

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

Par : Manon DUTRUC-ROSSET

Titre : Modélisation de la sinistralité des polices d'activités en assurance construction dans un contexte de forte inflation

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

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

Entreprise :

Nom : GENERALI

Signature : 

Membres présents du jury de Signature
l'Institut des Actuaires :

A. Derien

C. Moyne

A. Mantel

Directeur de mémoire en entre-
prise :

Nom : GUFFROY DESCHAMPS
Cindy

Signature : 

Membres présents du jury de
l'ISFA :

A. Couloumy

Invité :

Nom :

Signature :

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

Signature du responsable entreprise



Signature du candidat



Résumé

Mots clés : assurance construction, Responsabilité Civile Décennale, Responsabilité Civile Bâtiment, inflation, fréquence, sévérité, GLM

Le secteur de la construction, comme de nombreux autres secteurs, est confronté à une forte inflation, et en juin 2022, elle atteint 8,26 %. Ce surcoût se répercute logiquement sur celui des sinistres, et les garanties avec une duration plus longue sont plus spécifiquement concernées. En effet, pour l'assurance construction décennale, la prime est payée à l'ouverture du chantier, tandis que le sinistre peut survenir plus de dix ans après, ainsi un même sinistre peut avoir un coût très différent entre ces deux dates. Ce changement du contexte économique s'accompagne de nombreux changements sur le portefeuille Generali en construction et sur la gestion des sinistres. L'objectif de ce mémoire sera de réaliser des travaux sur la modélisation des charges pour les garanties Responsabilité Civile Décennale et Responsabilité Civile Bâtiment, en prenant en compte l'inflation.

La première partie des travaux consistera à traiter les données afin de supprimer les valeurs manquantes ou aberrantes, et de produire une base de modélisation de la charge. Celle-ci sera réalisée avec un modèle de fréquence et de sévérité en utilisant les modèles linéaires généralisés en premier lieu. En amont de la modélisation des coûts, une vérification de l'adéquation de la véracité du seuil historique de graves sera faite avec la théorie es valeurs extrêmes. Elle permettra de déduire que le montant historique est cohérent avec l'historique de sinistralité, cependant des seuils supplémentaires doivent être ajoutées, afin de créer quatre tranches distinctes de sévérité.

On utilisera de manière complémentaire les techniques de Machine Learning (Arbres de décision et Random Forest, Extrêmes gradients boosting) pour modéliser la fréquence et la première tranche de sévérité. Enfin, une étude et une prédiction de l'inflation sera effectuée avec les séries temporelles. La dernière partie permettra de mettre en avant l'importance de considérer l'inflation dans le contexte inflationniste, et les prédictions de la charge pour les prochaines années.

Abstract

Key words : Construction insurance, ten-year public liability, building civil liability, inflation, frequency, severity, GLM

The construction sector, like many other sectors, is facing with high inflation, which reached 8.26% in June 2022. This additional cost has logically an effect on the amount of claims and guarantees with a longer duration are more specifically concerned. Indeed, for ten-year construction insurance, the premium is paid at the start of the construction site, while the claim can occur more than ten years later, so the same claim can have a very different cost between these two dates. This change in the economic context is accompanied by several changes to the Generali portfolio under construction and to claims management. The objective of this thesis will be to carry out work on the modeling of the charges for the ten-year public liability and building civil liability guarantees, taking inflation into account.

The first part of the work will consist of processing the data to remove missing or outlying values, and to produce a database for modeling the cost. This will be carried out with a frequency and severity model using generalized linear models. Before modeling the costs, a verification of the adequacy of the veracity of the historical threshold of seriousness will be made with the theory of extreme values. This historical amount is consistent with the history of claims, however additional thresholds must be added, they will create four distinct group of severity.

We will also use Machine Learning techniques (Decision Trees and Random Forest, Extreme Gradient Boosting) to model the frequency and the first group of severity. Finally, a study and a prediction of inflation will be carried out with the time series. The last part will highlight the importance of considering inflation in the inflationary context, and the forecasts of the load for the coming years.

Remerciements

Je tiens à remercier, en premier lieu, Mme Cindy GUFFROY-DESCHAMPS, ma tutrice pendant mon année d'alternance chez Generali, pour sa confiance, son écoute et son aide dans la réalisation de ce mémoire. Je remercie également Alexandre ZULIANI, pour ses connaissances sur l'assurance construction et son aide précieuse pour le mémoire. Plus généralement, je souhaite remercier toute l'équipe de la TA-IARD et les alternants de cette équipe avec qui j'ai pu échanger tout au long de l'année.

Je remercie toute l'équipe pédagogique de l'Institut des Science Financière et d'Assurances (ISFA) pour leur accompagnement pendant ces trois années d'études, et plus particulièrement M. Frédéric PLANCHET, mon tuteur pédagogique pour les bons échanges que nous avons pu avoir et qui m'ont aidé dans la réalisation de ce mémoire.

Enfin, je remercie ma famille, et mes amis pour leur soutien, leurs conseils et les relectures.

Table des matières

Résumé	3
Abstract	5
Remerciements	7
Table des matières	7
Glossaire	11
Introduction	13
1 Présentation de l'assurance construction	15
1.1 Présentation générale	15
1.1.1 Les intervenants	15
1.1.2 Dispositions juridiques	16
1.1.3 Les différents types d'assurance	18
1.1.4 Les différentes garanties	18
1.1.5 Les dates de l'assurance construction décennale	19
1.2 Marché de l'assurance construction en France	20
1.3 Inflation dans l'assurance construction	21
2 Présentation du portefeuille	25
2.1 Présentation de l'entreprise et du portefeuille	25
2.2 Évolution du portefeuille	26
3 Préparation de la base Sinistre	29
3.1 Présentation des données initiales	29
3.2 Sélection des données	31
3.2.1 Choix des variables	31
3.2.2 Traitement des données manquantes	32
3.3 Revalorisation des sinistres	33
3.4 Segmentation des sinistres	35
3.4.1 Théorie des valeurs extrêmes	36
3.4.2 Détermination de seuils de gravité avec différents estimateurs	37
3.4.3 Choix des seuils de gravité	40
4 Préparation de la base Risque	43
4.1 Récupération des données initiales	43
4.2 Création de la base	43

4.2.1	Choix des variables	44
4.2.2	Traitement des valeurs manquantes et aberrantes	46
5	Modélisation de la fréquence et de la sévérité	49
5.1	Préparation à la modélisation	50
5.1.1	Création des bases fréquence	50
5.1.2	Création des bases sévérité	52
5.1.3	Création des bases Sans-Suite	53
5.1.4	Regroupement et création de classes des modalités	54
5.1.5	Corrélations	55
5.2	Mesure de la performance prédictive d'un modèle	57
5.2.1	Base d'apprentissage et de test	57
5.2.2	Critère de sélection d'un modèle	58
5.3	Modèles Linéaires Généralisés	59
5.3.1	Théorie	59
5.3.2	Sélection des variables et simplification	60
5.3.3	Validation des modèles	72
5.4	Modèles de Machine Learning	75
5.4.1	Arbres de décision - CART	75
5.4.2	Forêts aléatoires - Random Forest	76
5.4.3	XG Boost	77
5.5	Comparaison des modèles	78
6	Modélisation avec les séries temporelles de l'inflation	79
6.1	Historique de l'inflation du BT01	80
6.2	Approche avec les séries temporelles	81
6.2.1	Concept et propriétés des séries temporelles	81
6.2.2	Étude de la stationnarité	82
6.2.3	Étude des autocorrélations et sélection des paramètres p et q	82
6.2.4	Prédictions de l'inflation	83
7	Application et résultats	85
	Conclusion	91
	Table des figures	92
	Bibliographie	95
	Annexe A : Informations supplémentaires sur les données	97
	A.1 Répartition du portefeuille	97
	A.2 Répartition des activités	97
	Annexe B : Choix des lois de modélisation	99
	Annexe C : Regroupements et choix des variables dans les modélisations	101
	Annexe D : Séries Temporelles	105
	E.1 Étude de la tendance	105
	E.2 KPSS Test	105
	E.3 Choix des paramètres p et q	107

Glossaire

AIC : Akaike Information Criterion (Critère d'Information d'Akaike)
BCE : Banque Centrale Européenne
BF : Banque de France
BIC : Bayesian Information Criterion (Critère d'Information Bayésien)
BTP : Bâtiment et Travaux Publics
BT01 : Indice National du Bâtiment (tous corps d'état)
CA : Chiffre d'Affaires
CA_ST : Chiffre d'Affaires Sous Traité
DROC : Date Réglementaire d'Ouverture de Chantier (ou **DOC** : Date d'Ouverture de Chantier)
ETP : Équivalent Temps Plein
FABNG : Fabrication - Négoce
FFB : Fédération Française du Bâtiment
GEV : Generalized Extreme Value
GLM : Generalized Linear Models (Modèles linéaires généralisés)
GO : Gros Oeuvre
ICC : Indice du Coût de la Construction
ICPC : Indices des coûts de production dans la construction
IID : Indépendantes et Identiquement Distribuées
INSEE : Institut National de la Statistique et des Études Économiques
IPC : Indice des Prix à la Consommation
RC Bât : Responsabilité Civile Bâtiment
RCD : Responsabilité Civile Décennale
MOA : Maître d'Ouvrage
MOE : Maître d'Œuvre
ProPE : Professionnels et Petite Entreprise
SO : Second Oeuvre
ZC : Clos et couvert
ZD : Lots divisions - Aménagement
ZP : Préparation et aménagement du site
ZS : Structure et Gros Œuvre
ZT : Lots techniques
ZX : Travaux tous corps d'état

Introduction

Dans un objectif de confidentialité, les valeurs présentées dans ce mémoire ont été modifiées. Elles ont été multipliées par un même coefficient choisi arbitrairement, permettant ainsi de respecter la cohérence des résultats entre eux.

L'assurance construction a la particularité de compter des garanties décennales obligatoires. En effet, l'obligation d'assurance pendant 10 ans à la réception de travaux, et ainsi des périodes d'assurances pouvant dépasser 15 ans, ne se retrouvent dans aucune autre garantie. La première apparition de la mention de Responsabilité Civile Décennale (RCD) est présente dans le Code Civil de 1804 dans les articles 1792 et 2270. Elle fut éditée pour protéger les propriétaires ou habitants d'un ouvrage face aux aléas de la construction après la réception des travaux. Plusieurs réformes jusqu'à la loi Spinetta de 1978 vont permettre de compléter et de préciser ces articles, et ainsi la définition de la responsabilité décennale.

Les changements apportés instaurent l'obligation d'assurance du risque de responsabilité civile décennale pour les constructeurs, et une gestion en capitalisation de cette garantie dès 1983. Désormais, les sinistres survenus pendant les dix ans suivant la réception des travaux doivent être couverts par les primes perçues durant l'exercice pendant lequel le chantier a démarré.

Cette obligation d'assurance donne un rôle à jouer aux acteurs de ce secteur pour assurer une protection efficace du secteur de la construction, qui représente une part importante de l'économie française. En 2019, 486 876 entreprises avec près de 1,4 millions de salariés réalisent 331,7 milliards d'euros de chiffre d'affaires hors taxes, ce montant correspondant à plus de deux fois celui du secteur de l'assurance.

La RCD est une police d'activités, tout comme la Responsabilité Civile Bâtiment, qui est une assurance responsabilité civile professionnelle. Ces deux types de garantie seront les seules traitées dans ce mémoire, et ont la particularité d'être souscrites de manière simultanée chez Generali.

En 2020, pendant la crise sanitaire, l'activité du bâtiment a fortement chuté du fait des confinements et des restrictions de circulation qui ont limité l'accès aux matières premières. En 2021, l'activité reprend mais cette reprise est confrontée à une pénurie de matières premières, qui est due à un arrêt de production puis à une reprise intense. Une tension apparaît alors et les prix augmentent fortement. Ce contexte d'inflation s'installe en 2021 et se renforce en 2022 avec la guerre en Ukraine, il touche désormais tous les secteurs d'activités, et l'incertitude autour de l'évolution de la situation semble indiquer que l'inflation restera importante pendant plusieurs mois.

Les évolutions contextuelles s'accompagnent, au sein du groupe Generali France, d'évolutions structurelles du portefeuille avec de nouvelles politiques de souscription visant à privilégier certains assurés, ainsi que de nombreuses résiliations afin de rétablir une bonne solvabilité. Toutes ces modifications, accompagnées d'autres dans la gestion du portefeuille permettent d'expliquer la nécessité de redéfinir les méthodes de modélisations de la sinistralité. De plus, actuellement, il existe seulement un modèle pour la RCD avec une segmentation des sinistres se faisant autour d'un seuil de grave à 150 000 €.

L'objectif de ce mémoire est, premièrement, de vérifier l'adéquation du seuil de gravité avec l'historique de sinistralité et de définir d'éventuels seuils supplémentaires. Secondement, de reprendre le modèle de détermination de la sinistralité en ajoutant des variables explicatives telles que le chiffre d'affaires, l'effectif et une donnée géographique. Il conviendra également de créer un modèle pour la RC Bât, et de comparer les résultats afin de comprendre les éléments significatifs en fonction des différentes garanties. Enfin, l'inflation étant aujourd'hui en forte augmentation, il semble important de la considérer dans les modèles. Ainsi une méthode de prise en compte de l'inflation devra être étudiée.

La première partie de ce mémoire est une présentation de l'assurance construction, du marché en France et de ses évolutions dans un contexte particulier suite à la crise sanitaire du Covid-19 et la guerre en Ukraine. Elle contient notamment les définitions des garanties qui seront traitées dans ce mémoire : la responsabilité Civile Décennale (RCD) ainsi que la Responsabilité Civile Bâtiment (RC Bât).

La deuxième partie analyse les données disponibles. Elle met en avant les évolutions structurelles ayant impacté le portefeuille assuré, et les changements apportés dans la gestion des sinistres. Ces observations ont fortement modifié le profil de risque, ainsi il convient de redéfinir les méthodes et les modèles de provisionnement et de tarification. La troisième, ainsi que la quatrième parties traitent des travaux réalisés afin de créer les deux bases de données, l'une contenant les informations sur la sinistralité et l'autre les informations sur le risque. Pour l'étude des sinistres, une segmentation des sinistres est réalisée afin de vérifier la cohérence du seuil de gravité historique du groupe de 150 000 €, et elle permettra de définir d'autres seuils pour segmenter d'avantage ces sinistres. Tandis que pour la base Risque, les principales tâches sont le traitement de données et en particulier des valeurs manquantes.

La cinquième partie contient l'ensemble des travaux réalisés pour définir le modèle proposé. Plus particulièrement, les modèles linéaires généralisés sont utilisés pour estimer la fréquence et les coûts des sinistres en tenant compte des profils de risques présents en portefeuille au cours du temps.

La sixième partie traite de l'inflation et des choix effectués afin de la modéliser. Sa définition est rappelée ainsi que son évolution pendant les 50 dernières années.

La dernière partie est une application des modèles obtenus, elle permet de créer le modèle final combinant la fréquence, le coût et l'inflation.

Présentation de l'assurance construction

L'assurance construction permet d'assurer une protection du maître d'ouvrage et de l'utilisateur d'un bien face aux dommages pouvant être causés sur l'édifice construit. Le maître d'ouvrage bénéficie d'une couverture pour les dommages pouvant altérer la solidité de la construction ou pouvant la rendre impropre à destination. Le propriétaire de l'édifice obtient, quant à lui, des fonds pour financer les réparations nécessaires à la remise en état du bâtiment, sans avoir besoin d'attendre que le responsable du vice de construction soit désigné par la justice. Par la suite, la société d'assurance pourra se retourner vers la partie adverse pour récupérer l'équivalent des indemnités versées.

Cette première partie permet de présenter le contexte dans lequel s'inscrit ce mémoire, ainsi que les spécificités liées à ce type d'assurance. Une présentation générale sera, dans un premier temps, exposée. Elle décrira les différents intervenants dans un chantier en construction, les dispositions juridiques qui lui sont associées et les spécificités de l'assurance Construction. Dans un deuxième temps, le marché de la Construction en France sera décrit, enfin l'inflation sur ce marché et ses conséquences seront abordées.

1.1 Présentation générale

1.1.1 Les intervenants

La réalisation d'un projet en Construction vise à édifier un ouvrage dans un lieu défini selon le choix du maître d'ouvrage et conformément à la vision du maître d'œuvre. Elle met en relation différents intervenants, ayant des caractéristiques propres et liés selon différents contrats.

Le maître d'ouvrage (MOA) : personne morale ou physique, pour laquelle les travaux sont réalisés. Il est le commanditaire du projet et doit apporter le financement nécessaire aux différents travaux du chantier ainsi que définir le programme de construction grâce à un cahier des charges. Il est garant des délais de construction sur le chantier et doit valider les étapes de conception et de production du chantier. Enfin il assure la réception des travaux. Il a la possibilité de déléguer certaines missions à un tiers que l'on nomme le maître d'ouvrage délégué (MOD).

Le maître d'œuvre (MOE) : personne chargée, par le maître d'ouvrage, de la réalisation du projet de construction. Il est présent de la conception à la réception du chantier, il dirige et contrôle l'exécution des travaux. La maîtrise d'œuvre regroupe les architectes, les bureaux d'études et les ingénieurs qui les composent, et les autres techniciens spécialisés de la construction. Le maître d'œuvre et le maître d'ouvrage établissent un contrat de louage d'ouvrage pour le projet de construction.

Les réalisateurs : personnes physiques ou morales titulaires d'un marché de travaux. Il convient de distinguer deux types d'entreprises : l'entreprise générale réalisatrice unique de l'ensemble de l'ouvrage, et l'entreprise réalisatrice d'un lot et ainsi engagée dans la réalisation d'une partie de l'ouvrage (il s'agit souvent d'une entreprise spécialisée et qui interviendra seulement dans le domaine de ses compétences). Le marché de travaux est caractérisé par un contrat de louage d'ouvrage établi entre le maître de l'ouvrage et un des réalisateurs. Une entreprise peut sous-traiter en totalité ou une partie l'exécution de son marché, à une ou plusieurs entreprises. Un contrat de sous-traitance sera alors établi entre l'entreprise désignée par le contrat de louage d'ouvrage et l'entreprise sous-traitante.

Le schéma ci-dessous permet de résumer les relations entre les intervenants d'une opération de construction :

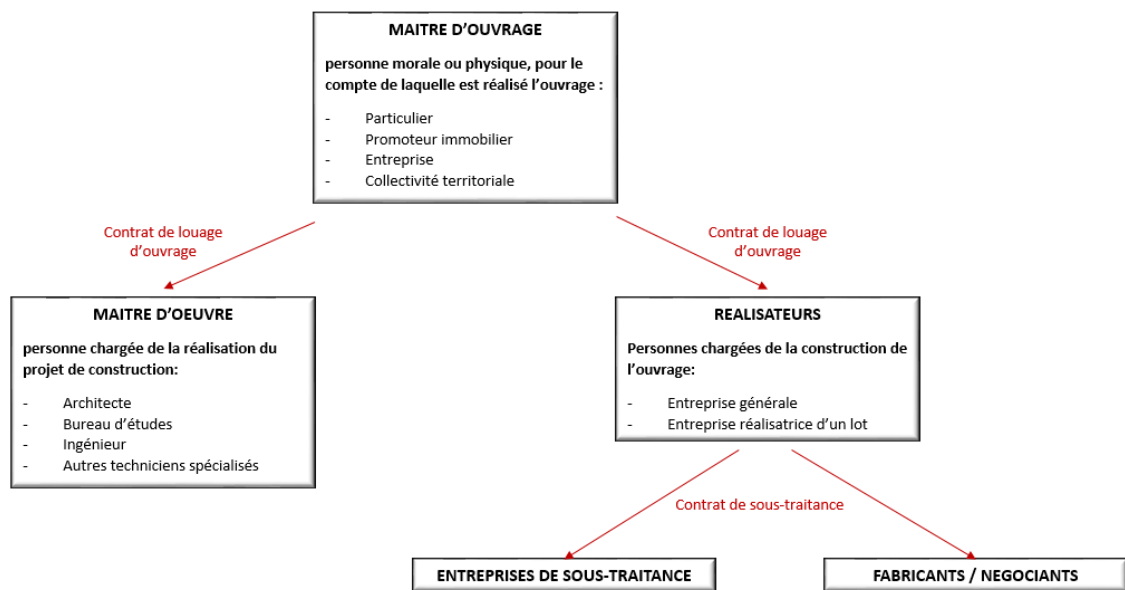


FIGURE 1.1 – Liens contractuels entre les différents intervenants sur une opération de construction

1.1.2 Dispositions juridiques

La première notion de Responsabilité Civile Décennale est définie dans les articles 1792 et 2270 du Code Civil de 1804. Ils définissent la responsabilité des architectes et entrepreneurs pendant 10 ans après la réception des travaux :

- **Article 1792 :** *Si l'édifice construit à prix fait, périt en tout ou en partie par le vice de la construction, même par le vice du sol, les architecte et entrepreneur en sont responsables pendant dix ans.*
- **Article 2270 :** *Après dix ans, l'architecte et les entrepreneurs sont déchargés de la garantie des gros ouvrages qu'ils ont faits ou dirigés.*

Plusieurs réformes vont avoir lieu et vont modifier ses articles en élargissant et en précisant cette notion de responsabilité civile décennale. En effet, ces articles n'interviennent que pour un effondrement total ou partiel, dans les autres cas la faute des constructeurs devait être prouvée. De plus, les types d'édifice concernés par cette responsabilité décennale ne sont pas clairement identifiés et le terme de "gros ouvrages" dans l'article 2270 n'est pas précis. Enfin, l'obligation d'assurance n'est pas indiquée.

La loi du 3 janvier 1967 modifia ces articles 1792 et 2270 du Code Civil afin d'étendre et de préciser les responsabilités engagées dans la construction. Cette loi impose l'obligation de garantie à raison des vices de construction.

Cependant, une nouvelle loi publiée au Journal Officiel du 4 janvier 1978, nommée **loi Spinetta**, du nom d'Adrien Spinetta, président de la commission interministérielle ayant rédigé des propositions de réformes pour l'assurance construction, impose de nouvelles modifications.

Loi Spinetta : *"Toute personne physique ou morale qui, agissant en qualité de propriétaire de l'ouvrage, de vendeur ou de mandataire du propriétaire de l'ouvrage, fait réaliser des travaux de bâtiment, doit souscrire avant l'ouverture du chantier, pour son compte ou pour celui des propriétaires successifs, une assurance garantissant, en dehors de toute recherche des responsabilités, le paiement des travaux de réparation des dommages de la nature de ceux dont sont responsables les constructeurs au sens de l'article 1792-1 du Code civil (...)"*

La loi Spinetta entraîne la prise en compte de tous les dommages ou malfaçons de nature décennale pouvant survenir dans le secteur de la construction. Elle rend également obligatoire la souscription d'un contrat d'assurance spécifique :

- Pour le maître d'oeuvre et le professionnel constructeur, l'assurance souscrite doit couvrir la garantie décennale (délai de 10 ans à compter de la réception des travaux)
- Pour tout maître d'ouvrage, la souscription à une assurance dommage-ouvrage est obligatoire

Les versions aujourd'hui en vigueur des articles 1792 et 2270 du Code Civil sont les suivantes :

- **Article 1792 :** *"Tout constructeur d'un ouvrage est responsable de plein droit, envers le maître ou l'acquéreur de l'ouvrage, des dommages, même résultant d'un vice du sol, qui compromettent la solidité de l'ouvrage ou qui, l'affectant dans l'un de ses éléments constitutifs ou l'un de ses éléments d'équipement, le rendent impropre à sa destination. Une telle responsabilité n'a point lieu si le constructeur prouve que les dommages proviennent d'une cause étrangère."*
- **Article 2270 :** *"Les architectes, entrepreneurs et autres personnes liées au maître de l'ouvrage par un contrat de louage d'ouvrage sont déchargés de la garantie des ouvrages qu'ils ont faits ou dirigés après dix ans s'il s'agit de gros ouvrages, après deux ans pour les menus ouvrages."*

Ces lois ont permis de définir les obligations d'assurance ainsi que les types de dommages auxquels sont exposés l'obligation d'assurance et donc de dédommagement des sinistrés. Cependant, le traitement des sinistres est parfois long car plusieurs assureurs peuvent être engagés, et il convient de déterminer les causes exactes du sinistre pour établir la responsabilité des assurés. Afin de faciliter l'instruction des sinistres, et ainsi d'accélérer le paiement de la victime, **la Convention de Règlement de l'Assurance Construction (CRAC)** a été mise en place en 1983. Cet agrément signé par les assureurs intervenant sur le marché français de l'assurance construction, intervient entre des assureurs dommage-ouvrage (DO) et/ou responsabilité civile décennale (RCD). L'assureur DO, signataire de la convention, est obligé de régler, dans un délai de trois mois, les frais du sinistre. Celui-ci pourra ensuite exercer un recours auprès de l'assureur RCD afin de partager les règlements du sinistre. Ces assureurs ont alors un délai de 2 mois pour se mettre d'accord sur le partage de la responsabilité, à défaut un partage provisoire de la charge sera mis en place. Cette convention permet de réduire les coûts externes, tels que les frais et les coûts de gestion, en imposant une expertise unique.

Enfin, **la réglementation pour la gestion de l'assurance construction** a évolué avec les différentes réformes. Dans un premier temps, la loi Spinetta avait défini un système en semi-répartition pour gérer les garanties obligatoires liées à la responsabilité civile décennale, ainsi les primes perçues au cours d'une année (proportionnelles à l'activité exercée dans l'année) constituaient les provisions pour le règlement des sinistres survenant cette même année. Cette gestion était donc sensible à l'activité du secteur du bâtiment et peu protectrice du maître d'ouvrage. L'arrêté du 27 décembre 1982 instaure alors un système en capitalisation pour la gestion de ces garanties à partir du 1^{er} janvier 1983. Désormais, les primes perçues au cours de l'année constituent des provisions pour le règlement des sinistres afférents aux chantiers ouverts cette même année. Un fonds de compensation des risques de l'Assurance Construction avait été instauré en 1982 afin de financer le passage entre la gestion en répartition et la gestion en capitalisation. Financé par une contribution sur les contrats d'assurance construction, il était chargé de rembourser les assureurs de sinistres survenus au titre de chantiers ouverts avant le 1^{er} janvier 1983. Le financement de ce fonds de garantie s'est achevé en 2005.

1.1.3 Les différents types d'assurance

Le marché de l'assurance Construction présente deux types de polices différentes :

Police de chantier : il s'agit d'un contrat couvrant la réalisation d'un chantier précis clairement identifié, ce contrat est donc limité à la durée du chantier. Afin de tarifier ces polices, il est nécessaire de disposer des informations et caractéristiques précises sur le chantier. Les polices de chantier sont les assurances Dommage-Ouvrage (DO), Tous Risques Chantier (TRC), complémentaire RCD, assurance RC maître d'ouvrage, assurance constructeur non réalisateur...

Police d'activité : il s'agit d'un contrat couvrant un assuré pour l'ensemble de son activité professionnelle. A l'inverse de la police de chantier, la police d'activité couvre l'ensemble des chantiers ouverts pendant la période de souscription. La tarification est faite en fonction des caractéristiques de l'entreprise et de l'activité (chiffre d'affaires, type d'activité, effectif...). Les contrats d'assurance étant des polices d'activité sont, par exemple, la Responsabilité Civile Décennale (RCD), Responsabilité Civile Bâtiment (RC Bât)...

Ce mémoire traitera seulement des contrats d'assurances Responsabilité Civile Décennale (RCD) et Responsabilité Civile Bâtiment (RC Bât), qui sont des polices d'activités. La section suivante présentera plus en détail ces deux garanties.

1.1.4 Les différentes garanties

Les polices d'activités qui seront traitées dans ce mémoire, sont présentées ci-dessous :

La garantie Responsabilité Civile Décennale (RCD) : Tout constructeur a l'obligation de souscrire une assurance responsabilité civile décennale. Cette garantie décennale est obligatoire pour tous les travaux de construction, d'extension et de rénovation sur la structure d'un bâtiment.

Elle doit être souscrite avant le démarrage des travaux, et s'applique uniquement aux travaux ayant fait l'objet d'une ouverture de chantier pendant la période de validité du contrat. Seuls les travaux déclarés dans le contrat d'assurance du constructeur seront couverts.

Elle couvre les dommages survenus, à partir du lendemain de la réception des travaux par le maître d'ouvrage et ce pendant 10 ans. Le constructeur est responsable des dommages qui peuvent survenir sur la construction pendant ces 10 années, et il engage sa responsabilité à l'égard du maître d'ouvrage, auquel il est lié par un contrat de louage d'ouvrage. L'assurance couvre les malfaçons

qui n'étaient pas décelables lors de la réception des travaux, des désordres qui affectent la solidité de l'ouvrage et le rendent inhabitable ou impropre à l'usage auquel il est destiné.

Enfin, cette assurance de responsabilité décennale n'est pas affectée par le dépôt de bilan de l'assuré. En effet, lorsque des malfaçons sont constatées dans les 10 ans de garantie, mais que le constructeur a déposé le bilan, l'assureur de ce dernier est toujours dans l'obligation de couvrir ces dommages. Les primes ont été payées en amont du dépôt de bilan, à la date d'ouverture du chantier, et de ce fait, le chantier est toujours couvert. De manière similaire, si le constructeur change d'assureur pendant la période de garantie de 10 ans, c'est l'assureur avec lequel le constructeur avait un contrat à la date d'ouverture du chantier qui reste garant des dommages, jusqu'à la fin de la période décennale.

La garantie Responsabilité Civile Bâtiment (RC Bât) : En tant que professionnel du bâtiment, exerçant une activité réglementée, l'assurance responsabilité civile professionnelle est obligatoire. Se définissant comme une obligation légale de réparation du dommage causé à autrui, cette garantie permet de couvrir tous les dommages que l'activité de l'artisan pourrait causer à n'importe quel tiers (client, fournisseur, salarié ou un passant) ; et quel que soit le responsable du dommage (le chef d'entreprise, un salarié, un sous-traitant, ou une machine du chantier) ; pour tous les types de dommages (corporels, matériels ou immatériels).

Il s'agit d'une garantie annuelle qui couvre tous les préjudices intervenus pendant la durée du chantier lorsque la preuve du préjudice subi, l'explication du fait dommageable et le lien de causalité entre le préjudice et le fait sont explicités, et mettant en cause l'assuré.

1.1.5 Les dates de l'assurance construction décennale

L'assurance décennale présente une période de garantie de 10 ans à partir de la réception des travaux. Bien que la période de garantie soit une décennie, le versement des primes a lieu au moment de l'ouverture du chantier. Les différentes dates sont résumées dans le graphique ci-dessous :

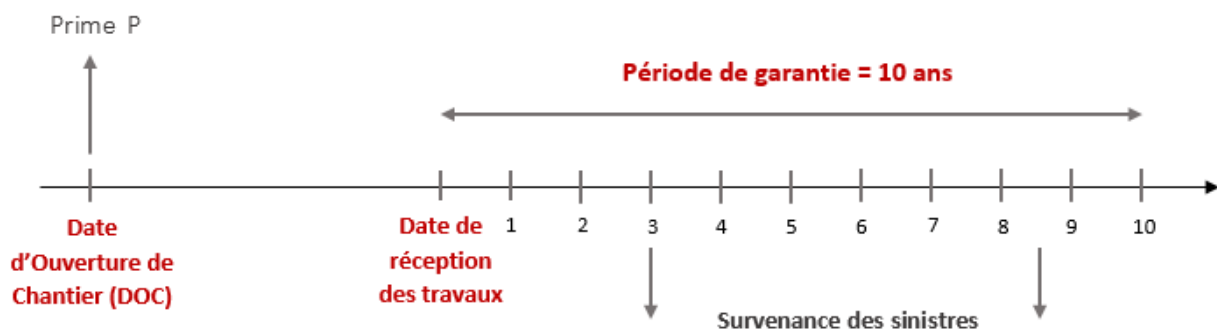


FIGURE 1.2 – Dates importantes de l'assurance décennale

L'assurance décennale se développe donc sur plusieurs dimensions qui seront considérées dans la modélisation :

- L'année de la date réglementaire d'ouverture de chantier,
- L'année de survenance du sinistre,
- Le délai de survenance du sinistre, étant la différence entre la date de survenance du sinistre et celle d'ouverture de chantier

1.2 Marché de l'assurance construction en France

Cette partie apporte des précisions sur le contexte général du marché de l'assurance construction en France. Les premiers éléments apportés proviennent de l'étude de la Fédération des Assurances parue en décembre 2021, s'intitulant : "Le marché de l'assurance construction en France en 2020".

L'assurance construction est une assurance de dommages aux biens et responsabilité, ce sont les contrats destinés à couvrir les biens de l'assuré, ainsi que la responsabilité de celui-ci envers les tierces victimes des dommages qu'il a causé. Cette famille se compose de différents types d'assurances comme l'automobile ou la responsabilité civile générale... Le graphique suivant permet de mettre en évidence le poids des cotisations de l'assurance construction dans l'ensemble des assurances de dommages aux biens et responsabilité en 2020 :

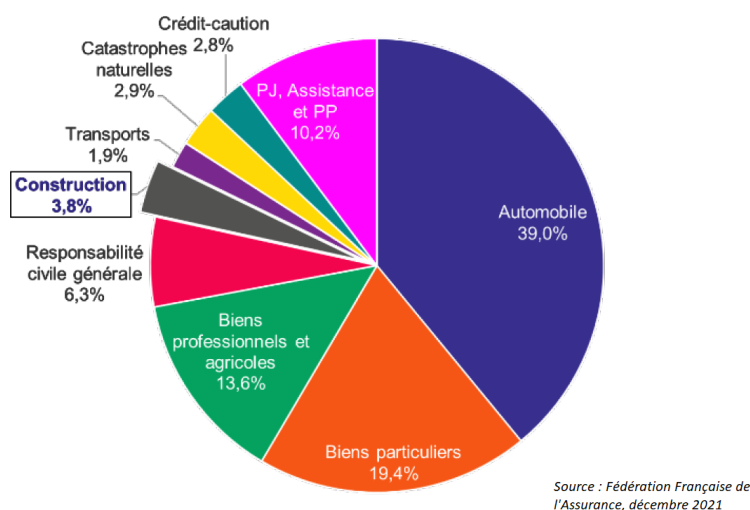


FIGURE 1.3 – Des cotisations de l'assurance construction en faible proportion

L'assurance construction pèse faiblement dans les cotisations des assurances de dommages et responsabilité. Destinés au professionnel et à un secteur d'activité précis, avec 3,8 % des cotisations, ce type d'assurance est bien loin derrière l'automobile (39 %) ou l'assurance des biens des particuliers (19,4 %).

L'année 2020 étant une année atypique avec une activité moindre, le niveau des cotisations a baissé de 3,3 % par rapport à l'année 2019, correspondant à une baisse de 9,3% de cotisations en Dommages-Ouvrages et une baisse de 0,8 % en Responsabilité Civile Décennale.

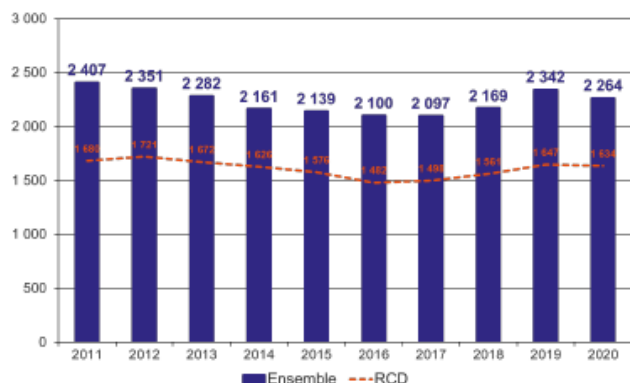


FIGURE 1.4 – Cotisations de l'assurance construction

Une baisse dans le niveau des cotisations est observé entre 2011 et 2017, le marché repart à la hausse en 2018 mais cet élan est arrêté par la crise sanitaire et une nouvelle baisse des cotisations est constatée en 2020. Ainsi le niveau des cotisations de l'assurance construction en 2020 reste inférieur de près de 6 % par rapport à celui de 2011. En 2021, l'activité reprend et le marché de l'assurance de dommages et de responsabilité bénéficie de la reprise économique et affiche une croissance de +3,8 % pour atteindre 62,4 milliards d'euros de cotisations.

L'année 2020 se traduit par une baisse du nombre de cotisations mais aussi par une forte chute de l'activité. En effet, une diminution de 15,3% en volume de l'activité en constatée, elle provient à la fois du secteur du logement neuf (-23,3 %), du secteur du non résidentiel neuf (-21,1 %) et, dans une moindre mesure, de l'activité en amélioration-entretien (-9,0 %). Le nombre de mises en chantier a également diminué (-7,8 %) pour tous les types de logements.

Avec la crise sanitaire et l'incapacité de réaliser certains chantier, une crise économique apparaît sur ce marché. Une telle crise peut faire augmenter la sinistralité car :

- Il existe une chute du nombre de mises en chantier et donc une augmentation de la concurrence. Celle-ci entraîne la baisse des prix et une concentration sur obtention de nouveaux marchés, et non pas sur le service après-vente
- On constate une augmentation de la sous-traitance, celle-ci permet de limiter les coûts pour l'entreprise engagée dans le chantier. Cependant les sous-traitants sont souvent peu ou pas assurés, et en cas de sinistre c'est l'assureur de l'entreprise principale qui se retrouve seul à régler le sinistre sans aucun recours possible
- Enfin les rapports sont plus conflictuels entre les différents intervenants à l'acte de construire. On constate généralement une hausse des dossiers contentieux au détriment d'une solution amiable (en moyenne les dossiers contentieux ont un coût 7,5 fois supérieur à la moyenne car il y a une multiplicité des réunions d'expertise, des intervenants et des délais)

La sinistralité peut donc augmenter en nombre mais également en montant. En effet, les innovations technologiques, peuvent être source de sinistralité quand elles sont insuffisamment expérimentées ou mal maîtrisées, cependant elles entraînent également des coûts supérieurs car elles nécessitent des outils plus développés et des connaissances supérieures. Enfin la multiplication des normes réglementaires peut être également une source de l'augmentation des coûts des sinistres.

Dans le contexte actuel, l'activité dans le secteur de la construction a repris, cependant elle est confrontée à une forte inflation qui touche tous les secteurs d'activités et de nombreux produits. La section suivante détaillera davantage l'inflation à laquelle est confrontée ce secteur.

1.3 Inflation dans l'assurance construction

Les coûts des travaux peuvent se décomposer selon quatre points : le coût des matériaux, celui de la main d'œuvre, les frais généraux et la marge. Les évolutions dans ces catégories sont différentes entre les années 2021 et 2022.

- Le prix des matériaux a fortement augmenté pendant les années 2021 et 2022 suite à la reprise de l'activité et à la guerre en Ukraine mais selon des proportions différentes en fonction du type de matériaux. Ce point sera détaillé d'avantage dans la section suivante.
- Le coût de la main d'œuvre est quand à lui stable en 2021 par rapport à 2020, mais avec le contexte d'inflation en 2022 une tension apparaît sur les salaires et ils pourraient être amenés à augmenter.
- Les frais généraux (coûts engagés par une entreprise dans le cadre de ses activités quotidiennes), les impôts et la TVA sont stables en 2021 et 2022.
- Enfin, pour son activité une entreprise a besoin de se dégager une certaine marge pour le futur mais aussi pour des besoins divers de reconstitutions post crise, avec le remboursement des Prêts Garantis par l'État (PGE) par exemple.

A la sortie de la crise sanitaire et à la suite de la guerre en Ukraine, une forte inflation apparaît dans le monde. Elle concerne tous les secteurs d'activités et donc celui de la Construction mais

également celui de l'assurance. En effet une hausse des coûts dans le bâtiment se répercute sur le coût des sinistres, et donc sur le montant de la prime.

