Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Madame Mame Aminata DIALLO

Titre du mémoire : Analyse de l'impact de la pandémie sur les flottes automobiles à travers les variables télématiques (appliqué à la flotte ALD-Italie)

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

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

Membres présents du jury de la filière : Signature :

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

Entreprise :

Nom : SOGECAP

Signature : 

Directeur de mémoire en entreprise

Nom : Paul SCHMITT-BAILER

Signature : 

Invité :

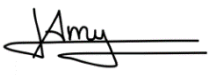
Nom : Thibault VAN-EVERBROECK

Signature : 

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

Signature du responsable entreprise :

Signature du candidat :



Résumé

Retour en 2020, une pandémie a bouleversé les sociétés et changé la façon de penser sur l'assurance automobile. Le monde a été plongé dans la tourmente pandémique de la Covid-19, et le télétravail est devenu l'un des changements notables qui composent la "nouvelle normalité". La réduction de la conduite automobile observée dans l'industrie a conduit les consommateurs et les assureurs à réévaluer la valeur des polices d'assurance automobile traditionnelles. Les consommateurs se demandent pourquoi ils devraient payer une prime d'assurance annuelle pour un véhicule stationné la plupart du temps et pourquoi les primes ne peuvent-elles pas refléter leurs habitudes de conduite.

A l'heure actuelle, les consommateurs deviennent de plus en plus tolérants quant au partage de leurs données personnelles, en particulier si cela est synonyme de recevoir une prime moins élevée. Cela constitue une bonne nouvelle pour l'assurance flotte automobile : l'acceptation accrue par les consommateurs des outils numériques dans le processus de souscription et de réclamation des polices, combinée à la nécessité pour les assureurs de tarifer les polices plus précisément dans un marché automobile très concurrentiel, et les changements dus au COVID-19, a créé les conditions idéales pour que les assureurs passent des modèles actuariels et de souscription traditionnels aux modèles télématiques.

Le but de ce mémoire est de capter l'impact de la pandémie croisé avec la contribution des paramètres télématiques dans les flottes automobiles. Pour ce faire, nous allons modéliser la fréquence et le coût moyen des sinistres matériels sur la garantie responsabilité civile dans la flotte ALD Italie en se basant sur les modèles linéaires généralisées (GLM).

Ainsi, dans la première partie, nous allons expliquer plus en détail le contexte ainsi que les différentes données qui vont être utilisées pour la modélisation. Dans la seconde partie, nous présenterons la tarification automobile et la théorie sur les GLM avant d'effectuer une analyse détaillée des données à notre disposition. Dans la troisième partie, dans le but de déterminer la prime de risque, nous étudierons l'impact de la pandémie en exposant les différents modèles mis en place pour modéliser la fréquence et le coût moyen des sinistres matériels. Pour finir, nous essayerons de concrétiser notre étude en effectuant des conjectures sur la prime pure avant d'exhiber la corrélation de cette analyse avec les effets macro-économiques.

Mots-clés *: Covid-19, Télématique, score, Assurance flotte automobile, modélisation, GLM, sinistralité, garantie responsabilité civile (RC)*

Ce mémoire se concentre principalement sur la télématique. La mise en place d'un score de conduite à déjà fait l'objet d'un mémoire réalisé par Christian CHOW. (URL disponible dans les références).

Abstract

Back in 2020, a pandemic upended societies and changed the way people think about car insurance. The world has been thrown into the pandemic turmoil of Covid-19 and working from home has become one of the notable changes that make up the “new normal”. The reduction in driving observed across the industry has led consumers and insurers alike to reevaluate the value of traditional motor insurance policies. Consumers wonder why they should pay an annual premium for a vehicle parked on their driveway most of the time, and why premiums can't reflect their driving habits.

At present, consumers are becoming more comfortable sharing their personal data, particularly if it means receiving a lower premium. This is good news for fleet insurance : the increased acceptance by consumers of digital tools in the policy binding and claims process, combined with a need for insurers to rate policies more precisely in a fiercely competitive auto market, and changes in society due to COVID-19, created the perfect conditions for insurers to shift from traditional actuarial and underwriting models to telematics models.

