selon le type de gestion

En ajoutant cet élément à la base de données, l'ensemble des informations nécessaires à l'étude sont désormais à disposition.

2.3.2 Description des variables

La base de données finalement créée est composée d'environ 23 millions d'observations et de 16 variables. Elle a été construite sur des extractions à date du 31/12/2023. Sur le tableau ci-dessous est présenté l'ensemble des variables de cette base :

Variable	Signification	Source
IDCORP	Numéro d'identification du contrat	Mouvements sinistres
num_sin	Numéro de sinistre	Mouvements sinistres
an_surv	Année de survenance du sinistre	Mouvements sinistres
IDEAN1	Numéro du contrat juridique	Cartographie
ANNCPT	Année comptable	Mouvements sinistres
MOISCPT	Mois comptable	Mouvements sinistres
mois_vision	Mois de vision comptable	Mouvements sinistres
mois_vision_del	Mois de vision délégataire	Facture
client	Nom du client	Mouvements sinistres
segment_daap	Segment d'inventaire	Cartographie
PB	Participation aux bénéfices	Cartographie
gestionsin	Gestion directe ou déléguée	Cartographie
DELEGATAIRE_SIN	Nom du délégataire	Cartographie
Segment_Gestion	Segment d'étude	Cartographie
CA	Chiffre d'affaire	Cartographie
montant_regle	Montant de règlement du sinistre	Mouvements sinistres

Après avoir établi cette base de données complète et structurée concernant les règlements en assurance santé collective, il est désormais possible de procéder à une analyse statistique approfondie de ces données. Ces données, comprenant des variables clés telles que les montants réglés, les différentes visions temporelles (comptable, délégataire) et les segments d'étude constituent une source précieuse d'information, tout en omettant les potentiels éléments confidentiels. Leur analyse statistique permettra d'explorer les tendances, d'identifier des schémas récurrents et d'évaluer la performance du portefeuille d'assurance santé. Ainsi, cette étape est essentielle pour tirer des conclusions pertinentes qui guideront la prise de décision stratégique concernant l'étude plus approfondie des anomalies et l'optimisation du processus actuariel de provisionnnement.

2.3.3 Statistiques sur la base

2.3.3.1 Type de gestion

L'une des variables fondamentales de la base de données est "gestionsin", qui distingue les contrats gérés en gestion déléguée et ceux en gestion directe. Cette distinction est primordiale pour comprendre les dynamiques de règlement des sinistres dans l'assurance santé collective. La gestion déléguée, où un tiers assure la gestion des contrats, et la gestion directe, où l'assureur gère directement les contrats, impliquent des mécanismes et des processus distincts qui peuvent influencer les résultats financiers et opérationnels.

Le graphique présenté ci-dessous met en évidence la répartition des règlements selon le type de gestion. Il apparaît clairement qu'environ un quart des règlements proviennent de la gestion directe, tandis que le reste, soit environ 75%, est issu de la gestion déléguée. La prépondérance de la gestion déléguée pourrait s'expliquer par plusieurs facteurs, tels

que la complexité des contrats, le besoin de spécialisation dans la gestion des sinistres, ou encore des considérations économiques favorisant l'externalisation.

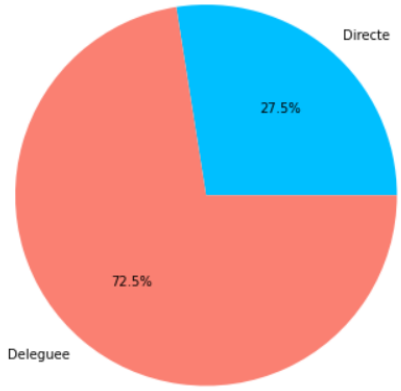


FIGURE 2.3 – Répartition des montants réglés des contrats santé collectifs en fonction du type de gestion

L'analyse de la variable "gestionsin" ne se limite pas à une simple répartition des règlements. Il est capital d'examiner comment ces deux modes de gestion influencent la performance globale des règlements. Par exemple, la gestion directe peut offrir une meilleure maîtrise des coûts et des délais, mais pourrait également exiger des ressources internes plus importantes. À l'inverse, la gestion déléguée pourrait permettre une plus grande efficacité opérationnelle, au prix d'une moindre flexibilité et d'une dépendance vis-à-vis du gestionnaire délégué. Le délai moyen de règlement sera donc étudié en fonction de cette variable pour en tirer des conclusions opérationnelles.

2.3.3.2 Segment d'étude

En complément de l'analyse de la variable "gestionsin", la variable "Segment_Gestion" offre une décomposition plus fine de la gestion des contrats. Cette segmentation permet une analyse granulaire des modes de gestion et des performances associées.

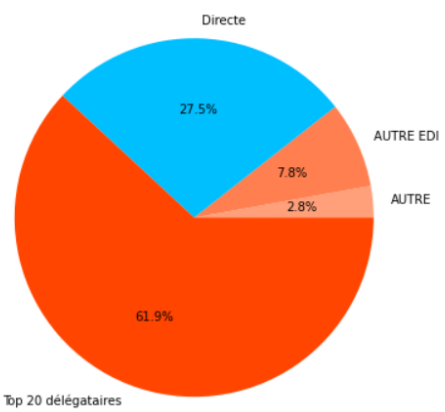


FIGURE 2.4 – Répartition des montants réglés des contrats santé collectifs en fonction des segments d'étude

Le diagramme démontre que le "top 20 délégataires" occupe la part la plus importante de la gestion déléguée, suivi par les délégataires EDI, et enfin, les règlements via les "autres" (autrement dit "autres délégataires non EDI"). Cette répartition met en lumière le rôle prépondérant des grands délégataires dans la gestion des sinistres en assurance santé collective, ainsi que l'émergence des plateformes d'échange automatisé.

Chaque segment présente des caractéristiques distinctes qui peuvent influencer la performance des règlements. Le "top 20 délégataires", par exemple, regroupe les acteurs les plus importants et souvent les plus expérimentés du marché, ce qui pourrait se traduire par une meilleure efficacité dans la gestion des sinistres, mais aussi par une négociation de conditions financières plus favorables. À l'inverse, les "autres délégataires EDI" pourraient offrir des gains d'efficacité par l'automatisation des échanges, mais nécessitent une analyse plus approfondie pour évaluer leur impact sur les délais et les coûts de règlement. Les "autres non EDI", quant à eux, pourraient représenter une plus grande diversité de pratiques et de résultats, avec des performances potentiellement plus hétérogènes.

Ainsi, il sera judicieux de sélectionner des exemples représentatifs de ces divers segments lors de l'évaluation des anomalies et la mise en place du modèle, pour distinguer l'impact de ces différences de gestion dans le résultat du provisionnement.

2.3.3.3 Montant, mois de vision et survenance

Dans cette section est exposée l'analyse des montants de règlement des sinistres en assurance santé collective sur les cinq dernières années de survenance, présentés par mois de vision. Cette approche permet d'examiner non seulement l'évolution temporelle des règlements, mais aussi d'identifier leur dynamique et les variations au fil du temps.

Une année de survenance se réfère à l'année au cours de laquelle un sinistre est survenu, indépendamment de la date à laquelle il a été réglé. Cette distinction est primordiale

dans l'analyse actuarielle car elle permet de comprendre la temporalité des sinistres par rapport à leur gestion et à leur règlement. En examinant les montants réglés par année de survenance, il est possible d'observer la vitesse de règlement des sinistres et d'identifier des tendances à long terme dans le comportement des règlements.

En présentant les montants de règlement par mois de vision, une vue détaillée de l'évolution des règlements est obtenue en tenant compte de la temporalité entre l'apparition du sinistre et son règlement effectif. Cette présentation permet d'identifier des variations saisonnières et des tendances dans la gestion des sinistres.

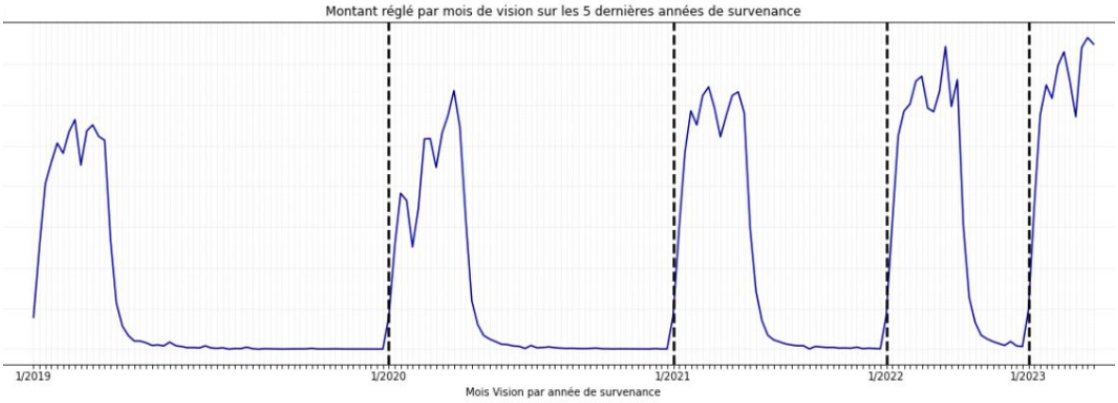


FIGURE 2.5 – Montant réglé total par mois de vision sur les 5 dernières années de survenance

Le graphique illustre cette dynamique en représentant les montants de règlement par mois de vision sur les survenances de 2019 à 2023. Cette représentation sera utilisée tout au long de l'étude car elle permet de saisir d'un coup d'œil les tendances générales et les variations spécifiques, facilitant ainsi l'analyse et l'interprétation des données. Plusieurs observations importantes peuvent être formulées. Globalement, une augmentation progressive des montants de règlement par survenance est notable, indiquant une tendance à la hausse dans la prise en charge des sinistres. De plus, l'analyse des mois de vision permet de repérer des cycles ou des schémas récurrents dans les règlements, tels que des augmentations en fin d'année (proche du 12ème mois de vision), souvent associées à des révisions de provisionnement ou à des clôtures de dossiers.

Se détache de manière distincte la survenance 2020 avec un montant de règlement particulièrement faible pour les mois de vision 5 et 6. Cela s'explique par les conséquences sur la gestion des contrats de la pandémie COVID-19. Des tendances croissantes et décroissantes apparaissent mais elles seront plus notables si un zoom est effectué sur l'évolution des règlements globaux par mois comptable.

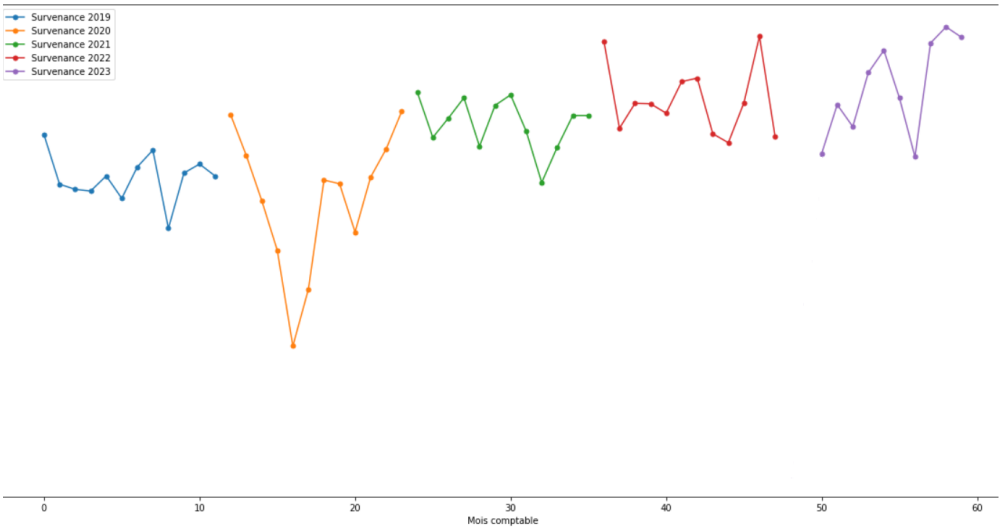


FIGURE 2.6 – Evolution des montants réglés par mois comptable et par survenance

Une nouvelle fois, l'année 2020 se distingue particulièrement des autres années de survenance. Cette période sera particulièrement intéressante à étudier dans le cadre d'une détection automatisée des anomalies.

Une chute des règlements au mois comptable 9 est un phénomène apparent. Cette occurence au mois de septembre peut s'expliquer par plusieurs facteurs. Les effets saisonniers, comme la rentrée scolaire et la fin des vacances d'été, ainsi que le comportement des assurés, peuvent entraîner une baisse temporaire des consultations et des règlements. Les facteurs administratifs, tels que les délais de traitement ou les ajustements administratifs, peuvent également retarder les règlements. De plus, les politiques internes des assureurs concernant le remboursement ou la gestion des sinistres peuvent influencer la répartition des règlements dans le temps.

A l'inverse, l'analyse des données révèle une augmentation significative des règlements en santé durant les périodes estivales et hivernales, suggérant une tendance saisonnière marquée par une hausse des sinistres et des demandes de remboursement pendant ces saisons.

On peut enfin apercevoir la tendance croissante globale des règlements au fur et à mesure des années de survenance. Cela peut aussi se traduire au travers des deux graphiques suivants :

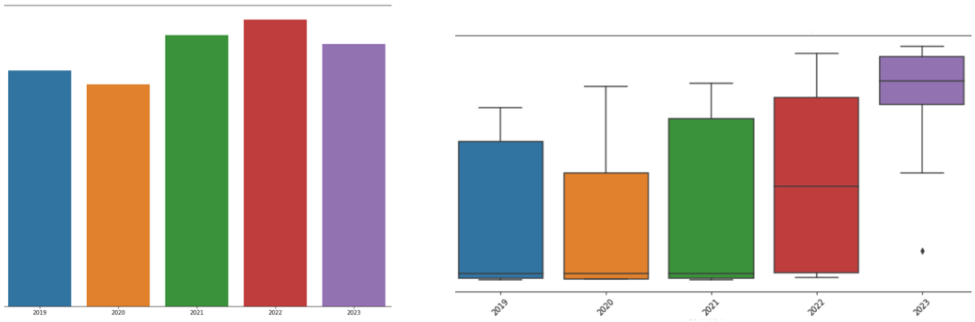


FIGURE 2.7 – Histogramme et *Boxplot* des règlements de santé en fonction de l'année de survenance

Le *boxplot* révèle un aspect important de l'exploration des données et la recherche de données anormales. En effet, en conservant l'ensemble des mois de vision de chaque survenance, de nombreuses valeurs non pertinentes seraient intégrées au repérage des anomalies et viendraient probablement fausser certains résultats. En prenant l'exemple de l'année 2019, 60 mois de vision seraient comptabilisés (jusqu'à 2023). Or, la trajectoire des règlements par survenance montre qu'après un certain temps les montants sont négligeables et n'apportent aucun intérêt à l'étude. De même, les premières boîtes à moustache présentent une médiane proche de 0 ce qui rend incomparable de prime abord les cinq années de survenance.

Il va donc falloir appronfondir quelques recherches avant de chercher à modéliser.

2.3.4 Questionnement et enjeux

2.3.4.1 Accélération de la gestion

La gestion des contrats d'assurance a évolué de manière significative au cours des dernières années, marquée par des améliorations notables en termes d'efficacité, de rapidité et de satisfaction client, grâce à l'intégration de nouvelles technologies et à l'adaptation aux exigences réglementaires.

Dans le contexte des règlements de sinistre en santé, il est prévu contractuellement qu'ils soient effectués au cours des 2 années suivant la survenance de l'aléa. Cependant, des retards peuvent apparaître et afin de les intégrer dans l'étude, il convient de définir un nombre de mois de vision limite.

Dans un premier temps, il faut regrouper les mois de vision par année de vision, c'est à dire que l'année 1 correspondra aux 12 premiers mois de vision, l'année 2 du 13ème au 24ème mois et ainsi de suite.

Puis, il faut déterminer le pourcentage de règlement. Cette nouvelle variable correspond à la somme des règlements par année de vision divisée par la somme totale des règlements pour chaque survenance. Ce pourcentage, à un mois de vision donné, représente la part de règlements effectuée jusqu'à ce mois de vision. En plus de cette variable est ajoutée la moyenne du pourcentage des règlements par année de vision facilitant la confrontation des résultats par survenance. Cela va aussi permettre de se rendre compte de la répartition des règlements des contrats santé collectifs au fur et à mesure du temps.

FIGURE 2.8 – Evolution du pourcentage de règlement par survenance et pour chaque année de vision

Ces courbes montrent que plus de 80% des règlements sont effectués au cours de la première année suivant la survenance du sinistre. Néanmoins, on peut toujours distinguer un pourcentage très faible de montant d'indemnisation à partir de la troisième année de vision.

Pour rendre d'autant plus explicite ces résultats, il est possible de les convertir sous forme de diagrammes par année de survenance pour faire ressortir d'autant plus l'accélération de la gestion des contrats d'assurance collectifs.

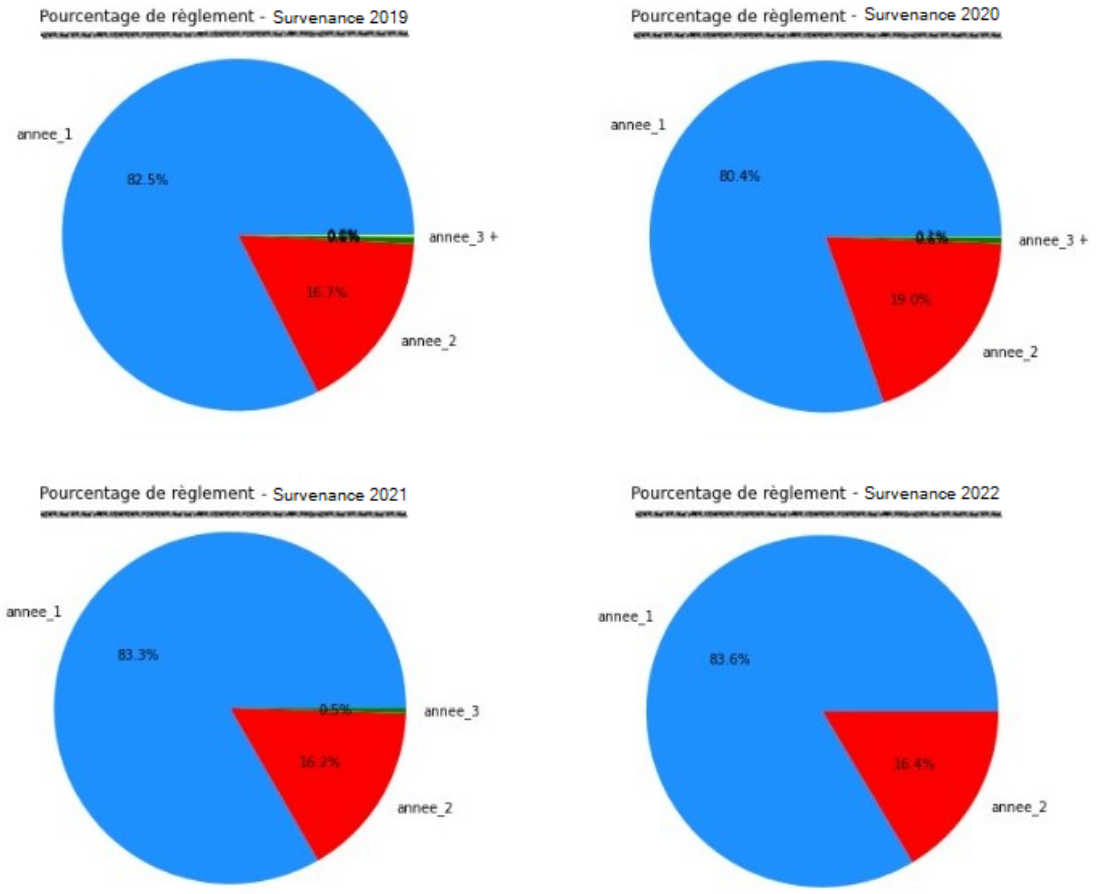


FIGURE 2.9 – Diagrammes sur le pourcentage de règlement par année de survenance

Hormis la survenance 2020 qui semble un peu "à part" dûe au phénomène pandémique, l'accélération de la gestion des contrats d'assurance santé collective est perceptible avec l'augmentation du pourcentage de règlement de la première année de vision.
Cet accroissement au cours des cinq dernières années s'explique par plusieurs facteurs. La transformation numérique et l'adoption de technologies automatisées ont réduit les délais de traitement des sinistres. L'amélioration de l'expérience client et la simplification des procédures administratives ont également contribué à cette accélération. Enfin, les changements réglementaires ont poussé les assureurs à adapter leurs processus, ce qui est visible dans les tendances observées.

Finalement, deux paramètres majeurs permettent de choisir le mois de vision limite : d'un côté, les règlements de santé doivent être réalisés au cours des deux années suivant la survenance des sinistres et cela se confirme grâce aux études statistiques réalisées. De l'autre, quelques montants après le 24ème mois de vision peuvent être décelés. Ils sont probablement dûs à des retards ou à d'autres problèmes spécifiques.

Pour examiner les anomalies visibles après ces 2 ans de projection et afin de garder une vision prudente lors du provisionnement, 36 mois de vision seront retenus.

2.3.4.2 Différences de règlement par type de gestion

En amont de l'exploration des données anormales, il est judicieux d'anticiper de manière statistique les distinctions à réaliser entre les différents segments d'étude afin de les confirmer ou de les réfuter lors de la détection d'anomalies.
Un paramètre important et différenciant est le type de gestion des contrats d'assurance, comme cité précédemment. En sélectionnant 36 mois de vision sur les 5 dernières années de survenance, il est possible de confronter l'évolution des indemnisations entre la gestion directe et la gestion déléguée, que ce soit les variations ou les disparités au niveau des dates de règlement. L'axe des ordonnées sera omis pour ne pas dévoiler d'informations confidentielles.

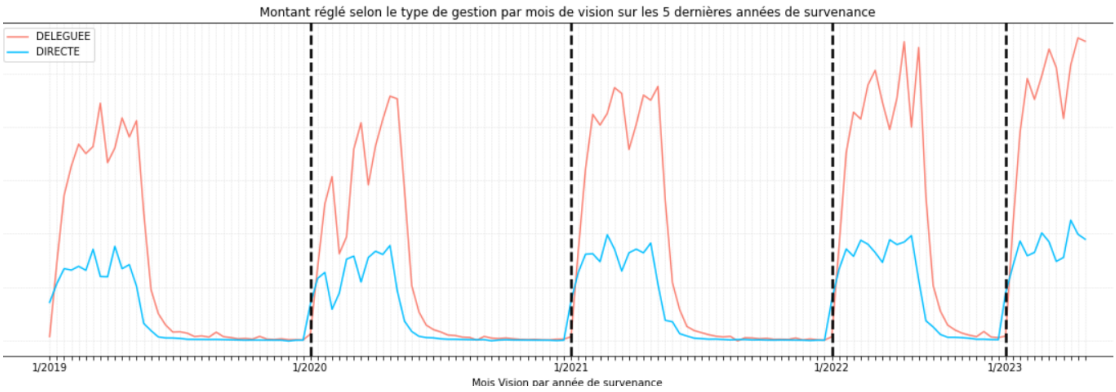


FIGURE 2.10 – Montant réglé sur les 5 dernières années de survenance par type de gestion

Ces deux courbes juxtaposées révèlent plusieurs aspects clés permettant une meilleure compréhension des divergences entre les deux ensembles de données. Le premier élément notable est la similarité concernant les motifs entre la gestion directe et déléguée. Des pics analogues sont visibles sur des périodes très proches pour les deux types de gestion. Bien que les écarts entre ces pics semblent plus importants pour ce qui est de la gestion déléguée, il faut tenir de compte de son volume de règlements plus important.

Cependant, un écart quasiment constant sur la période d'étude est apparent. Il correspond à une avance de deux mois pour les règlements de la gestion directe. En réalité, si un coefficient global sur les règlements en gestion directe était appliqué ainsi qu'une translation de deux mois vers la droite, il serait possible de retrouver de manière quasi équivalente la trajectoire de la courbe rouge. Aussi, plusieurs irrégularités sur la courbe rouge sont décelables après le 18ème mois de vision. Elles sont probablement provoquées par des retards ou des rattrapages au niveau des règlements ce qui semble être moins le cas en gestion directe.

Il faut alors se demander si la principale cause de ces distinctions n'est pas le contraste de vision. Ici, quelque soit le type de gestion sélectionné, la méthode de projection est le mois de vision comptable. Or, le délai de traitement de la gestion des affaires gérée par des intermédiaires est un potentiel facteur venant influencer la date prise en compte de ce règlement ainsi qu'être à l'origine d'erreurs supplémentaires lors des échanges de données.

Il est donc opportun de se pencher sur la vision délégataire, traduite par la variable créée au début de l'étude "mois_vision_delegataire".

2.3.4.3 Comparaison mois de vision comptable et délégataire

La gestion déléguée en assurance santé collective implique une distinction importante entre la vision comptable de l'assureur et celle du délégataire, générant des différences notables dans le traitement et la présentation des données.

Du point de vue comptable, les règlements sont enregistrés en fonction des périodes de reporting, reflétant ainsi le moment où les coûts sont réellement engagés et les transactions comptabilisées. En revanche, le délégataire, qui gère les sinistres au quotidien, enregistre les transactions selon le moment où les sinistres surviennent et sont traités, ce qui peut créer des décalages importants. Par exemple, un sinistre déclaré en fin d'année mais réglé au début de l'année suivante peut être comptabilisé différemment selon le point de vue adopté.

De plus, l'échange de données et la prise en compte comptable de la gestion des contrats par intermédiaire ajoutent un délai supplémentaire. Cette divergence pourrait entraîner des écarts dans la manière dont les données financières sont présentées et analysées.

Mettons en regard les points de vue comptable et délégataire concernant les montants réglés :

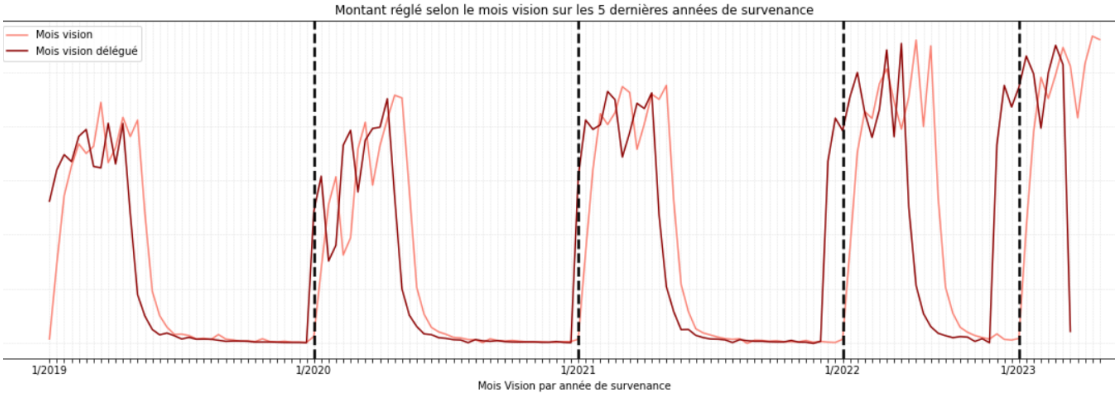


FIGURE 2.11 – Montant réglé par mois vision comptable et délégataire

Le premier élément perceptible sur le graphique, également observable dans le contexte de la gestion directe, est un écart de deux mois concernant les règlements. Il s'agit bien de la durée de traitement de gestion comptable des contrats gérés par les délégataires. Cette différence est donc compréhensible et ne vient pas modifier l'étude des anomalies dans les données.

Néanmoins, il est possible d'observer davantage de divergences du point de vue comptable par rapport à celui du délégataire après le 18ème mois de vision. Cela signifie que des erreurs de gestion viennent s'insérer dans les données du délégataire. Il sera alors essentiel que la détection des données anormales tienne compte de ces données qui pourraient fausser l'estimation finale des provisions.

Ainsi, ces nombreuses conjectures permettent de mieux cibler le périmètre d'analyse et les hypothèses de projection ainsi que d'identifier les théories qui seront ensuite validées ou infirmées lors des applications finales du modèle de provisionnement.