Comme vu précédemment, le prix des travaux se décomposent selon plusieurs catégories. Un indice permettant de regrouper toutes ces catégories et de refléter cette évolution globale est **l'indice BT01**. Calculé par l'INSEE (Institut National de la Statistique et des Études Économiques) et publié mensuellement, il a pour objectif de refléter au mieux l'évolution des coûts sur le marché de la construction. Il prend ainsi en compte les salaires et charges, le coût des matériaux, les frais divers, le matériel, les frais de transport et d'énergie. Son cours est représenté ci-dessous :

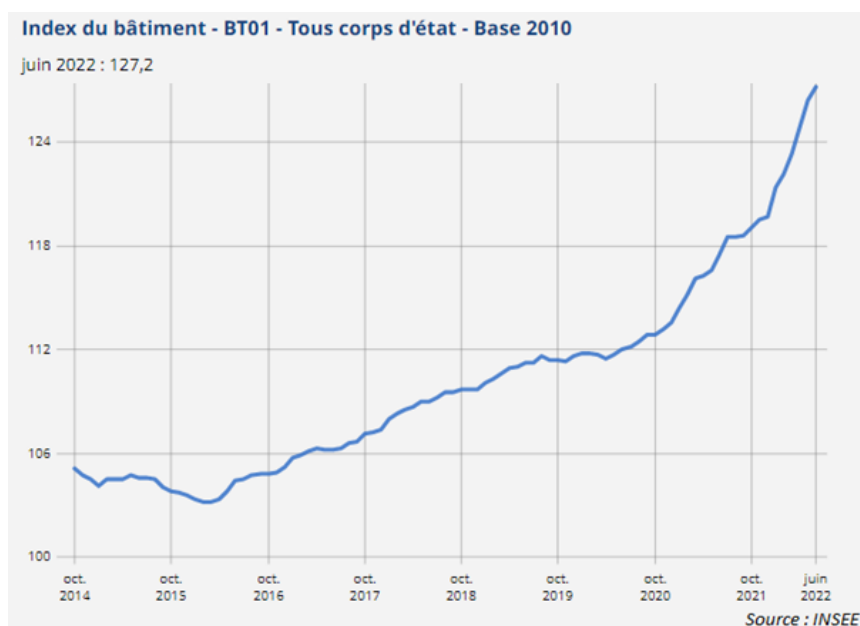


FIGURE 1.5 – Impact de la crise sanitaire sur l'indice BT01

L'indice est en hausse depuis 2016, cependant cette hausse s'est accentuée à partir de la fin d'année 2020. Enfin le début de la guerre en Ukraine a de nouveau entraîné une forte augmentation. Toutes ces évolutions peuvent être constatées sur le graphique de l'inflation annuelle du BT01.

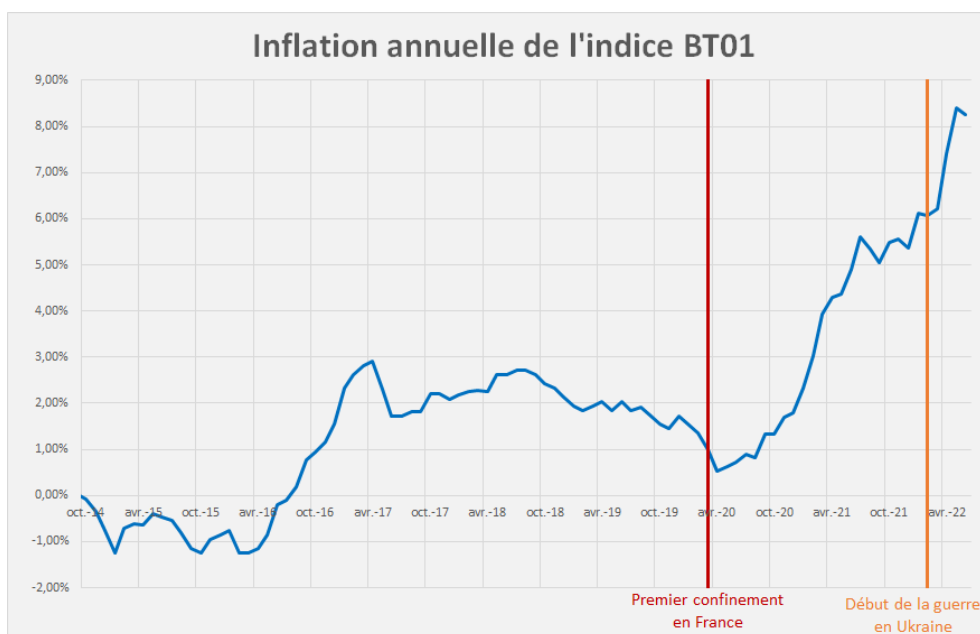


FIGURE 1.6 – Impact du contexte géopolitique sur l'inflation du BT01

Les tendances constatées sur le graphique des valeurs de l'indice BT01 se confirme sur celui de l'inflation. Elle est positive depuis juillet 2016 mais son cours augmente fortement à partir d'avril 2020. La guerre en Ukraine permet une reprise forte avec une inflation de 6,12 % en janvier 2022 et qui vaut 8,41 % en mai.

Cet indice donne une vision globale du marché de la construction, cependant toutes les branches de la construction ne sont pas impactées de la même manière. De même pour les prix des matières premières et des matériaux qui sont le plus durement touchés par le contexte économique actuel, où il existe des disparités entre les matériaux.

Afin de mettre en évidence ces différences entre les matériaux, le tableau suivant résume l'inflation constatée, à plusieurs échéances (1 mois, 3 mois, 6 mois, 1 an et 3 ans) pour les principaux matériaux utilisés en construction.

	Inflation				
Matériau	1 mois	3 mois	6 mois	1 an	3 ans
Acier	-4,79 %	20,98 %	31,77 %	56,51 %	133,89 %
Aluminium	-1,34 %	-1,16 %	23,42 %	50,84 %	91,17 %
Charpentes, menuiseries de bâtiment	2,24 %	5,60 %	9,08 %	15,78 %	24,51 %
Ciment, chaux et plâtre	4,42 %	8,70 %	21,92 %	27,43 %	29,17 %
Cuivre	-3,04 %	-1,51 %	7,39 %	14,98 %	58,11 %
Peinture Bâtiment	1,70 %	4,32 %	7,54 %	13,68 %	13,99 %
Plomb, zinc et étain	-3,08 %	2,46 %	14,32 %	32,96 %	59,69 %
PVC	-1,06 %	4,31 %	15,74 %	34,12 %	138,54 %
Tuiles et briques (terre cuite)	0,94 %	8,42 %	22,57 %	23,28 %	23,87 %

Le tableau précédent permet de mettre en évidence des différences d'inflation entre les matériaux utilisés en construction, ainsi le PVC et l'acier sont ceux ayant la plus forte augmentation de prix en 3 ans, tandis que le prix des peintures rencontre la plus faible augmentation. Enfin, pour plusieurs matériaux, l'inflation durant les trois dernier mois est négative, ainsi le climat de forte inflation semble ralentir, mais le contexte géopolitique, avec la guerre en Ukraine pourrait faire à nouveau remonter les prix.

Enfin une autre raison de la forte évolution de l'indice BT01, est l'augmentation des frais de transport et notamment du carburant. Après avoir baissé durant le premier confinement en France, le prix des carburants repart à la hausse avec la reprise de l'activité mondiale. Enfin comme pour les matériaux, la guerre en Ukraine entraîne à nouveau une forte augmentation du fait des sanctions à l'encontre de la Russie.

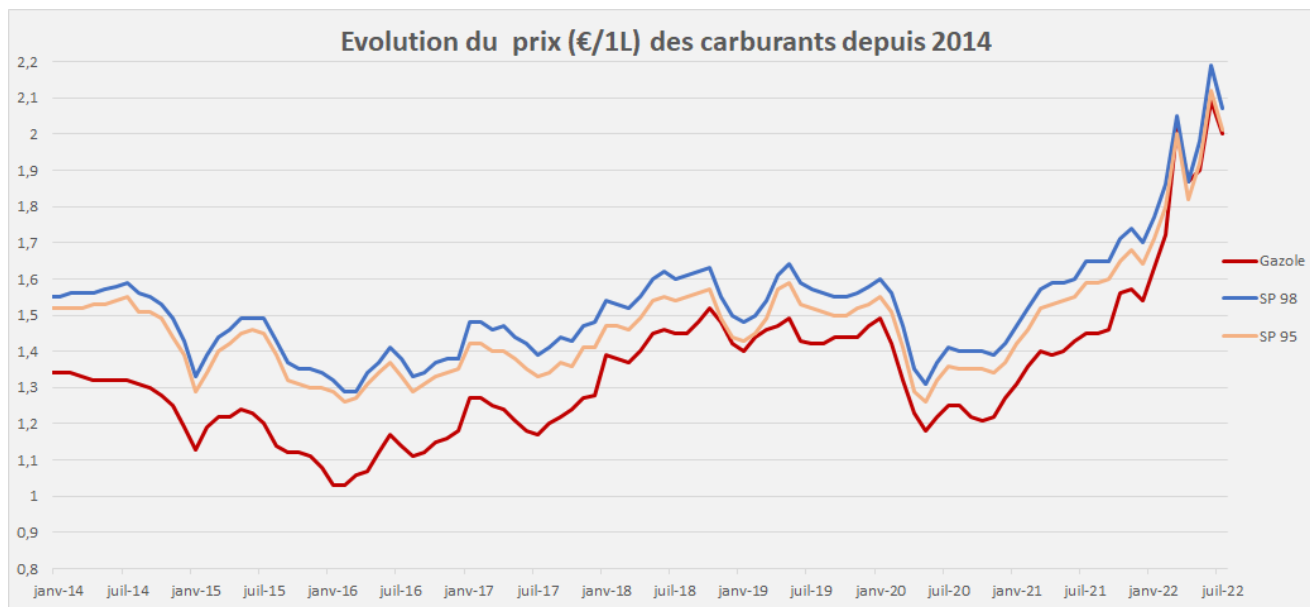


FIGURE 1.7 – Le prix des carburants, une source supplémentaire de dépenses pour les entreprises

Ce graphique sur les prix des carburants permet de constater une forte hausse depuis la sortie des premiers confinement mondiaux en juillet 2020. Celle-ci s'est accentuée depuis le début de la guerre en Ukraine, fin février 2022. Comme pour les matériaux, une baisse des prix est constatée durant le dernier mois mais cette tendance à la baisse pourrait ne pas durer.

Cette partie a permis de définir le domaine de la construction et de l'assurance construction : sa définition, ses spécificités, son évolution et le contexte actuel. Désormais, le portefeuille des contrats de Generali France sera étudié afin de présenter les données utilisées dans ce mémoire. Ensuite les travaux de création et de préparation des bases pour la modélisation seront réalisés.

Présentation du portefeuille

Generali s'implante en France en 1832 lorsque le premier mandat d'agent a été délivré. La succursale française a été ouverte à Paris en 1848, et deviendra en 1973 la société Generali France. Elle se développe au fil des années avec des acquisitions et des fusions de filiales, pour devenir aujourd'hui l'un des acteurs principal du marché de l'assurance. L'assurance construction représente une faible part de l'activité du groupe sachant que les garanties dommages aux biens et responsabilités des entreprises représentent 3% du chiffre d'affaire de Generali France en 2021. De plus, pour la responsabilité décennale, le montant des cotisations s'élève à plus de 30 millions ce qui ne permet pas à Generali de figurer parmi les 9 sociétés générant le plus de cotisations sur ce marché en France.

2.1 Présentation de l'entreprise et du portefeuille

Les polices d'activités proposées par le groupe sont la responsabilité civile décennale et la responsabilité civile Bâtiment. Ces contrats à destination des artisans et entreprises du BTP sont vendus ensemble de façon indissociable par Generali. En 2022, environ 10 000 contrats sont en vigueur pour plus de 30 millions d'euros de primes pour chacune des garanties.

Les artisans et entreprises du BTP souscrivant à ses garanties sont classés en deux catégories, selon leurs caractéristiques (chiffres d'affaires, effectif à la souscription) :

- **Entreprise** : lorsque l'effectif est supérieur à 20 ETP (Équivalent Temps Plein),
- **ProPE (Professionnels - Petite Entreprise)** : lorsque le chiffre d'affaires est inférieur à 2 millions d'euros.

Les caractéristiques de l'assuré sont également utilisée pour définir la typologie des primes que l'assuré peut payer. Celles-ci peuvent être

- **Forfaitaires** lorsque l'effectif est inférieur ou égal à 5 ETP, ou lorsque le chiffre d'affaires ne dépasse pas 600 000 euros. Dans ce cas, le paiement des primes est annuel et le montant est déterminé à partir de la valeur de base qui est modulée en fonction de l'activité, de l'effectif, de l'ancienneté et avec une éventuelle remise à niveau en fonction de la sinistralité observée et de l'inflation. Le montant de la prime est fixé en fonction des caractéristiques connues au moment du calcul.
- **Révisables**, elles dépendent du chiffres d'affaires et de l'activité, le paiement s'effectue sur deux ans, et tout comme les primes forfaitaires, elles peuvent être ajustées en fonction de la sinistralité observée. A chaque fin de période, l'assuré doit déclarer le montant réel du chiffre d'affaires et une régularisation de la prime peut alors être réalisée.

La répartition du portefeuille, en 2022, selon ces typologies de primes et d'assurés est la suivante :

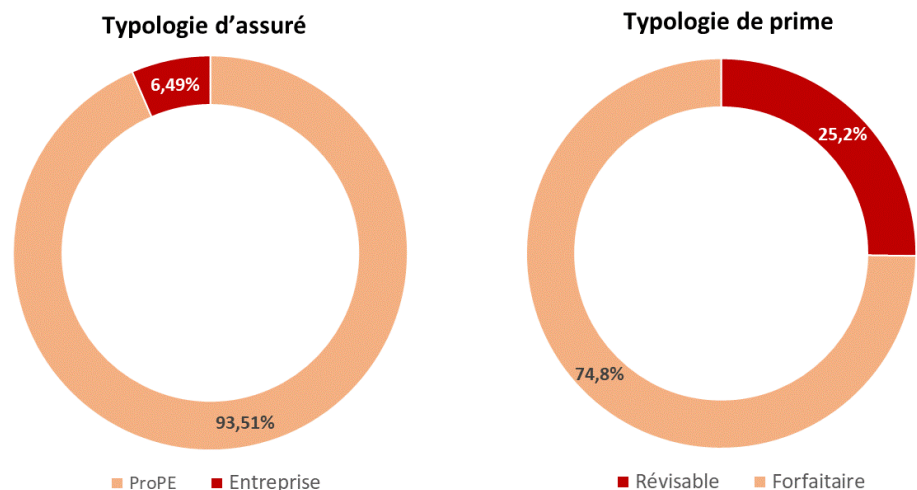


FIGURE 2.1 – Répartition des primes et des assurés selon leur typologie

Les assurés de type ProPE sont majoritairement présents dans le portefeuille (93,51%) tandis que les primes sont pour près de 75 % forfaitaire. En 2022, la répartition entre le type de contrat et la typologie d'assurés est détaillée dans les tableaux ci-dessous :

Type d'assuré	ProPE	Entreprise
Forfaitaire	79,61%	5,40 %
Révisable	20,39%	94,60 %
Total	100%	100 %

Type de prime	ProPE	Entreprise	Total
Forfaitaire	99,53%	0,47 %	100%
Révisable	24,34 %	75,66%	100%

On retrouve bien une cohérence entre les typologies d'assurés et de primes. Les ProPE ont majoritairement des primes forfaitaires (79,61 %), car elles sont destinées à des assurés avec un chiffre d'affaire et un effectif plus faible. Tandis que les primes révisables sont majoritairement payées par les entreprises (75,66%).

2.2 Évolution du portefeuille

Le portefeuille de contrats, pour les garanties RCD et RC Bât, a fortement évolué pendant les dernières années. Ces évolutions sont dues à des modifications de gestion et des innovations au sein du groupe, afin de maintenir un niveau de solvabilité correct.

Le nombre de contrats en vigueur est en légère hausse entre 2007 et 2010, cependant le nombre de contrats chute à partir de 2012 jusqu'en 2017, pour ensuite augmenter à nouveau. Aujourd'hui près de 10 000 contrats sont présents dans le portefeuille contre environ 19 000 en 2010. Le graphique ci-dessous représente l'évolution du nombre de contrats en vigueur pendant les 16 dernières années.

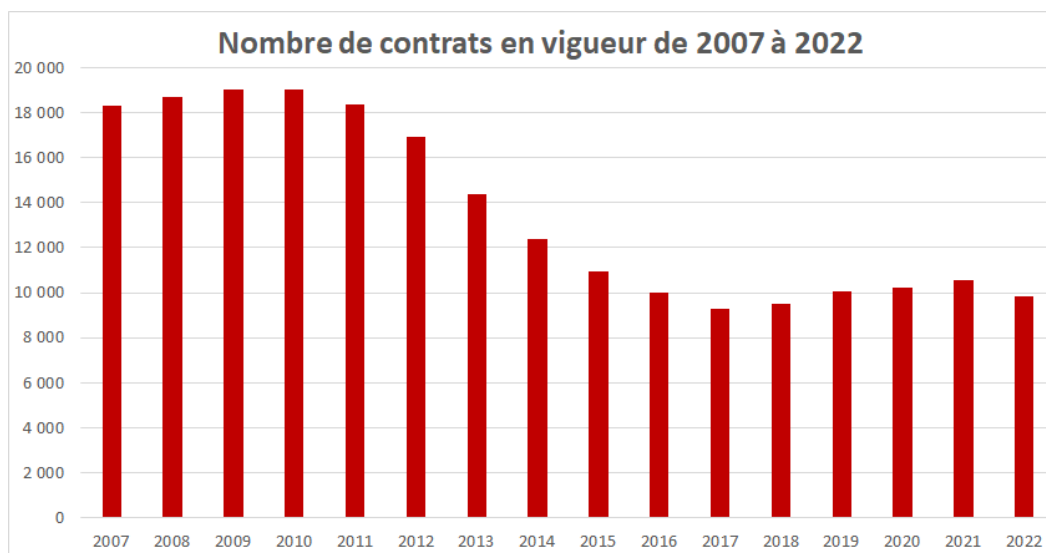


FIGURE 2.2 – Impact des modification de gestion sur le portefeuille

Cette forte chute s'explique par une vague de résiliations spéciales en 2013 suite à **la mise sous surveillance du portefeuille**. Cette mise sous surveillance s'explique par une volonté de contrôler la solvabilité du produit. En effet les ratios de solvabilité était supérieur à 1 entre 2006 et 2013, et de nombreux contrats présentaient une forte sinistralité. Une réduction de près de 50 % du portefeuille est observée entre cette date de mise sous surveillance et aujourd'hui.

En 2017, une nouvelle politique de souscription ProPE est mise en place, afin de privilégier ce type d'assurés qui gère de plus petits chantiers et qui génère des sinistres avec des délais de survenance plus faible (54% des sinistres ProPE ont des délais de survenance entre 0 et 3 ans, tandis que pour les Entreprises, ce nombre est de seulement 41 %) ¹. Cette évolution structurelle du portefeuille permet à la proportion d'assurés *ProPE* dans le portefeuille de croître entre 2007 et 2022, à l'inverse celle des assurés *Entreprise* diminue :

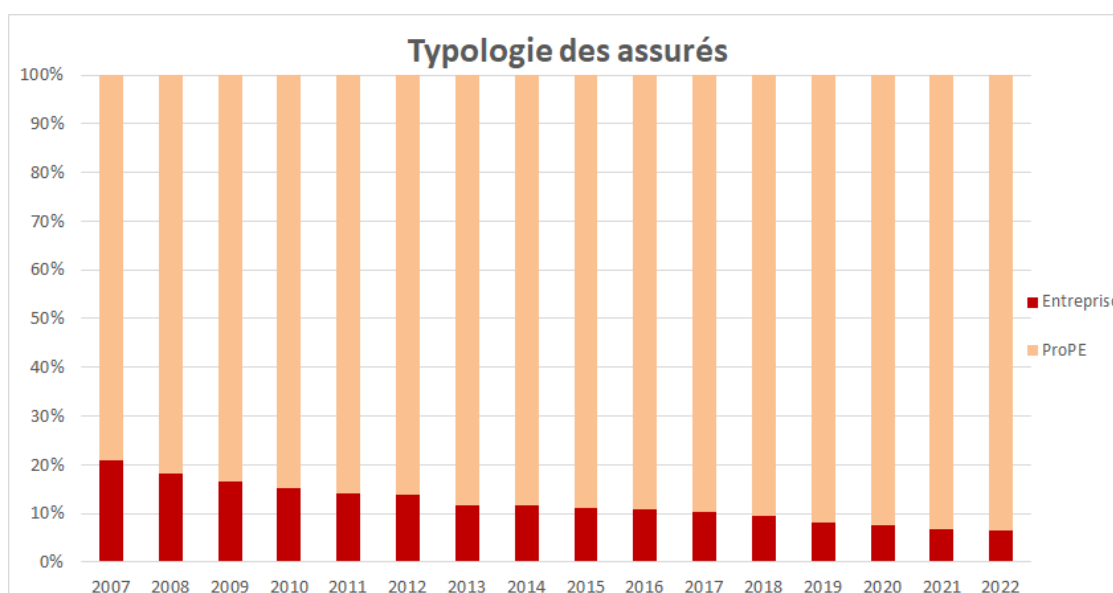


FIGURE 2.3 – Réduction du nombre d'assurés *Entreprise* dans le portefeuille

Cette évolution est en accord avec celle de la typologie de contrat. Une baisse de la proportion de contrat révisable est constatée, et les Entreprises détiennent majoritairement des contrats

1. Une répartition des délais de survenance en fonction de l'assuré est présentée en annexe A

révisables (94,6 % des assurés de type *Entreprise* possèdent un contrat révisable) :

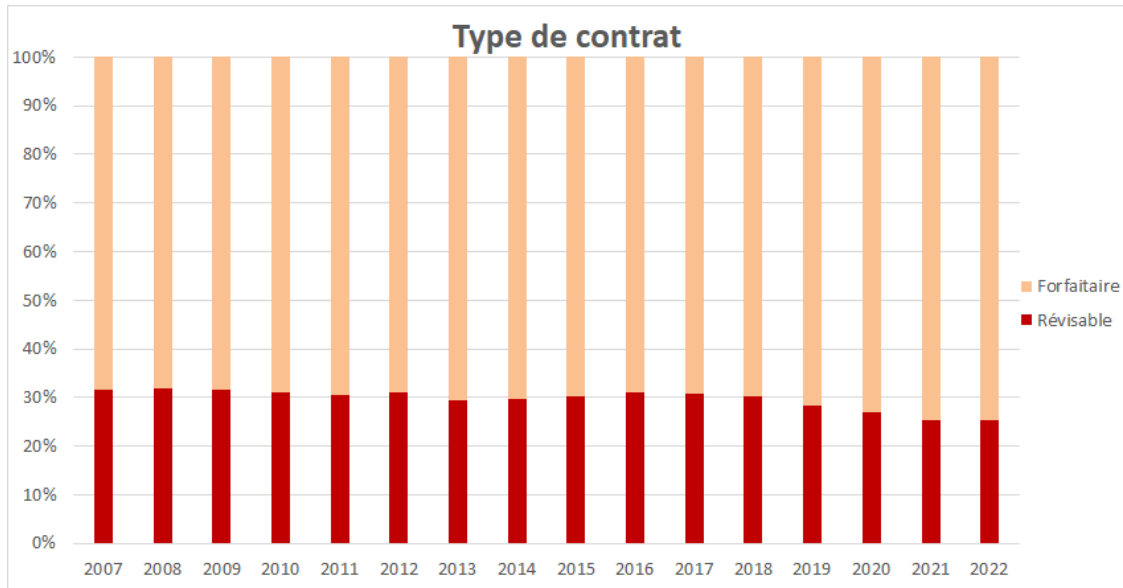


FIGURE 2.4 – Une légère baisse de la proportion de contrats *Révisable* de 2007 à 2022

Des modifications dans la gestion des sinistres et le traitement de la sinistralité ont également été opérées.

En 2019, le **multi-garantie** a été mis en place. Il est désormais possible d’associer plusieurs garanties à un seul sinistre. Il y a donc une ouverture simultanée de plusieurs dossiers sinistres avec le même numéro mais associés à des garanties différentes. Lorsque la typologie de la garantie est définie, les contrats non associés sont fermés et considérés sans suite. Cet outil peut entraîner une augmentation du nombre de sinistres et également du nombre de sans suite.

En 2021, **branchement à l’ECRAC**, qui est un dispositif de déclaration des sinistres. Les sinistres peuvent être déclarés plus facilement et rapidement. On pourrait constater une augmentation du nombre de sinistres déclarés, cependant la mise en place est récente (1 an) ; il est trop tôt pour constater un réel impact.

Toutes ces évolutions au sein du portefeuille ont fortement modifié son profil de risque, de même pour la sinistralité avec les modifications de gestion. Ainsi les méthodes de provisionnement utilisant les triangles de charges ne sont plus adaptées car il n’y a plus de stabilité dans le temps. De plus, dans le contexte inflationniste actuel, l’hypothèse de stabilité temporelle de l’inflation permettant d’appliquer les méthodes de développement des triangles n’est pas respectée.

Ce mémoire traitera de la modélisation de la sinistralité à l’aide de l’approche Fréquence-Coût, utilisant les modèles linéaires généralisés et des modèles de Machine Learning. Les parties suivantes permettront de présenter tous les travaux de préparation de la modélisation puis de la mise en place de celle-ci.

Préparation de la base Sinistre

Les informations sur la sinistralité sont multiples, et varient selon la garantie exercée. Il est nécessaire de regrouper toutes ces données, en veillant à avoir le même format pour toutes les modalités d'une variable, et en traitant les valeurs aberrantes ou manquantes.

L'objectif de cette partie est de détailler le processus ayant permis la création d'une base de données regroupant les informations sur la sinistralité. Dans un premier temps, les données initiales seront présentées, puis la sélection des variables sera abordée. Afin de tenir compte de l'inflation et de l'évolution des coûts des sinistres, ils seront revalorisés, et la technique utilisée sera expliquée dans la troisième section. Enfin les seuils de gravité permettant la segmentation des charges des sinistres seront déterminés, pour finalement créer la base de données finale.

3.1 Présentation des données initiales

Au sein du groupe, une table regroupe toutes les données sur la sinistralité. Il convient de sélectionner seulement celles se rapportant aux deux garanties étudiées, et nous nous placerons au 31 décembre 2021, ainsi seront étudiés tous les sinistres survenus et enregistrés avant cette date.

Le tableau ci-dessous résume la répartition des sinistres selon leur état et selon leur charge :

Etat du dossier	Ensemble des sinistres		Charges négatives ou nulles		Charges positives	
	Nombre	Taux	Nombre	Taux	Nombre	Taux
En cours	11 588	3.98%	219	0.23%	11 369	5,83%
Suspendu	0	0%	0	0	0	0
Clôturé	135 335	46.48%	6 972	7.26%	128 363	65,78%
Sans suite	138 725	47.64%	85 940	89.49%	52 785	27,05%
Annulé	2 838	0.97%	2 837	2.95%	1	0,001%
Réouvert	2 697	0.93%	62	0.06%	2 635	1,35%
Total	291 183	100%	96 030	100%	195 153	100%

L'état des dossiers sinistre est un indicateur permettant de connaître l'état d'avancement dans le traitement du sinistre. Il peut prendre les valeurs suivantes :

- En cours, cela signifie que le sinistre est en cours de gestion ;
- Suspendu, la gestion du sinistre est suspendue ;
- Clôturé, le sinistre est clos avec des possibles règlements de frais ou d'indemnités ;

- Sans suite, le sinistre est classé sans suite, la distinction sera faite entre les sans suites ayant engendré un règlement et ceux sans aucun règlement. Ce dernier type de sinistre peut exister pour différentes raisons :
 - sinistre ouvert à tort dans une catégorie (lié à une erreur de saisie ou à une ouverture multiple dans un premier temps) ;
 - sinistre ne permettant pas de procéder à une indemnisation (absence de préjudice, ou préjudice n'entrant pas dans le cadre des garanties) ;
- Annulé ;
- Réouvert, le sinistre a été clôturé une première fois, mais son évolution a nécessité une réouverture du dossier, ainsi il est actuellement en cours de gestion.

Les montants des sinistres positifs sont compris entre 0 et 4,2 millions d'euros. Les sinistres ayant une charge supérieure à 2 millions d'euros sont très faibles et 90 % des sinistres ont une charge inférieure à 30 000 euros. Le graphique ci-dessous représente la distribution de ces derniers sinistres :

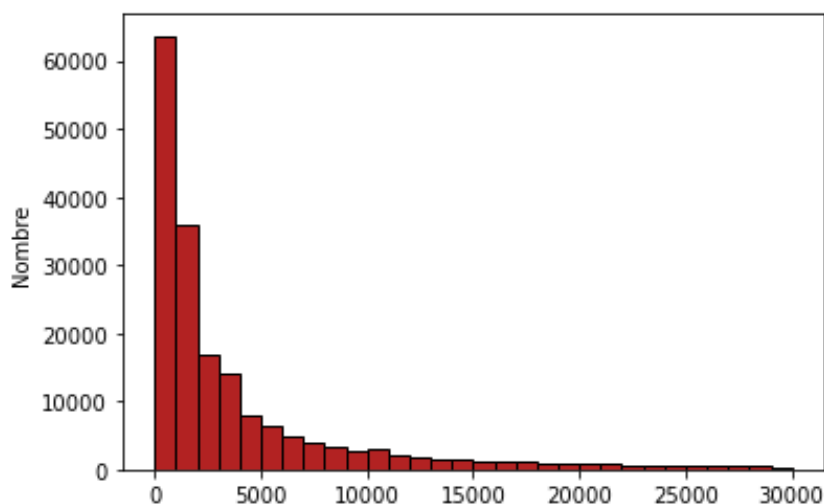


FIGURE 3.1 – Distribution des sinistres de charge inférieure à 30 000 euros

Une forte concentration de sinistres est présente entre 0 et 1 000 euros. Afin d'observer plus en détail la répartition des sinistres pour les charges inférieures à 3 000 euros (charges les plus représentées selon le graphique précédent), un zoom est effectué et représenté ci-dessous :

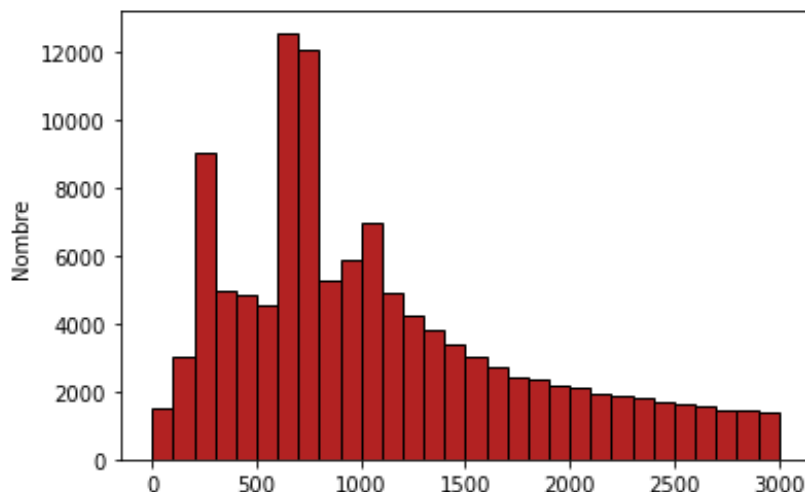


FIGURE 3.2 – Importance des coûts forfaitaire dans la distribution des sinistres

Ce graphique permet de constater qu'il existe un premier pic de sinistres de charges entre 200 et 300 euros mais également deux autres pics de charges entre 600 et 800 euros. Ces coûts correspondent à des frais d'ouverture de dossier ou d'expertise, ce sont des coûts forfaitaires fixes qui peuvent altérer les résultats obtenus lors de l'étude de la sinistralité ou lors de la modélisation. Ces coûts forfaitaires devront être retirés lors des travaux de modélisation de la sinistralité.

La répartition des sinistres est similaire pour les sinistres RCD et RC Bât, ainsi il ne semble pas nécessaire de traiter leur charge de manière différente. La section suivante permettra de vérifier cela, en définissant des seuils de gravité pour l'ensemble des sinistres et en considérant les sinistres RCD et RC Bât séparément.

Enfin, parmi les sinistres classés sans-suite, la distinction est faite entre ceux n'ayant entraîné aucune charge (pas de règlement, ni de provisions) et ceux de charge non nulle. Ils ne seront pas modélisés de la même manière par la suite, les parties suivantes permettront d'expliquer les différences qui seront opérées.

3.2 Sélection des données

Pour la réalisation future de la modélisation, il est nécessaire de créer une base de données comportant toutes les informations importantes et descriptives sur un sinistre. Pour cela, il faut sélectionner les données, les traiter pour enlever les valeurs manquantes ou aberrantes, permettant ainsi d'obtenir une base optimale.

3.2.1 Choix des variables

Différentes variables ont été sélectionnées pour créer la base de données sur la sinistralité. Une liste, ainsi qu'une présentation de celles-ci sont exposées ci-dessous :

Numéro de dossier sinistre : Un sinistre est identifié par un numéro de dossier sinistre. Ce numéro unique d'identification permet de distinguer les différents sinistres mais également de retrouver facilement les informations associées à ce sinistre. Chaque sinistre est associé au contrat auquel il se rapporte, ainsi le **numéro de police** est indiqué dans la base. Ce dernier permettra de faire le lien avec les éléments du contrat et les informations de l'assuré.

L'état des dossier sinistre : Cet indicateur permet de connaître l'état d'avancement dans le traitement du sinistre, les modalités de cette variable ont été détaillées dans la section précédente. Il sera utilisé dans la partie modélisation afin de sélectionner les sinistres pour les différentes bases.

La date de survenance correspond à la date à laquelle le sinistre est survenu. Elle est associée aux **dates de déclaration et d'enregistrement du sinistres**. Ce sont les dates auxquelles le sinistre est déclaré à l'assureur et enregistré au sein du groupe Generali. Ces dates permettent d'associer le sinistre à un exercice donné et à la période de garantie correspondante.

La garantie associée au sinistre permet de déterminer le type de contrat auquel le sinistre se rapporte. La variable présente les modalités : RC Bât ou RCD.

La charge du sinistre correspond à la somme des indemnités versées pour la réparation du sinistre, des frais de gestion des sinistres (frais d'expertise ou frais juridique, par exemple) et des provisions, à laquelle on retrace les recours encaissés.

Afin de compléter la base en ajoutant des informations supplémentaires liées aux évolutions de la gestion des sinistres, des variables indicatrices sont créées. L'une permet de déterminer si l'ouverture du dossier sinistre est réalisé par l'ECRAC. Pour cela, il faut déterminer le site et le code du gestionnaire ayant ouvert le sinistre. La deuxième variable créée indique s'il est possible d'ouvrir un dossier sinistre avec plusieurs garanties. Le multi garantie ayant été mis en place en 2019, la variable prendra la valeur 1, si la date d'enregistrement est postérieure au 1^{er} janvier 2019.

Dans le cas d'un sinistre RCD, un élément supplémentaire doit apparaître dans les données sur la sinistralité. Il s'agit de la date de **Déclaration Réglementaire d'Ouverture de Chantier (DROC)**. Celle-ci est essentielle pour déterminer si le contrat d'assurance est concerné par le sinistre (le contrat doit être signé avant cette date, selon les dispositions juridiques explicitées dans le premier chapitre).

Toutes ces variables seront utiles pour la modélisation, cependant pour pouvoir les utiliser les valeurs aberrantes ou manquantes doivent être corrigées. C'est l'objectif de la section suivante.

3.2.2 Traitement des données manquantes

Lors de la récupération des données sur la sinistralité, des données manquantes ou aberrantes ont pu être constatées sur la date de survenance du sinistre ou sur la date réglementaire d'ouverture de chantier. Lors de l'ouverture d'un dossier pour un sinistre, outre la date de survenance et d'ouverture de chantier, les dates de déclaration et d'enregistrement sont également notées. Elles serviront pour traiter les incohérences repérées sur la base.

La première incohérence constatée concerne des contrats avec une garantie décennale : la date de survenance du sinistre est antérieure à la date d'ouverture de chantier. Cependant les dates de déclaration et d'enregistrement du sinistre étaient cohérentes avec la réalisation du chantier (elles étaient bien postérieures à la DROC), ainsi les dates de survenance ont été forcées à la date de déclaration.

Plus généralement, plusieurs problèmes se sont posés sur les dates de survenance :

- **Valeur manquante** : pas d'informations disponibles sur cette date, pour ces contrats cette absence se retrouvait également sur les dates d'enregistrement ou de déclaration. Ne disposant pas d'informations permettant de déterminer la survenance des sinistres, il a été décidé de les supprimer.
- **Valeur aberrante à 1199** : l'année de déclaration de ce sinistre est 2000, nous supposons alors qu'il s'agit d'une erreur de frappe et la survenance est rehaussée à 1999.
- **Valeur entre 1900 et 1966** : Certains contrats présentaient une survenance avant 1966, cependant cela est incohérent avec la composition du portefeuille, ainsi elles sont rehaussées à la date de déclaration

Ce problème d'incohérence lié aux dates entre 1900 et 1966 fut également constaté sur les dates réglementaires d'ouverture de Chantier, ainsi, de manière équivalente, elles ont été rehaussées à la date de survenance.

3.3 Revalorisation des sinistres

Avant de segmenter les sinistres, il convient de revaloriser leur charge. L'objectif de la revalorisation est de déterminer le coût qu'aurait eu un sinistre s'il survenait aujourd'hui, cela revient à prendre en compte l'inflation perçue depuis la date de survenance d'un sinistre et jusqu'à aujourd'hui. Cette revalorisation permettra de comparer les sinistres entre eux. Pour la construction, il existe différents indices qui reflètent les évolutions des coûts de ce domaine :

- **Indice du coût de la construction par la FFB** : il s'agit d'un indice trimestriel établi par la Fédération Française du Bâtiment, basé sur le prix de revient d'un immeuble de type courant à Paris, et est calculé en base 1 au 1^{er} trimestre 1941. Il permet de tenir compte des évolutions de prix des éléments qui interviennent dans la construction d'un bâtiment (main d'œuvre, matériaux et fournitures, frais administratifs, taxes, prestations), en excluant la valeur du terrain.
- **Indice BT01** : il s'agit d'un indice calculé par l'INSEE et publié mensuellement, il a pour objectif de refléter au mieux l'évolution des coûts sur le marché de la construction. Il prend ainsi en compte les salaires et charges, le coût des matériaux, les frais divers, le matériel, les frais de transport et d'énergie. Les indices historiques sont en base 100 à janvier 1974, mais ont été modifiés depuis octobre 2014 et sont désormais en base 100 à janvier 2010. Afin de pouvoir utiliser simultanément des indices antérieurs et postérieurs à 2014, il convient de multiplier les indices postérieurs par un coefficient de raccordement, égal à 8,3802.
- **Indices des coûts de production dans la construction (ICPC)** : il s'agit d'indices permettant de suivre l'évolution des coûts de production dans le secteur de la construction, selon les types de travaux et les catégories de coûts (travail, matériaux, énergie, ...). Ils sont calculés sur l'ensemble du secteur de la construction (génie civil, bâtiments résidentiels ou non résidentiels...), leur usage est exclusivement statistique.
- **Indice du coût de la construction (ICC)** : fourni par l'INSEE, il permet de suivre l'évolution du prix des immeubles neufs à usage principal d'habitation. Calculé sur une base 100 au premier trimestre 1953, il est publié trimestriellement, et sert à la révision des loyers des baux commerciaux signés avant le 1^{er} septembre 2014.

Parmi ces indices, ceux pouvant être utilisés pour la revalorisation de sinistres sont **l'indice FFB** et **l'indice BT01**. En effet, ces deux indices reflètent les évolutions des coûts du marché de la construction, en prenant en compte aussi bien les coûts liés à la main d'œuvre que ceux liés aux matériaux. De plus, les autres indices ne sont pas destinés à être utilisés pour de la revalorisation.