The purpose of this thesis is to capture the impact of the pandemic crossed with the contribution of telematics parameters in car fleets. To do this, we are going to model the frequency and average cost of material claims on MTPL guarantee in the ALD Italy fleet based on generalized linear models (GLM).

Thus, in the first part, we will explain in more detail the context as well as the different data that will be used for the modeling. In the second part, we will present car pricing and theory on GLMs before carrying out a detailed analysis of the data at our disposal. In the third part, with the aim of determining the risk premium, we will study the impact of the pandemic by exposing the different models put in place to model the frequency and average cost of material losses. Finally, we will try to concretize our study by making conjectures on the pure premium before showing the correlation of this analysis with the macroeconomic effects.

***Mots-clés** : Covid-19, Telematics, score, car fleet insurance, modeling, GLM, claims, MTPL guarantee*

Note de synthèse

Notre objectif tout au long de ce mémoire est de modéliser la prime pure de la garantie RC sur des contrats d'assurance flotte automobile afin de relever d'une part l'impact de la crise sanitaire (COVID-19) et d'autre part l'apport des variables télématiques dans cette modélisation.

La télématique qui bénéficie de plus en plus d'une côte de popularité dans la gestion des flottes automobiles, est l'un des facteurs qui apporte une possibilité de renouvellement dans la manière de fonctionner : anticiper la probabilité de survenance d'un sinistre. Les usagers souhaitent de plus en plus des contrats personnalisés, connectés et adaptés à leurs modes de déplacement. L'intérêt pour les modèles d'assurance connectés a fortement augmenté en Europe depuis le début de la pandémie :

- L'assurance au kilomètre : également connue sous le nom d'assurance « Pay as you drive » (pour « Payez comme vous conduisez »).
- L'assurance à l'usage classique : basée essentiellement sur le comportement de conduite.
- L'assurance qui récompense : également basée sur le comportement, avec la différence que le score de conduite n'est pas directement lié à la prime.

Selon l'Institut de Contrôle des Assurances en Italie (IVASS – Istituto per la Vigilanza Sulle Assicurazioni), 21,2 % des contrats stipulés au deuxième trimestre 2022 comportent une clause ayant pour effet de réduire la prime liée à la présence d'une blackbox télématique.

Depuis le début de la crise de la COVID-19, les différentes périodes de confinement qu'a vécues le monde en 2020 et 2021 ont eu un impact important sur la mobilité des usagers, qui ont été nombreux à ne plus se servir de leur voiture quotidiennement. Les mois qui ont suivi la première vague de la COVID-19 ont découvert des défis que peu auraient pu prévoir et moins encore auraient pu se préparer. Les impacts de la pandémie se sont répercutés sur de nombreuses branches de l'assurance, mais l'une des conséquences les plus directes à l'échelle mondiale a été visible dans l'assurance automobile qui a enregistré un recul de sinistralité d'environ -20% en 2020.

Cela n'a pas laissé indifférent la sinistralité des portefeuilles de flottes automobiles. Selon le rapport (2020) de l'Ania (Association Nationale des compagnies d'assurance en Italie), les primes ont diminué (de près de 6 %). La fréquence des sinistres, a connu une baisse considérable à la suite du changement significatif de la mobilité et en raison des mesures restrictives. Le coût moyen relatif des sinistres a considérablement augmenté (+13 %) comparé à l'année précédente.