Chapitre 3

Détection des anomalies et correction

3.1 Analyse détaillée des anomalies

3.1.1 Contexte et détermination concrète des distorsions

Avant d'aborder l'étude plus pratique des modèles de détection et de correction des anomalies, il est essentiel de se concentrer sur une analyse approfondie des données à différents niveaux de granularité, notamment à la maille client et à la maille dossier. Cette étape permet non seulement d'identifier les types d'anomalies présentes, mais aussi de comprendre leurs origines. Cette analyse préliminaire est déterminante pour mettre en lumière les défis spécifiques posés par la validité des données dans le cadre de l'assurance santé collective.

Tout d'abord, il est pertinent de se focaliser sur la maille client. Dans le cas de l'assurance collective, le contrat couvre un groupe de personnes, généralement les employés d'une entreprise. Il faut donc regrouper les données à disposition par client pour mettre en évidence certaines données anormales.

Un client spécifique a été retenu car il présente des anomalies pertinentes à analyser. Pour cela, les cinq années de survenance ont été regroupées dans le but de mettre en exergue trois valeurs anormales de règlement sur les 36 mois de vision.

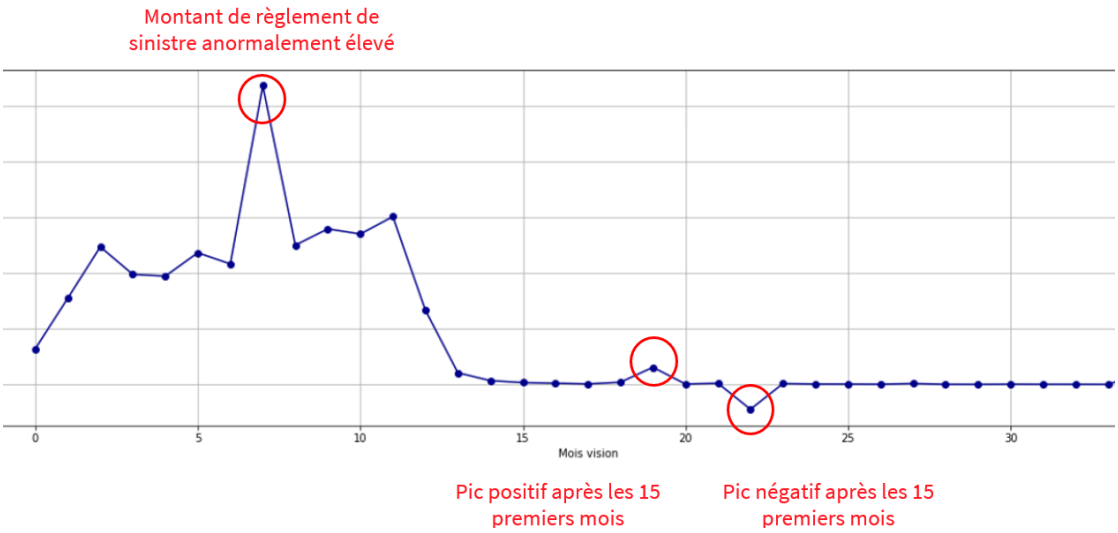


FIGURE 3.1 – Exemple d'anomalies pour un client sur les 5 dernières années de survenance groupées

Pour ce client, un pic environ deux fois supérieur se distingue du reste des montants réglés au cours des 15 premiers mois de vision. Cette valeur extrême semble étonnante à première vue. Cependant, si on détaille les règlements par année de survenance, les 6 premiers mois de vision de la survenance 2019 n'étaient pas comptabilisés. Cela indique que le 7ème mois de vision de la survenance 2019 correspond à un rattrapage des 6 premiers règlements et en fait ainsi une valeur anormale, bien qu'on puisse l'expliquer. De la même manière, un double comptage en gestion est à l'origine du pic positif au 19ème mois de vision. Pour corriger cette erreur, un nouveau rattrapage a été effectué faisant apparaître trois mois après un montant négatif venant annuler ce sur-comptage. Par conséquent, bien que plusieurs erreurs ont été produites puis rectifiées, elles viennent probablement biaiser les résultats du provisionnement.

Si le zoom est effectué à la maille contrat, il est envisageable de comprendre plus en détails l'origine des erreurs de gestion. Prenons l'exemple du contrat de santé suivant et essayons de corriger l'anomalie associée :

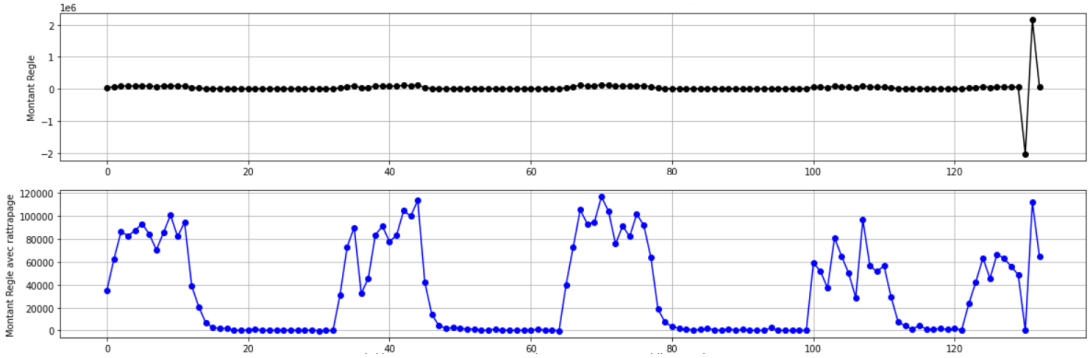


FIGURE 3.2 – Exemple d'anomalies sur un contrat de santé et correction avec rattrapage

Dans cet exemple, sur l'ensemble des règlements des 5 dernières survenances, deux valeurs sont extrêmes car elles sont dix fois plus importantes que le reste. Le premier pic est négatif et le suivant est positif, indiquant qu'il s'agit d'une mauvaise manipulation de la gestion qui a ensuite été rectifiée pour conserver la somme totale des règlements. Le problème est alors la conséquence néfaste sur le provisionnement en santé. Néanmoins, il est possible d'ajuster cet effet : le pic négatif est neutralisé et le pic positif correspond à la somme des deux extrêmes. Ainsi, la courbe bleue montre qu'après correction, aucune valeur incohérente n'est présente et viendrait engendrer des répercussions défavorables sur le calcul des provisions.

Cependant, d'autres cas d'anomalies peuvent être illustrées comme suit :

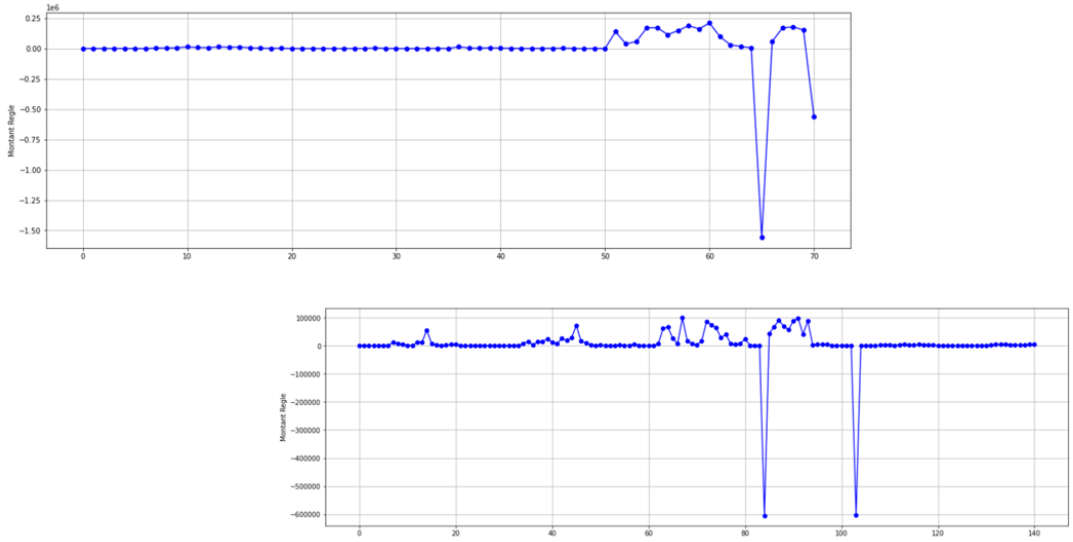


FIGURE 3.3 – Exemples d'anomalies complexes visibles sur des contrats de santé

En l'occurrence, le cas cité lors de l'analyse à la maille client est visible ici. Effectivement, pour le premier contrat, de nombreux montants plus élevés que les précédents sont comptabilisés et un pic vers le bas vient annuler ces règlements. Contrairement au cas précédent, il n'est pas aussi évident de corriger ce phénomène. De plus, ce phénomène peut être identifié pour d'autres contrats de santé, comme le montre le deuxième graphique. Deux pics négatifs corrigent plusieurs sur-comptabilisations de règlements. Les montants ainsi que les périodes de règlements associés diffèrent du premier contrat étudié, ce qui rend la tâche de correction d'autant plus complexe.

En définitive, les problèmes de fiabilité des données identifiés à travers cette analyse soulignent la nécessité de traiter ces anomalies avant l'application d'un modèle de provisionnement. En améliorant la qualité de la *data* à travers des corrections appropriées, il sera possible d'obtenir des provisions qui reflètent plus fidèlement les risques assurantiels.

Dans le cadre de la détection et correction d'anomalies, le périmètre d'étude retenu sera les segments définis par la variable "Segment_Gestion". Cette vision plus globale permettra de simplifier l'application des modèles et de remédier aux défaillances dans les données, tout en sélectionnant des segments représentatifs des problématiques évoquées.

3.1.2 Démonstration de l'importance de la *data* qualité dans la prédiction

En observant ces anomalies à divers niveaux d'analyse, il est pertinent de s'interroger sur leur impact réel sur le calcul du provisionnement actuariel. Dans un premier temps, il est essentiel de démontrer l'intérêt mathématique de la fiabilité des données dans le cadre d'une prédiction.

Soit un modèle de prédiction $\hat{Y}$ basé sur des données X, où $\hat{Y} = f(X) + \epsilon$, avec ϵ représentant l'erreur aléatoire. Supposons que les données X puissent être affectées par des anomalies ou des erreurs représentées par une variable aléatoire A, telles que les données altérées deviennent $X' = X + A$. Le modèle de prédiction sur les données altérées peut alors être exprimé comme $\hat{Y}' = f(X') + \epsilon$.

La variance totale de l'erreur de prédiction peut être scindée en deux composantes : la variance due au modèle et la variance due aux anomalies. Pour les données sans anomalies, la variance de l'erreur est donnée par :

$$\mathrm{Var}(\epsilon) = \sigma^2 \tag{3.1}$$

En présence d'anomalies, sous l'hypothèse que ϵ et A sont indépendants, la variance de l'erreur devient :

$$\mathrm{Var}(\epsilon + A) = \mathrm{Var}(\epsilon) + \mathrm{Var}(A) = \sigma^2 + \mathrm{Var}(A) \tag{3.2}$$

Considérons maintenant l'erreur quadratique moyenne (EQM) d'un modèle sur des données sans anomalies :

$$\mathrm{EQM} = \mathrm{E}[(Y - \hat{Y})^2] = \sigma^2 + \mathrm{Biais}(f)^2 \tag{3.3}$$

Lorsque des anomalies sont présentes, l'EQM devient :

$$\mathrm{EQM}' = \mathrm{E}[(Y - \hat{Y}')^2] = \sigma^2 + \mathrm{Var}(A) + \mathrm{Biais}(f)^2 \tag{3.4}$$

Les anomalies A introduisent alors un biais supplémentaire dans le modèle, cela entraîne : $\mathrm{EQM}' > \mathrm{EQM}$. L'hypothèse de correction des données anormales peut se traduire mathématiquement de la manière suivante : $\mathrm{Var}(A) \to 0$ et donc :

$$\mathrm{Var}(A) \to 0 \quad \sim \quad \mathrm{EQM}' \to \mathrm{EQM} \tag{3.5}$$

Cela montre que la réduction de $\mathrm{Var}(A)$ se traduit directement par une diminution de l'erreur quadratique moyenne, justifiant l'intérêt de la correction des anomalies pour améliorer la précision des prédictions.

Cette modélisation mathématique prouve donc l'intérêt potentiel d'une correction des données anormales et justifie la portée de l'étude dans le cas d'un provisionnement en assurance. L'intention est maintenant de mettre en pratique ces observations.

3.1.3 Zoom sur quatre segments d'étude

Tout au long de la modélisation et de l'application finale de provisionnement, il est nécessaire de choisir des éléments d'analyse pertinents pour apporter des conclusions révélatrices sur l'importance de la fiabilisation de la *data* dans ce contexte.

Pour atteindre cet objectif, quatre segments d'étude ont été retenus. Cette sélection permettra d'obtenir une vue d'ensemble complète et nuancée des anomalies potentielles dans le portefeuille d'assurance santé collective. Les segments choisis sont : un délégataire majoritaire, un délégataire présentant un nombre élevé d'anomalies, les autres délégataires non EDI (non gérés de manière automatisée), et la gestion directe.

Le délégataire majoritaire, que l'on va noté A, a été sélectionné en raison de son poids significatif dans le portefeuille. En effet, il représente une part substantielle des contrats d'assurance santé collective. L'analyse des anomalies pour ce délégataire permettra d'évaluer l'impact des données inexactes à grande échelle et d'obtenir des informations sur les éventuels problèmes systémiques ou généralisés. Il représentera aussi l'échantillon de référence pour paramétrer les modèles avancés de détection.

Le délégataire avec un nombre plus élevé d'anomalies, appelé B, a été choisi pour explorer en profondeur les types et la fréquence des anomalies rencontrées. Ce segment offre un terrain d'étude particulièrement riche pour comprendre la nature des anomalies et tester la robustesse des modèles de détection. L'analyse approfondie de ce segment peut fournir des exemples concrets des défis spécifiques liés à la qualité des données.

Les autres délégataires non EDI, hors du top 20 des délégataires par chiffre d'affaire, sont inclus dans cette étude pour examiner les anomalies dans un contexte où les données ne sont pas gérées de manière automatisée. Cette catégorie est essentielle pour identifier les problèmes spécifiques liés à la gestion manuelle des contrats. Il est important de vérifier comment les anomalies se manifestent dans ces contextes et de tester si les modèles de détection sont adaptés à ces particularités.

La gestion directe, enfin, représente un segment où les données sont traitées en interne sans passer par des délégataires. Ce segment va offrir ainsi une perspective contrastée avec les autres segments. En effet, il est attendu une quantité moindre d'anomalies ce qui validerait les premières conjectures réalisées.

Ces segments ont été choisis non seulement pour leur représentativité dans le portefeuille d'assurance santé collective, mais aussi pour leur capacité à offrir une compréhension détaillée et variée des anomalies potentielles. Ces quatre segments constituent une part non négligeable du portefeuille :

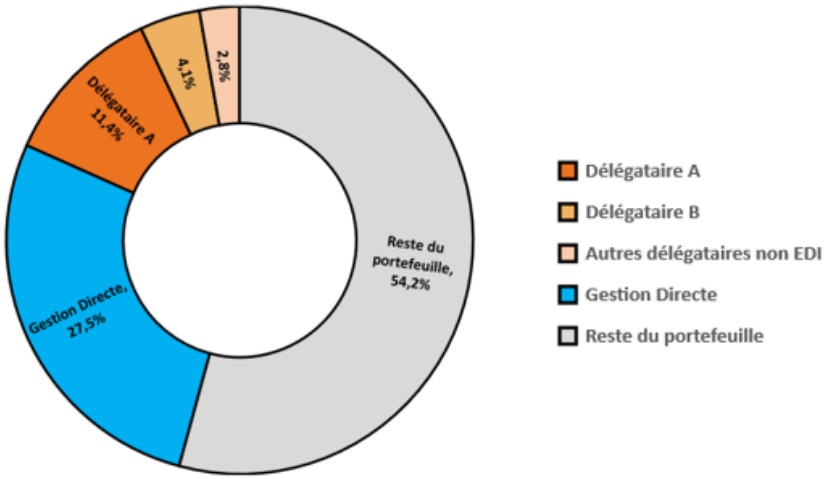


FIGURE 3.4 – Répartition du portefeuille en santé collective

Plus de 45% du portefeuille est couvert au travers de ces quatre exemples. L'étude portera alors sur une part conséquente des règlements en assurance, si bien qu'il faudra être prudent et rigoureux sur les hypothèses et les conclusions apportées.

C'est grâce aux caractéristiques diverses des segments d'étude choisis (une grande envergure, un nombre élevé d'anomalies, une gestion manuelle, et une gestion directe) que des modèles de détection pourront être testés dans des contextes variés afin d'identifier des anomalies de différentes natures.

Cette approche permettra de vérifier les hypothèses formulées lors de l'application des modèles de détection de données anormales.

Les applications seront menées avec des données modifiées afin de garantir la confidentialité des résultats réels. Bien que l'étude ciblera les quatre segments d'étude de manière à confronter les analyses, elle sera applicable sur l'ensemble des éléments constituant le portefeuille d'assurance santé collective.

3.1.4 Méthode statistique de détection des anomalies

Après avoir défini les quatre segments d'étude, il convient d'analyser de manière statistique l'évolution des montants pour chacun d'entre eux et de commencer à décrire les potentielles anomalies détectables. Représentons les montants associés aux exemples sélectionnés :

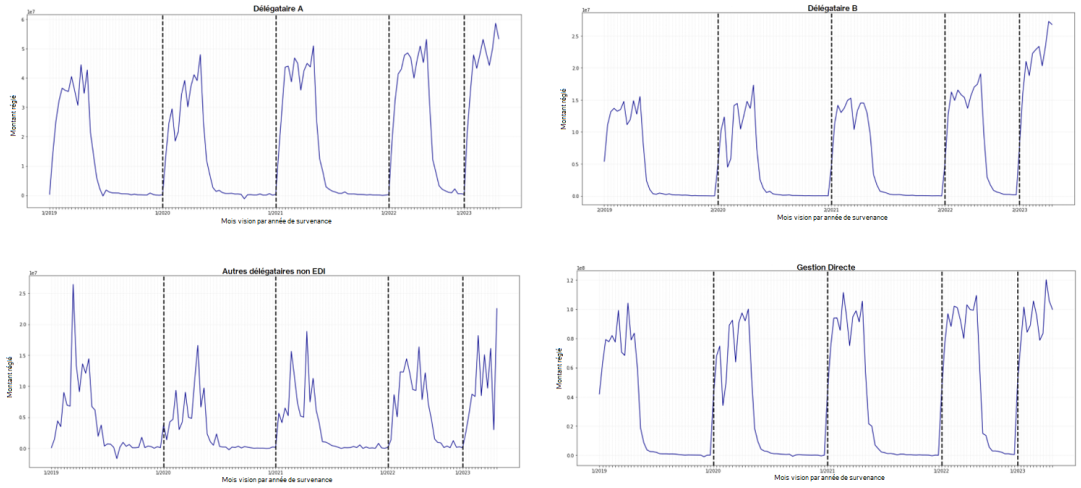


FIGURE 3.5 – Montants réglés sur 36 mois de vision pour les 5 dernières années de survenance, selon les quatre segments d'analyse

Ces quatre graphiques sont les sujets d'application de détection et correction développés dans le cadre de cette étude. Les survenances 2019, 2020 et 2021 sont composées de 36 mois de vision correspondant à la limite définie au préalable. L'extraction des données étant réalisée à date du 31/12/2023, seuls 24 et 12 mois de vision sont disponibles pour les survenances 2022 et 2023 respectivement.

La courbe d'évolution des règlements pour les "autres délégataires non EDI" se détache complètement des trois autres. De très grosses variations de montant sont notables et des irrégularités sont visibles quelque soit l'année de survenance. Cela confirme l'hypothèse d'une quantité plus importante d'anomalies dans ce cas, si bien qu'il faudra se poser la question de l'intérêt d'une détection automatisée pour cet exemple.

A partir de ces courbes, essayons de détecter par calculs statistiques simples les données "étonnantes". Dans ce but, la moyenne des 5 années de survenance permettrait d'obtenir un motif de référence sur 36 mois de vision. En superposant ce motif à ceux des cinq survenances, il serait possible de comparer les écarts par rapport à la moyenne. Cependant, la moyenne ne peut suffire. En effet, une tendance croissante des règlements est remarquable au fil des années quelque soit le segment d'étude choisi.

Pour tenir compte de cette progression, calculons un ratio "Sinistres sur Primes" pour chaque survenance et pour chaque segment d'étude, à l'aide de la variable "CA" représentant le chiffre d'affaire :

$$\text{Ratio de sinistre sur prime} = \frac{\text{Montant des sinistres survenus}}{\text{Primes encaissées}} = \frac{\text{Montant réglé}}{\text{Chiffre d'affaire}} \quad (3.6)$$

En appliquant ce ratio par survenance sur le motif de référence, il est désormais de possible de le mettre en parallèle avec les règlements réels.

L'objectif souhaité est de déterminer les valeurs anormales, qui sont suffisamment éloignées de la moyenne ajustée définie précédemment. Un moyen adapté pour vérifier cet écart est l'estimation d'un intervalle de confiance. Etant donné que la distribution des règlements ne suit pas une loi normale, l'intervalle de confiance peut être défini grâce à la loi de Student. Un intervalle de confiance de 99% sera retenu et il est défini de la façon suivante :

$$IC_{99\%} = \bar{X} \pm t_{0.995,n-1} \times \frac{s}{\sqrt{n}} \tag{3.7}$$

où :
— $\bar{X}$ est la moyenne de l'échantillon,
— $t_{0.995,n-1}$ est la valeur critique de la loi de Student pour 99% avec $n-1$ degrés de liberté (correspondant à 0.995 de la probabilité cumulée),
— s est l'écart-type de l'échantillon,
— n est la taille de l'échantillon.

Avec cet intervalle autour de la moyenne ajustée des règlements, il est possible de repérer quelles sont les valeurs qui en sont exclues et qui représentent donc des anomalies. Toutefois, cette méthode va être employée sur un nombre restreint de mois de vision. Pour repérer des retards, des annulations ou des rattrapages de règlement en santé, l'hypothèse de l'intervalle de confiance est trop forte.

Déterminons alors le mois de vision à partir duquel la stratégie de l'intervalle de confiance ne semble plus cohérente avec les données. Il faut alors calculer la somme des montants par mois de vision sur toutes les survenances afin d'évaluer le pourcentage cumulé par période. Prenons comme exemple de référence le délégataire A :

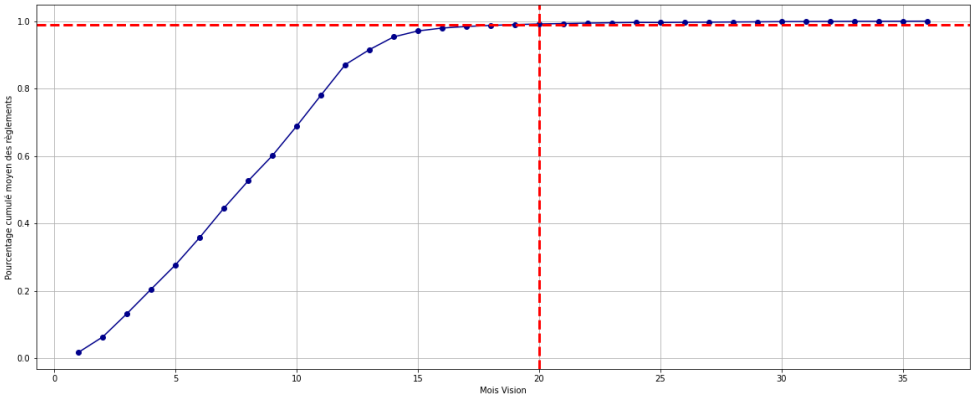


FIGURE 3.6 – Pourcentage cumulé des règlements sur 36 mois pour le délégataire A

De manière évidente, environ 99% des règlements totaux sont effectués avant le 20ème mois de vision. Donc, l'intervalle de confiance sera utilisé sur les 20 premiers mois, et lorsque la trajectoire des règlements s'écarte de la moyenne ajustée après cette période, alors il s'agira d'une anomalie. Pour le délégataire A, le résultat obtenu est le suivant :

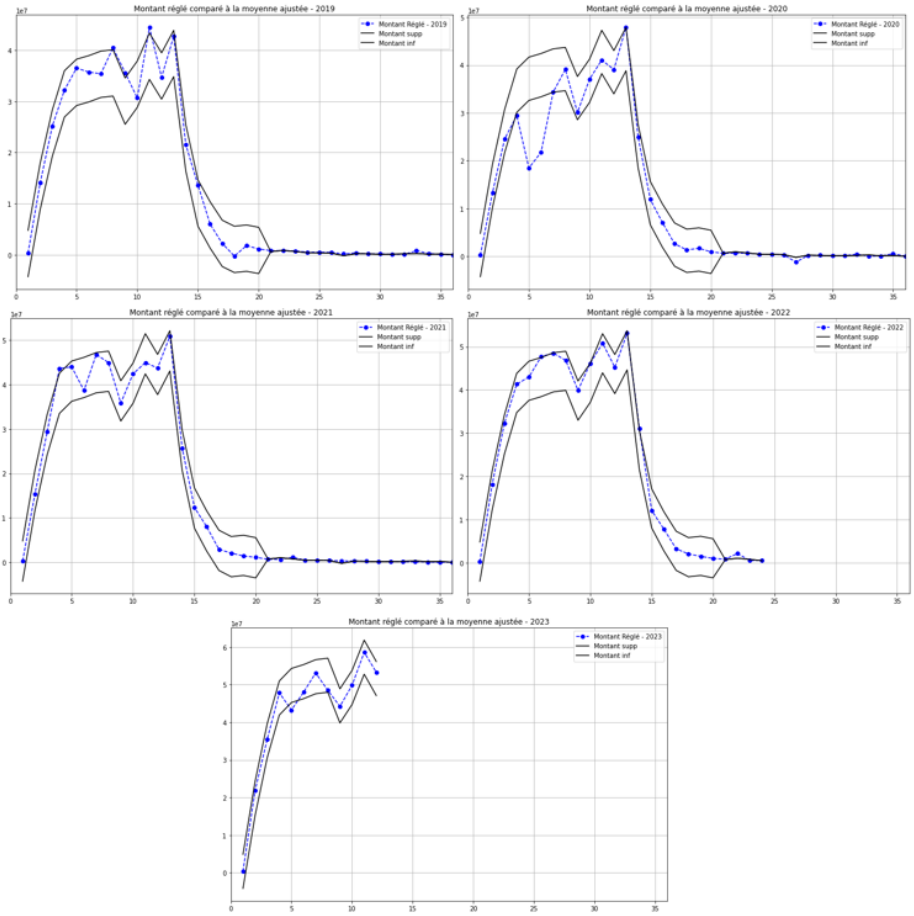


FIGURE 3.7 – Détection statistique des anomalies par survenance sur le segment A

L'intervalle de confiance défini précédemment met en évidence les points à considérer comme anomalie pour chacune des années de survenance. A partir de cette analyse, définissons une nouvelle variable binaire appelée "detect_stat" comme suit :

$$detect\_stat = \begin{cases} 1 & \text{si la valeur est une anomalie} \\ 0 & \text{sinon} \end{cases} \tag{3.8}$$

Pour simplifier la visibilité de ces anomalies, modélisons les à l'aide de cercles verts et l'application statistique sur le délégataire A donne alors le résultat ci-dessous :

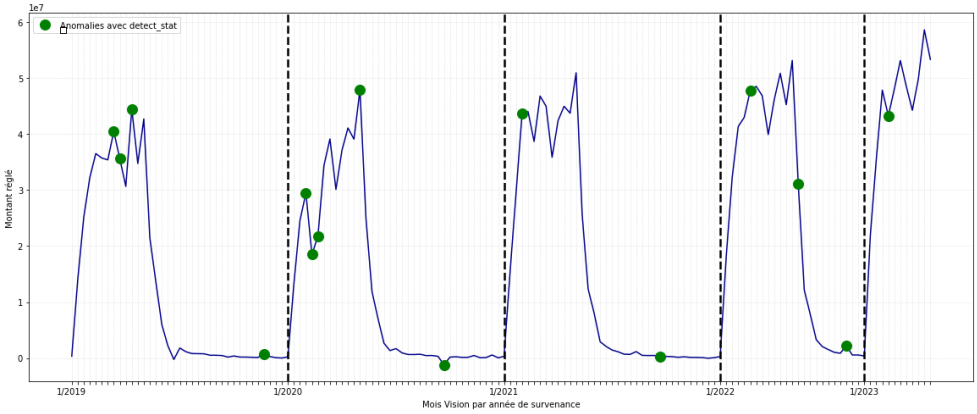


FIGURE 3.8 – Détection statistique des anomalies sur le segment d'étude A

Globalement, 15 anomalies ont été repérées sur les 144 points de données, soit environ 10%. De plus, l'année de survenance qui cumule le plus de données anormales est la survenance 2020 marquée par la pandémie ce qui confirme la théorie évoquée au préalable. Quatre cercles sont situés après la période forte des règlements, indiquant la prise en compte des éventuels retards ou annulation de règlement par le modèle statistique. Analysons alors les trois autres segments pour valider l'hypothèse et comparer la quantité d'anomalies entre ces exemples ;

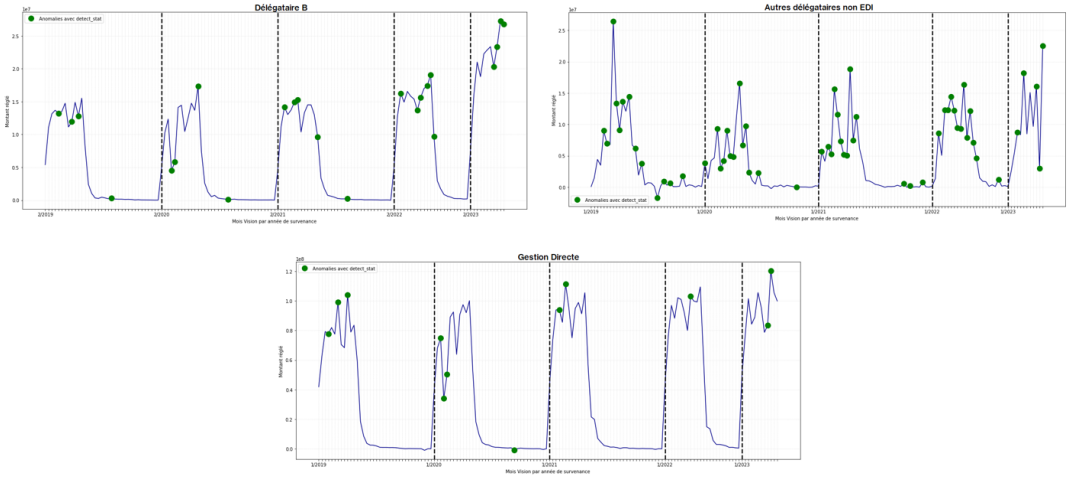


FIGURE 3.9 – Détection statistique des anomalies sur les trois autres segments d'étude

Les conjectures énoncées dans la partie précédente concernant la différence du nombre d'anomalies entre les quatre éléments d'étude sont corroborées. Le délégataire B présente 15% de données anormales, soit une dizaine de plus que le délégataire A. A l'inverse, la gestion directe présente 8 anomalies dont 4 en 2020 ce qui confirme l'augmentation non négligeable d'erreurs lors des échanges entre les mandataires et la comptabilité.