Désormais, une comparaison de ces indices va être effectuée afin de déterminer celui qui sera retenu pour l'étape de revalorisation. Dans un premier temps, leurs évolutions depuis janvier 2000 sont considérées :

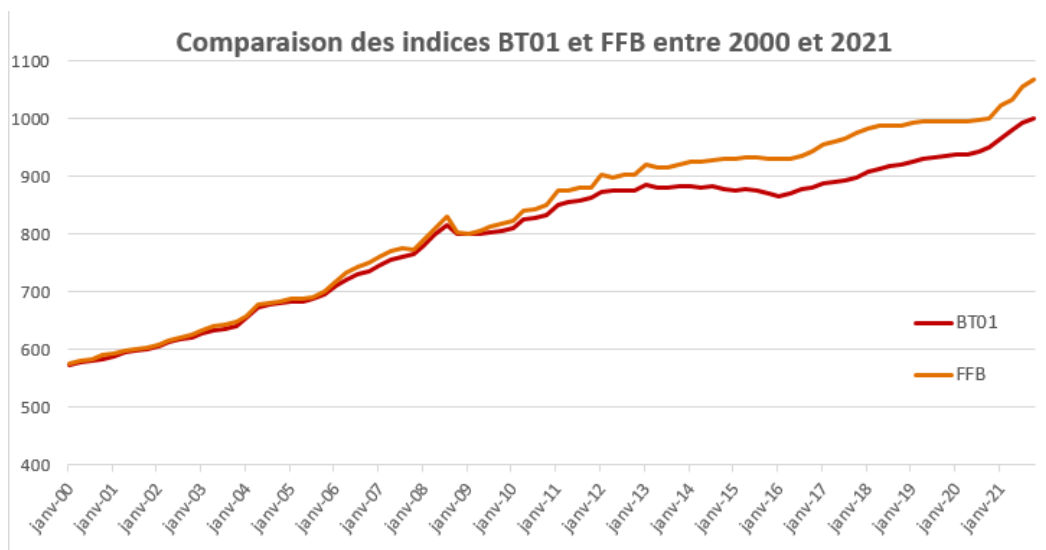


FIGURE 3.3 – Comparaison des indices BT01 et FFB entre 2000 et 2021

Les deux indices ont une tendance à la hausse et les mêmes changements de comportements sont constatés. Les valeurs des deux indices ont un faible écart entre 2000 et 2010, et celui-ci se creuse à partir de 2011. La corrélation entre les deux indices est de **99,48 %**.

Depuis janvier 2021, une forte augmentation des indices est constatée, celle-ci s'explique par la crise sanitaire mondiale, liée au Covid-19, qui a entraîné une pénurie de matériaux et de matières premières, ainsi que des retards d'approvisionnement. Cette augmentation se constate également sur l'inflation, qui est représentée sur le graphique suivant :

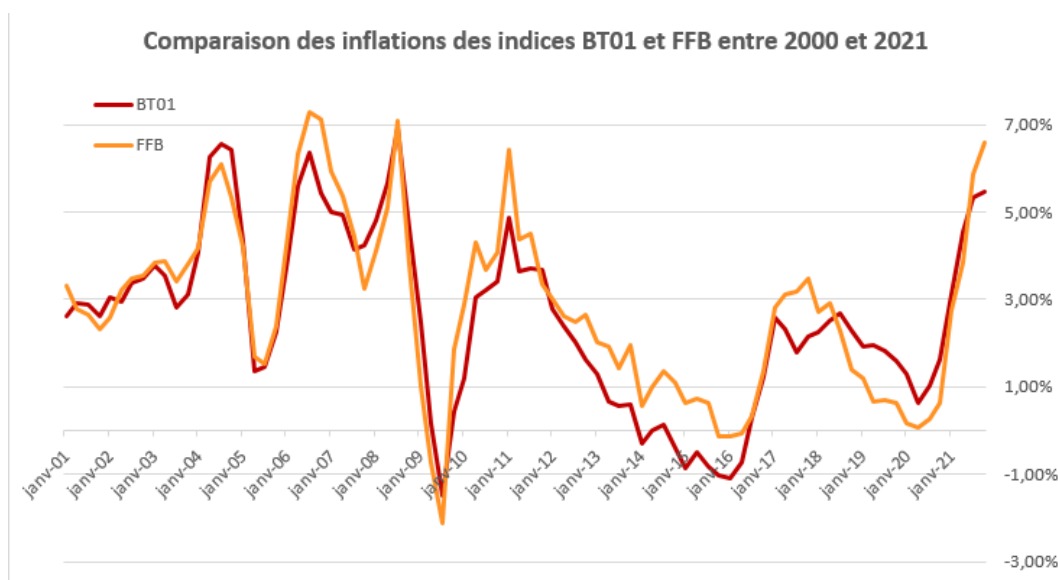


FIGURE 3.4 – Comparaison de l'inflation des indices BT01 et FFB entre 2000 et 2021

Les inflations suivent les mêmes tendances, cependant celle de l'indice FFB est souvent supérieure à celle du BT01. La corrélation est de **91,63 %** pour les inflations.

Remarques : Origines des fluctuations des inflations : De nombreuses fluctuations de l'inflation

sont constatées sur la graphique précédent. La crise des subprimes en 2007 entraîne une diminution du nombre de mises en chantier qui pèse sur le marché et qui se traduit par une forte mise en concurrence et une baisse des prix, d'où la forte diminution constatée entre le milieu de l'année 2008 et l'année 2010. Un rebond intervient à partir de 2010 avec une reprise de l'activité mais une chute progressive est constatée jusqu'en 2016. Celle-ci s'explique par la crise du bâtiment et la réduction des effectifs. Un niveau positif de l'inflation est retrouvé en 2017 avant de subir la crise du Covid en 2020. Cette crise sanitaire se traduit par une baisse de l'inflation en 2020 au moment du confinement, la reprise des activités et la forte demande entraîne des tensions sur le marché et une augmentation significative de l'inflation. Cette augmentation se poursuit aujourd'hui en sortie de crise sanitaire et en conséquence de la guerre en Ukraine.

Choix de l'indice pour la revalorisation des charges des sinistres : en constatant que les corrélations entre les valeurs des indices, et entre leur inflation sont supérieures à 90%, et afin d'être cohérent avec les choix faits au sein du groupe, il est décidé de choisir **l'indice BT01**.

Méthode de calcul de la revalorisation : Dans un second temps, connaissant l'indice choisi pour la revalorisation, il faut l'utiliser afin de calculer l'indice de revalorisation. La formule de calcul est la suivante :

$$\text{indice de revalorisation} = \frac{\text{indice du mois de l'étude}}{\text{indice du mois de survenance du sinistre}}$$

Finalement, cet indice de revalorisation est multiplié au montant de sinistre associé afin d'obtenir la charge revalorisée des sinistres.

La répartition des charges avant et après la revalorisation est la suivante :

	Charges non revalorisées	Charges revalorisées
RCD	1,076 Md	1,474 Md
RC Bât	443 M	640 M

3.4 Segmentation des sinistres

Le traitement des sinistres est différent selon sa gravité, son montant, il convient donc de déterminer un seuil permettant de classer les sinistres. Le seuil historique de gravité au sein du groupe est de 150 000 €, ainsi les sinistres dont le montant est supérieur à ce montant sont considérés comme des sinistres graves. Cependant certaines équipes, comme l'équipe Inventaire, ne conservent pas ce seuil historique mais définissent plusieurs tranches distinctes pour répartir les sinistres. Les sinistres sont répartis en fonction de leur coût, de la manière suivante : inférieur à 30 000 €, entre 30 000 € et 750 000 €, entre 750 000 € et 2 000 000 €, et supérieur à 2 000 000 €.

L'objectif de cette partie est de vérifier l'adéquation du seuil de 150 000 € avec l'historique de sinistralité pour les garanties RCD et RC Bât, mais également de définir d'éventuels nouveaux seuils pour segmenter d'avantage cette sinistralité. Pour cela, dans un premier temps, la théorie des valeurs extrêmes sera rappelée, puis les estimateurs utilisés explicités, enfin leurs utilisations et les résultats obtenus seront décrits.

3.4.1 Théorie des valeurs extrêmes

La théorie des Valeurs Extrêmes s'intéresse aux "grandes" valeurs d'une distribution de probabilité, à la queue de distribution. Elle permet l'étude de ces valeurs extrêmes associées à des événements extrêmes et très peu observés, en segmentant les données en fonction de leur sévérité.

Soit $X_1, \dots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées (IID) de fonction de distribution F et soit $M_n = \max(X_1, \dots, X_n)$. Alors

$$\begin{aligned}\mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, \dots, X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \dots \mathbb{P}(X_n \leq x) \\ &= [F(x)]^n.\end{aligned}$$

Si F n'est pas connue, cette formule est peu utile et son estimation se ferait à partir de la partie centrale de la distribution. De plus, nous nous intéressons souvent à M_n , lorsque la taille de l'échantillon est importante, et lorsque des approximations asymptotiques sont souhaitées.

Notons x^F le point extrême de F , défini par :

$$x^F = \sup\{x \mid F(x) < 1\}$$

alors $M_n \rightarrow x^F$ en probabilité quand $n \rightarrow \infty$. La distribution asymptotique de M_n est donc dégénérée : une normalisation de la suite est donc nécessaire.

Un des théorèmes fondamentaux de la théorie des valeurs extrêmes est celui de Fisher-Tippett, sa formulation est la suivante :

Théorème (Fisher - Tippett) : s'il existe des suites de réels $a_n > 0$ et b_n telles que quand $n \rightarrow \infty$,

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = [F(xa_n + b_n)]^n \rightarrow G(x)$$

pour une distribution non-dégénérée G , alors G est du même type que l'une des trois distributions suivantes :

$$\begin{aligned}\text{Fréchet } (\alpha > 0) : \phi_\alpha(x) &= \begin{cases} 0 & \text{si } x \leq 0 \\ \exp(-x^{-\alpha}) & \text{si } x > 0 \end{cases} \\ \text{Weibull } (\alpha > 0) : \Psi_\alpha(x) &= \begin{cases} \exp(-(-x)^\alpha) & \text{si } x \leq 0 \\ 0 & \text{si } x > 0 \end{cases} \\ \text{Gumbel} : \Lambda(x) &= \exp(-e^{-x}) \quad x \in \mathbb{R}\end{aligned}$$

Pour résoudre des problèmes statistiques, Von Mises (1954) et Jenkins (1955) ont proposé la loi d'extremum généralisée (notée GEV pour Generalized Extreme Value) qui unifie les distributions de Fréchet, Weibull et Gumbel.

Définition : La loi d'extremum généralisée $GEV(\mu, \sigma, \xi)$ est définie par la fonction de répartition :

$$G(x) = \begin{cases} \exp(-[1 + \xi(\frac{x-\mu}{\sigma})]_+^{-\frac{1}{\xi}}) & \text{si } \xi \neq 0 \\ \exp(-\exp(-(\frac{x-\mu}{\sigma}))) & \text{si } \xi = 0 \end{cases}$$

ξ est le paramètre de forme (également appelé paramètre de queue), μ le paramètre de position et σ le paramètre d'échelle.

La correspondance entre la GEV et les lois limites est déterminée par le paramètre de queue :

- La distribution de Gumbel correspond à $\xi = 0$: $GEV(0, 1, 0) = Gumbel$
- La distribution de Fréchet correspond à $\xi > 0$: $GEV(1, \alpha^{-1}, \alpha^{-1}) = Fréchet(\alpha)$
- La distribution de Weibull correspond à $\xi < 0$: $GEV(-1, \alpha^{-1}, -\alpha^{-1}) = Weibull(\alpha)$

3.4.2 Détermination de seuils de gravité avec différents estimateurs

La partie précédente, sur la théorie des valeurs extrêmes permet de proposer plusieurs méthodes graphiques dédiées à l'étude des sinistres extrêmes et à la détermination de seuils de gravité. Elles seront présentées dans cette partie.

Ces estimateurs seront appliqués à l'ensemble des sinistres, ainsi qu'aux sinistres RCD et RC Bât séparément, afin d'établir s'il est nécessaire de considérer des seuils différents selon le type de garantie. La partie 3.1 a permis de mettre en évidence une forte concentration des charges inférieures à 2 000 euros, avec un pic entre 600 et 800 euros (cela correspond majoritairement à des coûts forfaitaires d'ouverture de sinistres ou d'expertises). Afin de ne pas avoir un poids trop important sur les sinistres de coût faible, les sinistres de coût inférieur à 800 euros ne sont pas considérés pour la détermination des seuils de gravité. De plus, le nombre de sinistres au dessus de 2 millions d'euros est assez faible, ainsi l'étude de la base sera limitée à des sinistres de coût inférieur à ce montant.

Étude des quantiles :

Les quantiles sont utiles pour étudier la répartition, en nombre, des montants dans la base. Chaque quantile est associé à un niveau de charge et ils permettent d'observer, en particulier, comment sont répartis les sinistres de charge élevée. Ainsi le graphique suivant représente les variations, en pourcentage, de l'augmentation de 0,1 point de pourcentage d'un quantile. Une forte évolution permet de détecter un seuil de gravité.

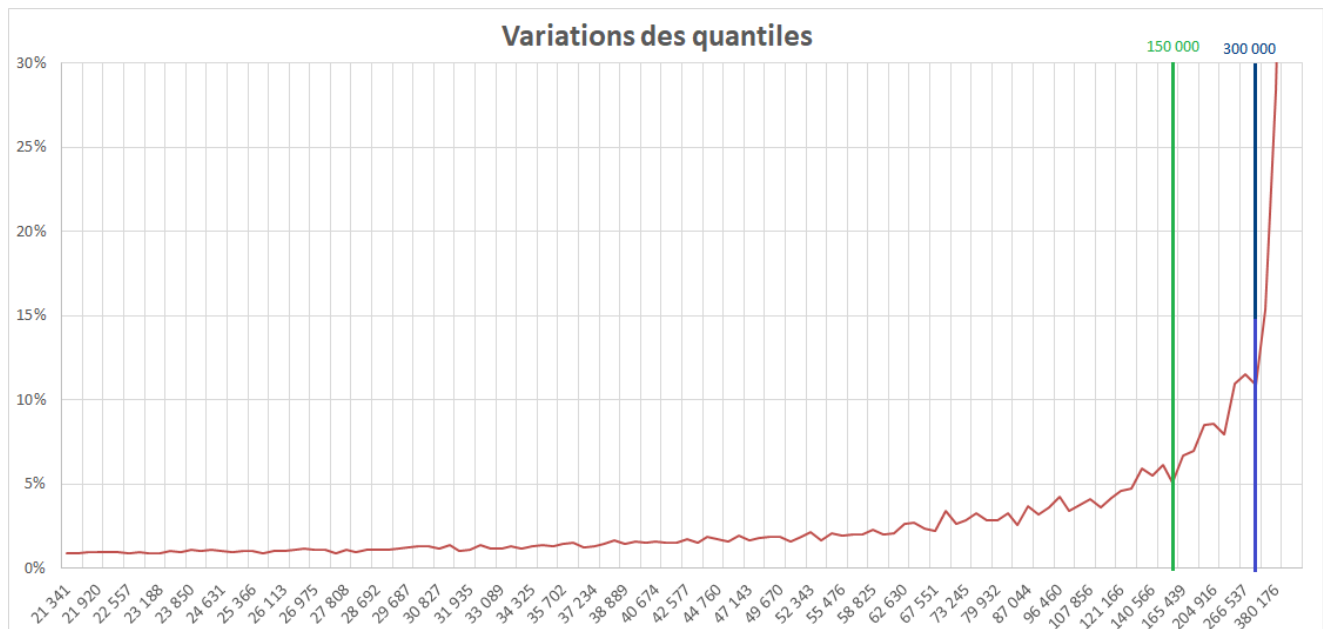


FIGURE 3.5 – Variations des quantiles pour l'ensemble des données

Ce graphique permet de distinguer deux seuils : autour de **150 000** et autour de **300 000**. L'étude des quantiles est également effectuée pour les sinistres RCD et RC Bât séparément et met en évidence des niveaux de seuils similaires.

La fonction de dépassement moyen (mean excess function) :

Cette fonction est un outil permettant d'étudier les extrêmes ainsi que le comportement au-dessus d'un seuil u . Elle est définie par la formule suivante :

$$e_n(u) = \mathbb{E}[X - u | X > u]$$

Sa fonction empirique est définie par :

$$\hat{e}_n(u) = \frac{\sum_{i=1}^n (X_i - u)^+}{\sum_{i=1}^n \mathbb{1}_{X_i > u}} = \frac{\sum_{i=1}^n X_i \mathbb{1}_{X_i > u}}{\sum_{i=1}^n \mathbb{1}_{X_i > u}} - u$$

La fonction de dépassement moyen est donc estimée par la somme des excès au delà du seuil u , divisée par le nombre de points excédant ce seuil u .

Un moyen de déterminer les seuils est de tracer le graphique de dépassement moyen : $\{X_{(i)}, e_n(X_{(i)}) : i = 1, \dots, n\}$. Le graphique doit être approximativement linéaire au delà de ce seuil. Pour l'ensemble des sinistres, le graphique obtenu est le suivant :

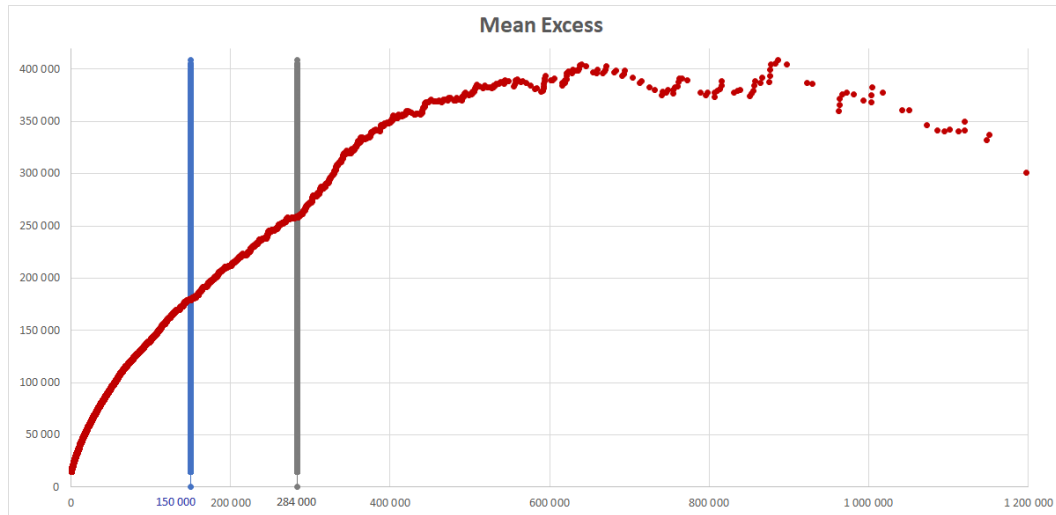


FIGURE 3.6 – Représentation de la fonction de dépassement moyen

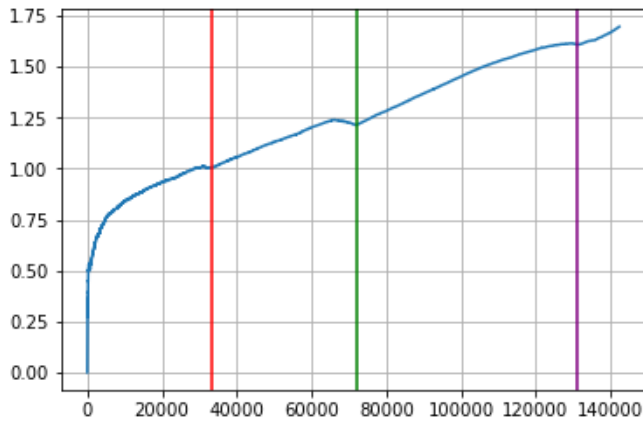
Sur ce graphique deux seuils peuvent être identifiés : **150 000** et **284 000**. Pour les bases RCD et RC Bât, le seuil autour de 150 000 est retrouvé, un second seuil se distingue autour de 300 000 pour la RCD et autour de 250 000 pour la RC Bât.

Estimateur de Hill :

Il s'agit de l'estimateur le plus connu permettant de déterminer les seuils au dessus desquels un changement de loi apparaît. En notant $X_{(1)} \geq \dots \geq X_{(n)}$ la statistique d'ordre et n le nombre d'observations, l'estimateur de Hill s'écrit, pour $k = 1, \dots, n - 1$:

$$\hat{H}_{k,n} = \frac{1}{n} \sum_{j=1}^k \log(X_{n-j+1,n}) - \log(X_{n-k,n}) \quad (3.4.1)$$

La représentation graphique de cet estimateur permet de faire apparaître plusieurs seuils au delà desquels l'estimateur est linéaire. Pour l'ensemble des données, RCD et RC Bât, le graphe obtenu est de la forme suivante, ici est représenté le graphique de l'estimateur de Hill pour les données de la garantie décennale :



Cet estimateur permet de faire 3 seuils de gravité pour chacune des bases. Les montants associés sont résumés dans le tableau ci-dessous :

	Seuil 1	Seuil 2	Seuil 3
Base Sinistre	1 465	3 544	32 506
RCD	1 933	5 334	31 131
RC Bât	1 790	8 156	31 426

FIGURE 3.7 – Détermination de plusieurs seuils avec l'estimateur de Hill

Estimateur de Gerstengarbe :

Proposé par Gerstengarbe et Werner en 1989, il permet de déterminer le point de départ de la région extrême et il fournit une estimation de ce seuil.

Pour un échantillon $x_1, \dots, x_n$ des coûts de sinistres, on considère la série des différences $\Delta_i = x_{[i]} - x_{[i-1]}$ pour $i = 2$ à n de l'échantillon ordonné, $x_{[1]} \leq \dots \leq x_{[n]}$.

L'idée de la procédure est qu'en entrant dans la zone de sinistres extrêmes, on peut s'attendre à un changement dans le comportement de la série des différences. Pour identifier le changement de comportement, nous utilisons le test de Mann-Kendall, et nous calculons, pour $i = 1$ à $n-1$, la série U_i définie par :

$$U_i = \frac{U_i^* - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i-1)(i+5)}{72}}} \quad (3.4.2)$$

où $U_i^* = \sum_{k=2}^i n_k$ où n_k est le nombre de valeurs de $\Delta_2, \dots, \Delta_k$ inférieures à Δ_k

De manière équivalente, une autre série U_i^f est calculée, pour la série décroissante des différences $\Delta_n, \dots, \Delta_2$. Le point d'intersection de ces deux séries détermine le seuil d'entrée dans la zone extrême.

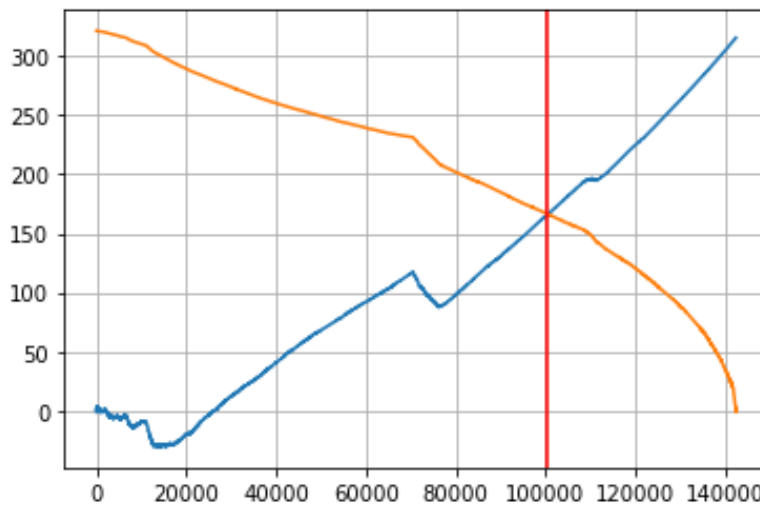


FIGURE 3.8 – Détermination d'un seuil avec l'estimateur de Gerstengarbe

Sur ce graphique, le point d'intersection se situe autour de la 100 000^{ème} observation de la base ordonnée contenant l'ensemble des données, elle correspond à une valeur de 7 553.

Le premier seuil est désormais déterminé, il est alors possible de reprendre cette méthode en ne considérant que les données supérieures à ce montant. Les deux séries U_i et U_i^f sont calculées et un nouveau point d'intersection peut apparaître et un deuxième seuil de gravité est déterminé. Cette méthode est répétée autant de fois que de points d'intersection peuvent être distingués. Lorsque le graphique obtenu ne présente aucun point d'intersection ou qu'il en présente plusieurs, l'estimateur n'est plus optimal et l'étude se termine.

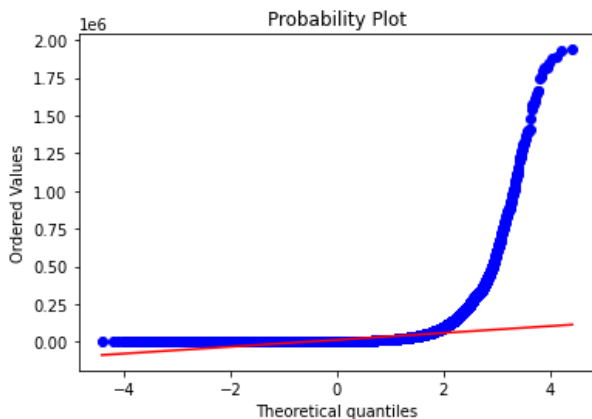
Cet estimateur a permis de faire apparaître les seuils suivants pour les différentes bases :

	Seuil 1	Seuil 2	Seuil 3	Seuil 4
Ensemble des données	7 553	31 792	97 574	247 751
Données RCD	9 716	35 876	106 231	255 600
Données RC Bât	6 132	27 418	94 258	214 759

Plusieurs seuils semblent se distinguer : 30 000 et 100 000.

QQ-Plot (Graphique Quantile-Quantile) :

Il s'agit d'une méthode graphique permettant de comparer deux distributions de probabilité en affichant leurs quantiles contre quantiles.



Un graphique quantile-quantile est une méthode graphique permettant de mesurer l'adéquation d'une variable observée à une loi théorique de fonction de répartition F continue. Ici une comparaison est faite avec la loi normale, le graphe permet de constater qu'il n'y a pas adéquation des lois. Il permet de constater une queue de distribution avec un poids important.

FIGURE 3.9 – QQ-Plot ensemble des données

3.4.3 Choix des seuils de gravité

Les estimateurs ont permis de faire apparaître plusieurs seuils de gravité. Le tableau ci-dessous résume les seuils déterminés dans la partie précédente :

Quantiles	Mean Excess	Hill	Gerstengarbe
-	-	1 465	-
-	-	3 544	-
-	-	-	7 533
-	-	32 506	31 792
-	-	-	97 574
150 000	150 000	-	-
-	-	-	247 751
300 000	284 000	-	-

L'utilisation des différents estimateurs a permis de faire apparaître différents seuils :

- 30 000 euros
- 150 000 euros
- 300 000 euros

Les seuils précédemment déterminés sont complétés avec un nouveau à 750 000, en lien avec le seuil déterminé par l'équipe Inventaire et celui à 2M déterminé plus tôt qui marque une diminution des sinistres. L'ajout de ces seuils supplémentaires permettra de diviser d'avantage la sinistralité, et de regarder comment les sinistres sont répartis entre tous ces seuils.

Le tableau suivant indique la répartition des sinistres en montant dans les différentes tranches.

	30k	150k	300k	750k	2M	2M et +
Charges revalorisées totales	34,64%	29,49%	12,46%	11,25%	7,42%	4,74%
Charges Revalorisées RCD	34,93%	31,94%	12,28%	10,16%	7,14%	3,53%
Charges Revalorisées RC Bât	33,95%	23,84%	12,87%	13,75%	8,05%	7,52%
Charges non Revalorisées totales	40,07%	29,59%	12,00%	8,68%	5,95%	3,41%
Charges non Revalorisées RCD	40,59%	31,28%	11,32%	8,20%	5,74%	2,55%
Charges non Revalorisées RC Bât	38,82%	25,49%	13,66%	9,386%	6,47%	5,50

Les tranches [0 – 30k], [30k – 150k] et [150k – 2M] représentent chacune environ 30% de la charge totale. Les observations sont réparties de la manière suivante dans ces tranches :

- **[0 – 30k]** : majorité des observations, plus de 90%
- **[30k – 150k]** : entre 3 et 7% des observations
- **[150k – 2M]** : environ 1% des observations
- **Plus de 2M** : moins de 0,5% des observations

Le nombre de sinistres dans les tranches décroît lorsque les coûts augmentent. Cette constatation est logique car les sinistres avec des coûts élevés sont moins fréquents que des sinistres de coûts plus faibles (les sinistres attritionnels).

Ces différentes observations permettent de conclure sur la sélection des seuils pour la suite et pour la modélisation. Ainsi sont retenus les seuils :

- 30 000
- 150 000 (seuil historique)
- 2 000 000

Cette division des charges permet de définir quatre tranches dont les modélisations seront effectuées dans la cinquième partie.

Lors de cette étude, l'ensemble des données sur la sinistralité ont été conservée, afin d'avoir l'historique le plus complet possible. Cependant lors de la modélisation, une sélection sera effectuée en fonction des contrats conservés. Ce choix sera fait dans la partie suivante lors de la création de la base Risque.

Préparation de la base Risque

Après avoir créé la base de données sur la sinistralité, il convient désormais de concevoir celle regroupant les informations sur les contrats et sur les risques. Un historique de 15 ans est choisi, et tous les contrats ayant été en vigueur pendant au moins 1 jour durant ces 15 dernières années sont conservés. La base créée sera historisée, cela signifie que pour un contrat donné il y aura autant de lignes que d'années pendant lesquelles le contrat a été en vigueur.

Durant les 15 dernières années, 40 571 contrats, RC Bât ou RCD, ont été en vigueur au moins 1 journée, avec 9 783 contrats actuellement en vigueur et 30 788 contrats résiliés. La table créée comporte 201 665 lignes.

4.1 Récupération des données initiales

Afin de créer la base risque, il convient de récupérer l'ensemble des données sur les contrats, les assurés et sur les caractéristiques des entreprises. Celles-ci sont stockées dans différentes tables :

- **Table contrat** contient les informations générales sur un contrat, comme le numéro de police, l'état du contrat, les dates d'effet du contrat, la garantie associée et le montant de la prime. Elle permet de récupérer tous les contrats actuellement en vigueur ou qui ont été en vigueur, pendant au moins une journée, dans les 15 dernières années, et qui ont une garantie RCD ou RC Bât.
- **Table entreprise** permet de récupérer les informations sur les entreprises. Les informations contenues dans cette base correspondent à celles déclarées à la souscription ou lors du dernier avenant. Enfin, elle contient le numéro de police du contrat que l'entreprise possède, qui permettra de faire la liaison avec la table précédente.
- Il existe ensuite une table contenant tous les historiques de données sur les différentes variables, comme, par exemple, le chiffre d'affaire, l'effectif ou l'activité. Elle permet d'associer à chaque année une valeur tarifante. Enfin, toutes ces tables contiennent le numéro de client ou le numéro de police afin de faire la liaison avec les tables précédentes.

4.2 Création de la base

Pour créer la base risque, des variables permettant d'obtenir des informations sur le contrat et sur l'entreprise sont considérées. La récolte des variables est réalisée en deux phases : dans un premier temps, les variables internes liées au contrat sont récupérées, puis celles sur l'entreprise et son activité sont considérées.

4.2.1 Choix des variables

Les variables retenues pour la base risque et donnant des informations sur le contrat sont les suivantes :

L'année permet de spécifier la date à laquelle le contrat est considéré. Elle est comprise entre 2007 et 2022, et est associée à un contrat lorsque celui-ci a été en vigueur au moins une journée pendant cette année.

L'exposition, notée E , du contrat sur un exercice donné est la proportion de l'année pendant laquelle le contrat est en vigueur. Par exemple, pour l'année n , l'exposition se calcule de la manière suivante :

$$E_n = \frac{\min(\text{Date résiliation}, 01/01/n + 1) - \max(\text{Date souscription}, 01/01/n)}{01/01/n + 1 - 01/01/n}$$

Le type de contrat indique le mode de calcul de la prime. Les contrats émis peuvent avoir une prime **Forfaitaire** ou **Révisable**.

La typologie de l'assuré indique les caractéristiques de l'assuré. Ainsi celui-ci peut être **Entreprise** ou **ProPE**.

La prime acquise de l'exercice n , notée P_n , correspond à la prime hors taxes payée au titre de l'exercice n . Un contrat pouvant être en vigueur sur deux années d'exercice consécutives, la prime devra être calculée afin de prendre en compte la proportion de présence du contrat sur une année d'exercice.

Soit un contrat émis au cours de l'année n , de prime annuelle émise la première année, P_1 , et de prime annuelle P_2 la deuxième année. La période d'acquisition de la prime est noté A , et correspond à la proportion de l'année n pendant laquelle la prime annuelle s'applique. Le graphique ci-dessous résume les informations précédentes :

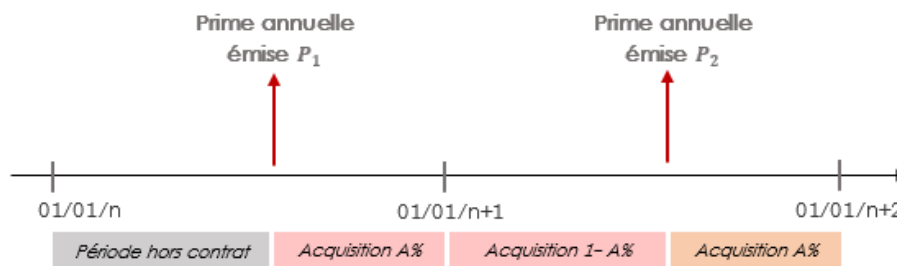


FIGURE 4.1 – Fractionnement de la prime par exercice

Le calcul des primes acquises des exercices n et $n + 1$ pour ce contrat est donc :

$$P_n = P_1 \times A\%$$

$$P_{n+1} = P_1 \times (1 - A\%) + P_2 \times A\%$$

L'ancienneté du contrat est calculée à partir de la date de souscription du contrat, il s'agit de la différence entre l'année de vision du contrat et l'année de souscription du contrat.

Multi équipement : On détermine si le client possède plusieurs contrats sur plusieurs produits différents pendant la période considérée. Cet indicateur est souvent utilisé par les équipes clients et souscription qui adaptent le tarif, et l'attention portée à un client lorsque ce dernier est multi équipé. Il est calculé à partir du **nombre de contrats** possédé par l'assuré au sein de Generali France et du **nombre de contrats en assurance construction** dont est pourvu ce même assuré. Ces deux autres variables sont également présentes dans la base Risque.

Désormais, les variables apportant des précisions sur l'entreprise et son activité sont considérées :

Le chiffre d'affaires est déclaré à la souscription ou lors du dernier avenant pour les contrats forfaitaires, tandis que pour les contrats révisables celui-ci est déclaré tous les ans.

Pour ce dernier type de contrat, et comme pour les primes précédemment, la période de validité du chiffre d'affaires déclaré peut être sur deux années, il convient alors de procéder comme précédemment afin de séparer le montant entre les exercices n et $n + 1$.

Pour les contrats forfaitaires, on récupère le dernier chiffre d'affaires déclaré et l'exercice auquel il est rattaché, et on l'utilisera pour déterminer les montants des années précédentes si celui-ci est manquant, la procédure utilisée sera détaillée ultérieurement.

Le pourcentage de chiffre d'affaires sous traité est déclaré par l'entreprise et permet de déterminer si une entreprise sous traite tout ou partie de son activité.

L'effectif est une donnée permettant de connaître la dimension de l'entreprise et de suivre l'évolution de son activité. C'est une variable tarifante pour les contrats forfaitaires, elle est déclarée au moment de la souscription ou de la re-souscription du contrat. Les données pour cette variable sont différentes, il est possible d'avoir l'historique des effectifs, le dernier connu ou celui déclaré à la souscription. Afin de déterminer les effectifs manquants, l'hypothèse de la stabilité de l'effectif peut être faite ou il est possible d'utiliser le chiffre d'affaires en calculant un montant moyen par salarié. Ces deux méthodes seront décrites dans la section suivante.

L'ancienneté de l'entreprise est calculée à partir de la date de création de l'entreprise qui est déclarée à la souscription du contrat. Il s'agit de la différence entre l'année de vision du contrat et l'année de la création de l'entreprise

L'activité est présente dans la base de données avec plusieurs variables permettant d'apporter des informations différentes. Les différents indicateurs sont détaillés ci-dessous :

- **Le code activité** composé, aujourd'hui, de 5 caractères permet de caractériser l'activité de l'assuré. Il a évolué durant les dernières années, afin de pouvoir comparer tous ces codes, il convient de les traiter pour qu'ils aient tous le même format.
- **Le type d'activité** est déterminé à partir du code activité faisant référence à la nomenclature de la Fédération des Assurances des activités du bâtiment. Elle définit cinq familles d'activités, et un code leur est associé :
 - Lots divisions - Aménagement, notée ZD
 - Lots techniques, notée ZT
 - Clos et couvert, notée ZC
 - Structure et Gros Œuvre, notée ZS
 - Préparation et aménagement du site, notée ZP

Ces activités peuvent être regroupées selon le type d'œuvre réalisé. Soit en Gros Œuvre (GO), correspondant aux travaux liés à la construction ou à la rénovation de l'ossature du

bâtiment, incluant les composantes qui participent à la stabilité et à la solidité de l'édifice. Ainsi sont classées dans cette catégorie les activités ZC, ZS et ZP. Soit en Second Œuvre (SO) correspondant à tout ce qui n'est pas du gros œuvre, ainsi tout ce qui ne concerne pas l'ossature d'un bâtiment. Cette catégorie regroupe les activités ZD et ZT.

Cependant d'autres activités, non détaillées dans la nomenclature de la Fédération des Assurances, sont présentes dans le portefeuille :

- Travaux tous corps d'état, noté ZX
- Fabrication – Négoces, noté FABNG
- Des professions libérales, notées Prof Lib
- D'autres professions, notées AUTRE

Ces activités ne font parties d'aucun des regroupements détaillés précédemment. Ainsi la variable **Regroupement d'activité** présentera comme modalité Gros Œuvre, Second Œuvre, Travaux tous corps d'état, Fabrication-Négoces, AUTRE, Prof Lib et "Non renseigné" pour les activités manquantes.

- Une entreprise pouvant réaliser plusieurs activités pour un exercice donné, **le nombre d'activités** sera rapporté afin d'indiquer cette multiplicité des activités. Cependant pour ne pas alourdir la base de données, seulement l'activité principale sera indiquée. Ce nombre d'activités est compris entre 1 et 22.
- Les activités peuvent être classées selon un critère d'**activité aggravante**. Pour les activités faisant parties de la nomenclature des activités du bâtiment de la fédération des assurances, un taux est déterminé en fonction de la sinistralité de l'activité. Les autres professions telles que les travaux tous corps d'états, prof lib, fabrique-négoces, AUTRE et "Non renseigné", ne sont pas concernés par ce classement et se verront affecter une lettre (respectivement "X", "L", "F", "AUTRE" et "N") pour les distinguer.

Informations géographiques : le département et la région de l'entreprise sont récupérés lors de la formation du contrat d'assurance. Les deux variables étant corrélées, on ne peut en garder qu'une seule pour la modélisation, la variable région sera privilégiée ayant une meilleure corrélation avec les variables à expliquer.

Remarque : Des recherches ont été réalisées afin d'inclure des variables externes telles que les certifications possédées par un assuré (certifications QUALIBAT¹ ou QUALIFELEC²) mais également pour obtenir des informations sur les risques géologiques tels que les inondations ou la sécheresse. Cependant les bases de données ne sont pas assez complètes, et comptent de trop nombreuses valeurs manquantes. Ces variables ne seront donc pas considérées pour la suite de l'étude.

4.2.2 Traitement des valeurs manquantes et aberrantes

Lors de l'élaboration de la base de données, la présence de valeurs manquantes ou aberrantes a été constatée. Cette partie explique les traitements, pour les différentes variables, qui ont été mis en place pour supprimer ces valeurs, enfin un tableau récapitulera les valeurs manquantes présentes dans la table finale.