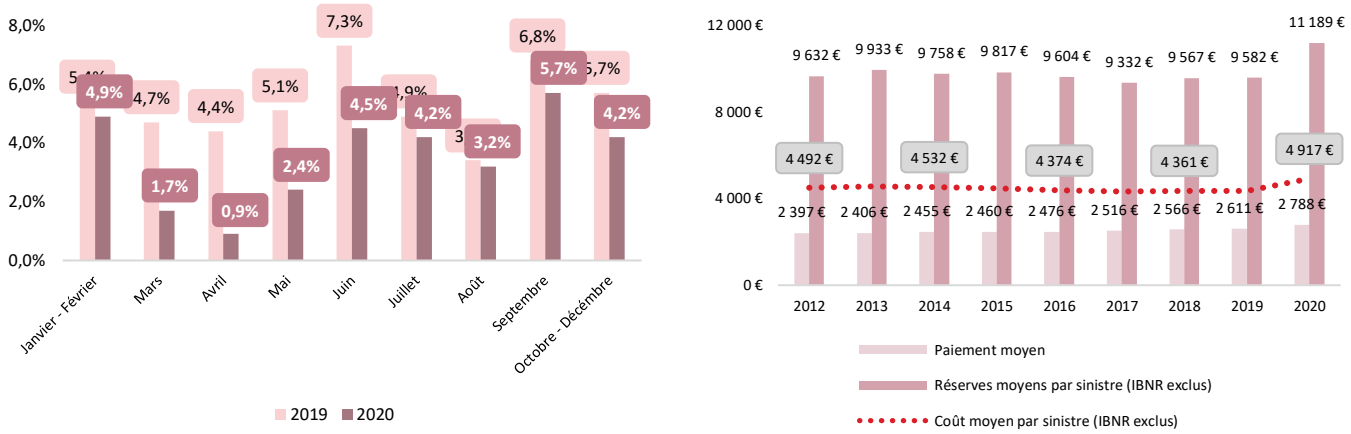


Figure 1 – Evolution de la fréquence de sinistralité entre 2019 et 2020 et du coût moyen de 2012 à 2020.

Le résultat, malgré une crise drastique baisse des bénéfices de placement, a été une amélioration du résultat du compte technique pour cette branche.

Données

Plusieurs bases de données de SGA (Société Générale Assurance) ont été utilisées pour créer la base de modélisation. Les données utilisées sont issues de la flotte automobile ALD Italie. C'est une grande flotte automobile comprenant plus de 120 000 véhicules assurés par Sogessur sur la garantie responsabilité civile.

Pour les données assurancielles, trois bases de données ont été utilisées : une base contrat historisée, et deux bases sinistres historisées. Ces bases sont fournies pas STI (Support Technique Infocentre de Sogessur), ce qui fait d'elles des bases très exploitables, bien renseignées qui ne présentent pas d'anomalies particulières. Ce sont des bases historisées c'est-à-dire qu'elles contiennent tous les détails des contrats depuis le début du partenariat avec ALD Italie.

	Y6	Y7	Y8	Y9	Y10
Nombre de véhicules	138 718	156 469	138 753	127 455	127 670
Exposition	135 580	159 787	144 715	130 056	52 917
Nombre de Sinistres Garantie	13 437	15 473	9 497	8 729	6 589

Tableau 1 – Nombre de véhicules et nombre de sinistres dans les bases assurancielles (contrats et sinistres)

**Y10 que sur 6 mois : de novembre 2021 à avril 2022

D'autres types de données sont aussi utilisées dans la créations de la base finale : il s'agit des données télématiques. Les données contenues dans la base télématique proviennent d'un prétraitement préalable fait sur des données brutes à la maille journalière mais aussi d'une construction d'un score de conduite journalier, mensuel et même annuel.

Ce prétraitement consiste à corriger les erreurs lors de la transmission des données car il arrive que le boîtier ou le récepteur présente des défauts de fonctionnement pouvant causer un manque de données dans la base. Il permet également de déterminer le score de conduite pour chaque véhicule. Le score de conduite est une variable, comprise entre 0 et 100, qui capte le profil de risque d'un individu en fonction de sa manière de conduire. Il est créé à partir de 4 sous scores comportementaux (par événement : accélération, décélération, virage et mouvement brusque) Le score est construit sur la base de la sinistralité empirique observée et des événements de conduites relevés sur la flotte équipée. Une représentation simplifiée du modèle de scoring est la suivante :