Pour ce qui est des autres délégataires non EDI, l'absence d'une plateforme automatisée génère un nombre d'anomalies considérable ce qui fait remettre en question le modèle statistique dans ce cadre. Trop d'anomalies ont été constatées à cause des variations irrégulières des règlements.

Globalement, les données anormales détectées par l'approche statistique correspondent aux conséquences des erreurs de gestion et des phénomènes conjoncturels sur les règlements de sinistre. Néanmoins, cette technique nécessite de nombreux choix arbitraires afin de sélectionner les anomalies et n'est donc pas réplicable parfaitement sur l'ensemble des segments d'étude.

Par conséquent, il serait plus efficace d'appliquer des modèles de *machine learning* appropriés à la recherche d'anomalies, notamment dans le but d'automatiser cette étape de détection. Il est alors intéressant d'en tester plusieurs pour voir lequel serait le plus adapté aux données de règlement de sinistre en les confrontant au modèle statistique.

3.2 Automatisation par *machine learning*

Les méthodes exposées sont des techniques d'apprentissage non supervisé adaptées à la détection d'anomalies dans un ensemble de données. L'objectif est alors de définir une nouvelle variable binaire, soit 1 lorsque le modèle repèrera une valeur anormale, 0 sinon.

3.2.1 Local Outlier Factor (LOF)

La méthode du Local Outlier Factor (LOF) mesure l'anomalie d'un point de données : si ce point a une densité locale significativement plus faible que ses voisins, alors il s'agira d'une donnée anormale.
Trois concepts clés sont nécessaires pour comprendre le fonctionnement de cette méthode :

- Densité Locale : la densité locale d'un point de données est déterminée par la distance entre ce point et ses voisins les plus proches.
- Densité d'Atteignabilité Locale (*Local Reachability Density*, LRD) : la densité d'atteignabilité locale mesure à quel point un point de données est accessible par ses voisins.
- Facteur d'Isolement Local : le LOF d'un point de données est le rapport de la densité locale moyenne de ses voisins à sa propre densité locale.

Le calcul du LOF se fait en plusieurs étapes :

- Détermination de la distance k-NN : pour un point de données p, la distance est calculée jusqu'à ses k-plus proches voisins. La distance jusqu'au k-ième voisin est notée $d(p, k)$. La distance euclidienne entre deux points $p = (p_1, p_2, \ldots, p_n)$ et $k = (k_1, k_2, \ldots, k_n)$ dans un espace de dimension n est donnée par :

$$d(p, k) = \sqrt{\sum_{i=1}^{n} (p_i - k_i)^2} \tag{3.9}$$

- Calcul de la distance d'atteignabilité : la distance d'atteignabilité d'un point p par rapport à un voisin o est définie comme :

$$\text{reach-dist}_k(p, o) = \max\left(d(o, k), d(p, o)\right) \tag{3.10}$$

où $d(o, k)$ est la distance jusqu'au k-ième plus proche voisin de o, et $d(p, o)$ est la distance entre p et o.

- Calcul de la densité d'atteignabilité locale : la densité d'atteignabilité locale de p est l'inverse de la moyenne des distances d'atteignabilité de ses k-plus proches voisins :

$$LRD_k(p) = \frac{k}{\sum_{o \in kNN(p)} \text{reach-dist}_k(p, o)} \tag{3.11}$$

où $kNN(p)$ est l'ensemble des k-plus proches voisins de p

- Calcul du LOF : le LOF de p est le rapport entre la densité moyenne de ses voisins et sa propre densité :

$$LOF_k(p) = \frac{\sum_{o \in kNN(p)} LRD_k(o)}{k \cdot LRD_k(p)} \tag{3.12}$$

En fonction du résultat obtenu avec le LOF, plusieurs interprétations sont possibles :
— $LOF \approx 1$: le point est dans une région de densité homogène, similaire à celle de ses voisins.
— LOF > 1 : le point a une densité locale inférieure à celle de ses voisins. Plus la valeur du LOF est grande, plus ce point est isolé. Cela indique donc que ce point est potentiellement une anomalie.
— LOF < 1 : le point est situé dans une région plus dense que ses voisins et donc il n'est pas considéré comme une donnée anormale.

La méthode du Local Outlier Factor possède de nombreux avantages :
— Elle prend en compte la densité locale, rendant la méthode efficace pour les données avec des distributions variées.
— Elle est flexible et peut être appliquée à différents types de données et de domaines.

Cependant, ce modèle peut faire apparaître quelques limites :
— La performance dépend du choix de k. Un choix inadéquat de k peut affecter les résultats. Un paramètre trop petit pourrait rendre le modèle trop sensible aux fluctuations locales, tandis qu'un paramètre trop grand pourrait diluer la sensibilité aux anomalies réelles.
— Le calcul peut être coûteux en termes de temps et de ressources pour des ensembles de données très volumineux.

Dans un premier temps, il est nécessaire d'hyperparamétrer le modèle Local Outlier Factor, c'est à dire sélectionner les meilleurs paramètres qui optimiseront le LOF. Pour cela, il faut choisir une métrique de référence pour comparer les modèles et opter pour le meilleur d'entre eux. La métrique de classification la plus adaptée reste le F1-score dans le cas présent. En pratique, il combine deux aspects importants : l'*accuracy* (la proportion de bonnes prédictions parmi les prédictions positives) et le *recall* (la proportion de bonnes prédictions parmi les cas réellement positifs). Il se calcule avec la formule :

$$F_1 = 2 \times \frac{\text{Accuracy} \times \text{Recall}}{\text{Accuracy} + \text{Recall}} \tag{3.13}$$

Un F1-score élevé indique que le modèle fait peu d'erreurs et qu'il détecte bien les éléments positifs. Concrètement, il faut faire varier le paramètre étudié, ici le 'k_neighbors' de la LOF, exécuter le modèle pour chacune de ses valeurs et estimer le F1-score à chaque fois afin de sélectionner le paramètre maximisant cette mesure.

Appliquons cette procédure au délégataire A en ajustant le paramètre k dans la plage 3 à 20 et mesurer le F1-score pour chacune de ces valeurs. Traçons alors la courbe d'évolution du F1-score en fonction de ce facteur :

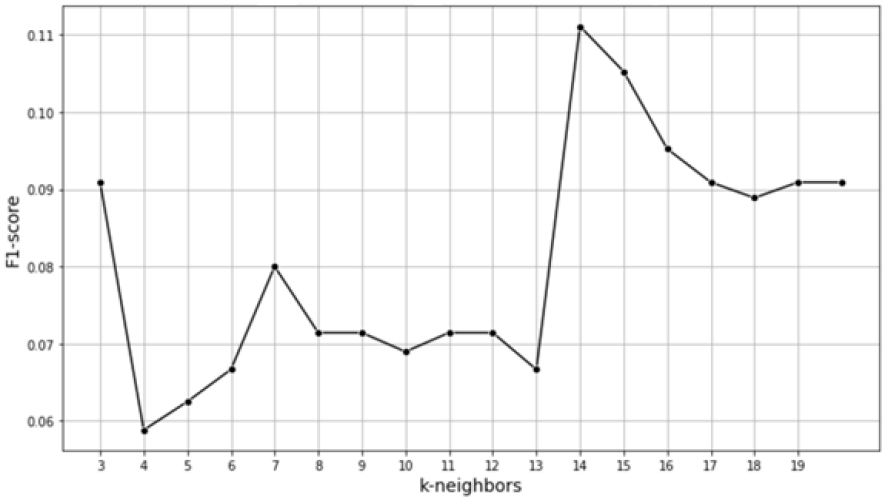


FIGURE 3.10 – Hyperparamétrage de la LOF sur le délégataire A

Bien que les valeurs de F1-score paraissent faibles, la métrique maximale est obtenue pour 'k_neighbors' équivalent à 14. Ce paramètre sera donc retenu pour la détection d'anomalies avec le modèle LOF.

Désormais, il est possible d'employer cet outil de détection sur les différents segments d'étude du jeu de données. Après son exécution, il renvoie une liste de valeurs qui lui semble anormales. En créant une nouvelle variable binaire appelée "detect_LOF" correspondant à 1 si le montant de règlement est dans cette liste, sinon 0, il est alors possible de

tracer la courbe d'évolution des règlements en faisant apparaître en rouge les anomalies. Analysons les résultats dans le cas de l'exemple de référence, le délégataire A.

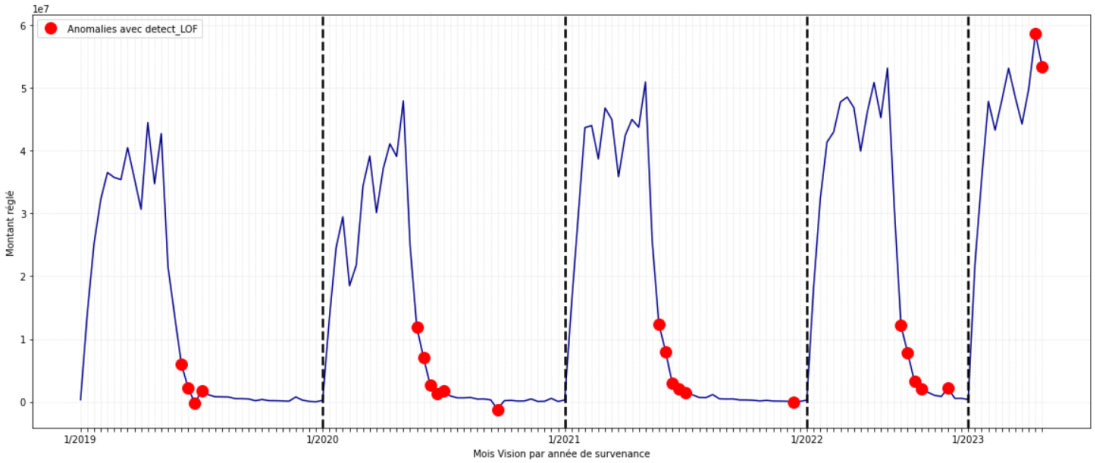


FIGURE 3.11 – Détection des anomalies avec Local Outlier Factor

Les anomalies se répètent presque systématiquement à la fin de chaque cycle, probablement en raison de la chute importante des montants réglés. Cela pourrait signifier que le modèle détecte ces chutes comme anormales par rapport aux montants élevés qui précèdent, bien qu'elles fassent partie d'un cycle naturel. La question pourrait être de savoir si ces périodes de baisse sont effectivement des anomalies du point de vue de l'analyse. Dans le cas du LOF, un point est considéré comme une anomalie s'il est moins dense que ses voisins, ce qui peut expliquer pourquoi les points en fin de cycle sont détectés comme des anomalies. Cependant, ils sont placés majoritairement lors de cycles annuels, ce qui signifierait que le modèle considère à tort ces données comme anormales, malgré l'ajustement de son paramètre 'k_neighbors'. Le caractère "local" de cette méthode empêche l'analyse de données anormales, notamment concernant la comparaison des années de survenance entre elles.

Le Local Outlier Factor est donc remis en question suite à son application sur le délégataire A car il ne semble pas adapté au contexte des règlements en assurance santé.

3.2.2 One-Class Support Vector Machine (OC-SVM)

La méthode des One-Class SVM est fondée sur le principe du *Support Vector Machine* : un algorithme d'apprentissage supervisé qui cherche à trouver un hyperplan séparant les données de manière optimale. Mais il est adapté pour gérer des ensembles disposant principalement de données "normales" sans étiquettes spécifiques pour les anomalies.

Le One-Class SVM cherche à apprendre une frontière qui entoure les données normales, en séparant ainsi les anomalies. Il repose sur la fonction de décision suivante :

$$f(x) = \langle w, \phi(x) \rangle - \rho \tag{3.14}$$

où :
- ϕ est une fonction de transformation qui projette les données dans un espace de haute dimension.
- w est un vecteur normal à l'hyperplan.
- ρ est un biais.

La frontière de décision est déterminée de manière à maximiser la distance entre l'hyperplan et les données normales, tout en minimisant le nombre de points de données en dehors de cette frontière. L'objectif est de trouver cette fonction de décision qui prend des valeurs positives pour les points "normaux" et négatives pour les anomalies. Elle est déterminée de manière à envelopper le plus grand nombre possible de points de données normales dans une région compacte.

D'un point de vue mathématique, cela revient à résoudre le problème d'optimisation suivant :

$$\begin{aligned} &\min_{w,\xi_i,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \\ &\text{sous contraintes} \\ &\langle w, \phi(x_i) \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n \end{aligned} \tag{3.15}$$

où :
- ν est un paramètre contrôlant la fraction de données considérées comme anomalies et l'erreur de classification.
- ξ_i sont les variables d'écart permettant de tolérer certaines erreurs.

One-Class SVM possède plusieurs intérêts :
- Capacité à apprendre une frontière de décision non linéaire grâce à l'utilisation de noyaux (*kernels*),
- Adaptable à des données de haute dimension et des distributions complexes.

Cependant, cette méthode présente quelques inconvénients :
- La performance dépend du choix du noyau et des paramètres (comme ν).
- Elle est sensible aux valeurs anormales si les données d'entraînement contiennent des anomalies.

Il est donc primordial d'hyperparamétrer correctement ce modèle. Tout d'abord, il faut sélectionner le noyau qui lui permet de considérer une frontière de décision non linéaire, adaptée aux données de règlements. Le noyau retenu est celui le plus couramment

utilisé dans ce cadre, le noyau gaussien. Aussi appelé noyau RBF (*Radial Basis Function*), il est formulé de la manière suivante :

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{3.16}$$

où :
— x_i et x_j sont des vecteurs de caractéristiques.
— $\|x_i - x_j\|^2$ est la distance euclidienne au carré entre les deux vecteurs.
— γ est un paramètre qui contrôle l'étendue de l'influence d'un seul point d'entraînement.

Pour maximiser la précision de ce One-Class SVM, il faut optimiser les deux paramètres ν et γ. Avant cela, il faut rajouter une étape supplémentaire : normaliser les données. La formule de la normalisation d'une valeur x_i d'une caractéristique X est donnée par :

$$z_i = \frac{x_i - \mu_X}{\sigma_X} \tag{3.17}$$

où x_i représente la valeur originale de la donnée pour la caractéristique X, μ_X est la moyenne de toutes les valeurs de la caractéristique, σ_X est l'écart-type des valeurs de cette caractéristique, et z_i est la valeur normalisée.

La normalisation permet de mettre toutes les caractéristiques sur une même échelle, ce qui évite que certaines dominent d'autres en raison de leur amplitude. Elle aide également à améliorer la performance des algorithmes de *machine learning*, notamment ceux qui se basent sur la distance ou la corrélation, et peut accélérer la convergence des algorithmes d'optimisation.

Tout comme pour le LOF, estimons le F1-score en faisant varier ses facteurs à l'aide d'une boucle sur les données du délégataire A. Le résultat obtenu est le couple $(\nu, \gamma) = (0.2, 0.05)$.

La variable binaire "detect_ocsvm" peut alors être estimée en considérant les valeurs négatives renvoyées par le modèle et les anomalies repérées peuvent être schématisées :

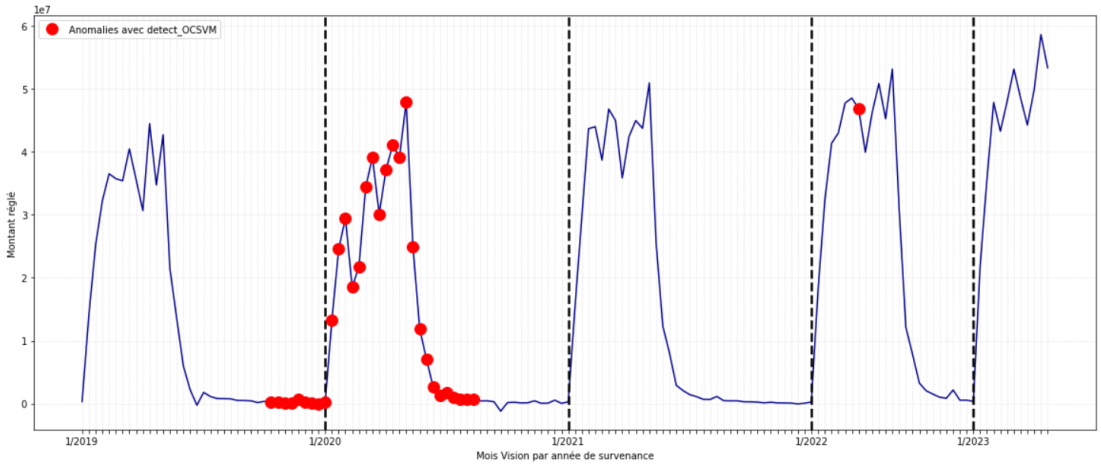


FIGURE 3.12 – Détection des anomalies avec One-Class SVM

Les données anormales définies par le One-Class SVM sont réparties de manière plus diffuse par rapport au modèle LOF. Celles-ci sont particulièrement présentes autour de la survenance 2020. Cela peut indiquer une déviation par rapport au comportement attendu ou à une normalité précédemment observée, probablement dû à l'effet pandémique de cette année sur l'évolution des montants réglés. Le modèle semble donc capturer des variations fines dans les données, ce qui peut être utile pour identifier des comportements inhabituels liés à des événements significatifs. Néanmoins, il reste trop sensible aux conséquences de la crise sanitaire et perturbe la détection d'anomalies attendue.

3.2.3 DBSCAN

DBSCAN (*Density-based Spatial Clustering of Applications with Noise*) est un algorithme de *clustering* fondé sur la densité qui est utilisé pour identifier des *clusters* de forme arbitraire et pour détecter des anomalies dans un ensemble de données.

DBSCAN nécessite deux paramètres principaux : ϵ (epsilon), le rayon de recherche autour d'un point donné et 'minPts', le nombre minimum de points requis pour former un *cluster* à l'intérieur de ce rayon.

Ce modèle fonctionne selon plusieurs concepts clés :

— Point de Base : un point est considéré comme un point de base si au moins 'minPts' se trouvent dans son voisinage de rayon ϵ.

— Point Frontalier : un point est frontalier s'il est dans le voisinage d'un point de base mais n'est pas lui-même un point de base.

— Point Isolé : un point est jugé comme isolé s'il n'est ni un point de base ni un point frontière, il est donc un outlier d'après le modèle.

— Voisinage Direct : le voisinage direct d'un point p est l'ensemble des points situés à une distance de p inférieure ou égale à ϵ.

Traduisons sous la forme d'un schéma ces principes fondamentaux du DBSCAN :

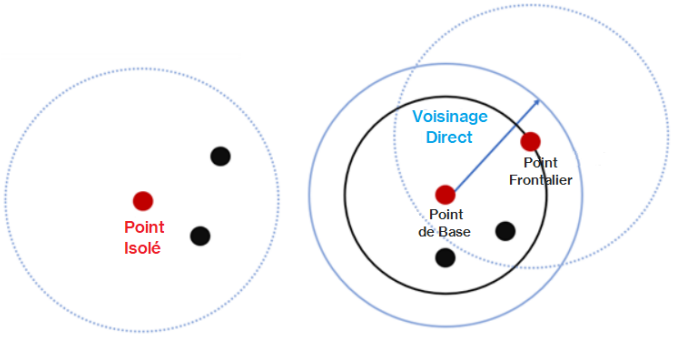


FIGURE 3.13 – Fonctionnement de l'algorithme DBSCAN

Le fonctionnement de l'algorithme DBSCAN se déroule en plusieurs étapes :

- Étape 1 : sélectionner un point non visité p dans l'ensemble de données.
- Étape 2 : récupérer le voisinage direct de p en utilisant la distance ϵ.
- Étape 3 : si le nombre de voisins est supérieur ou égal à 'minPts' , un nouveau *cluster* est créé et p devient un point de base.
- Étape 4 : ajouter tous les points du voisinage direct non encore assignés au cluster. Répéter l'étape 2 pour chaque nouveau point de base ajouté.
- Étape 5 : répéter jusqu'à ce que tous les points soient visités. Les points non assignés à un *cluster* sont considérés comme des points de bruit.

Le *clustering* DBSCAN possède divers intérêts :

— Il peut identifier des *clusters* de forme arbitraire (pas seulement sphériques).
— Il peut gérer les points isolés efficacement en les identifiant comme des anomalies.
— L'algorithme détermine le nombre de *clusters* automatiquement, sans avoir besoin de spécifier le nombre de *clusters* à l'avance.

Néanmoins, il compte plusieurs désavantages :

— Les résultats de DBSCAN dépendent fortement du choix des paramètres ϵ et 'minPts' . Un choix inadéquat peut affecter significativement la performance.
— Il peut avoir des difficultés avec des ensembles de données ayant des densités de *clusters* très différentes.
— La complexité temporelle de DBSCAN est $O(n^2)$ dans le pire des cas, ce qui peut être problématique pour des ensembles de données très volumineux.

Tout comme les modèles précédents, l'étape d'hyperparamétrage est fondamentale avant d'appliquer concrètement le *clustering*. Après avoir normalisé les données et fait une boucle parcourant différentes valeurs pour ϵ et 'minPts' , la matrice de valeurs du F1-score peut se dessiner :

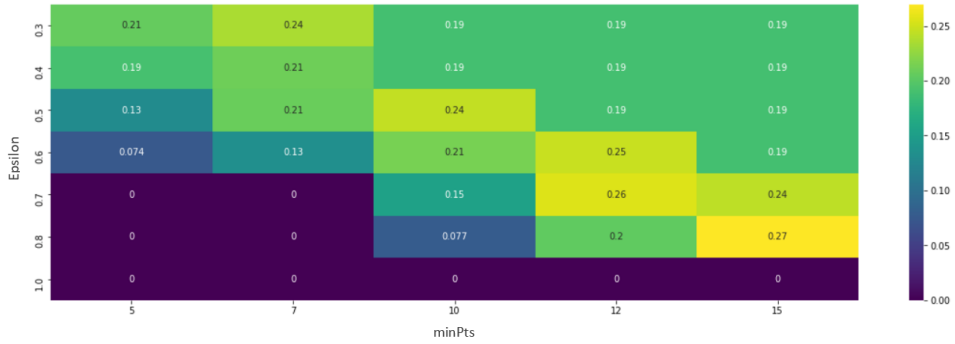


FIGURE 3.14 – Hyperparamétrage du modèle DBSCAN

Le couple $(\text{minPts}, \epsilon) = (15, 0.8)$ maximise la métrique d'évaluation. Cela signifie que DBSCAN recherche des régions de l'espace des données où chaque point a au moins 15 voisins proches pour former un *cluster*, en sachant que pour qu'un point soit considéré comme un voisin, il doit se situer à une distance inférieure ou égale à 0.8 unités.

Appliquons ce modèle hyperparamétré sur les données du délégataire A :

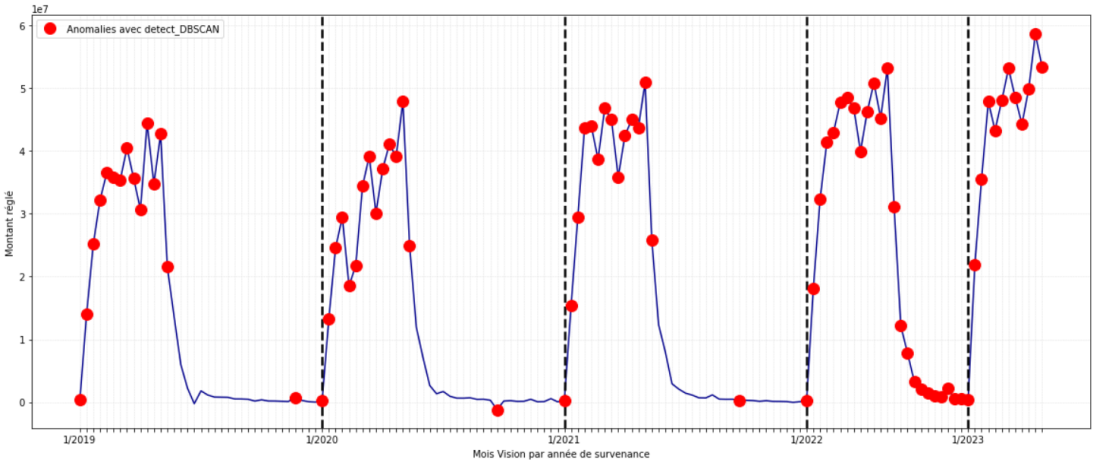


FIGURE 3.15 – Détection des anomalies avec DBSCAN

Les points rouges marquent les anomalies identifiées par DBSCAN. Ici, l'algorithme a identifié un grand nombre de points comme anomalies à travers la série chronologique, y compris plusieurs points dans les phases où les montants réglés sont élevés quelque soit la survenance. Les anomalies sont souvent regroupées près des points de transition où le montant réglé change brusquement. Cela inclut les périodes de forte hausse ou de baisse des montants réglés.

Par rapport à OCSVM et LOF, DBSCAN semble identifier un plus grand nombre d'anomalies, même dans les périodes où la courbe de montant réglé est stable. Cela peut

indiquer qu'il est plus sensible aux variations locales dans la densité des données, ce qui peut entraîner plus de faux positifs ou des détections plus subtiles.

DBSCAN est capable de gérer les formes de *clusters* plus complexes et peut aussi identifier des régions de bruit, ou des points qui ne correspondent à aucun *cluster* dense. Dans le graphique précédent, plusieurs points dans des phases de transitions et certains points isolés en bas de la courbe sont détectés comme du bruit par DBSCAN. Les anomalies identifiées par le modèle dans les phases de déclin pourraient indiquer des événements spécifiques ou des comportements inhabituels, comme des erreurs de paiement, des variations saisonnières inattendues ou des problèmes dans le traitement des données.

Le modèle DBSCAN semble être efficace pour détecter les anomalies dans les données où la densité est une caractéristique clé. Cependant, cette sensibilité accrue et les paramètres choisis rendent le modèle inadapté à une estimation des anomalies proche du modèle statistique réalisé précédemment, le nombre de données anormales repérées étant disproportionné.