1. Label attribué par l'association QUALIBAT pour tous les travaux de construction et de rénovation à l'exception de l'électricité

2. Label attribué par QUALIFELEC, l'association professionnelle et technique de qualification des entreprises du génie électrique et énergétique

Le Chiffre d'affaire présentait des valeurs manquantes ainsi que des valeurs pouvant sembler aberrantes, en effet certains assuré avait un chiffre d'affaires faible, inférieur à 1 000 €. Pour les contrats révisables, le montant déclaré est gardé mais pour ceux forfaitaires, le chiffre d'affaire sera considéré manquant. L'absence du chiffre d'affaire apparaît de plusieurs manières :

- Le montant est manquant pour l'ensemble des années où le contrat a été en vigueur dans le portefeuille
- Le montant est manquant pour certaines années et déclaré pour d'autres pendant la durée de présence du contrat dans le portefeuille

Dans le second cas, le montant du dernier chiffre d'affaires connu est récupéré et sera revalorisé avec l'indice du BT01 pour déterminer ceux manquants. Les chiffres d'affaires des contrats étant dans le premier cas, seront traités ultérieurement avec les traitements apportés à d'autres variables.

L'effectif présentait des valeurs manquantes étant des deux mêmes formes que pour le chiffre d'affaires (totalement manquant ou partiellement manquant sur les années de présence dans le portefeuille). Une première approche est d'utiliser le chiffre d'affaires connu pour en déduire l'effectif. Pour cela, un montant de chiffre d'affaires moyen par salarié est calculé pour les 15 années d'historique, et la valeur obtenue est de 110 000 € de chiffres d'affaires pour un salarié. Ce montant permet de déterminer un effectif mais il peut également être utilisé dans le sens inverse et permettre de déterminer le chiffre d'affaire lorsque celui-ci est manquant mais que l'effectif est connu. Cette méthode permet donc de compléter les informations sur les effectifs mais aussi sur le chiffre d'affaire. Cependant, si l'information sur le chiffre d'affaire est manquante et que l'effectif a été déclaré pour certaines années, le dernier effectif connu est récupéré et on le suppose constant dans le temps.

Les activités ne sont déclarées systématiquement que depuis 2013, ainsi avant cette date de nombreuses valeurs manquantes sont présentes. L'activité principale et le nombre d'activité sont déterminés pour chaque contrat et chaque année, il est alors possible de compléter ces informations lorsque celles-ci sont connues pour au moins une année et en les supposant constant au cours du temps.

Le schéma ci-dessous résume les étapes permettant d'obtenir les informations complètes sur l'activité :

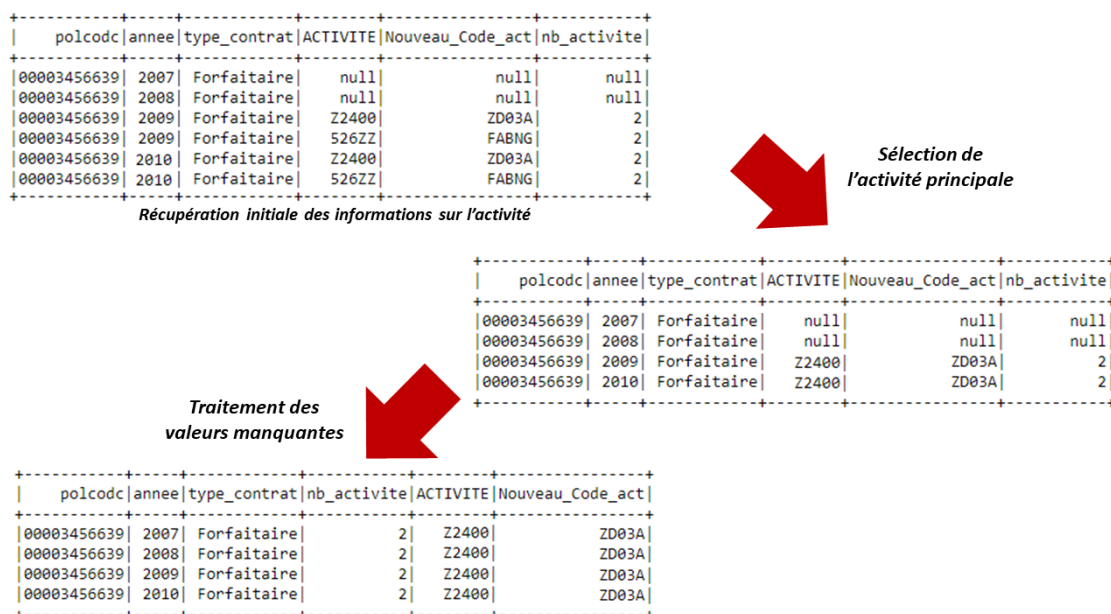


FIGURE 4.2 – Récupération et traitement des informations sur l'activité

La date de création de l'entreprise permet de déterminer son ancienneté et peut être une source de fiabilité. En effet, une entreprise présente sur le marché depuis plusieurs décennies révèle que les travaux réalisés sont de bonne qualité. Des valeurs aberrantes de différents types ont été constatées :

- Des valeurs inférieurs à 1900, comme par exemple 0009 ou 1193. Ces années sont en faible nombre et elles seront corrigées manuellement. Dans l'exemple, les années deviennent 2009 et 1993.
- L'année de création d'entreprise supérieure à celle de mise en vigueur du contrat, elle sera donc rehaussée par cette dernière. En effet, la date de création du contrat est une variable interne, et non une donnée rapportée par le client, ainsi l'erreur sur celle-ci est minime et elle ne présente pas de valeurs manquantes ou aberrantes.

Les primes acquises sont parfois manquantes, nulles ou négatives, afin de régler ce problème, elles seront rehaussées à la prime terme.

Lorsque tous les traitements précédents ont été appliqués, il reste pour certaines variables des valeurs manquantes. Le tableau ci-dessous permet de résumer ces informations :

Variables descriptives	Lignes manquantes	Taux	Choix de traitement
Chiffres d'affaires	7 709	3,82%	Considéré comme une modalité
Effectif	7 709	3,82%	Considéré comme une modalité
CA sous traité	120 688	59,85%	Considéré comme une modalité
Ancienneté contrat	0	0%	-
Ancienneté entreprise	9 249	4,59%	Considéré comme une modalité
Primes	0	0%	-
Activités	19 706	9,77%	Considéré comme une modalité
Multi équipement	0	0%	-
Région	595	0,30%	Regroupé avec la modalité la plus représentée (Auvergne Rhône-Alpes)

Une modalité "Non renseigné" est créée pour les variables où des valeurs manquantes sont présentes. Celles-ci ne sont pas supprimées car elles enlèveraient de l'information. La modélisation permettra de déterminer si cette modalité est significative ou non, et si elle apporte de l'information.

La création de cette base a permis de rassembler des données initialement présentes dans des tables différentes, et de récupérer les derniers chiffres déclarés. Un traitement des valeurs manquants a été nécessaire, cependant toutes n'ont pas pu être supprimées de la base, la modélisation permettra de choisir comment les appréhender.

Les deux bases de données sont désormais finalisées, la partie suivante décrira la mise en place de la modélisation et des travaux de préparation réalisés.

Modélisation de la fréquence et de la sévérité

Les bases de données ont été créées, elles contiennent l'ensemble des informations sur un assuré ainsi que sur sa sinistralité. Ces bases vont désormais pouvoir être utilisées afin de modéliser le montant attendu des sinistres. Plusieurs méthodes vont être testées et détaillées. La modélisation avec les modèles linéaires généralisés est effectuée avec Emblem.

La modélisation permet de déterminer la charge totale des sinistres, calculée, pour un assuré k , par la formule $S_k = \sum_{i=1}^{N_k} C_i$ avec

- N_k une loi de probabilité à valeur dans $\mathbb{N}$, correspondant au nombre de sinistres de l'assuré k . Ainsi $S_k = 0$ si $N_k = 0$
- C_i un ensemble de lois indépendantes et identiquement distribuées à support réel et strictement positif, correspondant au montant des N_k sinistres
- Une hypothèse supplémentaire est ajoutée, telle que les C_i sont indépendants des N_k et indépendants entre eux. Cette hypothèse théorique, très forte est discutable en réalité avec les comportements humains.

L'approche utilisée est un modèle Fréquence-Coût. Il se définit comme l'association de deux variables supposées indépendantes :

- **La fréquence** de sinistres correspondant au nombre de fois où un sinistre est survenu durant la période d'exposition du client :

$$fréquence = f_k = \frac{NB_SIN}{Exposition}$$

- **le coût moyen** correspondant au montant de sinistre moyen pour un assuré (si cet assuré est sinistré) :

$$coût = c_k = \frac{Charge\ totale\ des\ sinistres}{NB_SIN}$$

L'approche de ce modèle s'explique par le calcul :

$$\underbrace{\mathbb{E}\left(\sum_{i=1}^{N_k} C_i\right)}_{Prime\ Pure} = \mathbb{E}(S_k) = \mathbb{E}(\mathbb{E}(S_k|N_k)) = \mathbb{E}\left(\sum_{i=1}^{N_k} \mathbb{E}(C_i|N_k)\right) = \mathbb{E}(N_k \times \mathbb{E}(C)) = \underbrace{\mathbb{E}(N_k)}_{Fréquence} \times \underbrace{\mathbb{E}(C)}_{Coût\ Moyen}$$

Ce calcul suppose une indépendance de la fréquence et des coûts de sinistres. Le résultat de l'espérance du montant de sinistre total, ou prime pure, est égal au produit des espérances de

la fréquence et du coût moyen. Cela revient donc à modéliser la fréquence de sinistre et le coût moyen. Cette approche sera suivie dans ce mémoire.

Cette partie détaillera toutes les étapes suivies pour modéliser la sinistralité du portefeuille. Ainsi il conviendra, dans un premier temps, de créer les bases utiles pour la modélisation et d'effectuer les traitements fondamentaux. Seront ensuite présentés les éléments nécessaires à la sélection d'un modèle. Enfin les différents modèles seront exposés : les modèles linéaires généralisés et ceux de Machine Learning.

5.1 Préparation à la modélisation

Afin de réaliser une modélisation performante, il est nécessaire que les bases de données utilisées soient optimales, c'est-à-dire ne présentant pas d'incohérence et ayant le moins de valeurs manquantes possibles. Ces traitements ont été détaillés dans les sections précédentes. Il convient également de retirer de la base, des contrats résiliés qui pourraient ajouter un biais, ainsi ceux associés à deux entreprises sont exclus des bases de modélisation. Ces contrats étaient très sinistrogènes et présentaient un poids déstabilisant le portefeuille. Ils représentaient 2% des primes pour 20% de la sinistralité, avec des primes de plus d'un million d'euros tandis que le montant maximum de primes aujourd'hui est de 400 000 €. Ces contrats étant résiliés et arrivés à échéance, les risques qui leur sont associés ne sont plus présents dans le portefeuille, d'où le choix de supprimer ces contrats et les sinistres associés dans les bases dédiées.

Pour cette approche, les garanties RC Bât et RCD seront modélisées séparément. Pour chacune des garanties, il faut disposer d'une base fréquence et de plusieurs bases sévérité. Ces bases sont obtenues grâce à une jointure entre les bases sinistre et risque. Les mécanismes de création de ces bases seront décrits dans la suite de cette partie.

5.1.1 Création des bases fréquence

La base fréquence permet de modéliser l'occurrence des sinistres, il est alors possible de savoir, si pour un exercice donné, un contrat a été sinistré, ainsi que le nombre de sinistre qu'il a engendré. Dans cette base, les sinistres en cours, clos ou sans suites sont gardés, ainsi que tous les contrats à l'exception de ceux mentionnés précédemment. Deux bases fréquences sont créées, l'une pour la garantie RCD et une autre pour la garantie RC Bât. L'élaboration de ces bases est expliquée ci-dessous :

Base Fréquence RCD : La spécificité d'un contrat présentant une garantie décennale est que celui-ci peut être sinistré plusieurs années après la date d'ouverture du chantier. Ce délai entre la date réglementaire d'ouverture de chantier et la date de survenance du sinistre est appelé '*délai de survenance*'. Sur la base des sinistres, cette variable supplémentaire est ajoutée, en calculant la différence entre la date de survenance et la date d'ouverture de chantier. Sur l'historique de sinistralité, il est constaté que 99,66% des sinistres présentent un délai de survenance compris entre 0 et 16 ans. Ainsi, pour les 341 sinistres dont ce délai n'est pas compris dans l'intervalle donné précédemment, on l'imposera au délai maximal soit 16 années.

Ce délai de survenance doit également être ajouté dans la base risque. Pour cela, chaque contrat sera associé à chacune des 17 années possibles de survenance, ainsi pour un exercice donné, un contrat peut être présent jusqu'à 17 fois. Enfin une condition est imposée telle que, la somme de l'année d'exercice et du délai de survenance soit inférieure à 2022. Le schéma suivant permet de résumer le mécanisme d'ajout de la variable *délai de survenance* dans la base risque :

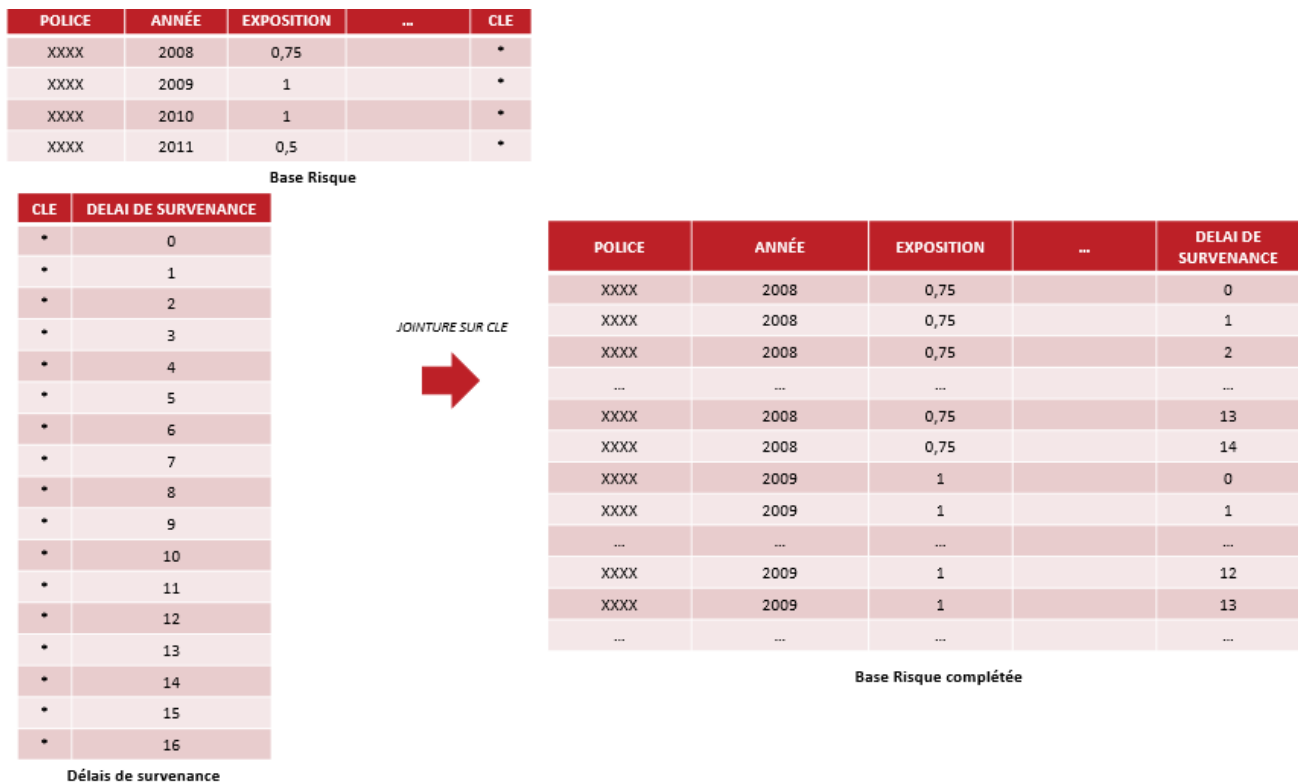


FIGURE 5.1 – Mécanisme d'ajout du délai de survenance sur la base risque

Il convient désormais de joindre les bases risque et sinistres selon le numéro de police, l'année et le délai de survenance pour obtenir la base fréquence. La création de cette base est explicitée dans le schéma ci-dessous :

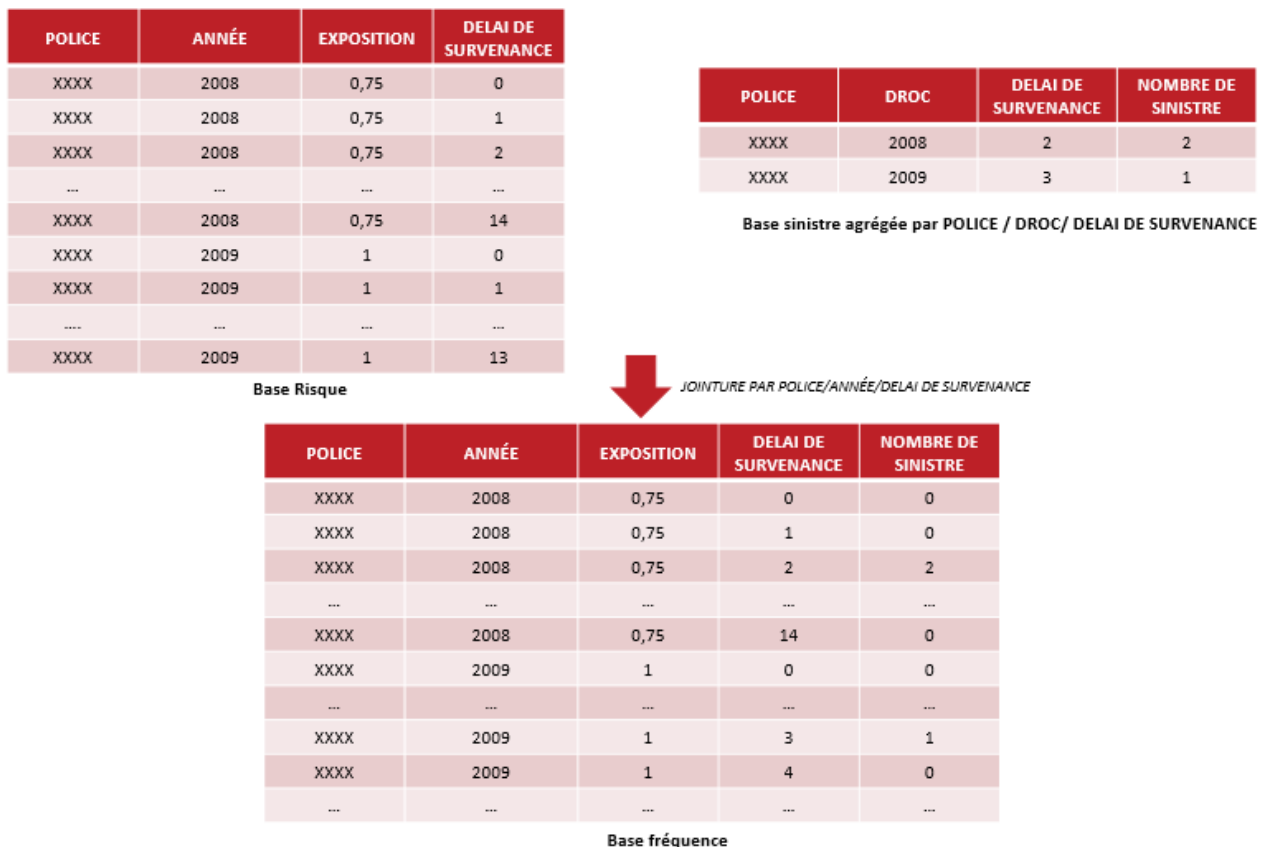


FIGURE 5.2 – Mécanisme de construction de la base fréquence

Lors de la jointure finale toutes les lignes sont conservées, ainsi la table créée compte 1 668 289 lignes. La variable réponse est le nombre de sinistres et le poids est l'exposition, la base comporte également 23 variables explicatives.

Base Fréquence RC Bât : Pour les sinistres RC Bât, le délai de survenance n'est pas considéré (il est égal à 0 pour l'ensemble des données), il n'est donc pas nécessaire de modifier la base risque. La construction de la base de fréquence est similaire à celle réalisée pour la garantie RCD, cependant la jointure entre les tables risque et sinistre se fait selon le numéro de police et l'année seulement. La variable réponse est le nombre de sinistres et le poids est l'exposition. La base créée comporte 182 175 lignes et 24 colonnes, avec la variable à expliquer, le nombre de sinistres, le poids et 22 variables explicatives.

5.1.2 Création des bases sévérité

Dans la partie 3.4, les sinistres ont été segmentés en plusieurs tranches selon leur coût. L'appartenance à une tranche traduit la sévérité d'un sinistre. Afin de modéliser différemment ces tranches, il faut créer une base pour chaque tranche de sévérité.

Dans un premier temps, sur les bases sinistres RCD et RC Bât, une filtration est réalisée afin de ne conserver que les sinistres de charge strictement positive, parmi ces sinistres, ne seront gardés ensuite que ceux clos ou en cours avec une date de survenance maximum de 2019. Ensuite, quatre colonnes sont créées correspondant aux quatre tranches de sévérité déterminées précédemment. Le montant total d'un sinistre est ensuite affecté à une tranche, par exemple un sinistre de 130 000€ sera affecté à la deuxième tranche. Les sinistres ne sont pas écrêtés afin d'éviter le problème d'effet de bord dans les tranches avec les sinistres de coût supérieur, on aurait, par exemple, une concentration des autour de 30 000€ pour la première tranche. Pour créer les différentes bases sévérité, il faut ensuite joindre les bases risque et sinistres selon le numéro de police, l'année, et le délai de survenance pour la RCD. Un exemple de jointure pour la création de la base sévérité RCD pour la première tranche est explicitée ci-dessous :

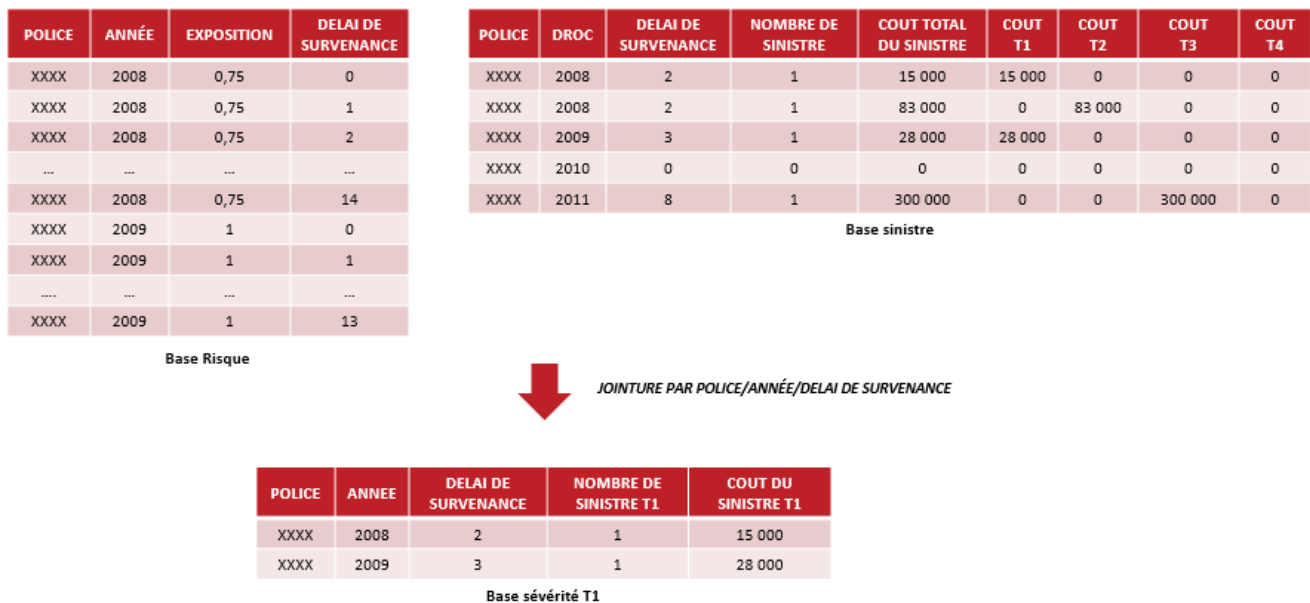


FIGURE 5.3 – Construction de la base sévérité pour les sinistres de la première tranche

Après la jointure, les lignes où le nombre de sinistres vaut 1 et où la charge de la tranche considérée est non nulle, sont gardées. Cette procédure est répétée pour créer les 8 bases de sévérité.

Le tableau ci-dessous résume les dimensions des tables de sévérité créées. Les tables RCD se composent de la variable à expliquer (la charge du sinistre), et des 24 variables explicatives, tandis que les tables RC Bât comptent une variable explicative en moins car le délai de survenance n'est plus considéré. Les tables présentent des dimensions différentes qui ne permettront pas d'utiliser les mêmes approches pour déterminer leur sévérité et leur fréquence.

	Sévérité	Nombre lignes
RCD	Tranche 1	31 356
	Tranche 2	1 871
	Tranche 3	344
	Tranche 4	4

	Sévérité	Nombre lignes
RC Bât	Tranche 1	26 384
	Tranche 2	786
	Tranche 3	206
	Tranche 4	6

La première tranche contient plus de 90 % des observations, ainsi il sera possible de la modéliser avec les modèles linéaires généralisés mais aussi avec des méthodes de Machine Learning. Pour les tranches 2 et 3, des méthodes paramétriques seront utilisées pour limiter le biais et la variance, et afin d'être le plus prédictif possible. Les quatrièmes tranches, pour la base de données RCD et pour celle RC Bât, ne comptent qu'un nombre très faible de lignes, ainsi une approche simple *coût moyen × nombre* sera réalisée.

5.1.3 Création des bases Sans-Suite

Plus de 30% des sinistres RCD et 20% des sinistres RC Bât, sont des sinistres sans-suite de charge nulle. Cependant ces sinistres sont pris en compte dans la modélisation en fréquence, ainsi un montant leur est associé avec la modélisation de la sévérité. Afin de ne pas surestimer la charge associée à ces sinistres, une nouvelle colonne est ajoutée dans la base sinistre présentant comme modalités 0 ou 1 permettant de déterminer si le sinistre est un sans suite de charge nulle.

Pour effectuer la modélisation de l'occurrence d'un sinistre sans suite de montant nul, une base a été créée selon le schéma suivant :

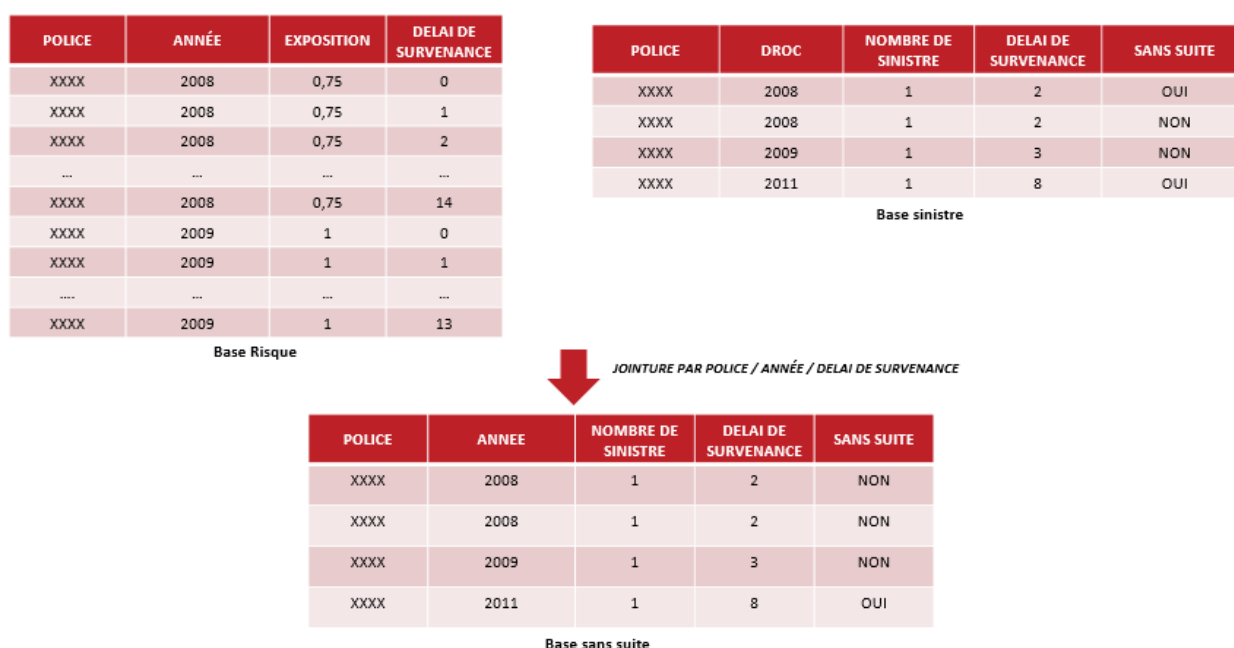


FIGURE 5.4 – Construction de la base des sans suites pour la garantie RCD

Dans un premier temps, sur la base sinistre, une nouvelle variable est créée permettant d'indiquer si le sinistre considéré est un sans-suite de charge nulle ou non. Cette variable nommée *Sans Suite* possède deux modalités : "OUI" ou "NON". Une jointure, sur les variables police, année et délai de survenance, est ensuite réalisée entre les bases risques et sinistres.

Après celle-ci, ne sont gardées que les lignes où le nombre de sinistres est égal à 1. La même procédure est répétée pour les sinistres RCD et pour ceux liés à la garantie RC Bât, permettant ainsi de créer deux bases de données sur les sans-suites. Elles comptent :

- 56 942 lignes pour la garantie décennale
- 34 882 lignes pour la garantie responsabilité civile bâtiment

Les différentes bases sont désormais construites, elles peuvent donc être utilisées pour réaliser les modélisations. Elles contiennent semblablement les mêmes variables explicatives, cependant avec la mise en place de la modélisation, l'étude des corrélations et les éventuels regroupement et création de classes, elles sont amenées à évoluer et possiblement de manières différentes. Les parties suivantes vont détailler la mise en place de la modélisation et les choix opérés.

5.1.4 Regroupement et création de classes des modalités

La modélisation sous Emblem impose que l'ensemble des variables explicatives soient sous forme qualitative, ainsi toutes les variables quantitatives sont transformées telles que chaque valeurs prises par une de ses variables deviennent une modalité. Cependant cette modification implique que pour certaines variables, comme le chiffre d'affaire par exemple, le nombre de modalités soit très élevé, il convient alors de créer des classes pour ces variables.

Les variables *année*, *délai de survenance* ne sont pas modifiées, les valeurs prises par celles-ci sont les modalités de la variable. De même, pour les variables *Nombre d'activité* et *Nombre de contrats en construction* qui prennent un nombre limité de valeurs, respectivement 23 modalités (de 1 à 22 activités, et la valeur "Non renseigné" pour les valeurs manquantes) et 9 modalités (de 1 à 9 contrats).

L'ancienneté du contrat est comprise entre 0 et 68 ans. Au dessus de 25 ans d'ancienneté, le nombre de contrats devient très faible. Il est alors décidé de créer une modalité "ancienneté supérieure à 25 ans", les anciennetés antérieures sont laissées selon une valeur unitaire. *L'ancienneté de l'entreprise* est quand à elle comprise entre 0 et 144 ans, sa distribution est représentée ci-dessous :

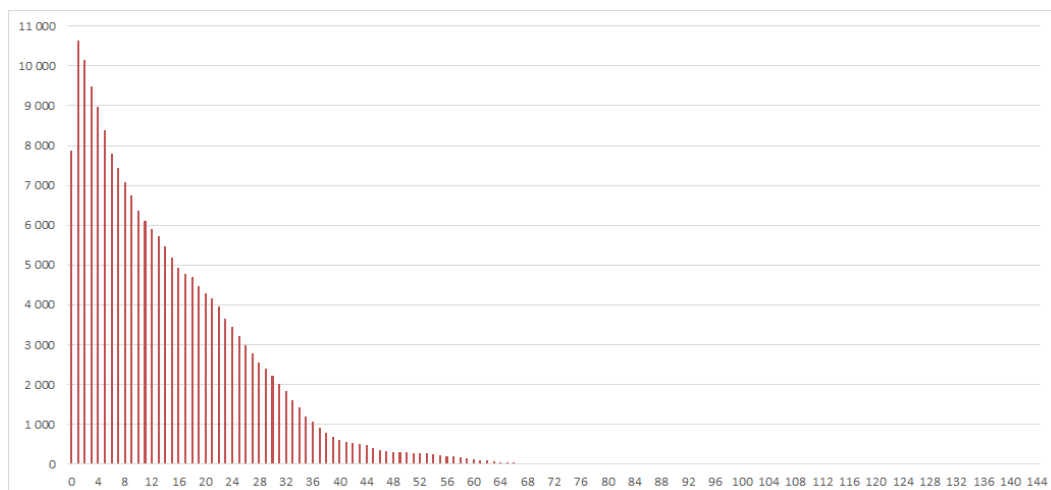


FIGURE 5.5 – Concentration des anciennetés des entreprises avant 60 ans

Elle possède des valeurs manquantes, non représentées sur le graphique précédent, qui sont renommées "Non renseigné". Après 60 ans d'ancienneté, certaines valeurs ne sont pas retrouvées et seulement 0,23 % des données sont considérées, ainsi une borne maximale est décidée, et tous les contrats ayant une ancienneté supérieure à 60 ans seront considérés dans la même modalité.

Pour le *pourcentage de chiffre d'affaires sous traités*, les valeurs 0 et 100 sont considérées comme deux modalités. Des classes sont ensuite créées permettant de rassembler les pourcentages de 10 en 10, les premières classes sont donc]0;10],]10;20]... Une modalité pour les valeurs manquantes est également conçue prenant la valeur "Non renseigné".

Certaines entreprises peuvent posséder plus d'une centaine de contrats au sein de Generali, en particulier si elle possède des contrats de Flottes automobiles. Ceci étant des exceptions, moins de 1% des assurés possèdent plus de 60 contrats. Les modalités de la variable *Nombre de contrats* supérieures à ce nombre seront regroupées. Entre 30 et 60 contrats des classes de 10 contrats sont créées afin de regrouper les données.

L'*effectif* d'une entreprise au sein du portefeuille peut dépasser 200 ETP. Les contrats pour lesquels le nombre de salariés déclarés sont les plus importants sont regroupés ensemble, afin de créer les modalités :]50;100],]100;200], "> 200". Environ 85 % des entreprises comptent entre 1 et 10 employés, ces valeurs resteront inchangées et considérées comme des modalités. Enfin des classes de 10 ETP seront définies pour des effectifs compris entre 11 et 50. La modalité "Non renseigné" pour les valeurs manquantes est également retrouvée.

Enfin, le *chiffre d'affaires* est la variable possédant le plus de modalités. Des classes de tailles différentes sont produites afin de retranscrire les différences apparaissant dans le portefeuille et pour les plus petites structures. Dans un premier temps, les chiffres d'affaires inférieurs à 10 000 euros sont regroupés, puis ceux de chiffre d'affaires supérieurs mais restant inférieurs à 50 000. Ensuite, des classes de largeur 50 000 sont créées jusqu'à 300 000 euros. Après cela, les chiffres d'affaires jusqu'à 500 000 euros sont présents dans la même classe, et enfin deux classes de largeur 250 000 euros sont créées pour arriver à 1 million d'euros. Au dessus de cette valeur, les classes sont les suivantes :]1 000 000;2 000 000],]2 000 000;5 000 000],]5 000 000;10 000 000],]10 000 000;50 000 000] et "> 50 000 000". La modalité "Non renseigné" se retrouve également, pour cette variable, pour les valeurs manquantes.

Les variables initialement qualitatives, telles que le *type de contrat*, le *type d'assuré* et le *multi-équipement* ne présentent que deux modalités et ne sont donc pas modifiées. Les variables donnant des informations sur l'activité ont été créées afin de regrouper les données. Ainsi la variable *Regroupement activité* est un regroupement du code activité, tandis que le *type d'œuvre* est défini à partir du regroupement d'activité. Enfin la variable sur l'activité aggravante n'est pas modifiée. L'unique variable qualitative modifiée est la *région* pour corriger certains noms, et l'orthographe, et pour que la forme de la modalité apparaisse correctement lors de la modélisation.

Lors de la modélisation, une modalité de chaque variable est mise en référence, ainsi les résultats obtenus seront à comparer avec cette modalité. Pour l'ensemble des variables, sauf pour l'année, la modalité la plus représentée est celle de référence.

5.1.5 Corrélations

Avant de commencer la modélisation, il est primordial de regarder les corrélations existantes entre les différentes variables. Elles permettent de déterminer et de quantifier le lien entre deux variables. Ces dépendances peuvent impacter la réalisation de la modélisation, et affaiblir les performances

des modèles ou algorithmes. Il faut donc fixer un seuil au dessus duquel la corrélation entre deux variables est considérée importante, et par conséquent il faudra supprimer une des variables. Dans notre cas d'étude, le seuil sera fixé à $\rho = 0.5$.

L'utilisation de la modélisation sous Emblem fait intervenir seulement des variables qualitatives, ainsi pour mesurer leur dépendance il sera possible d'utiliser le test du chi-deux ou le V de Cramer. Leurs définitions sont données ci-dessous :

Le test du chi-deux

Il permet de tester l'indépendance entre deux variables à partir du tableau de contingence construit pour ces variables. Le tableau de contingence étant un tableau permettant de présenter l'effectif présent pour la croisement des modalités de 2 variables, notons les A et B. Son expression est la suivante :

$$\chi^2 = \sum_i^M \sum_j^N \frac{n_{ij} - \frac{n_i n_j}{n}}{\frac{n_i n_j}{n}}$$

Où M et N sont respectivement le nombre de modalités de la variable A et de la variable B

n est le nombre total d'observations

n_i est le nombre d'observations présentant la modalité i pour la variable A

n_{ij} est le nombre d'observations présentant la modalité i pour la variable A et la modalité j pour la variable B

La statistique suit une loi du χ^2 à $(M - 1)(N - 1)$ degrés de liberté. Cet indicateur permet de déterminer l'indépendance ou non de deux variables mais ne permet pas de déterminer l'intensité de la liaison existant entre celles-ci.

Le V de Cramer

Cet indicateur basé sur le test du chi-deux permet de déterminer l'intensité de la liaison existante entre deux variables. Il se définit par la formule :

$$V = \sqrt{\frac{\chi^2/n}{\min(M - 1, N - 1)}}$$

Où (M, N) est la dimension du tableau de contingence, correspondant donc au nombre de modalités pour la variable A et pour la variable B.

Contrairement au χ^2 , cet indicateur reste stable si l'on augmente la taille de l'échantillon de valeur. L'indépendance entre les deux variables étudiées est d'autant plus forte que le V de Cramer est proche de zéro, il vaut 1 en cas de complète dépendance.

Pour observer ces valeurs, il est possible de représenter un tableau des corrélations rassemblant les corrélations entre toutes les variables. Sinon, le graphique des corrélations suivant peut être affiché, plus la ligne entre deux variables est épaisse, plus la corrélation est forte.