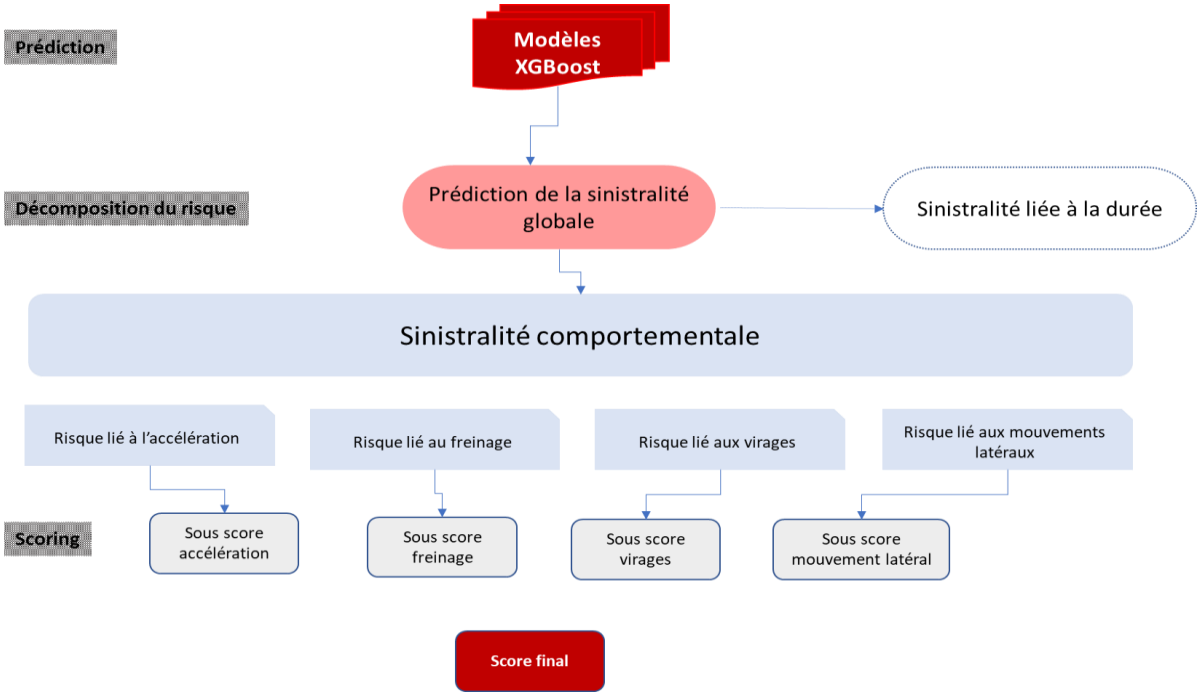


Figure 2 – Procédé de détermination du score de conduite final

La partie prédiction est réalisée suivant un modèle de XGBoost qui commence par créer un premier modèle de prédiction, donnant à chaque variable un coefficient d'ajout ou de diminution de la probabilité d'avoir un sinistre en fonction de la valeur de la variable. On peut remarquer que sur le schéma ci-dessus lors de la partie « décomposition du risque », l'impact de la durée sur la sinistralité est mis de côté. Le risque comportemental est ensuite scindé en quatre sous scores pour enfin produire un score global journalière, issue d'une moyenne pondérée des quatre sous scores en fonction de leur poids dans la sinistralité comportementale totale.

Le tableau suivant présente les principales variables explicatives constituant la base de modélisation :

Variables Contrat	**Variables Télématiques**
- Année contrat - Age du véhicule -Type d'énergie -Secteur d'activité - Zone* -Taille de la flotte -Nombre de chevaux fiscaux du véhicule -Marque du véhicule	-Distance annuelle parcourue -Durée annuelle effectuée -Nombre de jours annuel conduit -Score accélération -Score décélération -Score mouvement brusque -Score virage -Score globale annuelle -Taux de roulage urbain (TRU)