3.2.4 Isolation Forest

Isolation Forest est un algorithme d'apprentissage non supervisé qui se base sur l'idée que les anomalies sont plus faciles à isoler que les points normaux dans un ensemble de données.

Il fonctionne sur la base de 2 éléments fondamentaux :
- Isolation : l'algorithme fonctionne en construisant des arbres de décision pour isoler chaque point de données. Les points qui nécessitent moins de coupures pour être isolés sont considérés comme des anomalies.
- Chemin d'Isolation : la longueur du chemin d'isolation pour un point est le nombre de coupures nécessaires pour isoler ce point. Les points avec des chemins plus courts sont plus susceptibles d'être des anomalies.

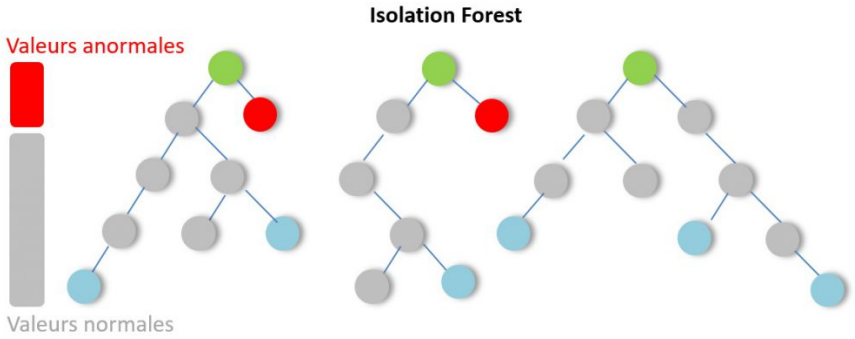


FIGURE 3.16 – Fonctionnement de l'algorithme Isolation Forest

L'algorithme de détection Isolation Forest passe par plusieurs phases :

- Étape 1 : sélectionner un sous-ensemble de l'ensemble de données de manière aléatoire.
- Étape 2 : construire un arbre d'isolation en sélectionnant de manière récursive une caractéristique aléatoire et une valeur de coupure aléatoire jusqu'à ce que chaque point soit isolé.
- Étape 3 : répéter les étapes 1 et 2 pour construire une forêt d'arbres d'isolation.
- Étape 4 : calculer le score d'anomalie pour chaque point basé sur la longueur moyenne du chemin d'isolation à travers les arbres de la forêt.

D'un point de vue mathématique, les notions de coupure et de score d'anomalie se définissent de la manière suivante.

Soit $h(x)$ la longueur du chemin pour un point x, c'est-à-dire la profondeur de x dans un arbre. La hauteur moyenne $\mathbb{E}(h(x))$ d'un arbre binaire complet avec n nœuds peut être approximée par :

$$\mathbb{E}(h(x)) \approx 2 \cdot \log(n) + \gamma - \frac{2 \cdot (n-1)}{n} \tag{3.18}$$

où γ est une constante d'Euler-Mascheroni ($\gamma \approx 0.577$).

La fonction d'anomalie pour un point x est formulée comme suit :

$$s(x, n) = 2^{-\frac{h(x)}{c(n)}} \tag{3.19}$$

où $c(n)$ est la hauteur moyenne des arbres, donnée par :

$$c(n) = 2 \cdot \log(n-1) + \gamma - \frac{2 \cdot (n-1)}{n} \tag{3.20}$$

Les valeurs de $s(x, n)$ sont comprises entre 0 et 1. Une valeur proche de 1 indique que le point x est une anomalie.

Le modèle Isolation Forest montre des atouts notables :

— Il est efficace pour les grands ensembles de données et les données à haute dimension.
— L'algorithme est simple à comprendre et à implémenter.
— Isolation Forest est spécifiquement conçu pour la détection d'anomalies, ce qui le rend robuste et efficace pour cette tâche.

Toutefois, des défauts doivent être soulignés :

— La performance de l'algorithme peut être sensible au choix des paramètres, tels que le nombre d'arbres et la taille des sous-échantillons.
— Comme pour beaucoup de méthodes d'ensemble, il peut être difficile d'interpréter pourquoi un point spécifique est considéré comme une anomalie.

— Isolation Forest peut ne pas être aussi efficace pour détecter des zones denses d'anomalies dans certaines distributions de données.

Pour sélectionner le modèle le plus optimisé aux données, prenons comme référence le délégataire A et faisons varier le nombre d'arbres dans la forêt de décision. La taille des sous-échantillons sera définie de manière automatique par l'Isolation Forest. Le F1-score alors obtenu se rapproche des 70%, ce qui indique une prédiction d'incohérence dans les données assez similaire au modèle statistique. Analysons les anomalies déterminées par l'Isolation Forest sur les données de règlements :

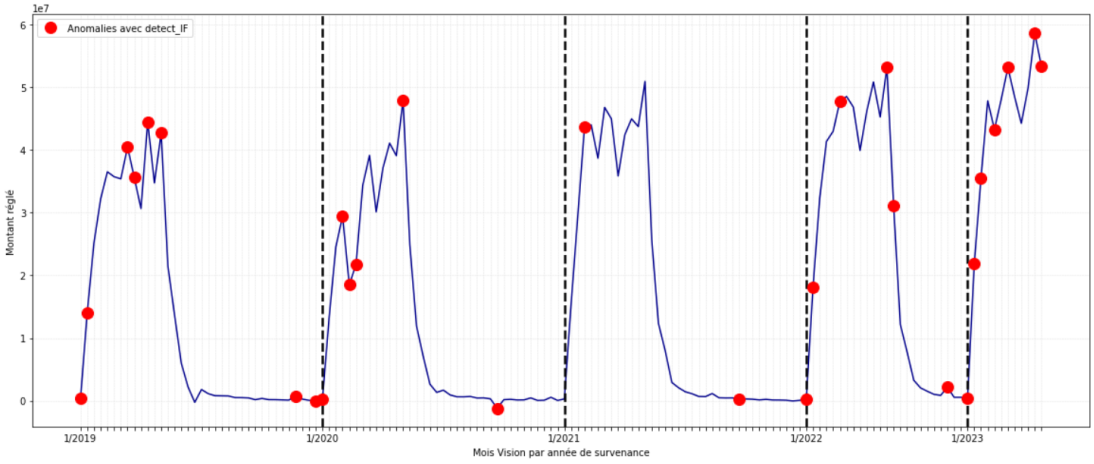


FIGURE 3.17 – Détection des anomalies avec Isolation Forest

Les anomalies sont détectées tout au long de la série temporelle, dans les différentes périodes de temps représentées sur le graphique. Ces points sont identifiés comme des anomalies parce qu'ils sont isolés plus rapidement par les arbres d'isolation, ce qui signifie qu'ils sont inhabituels à l'égard du reste des observations. Par rapport aux autres modèles non supervisés, on peut observer que l'algorithme Isolation Forest détecte des anomalies à des points différents et dans une proportion mesurée.

En effet, cette méthode est moins sensible aux paramètres tels que la distance ou la densité locale et ne nécessite pas de définir un espace ou une norme de distance spécifique. Au lieu de cela, elle utilise la structure hiérarchique pour identifier les anomalies, ce qui peut la rendre plus flexible dans certains scénarios. L'Isolation Forest semble alors identifier efficacement des points qui dévient de la norme de l'évolution des règlements, avec des anomalies détectées à travers plusieurs cycles. Comparé aux autres méthodes, elle peut être plus robuste dans des scénarios où les anomalies sont dispersées de manière non uniforme.

Afin d'explorer une alternative qui pourrait offrir une meilleure précision en tirant parti des données étiquetées grâce au modèle statistique, il est pertinent de se tourner vers des techniques supervisées telles que le XGBoost.

3.2.5 Approche supervisée : XGBoost

XGBoost, ou *eXtreme Gradient Boosting*, est un algorithme de *machine learning* largement utilisé pour les tâches de classification en raison de son efficacité et de sa capacité à gérer des ensembles de données complexes avec des interactions non linéaires. Comme l'objectif de détection d'anomalies revient à définir une variable binaire (1 si la donnée est anormale, 0 sinon), cette méthode semble convenir.

XGBoost est basé sur l'algorithme de *boosting* par gradient, une méthode qui améliore la performance des modèles en combinant plusieurs arbres de décision. Le processus peut être décrit en plusieurs étapes clés :

— *Ensemble Learning* : XGBoost construit un ensemble d'arbres de décision de manière itérative. Chaque nouvel arbre corrige les erreurs des arbres précédents. Les arbres sont construits en minimisant une fonction de perte, ce qui permet d'améliorer les performances globales du modèle.

— *Gradient Boosting* : chaque arbre est ajusté pour réduire l'erreur résiduelle des prédictions. La mise à jour des arbres est effectuée en suivant la direction du gradient de la fonction de perte.

— Fonction de Perte : elle dépend de la tâche spécifique. Pour la classification, on utilise typiquement la fonction de perte logistique pour les problèmes binaires :

$$l(y_i, \hat{y}_i) = -\left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right) \tag{3.21}$$

Chaque arbre dans XGBoost est construit pour minimiser une fonction objective qui combine la perte et la complexité du modèle. La fonction objective L est définie comme suit :

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{k} \Omega(f_j) \tag{3.22}$$

où :

— $l(y_i, \hat{y}_i)$ est la fonction de perte, mesurant l'écart entre la valeur réelle y_i et la prédiction $\hat{y}_i$.

— $\Omega(f_j)$ est un terme de régularisation qui pénalise la complexité du modèle.

— f_j est le j-ième arbre de décision.

XGBoost utilise une approche de "*greedy algorithm*" pour construire les arbres de décision. Cela signifie que l'algorithme sélectionne à chaque étape la division qui entraîne la plus grande réduction de l'erreur de prédiction. Il fait le choix optimal localement dans l'espoir d'obtenir une solution globale.

La construction des arbres implique les étapes suivantes :

- Initialisation : définir une prédiction initiale, souvent la moyenne.
- Calcul des résidus : à chaque itération, les résidus ou les gradients des erreurs prédictives sont calculés pour chaque observation.

- Construction des arbres : chaque arbre est construit en utilisant les résidus calculés pour partitionner les données. Le critère de partition est la réduction de la fonction de perte. La qualité d'une séparation est mesurée par :

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in L} g_i)^2}{\sum_{i\in L} h_i + \lambda} + \frac{(\sum_{i\in R} g_i)^2}{\sum_{i\in R} h_i + \lambda} - \frac{(\sum_{i\in S} g_i)^2}{\sum_{i\in S} h_i + \lambda}\right] \tag{3.23}$$

  où g_i et h_i représentent respectivement les gradients et les hessians pour les données dans le sous-ensemble i. Le gradient g_i est la dérivée première de la fonction de perte par rapport à la prédiction, tandis que le hessian h_i est la dérivée seconde de cette fonction. L et R sont les ensembles de données de gauche et de droite après une séparation, et λ est le terme de régularisation pour contrôler la complexité de l'arbre.
- Mise à jour des prédictions : les prédictions sont mises à jour en ajoutant les nouvelles prédictions des arbres construits.

Les avantages de XGBoost incluent une précision élevée, car il optimise la fonction de perte et réduit les erreurs résiduelles. Il gère bien la complexité des modèles grâce à la régularisation intégrée, ce qui aide à éviter le surapprentissage. XGBoost est également évolutif, capable de traiter de grands ensembles de données de manière efficace grâce à des techniques d'optimisation et de parallélisation. De plus, il offre une grande flexibilité en supportant diverses fonctions de perte, ce qui le rend adapté à différents types de tâches de classification et de régression.

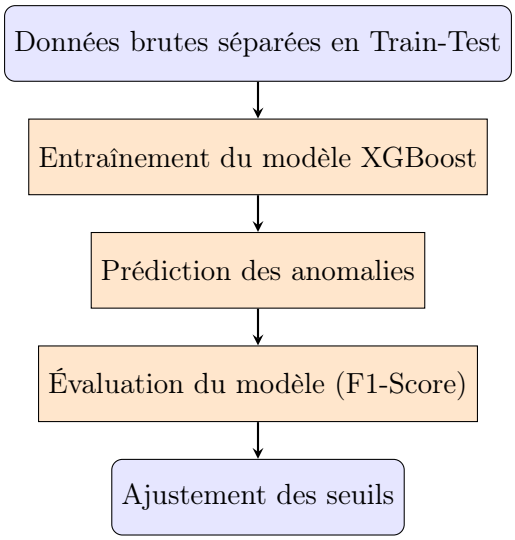
Cependant, XGBoost présente également des inconvénients. L'implémentation et le réglage des paramètres peuvent être complexes ce qui demande une recherche approfondie pour optimiser le modèle. Le temps de calcul peut être élevé, surtout pour des ensembles de données très grands ou très complexes. Enfin, bien que des techniques d'interprétation des caractéristiques soient disponibles, les modèles XGBoost peuvent être moins interprétables comparés à des modèles plus simples comme les arbres de décision uniques.

La détection d'anomalies dans les données de règlement en assurance santé collective à l'aide de XGBoost implique les étapes suivantes :

— Entraînement du modèle : XGBoost est entraîné pour classifier les observations comme normales ou anormales. La détection d'anomalies se fait en apprenant à différencier les observations normales des anomalies à partir des données historiques.

— Évaluation du modèle : la performance du modèle est évaluée en utilisant des métriques, ici le *recall* et le F1-score. En particulier, la capacité du modèle à identifier les anomalies avec un faible taux de faux positifs est cruciale.

Afin de modéliser de manière simplifiée les principes de XGBoost, schématisons les étapes fondamentales pour son bon fonctionnement :



Il est donc nécessaire de faire un split des données en bases d'entraînement et de test. La séparation s'est faite arbitrairement à 75% - 25% respectivement, en vérifiant que la proportion d'anomalies dans la base d'entraînement était non négligeable. Cette étape va permettre d'ajuster le modèle en lui donnant comme référence la variable "detect_stat". Ainsi, avec la base de test, il est possible de vérifier que le modèle prédit de manière assez similaire les anomalies du modèle statistique de référence. Tout comme les modèles non supervisés, il faut sélectionner les paramètres qui optimise le XGBoost. Seuls 3 de ses hyperparamètres seront étudiés : 'learning_rate' (définit la vitesse à laquelle le modèle apprend et ajuste ses prédictions), 'n_estimators' (nombre total d'arbres de décision à construire) et 'max_depth' (nombre maximal de niveaux que chaque arbre peut avoir).

Grâce à ces étapes de mise en place des données et d'hyperparamétrage, le F1-score maximal trouvé est supérieur à 80% pour les paramètres (learning_rate, max_depth, n_estimators) = (0.5, 5, 200). Appliqué aux règlements en santé du délégataire A, le graphique suivant se dessine :

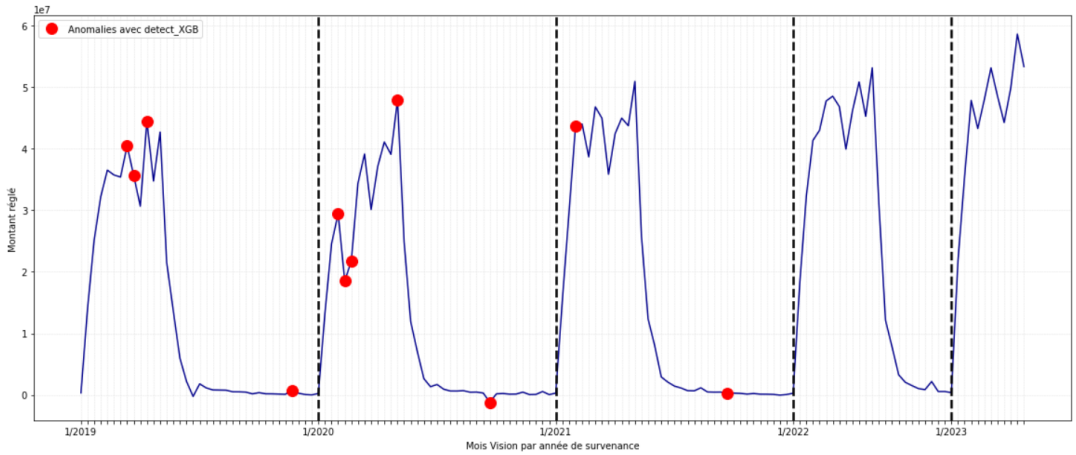


FIGURE 3.18 – Détection des anomalies avec XGBoost

D'après ce graphique, les anomalies sont dispersées à travers différentes périodes, mais elles semblent se concentrer plus particulièrement à certains moments spécifiques, comme au début des années de survenance 2020, 2021 et 2022.

Par rapport aux méthodes non supervisées, XGBoost pourrait détecter moins d'anomalies car il ne décèle que celles pour lesquelles il a été entraîné. En considérant que les données d'entraînement sont bien étiquetées par le modèle statistique, cette approche semble efficace bien que moins flexible dans des situations où les anomalies sont imprévues ou non représentées dans les données d'entraînement.

Comparons alors l'ensemble des modèles testés sur les exemples d'application choisis pour convenir de la meilleure approche de détection.

3.2.6 Comparatif et choix du modèle

Suite à l'hyperparamétrage de chacun des modèles à partir du délégataire A, il est envisageable de répliquer la détection d'anomalies sur les trois autres segments. En calculant le F1-score pour chacun entre eux, il est possible d'établir le tableau récapitulatif suivant :

	Local Outlier Factor	One-Class SVM	DBSCAN	Isolation Forest	XGBoost
Délégataire A	10,53%	21,28%	31,25%	68,18%	84,62%
Délégataire B	28,07%	14,81%	51,69%	70,59%	66,67%
Autres délégataires non EDI	23,19%	22,22%	77,85%	36,78%	93,44%
Gestion Directe	14,29%	13,64%	36,36%	58,54%	66,67%

FIGURE 3.19 – Tableau du score de prédiction des anomalies

Le tableau montre que deux modèles se détachent de manière assez évidente : le XGBoost, représentant l'approche supervisée et donc le modèle avec les meilleurs scores de prédiction mais aussi l'Isolation Forest qui semble être la méthode non-supervisée la plus adaptée à la caractérisation des données dans le contexte des règlements en assurance santé collective. Confirmons l'hypothèse de sélection des méthodes de détection avec la courbe ROC (décrite dans l'annexe A.3) sur le délégataire B qui présente 15% d'anomalies dans ses données :

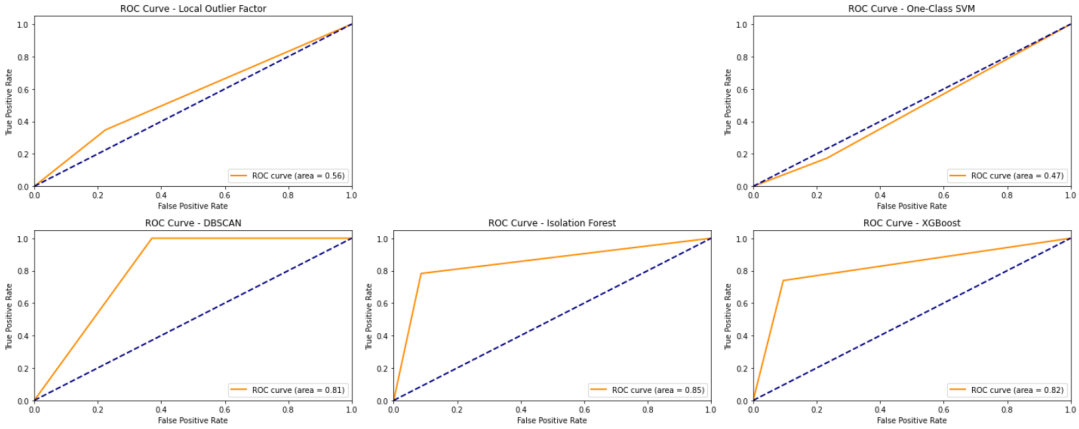


FIGURE 3.20 – ROC-curve des modèles de détection pour le délégataire B

Le but principal de l'étape de détection est l'estimation des vrais et faux positifs, c'est à dire vérifier que le modèle est capable de repérer les anomalies souhaitées. Cela se traduit ici par une courbe qui monte le plus tôt possible, ce qui est le cas du XGBoost ainsi que de l'Isolation Forest.

Finalement, ce sont ces deux approches qui seront conservées pour la phase de détection des anomalies car elles semblent les plus optimales dans le contexte des règlements de sinistres en santé. Leurs différences seront aussi intéressantes à confronter lors du provisionnement, suite à l'étape de retraitement des données qu'il faut alors expliciter.

3.3 Ajustement des données par lissage

Différentes méthodes de lissage ont été testées afin d'atténuer les disparités dans les données et d'éviter un provisionnement "mal" ajusté. L'objectif est de venir diminuer les effets provoqués par les anomalies en remplaçant le montant réel par la valeur lissée. Cette valeur ne doit donc pas effacer les tendances dans les données afin que cette correction reste cohérente avec le contexte des règlements en santé.

Concernant l'un des segments d'étude, les autres délégataires non EDI, les courbes de détection présentées dans l'annexe A.6 prouvent qu'une quantité d'anomalies très importante est repérée. Cependant, dans un contexte de correction pour le provisionnement, cela reviendrait à modifier de manière conséquente les données réelles ce qui n'est pas le but souhaité. L'exemple des autres délégataires, dont la gestion n'est pas automatisée, ne peut donc être utilisé dans cette étude. Dorénavant, l'attention sera exclusivement portée sur les délégataires A et B ainsi que sur la gestion directe, qui représentent plus de 40% du portefeuille des contrats santé, garantissant ainsi la pertinence de cette étude.

3.3.1 Moyenne mobile

La moyenne mobile simple est une technique de lissage qui consiste à calculer la moyenne arithmétique des valeurs d'une série temporelle sur une fenêtre glissante de taille fixe. La formule générale pour la moyenne mobile simple au temps t est donnée par :

$$MMS_t = \frac{1}{n} \sum_{i=0}^{n-1} x_{t-i} \tag{3.24}$$

où x_t est la valeur de la série temporelle au temps t et n est la taille de la fenêtre.

La moyenne mobile simple présente plusieurs avantages, notamment sa simplicité de mise en œuvre et son efficacité pour lisser des séries temporelles avec des tendances ou des cycles. Toutefois, cette méthode ne réagit pas bien aux changements soudains ou aux pics dans les données et nécessite un nombre suffisant de données historiques pour le calcul initial. De plus, les valeurs des premières observations (avant la taille de la fenêtre) ne peuvent pas être lissées.

Dans cette étude, la moyenne mobile simple a été testée avec des fenêtres de 2, 3 et 4 mois. Pour une fenêtre de 2 mois ($n = 2$), la moyenne mobile simple est calculée en prenant la moyenne des deux dernières valeurs. Pour une fenêtre de 3 mois ($n = 3$), la moyenne est calculée sur les trois dernières valeurs, et ainsi de suite pour la fenêtre de 4 mois ($n = 4$). En appliquant ces différentes moyennes mobiles aux règlements totaux de santé sur les 5 dernières années de survenance, les quatre graphiques suivants peuvent être comparés :

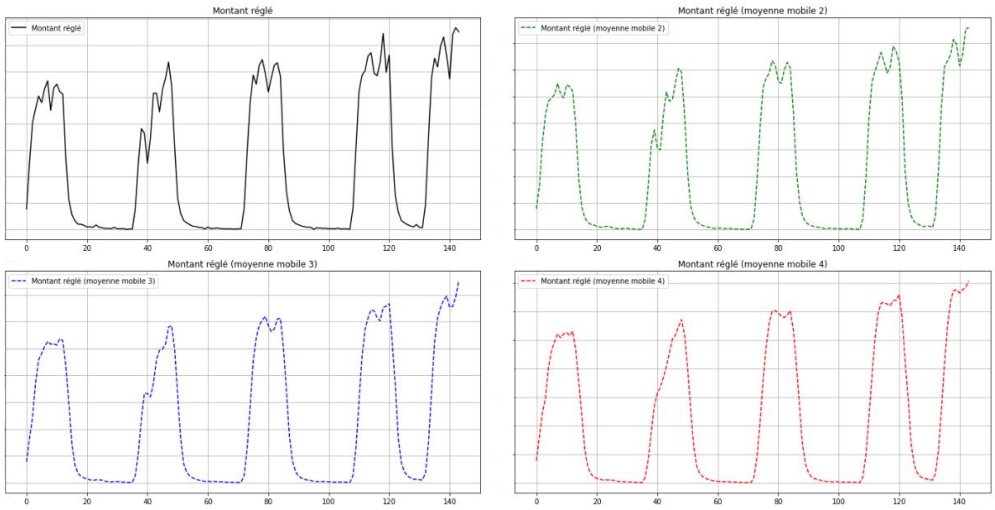


FIGURE 3.21 – Montants réglés lissés par moyenne mobile

Avec la moyenne mobile, des tendances dans les données disparaissent plus la fenêtre de lissage est importante. Même avec une fenêtre de 2 mois, certains pics sont effacés et la structure des règlements est modifiée ce qui n'est pas le but voulu.
En effet, la finalité de l'étape de lissage des données étant de corriger les effets déviants, il est nécessaire que la méthode sélectionnée les gomme tout en conservant les trajectoires des règlements. L'idée n'est pas de venir modifier les données réelles mais simplement d'atténuer les potentielles anomalies qui viendraient impacter les calculs du provisionnement mensuel. Il faut donc expérimenter d'autres modèles de lissage plus complexes et probablement mieux adaptés au contexte de l'étude.

3.3.2 Splines

La méthode des splines est une technique de lissage mathématique qui consiste à ajuster des morceaux de polynômes lisses, généralement de degré faible, à des segments consécutifs de la série temporelle. Les splines peuvent être définies de différentes manières, notamment par des splines naturelles qui imposent des conditions de continuité et de régularité supplémentaires aux extrémités.

Formellement, les splines cubiques naturelles peuvent être définies par morceaux sur des intervalles $[x_i, x_{i+1}]$ comme suit :

$$S(x) = \sum_{i=1}^{n} \left(a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3\right) \cdot I(x \in [x_i, x_{i+1}]) \qquad (3.25)$$

où a_i, b_i, c_i, d_i sont les coefficients du polynôme cubique sur chaque segment $[x_i, x_{i+1}]$, et $I(x \in [x_i, x_{i+1}])$ est la fonction indicatrice qui vaut 1 si x est dans l'intervalle $[x_i, x_{i+1}]$ et 0 sinon.

Les coefficients a_i, b_i, c_i, d_i sont déterminés en imposant des conditions de continuité et de régularité aux nœuds x_i et en minimisant une fonction de pénalisation qui contrôle la souplesse de la courbe.

Les splines offrent plusieurs avantages, notamment leur flexibilité pour ajuster des séries temporelles présentant des variations complexes. En utilisant des morceaux de polynômes cubiques lisses, les splines peuvent capturer efficacement les tendances locales sans sur-ajuster les données. De plus, les splines naturelles imposent des conditions de bord qui garantissent une courbe lisse et continue sur toute la plage des données.

Toutefois, le choix du nombre de nœuds et la gestion des degrés de liberté peuvent influencer la performance des splines. Un nombre excessif de nœuds peut conduire à un sur-ajustement, tandis qu'un nombre insuffisant peut ne pas capturer toutes les variations locales importantes.

Après plusieurs essais, l'application choisie des splines cubiques permet de générer une trajectoire proche de celle des règlements. Confrontons alors la courbe réelle (en pointillés rouge) et celle obtenue après lissage par spline (en bleu) :

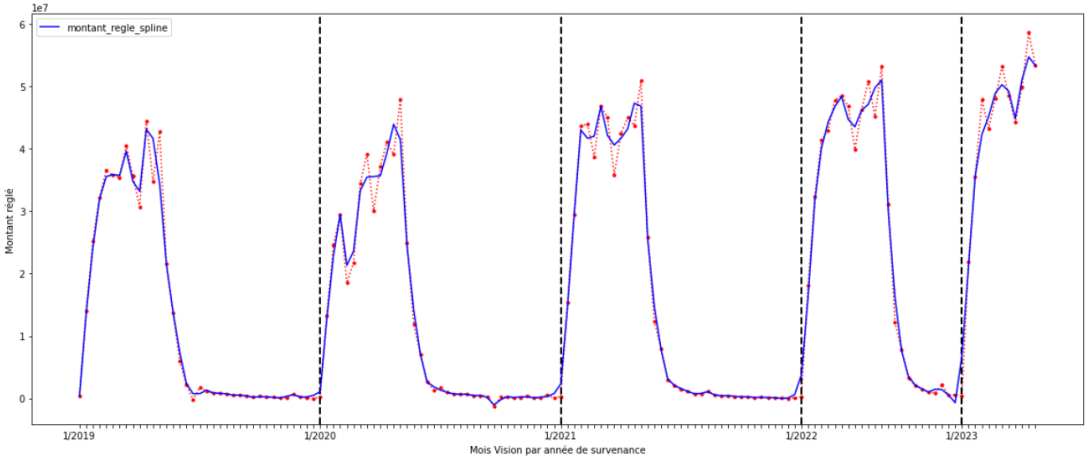


FIGURE 3.22 – Lissage des règlements avec spline

Le lissage par spline semble suivre de très près les valeurs observées. Les courbes lissées s'adaptent bien aux fluctuations des données, y compris les variations brusques. Cette méthode étant très flexible, elle tend à minimiser l'écart avec les données réelles, ce qui est visible ici avec une courbe qui épouse bien les pics et creux. Le lissage par spline pourrait donc atténuer certaines anomalies (valeurs extrêmes) en lissant les points qui dévient de manière importante de la tendance générale. Cependant, il est également capable de conserver certaines anomalies si celles-ci sont soutenues par plusieurs points consécutifs.