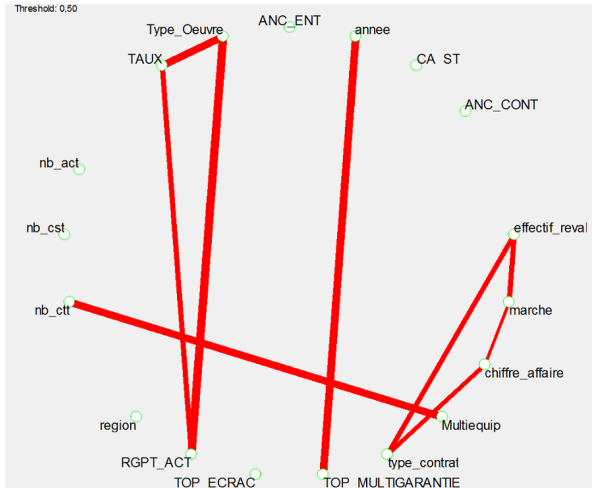


FIGURE 5.6 – Graphique des corrélations entre les variables au seuil de 0,5

Le graphique ci-contre permet de faire apparaître des liaisons importantes entre les variables pour la modélisation du taux des sans suite en RC Bât. Les variables Taux, type d'oeuvre et regroupement d'activités sont fortement corrélées, ce qui est logique. En effet, les variables type d'oeuvre et taux ont été créées à partir du regroupement d'activité. Ainsi pour décider de la variable à garder dans le modèle, il conviendra de regarder l'information apportée par chacune de ces variables ainsi que de comparer les déviations, et critères AIC et BIC pour l'ajout de chacune de ces variables dans le modèle.

Une forte corrélation est également constatée entre le fait d'être multi-équipé et le nombre de contrats possédé par un assuré. Ce qui est toujours cohérent car la variable du nombre de contrat est utilisée pour déduire la valeur du multi-équipement. Tout comme les variables année et Top_multigarantie.

Enfin la définition du type de contrat (Forfaitaire ou Révisable) ou du marché (Entreprise ou ProPE) dépend du chiffre d'affaires et/ou de l'effectif de l'assuré ainsi on retrouve bien des corrélations importantes entre ces quatre variables.

Les corrélations détaillées ci-dessus se retrouvent dans l'ensemble des bases de données créées. Cependant les variables gardées lors des modélisations seront différentes, cela dépendra de l'apport d'informations de chacune pour la prédiction. Par exemple, la variable "Type_Oeuvre" sera conservée pour la modélisation de la fréquence RCD tandis que la variable "TAUX" sera retenue pour les sans-suites RCD.

5.2 Mesure de la performance prédictive d'un modèle

5.2.1 Base d'apprentissage et de test

L'évaluation d'un modèle implique de réserver une partie des données pour tester le modèle. Cette séparation des données permet de limiter les effets de surapprentissage qui peuvent survenir lorsque le modèle est évalué sur les mêmes données que celles qui ont servi à son ajustement, et pouvant donner une performance artificiellement bonne.

La base de données initiale de modélisation est divisée en deux sous bases : la base d'apprentissage et la base de test comprenant respectivement 80% et 20% des données. Pour réaliser cette division, une variable supplémentaire a été ajoutée lors de la réalisation des bases de données pour la modélisation. Nommée "Z_Random", il s'agit d'une variable pouvant prendre comme valeur les entiers entre 1 et 20 inclus et repartis uniformément entre les lignes.

Elle permet de sélectionner les lignes à mettre dans les bases d'apprentissage et de test. En effet, pour chaque modalité de la variable *Z_Random*, le poids ainsi que la moyenne de la valeur à expliquer sont étudiés, afin de répartir convenablement les observations. Les bases sont alors créées de sorte que les poids entre l'apprentissage et le test soient proportionnels et que les moyennes des valeurs prédites soient équivalentes.

5.2.2 Critère de sélection d'un modèle

Lors de la modélisation, il convient de choisir les variables explicatives qui seront pertinentes pour l'étude. Il s'agira alors de mesurer la qualité d'ajustement du modèle et ainsi de vérifier l'adéquation entre les observations et les estimations. Pour ce faire, il faut s'assurer de la significativité des paramètres, et ensuite de la qualité d'ajustement du modèle. Plusieurs indicateurs sont alors utilisés, une définition ainsi que leur utilisation seront détaillées ci-dessous.

La Déviance

Elle compare le modèle estimé au modèle dit *saturé*, ce dernier étant le modèle possédant autant de paramètres que d'observations et estimant donc exactement les données. La déviance se calcule comme étant la différence des log-vraisemblances des deux modèles :

$$D = -2(L - L_{sat})$$

D suit asymptotiquement une loi du χ^2 à $n - p$ degrés (n étant le nombre de paramètres du modèle saturé, et p le nombre de paramètres du modèle estimé). Un test de rejet ou d'acceptation du modèle peut donc être construit selon si la déviance est significative ou non. Elle permet donc de juger de l'ajustement d'un modèle.

La statistique de Pearson

Pour calculer l'ajustement d'un modèle, un autre test peut être utilisé, il s'agit de la statistique de Pearson calculée par :

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{Var(Y_i)}$$

Chaque terme de la somme est un écart au carré entre la valeur observée et la valeur prédite divisée par la variance de Y_i , où les Y_i sont les variables explicatives.

Elle admet asymptotiquement la même loi que la déviance asymptotique, il s'agit d'une loi de χ^2 , et la règle de décision est donc également la même que celle de la déviance.

Critères AIC et BIC

D'autres critères numériques peuvent être utilisés pour comparer plusieurs modèles. Ainsi il est possible de calculer, pour un modèle précis, le critère d'information d'Akaike (AIC) :

$$AIC = 2k - 2\ln(L)$$

où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle.

Lors de l'ajout d'une variable dans le modèle, le critère AIC est calculé et il est comparé à celui du modèle sans cette nouvelle variable. L'ajout sera considéré pertinent si la valeur de l'AIC est plus faible. Lorsque la pertinence du retrait d'une variable veut être testée, il est possible d'utiliser le critère d'information Bayésien (BIC) :

$$BIC = 2\ln(L) + \ln(n)k$$

Où n est le nombre d'observations dans la base étudiée et k le nombre de paramètres à estimer.

Ce critère est plus pénalisant lorsque le nombre de paramètres est élevé, à cause de la présence du terme $\ln(n)k$. C'est pour cela que ce critère est préféré à l'AIC dans le cas du retrait des variables, et inversement lors de l'ajout de variables.

5.3 Modèles Linéaires Généralisés

5.3.1 Théorie

Les modèles linéaires généralisés (GLM) sont une généralisation du modèle linéaire. Celui-ci permet de prédire une variable $Y = (Y_1, \dots, Y_n)$, dite **variable réponse** ou **variable à expliquer**, à partir de p **variables explicatives** $(X_1, \dots, X_p)$, il prend la forme :

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ji} + \epsilon_i \quad i = 1, \dots, n$$

Où n est la taille de l'échantillon, les β sont les **paramètres du modèle**, des coefficients de régression qu'il faudra déterminer, les ϵ_i sont **les erreurs**, ce sont des variables gaussiennes indépendantes et identiquement distribuées, centrées, présentant de l'homoscedasticité et de l'incorrélation.

Les GLMs sont des modèles expliqués qui présentent trois composantes :

- Y est le **vecteur à expliquer** où les Y_i sont indépendants et leur loi appartient à la famille exponentielle, c'est-à-dire que leur loi peut se mettre sous la forme suivante :

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

Où θ est le paramètre canonique ou paramètre de la moyenne, $\theta \in \mathbb{R}$

ϕ est le paramètre de dispersion, $\phi \in \mathbb{R}$

a est une fonction définie sur les réels et non nulle

b est une fonction définie sur les réels et deux fois dérivable

c est une fonction définie sur $\mathbb{R}^2$

Pour une variable Y dont la loi peut s'écrire sous la forme exponentielle ci-dessus, nous obtenons :

$$\mu = \mathbb{E}(Y) = b'(\theta)$$

$$\sigma^2 = \text{Var}(Y) = a(\phi)b''(\theta)$$

- Pour chaque Y_i , la valeur du p-uplet $(X_{1i}, \dots, X_{pi})$ de variables décrivant Y_i est connue.
- Une **fonction lien** g , réelle, déterministe et strictement monotone telle que $g_n(\mathbb{E}(Y)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$

Le GLM s'écrit donc $g(\mathbb{E}(Y_i)) = g(\mu_i) = x_i' \beta = \eta_i$, où η est le **prédicteur linéaire**. La fonction lien g fournit donc une relation fonctionnelle entre le prédicteur linéaire et la moyenne de la fonction de distribution : $\mathbb{E}(Y) = \mu = g^{-1}(X^T \beta)$.

Les principales fonctions lien sont présentées dans le tableau suivant

Loi	Fonction lien	Formule
Normale	Identité	$g(y) = y$
Poisson	Logarithme	$g(y) = \ln(y)$
Gamma	Inverse	$g(y) = \frac{1}{y}$
Binomiale	Logit	$g(y) = \ln\left(\frac{y}{1-y}\right)$
Binomiale Négative	Logarithme	$g(y) = \ln(y)$

Significativité des paramètres

Lors de l'ajout d'une variable dans le modèle, il convient de déterminer si l'ajout de cette variable est significatif pour l'étude, et d'étudier la significativité des modalités. Ainsi il convient de comparer deux hypothèses différentes :

$$H_0 : \beta = \beta_{H_0} = (\beta_0, \dots, \beta_q)$$

contre

$$H_1 : \beta = \beta_{H_1} = (\beta_0, \dots, \beta_p)$$

avec $0 < p < q < n$.

Différents tests et estimateurs peuvent être utilisés. Deux d'entre eux sont détaillés ci-dessous.

Rapport de vraisemblance

Il permet de comparer deux modèles : l'un possédant toutes les variables significatives voulues et l'autre étant un modèle réduit du premier. Il se calcule selon la formule :

$$\lambda = \Lambda = \frac{L_{\hat{\beta}_{H_0}}(y)}{L_{\hat{\beta}_{H_1}}(y)}$$

où la différence de déviance

$$\delta = D_0 - D_1 = 2(\ln(L_{\hat{\beta}_{H_0}}(y)) - \ln(L_{\hat{\beta}_{H_1}}(y))) \sim \chi^2_{p-q,ddl}$$

est une mesure de l'apport des $p - q$ variables supplémentaires.

Test de Wald

Ce test permet de tester la significativité d'un paramètre et donc la nullité de l'estimateur du paramètre, ainsi l'hypothèse du test est :

$$H_0 : \beta_j = 0$$

La statistique de test se calcule par :

$$z = \frac{\hat{\beta}_j}{\text{Var}(\hat{\beta}_j)}$$

Cette statistique tend en loi vers une variable de loi gaussienne centrée et réduite quand le nombre d'observations tend vers l'infini.

5.3.2 Sélection des variables et simplification

La méthode adoptée pour la modélisation avec les modèles linéaires généralisés est la même pour chaque base. Les étapes suivantes sont réalisées sur la base d'apprentissage :

- Dans un premier temps, une étude des corrélations est réalisée : lorsque deux variables sont corrélées, celle ayant la corrélation la plus faible avec la variable à expliquer est supprimée de la base de modélisation.
- Ensuite, la méthode **Forward** est utilisée : le modèle initial est un modèle minimal (sans variable ou avec les premières variables que l'on veut retrouver dans le modèle), puis les variables sont ajoutées une à une en fonction de celles qui apportent le plus d'information au modèle et qui permettent de réduire le critère AIC (celui-ci est moins sensible que le critère BIC lors de l'ajout de variables).

A chaque ajout de variables, il faut vérifier la bonne définition des bêtas de chaque modalité. Si ce n'est pas le cas, des créations de classes ou des regroupement peuvent être effectués.

- Lorsqu'il n'est plus possible d'ajouter des variables avec la méthode Forward, une seconde méthode est utilisée, il s'agit de la méthode **Backward**. Désormais, le modèle initial est celui composé de l'ensemble des variables ajoutées avec la méthode Forward, et elles sont retirées une à une en fonction de celles qui apportent le moins d'information au modèle et qui permettent de réduire le critère BIC.

La bonne définition des bêtas est toujours vérifiée à chaque suppression de variables.

- Éventuellement, les méthodes Forward et Backward peuvent à nouveau être utilisées afin d'obtenir le modèle final avec les valeurs les plus basses pour chacun des critères AIC et BIC.

Des vérifications sont effectuées sur la base de test afin de confirmer la bonne définition des bêtas, et ainsi la bonne réalisation des regroupements et des ajouts de variables.

Désormais les modèles obtenus pour chaque base vont être détaillés, en expliquant les étapes effectuées et les choix faits :

MODÈLES DE FRÉQUENCE :

Les modèles permettront de prédire la fréquence des sinistres. Celle-ci correspond au nombre de fois où un sinistre est survenu durant la période d'exposition du client :

$$fréquence = f_k = \frac{NB_SIN}{Exposition}$$

Responsabilité Civile Décennale

Dans la distribution du nombre de sinistres, un pic est observé en 0, ainsi plus de 95 % des lignes ne comptent aucun sinistres. Cette masse en 0 s'explique par l'ajout de l'ensemble des délais de survenance pendant lesquels un assuré peut déclarer un sinistre. Cependant certains contrats présentent, pour une année et un délai de survenance donnés, plus de 300 sinistres.

La modélisation sera effectuée avec une fonction lien **log**, qui permet d'obtenir un modèle multiplicatif, la structure d'erreur sera une **loi Binomiale Négative** et la métrique d'ajustement sera **la statistique de test de Pearson**.

Les résultats de la procédure Forward pour la fréquence RCD sont donnés dans la figure ci-dessous.

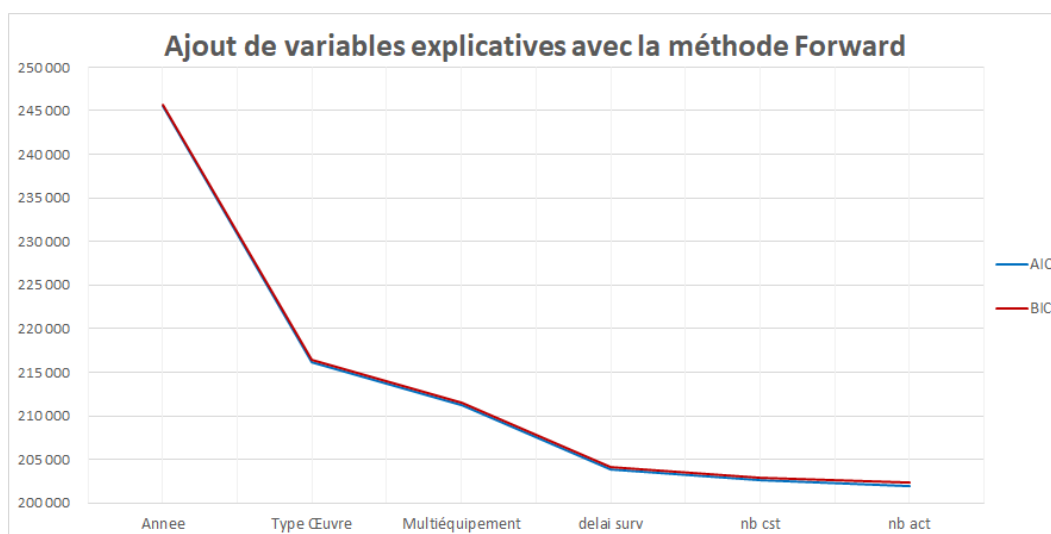


FIGURE 5.7 – Ajout de six variables explicatives avec la méthode Forward, pour la modélisation de la fréquence RCD

Lorsque les variables explicatives du modèle ont été choisies, il faut déterminer si les coefficients sont significatifs. Pour cela, Emblem fournit un tableau avec différents indicateurs, celui-ci est présenté ci-dessous pour les premières variables de la modélisation :

Nombre de paramètres	Nom	Valeur	Erreur Standard	Erreur Standard (%)	Poids	Poids (%)	Exp(Valeur)
1	Mean	-3,74	0,0854	2,29	1 202 338	100,00	0,0238
2	Multiequip (Non)	-0,31	0,0198	6,36	369 336	30,72	0,7323
-	Multiequip (Oui)				833 001	69,28	
-	nb_cst (1)				1 016 830	84,57	
3	nb_cst (2)	-0,01	0,0289	280,98	163 972	13,64	0,9898
4	nb_cst (3)	0,40	0,0597	14,78	20 056	1,67	1,4971
5	nb_cst (4)	0,58	0,2427	42,15	1 087	0,09	1,7787
6	nb_cst (5)	-2,20	1,1083	50,37	198	0,02	0,1108
7	nb_cst (6)	-1,06	1,4122	132,80	76	0,01	0,3453
8	nb_cst (7)	-3,71	3,5557	95,89	70	0,01	0,0245
9	nb_cst (8)	-3,66	4,2057	115,06	40	0,00	0,0259
10	nb_cst (9)	-3,65	6,9867	191,28	10	0,00	0,0259
11	Type_Oeuvre (AUTRE)	-1,41	0,9368	66,28	355	0,03	0,2433
12	Type_Oeuvre (Gros Oeuvre)	0,79	0,0190	2,41	338 056	28,12	2,2005
13	Type_Oeuvre (Non renseigné)	0,00	0,0000	2,41	150 669	12,53	1,0000
14	Type_Oeuvre (Prof Lib)	0,01	0,1393	945,04	5 392	0,45	1,0148
-	Type_Oeuvre (Second Oeuvre)				689 930	57,38	
15	Type_Oeuvre (Travaux tous corps d'état)	1,75	0,0697	3,97	11 137	0,93	5,7820
16	Type_Oeuvre (Fabrication-Négoce)	0,22	0,1153	53,49	6 800	0,57	1,2407

Pour chaque variable, la modalité de référence est affichée en bleu dans le tableau, l'individu possédant chacune de ces modalités est considéré comme l'individu de référence du modèle. Sa fréquence de sinistre est l'exponentielle de la valeur moyenne, ce qui équivaut à 2,38 %.

Afin de juger de la significativité d'un coefficient, *l'erreur standard* et *le pourcentage d'erreur standard* peuvent être regardés. Plus ce pourcentage est élevé, moins la prédiction est bonne. Il existe un code couleur pour aider à l'interprétation de ce pourcentage, si il est inférieur à 50% il sera de couleur vert, entre 50% et 75% de couleur gris et au-delà de 75% de couleur rouge. Si toutes les erreurs standards sont supérieures à 75% alors les modalités de la variable ne sont pas significatives.

Afin de déterminer les regroupements pouvant être effectués et comment les associer, il est également possible de regarder le graphique des observations et des prédictions. Est représenté ci-dessous une comparaison des observations (en rose) et des prédictions (en vert) pour la variable indiquant le nombre de contrats en assurance construction.

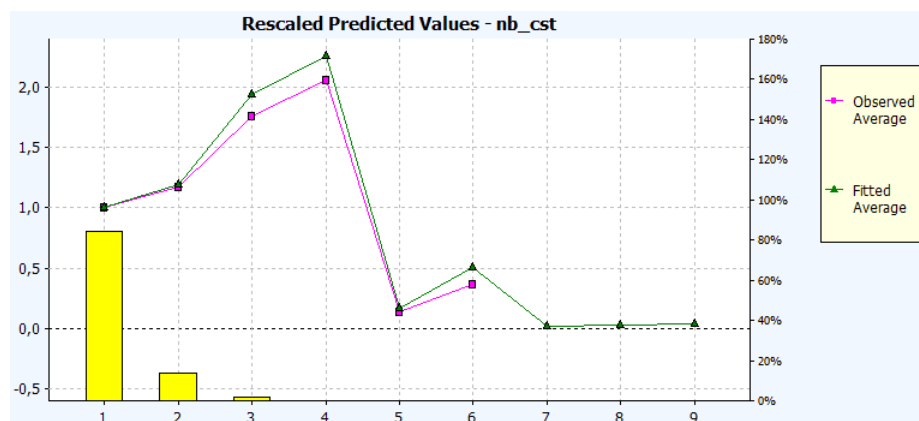


FIGURE 5.8 – La faible représentation des modalités détériore les erreurs

La variable donnant le *nombre de contrats en assurance construction* est ajoutée, les modalités sont regroupées afin d'obtenir les suivantes : "1 contrat" (modalité de référence) ou "2 ou plus". En effet, les dernières modalités ne sont pas très peu représentées, et la création de regroupements avec moins de modalités ne permet pas de réduire suffisamment l'erreur standard. La détention de plusieurs contrats en assurance construction apparaît comme un facteur influençant la fréquence de sinistralité en la faisant augmenter.

Le type d'œuvre apparaît comme étant le facteur le plus influent de la sinistralité, et est la première variable ajoutée dans le modèle. Il traduit le type d'activité réalisé par l'assuré, qui sera la source de sinistralité. La modalité de référence, étant celle la plus représentée est le "Second Œuvre", celle-ci sera regroupée avec les modalités : "Prof Lib" et "Autre" qui ne sont pas significatives et qui ne présentent qu'un faible nombre d'observations. Enfin, les autres modalités ("Gros Œuvre" et "Travaux tous corps d'état") seront traitées séparément tandis que les valeurs manquantes seront regroupées avec les activités de fabrication et de négoce (FABNG). Ces regroupements ont des impacts différents sur la fréquence, ainsi les travaux tous corps d'états sont ceux faisant augmenter le plus la fréquence. Ensuite vient les activités de gros œuvre et enfin la fabrication - négoce avec les activités non renseignées.

Une autre variable donnant des informations sur l'activité sera également ajoutée au modèle, il s'agit du *nombre d'activités*. Elle possède de nombreuses modalités qui ne sont pas toutes significatives, ainsi des regroupements sont effectués, les modalités retenues dans le modèle sont : "de 1 à 5", "de 6 à 10", "plus de 10" et "Non renseigné". Ces regroupements permettent de baisser le critère AIC d'environ 900 points et le BIC de 200 points. Plus le nombre d'activités est élevé et plus les coefficients de modélisation sont élevés.

Ensuite, la variable sur le *multi-équipement* est ajoutée, celle-ci ne possède que deux modalités "Oui" (qui est celle de référence) ou "Non". Enfin, la dernière variable ajoutée dans le modèle est le *délai de survenance*, ajoutée est troisième position, elle permet de faire baisser les deux critères de plus de 7 000 points chacun. Pour cette dernière variable, la pic de fréquence se constate pour un délai de 4 ans, avant cette valeur les coefficients de la modélisation augmentent, puis diminuent linéairement après cette valeur et jusqu'à 15 ans.

Désormais les regroupements sont effectués, le tableau fourni par Emblem peut être à nouveau consulté et tous les pourcentages d'erreur standard sont bien inférieurs à 50 % :

Nombre de paramètres	Nom	Valeur	Erreur Standard	Erreur Standard (%)	Poids	Poids (%)	Exp(Valeur)
1	Mean	-3,74	0,0854	2,29	1 202 338	100,00	0,0238
2	Multiequip (Non)	-0,31	0,0198	6,36	369 336	30,72	0,7323
-	Multiequip (Oui)				833 001	69,28	
-	New nb_cst #2 (1)				1 016 830	84,57	
3	New nb_cst #2 (2 +)	0,07	0,0271	36,55	185 508	15,43	1,0771
4	New Type_Oeuvre (Gros Oeuvre)	0,79	0,0190	2,40	338 056	28,12	2,2134
5	New Type_Oeuvre (FABNG - NR)	0,24	0,1160	49,34	157 469	13,10	1,2650
-	New Type_Oeuvre (Second Oeuvre + Prof Lib + AUTRE)				695 677	57,86	
6	New Type_Oeuvre (Travaux tous corps d'état)	1,74	0,0698	4,01	11 137	0,93	5,7076

L'ajout de variables supplémentaires n'apporterait pas plus d'informations et le modèle ne s'adapterait pas aussi bien sur notre base de test, ainsi pour la fréquence RCD il n'y aura pas d'autres variables ajoutées. De plus, l'utilisation de la méthode Backward ne permet pas de distinguer des variables à retirer du modèle.

Responsabilité Civile Bâtiment

La distribution du nombre de sinistres pour la garantie RC Bât est semblable à celle de la garantie RCD. Ainsi un pic est constaté en 0 car tous les contrats ne sont pas sinistrés chaque année. 85 % des lignes de la base ne sont associées à aucun sinistre. La fréquence maximale obtenue pour un assuré pendant une période d'exposition est de 355.

La modélisation sera effectuée avec une fonction lien **log**, la structure d'erreur sera une **loi Binomiale Négative** et la métrique d'ajustement sera **la statistique de test de Pearson**.

La procédure **Forward** est mise en place, et la première variable ajoutée est *l'année*. Ensuite est ajoutée la variable indiquant le *nombre de contrats* possédé par un assuré. Cet ajout permet de réduire l'AIC et le BIC, passant, respectivement de 86 229 à 82 598 et de 86 387 à 83 083. Des regroupements sont nécessaires car toutes les modalités ne sont pas significatives, ceux-ci permettent de réduire les deux critères de plus de 300 points chacun. Le nombre de contrats en assurance construction influence la fréquence, l'appartenance à un groupe de modalités élevé traduit d'une plus grande fréquence de sinistres.

Une troisième variable est ajoutée dans le modèle, il s'agit du *nombre de contrats en assurance construction*. Cette variable, comme pour la garantie RCD, ne possède pas toutes ses modalités significatives. Ainsi les modalités retenues sont : "1 contrat" et "2 et plus". L'ajout de cette variable et les regroupements associés permettent d'obtenir un modèle final avec un AIC de 82 066 et un BIC de 82 284. Posséder plus d'un contrat en assurance construction accroît la fréquence de sinistralité.

Le tableau fourni par Emblem sur la significativité des paramètres permet de constater que les regroupements effectués sont significatifs.

Nombre de paramètres	Nom	Valeur	Erreur Standard	Erreur Standard (%)	Poids	Poids (%)	Exp(Valeur)
1	Mean	-2,5319	0,0584	2,3	132 969	100,00	0,0795
-	New nb_cst (1)				124 012	93,30	
2	New nb_cst (2 +)	0,4383	0,0417	9,5	8 958	6,70	1,5501
-	New nb_ctt (1)				38 727	29,10	
3	New nb_ctt (2)	0,3121	0,0373	12	22 422	16,90	1,3663
4	New nb_ctt (3-5)	0,476	0,0309	6,5	45 578	34,30	1,6096
5	New nb_ctt (6-10)	0,7555	0,0365	4,8	20 764	15,60	2,1286
6	New nb_ctt (11-19)	1,0973	0,0585	5,3	4 708	3,50	2,9960
7	New nb_ctt (20 +)	1,6269	0,1239	7,6	771	0,60	5,0882

Aucune variable supplémentaire n'est ajoutée au modèle, elle détériorerait les critères AIC et BIC, et n'apporterait pas d'informations supplémentaires. Enfin, la méthode Backward n'indique pas de variables à supprimer du modèle.

La modélisation de la fréquence par les GLMs a permis de déterminer différentes variables explicatives pour les responsabilités civiles décennale et bâtiment. Deux sont communes aux deux modélisations : l'année et le nombre de contrats en assurance construction. Cette dernière variable influe de la même manière la survenance d'un sinistre pour les deux garanties, un nombre plus élevé de contrats entraîne une fréquence plus importante.

MODÈLES DE SÉVÉRITÉ :

Désormais les coûts des sinistres sont modélisés et les tranches de sévérité seront considérées séparément. Chaque sous-partie détaillera les études menées ainsi que les choix opérés lors de la modélisation. La quatrième tranche de sinistralité ne sera pas modélisée car elle ne présente que peu de sinistres.

Première tranche de sinistralité :

La modélisation des coûts de la première tranche sera effectuée avec une fonction lien **log**, ainsi qu'une **loi Gamma** pour la structure d'erreur et la métrique d'ajustement sera **la Déviance**.

Pour la première tranche, il convient de représenter, dans un premier temps, la répartition des coûts pour les deux garanties. En effet, lors de la présentation des données initiales dans la troisième partie, plusieurs pics de charges s'observaient. Ils correspondent à des coûts forfaitaires d'ouverture de dossier ou d'expertise, et doivent être retirés de la base de modélisation car ils introduisent un biais. Il faut donc vérifier la présence de ces coûts pour pouvoir les traiter ensuite. Les graphiques ci-dessous représentent la distribution initiale des coûts des sinistres.

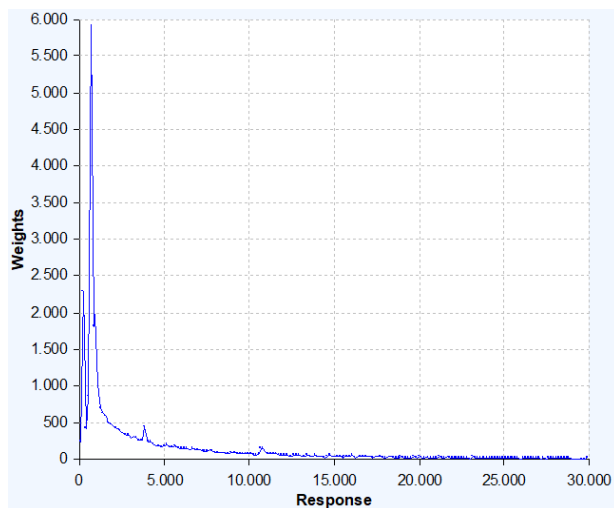


FIGURE 5.9 – Distribution des sinistres (RCD) de la première tranche de sévérité

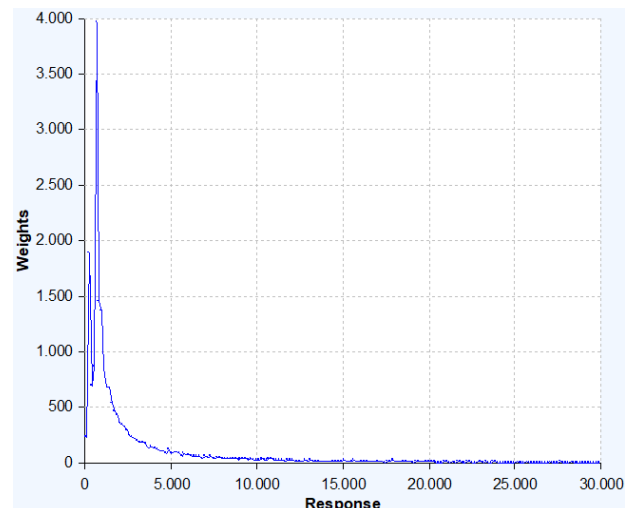


FIGURE 5.10 – Distribution des sinistres (RC Bât) de la première tranche de sévérité

Le premier graphique montre la présence d'un pic autour de 1 000 euros pour la garantie RCD. Les montants fortement représentés dans la base sont les mêmes que ceux présentés dans la partie 3. Ainsi, les pics sont présents pour des charges de 200 euros, entre 600 euros et 800 euros, et de 1000 euros. Comme pour la RCD, plusieurs pics de charges avant et autour de 1 000 euros sont constatés pour la RC Bât. Ainsi les charges surreprésentées sont exclues de la base de modélisation de la première tranche de sévérité de cette garantie.

La suppression des sinistres dont les charges sont des coûts forfaitaires permet d'obtenir une nouvelle distribution des sinistres, avec une réduction du nombre de sinistres ayant une charge inférieure à 1 000 euros. Cependant la forme de la distribution reste semblable à celle de la base contenant l'ensemble des observations.

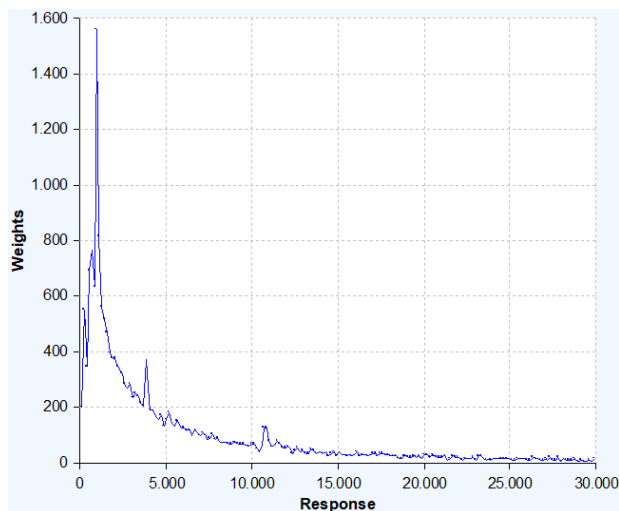


FIGURE 5.11 – Première tranche de sévérité RCD retraitée

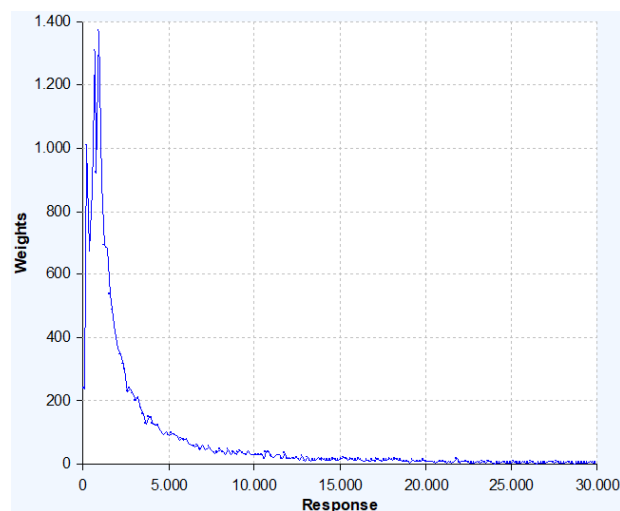


FIGURE 5.12 – Sévérité RC Bât de la première tranche retraitée

Ces nouvelles distributions de sinistres devront être prises en compte lors de la réalisation du modèle final. En effet, les montants moyens ont augmenté et ceux estimés pourrait être surestimés. L'ajout d'un coefficient pour répondre à cette problématique sera détaillé dans la septième partie.

Ces suppressions de sinistres entraînent une réduction des dimensions des bases de sévérité pour la première tranche. Le nouveau nombre de lignes est le suivant :

- 24 623 pour la Responsabilité Civile Décennale
- 21 376 pour la Responsabilité Civile Bâtiment

Responsabilité Civile Décennale :

↓ Délai de survenance
 ↓ Regroupement activités
 ↓ Année
 ↓ Nombre de contrats
 ↓ Effectif
 ↓ TOP_ECRAC

La procédure Forward a permis d'ajouter 6 variables dans le modèle de prédiction de la sévérité de la première tranche. Les variables *année* et *TOP_ECRAC*, celle-ci ne possédant que deux modalités, n'ont pas nécessité de regroupements supplémentaires. Pour les autres variables des traitements seront nécessaires afin d'avoir des modalités significatives.

Les variables supplémentaires ajoutées dans le modèle sont le code de regroupement de l'activité, le nombre de contrats, l'effectif et le délai de survenance. Pour cette dernière variable, les observations sont les suivantes :

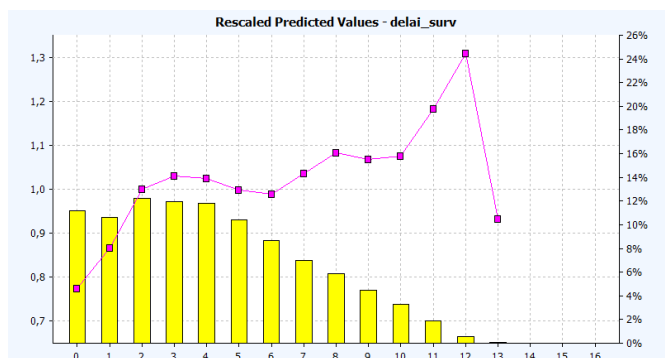


FIGURE 5.13 – Une tendance polynomiale sur les modalités

Une forme polynomiale des observations est constatée, hormis le dernier point correspondant à la modalité de 13 ans qui se détache du reste. De plus, l'ensemble des modalités ne sont pas significatives et leur grand nombre détériore les critères AIC et BIC. Ainsi il est décidé d'approximer cette variable par un courbe polynomiale $X^1 + X^2 + X^3 + X^6$, et toutes les modalités à partir de 12 ans de délai de survenance sont regroupées.

La prédiction obtenue est la suivante :

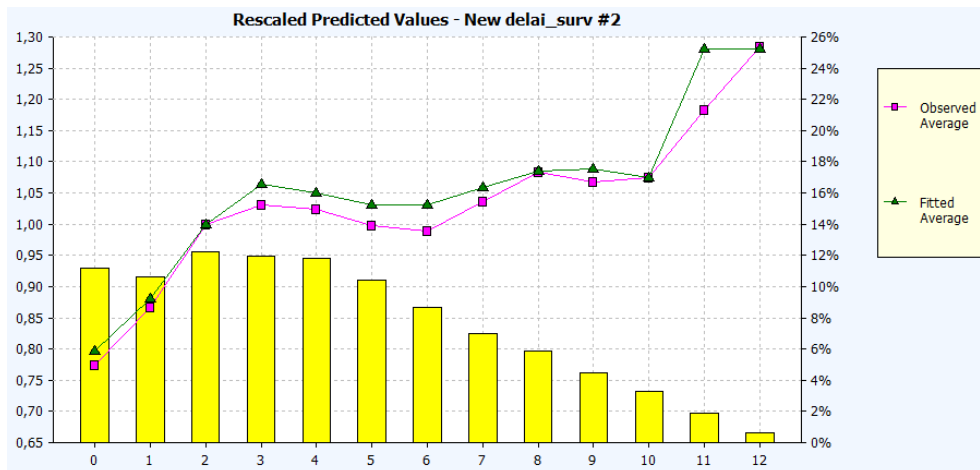


FIGURE 5.14 – Approximation du délai de survenance par une fonction polynomiale

Les coefficients de la régressions augmentent avec le délai de survenance, ainsi un coût croissant avec le délai de survenance est constaté. Cette observation est cohérente car une sinistralité plus tardive implique une inflation cumulée sur plusieurs années et donc un coût d'un sinistre supérieur. De plus, une déclaration plus tardive peut aussi être associée à des sinistres d'une volumétrie plus importante et de ce fait d'un coût supérieur.

Responsabilité Civile Bâtiment :

- + Taux
- + Année
- + Chiffre d'affaires sous traité
- + Ancienneté du contrat
- + nombre d'activités
- + Effectif
- + Région
- + Nombre de contrats
- + Ancienneté de l'entreprise
- Ancienneté de l'entreprise
- Ancienneté du contrat
- Région
- + Ancienneté du contrat

Dans un premier temps, la procédure Forward permet d'ajouter 9 variables dans le modèle. Cependant certains ajouts dégradent fortement le critère BIC. Ainsi la procédure Backward est mise en place et met en évidence 3 variables à supprimer, ce sont les anciennetés contrat et entreprise ainsi que la région. Elles présentaient de nombreuses modalités non significatives ainsi que des regroupements difficiles à effectuer. Enfin la procédure Forward est réutilisée et permet d'ajouter à nouveau la variable sur l'ancienneté du contrat. Le modèle final retenu présente donc 7 variables explicatives permettant de donner des informations sur l'entreprise et son activité.

Pour la modélisation de cette première tranche, les variables *année*, *effectif*, *nombre de contrats* ainsi qu'une variable sur le type d'activités de l'assuré sont explicatives pour chacune des garanties. Le délai de survenance, n'étant considéré que pour la RCD, et le branchement à l'ECRAC ne sont présents que dans la modélisation de la garantie décennale. Pour la RC Bât, le nombre d'activité, le chiffre d'affaires sous traité et l'ancienneté du contrat apportent de l'information sur le coût de la sinistralité.

Deuxième tranche de sinistralité :

Désormais, il convient de s'intéresser à la deuxième tranche de sinistralité. Il s'agit des sinistres ayant un coût compris entre 30 000 et 150 000 euros. La distribution de cette tranche pour la garantie RC Bât est représentée ci-dessous. La distribution RCD n'est pas représentée car elle présente la même forme, une différence existe sur le nombre d'observations dans la tranche.

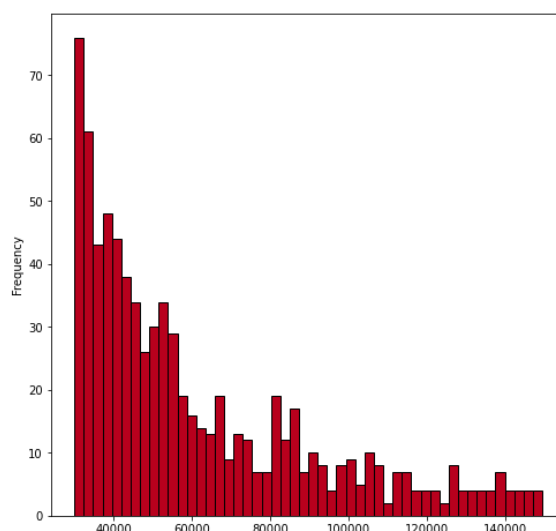


FIGURE 5.15 – Distribution de la deuxième tranche de sévérité RC Bât

La modélisation utilisera une fonction lien **log**, ainsi qu'une **loi Log-normale** pour la structure d'erreur et la métrique d'ajustement sera **la Déviance**.