Tableau 2 – Les principales variables explicatives dans la base de données créée

La réalisation de statistiques descriptives de la base créée a permis de relever les corrélations existantes entre les variables télématiques. Par exemple, le nombre de jours roulés par un véhicule dans l'année est fortement corrélée à la durée et à la distance parcourues par année, qui sont elles-mêmes corrélées entre elles. Ou encore une dépendance entre le score globale de conduite et le Taux de Roulage Urbain (TRU, rapport de la distance effectuée sur les routes urbaines sur la distance globale) : plus le taux de roulage urbain est bas, plus le score global de conduite est élevé (en particulier, on note qu'un TRU inférieur à 20, avec de fortes chances, correspondrait à un score global supérieur à 80 et un TRU supérieur à 70 à un score global inférieur à 30).

Modélisation GLM

Dans ce mémoire, nous avons modélisé la fréquence de sinistralité et la sévérité suivant les variables contrats et suivant les variables télématiques pour mettre en exergue les effets de la pandémie mais aussi la contribution des variables télématique dans la détermination de la prime pure sur la garantie RC flotte automobile. Les modèles étudiés sont les suivants :

Modèles		Descriptions
Fréquence	F1	Prend en compte que les variables contrat (âge véhicule, marque, segmentation car-van…)
	F1-bis	Modèle F1 + variable « *Covid* »
	F2	Prend en compte les variables télématiques (score, distance…) plus quelques variables contrat
	F2-bis	Modèle F2 + variable « *Covid* »
Coût moyen	C1	Variables contrat + variable « *Covid* »
	C2	Variables télématiques plus quelques variable contrat + variable « *Covid* »

Tableau 3 – Les différents modèles de fréquence et de sévérité étudiés

En ajoutant la variable « Covid » aux modèles F1 et F2, nous créons respectivement les modèles F1-bis et F2-bis. L'objectif étant de faire un choix entre F1 et F1-bis mais aussi entre F2 et F2-bis. Le choix se fait suivant plusieurs critères :

- Critères d'évaluation de l'adéquation par AIC et BIC
- Mesure du RMSE de chaque modèle
- Pouvoir prédictif de chaque modèle dans une base test créé au préalable

Dans le modèle de fréquence matérielle télématique, toutes les variables sont significatives et le score de conduite permet de capter une partie de l'information sur le conducteur. En effet dans les flottes automobiles, nous ne disposons pas d'informations relatives aux conducteurs dans la base de données (absence d'informations sur le conducteur).

Suivant le *RMSE*, *l'AIC* et le *BIC*, les choix portent sur les modèles F1-bis et F2-bis :

	RMSE apprentissage	*RMSE* test	*AIC*	*BIC*
F1	0,28106	0,26877	171 639	171 896
F1-bis	0,28087	0,26856	171 171	171 439
F2	0,28110	0,26834	170 744	170 022
F2-bis	0,28055	0,26793	169 794	169 084

Tableau 4 – Comparaison des différents modèles de fréquence

Ces choix prouvent l'effet positif de l'ajout de la variable « *Covid* » dans les deux modèles.

Aussi, segmenter selon les variables télématiques constitue une hypothèse plausible étant donné le fort pouvoir explicatif de ces dernières sur la sinistralité matérielle. En effet, en estimant le nombre de sinistres dans la base test selon les modèles F1-bis et F2-bis, nous obtenons une différence négligeable entre la sinistralité observée et la sinistralité prédite par le modèle :