Les splines représentent donc un moyen plus complexe mais aussi plus approprié que la moyenne mobile pour lisser les données de règlement en santé, bien qu'il soit difficile à paramétrer pour éviter un sur-ajustement. Une approche souvent utilisée en actuariat, le lissage de Whittaker-Henderson, pourrait être plus adaptée à ce contexte.

3.3.3 Whittaker-Henderson

La méthode de lissage Whittaker-Henderson est une technique utilisée pour ajuster les données de séries temporelles, en particulier dans les modèles de durée. Elle est particulièrement utile pour lisser des séries de valeurs en présence de bruit tout en préservant les tendances sous-jacentes. Cette méthode repose sur la minimisation d'une fonction de pénalisation qui équilibre la fidélité aux données observées et la lissité de la courbe ajustée.

Formellement, le problème d'optimisation peut être exprimé comme suit :

$$\min_{y} \left(\sum_{t=1}^{T} (x_t - y_t)^2 + \lambda \sum_{t=2}^{T-1} (y_{t-1} - 2y_t + y_{t+1})^2 \right) \tag{3.26}$$

où x_t est la valeur observée à l'instant t, y_t est la valeur lissée à l'instant t, et λ est un paramètre de lissage qui contrôle la balance entre la fidélité aux données observées et la lissité de la courbe. Le premier terme de la fonction de pénalisation mesure l'ajustement aux données, tandis que le second terme impose une pénalité sur les variations de la courbe ajustée.

Le paramètre de lissage λ joue un rôle essentiel dans cette méthode. Un faible λ favorise une courbe ajustée qui suit de près les données observées, au risque de capturer également le bruit. À l'inverse, un λ élevé produit une courbe plus lisse qui peut ignorer certaines des variations locales présentes dans les données.

Cette méthode présente plusieurs avantages, notamment sa capacité à produire des courbes ajustées lisses tout en s'adaptant aux tendances sous-jacentes des données. De plus, elle offre une flexibilité grâce au paramètre λ, permettant de contrôler le degré de lissage souhaité. Toutefois, le choix de λ peut être subjectif et nécessite souvent une validation croisée ou d'autres techniques empiriques pour déterminer la valeur optimale.

Pour sélectionner un paramètre adapté aux données, la métrique retenue est l'erreur quadratique moyenne, la MSE (*Mean Squared Error*). La formule de l'erreur quadratique moyenne est donnée par :

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.27}$$

où : y_i représente les valeurs observées, $\hat{y}_i$ représente les valeurs prédites et n est le nombre total de points de données.

Après plusieurs essais, $\lambda = 0.5$ optimise le lissage de ce modèle. Représentons alors les règlements du délégataire A, lissés par Whittaker-Henderson :

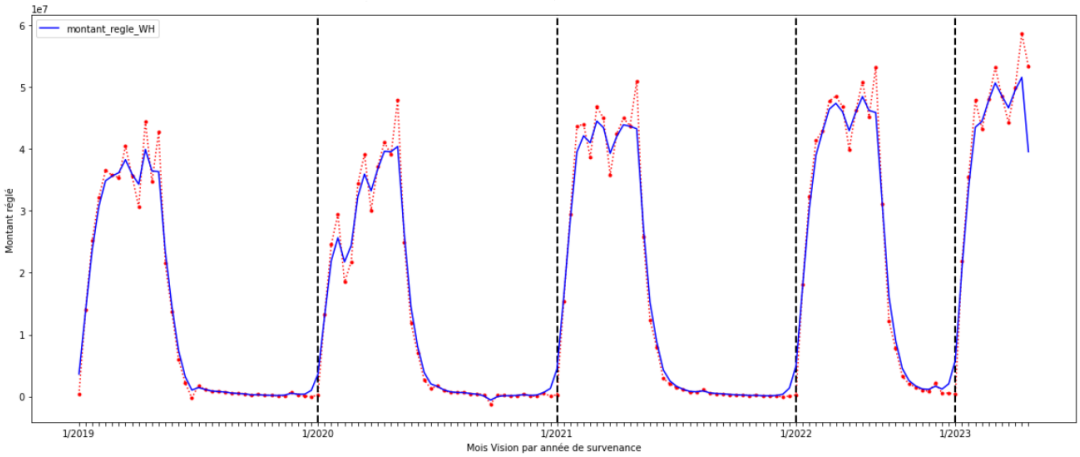


FIGURE 3.23 – Lissage des règlements avec Whittaker-Henderson

La courbe en bleu est beaucoup plus lisse et fluide que celle en pointillés rouges, ce qui indique que la méthode Whittaker-Henderson a réussi à réduire les fluctuations irrégulières et à révéler une tendance générale plus claire. Elle permet de réduire le bruit (les fluctuations irrégulières) tout en préservant les tendances globales. C'est particulièrement utile pour des séries temporelles avec des tendances sous-jacentes sur le long terme.

Néanmoins, en focalisant l'analyse sur les périodes de règlements majeurs des survenances 2019, 2021 et 2022, certains pics sont complètement effacés. En lissant trop les données, Whittaker-Henderson a pour conséquence que certaines variations significatives mais irrégulières soient atténuées ou perdues, ce qui masque des événements ou des anomalies importantes.

Whittaker-Henderson ne semble donc pas optimisée en assurance santé. De plus, au vu de la trajectoire des courbes de règlement, il est cohérent d'évaluer une méthode de lissage adaptée aux séries temporelles.

3.3.4 Triple lissage exponentiel de Holt-Winters

La méthode de lissage exponentiel de Holt-Winters est une technique utilisée pour lisser des séries temporelles et effectuer des prévisions. La méthode se base sur trois équations principales qui mettent à jour les estimations de niveau, de tendance et de saisonnalité.

Pour une série temporelle x_t, les équations de mise à jour de Holt-Winters sont les suivantes :

$$L_t = \alpha(x_t/S_{t-p}) + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{3.28}$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \tag{3.29}$$
$$S_t = \gamma(x_t/L_t) + (1-\gamma)S_{t-p} \tag{3.30}$$

où L_t est le niveau au temps t, T_t la tendance au temps t, S_t la composante saisonnière au temps t, p la période de la saisonnalité, et α, β, et γ sont les paramètres de lissage pour le niveau, la tendance et la saisonnalité respectivement, avec $0 \leq \alpha, \beta, \gamma \leq 1$.

Les prévisions h pas de temps à l'avance sont données par :

$$\hat{x}_{t+h} = (L_t + hT_t)S_{t-p+h} \tag{3.31}$$

Le paramètre α contrôle la réactivité du niveau aux nouvelles observations, β contrôle la réactivité de la tendance, et γ contrôle la réactivité de la saisonnalité.

La méthode de Holt-Winters présente plusieurs avantages, notamment sa capacité à gérer des séries temporelles avec des tendances et des composantes saisonnières. Elle est également flexible et peut être ajustée en fonction des caractéristiques spécifiques des données en modifiant les paramètres de lissage.

Cependant, le choix des paramètres α, β, et γ est fondamental pour obtenir des prévisions précises. Ces paramètres peuvent être déterminés par des techniques d'optimisation ou de validation croisée. De plus, la méthode suppose que la saisonnalité reste constante au fil du temps, ce qui peut ne pas être le cas pour toutes les séries temporelles.

Premièrement, décomposons les règlements du délégataire A afin de faire ressortir les éléments caractéristiques d'une série temporelle.

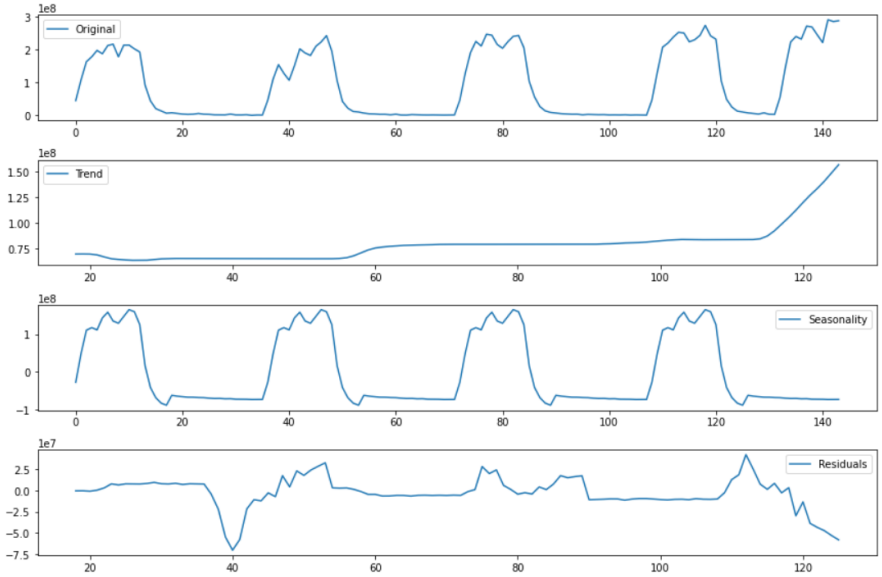


FIGURE 3.24 – Analyse temporelle des règlements du délégataire A

En découpant la série temporelle, il est possible d'identifier la tendance croissante des règlements qui s'explique par l'évolution du chiffre d'affaire associé au délégataire. Un motif cyclique tous les 36 mois de vision est perceptible grâce à la troisième courbe démontrant la saisonnalité des règlements par survenance. Néanmoins, la courbe des résidus justifie une non-stationnarité, avec une moyenne et une variance qui ne sont pas constantes en fonction du temps.

Essayons toutefois d'appliquer ce modèle afin de lisser les données :

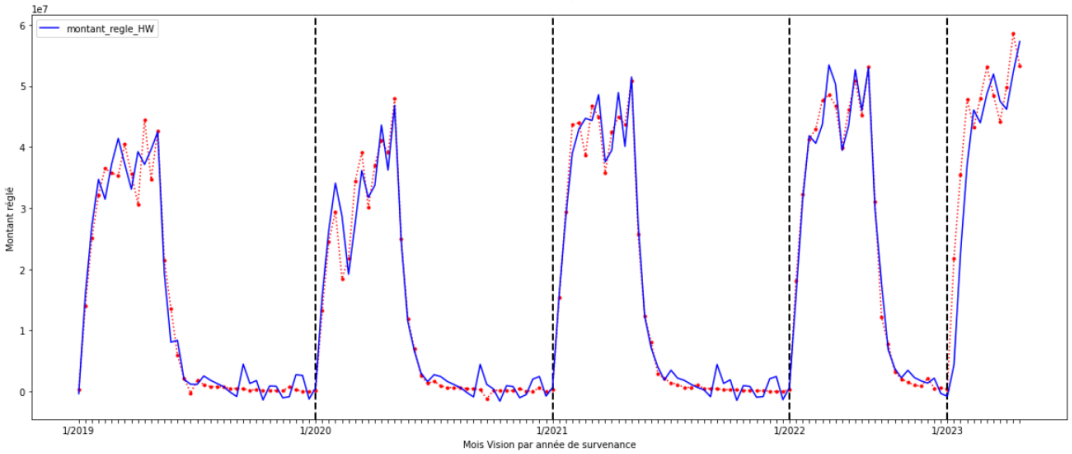


FIGURE 3.25 – Lissage des règlements avec Holt-Winters

Cette méthode capte la tendance générale et la saisonnalité, tout en conservant une certaine sensibilité aux variations à court terme. Comparée aux autres méthodes, la courbe semble suivre les données brutes de plus près, en particulier dans les périodes où les règlements sont les plus importants. Bien qu'elle capture les variations saisonnières, elle peut parfois être sensible aux changements soudains dans les données, ce qui pourrait rendre la prévision plus difficile en présence d'anomalies ou de ruptures structurelles.

Aussi, des pics très faibles sont générés par le modèle après 15 mois de vision pour chaque survenance ce qui n'est pas souhaité et viennent même créer de nouvelles anomalies dans les données. Cela peut être dû aux variations de saisonnalité, étant donné que 2022 et 2023 ont 24 et 12 mois de vision respectivement.

Finalement, l'approche avec des séries temporelles ne semblent pas adaptée au contexte des règlements en santé. Toutefois, le lissage souhaité pourrait être assimilé à un filtre, qui conserverait les tendances globales et limiterait les pics trop importants.

3.3.5 Savitzky-Golay

Le filtre de Savitzky-Golay est un type de filtre passe-bas (utilisé notamment en traitement du signal) qui ajuste une série de points de données à un polynôme local pour réduire le bruit tout en conservant les caractéristiques importantes, Cette méthode de lissage repose donc sur l'ajustement d'un polynôme de degré d sur une fenêtre glissante de taille $2m+1$ autour de chaque point de la série temporelle. Les coefficients du polynôme sont choisis de manière à minimiser l'erreur quadratique entre les valeurs du polynôme et les valeurs observées dans la fenêtre.

Ainsi, pour une fenêtre centrée autour du point t, les coefficients du polynôme $p(t)$ sont déterminés en résolvant le problème d'optimisation suivant :

$$\min_{a_0,a_1,\ldots,a_d} \sum_{i=-m}^{m} \left(x_{t+i} - \sum_{k=0}^{d} a_k (i)^k \right)^2 \tag{3.32}$$

où x_{t+i} est la valeur observée à l'instant $t+i$, et a_k sont les coefficients du polynôme de degré d.

Une fois les coefficients déterminés, la valeur lissée de y_t, notée $\hat{y}_t$, est obtenue en évaluant le polynôme au point central de la fenêtre. La formule générale du lissage Savitzky-Golay s'écrit :

$$\hat{y}_i = \sum_{j=-m}^{m} a_j y_{i+j} \tag{3.33}$$

où :
— $\hat{y}_i$ est la valeur lissée au point i.
— y_{i+j} est la valeur des données à la position $i+j$ dans la fenêtre de lissage.
— a_j sont les coefficients du filtre Savitzky-Golay.
— m est le paramètre qui définit la taille de la fenêtre autour du point i.

L'avantage principal de la méthode Savitzky-Golay est sa capacité à lisser les données tout en préservant les caractéristiques locales importantes telles que les maxima et les minima. Cela est particulièrement utile dans des applications où il est important de conserver la forme des signaux, comme en traitement du signal et en analyse de séries temporelles complexes.

Cependant, le choix des paramètres m (demi-largeur de la fenêtre) et d (degré du polynôme) est déterminant pour la performance de la méthode. Une fenêtre trop large ou un degré de polynôme trop élevé peut conduire à un sur-ajustement, tandis qu'une fenêtre trop étroite ou un degré de polynôme trop faible peut ne pas lisser suffisamment les données.

Après plusieurs applications du modèle avec différentes variantes des paramètres, la MSE minimale a été atteinte pour le couple (m, d) = (5, 4). Vérifions que le filtre s'adapte bien aux montants réglés associés au délégataire A :

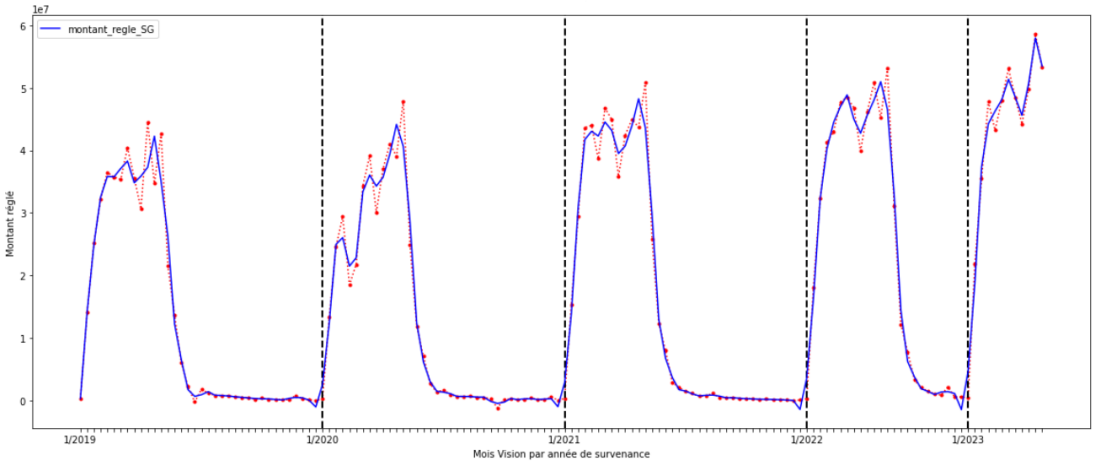


FIGURE 3.26 – Lissage des règlements avec Savitzky-Golay

La courbe lissée (en bleu) est plus souple que les données réelles, mais corrige de manière moins brutale que le modèle de Whittaker-Henderson. Cela indique que la méthode Savitzky-Golay conserve plus le détails des fluctuations locales tout en atténuant le bruit. Bien que le filtre présente un lissage moins rigide, le filtre ajuste les données tout en préservant les éléments significatifs, comme les pics et les creux, ce qui est essentiel pour identifier des fluctuations récurrentes ou des anomalies locales.

La méthode de lissage Savitzky-Golay, appliquée à ces données de règlement en santé, permet de repérer des changements soudains ou des irrégularités ce qui justifie son utilisation dans le cadre de la correction des anomalies. Le filtre de Savitzky-Golay reste moins efficace pour lisser les tendances globales mais offre une meilleure réactivité aux variations locales et conserve une fidélité aux données réelles. Il paraît être le choix le plus cohérent avec l'objectif de correction de données anormales.

3.3.6 Comparatif des méthodes de lissage des données

Avant de vérifier l'application du lissage sur le délégataire B et la gestion directe, confirmons en premier lieu les conclusions réalisées pour chacun des modèles. Pour cela, la qualité d'un modèle de lissage peut être mesurée par différentes métriques présentées en annexe B.1. En les calculant pour chacun des modèles, le tableau de valeurs du délégataire A peut être représenté :

	methode_lissage	MAE	RMSE	R2	MAPE
1	Moyenne Mobile	1.34e+06	2.36e+06	0.90	52.96
2	Whittaker-Henderson	5.82e+05	1.07e+06	0.98	938.59
3	Savitzky-Golay	4.52e+05	7.78e+05	0.99	458.17
4	Holt-Winters	1.54e+06	2.31e+06	0.91	1132.25
5	Spline	3.73e+05	6.70e+05	0.99	755.76

FIGURE 3.27 – Tableau des métriques de performance pour le lissage des données du délégataire A

Le filtre Savitzky-Golay semble, en lissant les pics et creux des données, conserver des données proches et fidèles aux règlements réels de santé. Whittaker-Henderson et les splines sont aussi des approches dont les résultats sont satisfaisants dans le cadre de cette étude. Si cette analyse est répliquée sur les deux autres segments, il pourra être intéressant de confronter les résultats obtenus. Pour le délégataire B, traçons le graphique de dispersion des données après lissage pour chacune des méthodes testées :

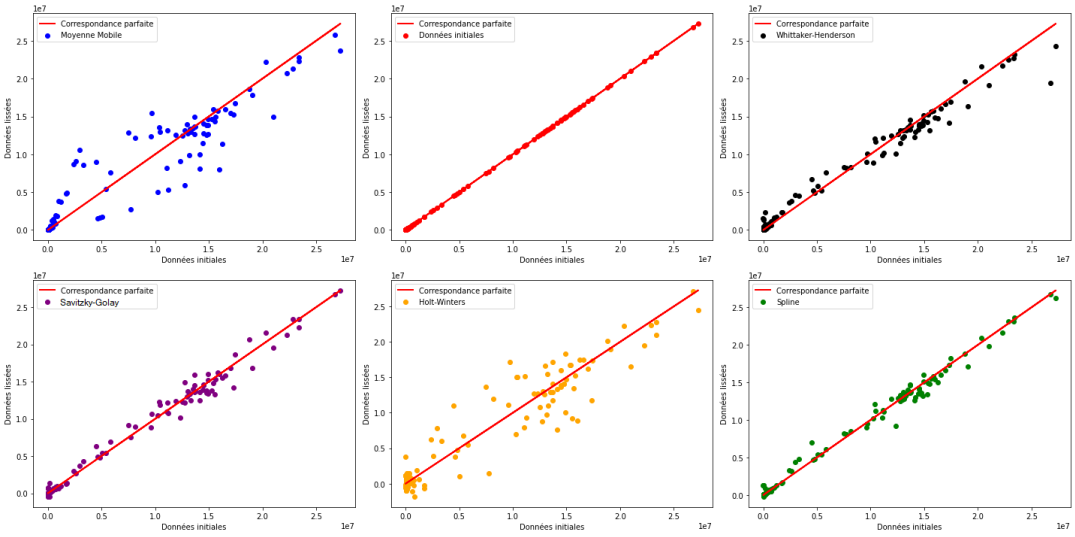


FIGURE 3.28 – Graphiques de dispersion du lissage des règlements du délégataire B

Un graphique de dispersion permet d'analyser visuellement les relations entre deux variables quantitatives. En observant la disposition des points, on peut déduire des corrélations, des tendances, et identifier des valeurs étonnantes ou des schémas complexes qui méritent une analyse plus approfondie. Si les points sont proches de la diagonale en rouge, cela indique une forte relation linéaire entre les données réelles et lissées. Le résultat attendu est des points qui sont globalement proches de cette ligne, ce qui signifie que le modèle de lissage ne crée pas d'écarts importants avec les données réelles. Il

est souhaité aussi que ces points dispersés soient autour de la ligne et qu'il y en ait un nombre équilibré au dessus et en dessous de la diagonale pour éviter un lissage qui vient en moyenne diminuer ou augmenter les montants de règlement.

Cette vision permet de confirmer une nouvelle fois l'efficacité du filtre ainsi que des splines, alors que Whittaker-Henderson possède de nombreux points en dessous de la courbe. La moyenne mobile et le lissage exponentiel des séries temporelles viennent altérer les données de manière trop importante.

Enfin, mettons en parallèle les graphiques des règlements de la gestion directe pour vérifier la performance des modèles :

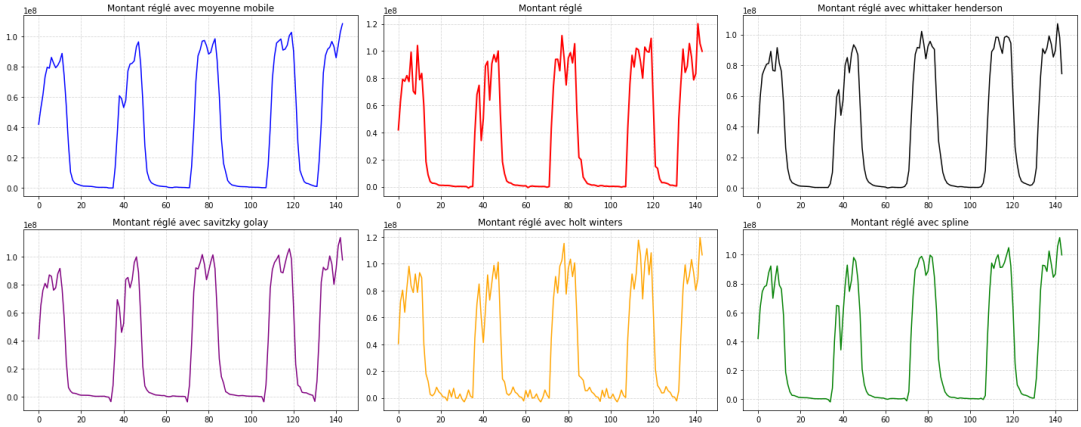


FIGURE 3.29 – Comparaison des modèles de lissage sur les montants associés à la gestion directe

Des contrastes apparaissent entre chacune de ces courbes et viennent conforter le choix final du modèle de correction :

- La moyenne mobile (3 mois) gomme de nombreux pics et affectent de manière disproportionnée l'évaluation des montants
- Whitakker-Henderson efface certaines tendances lors des survenances 2021 et 2022
- Le lissage exponentiel de Holt-Winters crée une nouvelle fois du bruit après 15 mois de vision, probablement dû aux changements de saisonalité
- Les splines et le filtre de Savitzky-Golay semblent adéquats pour le lissage des données, en supprimant certaines variations trop importantes tout en gardant une certaine authenticité des règlements.

Finalement, le modèle de correction sélectionné est le filtre de Savitzky-Golay, qui semble être adapté à la diminution des bruits dans les règlements en santé.

Après avoir modélisé les phases de détection (Isolation Forest pour la méthode non supevisée et XGBoost pour une approche supervisée) et de correction (filtre de Savtizky-Golay), il est désormais possible de lisser les anomalies et d'analyser la fiabilisation des données dans le cas pratique du provisionnement actuariel.

Chapitre 4

Effets du retraitement des données sur le provisionnement

4.1 Théories sur l'estimation des provisions

Le provisionnement en assurance, en particulier en assurance santé, est un processus indispensable qui consiste à évaluer et à mettre de côté des fonds pour couvrir les engagements de l'assureur. Ce processus est essentiel pour assurer la stabilité financière de l'assureur et pour honorer les engagements envers les assurés. En assurance santé, où les marges sont souvent étroites en raison de la forte concurrence, du contrôle des coûts et de la nature imprévisible des dépenses de santé, un provisionnement précis est d'autant plus critique.

Le provisionnement actuariel constitue un exercice difficile car il faut donc estimer de manière rigoureuse les provisions techniques nécessaires pour couvrir les engagements futurs des assureurs. Ce processus repose sur l'analyse statistique des données historiques et sur l'application de modèles mathématiques complexes pour prédire les coûts futurs des sinistres et des prestations. Cette prédiction doit se révéler fidèle à la réalité mais aussi prudente, pour éviter un sous-provisionnement.

En effet, un sous-provisionnement (lorsque les provisions sont inférieures aux montants nécessaires pour couvrir les engagements) peut mettre l'assureur en difficulté, compromettant sa capacité à payer les sinistres. Cela peut conduire à des pertes financières importantes, une érosion de la solvabilité et, en conséquence, à une perte de confiance des assurés et des régulateurs.

À l'inverse, un sur-provisionnement (lorsque les provisions sont excessivement prudentes, dépassant les montants nécessaires) peut sembler être une approche plus sûre, mais il a également des impacts négatifs. Un sur-provisionnement immobilise inutilement des fonds qui pourraient être utilisés plus efficacement, par exemple pour des investissements, le développement de produits ou des améliorations de services. Cela peut aussi entraîner une réduction artificielle des bénéfices, affectant la rentabilité apparente de

l'entreprise, ce qui peut à son tour avoir des implications pour les actionnaires, la valorisation de l'entreprise et sa capacité à attirer des capitaux.

La théorie du provisionnement repose sur plusieurs méthodes actuarielles, l'objectif principal étant d'estimer les provisions pour sinistres survenus mais non déclarés, désignées par l'acronyme IBNR (*Incurred But Not Reported*). Les IBNR sont particulièrement importantes en assurance santé et dans d'autres branches où il peut y avoir un délai entre la survenance de l'événement et la déclaration du sinistre. Sa formule générale de calcul s'écrit :

$$\text{IBNR} = \sum_{i=1}^{n} (C_i - P_i) \tag{4.1}$$

où C_i est l'estimation des sinistres cumulés à la date d'évaluation pour la période i, et P_i est le montant des sinistres déclarés à la même date pour la période i.

Pour mener à bien ces calculs, la qualité des données doit être garantie. En effet, il est primordial que les données soient les plus fidèles aux risques associés afin d'estimer au mieux les provisions. Modélisons alors les processus de détection-correction pour obtenir des triangles de règlement corrigés.