Les tableaux ci-dessous résument les variables explicatives ajoutées au modèle, pour la responsabilité civile décennale et pour la responsabilité civile bâtiment :

RC Décennale
1. année
2. nombre d'activités
3. Taux

RC Bâtiment	
1. année	4. effectif
2. ancienneté de l'entreprise	5. Taux
3. chiffre d'affaires sous traité	

Les variables explicatives sélectionnées pour le modèle RCD se retrouvent dans le modèle RC Bât. Comme pour les modélisations précédentes, une variable donnant des informations l'activité de l'entreprise est incluse dans les modèle, ici il s'agit de la variable *Taux*.

Troisième tranche de sinistralité :

Enfin, les sinistres appartenant à la troisième tranche sont modélisés, ce sont ceux ayant une charge entre 150 000 et 2 000 000 euros. Désormais, la répartition des observations dans cette tranche sont représentée ci-dessous pour la garantie RCD :

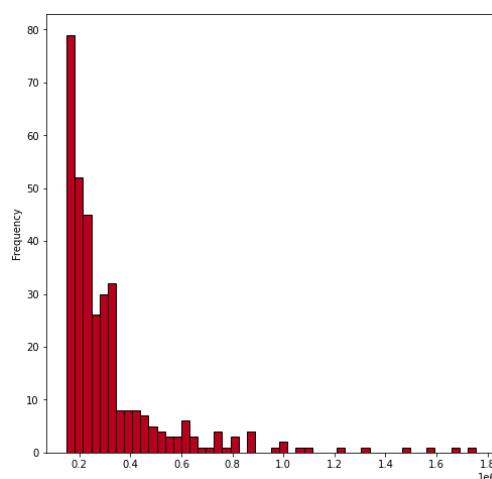


FIGURE 5.16 – Distribution de la troisième tranche de sévérité RCD

Comme précédemment, la distribution des sinistres dans cette tranche pour la garantie RC Bât n'est pas représentée car elle est similaire à celle de la RCD, mais avec un nombre plus faible d'observations.

Pour la modélisation, une fonction lien **log** sera utilisée, ainsi qu'une **loi Gamma** pour la structure d'erreur et la métrique d'ajustement sera **la Déviance**. Cette dernière modélisation présente un nombre plus réduit de variables explicatives ajoutées aux deux modèles.

Pour la responsabilité civile décennale, les variables ajoutées sont *l'année* et *le pourcentage de chiffres d'affaires sous traité*. Toutes les modalités de cette dernière variable n'étant pas significatives, des regroupements sont effectués. Les nouvelles catégories obtenues sont : "0 - [10;30]", "[0;10[-[30;50[" et "[50-100] - Non renseigné"

Pour la responsabilité civile Bâtiment, *l'année* et *l'ancienneté du contrat* sont incluses dans le modèle. Des regroupements sur les modalités de l'ancienneté sont définies : "[0-6]", "[7-15]" et " ≥ 16 "

La modélisation de la sévérité par les GLMs a permis de déterminer différentes variables explicatives pour les responsabilités civiles décennale et bâtiment. Hormis la variable *année*, il n'existe pas de variables communes à l'ensemble des modélisations faites. Pour les deux premières tranches, une variable spécifiant le type d'activité de l'entreprise est incluse dans tous les modèles. Ainsi la variable *Regroupement activités* est présente dans le modèle de coût de la première tranche de sévérité pour la RCD, et la variables *Taux* est incluse dans les trois autres modèles. L'activité de l'assuré semble alors être un élément impactant le coût d'un sinistre, et les activités de gros et second œuvres sont plus susceptibles d'avoir un coût supérieur aux autres activités. Cette constatation semble cohérente avec les travaux de construction, car ces deux catégories regroupent des travaux d'ampleurs supérieures à ceux pouvant être fait dans les autres catégories (structure du bâtiment, construction d'un édifice...) et donc relevant de coûts supérieurs.

Le graphique ci-dessous permet de confirmer un coût moyen de sinistre supérieur pour les gros et second œuvres par rapport aux autres catégories.

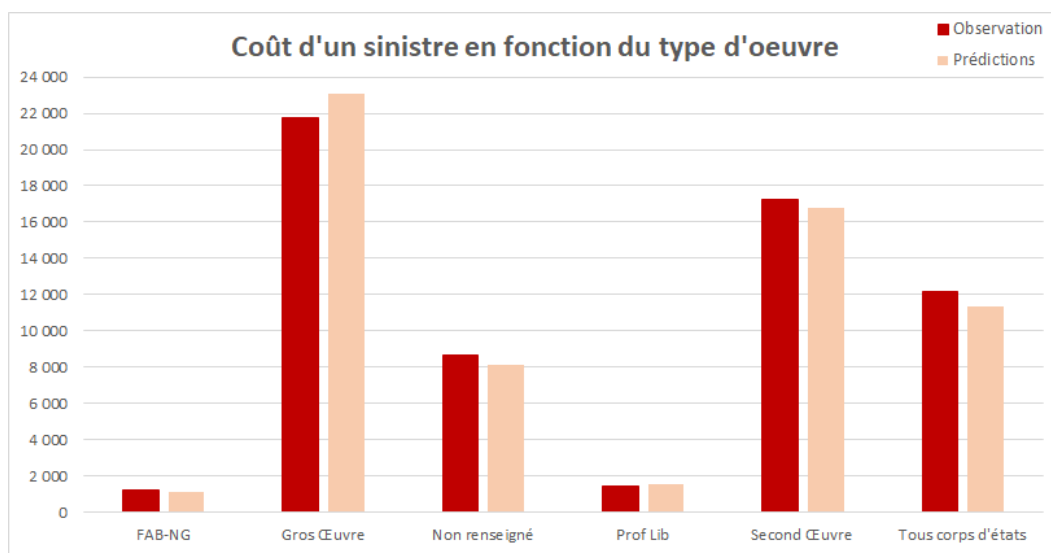


FIGURE 5.17 – Le gros et le second œuvres présentent un coût de sinistre supérieur

Cette tendance supérieure s'observe sur les données réelles et est bien retrouvée sur les prédictions. En moyenne un sinistre gros œuvre a un coût d'environ 22 000 euros, et un sinistre second œuvre de 17 000 euros. A l'inverse, les sinistres liés aux activités de fabrication-négoce et aux professions libérales ont des montants inférieurs à 2 000 euros.

MODÈLES DES SANS SUITES RC BAT ET RCD :

Les bases sans-suites ont été créées afin d'ajouter une information dans la modélisation sur la part des sans-suite dans la sinistralité. La variable à expliquer est une variable binaire prenant 1 comme valeur lorsque le sinistre est un sans suite n'ayant généré aucune charge et 0 sinon.

Le taux de sans-suite dans les deux bases est de :

- 26,9% pour la base RCD
- 14,5 % pour la RC Bât

Lors du lancement de la modélisation, le logiciel Emblem demande de spécifier les différents composants du GLM dans une fenêtre comme ci-contre. Il faut indiquer la fonction lien, la loi de la structure d'erreur et la métrique d'ajustement du modèle (Déviance ou Pearson, en fonction du modèle choisi).

Cette première fenêtre permet également de définir les bases d'apprentissage et de test. Pour cela, il faut définir dans la partie *Sample Set*, les échantillons associés à *Modelling* et à *Validation* grâce à la variable *Z_Random*.

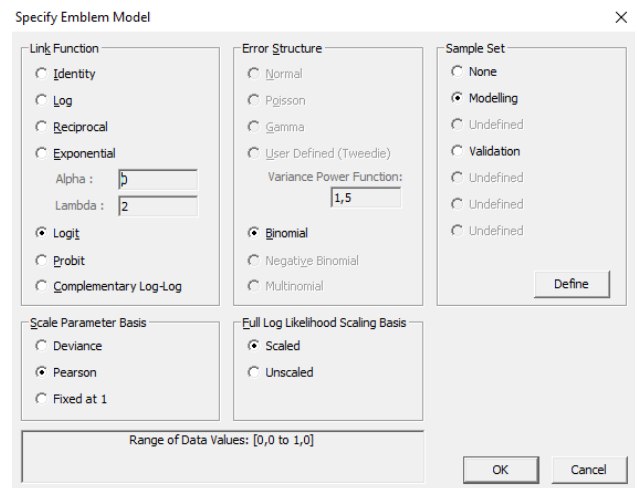


FIGURE 5.18 – Spécification du modèle des sans suites sur Emblem

Dans le cadre de la modélisation du taux de sans-suite, pour les bases RCD et RC Bât, la structure d'erreur sera la **loi Binomial**, la métrique d'ajustement sera la **statistique de test de Pearson** et la fonction lien sera la **loi Logit**, permettant de prédire la réalisation d'une action ou non. Cette dernière loi est utilisée car permettant une meilleure interprétabilité des résultats.

Responsabilité Civile Décennale

Afin de choisir les variables explicatives à ajouter dans le modèle, les méthodes Forward et Backward sont mises en place. Le graphique ci-dessous résume les variables ajoutées et retirées ainsi que les évolutions des métriques AIC et BIC associées :

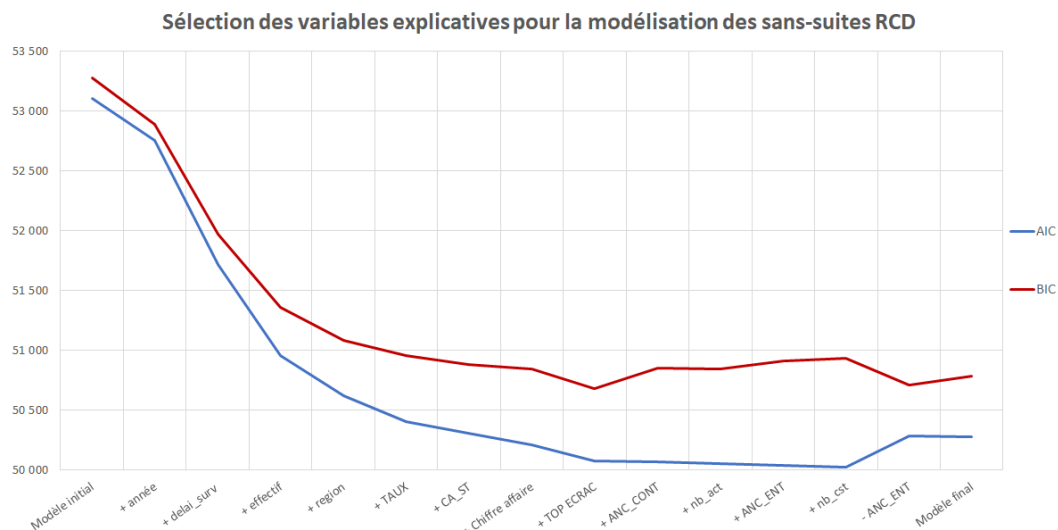


FIGURE 5.19 – Choix des variables retenues dans la modélisation des sans-suites RCD

Les méthodes Backward et Forward permettent d'obtenir un modèle final comprenant 11 variables explicatives. La variable ANC_ENT fait réduire la métrique AIC mais augmenter le BIC lors de son ajout et inversement lors de son retrait. Le choix est fait de la retirer de la modélisation car de nombreuses modalités ne sont pas significatives et les regroupements effectués détérioreraient davantage les critères AIC et BIC.

Le modèle final correspond au modèle où les regroupements entre modalités ont été effectués afin d'avoir la bonne significativité de tous les paramètres.

Responsabilité Civile Bâtiment

Pour cette police d'activité, la méthode Forward est dans un premier temps mise en place, et ensuite la méthode Backward vient compléter la procédure de sélection des variables. Le tableau ci-dessous détaille l'ajout et le retrait des variables :

Sans Suites RC Bât		
Variables	AIC	BIC
Modèle initial	25 181,09	25 189,32
+ annee	22 663,60	22 787,18
+ Type_Oeuvre	22 635,29	22 808,29
+ region	22 608,42	22 872,04
+ ANC_ENT	22 586,11	23 030,98
+ nb_ctt	22 592,14	23 119,39
+ effectif	22 579,75	23 247,05
+ nb_act	22 572,87	23 388,46
- ANC_ENT	22 610,28	22 873,91
Modèle final	22 602,72	22 849,87

La même problématique que pour les sans-suite RCD apparaît pour la variable ANC_ENT, ainsi pour les mêmes raisons il est décidé de ne pas la garder pour la suite de la modélisation. Le modèle final, pour les sans-suites de la garantie RC Bât, correspond au modèle où les regroupements entre modalités ont été effectués afin d'avoir la bonne significativité de tous les paramètres.

Des variables telles que l'effectif, la région et le nombre d'activité se retrouvent dans le modèle de chacune des garanties, à l'inverse d'autres sont plus spécifiques aux modèles. En effet les polices d'activités sont différentes et ne présentent pas les mêmes caractéristiques, il est donc cohérent de constater des différences. Enfin, le branchement à l'ECRAC, influence le taux de sans-suite pour la RCD, cet ajout dans le modèle vient confirmer le fait les hypothèses faites dans la deuxième partie. L'ouverture rapide par ce système viendrait augmenter le nombre de sans-suite pour la garantie décennale.

Les variables explicatives du modèle des sans-suites sont désormais choisies, ainsi que pour les modèles de fréquence et de sévérité. La partie suivante traitera de la validation des modèles, permettant ensuite de créer le modèle final combinant ceux définis précédemment.

5.3.3 Validation des modèles

Des modèles ont été estimés sur les bases d'apprentissage, pour les valider plusieurs moyens peuvent être utilisés : le coefficient Gini, l'analyse des résidus et l'étude des prédictions faites sur la base de test.

Coefficient Gini

Il s'agit d'une mesure de performance de la modélisation. Plus ce coefficient est proche de 1 et plus la prédiction est bonne. Ce coefficient est calculé sur la base de test permettant la validation du modèle, les résultats réels et ceux prédits sont comparés.

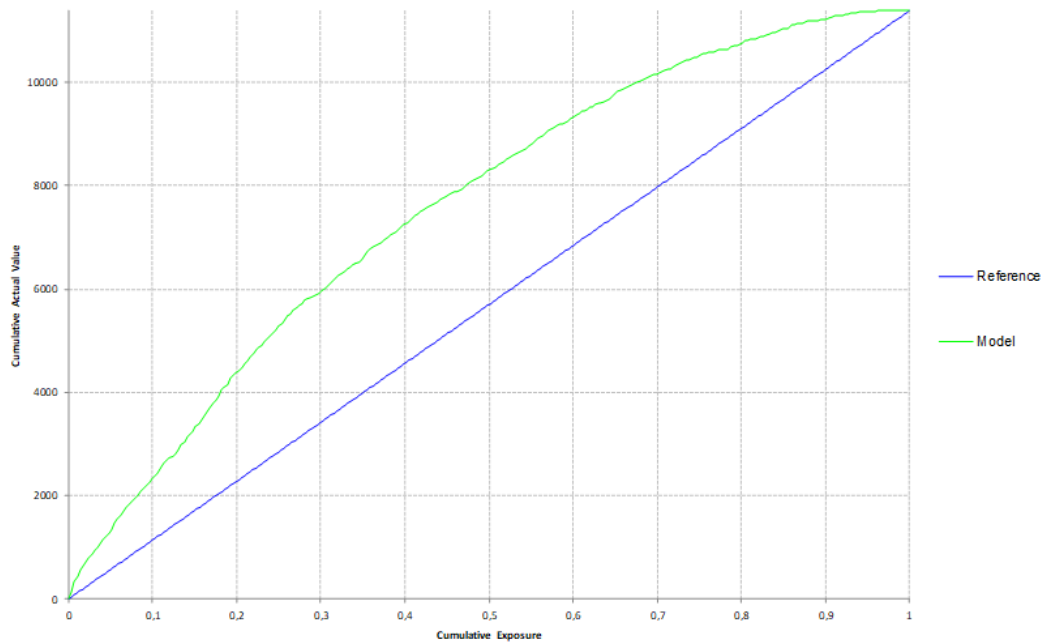


FIGURE 5.20 – Le coefficient Gini de la modélisation de la fréquence RCD est de 0,3302

Le graphique précédent est la courbe de gain, représentant la part de risque (le risque étant le nombre de sinistres observés sur la période d'exposition) en fonction de la part cumulée des assurés.

La diagonale bleue est la référence, il s'agit d'un modèle de prédiction aléatoire ne prenant aucune variables explicatives et la courbe verte est déduite du modèle retenu. Il permet de comparer deux modèles, en choisissant celui s'éloignant davantage de la diagonale.

La modélisation de la fréquence RCD est celle obtenant le coefficient Gini le plus élevé de **0,3302**, tandis que pour la RC Bât il est de **0,2821**. Pour la modélisation des sans suites, le coefficient Gini obtenu se situe autour de **0,21**. Enfin pour les différentes tranches de sévérité, le tableau ci-dessous résume les valeurs obtenues :

Sévérité	Responsabilité Civile Décennale	Responsabilité Civile Bâtiment
Tranche 1	0,1164	0,1187
Tranche 2	0,1344	0,1074
Tranche 3	0,1725	0,1937

Analyse des résidus d'un modèle généralisé linéaire

Les résidus sont des mesures de distance entre les valeurs observées et les valeurs prédites. Plus celui-ci est élevé et plus la prédiction s'éloigne de la valeur réelle. Il existe différents types de résidus pouvant être calculés :

- Les résidus classiques sont la différence entre les valeurs observées et les valeurs prédites :

$$r_i = y_i - \hat{y}_i$$

- Les résidus de Pearson mesurent l'écart entre les valeurs observées et les valeurs prédites pondérées par leur précision estimée par l'écart-type

$$r_i^{Pearson} = \frac{y_i - \hat{y}_i}{\sqrt{Var(\hat{y}_i)}}$$

Une représentation graphique des résidus peut être effectuée, ainsi pour que l'erreur de modélisation soit faible, il faut que tous les résidus soient centrés en 0 en ne possédant pas de structure non aléatoire.

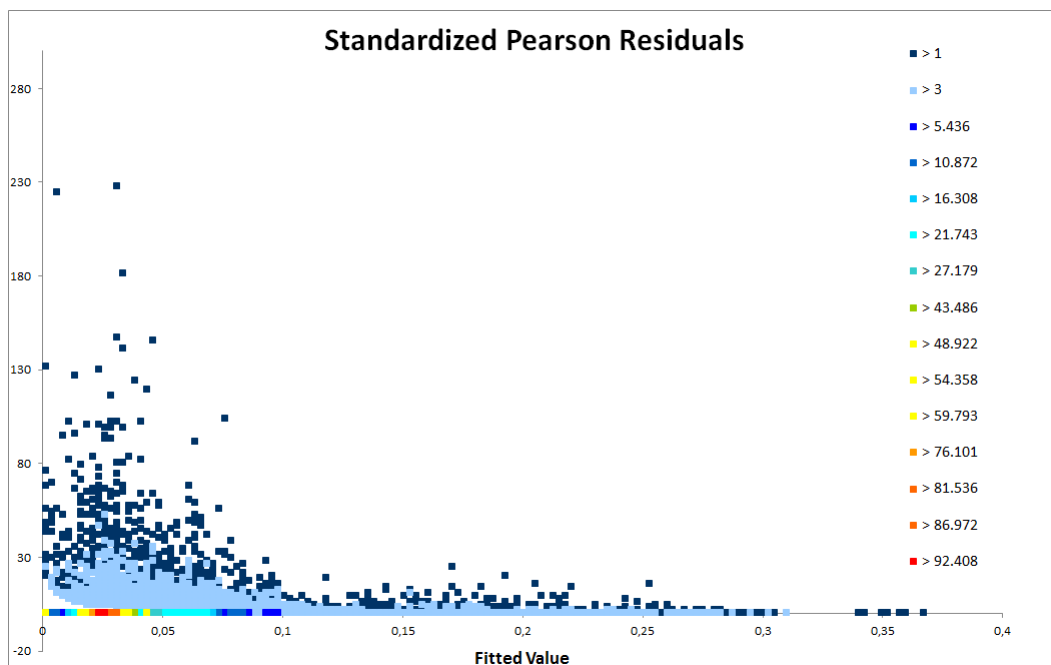


FIGURE 5.21 – Résidus centrés en zéro pour le modèle de fréquence RCD

Dans le modèle de fréquence, la majorité des observations ont une fréquence nulle, correspondant à une année associée à un délai de survenance sans sinistre, et certaines auront une fréquence strictement positive parfois importante.

La plupart des résidus sont centrés en zéro et il n'y a pas de structure non aléatoire.

Les résidus pour le coût moyen se comportent de manière similaire avec une concentration autour de zéro, les points les plus éloignés du 0 sont les sinistres dont la charge est proche du seuil maximal des différentes tranches.

Étude des prédictions

Le modèle doit prédire correctement les modalités de chaque variable que celle-ci soit incluse dans le modèle ou non. Pour vérifier cette adéquation, il est possible de regarder les graphiques Emblem comparant les observations et les prédictions pour chaque variable des différents modèles.

La variable sur l'ancienneté de l'entreprise, est une variable incluse dans la modélisation de la sévérité de la deuxième tranche de sévérité pour la garantie RC Bât. Le graphique ci-dessous permet de constater la bonne adéquation des valeurs observées et de celles prédites.

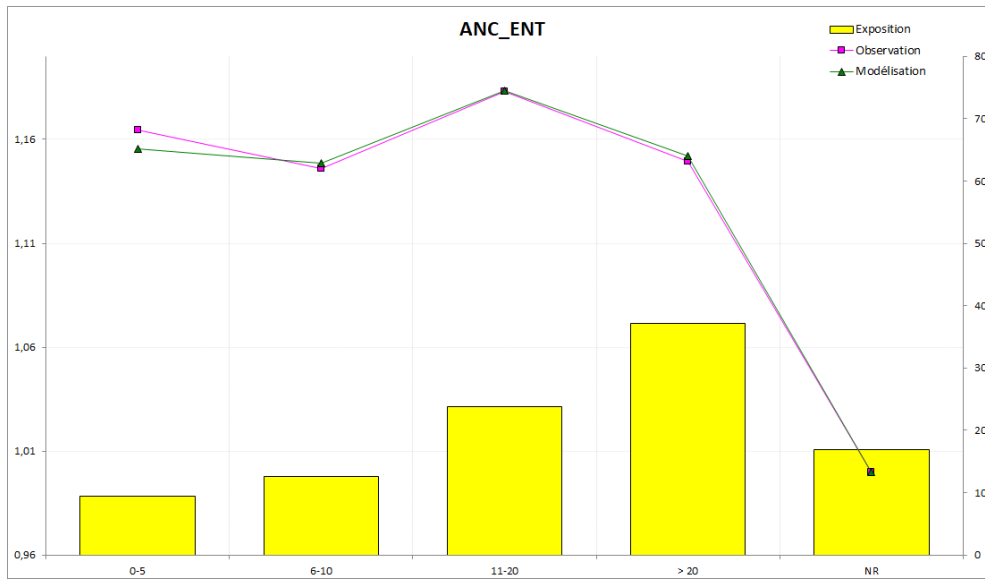


FIGURE 5.22 – Bonne prédiction de la variable ancienneté entreprise incluse dans le modèle

Un bon modèle doit également permettre un bon ajustement des variables non incluses dans le modèle. En effet, celles-ci influent sur les paramètres globaux du modèle. La variable *effectif*, non incluse dans le modèle pour la deuxième tranche de sévérité RCD, est représentée ci-dessous :

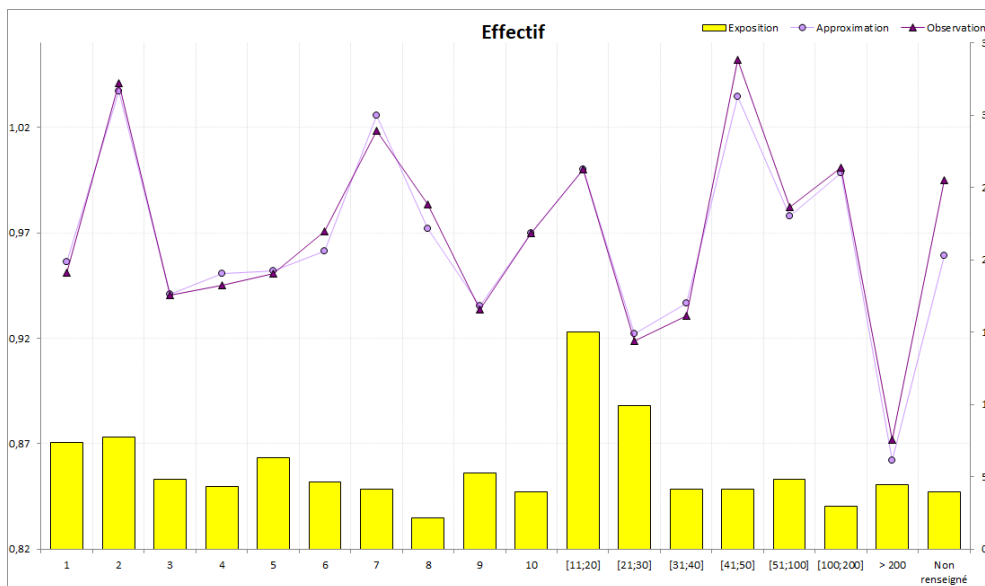


FIGURE 5.23 – Bonne prédiction de la variable effectif non incluse dans le modèle

Un indicateur supplémentaire pour la comparaison et la validation des modèles est le RMSE (Root Mean Square Error). Il s'agit de l'erreur quadratique moyenne des prédictions, définie par la formule suivante :

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

5.4 Modèles de Machine Learning

La méthode classique de modélisation par les GLMs peut être comparée à des méthodes de Machine Learning. Elles ont la capacité d'apprendre par elles-mêmes, et de s'adapter aux données. Ces modèles demandent donc de nombreuses observations, ainsi seules la fréquence et la sévérité de la première tranche seront modélisées avec ces méthodes.

Cette modélisation reprend les premiers regroupements effectués, afin d'avoir les mêmes modalités sur les variables que pour les modèles linéaires généralisés. Ces modèles ne pouvant pas être implémentés sur *Emblem*, ils seront faits sur *Python*. Plusieurs types de modèles en Machine Learning ont été testés, seuls les modèles les plus pertinents seront détaillés. Leurs caractéristiques et les résultats obtenus seront expliqués.

5.4.1 Arbres de décision - CART

L'objectif de l'arbre est de classer des individus, c'est-à-dire de regrouper des individus hétérogènes en classes homogènes de risque pour résumer l'information d'une grande base de données. Cet arbre se compose d'une racine contenant l'ensemble de la population, d'un tronc et de branches qui contiennent les règles de divisions de la population, et se terminent par des feuilles qui contiennent les sous-populations homogènes et fournissent une estimation de la variable à expliquer. N'importe quel individu de la population initiale appartient à une unique feuille.

L'objectif de l'arbre de décision est de déterminer les meilleures modalités à placer à chaque nœud pour que l'arbre soit le plus petit possible et pour que ces prédictions soient bonnes.

La modélisation CART a la particularité d'être une méthode simple à mettre en place, avec des règles de décisions binaires (à partir d'un nœud seulement deux branches peuvent être construites et la règle de décision n'est basée que sur un seul facteur de risque). De plus, il n'est pas nécessaire de disposer d'hypothèses sur les variables, elles peuvent être discrètes ou continues. Ci-dessous est représenté l'arbre de décision obtenu pour la prédiction de la fréquence RC Bât :

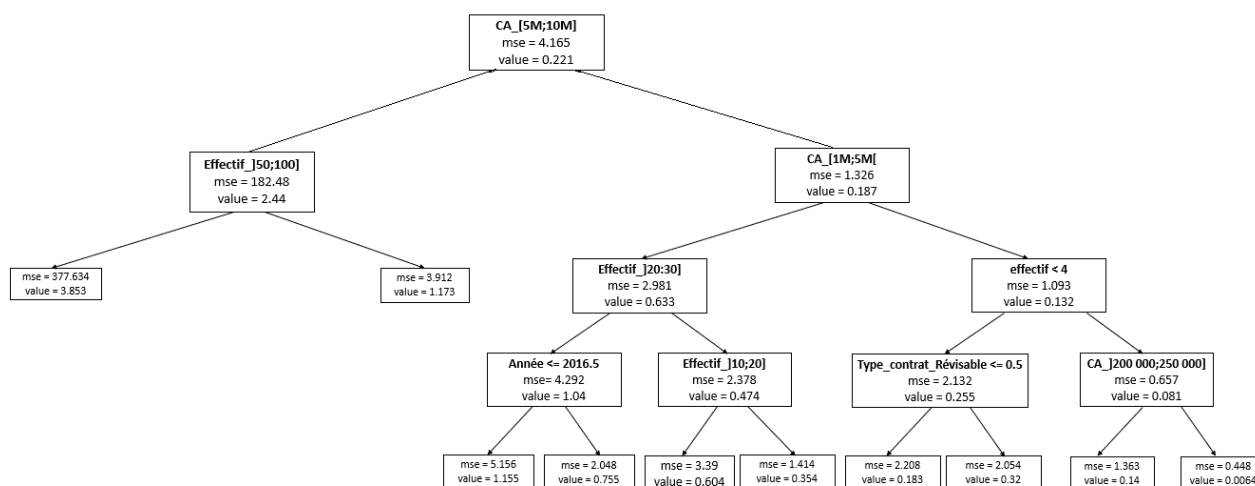


FIGURE 5.24 – Arbre de décision pour la fréquence RC Bât

L'arbre obtenu contient 10 feuilles, correspondant à 10 sous-populations associées à 10 fréquences de sinistralité différentes. Les règles de décision ne concernent que 4 variables différentes : le chiffre d'affaire, l'effectif, le type de contrat, et l'année, mais avec des choix sur des modalités différentes.

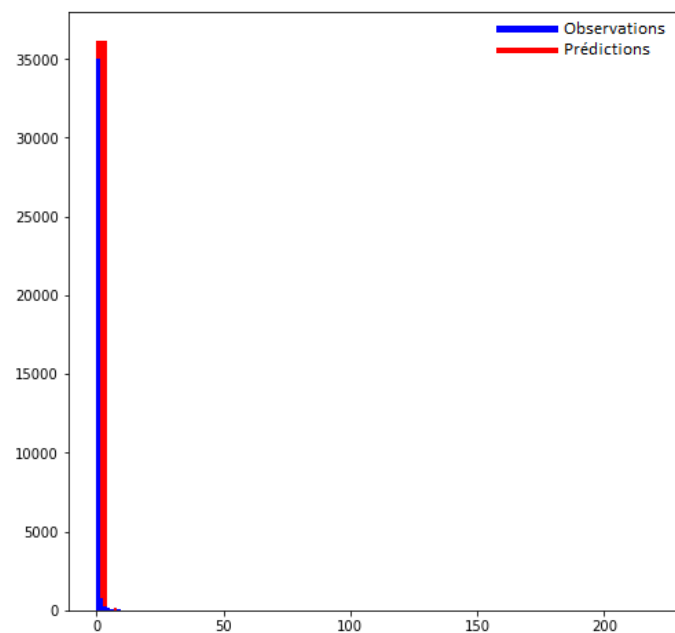
Ainsi, par exemple, un individu n'ayant pas son chiffre d'affaire appartenant à la modalité [5M ; 10M], mais étant inclus dans l'intervalle [1M ; 5M], avec un effectif étant entre 11 et 20 ETP (et ainsi n'étant pas entre 21 et 30 ETP) aura une fréquence annuelle de sinistre de 0,604.

Cette méthode ne permet de définir qu'un nombre fini de prédictions de la variable à expliquer. Ainsi, dans le cas de l'arbre précédent, le fréquence annuelle maximale de sinistre est de 3,855. Cependant, dans les observations ce nombre maximal est de plus de 300 sinistre pour une année, avec 2% des observations ayant un nombre de sinistres supérieurs à 3,855. De plus, près de 85% des observations ne présentent aucun sinistres, et la valeur nulle n'est pas prédite dans ce modèle. Ainsi cette méthode de régression ne permet pas de reproduire la distribution de la fréquence.

Des conclusions similaires pour les modélisations de la fréquence RCD, et des premières tranches de sévérité peuvent être émises. Tous les modèles présentent un nombre limité de valeurs prédites, et une valeur maximale bien inférieure à la réalité.

5.4.2 Forêts aléatoires - Random Forest

Le Random Forest est constitué d'un ensemble d'arbres de décision indépendants. Il utilise plusieurs estimateurs simples ayant une vision parcellaire du problème, ils ont souvent un pouvoir de prédiction de moins bonne qualité. L'ensemble des estimateurs sont réunis afin d'avoir une vision globale, et permet d'obtenir un estimateur final avec une prédiction plus performante.



Le graphique ci-contre permet de comparer la distribution réelle de la fréquence pour la garantie RC Bâtiment (en bleue) et celle de la prédiction du Random Forest (en rouge).

Cette nouvelle modélisation permet d'avoir une meilleure prédiction de la fréquence quand celle-ci est inférieure à 1, et d'obtenir des prédictions de fréquence nulle. Elle permet également d'avoir des fréquences prédites supérieures à 3,855 (valeur maximale des prédictions avec le CART). Il existe cependant un nombre limité d'individus avec une sinistralité supérieure à 10, et la prédiction maximale est de 80 sinistres.

FIGURE 5.25 – Performance du modèle Random Forest pour la fréquence RC Bât

Ainsi, la modélisation avec les forêts aléatoires permet une meilleure prédiction de la fréquence (un tableau récapitulatif avec une comparaison des modèles sera détaillé dans la partie 5.5). Cependant, il existe toujours un maximum de prédictions qui est inférieur à celui des observations.

Comme pour la modélisation CART, des conclusions similaires peuvent être apportées pour la fréquence RCD, ainsi que pour les modélisations des premières tranches de sévérité.

5.4.3 XG Boost

Un dernier modèle de prédiction peut être étudié, il s'agit de l'XGBoost. Correspondant à l'abréviation de *eXtreme Gradient Boosting*, ce modèle est utilisé pour des régressions et classifications, et il est réputé pour sa performance et sa vitesse.

Il s'agit d'un modèle de boosting, il adopte une stratégie adaptative qui permet de réduire la variance et de traiter le problème du biais. En effet, chaque modèle créé est une adaptation du précédent modèle, permettant de combiner les estimateurs afin d'éviter le surapprentissage.

La procédure à mettre en place lors de la réalisation d'un modèle de Boosting est la suivante :

- Estimation d'un premier modèle m_1 , et déduction d'un vecteur d'erreurs ϵ_1
- Estimation d'un deuxième modèle m_2 pour ϵ_1 à partir des données, et déduction d'un vecteur d'erreurs ϵ_2
- Itération k fois de cette méthode pour obtenir l'estimateur agrégé (la valeur de k est un paramètre du modèle).

Pour s'adapter de proche en proche dans le nouveau modèle, un poids plus important est donné aux observations mal prédites précédemment.

L'XGBoost correspond à l'utilisation de l'algorithme du Gradient Boosting avec des arbres de décisions. Le boosting de gradient (Gradient Boosting) est un cas particulier de boosting où les erreurs sont minimisées par l'algorithme de descente de gradient.

Pour ce modèle, il est possible d'étudier l'importance des variables dans la modélisation. Ainsi ci-dessous est représenté le diagramme en barre de l'importance des premières variables dans le modèle.

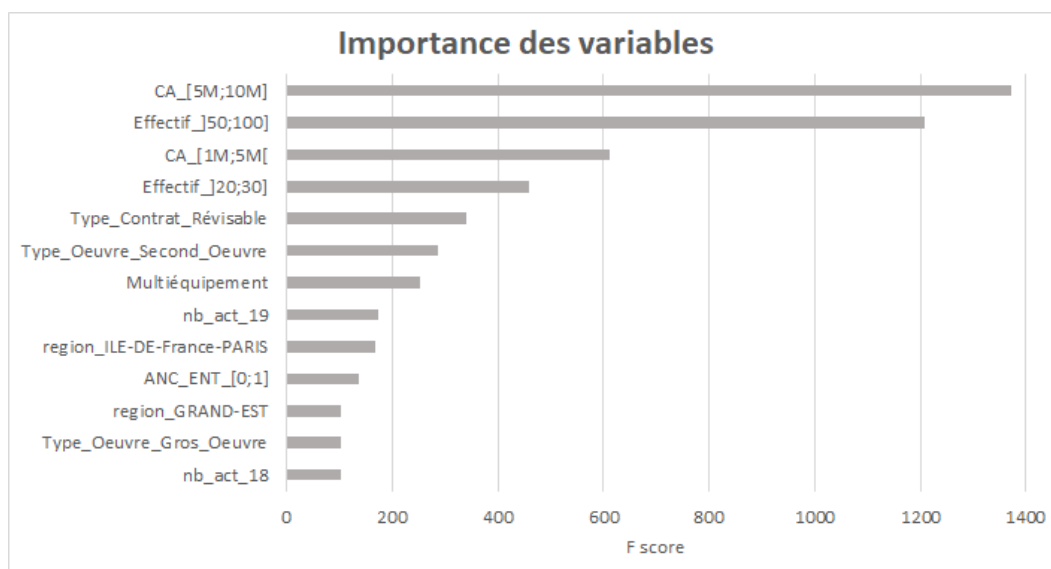


FIGURE 5.26 – Importance des variables dans la modélisation XGBoost de la fréquence RC Bât

Le chiffre d'affaires apparaît comme la variable influençant le plus la fréquence RC Bât, car deux modalités sont présentes dans les trois premières places du graphique précédent. Ensuite vient l'effectif avec également deux sous-variables créées à partir de deux modalités qui sont à la deuxième et à la quatrième place. Enfin, résider dans les régions Ile-de-France et Grand-Est est source d'une fréquence de sinistralité supérieure par rapport aux autres régions, tout comme le fait d'avoir un type œuvre de Second ou de Gros Œuvre, par rapport aux autres types d'œuvres possibles.

Cette modélisation permet d'obtenir des prédictions plus proches des observations, par rapport aux modèles précédents. Ainsi un pic de contrats n'ayant pas de sinistres est constaté en 0, et une décroissance progressive s'opère. Le nombre maximal de sinistres est de 300, ce qui est bien supérieur aux prédictions précédentes et qui s'approche de la réalité.

Cette modélisation permet d'obtenir de meilleurs résultats pour l'étude de la fréquence et de la sévérité de la première tranche. La partie suivante présentera une comparaison des résultats obtenus avec les différents modèles testés et permettra de confirmer cette observation.

5.5 Comparaison des modèles

Pour comparer les différents modèles, le coefficient Gini et le RMSE, présentés dans la partie 5.3.3, sont utilisés. Les parties précédentes ont permis de développer différents modèles pour la fréquence et la première tranche de sévérité pour les garanties RCD et RC Bât. Le tableau ci-dessous permet de comparer les résultats obtenus :

	Freq RCD		Freq RC Bât		Sev T1 RCD		Sev T1 RC Bât	
	Gini	RMSE	Gini	RMSE	Gini	RMSE	Gini	RMSE
GLM	0,3302	0,8318	0,2821	1,1854	0,1164	6383,61	0,1187	4658,95
CART	0,3154	0,8744	0,2635	1,2520	0,1076	6520,13	0,1023	5200,07
Random Forest	0,3370	0,8195	0,2977	1,1253	0,1128	6448,94	0,1127	4775,80
XGBoost	0,3441	0,8014	0,3175	1,0333	0,1287	6294,41	0,1275	4421,98

Le modèle XGBoost présente toujours le coefficient Gini le plus haut et le RMSE le plus bas, et l'inverse est constaté pour la modélisation CART. Cependant, les résultats obtenus restent assez proches et il n'y a pas de modèle qui présente des prédictions bien meilleures que d'autres.