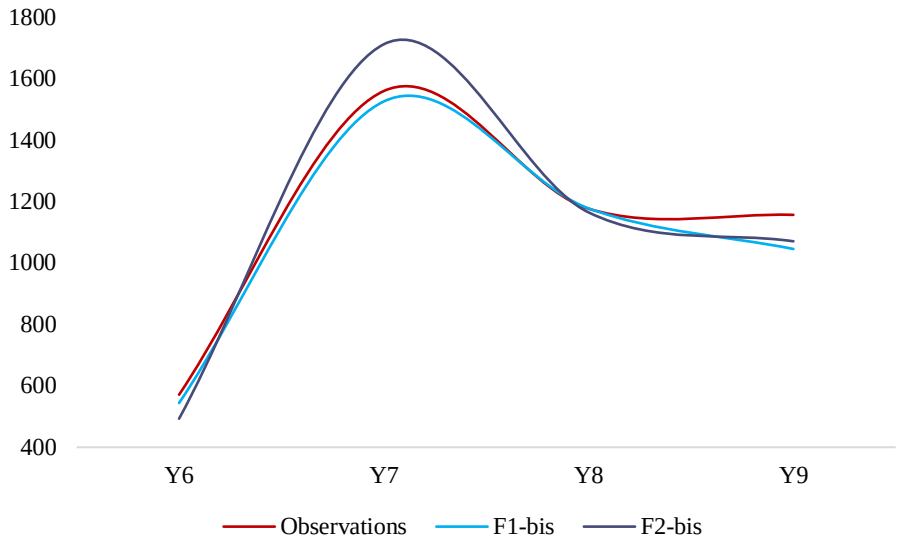


Figure 3 – Nombre de sinistres observés vs prédits par les modèle F1-bis et F2-bis

Dans les modèles de coût moyen matériel, certaines modalités de certaines variables ne sont pas significatives. Le coût moyen estimé par les modèles, révèle que le modèle télématique appliqué au portefeuille estime plus de coûts élevés que le modèle contrat. Le modèle C2 est donc un modèle qui surestime la charge. Par conséquent il peut être considéré comme un modèle prudent, ce qui est un atout en assurance, même si l'objectif principal est de rester au plus près du coût réel des sinistres. Aussi, le score de conduite représente un effet levier quant aux informations contenues dans le modèle.

En nous intéressant maintenant à la valeur de la prime pure, nous remarquons une baisse de la prime pendant les années Covid ; le modèle C1 révèle une baisse de 23% tandis que le modèle C2 laisse paraître une baisse plus grande de 27%. Suivant les années et suivant chaque modèle, nous remarquons que la prime pure obtenue sur le modèle télématique est presque identique à la prime pure observée. Le modèle contrat quant à lui estime bien aussi la prime pure. En même temps, nous observons cette baisse de la prime de risque évoquée pendant les années COVID :

	Contrat	Télématique	Observée
Y6	215 €	216 €	215 €
Y7	200 €	203 €	203 €
Y8	150 €	148 €	148 €
Y9	150 €	152 €	156 €

Tableau 5 – Prime de risque par année selon les modèles contrat et télématique et selon l'observée

Au vu de ces analyses nous avons pu capter l'impact de la crise de la COVID-19 dans la prime pure en flaguant les années mais aussi l'apport des variables télématiques dans celle-ci. Nous avons mis en place deux modèles pour la fréquence et deux modèles pour le coût moyen. Selon ces modèles, ceux de la télématique ont été retenus au vu du pouvoir de prédiction.

Après cette étape, nous avons projeté les modèles contrat (F1-bis) et télématique (F2-bis) sur l'année Y10. En effet, jusqu'ici, nous avons entraîné et testé les différents modèles construits sur les années Y6 à Y9, nous allons maintenant essayer d'indiquer la prime de risque sur l'année Y10 (6 premiers mois : de novembre 2021 à avril 2022) en testant deux scénarios : d'abord nous regardons ce que donnerait la prime si on considère Y10 comme une année COVID puis on regarde ce que cela donnerait dans le cas contraire. Nous obtenons dès lors les résultats suivants en fonction de la segmentation car/van :

		Contrat	Télématique	Observée
COVID=0	**Car**	174 €	169 €	127 €
	Van <2100kg	231 €	217 €	150 €
	Van >=2100kg	439 €	431 €	346 €
	PP Globale	191 €	185 €	140 €

		Contrat	Télématique	Observée
COVID=1	**Car**	136 €	123 €	127 €
	Van <2100kg	181 €	158 €	150 €
	Van >=2100kg	344 €	314 €	346 €
	PP Globale	150 €	135 €	140 €

Tableau 6 – Comparaison de scénarios pour l'année Y10

Selon la prime pure observée et les primes pures obtenues avec les modèles contrat et télématique, nous pouvons voir que l'année Y10 ressemble plus à une année COVID qu'à une année non-COVID. En effet, si l'année Y10 était considérée comme une année non-COVID, cela signifierait que les deux modèles n'estiment pas bien la prime pure puisqu'ils prédisent des primes de risques très élevées comparées à l'observé.