4.2 Arbre final d'ajustement des données

Tel que validés par l'inventaire, les montants réglés cumulés par survenance ne peuvent être modifiés. L'étape de retraitement des données après lissage doit alors atténuer les effets des données anormales tout en conservant les règlements obtenus en fin de période de vision. Pour mieux identifier cette problématique, formulons-le mathématiquement :

Soit T un tableau de données constituées des montants non cumulés $c_{i,j}$ par année de survenance i et par mois de vision j. Soit T' le tableau des données corrigées après application du modèle de détection-correction. Alors, pour certains couples (i,j) donnés, $c_{i,j}$ est remplacé par sa valeur lissée $c'_{i,j}$ tel que :

$$c'_{i,j} = c_{i,j} + \epsilon_{i,j} \quad \text{avec} \quad \epsilon_{i,j} \neq 0 \quad \text{si} \quad c_{i,j} \text{ est une anomalie} \tag{4.2}$$

Cependant, la détection-correction des données anormales vient modifier certaines valeurs et entraîne :

$$\forall i \in [\![1, n]\!], \quad \sum_{j=1}^{n} c'_{i,j} \neq \sum_{j=1}^{n} c_{i,j} \tag{4.3}$$

Cette inégalité implique que les montants cumulés de chaque année de survenance seront différents du cumul des règlements initiaux après correction. Ceci n'est pas souhaitable pour le provisionnement car cela reviendrait à reconsidérer la valeur comptable des règlements. Il est donc nécessaire d'appliquer un coefficient d'ajustement, noté k_i tel que :

$$\forall i \in [\![1, n]\!], \quad \sum_{j=1}^{n} k_i \times c'_{i,j} = \sum_{j=1}^{n} c_{i,j} \quad \Leftrightarrow \quad k_i = \frac{\sum_{j=1}^{n} c_{i,j}}{\sum_{j=1}^{n} c'_{i,j}} \tag{4.4}$$

Ce facteur d'ajustement doit alors être estimé pour chacun des segments d'étude du portefeuille d'assurance santé collective et pour chaque année de survenance. Cela permet d'obtenir des triangles de règlements cumulés qui sont appropriés et comparables à ceux obtenus à partir des données initiales.

Le schéma suivant synthétise les étapes de modélisation :

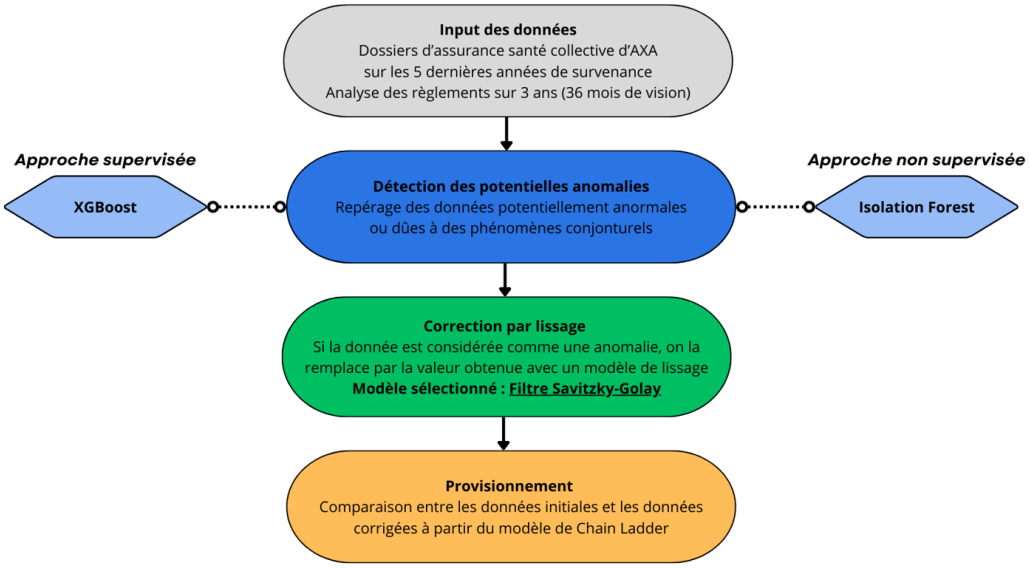


FIGURE 4.1 – Fonctionnement du modèle de détection-correction des données

En reprenant l'exemple de référence avec le délégataire A, l'application des méthodes ci-dessus donne pour les deux approches :

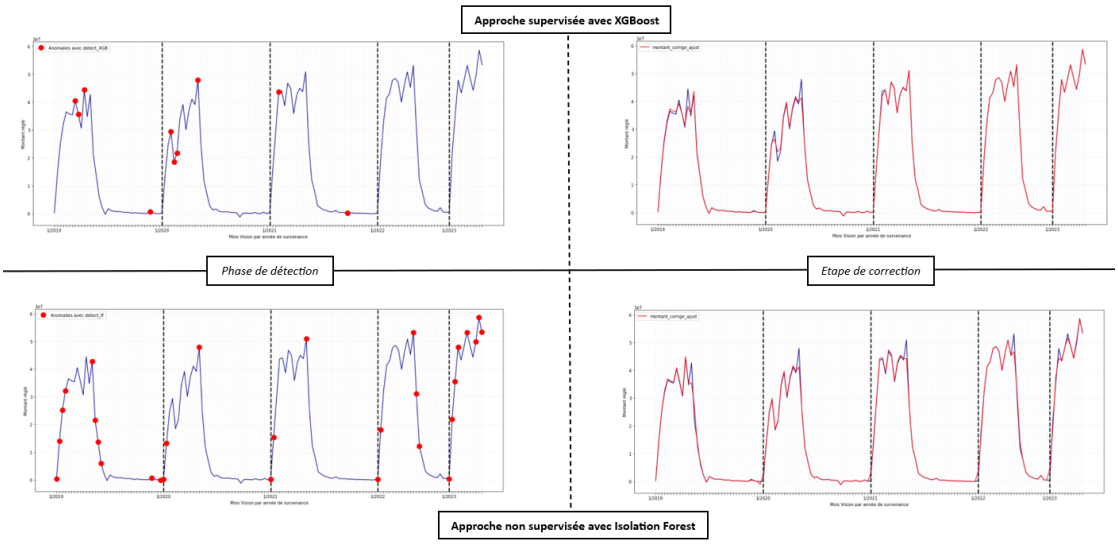


FIGURE 4.2 – Application du modèle de détection-correction sur le délégataire A

Le but est maintenant d'appliquer cette maquette aux deux autres sujets d'étude, d'en déduire le triangle de règlements cumulés, d'évaluer les provisions à l'aide de la méthode Chain Ladder et enfin de de comparer les résultats avec les données initiales.

4.3 Modèle de provisionnement

4.3.1 Modélisation théorique Chain Ladder

La méthode de provisionnement Chain Ladder est largement utilisée en actuariat pour estimer les réserves nécessaires en cas de sinistres reportés. Elle repose sur une technique récursive qui utilise les informations historiques sur les sinistres pour estimer les montants futurs. Dans le cas de l'assurance santé, le provisionnement mensuel est une pratique stratégique qui offre des avantages significatifs en termes de précision financière, gestion des coûts et flux de trésorerie car les bénéfices cumulés contribuent à la compétitivité des compagnies d'assurance dans un environnement en constante évolution.

Tout d'abord, les données historiques des sinistres sont organisées sous forme d'un triangle de développement, noté C_{ij}, où i représente l'année de survenance et j le mois de vision (correspondant à la période de développement). Chaque élément C_{ij} correspond au montant cumulé des sinistres survenus à l'année i et rapportés au mois de vision j. Ainsi, C_{ij} représente les sinistres survenus à l'année i et payés au mois j.

Pour chaque mois j, les facteurs de développement f_j sont estimés à partir des données historiques. Le calcul de f_j se présente comme suit :

$$f_j = \frac{C_{i,j+1}}{C_{i,j}} \tag{4.5}$$

L'estimation des sinistres totaux $\hat{C}_{i,n}$ pour l'année de survenance i à la fin de l'année n est obtenue en appliquant les facteurs de développement estimés :

$$\hat{C}_{i,n} = C_{i,n-1} \cdot f_{n-1} \cdot f_{n-2} \cdot \ldots \cdot f_i$$

où $C_{i,n-1}$ est le montant cumulé des sinistres observés à la fin de l'année $n-1$, et les f_j sont les facteurs de développement estimés.

Les provisions pour les sinistres non encore déclarés (IBNR) peuvent être calculées dans le modèle Chain Ladder comme la différence entre l'estimation totale des sinistres et les sinistres déjà observés :

$$IBNR_{i,n} = \hat{C}_{i,n} - C_{i,n-1} \tag{4.6}$$

Ce sont ces montants de provisions qui sont l'objet de l'étude faite ici. La confrontation entre les IBNR estimés avec les données initiales et les IBNR calculés à l'aide des valeurs corrigées par le modèle sera réalisée afin d'apporter les conclusions attendues par la problématique décrite en introduction.

Avant d'employer l'approche de provisionnement Chain Ladder, il est fondamental de vérifier les hypothèses justifiant son application.

4.3.2 Vérification des hypothèses du Chain Ladder

Pour appliquer correctement le modèle Chain Ladder en provisionnement actuariel, il est nécessaire de vérifier les hypothèses suivantes, qui peuvent être formulées mathématiquement :

1. Stabilité des schémas de développement :

   Pour $j = 0, 1, \ldots, n-1$, les facteurs de développement $f_j = \frac{C_{i,j+1}}{C_{i,j}}$ sont indépendants de l'année de survenance i. Ainsi, $\forall j \in [\![1, n-1]\!]$:

   $$\frac{C_{0,j+1}}{C_{0,j}} = \frac{C_{1,j+1}}{C_{1,j}} = \cdots = \frac{C_{i,j+1}}{C_{i,j}} = \frac{C_{n-j-1,j+1}}{C_{n-j-1,j}}. \tag{4.7}$$

   Cette hypothèse implique que les schémas de développement observés dans le passé se répéteront à l'avenir.

2. Homogénéité des sinistres :

   Les sinistres au sein de chaque groupe de survenance doivent être homogènes. Autrement dit, les sinistres doivent provenir d'une population avec des caractéristiques de risque similaires, ce qui permet de supposer que :

   $$E[C_{i,j+1} \mid C_{i,j}] = C_{i,j} \cdot f_j.$$

   L'hétérogénéité au sein d'un groupe pourrait biaiser les estimations de f_j.

3. Indépendance des sinistres d'une année à l'autre :

   Les sinistres observés pour une année de survenance i sont indépendants de ceux observés pour une autre année k ($k \neq i$). Mathématiquement, cela se traduit par l'hypothèse d'indépendance conditionnelle :

   $$\mathrm{Cov}(C_{i,j}, C_{k,l}) = 0 \quad \text{pour tout } i \neq k \text{ et pour tout } j, l.$$

   Cette indépendance assure que les données de différentes années de survenance ne sont pas corrélées, ce qui est nécessaire pour une estimation non biaisée des provisions.

Vérifions notamment les hypothèses fortes de la constance des cadences de règlements et de l'indépendance des années de survenance dans le cas du délégataire de référence A. Premièrement, il faut évaluer les montants réglés cumulés sous la forme d'un triangle de provisionnement. Le graphique ci-dessous présente ces règlements cumulés par mois de vision et pour chaque année de survenance :

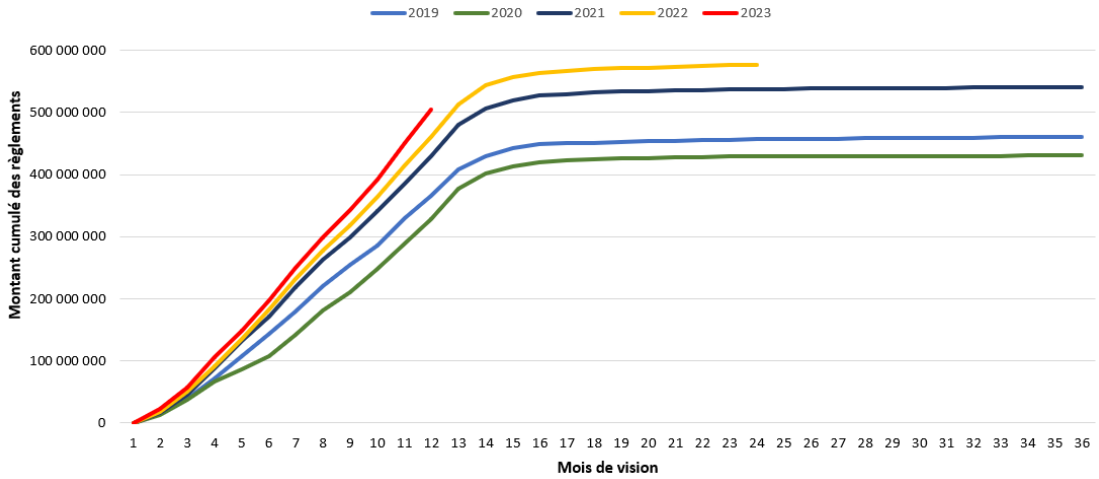


FIGURE 4.3 – Règlements cumulés du délégataire A par mois de vision, pour chaque année de survenance

Les survenances 2022 et 2023 seront tout le sujet de l'étude de provisionnement et de calcul des IBNR dans le contexte de l'apport de la fiabilité des données. A partir de ces données, il est envisageable de démontrer l'hypothèse de la stabilité des facteurs de développement.

Pour vérifier cette condition, il s'agit de montrer qu'en fixant j, les $(n-j)$ couples $(C_{i,j}, C_{i,j+1})$, pour tout $i \in \{0, \ldots, n-j-1\}$, sont "sensiblement" alignés suivant une droite passant par l'origine. Cela permettrait de justifier la relation linéaire entre les montants de deux mois consécutifs, confirmant ainsi une constance dans la cadence des paiements.

L'analyse étant mensuelle, cette étude est réalisée pour deux mois correspondant à la fin d'une année de vision, c'est-à-dire les mois 12 et 24. La relation entre $C_{i,j}$ et $C_{i,j+1}$ pour ces deux périodes est représentée graphiquement :

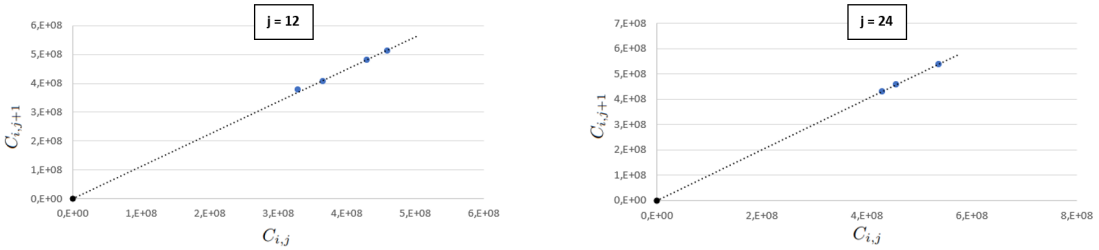


FIGURE 4.4 – Vérification de l'alignement de Ci,j et Ci,j+1 sur 2 périodes

L'alignement des points est clairement lisible sur les 2 graphiques car la droite en pointillés passe tout proche de chacune des valeurs, ce qui signifie que la première hypothèse est respectée. Justifions maintenant l'indépendance des années de survenance à l'aide de ces données.

Un moyen vérifiant ce postulat est le calcul du D-triangle. Ce triangle est consitué des facteurs individuels définis de la manière suivante :

$$\forall\ (i,j) \quad \text{tel que} \quad i+j \leq n-1, \quad f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}} \tag{4.8}$$

A partir de ce tableau de facteurs, il est possible d'évaluer plusieurs statistiques permettant de valider l'indépendance des survenances. Le tableau ci-dessous en présente certaines pour les douze premiers mois de vision :

	1	2	3	4	5	6	7	8	9	10	11	12
Moyenne	55,4458	2,7599	1,8402	1,4311	1,3101	1,2746	1,2207	1,1512	1,1456	1,1483	1,1161	1,1241
Ecart-Type	9,2364	0,1046	0,0748	0,0958	0,0389	0,0269	0,0326	0,0125	0,0197	0,0132	0,0116	0,0145
Ecart-Type / Moyenne	***16,66%***	***3,79%***	***4,06%***	***6,70%***	***2,97%***	***2,11%***	***2,67%***	***1,09%***	***1,72%***	***1,15%***	***1,04%***	***1,29%***

FIGURE 4.5 – Statistiques sur les cadences de règlement pour le délégataire A

Hormis le 1er mois de vision pour lequel le facteur de développement est très élevé, la variation maximale par rapport à la valeur moyenne ne dépasse pas 7% sur cette période. De plus, l'écart-type tend à diminuer au fur et à mesure tandis que la moyenne semble converger vers 1. Cette variation restera alors assez faible ce qui permettra de supposer l'hypothèse de l'indépendance des années de survenance vérifiée.

Ainsi, les hypothèses du modèle de Chain Ladder sont valides dans le cas du délégataire A. Par similarité de structure des données entre les segments d'étude, ces conjectures seront considérées valides pour le délégataire B ainsi que pour la gestion directe et il est désormais possible d'appliquer la méthode de provisionnement Chain Ladder.

4.3.3 Estimation des provisions

Afin de calculer les provisions avec Chain Ladder, représentons schématiquement le triangle de règlements cumulés spécifique à l'étude réalisée. Les notations suivantes sont rappelées : soit i l'année de survenance (l'intervalle de temps comprend les années 2019 à 2023) et j le mois de vision (allant de 1 à 36). $C_{i,j}$ est le montant cumulé de règlements. La matrice suivante apparaît :

$i \setminus j$	1	2	$\cdots$	12	13	$\cdots$	24	25	$\cdots$	36
2019	$C_{2019,1}$	$C_{2019,2}$	$\cdots$	$C_{2019,12}$	$C_{2019,13}$	$\cdots$	$C_{2019,24}$	$C_{2019,25}$	$\cdots$	$C_{2019,36}$
2020	$C_{2020,1}$	$C_{2020,2}$	$\cdots$	$C_{2020,12}$	$C_{2020,13}$	$\cdots$	$C_{2020,24}$	$C_{2020,25}$	$\cdots$	$C_{2020,36}$
2021	$C_{2021,1}$	$C_{2021,2}$	$\cdots$	$C_{2021,12}$	$C_{2021,13}$	$\cdots$	$C_{2021,24}$	$C_{2021,25}$	$\cdots$	$C_{2021,36}$
2022	$C_{2022,1}$	$C_{2022,2}$	$\cdots$	$C_{2022,12}$	$C_{2022,13}$	$\cdots$	$C_{2022,24}$			
2023	$C_{2023,1}$	$C_{2023,2}$	$\cdots$	$C_{2023,12}$						

Ce triangle met en évidence l'objectif du modèle de provisionnement : l'estimation des montants des mois de vision 25 à 36 pour la survenance 2022 et ceux des mois 13 à 36 pour l'année 2023. Ce calcul fera intervenir les règlements des 5 dernières années de

survenance bien que le triangle mensuel de provisionnement s'écrirait sur les 3 dernières survenances (étant donné que 36 mois de vision, soit 3 années de développement, sont considérés).

Avec la méthode Chain Ladder, il est alors possible d'évaluer les facteurs de développement et de compléter le triangle cumulé précédent, afin d'obtenir notamment le montant réglé projeté à l'ultime. En reprenant les données du délégataire A, il est possible de synthétiser les résultats dans le tableau suivant :

	Année	Dernier Règlement cumulé connu	Projection des règlements à l'ultime	Pourcentage réglé	Montant provision restant (IBNR)
1	2021	540,229,371.09	540,229,371.09	100.00%	nan
2	2022	576,374,868.41	579,377,868.95	99.48%	3,003,000.54
3	2023	504,444,418.91	640,353,999.95	78.78%	135,909,581.05

FIGURE 4.6 – Tableau de provisionnement sur les données initiales du délégataire A

Ce tableau présente pour les 3 dernières années de survenance les derniers montants connus , c'est à dire le règlement des 36ème, 24ème et 12ème mois pour les survenances 2021, 2022 et 2023 respectivement. Puis, est affiché la valeur de règlement du 36ème mois de vision de chaque année, projeté par Chain Ladder pour 2022 et 2023. Le pourcentage de règlement effectué correspond au ratio par survenance :

$$\text{Pourcentage réglé} = \frac{\text{Dernier règlement connu}}{\text{Règlement à l'ultime}}$$

Cela permet d'identifier la part de règlement déjà effectué avant la projection par Chain Ladder. Enfin, l'estimation des IBNR équivaut à la différence entre la valeur du montant à l'ultime et le dernier règlement connu.

Avec les données initales du délégataire A, il convient de relever que 3M€ reste à provisionner pour indemniser les sinistres en santé survenus lors de l'année 2022. Moins de 80% des règlements survenus en 2023 ont été payés, il resterait alors plus de 135M€ à régler, répartis sur les 24 mois suivants.

Cependant, le provisionnement Chain Ladder tient compte de la survenance 2020 dans ses projections. Or, comme analysé dans les parties précédentes, le contexte pandémique à cette période a entraîné des bouleversements dans les règlements en assurance santé si bien que sa prise en compte dans le modèle est remise en question.

4.3.4 Analyse sans 2020

Exclure l'année de survenance 2020 des calculs de projection pour le provisionnement en santé peut se révéler être une stratégie avantageuse dans plusieurs contextes. En effet, cette année a été marquée par la pandémie du COVID-19 qui a profondément perturbé le secteur de la santé et a généré des données atypiques à certains moments. Voici les raisons pour lesquelles ne pas inclure l'année 2020 dans les projections pourrait être pertinent :

- Représentativité des données : l'année 2020 a vu une réduction drastique de l'utilisation des services de santé non urgents, en raison des confinements et des restrictions de déplacement. Beaucoup d'interventions médicales ont été reportées, tandis que la demande pour des soins urgents liés au COVID-19 a explosé. Ces perturbations créent des données anormales qui ne reflètent pas les besoins habituels en santé.
- Risque d'une évaluation biaisée des provisions : cette année est perçue comme représentative d'un avenir où des crises sanitaires similaires se répètent. Or, les pics de dépenses liés à la COVID-19 sont probablement des événements ponctuels, qui ont un impact fort sur une étude faite sur 5 années. Exclure 2020 permet de créer un modèle de provisionnement plus réaliste, en évitant d'intégrer des niveaux de dépenses qui pourraient ne pas se reproduire.
- Moindre volatilité des projections : l'intégration de 2020 pourrait introduire une forte volatilité dans les prévisions, rendant les résultats plus sensibles à des événements imprévisibles et extrêmes. Son exclusion garantirait une stabilité dans les prévisions qui seraient plus susceptibles de capturer les comportements futurs, notamment en termes de consommation de soins réguliers.

Pour se rendre compte de l'influence du caractère atypique de la survenance 2020, il est possible de représenter la dispersion des résidus par année de survenance, les résidus correspondant à la différence entre les facteurs indiviuels et les facteurs de développement :

$$\text{Résidu}_{i,j} = \frac{C_{i,j+1}}{C_{i,j}} - f_j \tag{4.9}$$

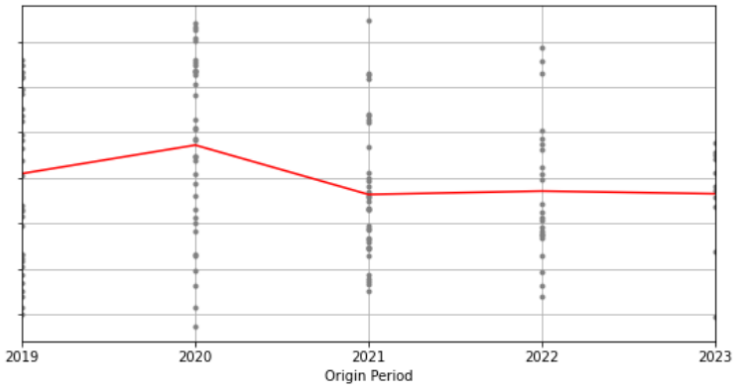


FIGURE 4.7 – Dispersion des résidus par survenance sur le délégataire A

La courbe rouge indique la dispersion moyenne des résidus. Ainsi la survenance 2020 présente un pic justifiant des variations plus importantes que les autres années. Cela confirme l'effet préjudiciable de l'année 2020 sur le provisionnement. Cette vision du

provisionnement sans 2020 étant pertinente à intégrer pour notre étude, il faut alors calculer les provisions Chain Ladder à partir du triangle suivant :

$i \setminus j$	1	2	$\cdots$	12	13	$\cdots$	24	25	$\cdots$	36
2019	$C_{2019,1}$	$C_{2019,2}$	$\cdots$	$C_{2019,12}$	$C_{2019,13}$	$\cdots$	$C_{2019,24}$	$C_{2019,25}$	$\cdots$	$C_{2019,36}$
2021	$C_{2021,1}$	$C_{2021,2}$	$\cdots$	$C_{2021,12}$	$C_{2021,13}$	$\cdots$	$C_{2021,24}$	$C_{2021,25}$	$\cdots$	$C_{2021,36}$
2022	$C_{2022,1}$	$C_{2022,2}$	$\cdots$	$C_{2022,12}$	$C_{2022,13}$	$\cdots$	$C_{2022,24}$			
2023	$C_{2023,1}$	$C_{2023,2}$	$\cdots$	$C_{2023,12}$						

Cette projection sans la survenance 2020 est à mettre en parallèle avec les estimations réalisées sur les données corrigées par le modèle de détection-correction. En effet, le XGBoost et l'Isolation Forest ont perçu un plus grand nombre d'anomalies durant l'année 2020 sur les différents segments d'étude, il sera donc judicieux d'interpréter l'effet du lissage des données anormales sur le calcul des IBNR.

4.4 Comparaison des estimations du provisionnement

Quatre applications du Chain Ladder ont donc été réalisées pour estimer les provisions IBNR selon différents contextes : les données initiales, les données corrigées suite à la détection avec une approche non-supervisée avec Isolation Forest, celles modifiées avec l'approche supervisée à l'aide de XGBoost et enfin la projection des règlements sans tenir compte de la survenance 2020. Analysons en détails chacun des trois segments d'étude.

4.4.1 Délégataire A

Le délégataire A est l'exemple de référence de l'étude, possédant pour rappel 15 anomalies sur 144 données selon le modèle statistique réalisé. Le meilleur moyen de comparer chacune des applications réalisées est de confronter le résultat de projection des IBNR pour les survenances 2022 et 2023. Le tableau ci-dessous présente les résultats condensés obtenus suivant les quatre visions :

	Données initiales	Données corrigées avec XGBoost	Données corrigées avec Isolation Forest	Données initiales sans 2020
2022	3 003 001	2 945 301	2 481 897	3 409 597
delta		*-1,92%*	*-17,35%*	*13,54%*
2023	135 909 581	134 506 715	128 353 867	130 495 855
delta		*-1,03%*	*-5,56%*	*-3,98%*
TOTAL	**138 912 582**	**137 452 016**	**130 835 764**	**133 905 452**
delta		*-1,05%*	*-5,81%*	*-3,60%*

FIGURE 4.8 – Estimation des IBNR du délégataire A suivant les quatre approches

Le tableau présente les valeurs des IBNR ainsi que le delta noté Δ obtenu avec le calcul suivant :

$$\Delta = \frac{\text{IBNR}_{\text{modèle}} - \text{IBNR}_{\text{initial}}}{\text{IBNR}_{\text{initial}}} \tag{4.10}$$

Ce delta correspond à un ratio représentatif de l'augmentation ou de la diminution des provisions évaluées par le modèle par rapport aux données initiales. Il faut cependant différencier sa valeur lorsqu'il est estimé pour l'année 2022 ou pour 2023. Le fait de projeter 12 ou 24 mois de visions respectivement crée un biais à ne pas considérer dans l'analyse finale.

Ainsi, une revue du provisionnement à la baisse aurait été réalisée si la correction des anomalies avait été effectuée à partir d'un des modèles de *machine learning*. Dans le cas de l'approche non-supervisée, l'Isolation Forest aurait sous-estimé de 8M€ les montants à provisionner sur les survenances 2022 et 2023. En revanche, La détection d'anomalies par XGBoost aurait entraîné quant à lui une sous-estimation de 1,5M€. L'hypothèse forte de ne pas prendre en compte l'année 2020 entraîne une prévision montrant un écart de 5M€ entre les deux types de correction, indiquant un montant de provisions moins élevé qu'avec la projection avec les données initiales.