Les GLMs permettent de disposer de modèles facilement interprétables, et compte tenu que leurs résultats ne sont pas très éloignés de ceux obtenus avec des méthodes de Machine Learning, ils seront utilisés dans la dernière partie pour modéliser l'approche Fréquence-Coût.

Modélisation avec les séries temporelles de l'inflation

L'inflation globale est une donnée très importante dans le contexte actuel. En effet, son taux, en France, atteint plus de 6% en juillet 2022, une valeur qui n'a plus été atteinte depuis le milieu des années 1980. Ainsi les coûts augmentent et donc ceux des sinistres également. D'après les premières estimations de la banque de France, pour l'année 2022, ce taux serait de 5,6% et baisserait progressivement avec 3,4% en 2023 et 1,9% en 2024. Ces conjectures sont revues trimestriellement, et à la hausse depuis le début de l'année 2022. Ainsi l'évolution de l'inflation semble difficile à appréhender et dépend fortement des contextes économique et géopolitique. Le secteur de la construction présente son propre indice de mesure des coûts sur ce secteur, il s'agit de l'indice BT01. Il permet de rendre compte des coûts des matériaux et matières premières, de la main d'œuvre ou des énergies. L'inflation globale en France calculée à partir de l'indice des prix à la consommation est corrélée à l'inflation de l'indice du BT01 à 92,73 %.

Pour un risque long, comme la responsabilité civile décennale, l'inflation est importante à considérer car elle permet d'estimer les coûts des sinistres des années futures.

Dans ce mémoire, l'inflation sera considérée comme une variable supplémentaire et ne sera donc pas incluse dans la modélisation. Elle sera utilisée pour revaloriser les coûts prédits par les modèles. Ce choix s'explique par une volonté de ne pas devoir reprendre l'ensemble des travaux de modélisation en cas d'augmentation ou de diminution de l'inflation supérieure à celle prédite. Seule la prédiction de l'inflation sera alors ajustée et les charges seront revalorisées en fonction du changement constaté.

6.1 Historique de l'inflation du BT01

Le niveau d'inflation du BT01 a fluctué au cours des dernières décennies. Le graphique ci-dessous résume son évolution, et permet de constater l'impact des différentes crises et événements géopolitiques sur son cours :

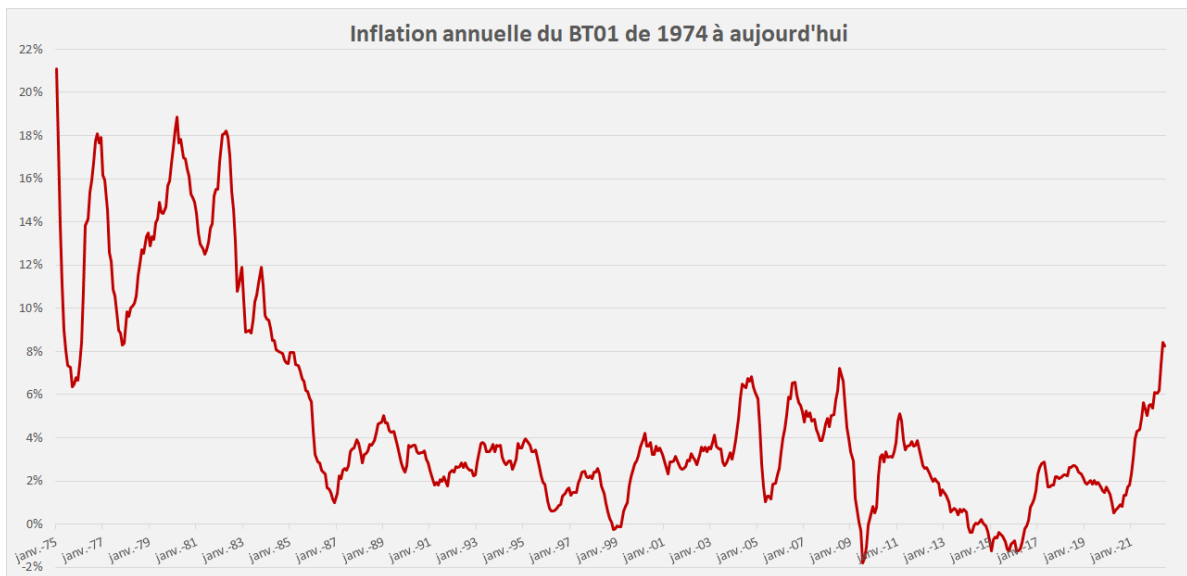


FIGURE 6.1 – Impact des différentes crises (pétrolière, sanitaire ou géopolitique) sur le BT01

Le choc pétrolier de 1974 entraîne un premier pic d'inflation en 1975, le taux atteint plus de 21%. Il fluctue fortement entre 1975 et 1983, en oscillant entre des inflations de 6% et 18%. La crise pétrolière entraîne une forte instabilité sur le secteur de la construction et la sortie de ce contexte est longue. A partir de l'année 1983, l'inflation décroît progressivement pour atteindre 1,5 % en 1987. Le niveau restera stable autour de 3% d'inflation annuelle, malgré quelques fluctuations à la fin du 20^{ème} siècle. Le début des années 2000 est marqué par de nombreuses oscillations de l'inflation. Une diminution des mises en chantier à la suite de la crise des subprimes de 2007 pèse sur la marché. Un rebond intervient en 2011 mais le secteur de la construction est touché par une nouvelle crise se traduisant par une réduction des effectifs, et une réduction du nombre de chantiers. La sortie de cette crise en 2017 permet de retrouver un niveau positif d'inflation. Le confinement en 2020, entraîne une légère baisse de l'inflation suite à l'arrêt de production mais la reprise des activités crée des tensions sur le marché, et en particulier sur celui des matières premières. Une augmentation significative de l'inflation est alors constatée et elle se poursuit aujourd'hui avec la guerre en Ukraine. Actuellement, le pic d'inflation atteint à la suite de la crise des subprimes (7%) est dépassé, et l'inflation en juin 2022 est de 8,26%. La dernière valeur connue dépasse toutes celles qui ont pu être constatées durant les 35 dernières années.

Cet historique permet de constater une tendance à la baisse, avec plusieurs pics qui interrogent la présence d'une certaine saisonnalité, cependant les valeurs maximales atteintes sont différentes et le temps entre deux pics varie. Aujourd'hui une tendance à la hausse s'aperçoit et l'objectif de la partie suivante sera de déterminer le maximum pouvant être atteint et la période pendant laquelle la hausse va se poursuivre.

6.2 Approche avec les séries temporelles

A partir de l'historique de l'inflation durant les cinquante dernières années, le développement dans les prochaines années doit être prédit. Pour cela, une approche avec les séries temporelles sera abordée.

Une série temporelle est une suite de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps.

Soient $t_1, t_2, \dots, t_n$ les temps successifs des observations, et $X_{t_1}, X_{t_2}, \dots, X_{t_n}$ les observations, ainsi x_{t_i} est l'observation réalisée au temps t_i . La série temporelle peut se noter : $\{X_t\}_{t \in T}$

Où T est l'ensemble des temps ordonnés, $T = \{t_1, t_2, \dots, t_n\}$

Dans notre étude, les observations sont les valeurs de l'inflation annuelle du BT01, et les temps sont les mois successifs de janvier 1975 à juillet 2022. L'étude de l'inflation peut donc bien se rapprocher de l'étude des séries temporelles.

6.2.1 Concept et propriétés des séries temporelles

Les séries temporelles peuvent être étudiées sous la forme suivante :

$$X_t = m_t + S_t + Y_t$$

Où m_t est une tendance déterministe

S_t est une saisonnalité déterministe. On suppose S_t de période $d > 1$ ainsi $S_{t+d} = S_t$, et les valeurs possibles de S_t sont notées $s_1, \dots, s_d$

Y_t est une variable/perturbation aléatoire

La modélisation de cette variable consiste à trouver une représentation du type suivant

$$Y_t = h(Y_{t-1}, \dots, Y_{t-p}, \epsilon_t, \dots, \epsilon_{t-q})$$

Où h est une fonction "simple" (généralement linéaire et/ou multiplicative)

p et q ne sont pas trop "grands" pour que h ne soit pas trop difficile à estimer,

(ϵ_t) est un bruit blanc faible

La modélisation adoptée dans ce mémoire, est celle des processus **ARIMA(p, d, q)** où q est l'ordre de la moyenne mobile, p est l'ordre d'un processus autorégressif et d est le paramètre de stationnarité. On dit qu'un processus stochastique $(X_t)_{t \geq -p-d}$ est un processus *ARIMA*(p, d, q) s'il satisfait l'équation suivante :

$$(1 - L)dX_t = \mu + \phi_1(1 - L)dX_{t-1} + \dots + \phi_p(1 - L)dX_{t-p} + \epsilon_t - \theta_1\epsilon_{t-1} - \dots - \theta_q\epsilon_{t-q}$$

Où L est l'opérateur retard, défini par $L(X_t) : (X_t)_{t \in \mathbb{Z}} \mapsto (Y_t)_{t \in \mathbb{Z}}$ tel que $Y_t = LX_t = X_{t-1}$

$\phi_j, j = 1, \dots, p$, et $\theta_j, j = 1, \dots, q$, sont des nombres réels (avec $\phi_p \neq 0$ et $\theta_q \neq 0$)

$(\epsilon_t)_{t \in \mathbb{Z}}$ est un bruit blanc faible de variance σ_ϵ^2

Les conditions initiales $Z_{-1} = \{X_{-1}, \dots, X_{-p-d}, \epsilon_{-1}, \dots, \epsilon_{-q}\}$ vérifient : $Cov(\epsilon_t, Z_{-1}) = 0 \quad \forall t \geq 0$

Plusieurs étapes sont mises en place afin de déterminer les paramètres (p, d, q) du processus. Il convient dans, un premier temps d'étudier la stationnarité de la série temporelle, permettant de déduire la valeur de d . Ensuite, une étude des autocorrélations est effectuée afin de définir p et q .

6.2.2 Étude de la stationnarité

On dit qu'une suite de variables aléatoires $(\dots, X_{-1}, X_0, X_1, \dots)$ est stationnaire si les espérances sont constantes :

$$\forall n, \mathbb{E}(X_n) = \mu$$

et si les covariances sont stables par translation du temps c'est-à-dire, pour tout h ,

$$\forall n, \text{Cov}(X_n, X_{n+h}) = \sigma(h)$$

où la covariance est définie par

$$\text{Cov}(X_n, X_{n+h}) = \mathbb{E}(X_n X_{n+h}) - \mathbb{E}(X_n)\mathbb{E}(X_{n+h}) = \mathbb{E}((X_n - \mathbb{E}(X_n))(X_{n+h} - \mathbb{E}(X_{n+h})))$$

La stationnarité est évaluée avec le KPSS-Test, et validée lorsque la p-value associée au test est supérieure à 5%. La réalisation de ce test sur la série temporelle de l'inflation permet de rejeter l'hypothèse de stationnarité. En effet, la p-value obtenue est de **0,01**, ainsi une nouvelle série est créée correspondant à la différenciation de la série initiale : $\Delta X_t = X_t - X_{t-1}$.

Le KPSS-Test est appliqué sur cette nouvelle série, et la p-value obtenue est **0,1**. Cette valeur est supérieure à 0,05 et l'hypothèse de stationnarité est retenue.

La différenciation d'ordre 1 a permis de rendre la série stationnaire, ainsi dans le modèle $ARIMA(p, q, d)$ le paramètre d vaut 1. Il convient désormais de déterminer les autres paramètres de ce modèle.

6.2.3 Étude des autocorrélations et sélection des paramètres p et q

Elle permet de déterminer les paramètres p et q du modèle ARMA. Pour cela, les graphiques d'autocorrélation et d'autocorrélation partielle sont observés.

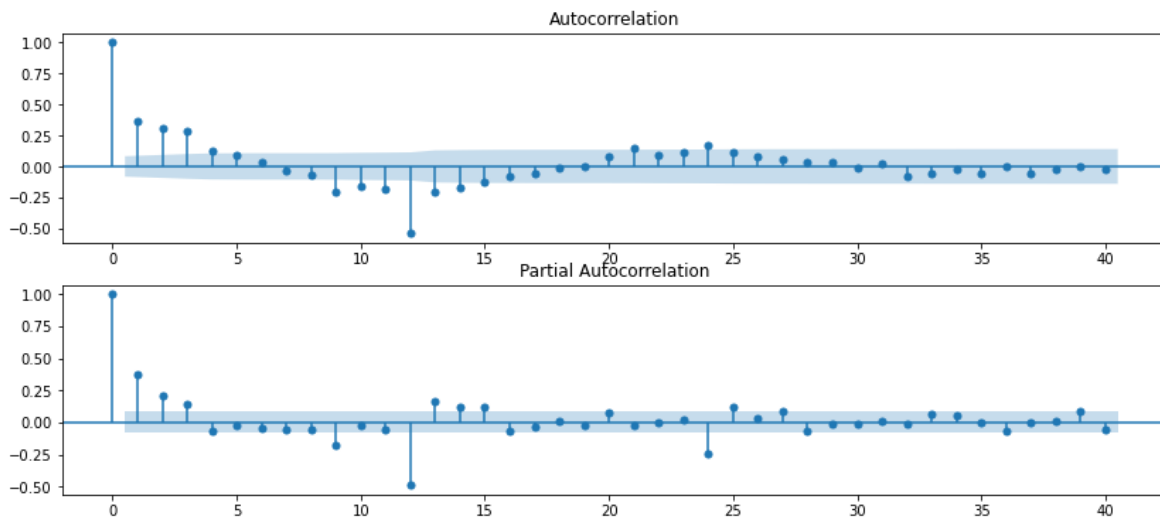


FIGURE 6.2 – Autocorrélation et autocorrélation partielle

Le premier graphique est l'autocorrélogramme permettant de déterminer le nombre de termes moyens mobiles (**ordre MA**) dans le processus ARMA.

Les valeurs q possibles sont celles pour lesquelles l'autocorrélation sort de l'intervalle de confiance à 95%, représenté par une bande bleue sur le graphique. Plusieurs valeurs sortent de cette intervalle : 1, 2, 3, 9, 10, 11, 12.

Le second graphique représente l'autocorrélation partielle et permet de définir le nombre de termes

autorégressifs (**ordre AR**) du modèle.
 La méthode de détermination de ces valeurs est la même que pour l'autocorrélation, ainsi le paramètre p peut prendre comme valeur : 1, 2,3, 9, 12.

Les données initiales sont séparées en deux jeux de données : la base d'apprentissage et la base de test. Ils seront utilisés pour définir le modèle adéquat et pour le tester. Dans un premier temps, la base de test se compose des dix dernières années d'inflation et la base d'apprentissage contient toutes les données antérieures.
 Cependant, la crise sanitaire et le contexte géopolitique influe fortement sur l'inflation. Ces baisses et augmentations ne sont pas prédites par le modèle lorsque l'historique de test est de dix ans. Ainsi d'autres bases d'apprentissage et de test sont créées avec des historiques de test de 5 et 3 ans. Les résultats obtenus sur avec les différents historiques sont comparés afin de déterminer les paramètres p et q optimaux pour le modèle ARIMA.¹.
 Les résultats permettent de sélectionner le modèle **ARIMA(3,1,10)**.

Une vérification de la bonne définition des paramètres du modèle est effectuée. Pour chaque coefficient du processus, la p-value est toujours inférieure à 0,05. Le modèle est valide, il n'y pas de réduction de l'ordre p ou q a effectué.

6.2.4 Prédictions de l'inflation

Le modèle déterminé précédemment est donc utilisé pour déterminer l'inflation dans les quatre prochaines années en utilisant l'historique de données disponible.

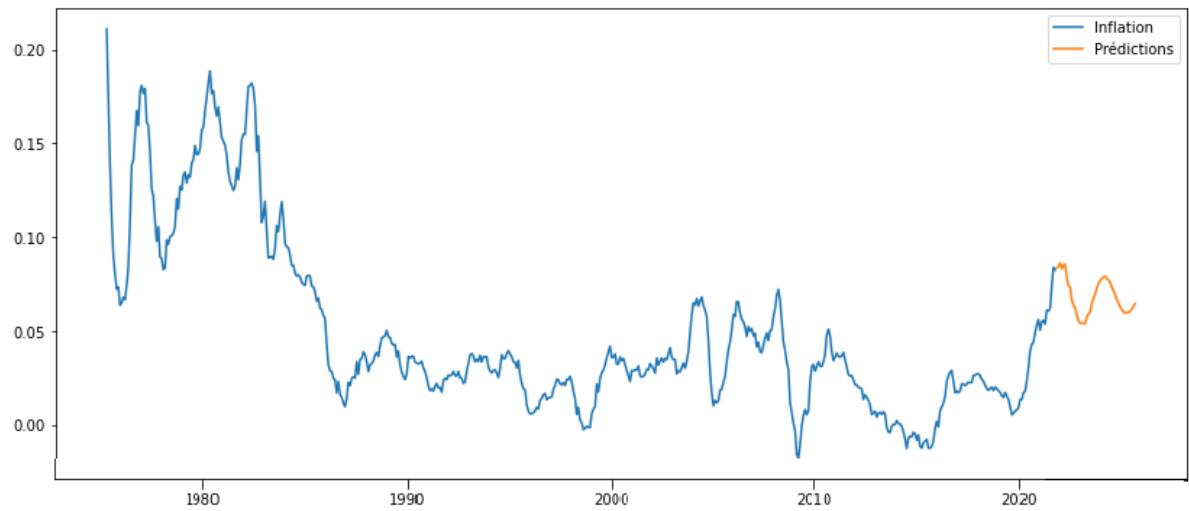


FIGURE 6.3 – Prédiction de l'inflation pendant 4 ans

Les prédictions indiquent que le maximum sera atteint dans 3 mois à 8,86% d'inflation. Une première baisse intervient à la fin de l'année 2022 pour atteindre un minimum de 5% en décembre 2023. En accord avec les crises passées, un deuxième pic d'inflation est constaté en 2024, avant de décroître en 2025 et 2026.
 Le tableau ci-dessous indique l'inflation moyenne annuelle prédite :

	2022	2023	2024	2025	2026
Moyenne	7,78%	6,25%	7,00%	6,85%	6,16%

1. La comparaison des modèles, ayant un historique de données différent, est indiquée en annexe

L'inflation du BT01 est fortement corrélée à celle calculée à partir de l'IPC, qui est l'inflation générale constatée en France. Il n'existe pas de publications sur l'évolution probable du BT01 dans les prochaines années, mais la Banque Centrale Européenne (BCE) et la Banque de France établissent des prédictions sur le développement de l'inflation générale.

Ainsi la prédiction à la hausse semble plausible et en accord avec les publications de ces instances. Cependant elles prédisent une baisse de l'inflation beaucoup plus rapide que celle prédite avec les séries temporelles, et pas de deuxième pic. En effet, d'après les prévisions macroéconomique de la banque de France, l'inflation se situerait à 4,7 % en moyenne sur l'année 2023 et elle se replierait à 2,7% en moyenne annuelle en 2024. La BCE prévoit, quant à elle, une inflation de 5,5% en 2023 et de 2,3% en 2024.

Un "double pic" d'inflation est plus souvent constaté pour le BT01 que pour l'IPC, ainsi nous garderons cette éventualité, cependant afin de s'adapter davantage aux prévisions de la BCE et de la BF, les résultats prédits par les séries temporelles seront réduits. Ainsi une réduction de 2 points de pourcentage sera effectuée sur les résultats de prédiction des années 2023 et en 2024, et elle sera de 3 points pour les années suivantes.

Application et résultats

Les parties précédentes ont permis de modéliser la fréquence ainsi que les différentes tranches de sévérités, qui sont nécessaires pour mettre en place l'approche *Fréquence × Coût*. Cette dernière partie permettra de modéliser cette approche avec les modèles GLMs déterminés en amont.

Pour calculer la charge totale, il faut additionner celle calculée avec l'approche *Fréquence × Coût*, avec la charge des sinistres de la quatrième tranche de sévérité. Concernant ces sinistres ayant un coût au-dessus de 2 millions d'euros, une approche simple est réalisée en divisant la somme des montants des sinistres par le nombre de sinistres.

Pour une garantie, le calcul de la charge est effectué selon la formule suivante :

$$Charge = freq \times \left[p_{T_1} \times Sev_{T_1} + p_{T_2} \times Sev_{T_2} + p_{T_3} \times Sev_{T_3} \right] + \frac{\sum Charge_{T_4}}{\text{nombre de sinistres } T_4}$$

Où

p_{T_1} , p_{T_2} et p_{T_3} sont respectivement les probabilités que le coût du sinistre appartienne à la tranche 1, 2 ou 3 de sévérité. Les probabilités d'appartenir aux différentes tranches sont détaillées dans le tableau ci-dessous :

	Responsabilité Civile Décennale	Responsabilité Civile Bâtiment
Tranche 1	93,39%	96,36%
Tranche 2	5,57%	2,87%
Tranche 3	1,02%	0,75%
Tranche 4	0,02%	0,02%

De plus, les sinistres de la quatrième tranche ont une charge moyenne de 3,7 millions d'euros pour la RCD et de 3 millions d'euros pour la RC Bât.

La définition précédente est générale, il faut désormais prendre en compte les spécificités liées aux données, et aux modifications faites pendant les étapes de modélisation.

Premièrement, lors de la modélisation de la première tranche de sévérité les sinistres de coûts forfaitaires ont été retirés. Cette suppression vient modifier la charge moyenne des sinistres de la tranche, ainsi il faut ajouter un coefficient multiplicatif C afin de revenir au coût moyen initial de la tranche. Les charges diffèrent selon la garantie, ainsi deux coefficients multiplicatifs seront déterminés. Le tableau ci-dessous résume la charge des sinistres avant et après la suppression des coûts forfaitaires, et permet ainsi de calculer les deux coefficients.

	Comparaison des charges moyennes de la première tranche de sévérité		
	Avant la suppression	Après la suppression	Coefficient multiplicatif
RCD	4 101,2	5 433,7	32,5%
RC Bât	2 792,4	3 405,1	22,0 %

La formule de calcul, pour la partie modélisée et pour la garantie RCD, devient donc :

$$Charge_{modélisée}^{RCD} = freq^{RCD} \times \left[\frac{p_{T_1}^{RCD} \times Sev_{T_1}^{RCD}}{1 - C_{RCD}} + p_{T_2}^{RCD} \times Sev_{T_2}^{RCD} + p_{T_3}^{RCD} \times Sev_{T_3}^{RCD} \right]$$

La formule de calcul de la charge pour la garantie RC Bât est similaire. Il faut considérer les modèles de fréquence et de sévérités de cette garantie ainsi le coefficient $C_{RC \text{ Bât}}$

Secondement, lors de la création des bases de fréquence et de sévérité les sinistres sans-suites de charge nulle n'ont pas été retirés. Ainsi il convient désormais d'inclure leur modélisation pour réduire la charge calculée précédemment, en affectant à zéro la charge à partir des caractéristiques de l'assuré. L'ajout de cet modélisation réduit donc la charge initialement calculée, et elle prédit le taux de sans-suite suivant :

- 25,1 % des sinistres pour la base RCD
- 13,7 % des sinistres pour la base RC Bât

Les taux de sans-suites modélisés sont proches de ceux initiaux présents dans les bases de données. Ces taux seront supposés constants dans le temps pour la suite de la prédiction.

Responsabilité Civile Décennale :

La charge des années passées peut être calculée à l'aide des modélisations précédentes. L'étude des résidus permet de juger de son adéquation, et de la fiabilité du modèle. Pour les années entre 2016 et 2019 et pour la garantie RCD, des prédictions à l'aide du modèle Fréquence-Coût sont réalisées, elles permettent de définir la présence d'un sinistre et sa charge associée pour chaque observation de l'année considérée.

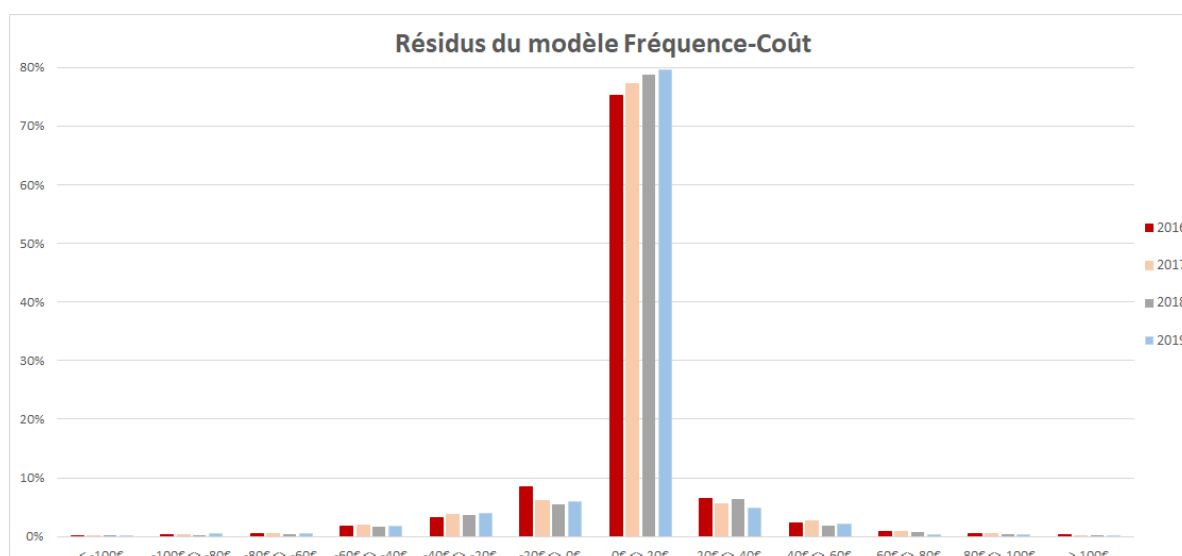


FIGURE 7.1 – Résidus du modèle Fréquence-Coût pour la RCD

Cet histogramme illustre la propagation des résidus pour plusieurs années, montrant le pourcentage d'observations pour lesquelles les résidus se situent dans les limites indiquées sur l'axe des abscisses. Les résidus sont concentrés à 75,7% autour de 0, avec un minimum de 70,2% pour l'année 2020.

La formule finale de calcul des charges est désormais définie, ainsi les montants futurs peuvent être déterminés. Ils correspondent aux dépenses estimées pour l'année considérée, ils comptent les dépenses liées aux chantiers dont la DROC est cette même année, et tous les sinistres dont la DROC est antérieure mais dont la garantie est toujours en vigueur. Les montants estimés, pour l'année 2022 ainsi que pour les quatre prochaines années, sont les suivants :

- 2022 : 18 885 236 €
- 2023 : 16 933 637 €
- 2024 : 13 735 707 €
- 2025 : 12 186 084 €
- 2026 : 11 641 099 €

Une baisse des charges à payer pour les années futures est constatée. Elle s'explique par la résiliation de nombreux contrats entre 2013 et 2017 qui arrivent désormais à échéance. Compte tenu de la durée d'un chantier et du délai de survenance d'un sinistre, cette observation est cohérente.

Le contexte inflationniste actuel nécessite de prendre en compte l'inflation dans les modélisations. Celle-ci a été étudiée et projetée indépendamment des travaux sur les GLMs. Les inflations déterminées sont donc ajoutées aux montants précédemment calculés. Il convient de cumuler les inflations successives, par exemple pour l'année 2023, il faut prendre en compte les inflations de 2022 et de 2023. En effet, les sinistres utilisés pour les travaux de modélisation sont antérieurs à 2022 et ne tiennent donc pas compte de l'augmentation des coûts actuelle.

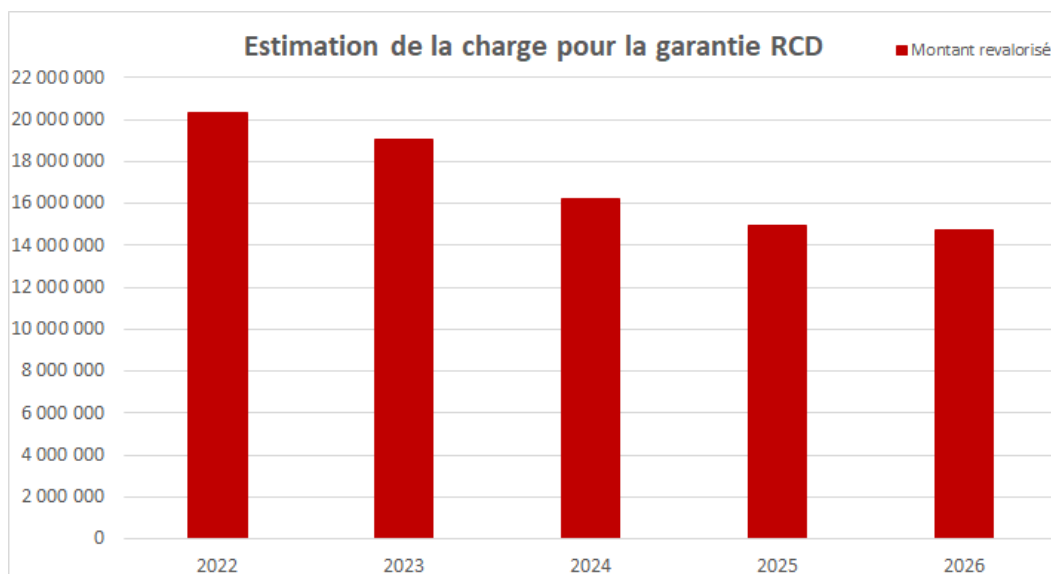


FIGURE 7.2 – Revalorisation des dépenses pour les prochaines années pour la garantie RCD

La revalorisation avec l'inflation prédite implique une augmentation des charges prédites. La baisse de la charge dans le temps est toujours constatée, mais l'inflation cumulée en 2026 détermine un montant proche de celui de 2025. La hausse des charges avec l'inflation est comprise entre 1,5 millions d'euros en 2022 et 3,1 millions en 2026 (hausse comprise entre 7,78% et 26%).

On constate que dans le contexte économique actuel, considérer l'inflation est primordiale, et que dans le cas contraire une sous-estimation des sinistres et des montants pourrait être effectuée.

Le second objectif de ce mémoire était de déterminer les éléments impactant la sinistralité, et par répercussion la prime. Plusieurs variables ont été ajoutées aux différents modèles avec des influences diverses. Le type d'activité effectuée par un assuré est une variable incluse dans la

majorité des modèles et apparaît ainsi comme une variable à considérer afin de segmenter les assurés, et pouvant être utilisée par la suite dans des travaux de tarification.

Afin de comparer la prédiction et l'observation de la charge sur la variable *Type d'Œuvre*, l'assuré de référence est considéré pour toutes les autres variables. Pour le graphique ci-dessous, la légende est la suivante :

- les bâtons jaunes représentent l'exposition ;
- La courbe rose représente la charge observée ;
- La courbe verte correspond à la charge modélisée ;

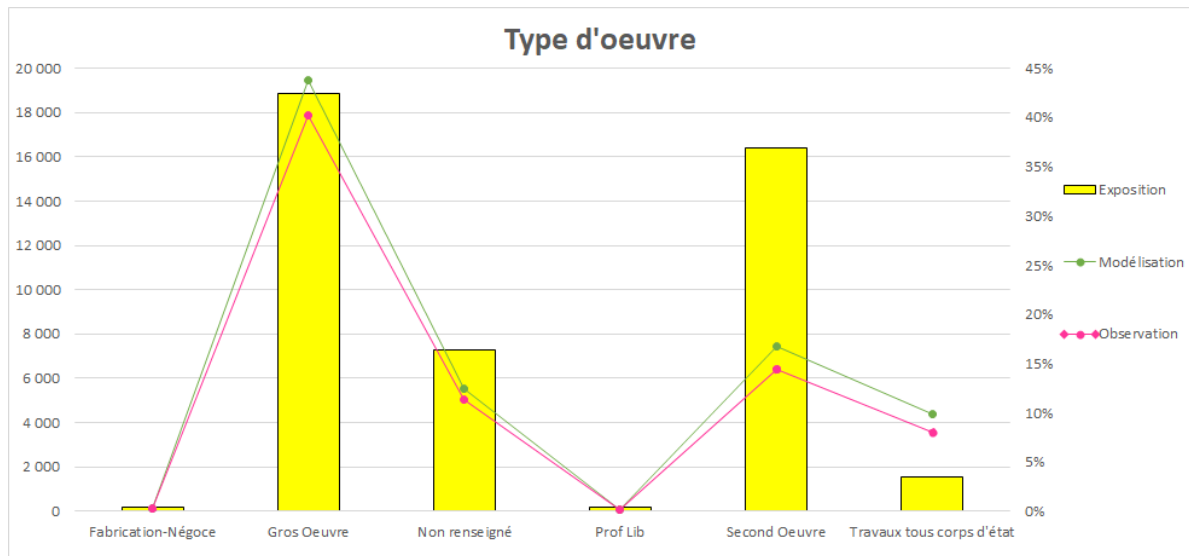


FIGURE 7.3 – Comparaison de la charge observée et de la charge modélisée sur la variable Type d'œuvre

La modalité "AUTRE" n'est pas représentée, sur le graphique précédent, car son exposition est faible (0,08%), tout comme la charge prédite.

Ce graphique permet de constater une légère sur-estimation des charges par le modèle par rapport aux observations. Il met en avant une grande différence sur les charges en fonction du type d'œuvre, ainsi les gros et second œuvres sont plus risqués. Cette constatation est cohérente avec les observations faites précédemment sur les coûts moyens des sinistres en fonction du type d'œuvre.

Responsabilité Civile Bâtiment :

Tout comme pour la RCD, les charges pour les années entre 2016 et 2019 sont calculées. Les résidus obtenus ne sont pas représentés, mais les résultats sont similaires à ceux obtenus pour la RCD. La concentration autour de 0 est plus faible, avec une moyenne sur les quatre années de 74,8%.

Les charges prédites pour les prochaines années sont représentées sur le graphique suivant :

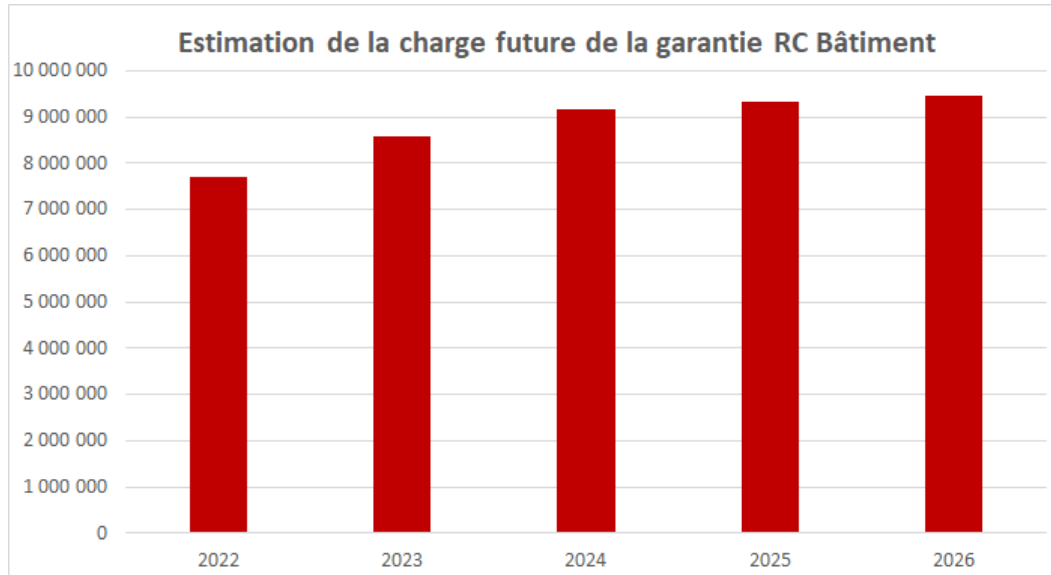


FIGURE 7.4 – Évolution à la hausse des charges futures pour la garantie RC Bât

La crise sanitaire avait fait diminué le nombre de mises en chantier et ainsi avait impacté la sinistralité pour la garantie Responsabilité Civile Bâtiment. Un retour à un nombre de chantier similaire à celui de 2019 est constaté dès 2021, et ce nombre se stabilise. Considérant ce nombre de mise en chantier constant dans les années de projection et compte tenu de la forte inflation prédite, la charge calculée est en augmentation pour les prochaines années.

Comme pour la garantie RCD, l'inflation entraîne une hausse de la charge initialement prédite avec les modèles GLMs. Elle est comprise entre environ 556 000 euros en 2022 et près de 2 millions d'euros en 2026.

Comme pour la RCD, des variables apparaissent comme influençant davantage la charge. Le type d'activité de l'assuré est également une source de sinistralité, mais pour cette garantie la deuxième variable est l'effectif.

Pour cette première variable, les conclusions pouvant être apportées sont similaires à celles formulées pour la RCD, ainsi les gros et second œuvres apparaissent comme étant les modalités avec la charge et l'exposition les plus élevées. Ainsi il conviendrait de porter une attention particulière à ces modalités.

Comme précédemment, pour comparer la prédiction et l'observation de la charge sur la variable *Effectif*, l'assuré de référence est considéré pour toutes les autres variables. La légende est également conservée.

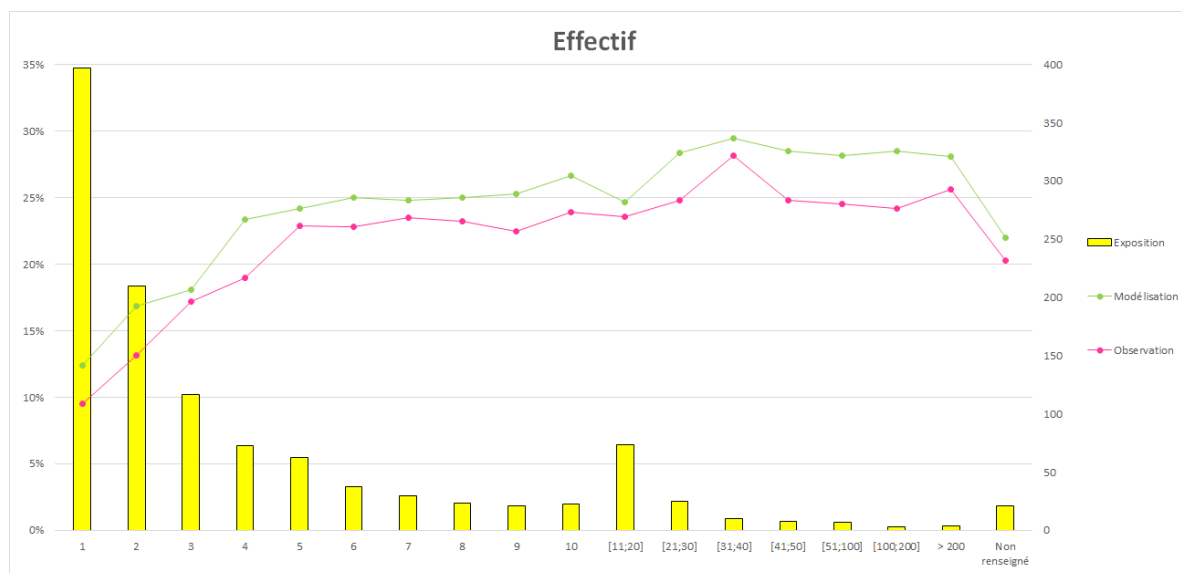


FIGURE 7.5 – Comparaison de la charge observée et de la charge modélisée sur la variable Effectif

La modélisation présente une estimation de la charge supérieure à celle observée pour l'ensemble des modalités de la variable "*Effectif*", avec une charge croissante avec le nombre d'ETP. Cette constatation paraît cohérente car une entreprise avec un nombre plus important de salarié peut prétendre à des chantiers plus importants, et donc avec des coûts plus élevés.

Cette partie a donc permis de mettre en évidence l'intérêt de considérer l'inflation dans le contexte économique actuel, et les risques de sous-estimation associés dans le cas contraire. Enfin, certaines variables sont ajoutées dans plusieurs des modèles GLMs réalisés dans les parties précédentes, et sont donc une source importante d'information sur la sinistralité, et donc sur la charge. Les résultats obtenus sur ces variables pourront être utilisés dans des travaux supplémentaires de mise en place d'une tarification plus segmentarisée.