Bilan et perspectives

Ce mémoire montre qu'il est possible de prédire la prime de risque en modélisant la fréquence de sinistralité et le coût moyen sur une période donnée. Cette modélisation de la fréquence et du coût moyen des sinistres matériels pour la garantie RC s'est faite suivant deux approches. Nous sommes partis d'un modèle contenant uniquement des variables contrats, puis à ce modèle, nous avons ajouté la variable « COVID » afin de relever l'effet global de la pandémie sur la sinistralité et la sévérité. Par la suite, dans l'optique de relever l'efficacité de l'intégration des variables télématiques dans la modélisation, nous nous sommes penchés sur l'étude des modèles prenant en compte ces variables.

Sur la modélisation de la fréquence matérielle, les modèles contrat et télématique ajustent bien tous les deux la sinistralité (rapprochement des sinistralités observée et estimée). Néanmoins Il serait bénéfique d'approfondir le modèle télématique puisqu'il contient une information capitale, la seule d'ailleurs (dans le cas de nos flottes), sur le conducteur qu'est le score global de conduite. Cette variable reflète parfaitement le comportement conducteur au volant. Quant à la modélisation du coût moyen matériel, nous avons vu que le modèle télématique surestime la charge et que par conséquent il peut être considéré comme un modèle prudent.

Il est donc possible d'imaginer diverses manières d'intégrer les variables télématiques dans la tarification afin de construire une politique tarifaire adaptée à l'offre souhaitée. Comme nous l'avons vu, nous pouvons envisager une hausse du tarif pour les conducteurs avec des scores bas et une baisse pour les bons scores. Néanmoins, il n'est pas question de pénaliser les mauvais conducteurs mais d'appliquer des abattements tarifaires uniquement aux conducteurs ayant obtenu de bons scores.

Dans le dernier chapitre, nous avons mis l'accent sur les la corrélation du tarif (de la prime) avec les effets macro-économiques. Compte tenu des cicatrices laissées par la COVID-19 et des retombées de la guerre en Ukraine, une croissance à long terme serait difficile à envisager, même pour l'assurance connectée. Il serait donc intéressant d'envisager des mesures en tenant compte de tous les facteurs qui pourraient avoir une portée dans le tarif. L'adoption d'une nouvelle convention en serait une très bonne illustration.

Synthesis Note

Our objective throughout this thesis is to model the pure premium of the MTPL guarantee on car fleet insurance contracts to identify on the one hand the impact of the pandemic (COVID-19) and on the other hand the contribution of telematics variables in this modeling.

Telematics, which enjoys more and more popularity in the management of automobile fleets, is one of the factors that brings a possibility of renewal in the way of operating: anticipating the probability of occurrence of a claim. Users increasingly want personalized contracts, connected, and adapted to their modes of travel. Interest in connected insurance models has increased sharply in Europe since the start of the pandemic:

- Insurance per kilometer also known as “Pay as you drive” insurance.
- Classic use insurance: based essentially on driving behavior.
- Insurance that rewards: also based on behavior, with the difference that the driving score is not directly linked to the premium.

According to the Insurance Control Institute in Italy (IVASS – Istituto per la Vigilanza Sulle Assicurazioni), 21.2% of contracts stipulated in the second quarter of 2022 include a clause having the effect of reducing the premium linked to the presence of a blackbox.