Analysons plus en détaillant les estimations réalisées en représentant graphiquement les provisions pour chacun des modèles :

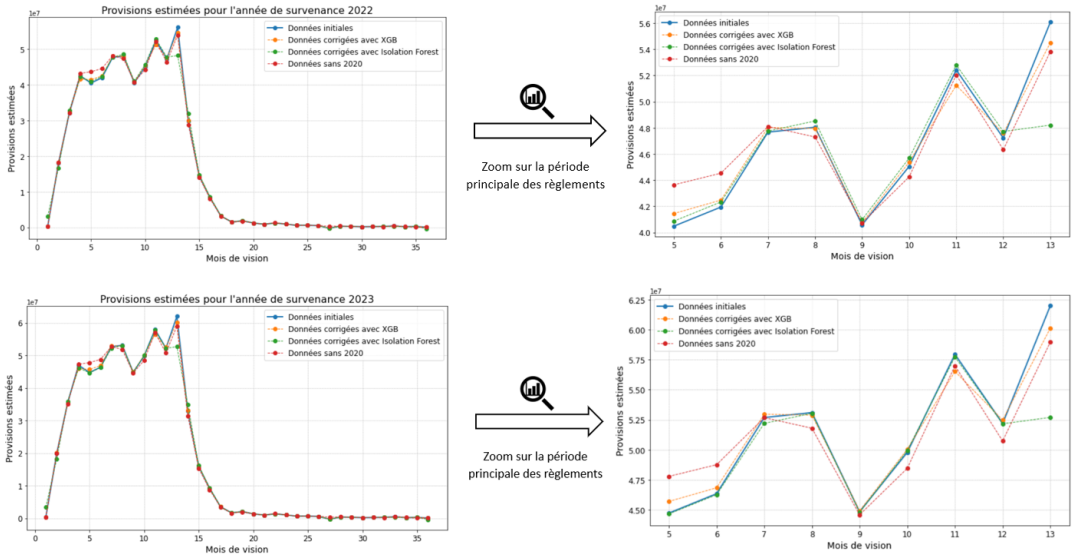


FIGURE 4.9 – Projection des règlements du délégataire A selon les quatre approches

Pour mieux comprendre les disparités entre les prédictions, un zoom a été réalisé sur les mois de vision 5 à 13, correspondant à la période de règlements la plus importante. L'écart le plus significatif se manifeste aux 5ème et 6ème mois de vision. Ces périodes, étant très affectées par la pandémie en 2020, créent des divergences de montant entre les données initiales et les données sans 2020. Les deux approches de détection-correction atténuent cet effet. Plus globalement, le Chain Ladder sans 2020 estime des provisions plus faibles en dehors des mois de vision 5 et 6 ce qui explique une revue globale à la baisse du provisionnement. Néanmoins, le lissage après détection avec Isolation Forest sous

estime fortement le règlement au mois de vision 13, car le modèle a repéré une anomalie à ce niveau à plusieurs reprises contrairement au XGBoost. Cela explique majoritairement l'écart de provisionnement entre ces deux méthodes par rapport aux données initiales.

Vérifions maintenant si ces conjectures s'appliquent dans le cas d'un autre délégataire.

4.4.2 Délégataire B

Le délégataire B présente plus d'anomalies que le délégataire A, plus exactement 15% d'anomalies dans ses données. Tout comme le premier exemple, évaluons les IBNR avec chacune des méthodes :

	Données initiales	Données corrigées avec XGBoost	Données corrigées avec Isolation Forest	Données initiales sans 2020
2022	751 313	750 741	834 748	780 390
delta		*-0,08%*	*11,11%*	*3,87%*
2023	53 930 469	51 361 329	52 778 578	51 891 443
delta		*-4,76%*	*-2,14%*	*-3,78%*
TOTAL	**54 681 782**	**52 112 070**	**53 613 326**	**52 671 833**
delta		*-4,70%*	*-1,95%*	*-3,68%*

FIGURE 4.10 – Estimation des IBNR du délégataire B suivant les quatre approches

Une revue à la baisse des provisions est aussi constatée quelque soit la vision appliquée. Toutefois, l'approche XGBoost sous-estime de 2.5M€ alors que l'Isolation Forest uniquement de 1M€. Cela peut s'expliquer par le montant plus élevé de provisions évalué par le modèle non-supervisé pour la survenance 2022. Quant aux calculs réalisés sans la survenance 2020, les résultats se trouvent une nouvelle fois entre les deux estimations avec correction.

Zoomons sur la période forte de règlements pour les deux survenances afin d'expliciter ces analyses :

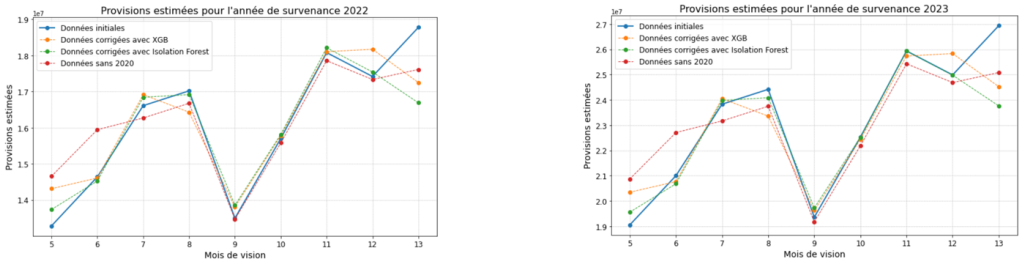


FIGURE 4.11 – Projection des règlements du délégataire B selon les quatre approches

Dans l'ensemble, il est possible de retrouver les mêmes justifications que pour le délégataire A. Les mois de vision 5 et 6 sont lissés par le modèle de détection-correction et présente des valeurs comprises entre l'estimation avec données initiales et sans l'année 2020. De même, le mois de vision 13 démontre une sous-évaluation des données de provisionnement, plus forte pour l'approche non-supervisée. Mais cela ne justifie pas la

sous-valorisation des provisions par XGBoost en comparaison avec les données initiales. Il faut en revenir à l'étude de détection pour identifier la source de cette différence. Ici, le XGBoost avait détecté plus d'anomalies sur des pics élevés que le modèle non supervisé sur les données du délégataire B, comme le démontre les graphiques de l'annexe A.6. Le modèle supervisé a donc provoqué un lissage globalement plus important sur ces pics ce qui a généré une sous-estimation des provisions.

Malgré tout, les hypothèses repérées lors de l'étude du délégataire A se retrouvent pour le délégataire B. Il est alors légitime de se demander si la gestion directe se comporte de la même manière que la gestion déléguée.

4.4.3 Gestion directe

La gestion directe a présenté une quantité plus faible de données anormales, s'expliquant notamment par une gestion simplifiée des contrats étant donné l'absence d'intermédiaires dans ce cadre. Observons de plus près les provisions estimées par les quatre visions réalisées dans ce contexte :

	Données initiales	Données corrigées avec XGBoost	Données corrigées avec Isolation Forest	Données initiales sans 2020
2022	2 681 972	2 714 637	2 965 898	2 896 403
delta		*1,22%*	*10,59%*	*8,00%*
2023	120 550 025	121 805 753	133 950 425	120 146 591
delta		*1,04%*	*11,12%*	*-0,33%*
TOTAL	**123 231 997**	**124 520 389**	**136 916 323**	**123 042 994**
delta		*1,05%*	*11,10%*	*-0,15%*

FIGURE 4.12 – Estimation des IBNR de la gestion directe suivant les quatre approches

Un résultat global dénote des analyses précédentes : la détection-correction des anomalies a généré un montant de provisions plus élevé qu'avec les données initiales, que l'approche soit supervisée ou non-supervisée. Bien qu'il soit faible (de l'ordre de 1M€) pour le modèle XGBoost, la détection avec Isolation Forest a relevé une quantité d'anomalies plus importante, et à des périodes moins fortes de règlement. Le lissage par filtre a provoqué pour ces anomalies une correction à la hausse de ces montants, si bien que l'approche non-supervisée a sur-évalué les provisions de plus de 13M€.

En remettant ces données dans leur contexte, la gestion directe est associée à des règlements de l'ordre du milliard d'euros (données modifiées dans le contexte de l'étude). Donc, un revue du provisionnement à la hausse de 1M€ est plutôt négligeable au regard d'un montant supérieur à 10M€. Représentons graphiquement les trajectoires des règlements spécifiques à la gestion directe et zoomons sur la période forte des règlements :

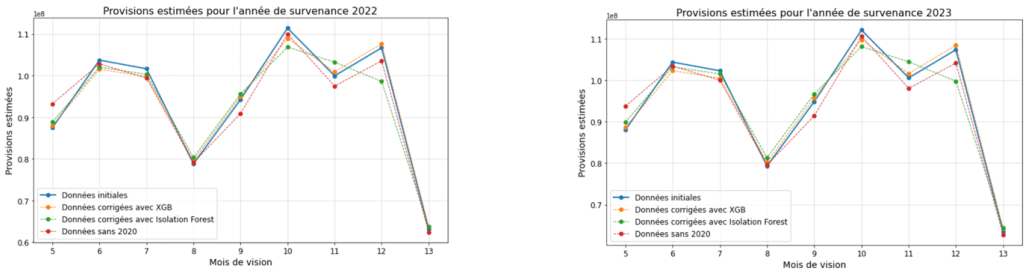


FIGURE 4.13 – Projection des règlements de la gestion directe selon les quatre approches

Les mois de vision 8, 9 et 11 justifient notamment la sur-valorisation des provisions par les deux modèles de détection-correction. Ces mois n'étant pas vraiment concernés par l'effet pandémique du COVID-19, les calculs de provisionnement sans prise en compte de la survenance 2020 ne sont pas affectés et engendrent même une revue à la baisse du provisionnement sur les derniers mois de vision du zoom effectué.

En corrigeant quelques anomalies, les modèles ont donc sur-évalué les provisions associées à la gestion directe et l'approche non-supervisée n'a fait qu'intensifier cette réaction.

4.4.4 Bilan sur l'apport de la fiabilité des données

Essayons de résumer l'application finale du provisionnement et l'intérêt du retraitement des données dans ce contexte, à partir des trois exemples analysés :

- Correction globale des anomalies : les données anormales détectées par les modèles de *machine learning* ont été lissées, les pics positifs exceptionnels ont donc été atténués et certains retards ou annulations de règlements de contrat d'assurance santé ont été rectifiés. Cela impacte alors les provisions estimées par Chain Ladder, mais il faut contextualiser les résultats pour identifier l'influence concrète de la qualité des données.

- Cas particulier de l'année 2020 : le contexte pandémique de cette année a produit de nombreuses valeurs étonnantes, déclarées majoritairement comme anomalies par les modèles de *machine learning*. Les provisions estimées après lissage permettent de tenir compte de la survenance 2020 tout en diminuant l'effet perturbateur de cette année dans le calcul.

- Confrontation entre la gestion déléguée et directe : la correction des données anormales ont provoqué une revue à la baisse du provisionnement global sur les 2 délégataires étudiés, alors qu'elle a sur-évalué les provisions pour la gestion directe. La rectification d'une plus grande quantité d'anomalies semble tendre vers une sous-valorisation des provisions. Lorsque les modèles ont tendance à détecter une

plus grande proportion de données anormales sur des pics descendants, le lissage vient réhausser les montants provisionnés. Ce phénomène se produit lorsque peu d'anomalies sont repérées.

- Disparités entre l'approche supervisée et non-supervisée : l'Isolation Forest a généré dans l'ensemble plus d'écarts de provisions que l'approche supervisée avec XGBoost. Le fait que le modèle travaille sur des données sans étiquettes prédéfinies génère plus d'erreurs dans la détection d'anomalies. Le lissage vient alors modifier ces "non-anomalies" et est alors à l'origine de la déviation des provisions estimées.

La fiabilité des données reste primordiale avant d'appliquer des modèles de projection, elle doit être assurée avant de calculer les provisions pour les contrats d'assurance santé collectifs. L'automatisation par *machine learning* pour corriger les anomalies dans les données est un bon moyen d'atténuer les potentielles erreurs de gestion. Seulement, si ces modèles viennent à corriger des données qui ne devraient pas être jugées comme anormales, alors l'estimation du provisionnement pourrait en être impactée.

L'application du Chain Ladder en gestion déléguée est légèrement plus prudente avec les données initiales même si elle ne reflète pas fidèlement le risque car la survenance 2020 est utilisée dans la prévision. La détection-correction des données anormales paraît plus risquée dans ce cadre mais semble plus juste et représentative du risque que le provisionnement avec ou sans 2020. Cette vision permet alors de prendre en compte un intervalle de temps plus important pour la projection, en atténuant l'impact de certaines anomalies probables qui viendraient modifier les calculs de provision.

Ainsi, des données fiables assurent l'intégrité et la solidité des modèles de provisionnement, en garantissant l'exactitude des prévisions de provisions, à condition que les modèles de détection soient justes dans l'identification des anomalies.

Conclusion

L'étude menée souligne la nécessité de l'exactitude des données pour garantir des estimations précises et fiables des provisions. L'utilisation de techniques de détection des anomalies par *machine learning*, tant supervisées que non supervisées, a permis d'identifier et de traiter les données incohérentes et les erreurs potentielles qui à défaut pourraient compromettre l'intégrité des modèles de provisionnement.

D'une part, l'approche supervisée utilisant le modèle XGBoost a démontré une grande efficacité dans la détection des anomalies, en s'appuyant sur un modèle statistique préalablement élaboré. Ce modèle combine des techniques d'apprentissage supervisé et des méthodes de pondération optimisées, permettant ainsi de détecter avec précision les anomalies tout en minimisant les erreurs. D'autre part, les méthodes non supervisées, comme les algorithmes de *clustering*, les techniques de réduction de dimensionnalité ou les forêts aléatoires avec Isolation Forest, ont offert une perspective complémentaire en identifiant des anomalies non étiquetées et en révélant des schémas sous-jacents dans les données.

L'intégration d'une correction par lissage mathématique après la détection des anomalies s'est révélée être un atout majeur. L'application d'un filtre avec Savitzky-Golay a contribué à minimiser les impacts des anomalies sur les prévisions en atténuant les fortes variations. Cette étape est essentielle pour obtenir des estimations de provisionnement plus cohérentes et fiables.

L'application actuarielle du provisionnement a confirmé plusieurs des hypothèses évoquées tout au long de l'analyse. Tout d'abord, le modèle de détection-correction atténue les effets pandémiques de 2020 tout en prenant en compte l'année dans la projection. Ensuite, l'approche non-supervisée crée plus de déviation sur l'estimation des provisions car elle repère plus de "non-anomalies". Dans le cas de la gestion déléguée, la fiabilité des données a diminué le montant de provisions. A l'inverse, une sur-valorisation des provisions est estimée pour la gestion directe, dûe au lissage qui réhausse le peu de montants anormaux dans ce contexte.

Dans l'ensemble, la prévision des provisions, plus fidèle au passé et donc aux risques actuels associés aux contrats d'assurance santé collectifs, est plus fiable tant que les anomalies détectées par le modèle sont vérifiées.

Toutefois, plusieurs axes d'amélioration peuvent être envisagés pour optimiser davantage le processus et pour approfondir l'analyse :

— Apport du *deep learning* : les modèles de *deep learning*, tels que les réseaux neuronaux convolutifs (CNN) et les réseaux neuronaux récurrents (RNN), pourraient offrir une meilleure performance dans la détection des anomalies en capturant des relations complexes dans les données. Leur capacité à apprendre des représentations plus riches pourrait améliorer la précision des prévisions de provisionnement.

— Correction des erreurs par maille fine de contrat : une approche plus granulaire, qui consiste à corriger les erreurs au niveau de chaque contrat individuel plutôt qu'à un niveau agrégé, permettrait une meilleure personnalisation des corrections et une précision accrue dans l'évaluation des provisions. Cela demanderait un temps de calcul considérable pour parcourir chaque ligne des dossiers et pour détecter les anomalies associées, mais contribuerait à un ajustement plus exact.

— Étude manuelle en absence d'automatisation : pour les cas où la gestion automatisée des données n'est pas encore implémentée, comme l'exemple des autres délégataires non EDI, une étude manuelle permettrait de mieux comprendre les types d'erreurs courantes et d'affiner les algorithmes de détection des anomalies. La vision automatisée et la régularisation des données ne semblent pas adaptées à des données très variables.

— Test d'autres modèles de provisionnement : l'exploration de nouveaux modèles de provisionnement, que ce soit Bornhuetter-Ferguson ou ceux basés sur des techniques avancées de séries temporelles, pourrait offrir des perspectives nouvelles sur la gestion des provisions et permettre une meilleure adaptation aux fluctuations des données. La fiabilisation des données dans ce cadre pourrait donner de nouvelles conclusions à l'étude réalisée.

— Application à diverses tâches actuarielles au-delà du provisionnement : par exemple, dans la tarification, des données fiables permettent de mieux segmenter les assurés et d'affiner les primes, ou dans la gestion du capital et l'évaluation des solvabilités pour anticiper les besoins en fonds propres et optimiser l'allocation des ressources.

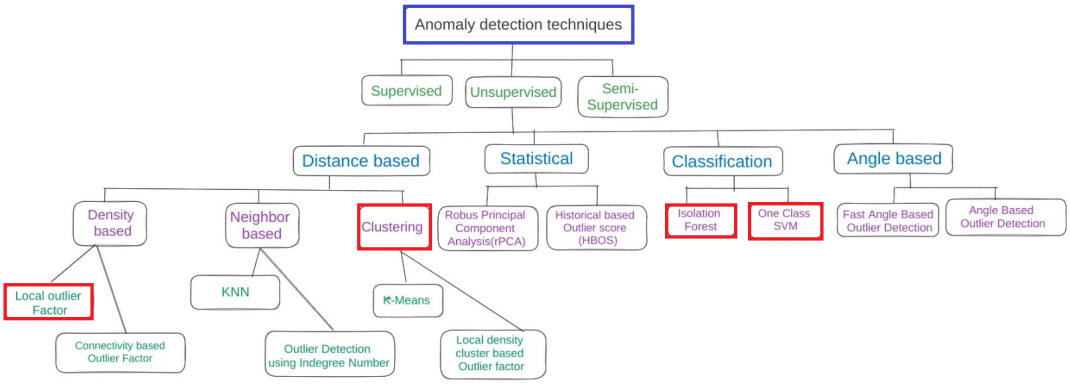
Finalement, le modèle de détection-correction mis en place a permis d'assurer une représentativité des données aux risques passées, entraînant une prévision plus robuste des provisions d'assurance santé collective. Bien que l'application des techniques de *machine learning* et de lissage mathématique ait considérablement amélioré la gestion des provisions, il est essentiel de continuer à explorer et à intégrer des méthodes avancées pour faire face à la complexité croissante des données en assurance santé collective. Une approche innovante sera déterminante pour maintenir la précision et la fiabilité des modèles de provisionnement face aux évolutions futures du secteur.

Annexes

Annexe A

Compléments sur les modèles de détection

A.1 Arbre des méthodes *machine learning* de détection



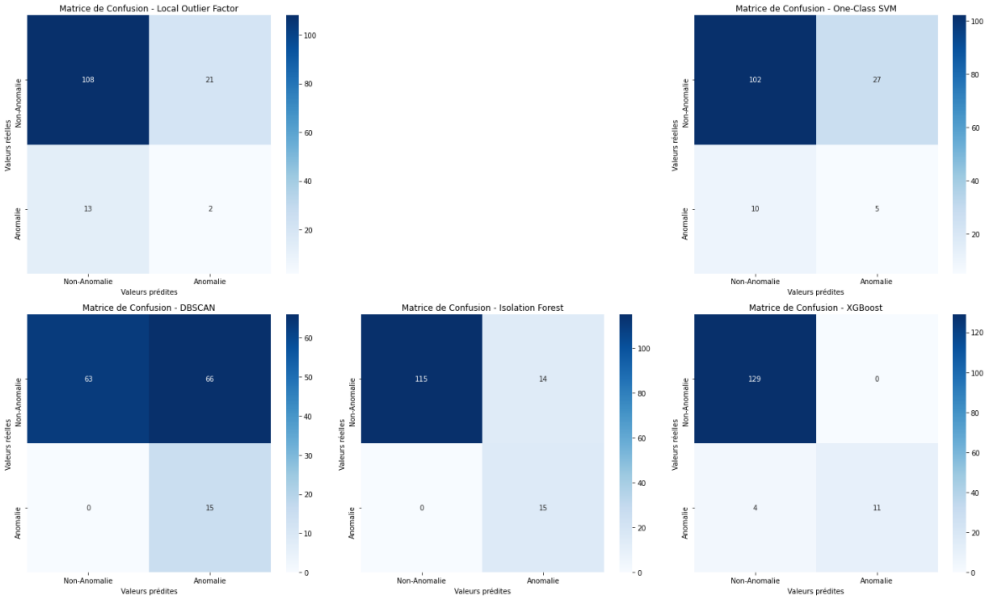
A.2 Tableau des méthodes *machine learning* non supervisées de détection

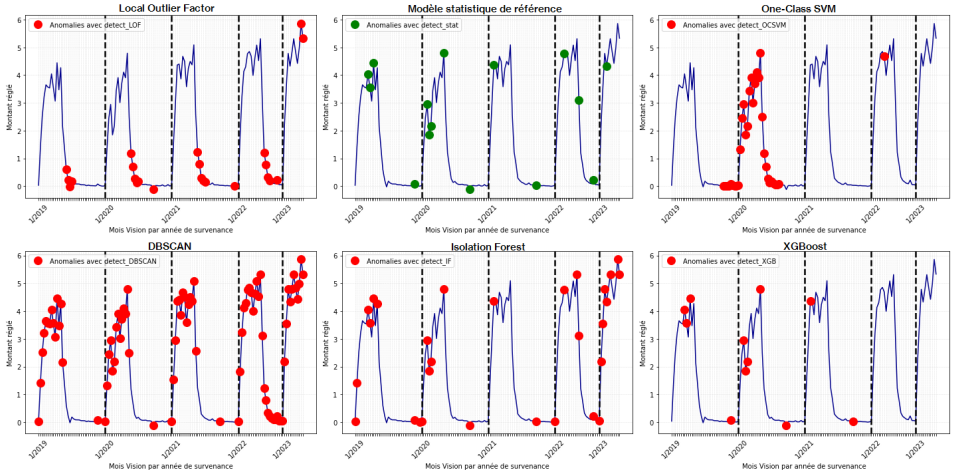
METHODE DE DETECTION D'ANOMALIES	UTILITE	FONCTIONNEMENT	PARAMETRES	AVANTAGES	INCONVENIENTS
LOCAL OUTLIER FACTOR (LOF)	Détection des anomalies basée sur la densité locale	Calcul de la densité locale de chaque échantillon par rapport à ses voisins, puis comparaison avec la densité de ses voisins	Nombre de voisins	- Adapté aux données de densité variable et pour les grandes dimensions	- Sensible au choix du nombre de voisins - Sensible aux dimensions des données
ONE-CLASS SVM	Détection des anomalies en apprenant à partir de données normales	Apprentissage d'un modèle basé sur les données normales, puis détection des échantillons rares	Largeur de la marge	- Lorsque les données normales sont bien représentées	- Sensible au choix de la largeur de la marge - Inefficace si les données normales sont mal représentées
DBSCAN	Détection des anomalies basée sur la densité	Partitionnement des données en sous-groupes de densité similaire, puis identification des échantillons isolés	Nombre minimum d'échantillons, distance maximale entre les échantillons considérés dans le même cluster	- Adapté aux structures de données de densité variable	- Nécessite la spécification de plusieurs paramètres - Sensible à la dimensionnalité élevée des données
ISOLATION FOREST	Détection des anomalies basée sur les arbres	Partitionnement récursif des données en utilisant des arbres de décision pour isoler les anomalies	Profondeur maximale de l'arbre, nombre d'échantillons pour l'échantillonnage	- Efficace pour détecter les anomalies sur des données de grande dimension	- Exige un ajustement minutieux des paramètres pour obtenir des résultats optimaux

A.3 Métriques d'évaluation pour la classification

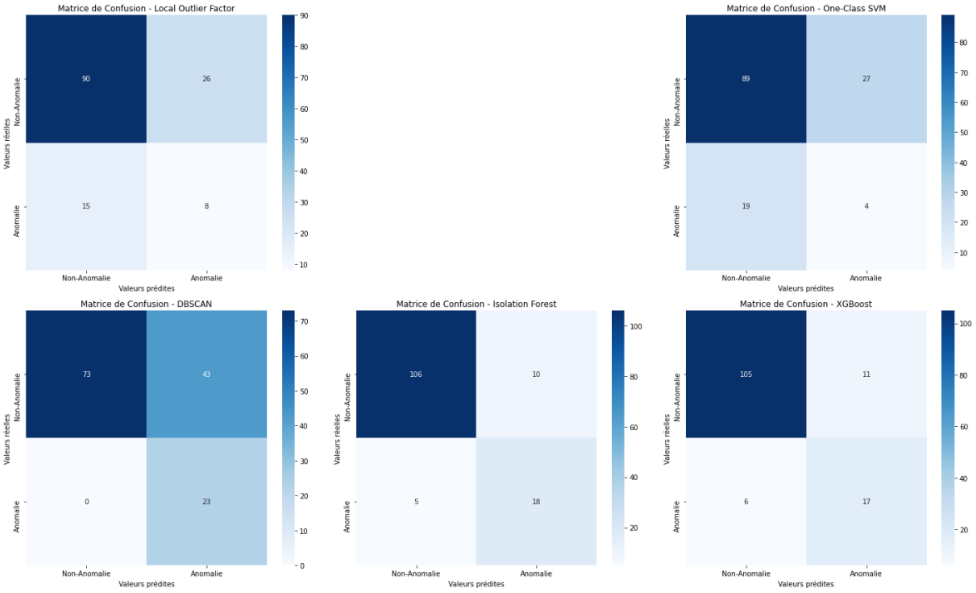
Métrique	Définition	Formule de calcul
Matrice de confusion	Tableau qui résume les performances d'un modèle en présentant les vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN)	Une matrice montrant VP, FP, FN, VN
Accuracy (Précision)	Proportion des observations prédites comme positives qui sont réellement positives	$\text{Accuracy} = \frac{VP}{VP + FP}$
Recall (Rappel)	Proportion des observations positives correctes parmi toutes les observations réellement positives	$\text{Recall} = \frac{VP}{VP + FN}$
F1-Score	Moyenne harmonique entre l'accuracy et le recall, fournissant un équilibre entre les deux	$F_1 = 2 \times \frac{\text{Accuracy} \times \text{Recall}}{\text{Accuracy} + \text{Recall}}$
Courbe ROC	Graphique qui montre le compromis entre le recall et le taux de faux positifs (FPR) pour différents seuils de décision.	$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{VN}}$

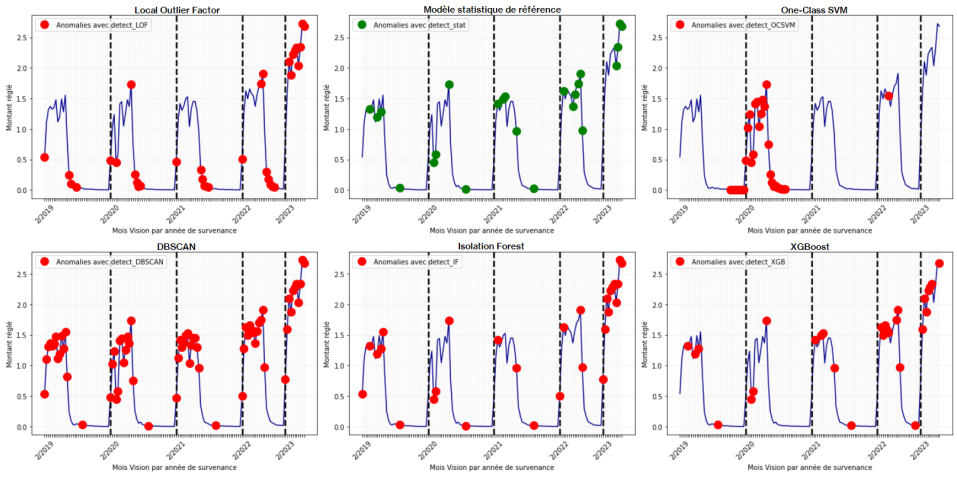
A.4 Détection des anomalies pour le délégataire A