Conclusion

Les polices d'activités permettent de couvrir un assuré pour l'ensemble de son activité professionnelle et donc pour éventuellement, plusieurs chantiers, à l'inverse des polices de chantier. Ce mémoire s'est intéressé à deux types de garanties : la Responsabilité Civile Bâtiment et la Responsabilité Civile Décennale. Cette dernière est un risque long avec des sinistres pouvant survenir plus de 10 ans après le paiement des primes. Il s'agit donc d'un risque particulier et différent de ceux habituellement rencontrés en assurance, qui demande de prendre en compte le délai de survenance d'un sinistre.

Au sein du groupe Generali France, depuis 2013, de nombreux changements avaient été opérés dans la structure du portefeuille et dans la gestion des sinistres, amenant à reconsidérer les modélisations alors existantes pour la responsabilité civile décennale. Afin de compléter l'étude sur les polices d'activités, une modélisation est mise en place pour la responsabilité civile bâtiment, aucuns travaux n'avaient été effectués sur cette garantie auparavant.

Dans un premier temps, la théorie des extrêmes et l'utilisation d'estimateurs tels que celui de Gerstengarbe ont permis de confirmer le seuil historique de gravité de 150 000 euros mais ils mettent en évidence d'autres seuils permettant de segmenter davantage la sinistralité. Ainsi quatre tranches ont été créées : $[0 - 30\,000]$, $[30\,000 - 150\,000]$ et $[150\,000 - 2M]$ et "Plus de 2M". Les trois premières ont la particularité de concentrer presque la même proportion de sinistres en montant (environ 30 % de la charge totale), mais la première tranche concentre plus de 90% des observations.

La segmentation réalisée, et les bases de données créées, regroupant séparément les informations sur la sinistralité et sur le profil de risque, la modélisation peut alors commencer. Plusieurs modèles sont déterminés afin de prédire la fréquence, le coût ainsi que le taux de sans-suite des deux garanties. Ces modélisations ont permis de mettre en avant des différences importantes dans choix des variables explicatives, mais une information sur l'activité se retrouve dans la majorité des modèles. Ainsi l'activité semble être une source majeur d'explication de la sinistralité, pour la fréquence ainsi que pour le coût. Cette constatation est cohérente avec les observations sur les données initiales car les activités de gros et de second œuvres présentent des coûts supérieurs aux autres type d'activités.

L'inflation est une donnée primordiale à considérer dans le contexte économique actuel. En effet, les niveaux d'inflation constatés aujourd'hui atteignent des valeurs bien supérieures à celles des dernières années. La projection de cette inflation, avec les séries temporelles, indique que le taux se maintiendra dans des valeurs supérieures à 2 % dans les 5 prochaines années. La dernière partie permet de constater que la considération de cette donnée permet de réévaluer la charge prédite par les modèles $Fréquence \times Coût$, et de limiter une sous-évaluation de la charge future dans ce contexte économique inflationniste. Enfin, l'estimation des charges futures pourra être réévaluée

facilement si l'évolution de l'inflation est différente de celle prédite, et il est également possible de déterminer plusieurs évolutions probables de l'inflation et ainsi de calculer les montants de charges futures associés.

En conclusion, ce mémoire a permis de mettre en évidence des différences mais aussi des similarités dans la modélisation de la charge des deux polices d'activités. Des similitudes sur les seuils de gravité retenus ou sur les lois de modélisation, mais des discordances sur les variables explicatives retenues. Il pourra être utilisé pour des travaux complémentaires de tarification, en considérant les variables influençant le montant de la charge (ou de la fréquence ou de la sévérité, plus particulièrement).

Table des figures

1.1	Liens contractuels entre les différents intervenants sur une opération de construction	16
1.2	Dates importantes de l'assurance décennale	19
1.3	Des cotisations de l'assurance construction en faible proportion	20
1.4	Cotisations de l'assurance construction	20
1.5	Impact de la crise sanitaire sur l'indice BT01	22
1.6	Impact du contexte géopolitique sur l'inflation du BT01	22
1.7	Le prix des carburants, une source supplémentaire de dépenses pour les entreprises	24
2.1	Répartition des primes et des assurés selon leur typologie	26
2.2	Impact des modification de gestion sur le portefeuille	27
2.3	Réduction du nombre d'assurés <i>Entreprise</i> dans le portefeuille	27
2.4	Une légère baisse de la proportion de contrats <i>Révisable</i> de 2007 à 2022	28
3.1	Distribution des sinistres de charge inférieure à 30 000 euros	30
3.2	Importance des coûts forfaitaire dans la distribution des sinistres	30
3.3	Comparaison des indices BT01 et FFB entre 2000 et 2021	34
3.4	Comparaison de l'inflation des indices BT01 et FFB entre 2000 et 2021	34
3.5	Variations des quantiles pour l'ensemble des données	37
3.6	Représentation de la fonction de dépassement moyen	38
3.7	Détermination de plusieurs seuils avec l'estimateur de Hill	39
3.8	Détermination d'un seuil avec l'estimateur de Gerstengarbe	39
3.9	QQ-Plot ensemble des données	40
4.1	Fractionnement de la prime par exercice	44
4.2	Récupération et traitement des informations sur l'activité	47
5.1	Mécanisme d'ajout du délai de survenance sur la base risque	51
5.2	Mécanisme de construction de la base fréquence	51
5.3	Construction de la base sévérité pour les sinistres de la première tranche	52
5.4	Construction de la base des sans suites pour la garantie RCD	53
5.5	Concentration des anciennetés des entreprises avant 60 ans	54
5.6	Graphique des corrélations entre les variables au seuil de 0,5	57
5.7	Ajout de six variables explicatives avec la méthode Forward, pour la modélisation de la fréquence RCD	61
5.8	La faible représentation des modalités détériore les erreurs	62
5.9	Distribution des sinistres (RCD) de la première tranche de sévérité	65
5.10	Distribution des sinistres (RC Bât) de la première tranche de sévérité	65
5.11	Première tranche de sévérité RCD retraitée	66
5.12	Sévérité RC Bât de la première tranche retraitée	66
5.13	Une tendance polynomiale sur les modalités	66
5.14	Approximation du délai de survenance par une fonction polynomiale	67

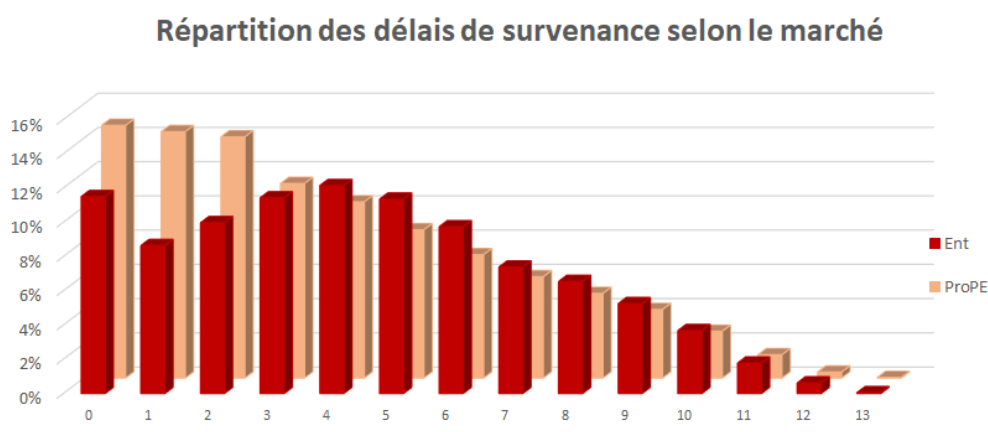
5.15	Distribution de la deuxième tranche de sévérité RC Bât	68
5.16	Distribution de la troisième tranche de sévérité RCD	68
5.17	Le gros et le second œuvres présentent un coût de sinistre supérieur	69
5.18	Spécification du modèle des sans suites sur Emblem	70
5.19	Choix des variables retenues dans la modélisation des sans-suites RCD	70
5.20	Le coefficient Gini de la modélisation de la fréquence RCD est de 0,3302	72
5.21	Résidus centrés en zéro pour le modèle de fréquence RCD	73
5.22	Bonne prédiction de la variable ancienneté entreprise incluse dans le modèle	74
5.23	Bonne prédiction de la variable effectif non incluse dans le modèle	74
5.24	Arbre de décision pour la fréquence RC Bât	75
5.25	Performance du modèle Random Forest pour la fréquence RC Bât	76
5.26	Importance des variables dans la modélisation XGBoost de la fréquence RC Bât . .	77
6.1	Impact des différentes crises (pétrolière, sanitaire ou géopolitique) sur le BT01 . .	80
6.2	Autocorrélation et autocorrélation partielle	82
6.3	Prédiction de l'inflation pendant 4 ans	83
7.1	Résidus du modèle Fréquence-Coût pour la RCD	86
7.2	Revalorisation des dépenses pour les prochaines années pour la garantie RCD . . .	87
7.3	Comparaison de la charge observée et de la charge modélisée sur la variable Type d'œuvre	88
7.4	Évolution à la hausse des charges futures pour la garantie RC Bât	89
7.5	Comparaison de la charge observée et de la charge modélisée sur la variable Effectif	90

Bibliographie

- [1] Wassil BENTARKA. *Evaluation des provisions techniques et estimation du risque de réserve en assurance construction*. Mémoire d'actuariat ISUP, 2021.
- [2] Clément BOURRY. *Évaluation des provisions techniques et du capital économique associé au risque de réserve en assurance construction*. Mémoire d'actuariat ISFA, 2016.
- [3] Banque de France. *Prévisions économiques - projections macroéconomiques*, septembre 2022.
- [4] Fédération Française de l'Assurance. *Nomenclature des activités du bâtiment*.
- [5] Fédération Française de l'Assurance. *Assurance construction en 2020*, décembre 2021.
- [6] Fédération Française du Bâtiment (FFB). *Indices et index du bâtiment*.
- [7] Banque Centrale Européenne. *Projections macroéconomiques pour la zone euro*, septembre 2022.
- [8] DEBOUDT Hélène. *Estimer l'impact du changement de politique de souscription sur la rentabilité du risque de Responsabilité Civile Décennale*. Mémoire d'actuariat CEA, 2017.
- [9] INSEE. *Caractéristiques de la construction par activité - données annuelles 2019*.
- [10] INSEE. *Prix moyens de vente de détail (carburants)*.
- [11] Franck PEKA KADJE. *Revalorisation technique des garanties du produit d 'Assurance Multirisque Professionnelle*. Mémoire d'actuariat ISUP, 2018.
- [12] Fabrice KIEMA. *Méthodes de machine learning en provisionnement non-vie : Constitution de provisions dossier/dossier pour des sinistres de protection juridique*. Mémoire d'actuariat DUAS, 2018.
- [13] Pascale QUENNELLE Mickael ITTAH, Xavier LAUREILLE. *Détermination des seuils de sinistres graves et optimisation de la réassurance*. *Institut des Actuaire*s, novembre 2019.
- [14] LAVERNY Oskar. *Provisionnement et risque de réserve en assurance construction*. Mémoire d'actuariat ISFA, 2018.
- [15] Charles TREMBLAY. *Prédire les sinistres graves en assurance : les apports de l'apprentissage statistique aux modèles linéaires*. Mémoire d'actuariat ENSAE, 2017.
- [16] Isabelle VIDAL. *Prédiction de l'acte de résiliation et tarification de la garantie RC des jeunes conducteurs*. Mémoire d'actuariat ISFA, 2014.

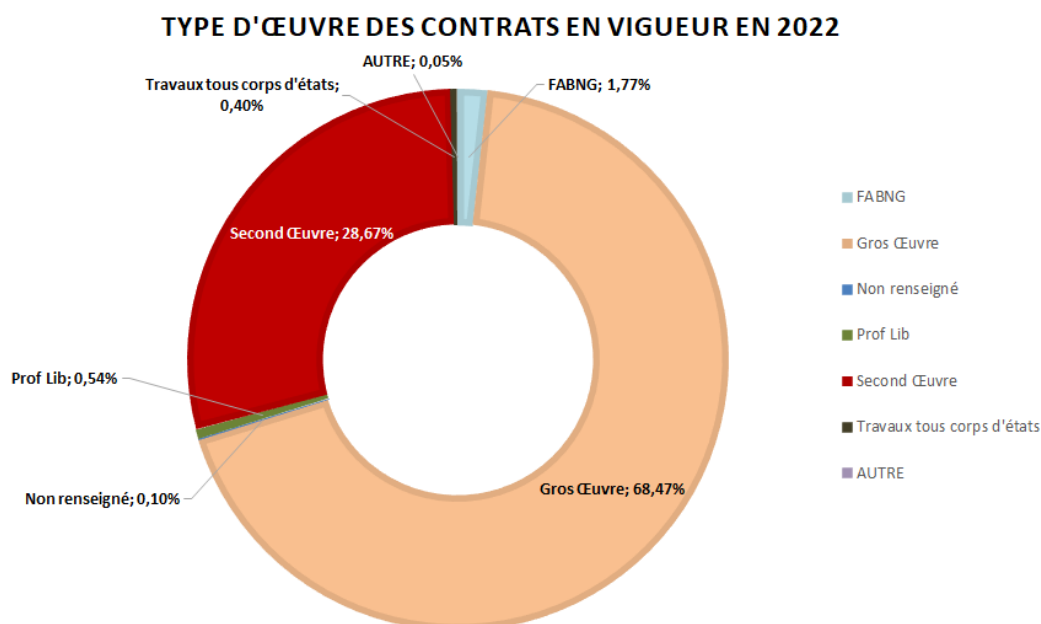
Annexe A : Informations supplémentaires sur les données

A.1 Répartition du portefeuille



Les assurés *ProPE* déclarent des sinistres plus rapidement que les assurés *Entreprise*

A.2 Répartition des activités



Répartition du *Type d'Œuvre* en 2022

	TYPE D'ŒUVRE						
Année	Gros Œuvre	Second Œuvre	ZX	FABNG	Prof Lib	AUTRE	Non renseigné
2007	41,20%	19,78%	1,98%	0,92%	0,60%	0,11%	35,41%
2008	47,06%	23,05%	1,88%	0,99%	0,67%	0,10%	26,25%
2009	51,09%	25,05%	1,66%	1,12%	0,74%	0,09%	20,26%
2010	54,57%	26,45%	1,43%	1,27%	0,86%	0,08%	15,32%
2011	57,90%	27,14%	1,18%	1,38%	0,98%	0,08%	11,34%
2012	59,68%	27,76%	1,13%	1,54%	1,11%	0,09%	8,69%
2013	62,24%	29,26%	1,09%	1,87%	1,31%	0,09%	4,14%
2014	63,08%	30,11%	1,02%	2,12%	1,51%	0,12%	2,03%
2015	63,68%	30,44%	0,98%	2,41%	1,33%	0,08%	1,08%
2016	64,14%	30,44%	0,95%	2,39%	1,21%	0,07%	0,79%
2017	64,82%	30,33%	0,74%	2,48%	1,10%	0,08%	0,46%
2018	65,24%	30,57%	0,64%	2,30%	0,91%	0,07%	0,27%
2019	66,12%	30,22%	0,58%	2,10%	0,74%	0,05%	0,19%
2020	67,15%	29,62%	0,48%	1,95%	0,60%	0,05%	0,15%
2021	68,16%	28,91%	0,43%	1,81%	0,54%	0,05%	0,11%
2022	68,47%	28,67%	0,40%	1,77%	0,54%	0,05%	0,10%

Le graphique précédent permet de constater une représentation à plus de 90% des gros et second œuvres. Ces deux types d'œuvre ont toujours été majoritairement représentés dans le portefeuille pendant les 15 dernières années, cependant leur proportion a augmenté significativement entre 2007 et aujourd'hui. Cette hausse s'accompagne d'une baisse des valeurs "Non renseigné" passant de 35,41% en 2007 à 0,10% en 2022. Elle s'explique par une obligation de déclaration de l'activité en 2013.

Les professions libérales et les activités classés dans "AUTRE" ont une représentation constante dans le portefeuille. Les activités tous corps d'état sont en baisse, avec une chute de 1,58 points de pourcentage en 15 ans. A l'inverse, les activités de fabrication et de négoce sont en hausse de 0,85 point de pourcentage.

Annexe B : Choix des lois de modélisation

Modélisation de la fréquence :

	RC Décennale			RC Bâtiment		
Critères de sélection	AIC	BIC	Deviance	AIC	BIC	Deviance
Loi de Poisson	252 691	253 117	118 395	90 431	90 659	45 763
Loi Binomiale Négative	245 540	245 733	110 269	87 312	87 483	42 915

La loi Binomiale Négative est donc utilisée pour la modélisation de la fréquence pour les deux garanties.

Modélisation de la sévérité :

Première tranche	RC Décennale			RC Bâtiment		
Critères de sélection	AIC	BIC	Deviance	AIC	BIC	Deviance
Loi Gamma	380 010	380 018	147 810	314 499	314 506	124 054
Loi Log-normale	411 560	411 718	163 278	351 667	352 103	128 159

Pour les deux garanties, la modélisation de la première tranche de sévérité sera effectuée avec une loi Gamma.

Deuxième tranche	RC Décennale			RC Bâtiment		
Critères de sélection	AIC	BIC	Deviance	AIC	BIC	Deviance
Loi Gamma	45 811	45 904	9 415	24 532	24 987	2 651
Loi Log-normale	34 752	34 757	6 358	12 049	12 053	1 156

La modélisation de la deuxième tranche de sévérité sera effectuée avec une loi log-normale pour les deux garanties.

Troisième tranche	RC Décennale			RC Bâtiment		
Critères de sélection	AIC	BIC	Deviance	AIC	BIC	Deviance
Loi Gamma	7 334	7 380	829	4 419	4 422	510
Loi Log-normale	21 007	21 409	2 281	15 687	16 241	1 451
Loi de Pareto	15 768	15 822	1 630	10 073	11 185	1 041

Pour les deux garanties, la modélisation de la troisième tranche de sévérité sera effectuée avec une loi Gamma.

Annexe C : Regroupements et choix des variables dans les modélisations

Ici seront détaillés les regroupements effectués dans les différentes modélisations présentées dans la cinquième partie. Les regroupements pour la fréquence ont été détaillés dans la partie associée, ainsi ils ne seront pas définis à nouveau ici.

Première tranche de sinistralité :

Responsabilité Civile Décennale : La modélisation compte 5 variables explicatives, la variable *année* et la variable *TOP ECRAC* qui ne possède que 2 modalités "Oui" ou "Non", ainsi que les 3 variables dont les regroupements sont détaillées ci-dessous :

- *délai de survenance* : les délais supérieurs à 11 ans sont regroupés permettant de créer la modalité " ≥ 11 ". Les coefficients de la régressions augmentent avec le délai de survenance.
- *nombre de contrats* : la variable présente désormais comme modalité "1", "2", "3", "4", "[5;9]", "[10;19]", " ≥ 20 ". La modalité "1" est la plus représentée et tous les coefficients de régressions des autres modalités sont inférieurs à celui de cette modalité.
- *Type d'activités* : deux regroupements sont effectués : les modalités "ZC", "FAB" et "AUTRE" sont regroupées, ainsi que "ZX" avec la modalité la plus représentée "ZD". Les autres modalités "ZP", "ZS", "ZT" et "Non renseigné" ne sont pas modifiées. Sauf pour "ZT", les coefficients de régression des différentes modalités sont supérieurs à 1 (correspondant au coefficient de la modalité la plus représentée).

Responsabilité Civile Bâtiment : Le modèle se compose de 7 variables explicatives qui sont *l'année* et les 6 variables suivantes dont les regroupements sont détaillés ci-dessous.

- *Ancienneté du contrat* : les modalités retenues pour la modélisation sont "[0;5]", "[6;10]" et " > 10 ". Les coefficients de la régression sont décroissants avec l'ancienneté du contrat. Ainsi un contrat présent dans le portefeuille depuis 1 an présentera un coût supérieur à un contrat présent dans le portefeuille depuis 10 ans, en considérant que ces deux contrats présentent les mêmes autres caractéristiques.
- *Chiffre d'affaires sous traité* : les pourcentages de sous-traitances sont associés permettant d'obtenir 4 catégories : "[0;20]", "[20;80]", "[80;100]" et "Non renseigné". Les modalités "[0;20]" et "[80;100]" ont des coefficients de régression inférieurs à ceux des autres modalités avec le coefficient le plus faible pour la première modalité.
- *Effectif* : les valeurs d'ETP dont se compose la variable sont "1", "[2;5]", "[6;20]", "[21;50]", " > 50 " et "Non renseigné". Les coefficients de la régression sont croissants avec l'effectif.
- *Taux* : Seules les modalités "F", "L", "N" et "X" sont regroupées, les autres ne sont pas impactées. Les coefficients de la régression sont croissants avec le taux.
- *nombre de contrats* : la modalité la plus représentée est "1", elle n'est pas regroupée avec d'autres valeurs possibles de la variable, tout comme la modalité "2". Ensuite des regroupe-

ments sont effectués permettant d'obtenir "[3;5]", "[6;10]", "[11-19]" et " ≥ 20 ". Les coefficients de régression sont décroissants avec le nombre de contrats possédé par un assuré.

- *nombre d'activités* : comme la variable précédente la modalité "1" est la plus représentée et n'est pas incluse dans un regroupement. Les modalités "2" et "3" sont associées pour obtenir la catégorie "[2;3]" et ensuite toutes les valeurs supérieures sont regroupées dans la modalité " ≥ 4 ", et les "Non renseigné" ne sont pas regroupés davantage. Toutes choses étant égales par ailleurs, un assuré effectuant 2 ou 3 aura un coût de sinistre prédit plus faible qu'un assuré avec une seule activité. A l'inverse, effectuer 4 activités ou plus, présentent un coût supérieur, tout comme pour l'absence d'information.

Deuxième tranche de sinistralité :

Responsabilité Civile Décennale : La modélisation compte 3 variables explicatives : *l'année*, *le nombre d'activités* et *le Taux*. Pour cette dernière, les modalités "ProfLib" et "Non renseigné" sont regroupées, tandis que les autres modalités sont associées. Pour le nombre d'activités les modalités obtenues sont "[1;2]" et "Non renseigné" et " ≥ 3 ". Cette dernière valeur, ainsi que la modalité regroupant les professions libérales et celles non renseignées pour la variable *Taux* ont des coefficients de régression supérieurs à ceux des autres modalités de leur variable respective.

Responsabilité Civile Bâtiment : Les variables incluses dans le modèle, ainsi que les regroupements effectués sont détaillés ci-dessous :

- *Ancienneté de l'entreprise* : "[0;5]", "[6;10]", "[11;20]", " ≥ 20 " et "non renseigné". Pour deux assurés présentant les mêmes caractéristiques, celui n'indiquant pas l'ancienneté de l'entreprise aura une prédiction de la charge dans la deuxième tranche plus faible que celui ayant indiqué l'année de création de son entreprise. Pour les différentes modalités autres que "Non renseigné", les coefficients de régression sont similaires, avec celui le plus élevé pour "[11;20]".
- *Chiffre d'affaires sous traité* : Deux modalités sont créées "[0;20[" et "[20;100] + NR", cette dernière est la plus représentée et présente un coefficient de régression supérieur à celui de l'autre modalité.
- *Effectif* : la modalité la plus représentée, "[11;20]", est associée aux valeurs non renseignées. Ce regroupement présente un coefficient de régression inférieur à celui de l'autre modalité.
- *Taux* : les modalités F (Fabrication-Négoce), L (Professions Libérales) et X (Tous corps d'états) sont regroupées, cette catégorie présente un coefficient de régression plus faible que celui de l'autre modalité.

Sans suite :

Responsabilité Civile Décennale : Hormis la variable indiquant si le sinistre a été déclaré via ECRAC, il a été nécessaire de réaliser des regroupements de modalités pour avoir une bonne significativité du modèle. Une explication des regroupements effectués est détaillée ci-dessous :

- *Le branchement à l'ECRAC* augmente la probabilité d'être un sinistre sans-suite de charge nulle par rapport à un sinistre non déclaré par l'ECRAC.
- Pour la variable *Région*, plusieurs regroupements sont effectués : la Corse avec la modalité la plus représentée (la région Auvergne Rhône-Alpes et les valeurs non renseignées), et les régions Occitanie et Pays de la Loire. La région Ile-de-France est la modalité augmentant le plus la probabilité d'être un sans suite par rapport à la modalité de référence, tandis que la région Hauts-de-France est celle l'augmentant le moins.

- *nombre de contrats en assurance construction* : les assurés présentant plusieurs contrats en assurance construction sont plus susceptibles d'avoir des sans-suite. Deux modalités sont obtenues dans le modèle final : "1" et " ≥ 2 ".
- *nombre d'activités*, la modalité la plus représentée est "1", elle est regroupée avec les valeurs non renseignées. D'autres regroupements sont effectués pour obtenir les modalités "[2;5]" et " ≥ 6 ". Le fait de réaliser plusieurs activités augmente la probabilité d'avoir des sinistres sans-suite, toutes choses étant égales par ailleurs.
- *Effectif* : les modalités avant 10 effectifs sont regroupés, ainsi que ceux au-dessus de 100. Les coefficients de la régression augmentent avec le nombre d'ETP.
- *délai de survenance* : les délais supérieurs ou égaux à 10 ans sont regroupés pour définir la modalité " ≥ 10 ", les modalités inférieures conservent leur forme initiale (1 modalité par an). A partir de 1 an de délai, les coefficients de régression sont croissants avec le délai de survenance.
- *Chiffre d'affaires* : Divers regroupements ont été effectués afin d'obtenir des modalités significatives. Le modèle final contient les catégories suivantes : "[0;150k]", "[150k;300k]", "[300k;1M]", "[1M;5M]", " $\geq 5M$ " et "Non renseigné". Le coefficient de la régression est croissant avec l'augmentation du chiffre d'affaires.
- Certaines modalités du *chiffre d'affaires sous traités* sont regroupées pour obtenir les valeurs : "[0;40]", "[40;80]", "[80;100]" et "Non renseigné". La modalité "[40;80]" présente le coefficient de régression le plus élevé, tandis qu'ils sont similaires pour les autres modalités.
- *Ancienneté du contrat*, les coefficients sont croissants avec l'ancienneté du contrat.
- *Taux* : les modalités F (Fabrication-Négoce), L (Professions Libérales), N (Non renseigné) et X (Tous corps d'états) sont regroupées, cette catégorie présente un coefficient de régression plus haut que ceux des autres modalités. Ensuite, toutes choses étant égales par ailleurs, un taux plus élevé présentera une probabilité de sans-suite plus importante.

Responsabilité Civile Bâtiment : Le modèle final compte 6 variables explicatives, qui ont nécessité des regroupements de modalités détaillés ci-dessous :

- Pour la variable *Type d'Œuvre*, la modalité Second-Œuvre qui est la plus représentée, est regroupée avec les modalités Fabrication-Négoce et Autre, tandis que le Gros Œuvre est regroupé avec les travaux tous corps d'états. Cette dernière modalité présente un coefficient de régression supérieur à celui de la modalité de référence. Ce coefficient est maximum pour les professions libérales, tandis qu'il est proche de celui de référence pour les valeurs non renseignées.
- *Région* : plusieurs regroupements sont effectués pour avoir toutes les modalités significatives. Premièrement, avec la modalité la plus représentée, la région "Auvergne Rhône-Alpes", les régions "Normandie", "PACA", et "Corse" sont associées. Ce regroupement présente le coefficient de régression le plus faible, tous les autres sont légèrement supérieurs et proches entre eux. Les regroupements finaux sont "Bretagne - Grand-Est", "Centre Val de Loire - Ile-de-France", "Hauts-de-France - Occitanie - Nouvelle Aquitaine" et "Bourgogne - Pays de la Loire".
- *nombre de contrats* : les modalités finales conservées dans le modèle sont "1", "[2;5]", "[6;9]" et " ≥ 10 ". Les coefficients de régression sont décroissants avec le nombre de contrats.
- *L'effectif* : Le coefficient de régression est croissant avec l'effectif. La modalité de référence "[11;20]" est regroupée avec toutes celles associées à un nombre d'ETP supérieur à 10, permettant d'obtenir la catégorie " > 10 ". Avant cette valeur, deux regroupements sont constitués : "[1;3]" et "[4;10]". Enfin une dernière modalité est présente dans le modèle final, il s'agit des valeurs non renseignées.

- *Nombre d'activités* : La modalité la plus représentée "1" est regroupées avec les modalités "2" et "3" permettant d'obtenir la catégorie "[1 ;3]", tandis que toutes les valeurs supérieures sont associées dans la modalité " ≥ 4 ", et les valeurs non renseignées sont la dernière modalité. Toutes choses étant égales par ailleurs, l'absence d'informations sur le nombre d'activités pour un assuré engendrera une prédiction de probabilité de sans-suite plus importante que pour un assuré avec l'information fournie. Parmi les assurés avec un nombre d'activités connu, le coefficient de régression est décroissant avec l'augmentation de ce nombre.

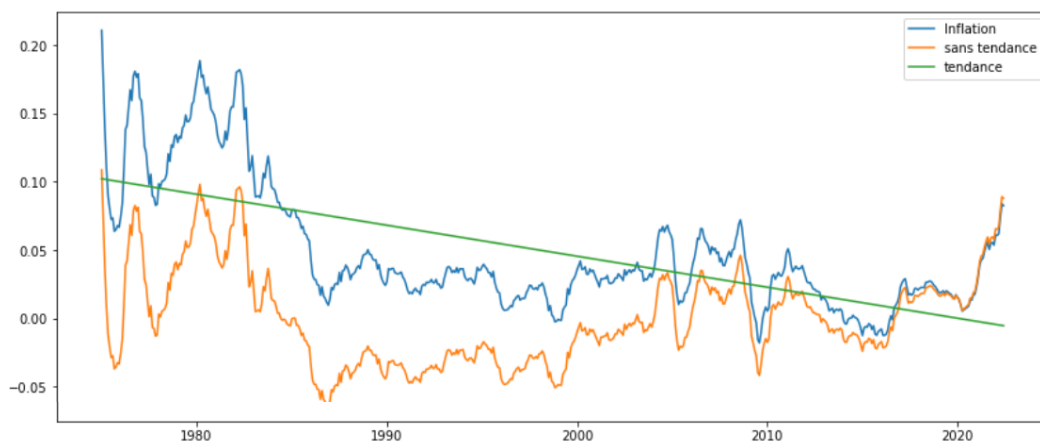
Annexe D : Séries Temporelles

Cette annexe permet de compléter les informations fournies dans la partie 6 sur la modélisation de l'inflation.

Un bruit blanc (faible) est une suite (ϵ_t) de variables centrées, de variances constantes et non corrélées :

$$\mathbb{E}[\epsilon_t] = 0, \quad \text{Var}(\epsilon_t) = \sigma^2, \quad \text{cov}(\epsilon_t, \epsilon_{t'}) = 0 \text{ pour } t \neq t'$$

E.1 Étude de la tendance



Le graphique précédent permet de constater une tendance à la baisse de l'inflation du BT01 depuis les années 1960.

E.2 KPSS Test

Le test de Kwiatkowski–Phillips–Schmidt–Shin (KPSS) détermine si une série temporelle est stationnaire ou non. Une telle série présente des propriétés statistiques, telles que la moyenne et la variance, constantes dans le temps.

Procédure du test :

Le test permet de vérifier l'hypothèse (H_0) : "La série temporelle est stationnaire".

Le test KPSS est basé sur la régression linéaire. Il décompose une série en trois parties : une tendance déterministe (β_t) , une marche aléatoire (r_t) et une erreur stationnaire (ϵ_t) , avec l'équation

de régression :

$$y_t = r_t + \xi_t + \epsilon_t$$

$$r_t = r_{t-1} + u_t.$$

u_t désigne un terme d'erreur de la seconde équation, et il s'agit d'une série de variables aléatoires iid de valeur attendue égale à zéro et à variance constante $\hat{\sigma}_u^2$. Par hypothèse, une valeur initiale de la seconde équation, r_0 , est une constante.

L'hypothèse nulle de stationnarité équivaut à l'hypothèse selon laquelle la variance du processus de marche aléatoire dans l'équation est égale à zéro.

Lorsque, $\xi = 0$, cela signifie que y_t est stationnaire autour de r_0 . Si $\xi \neq 0$, y_t est stationnaire autour d'une tendance linéaire. Si la variance σ_u^2 est supérieure à zéro, alors y_t est non stationnaire (somme d'une tendance et d'une marche aléatoire), en raison de la présence d'une racine unitaire.

En soustrayant y des deux côtés de la première équation de l'équation, nous obtenons :

$$\Delta y = \xi + u_t + \Delta \epsilon_t = \xi + w_t$$

où w_t , en raison de l'hypothèse que ϵ_t et u_t , sont des variables aléatoires indépendantes et identiquement distribuées, est généré par un processus autorégressif AR(1) : $w_t = v_t + \theta v_{t-1}$. Ainsi, le modèle KPSS peut être exprimé sous la forme suivante :

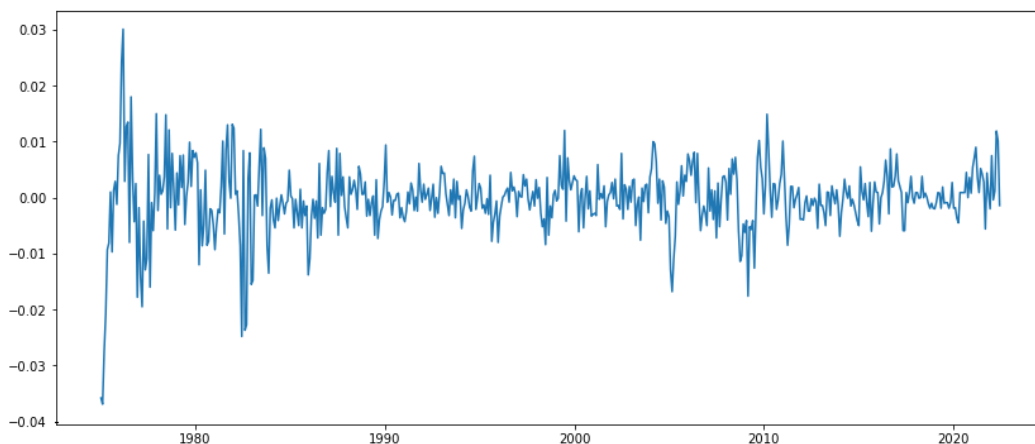
$$y_t = \xi + \beta y_{t-1} + w_t$$

$$w_t = v_t + \theta v_{t-1}, \quad \beta = 1$$

Interprétation des résultats :

Afin de déterminer si une série est stationnaire ou non, il faut consulter la p-value renvoyée par le test et la comparer au niveau alpha de signification choisi pour le test. Par exemple, une p-value de 0,02 (2 %) entraînerait le rejet de l'hypothèse H_0 à un niveau alpha de 0,05 (5 %).

La série différenciée de l'inflation présente la forme suivante :



E.3 Choix des paramètres p et q

Définition des critères de sélection des paramètres :

- **MAPE** : Mean Absolute Percentage Error

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i}$$

- **MAE** : Mean Absolut Error

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- **MSE** : Mean Square Error

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

- **RMSE** : Root Mean Square Error

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

Comparaison des résultats obtenus selon la profondeur d'historique de la base de test choisie :

Prédictions sur 3 ans					
p	q	MAPE	MAE	RMSE	MSE
1	1	0.7343	0.0199	0.0255	0.000652
1	2	0.7190	0.0204	0.0263	0.000694
1	3	0.7173	0.0204	0.0264	0.000699
1	10	0.5977	0.0208	0.0287	0.000823
1	12	0.6836	0.0205	0.0269	0.000728
2	1	0.7374	0.0197	0.0250	0.000630
2	2	0.7198	0.0203	0.0263	0.000691
2	3	0.7185	0.0204	0.0264	0.000697
2	10	0.5893	0.0206	0.0285	0.000816
2	12	0.6777	0.0206	0.0273	0.000743
3	1	0.7249	0.0203	0.0262	0.000684
3	2	0.7150	0.0204	0.0264	0.000698
3	3	0.7208	0.0204	0.0263	0.000693
3	10	0.9598	0.0194	0.0228	0.000518
3	12	0.6782	0.0207	0.0274	0.000750
9	1	0.8026	0.0206	0.0258	0.000665
9	2	0.8229	0.0207	0.0257	0.000661
9	3	0.8329	0.0207	0.0256	0.000654
9	10	0.9851	0.0209	0.0246	0.000607
9	12	0.6848	0.0210	0.0277	0.000767
12	1	0.8404	0.0212	0.0263	0.000693
12	2	0.8770	0.0214	0.0261	0.000682
12	3	0.8909	0.0214	0.0259	0.000671
12	10	0.8534	0.0209	0.0257	0.000663
12	12	0.7942	0.0209	0.0262	0.000692

Prédictions sur 5 ans					
p	q	MAPE	MAE	RMSE	MSE
1	1	1.2030	0.0330	0.0392	0.001542
1	2	0.8347	0.0258	0.0324	0.001057
1	3	0.7583	0.0243	0.0311	0.000969
1	10	1.6094	0.0410	0.0462	0.002136
1	12	0.557	0.0195	0.0272	0.000738
2	1	0.9309	0.0277	0.0341	0.001163
2	2	0.8583	0.0263	0.0328	0.001079
2	3	0.8759	0.0266	0.0331	0.001094
2	10	1.5105	0.0391	0.0443	0.001964
2	12	0.5891	0.0205	0.0280	0.000786
3	1	0.9624	0.0283	0.0346	0.001198
3	2	0.7994	0.0251	0.0319	0.001015
3	3	0.8617	0.0264	0.0329	0.001080
3	10	0.9069	0.0162	0.0214	0.000713
3	12	0.6853	0.0229	0.0300	0.000902
9	1	0.8568	0.0263	0.0327	0.0010708
9	2	0.9192	0.0275	0.0337	0.0011372
9	3	1.0127	0.0294	0.0354	0.0012507
9	10	1.1153	0.0312	0.0372	0.001385
9	12	0.8260	0.0258	0.0324	0.001050
12	1	0.7569	0.0233	0.0303	0.000920
12	2	0.7809	0.0237	0.0307	0.000944
12	3	0.7892	0.0239	0.0310	0.000959
12	10	0.6121	0.0212	0.0291	0.000845
12	12	0.6468	0.0217	0.0293	0.000859

Prédictions sur 10 ans					
p	q	MAPE	MAE	RMSE	MSE
1	1	2.8521	0.0148	0.0205	0.000418
1	2	3.2328	0.0147	0.0202	0.000409
1	3	3.6460	0.0148	0.0202	0.000410
1	10	3.2507	0.0278	0.0345	0.001195
1	12	5.1891	0.0175	0.0219	0.000483
2	1	1.5279	0.0171	0.0240	0.000572
2	2	3.1223	0.0146	0.0203	0.000410
2	3	3.4674	0.0147	0.0202	0.000408
2	10	2.1931	0.0234	0.0307	0.000942
2	12	4.4068	0.0157	0.0208	0.000432
3	1	3.0604	0.0147	0.0203	0.000411
3	2	3.2680	0.0147	0.0202	0.000408
3	3	3.3663	0.0147	0.0202	0.000408
3	10	1.5164	0.0145	0.0201	0.000407
3	12	4.0466	0.0151	0.0204	0.000415
9	1	4.7242	0.0166	0.0213	0.000455
9	2	4.6864	0.0165	0.0213	0.000452
9	3	5.4685	0.0190	0.0232	0.000538
9	10	4.3484	0.0158	0.0208	0.000433
9	12	5.0946	0.0166	0.0214	0.000459
12	1	5.3592	0.0188	0.0230	0.000531
12	2	5.3468	0.0187	0.0230	0.000527
12	3	5.2453	0.0184	0.0227	0.000517
12	10	4.2434	0.0158	0.0209	0.000436
12	12	4.6626	0.0162	0.0212	0.000449