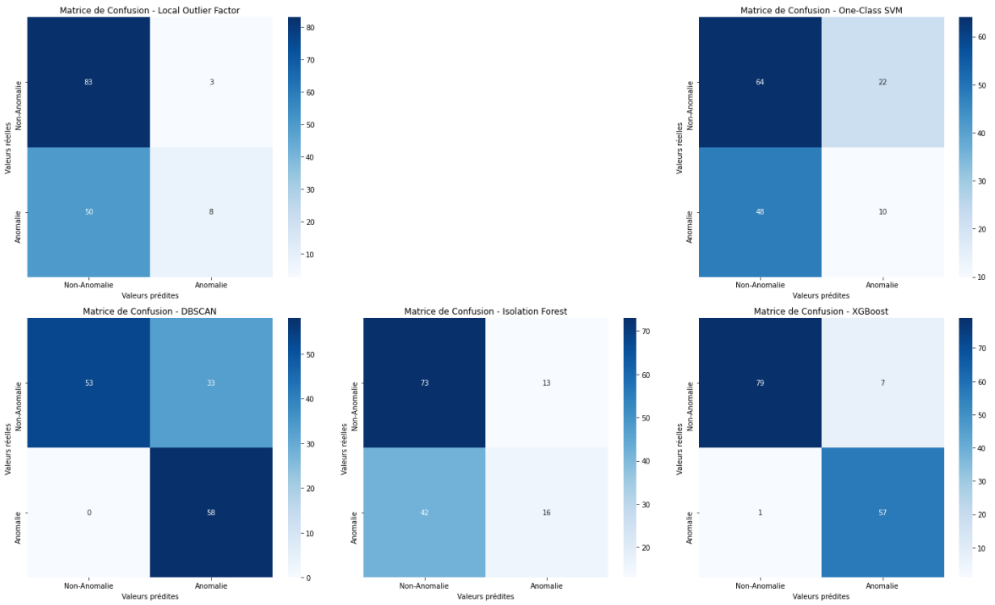


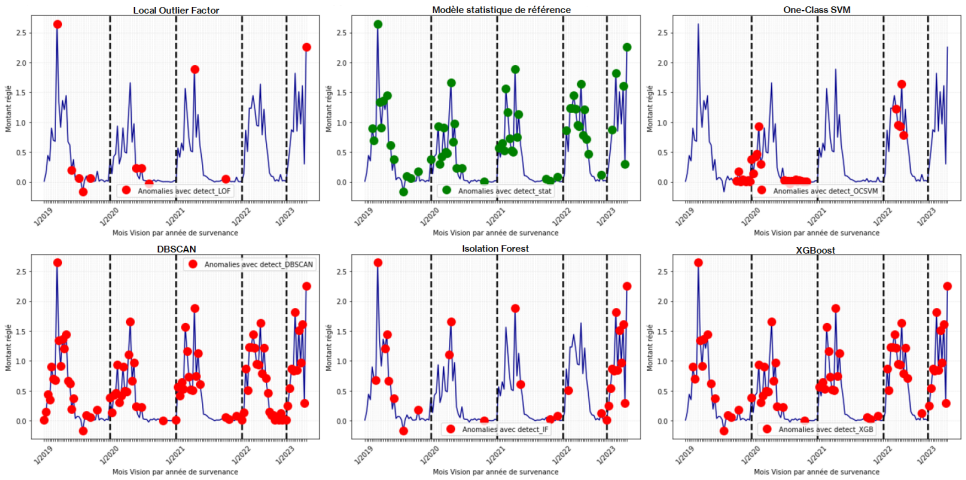
A.5 Détection des anomalies pour le délégataire B



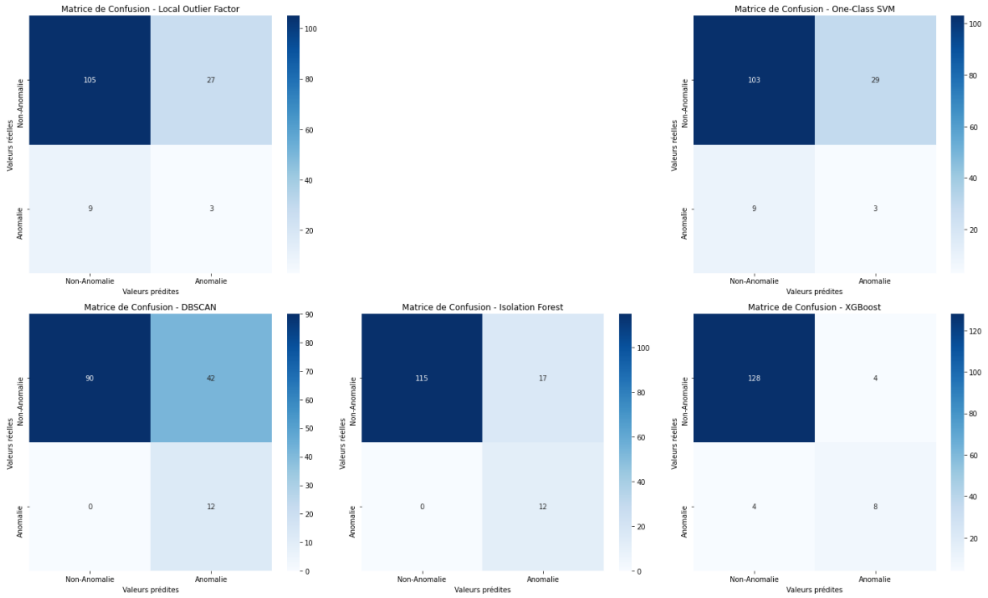


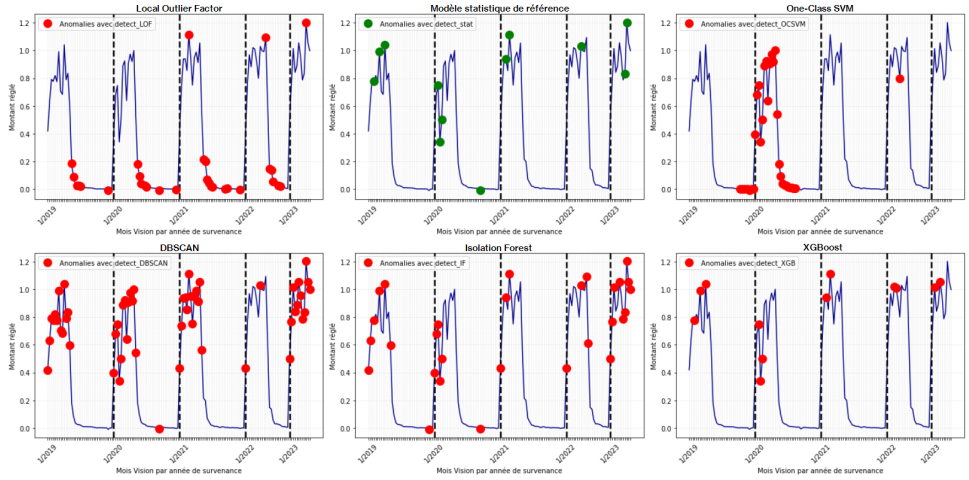
A.6 Détection des anomalies pour les autres délégataires non EDI





A.7 Détection des anomalies pour la gestion directe





A.8 Résultats des modèles de détection sur les quatre segments d'étude

Délégataire A

	methode_detection	Nb anomalies	Précision	Recall	F1-score
1	Local Outlier Factor	23	8.70%	13.33%	10.53%
2	One-Class SVM	32	15.62%	33.33%	21.28%
3	DBSCAN	81	18.52%	100.00%	31.25%
4	Isolation Forest	29	51.72%	100.00%	68.18%
5	XGBoost	11	100.00%	73.33%	84.62%

Délégataire B

	methode_detection	Nb anomalies	Précision	Recall	F1-score
1	Local Outlier Factor	34	23.53%	34.78%	28.07%
2	One-Class SVM	31	12.90%	17.39%	14.81%
3	DBSCAN	66	34.85%	100.00%	51.69%
4	Isolation Forest	28	64.29%	78.26%	70.59%
5	XGBoost	28	60.71%	73.91%	66.67%

Autres délégataires non EDI

	methode_detection	Nb anomalies	Précision	Recall	F1-score
1	Local Outlier Factor	11	72.73%	13.79%	23.19%
2	One-Class SVM	32	31.25%	17.24%	22.22%
3	DBSCAN	91	63.74%	100.00%	77.85%
4	Isolation Forest	29	55.17%	27.59%	36.78%
5	XGBoost	64	89.06%	98.28%	93.44%

Gestion directe

	methode_detection	Nb anomalies	Précision	Recall	F1-score
1	Local Outlier Factor	30	10.00%	25.00%	14.29%
2	One-Class SVM	32	9.38%	25.00%	13.64%
3	DBSCAN	54	22.22%	100.00%	36.36%
4	Isolation Forest	29	41.38%	100.00%	58.54%
5	XGBoost	12	66.67%	66.67%	66.67%

Annexe B

Compléments sur les méthodes de lissage

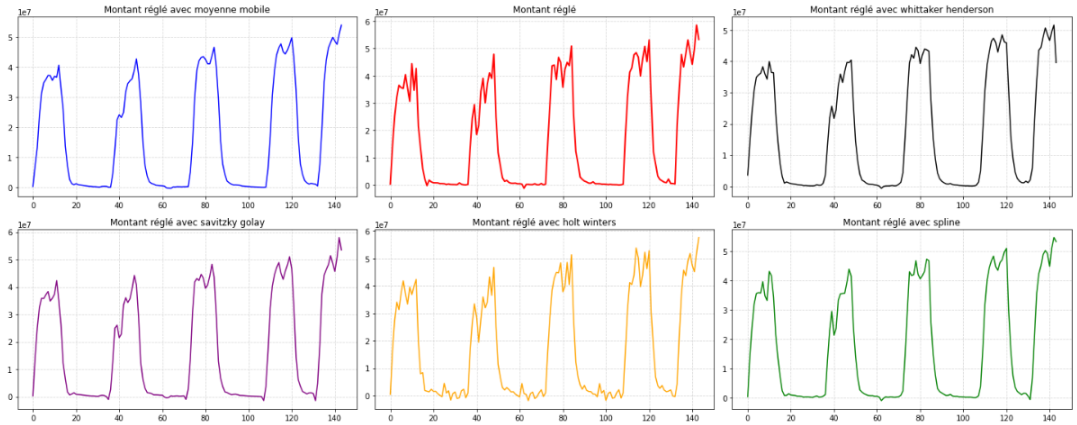
B.1 Mesures de performance du lissage

Métrique	Définition	Formule de calcul
MAE (Mean Absolute Error)	Moyenne des erreurs absolues entre les valeurs réelles et les valeurs prédites.	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
RMSE (Root Mean Squared Error)	Racine carrée de la moyenne des erreurs quadratiques, plus sensible aux grandes erreurs, ce qui la rend utile pour évaluer la dispersion.	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
R^2 (Coefficient of Determination)	Proportion de la variance des données observées qui est expliquée par le modèle de prédiction. Il varie entre 0 et 1, où 1 indique une parfaite prédiction.	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ avec $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
MAPE (Mean Absolute Percentage Error)	Moyenne des erreurs absolues en pourcentage, permettant d'évaluer l'erreur de prédiction en termes relatifs.	$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \times 100\%$

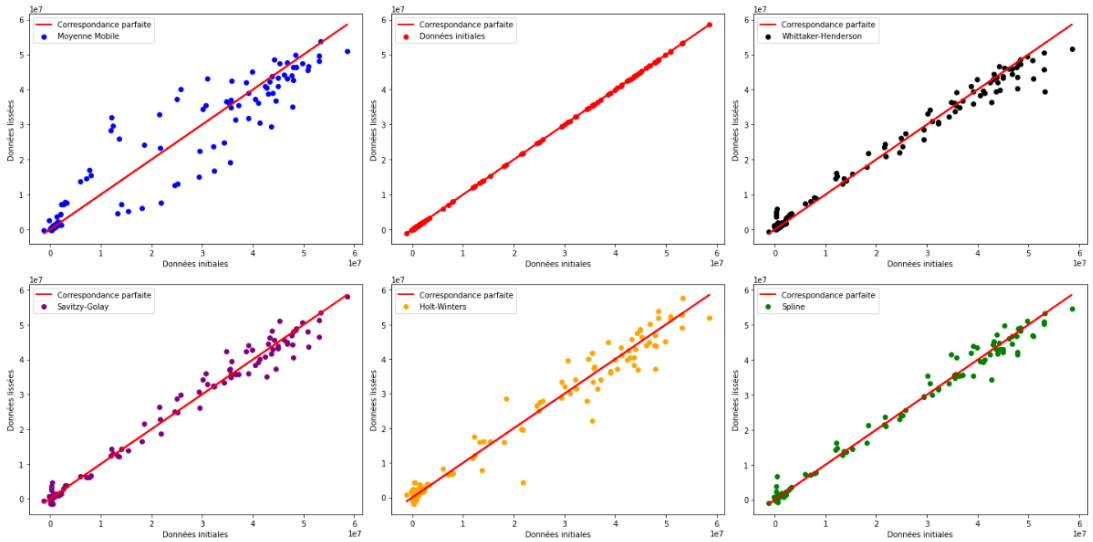
B.2 Tableau récapitulatif des méthodes de lissage testées

Méthode de lissage	*Utilité*	*Fonctionnement*	*Paramètres*	*Avantages*	*Inconvénients*
Moyenne mobile	Lissage des variations aléatoires dans une série temporelle	Calcul de la moyenne des observations dans une fenêtre mobile glissante	Taille de la fenêtre de la moyenne mobile	- Facile à comprendre et à mettre en œuvre - Efficace pour lisser les tendances à court terme	- Ne tient pas compte des variations saisonnières - Sensible aux valeurs aberrantes
Whittaker-Henderson	Lissage pour minimiser les variations brusques dans une série temporelle	Lissage par interpolation polynomiale pour minimiser une fonction coût pénalisant les variations brusques	Choix de la fonction coût, spécification du degré du polynôme d'interpolation	- Peut atténuer les variations brusques tout en conservant les tendances lentes	- Nécessite le choix de la fonction coût et du degré du polynôme - Sensible aux valeurs aberrantes
Savitzky-Golay	Lissage pour minimiser l'effet du bruit tout en préservant les caractéristiques importantes	Ajustement polynomial local pour lisser les données tout en préservant les caractéristiques importantes du signal	Longueur de la fenêtre, ordre du polynôme	- Peut lisser les données tout en préservant les infos importantes - Robuste aux valeurs aberrantes	- Sensible au choix de la longueur de la fenêtre et de l'ordre du polynôme - Rigueur pour ajuster les paramètres
Holt-Winters	Modélisation et prévision de séries temporelles avec des tendances et des saisons	Utilisation de composantes de tendance, de saisonnalité et de niveau pour modéliser les variations	Paramètres de lissage pour la tendance, la saisonnalité et le niveau, période saisonnière	- Peut capturer les tendances et les variations saisonnières - Peut être utilisé pour la prévision	- Peut être sensible aux fluctuations extrêmes et aux schémas inattendus
Spline	Lissage en ajustant des polynômes locaux à des sous-ensembles de données	Ajustement local de polynômes pour lisser les données tout en préservant les caractéristiques importantes	Facteur de lissage	- Peut lisser les données tout en préservant les infos importantes - Flexible et adaptable	- Nécessite le choix adéquat du facteur de lissage pour éviter le surajustement ou le sous-ajustement

B.3 Comparaison des modèles de lissage sur le délégataire A



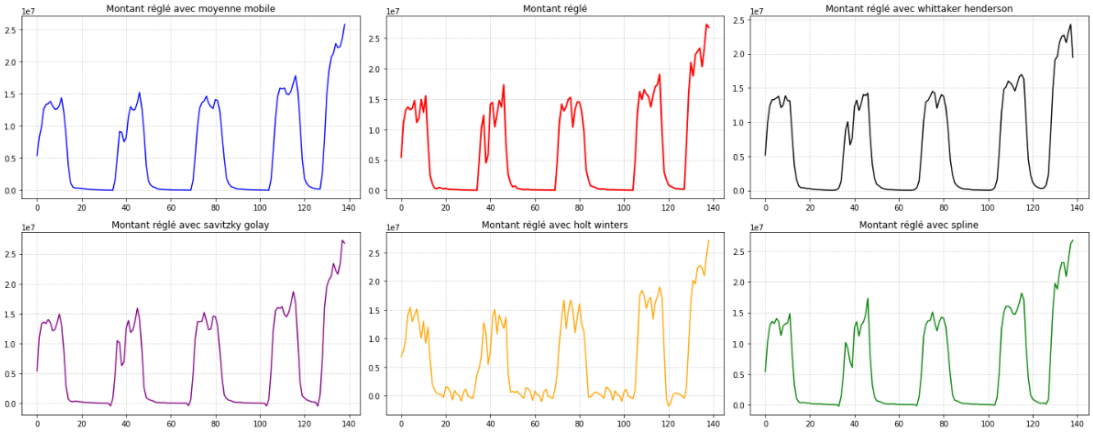
B.4 Graphiques de dispersion du lissage des règlements du délégataire A



B.5 Tableau des métriques du lissage des données du délégataire B

	methode_lissage	MAE	RMSE	R2	MAPE
1	Moyenne Mobile	1.34e+06	2.36e+06	0.90	52.96
2	Whittaker-Henderson	5.82e+05	1.07e+06	0.98	938.59
3	Savitzky-Golay	4.52e+05	7.78e+05	0.99	458.17
4	Holt-Winters	1.54e+06	2.31e+06	0.91	1132.25
5	Spline	3.73e+05	6.70e+05	0.99	755.76

B.6 Comparaison des modèles de lissage sur le délégataire B



B.7 Tableau des métriques du lissage des données de la gestion directe

	methode_lissage	MAE	RMSE	R2	MAPE
1	Moyenne Mobile	8.04e+06	1.37e+07	0.89	49.39
2	Whittaker-Henderson	3.45e+06	5.58e+06	0.98	290.41
3	Savitzky-Golay	3.39e+06	5.25e+06	0.98	244.92
4	Holt-Winters	5.73e+06	9.27e+06	0.95	261.42
5	Spline	2.92e+06	5.23e+06	0.98	136.26

B.8 Graphiques de dispersion du lissage des règlements de la gestion directe

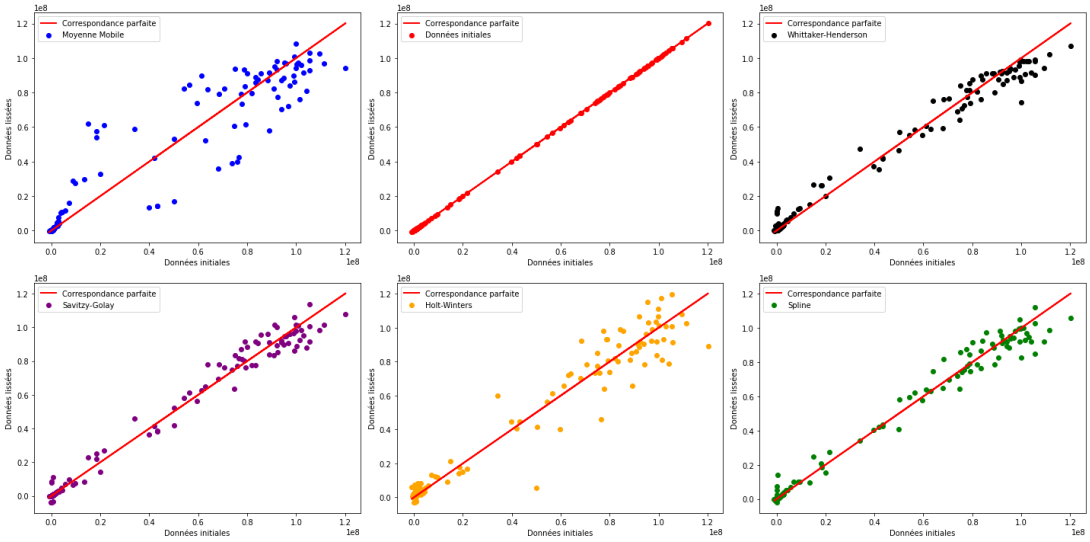


Table des figures

1.1 Fonctionnement de l'assurance santé en France 26
1.2 Illustration du phénomène de transfert de charge 30

2.1 Indicateurs clés et objectifs de la *data* qualité 35
2.2 Classification des segments d'étude selon le type de gestion 40
2.3 Répartition des montants réglés des contrats santé collectifs en fonction du type de gestion 42
2.4 Répartition des montants réglés des contrats santé collectifs en fonction des segments d'étude 43
2.5 Montant réglé total par mois de vision sur les 5 dernières années de survenance 44
2.6 Evolution des montants réglés par mois comptable et par survenance . . . 45
2.7 Histogramme et *Boxplot* des règlements de santé en fonction de l'année de survenance 46
2.8 Evolution du pourcentage de règlement par survenance et pour chaque année de vision 47
2.9 Diagrammes sur le pourcentage de règlement par année de survenance . . 48
2.10 Montant réglé sur les 5 dernières années de survenance par type de gestion 49
2.11 Montant réglé par mois vision comptable et délégataire 50

3.1 Exemple d'anomalies pour un client sur les 5 dernières années de survenance groupées 54
3.2 Exemple d'anomalies sur un contrat de santé et correction avec rattrapage 55
3.3 Exemples d'anomalies complexes visibles sur des contrats de santé 55
3.4 Répartition du portefeuille en santé collective 59
3.5 Montants réglés sur 36 mois de vision pour les 5 dernières années de survenance, selon les quatre segments d'analyse 60
3.6 Pourcentage cumulé des règlements sur 36 mois pour le délégataire A . . . 61
3.7 Détection statistique des anomalies par survenance sur le segment A . . . 62
3.8 Détection statistique des anomalies sur le segment d'étude A 63
3.9 Détection statistique des anomalies sur les trois autres segments d'étude . 63
3.10 Hyperparamétrage de la LOF sur le délégataire A 66
3.11 Détection des anomalies avec Local Outlier Factor 67

3.12 Détection des anomalies avec One-Class SVM 70
3.13 Fonctionnement de l'algorithme DBSCAN 71
3.14 Hyperparamétrage du modèle DBSCAN 72
3.15 Détection des anomalies avec DBSCAN 72
3.16 Fonctionnement de l'algorithme Isolation Forest 73
3.17 Détection des anomalies avec Isolation Forest 75
3.18 Détection des anomalies avec XGBoost 79
3.19 Tableau du score de prédiction des anomalies 80
3.20 ROC-curve des modèles de détection pour le délégataire B 80
3.21 Montants réglés lissés par moyenne mobile 82
3.22 Lissage des règlements avec spline 83
3.23 Lissage des règlements avec Whittaker-Henderson 85
3.24 Analyse temporelle des règlements du délégataire A 87
3.25 Lissage des règlements avec Holt-Winters 88
3.26 Lissage des règlements avec Savitzky-Golay 90
3.27 Tableau des métriques de performance pour le lissage des données du délégataire A 91
3.28 Graphiques de dispersion du lissage des règlements du délégataire B . . . 91
3.29 Comparaison des modèles de lissage sur les montants associés à la gestion directe 92

4.1 Fonctionnement du modèle de détection-correction des données 96
4.2 Application du modèle de détection-correction sur le délégataire A 97
4.3 Règlements cumulés du délégataire A par mois de vision, pour chaque année de survenance 100
4.4 Vérification de l'alignement de Ci,j et Ci,j+1 sur 2 périodes 100
4.5 Statistiques sur les cadences de règlement pour le délégataire A 101
4.6 Tableau de provisionnement sur les données initiales du délégataire A . . . 102
4.7 Dispersion des résidus par survenance sur le délégataire A 103
4.8 Estimation des IBNR du délégataire A suivant les quatre approches 104
4.9 Projection des règlements du délégataire A selon les quatre approches . . 105
4.10 Estimation des IBNR du délégataire B suivant les quatre approches 106
4.11 Projection des règlements du délégataire B selon les quatre approches . . 106
4.12 Estimation des IBNR de la gestion directe suivant les quatre approches . . 107
4.13 Projection des règlements de la gestion directe selon les quatre approches 108

Bibliographie

[1] M. Asad Iqbal Khan. *Anomaly Detection with Isolation Forest and Kernel Density Estimation*, 2022. URL : https://machinelearningmastery.com/anomaly-detection-with-isolation-forest-and-kernel-density-estimation/.

[2] D. Bardou. *Reste à charge zéro : quel fonctionnement en 2024 ?*, 2024. URL : https://reassurez-moi.fr/guide/mutuelle-sante/reste-a-charge-zero.

[3] B. Benjamin and J. H. Pollard. *The Analysis of Mortality and Other Actuarial Statistics*, 1993.

[4] G. Biessy. *Une vision moderne du lissage de Whittaker-Henderson*, 2023. URL : https://hal.science/hal-04124043.

[5] T. Bomtems and S. Goulin. *Qualité de l'information*, 2013.

[6] R. Brunet. *Comparaison des méthodes de lissage de tables de mortalité périodiques et étude de leur impact sur le provisionnement*, mémoire d'actuariat, 2018.

[7] H. Chen, J. Chen, and J. Ding. *Data Evaluation and Enhancement for Quality Improvement of Machine Learning*, 2021.

[8] DataValueConsulting. *Comment appliquer une stratégie de qualité des données à l'échelle de toute l'entreprise ?*, 2021. URL : https://datavalue-consulting.com/strategie-qualite-donnees-entreprise/.

[9] DelftStack. *How to Smooth Data in Python*, 2023. URL : https://www.delftstack.com/fr/howto/python/smooth-data-in-python/.

[10] L. Dierickx. *Cours de l'Université Libre de Bruxelles (ULB) sur l'apprentissage automatique : les challenges de la qualité des données dans la perspective d'une adéquation aux usages*, 2019. URL : https://mastic.ulb.ac.be/wp-content/uploads/2022/02/expose_final_dq.pdf.

[11] EURIA. *Estimation de l'erreur de prédiction dans le cas de l'utilisation d'une combinaison de méthodes pour le calcul de provisions en assurance IARD*, 2014. URL : https://www.univ-brest.fr/euria/sites/euria.nouveau.univ-brest.fr/files/2022-06/estimation_erreur_2013-2014.pdf.

[12] D. Fabre Rudelle. *Apport des méthodes d'apprentissage statistique pour le provisionnement individuel en assurance non-vie*, mémoire d'actuariat, 2018.

[13] P. Hairy. *La détection d'anomalies en Machine Learning non supervisé*, 2022. URL : https://metalblog.ctif.com/2022/10/03/la-detection-danomalies-en-machine-learning-non-supervise/.

[14] Hiraltalsaniya. *Anomaly Detection with Unsupervised Machine Learning*, 2023. URL : https://medium.com/simform-engineering/anomaly-detection-with-unsupervised-machine-learning-3bcf4c431aff.

[15] S. A. Klugman, H. H. Panjer, and G. E. Willmot. *Loss Models : From Data to Decisions*, 2004.

[16] Kobia. *Sensitivity, Specificity et ROC curve*, 2018. URL : https://kobia.fr/classification-metrics-sensitivity-specificity-roc/.

[17] A. Krishnan. *Anomaly Detection with Isolation Forest & Visualization*, 2019. URL : https://towardsdatascience.com/anomaly-detection-with-isolation-forest-visualization-23cd75c281e2.

[18] A. Lagnoux. Séries chronologiques - prévision par lissage exponentiel, 2013. URL : https://perso.math.univ-toulouse.fr/lagnoux/files/2013/12/Chap6.pdf.

[19] T. Maurras Ulbricht, Y. Chabchoub, A. Boly, and R. Chiky. *Étude comparative des méthodes de détection d'anomalies*, 2020. URL : https://hal.science/hal-02874904/document.

[20] MBB Assurances. *Tout savoir sur l'assurance santé collective en entreprise*, 2022. URL : https://www.mbb-assurances.fr/tout-savoir-sur-lassurance-sante-collective-en-entreprise/.

[21] N. Nomaou Windiá. *Modélisation de l'impact de la réforme de la protection sociale complémentaire dans la fonction publique de l'Etat*, mémoire d'actuariat, 2023.

[22] J. Robert. *Isolation Forest : Comment détecter les anomalies dans une dataset ?*, 2021. URL : https://datascientest.com/isolation-forest.

[23] J. Tardy. *Amélioration de la qualité des données en assurance par apprentissage automatique*, mémoire d'actuariat, 2018.

[24] V. Teissier. *Analyse et standardisation de méthodes de lissage pour une application globale à travers tous les risques d'un réassureur vie et santé*, mémoire d'actuariat, 2023.

[25] V. Uthayasooriyar. *Analyse et traitement du risque des valeurs extrêmes en prévoyance collective*, mémoire d'actuariat, 2015.

[26] G. Welterlin. *Calibration du choc prime sur un portefeuille de santé collective*, mémoire d'actuariat, 2020.