Since the start of the COVID-19 crisis, the various periods of confinement experienced by the world in 2020 and 2021 have had a significant impact on the mobility of users, many of whom have no longer used their cars. daily. The months since the first wave of COVID-19 have uncovered challenges few could have foreseen and fewer still could have prepared for. The impacts of the pandemic have reverberated across many lines of insurance, but one of the most direct consequences globally was visible in motor insurance which recorded a decline in claims of approximately - 20% in 2020.

This did not leave the loss experience of automobile fleet portfolios indifferent. According to the report (2020) of Ania (National Association of Insurance Companies in Italy), premiums have decreased (by almost 6%). The frequency of claims has fallen considerably following the significant change in mobility and because of the restrictive measures. The relative average cost of claims increased considerably (+13%) compared to the previous year.

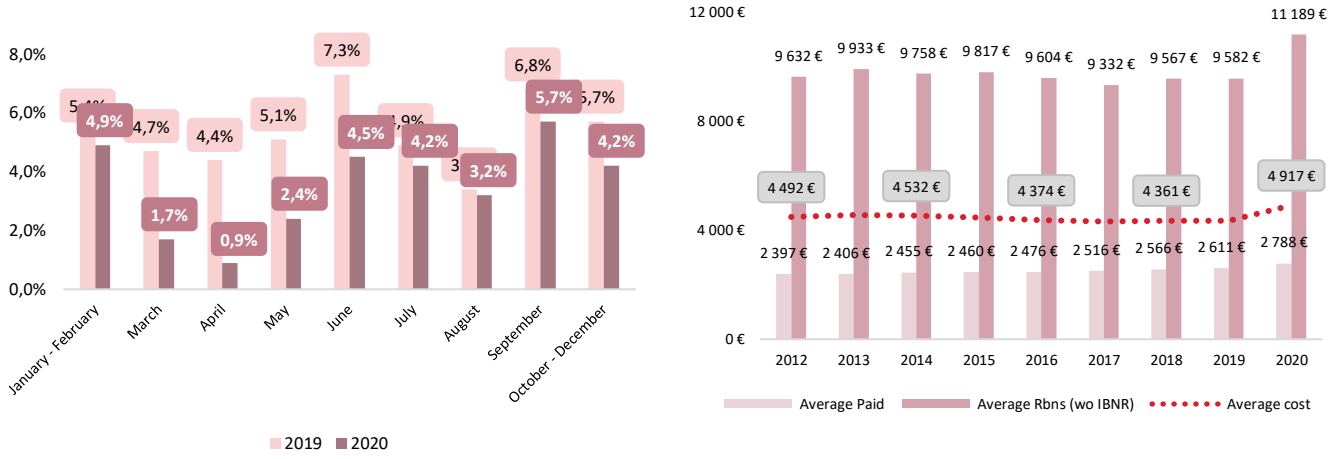


Figure 1 – Evolution of the claims frequency between 2019 and 2020 and the average cost from 2012 to 2020.

The result, despite a drastic drop in investment profits, was an improvement in the technical account result for this branch.

Data

Several SGI (Société Générale Insurance) databases were used to create the modeling database. The data used comes from the ALD Italy car fleet. It is a large car fleet having more than 120 000 vehicles insured by Sogessur on the MTPL guarantee.

For the insurance data, three databases were used: a historical contract database, and two historical claims databases. These databases are provided by STI (Technical Support Info Centre of Sogessur), which makes them very usable, well-informed databases that do not present any anomalies. These are historical databases, i.e., they contain all the details of the contracts since the start of the partnership with ALD Italy.

	Y6	Y7	Y8	Y9	Y10
Number of vehicles	138 718	156 469	138 753	127 455	127 670
Exposure	135 580	159 787	144 715	130 056	52 917
Guaranteed claims	13 437	15 473	9 497	8 729	6 589

Table 6 – Nombre de véhicules et nombre de sinistres dans les bases assurancielles (contrats et sinistres)

** Y10 only over 6 months: from November 2021 to April 2022

Other types of data are also used in the creation of the final database: these are telematics data. The data contained in the telematics database come from a prior pre-processing done on raw data at the daily mesh but also from the construction of a daily, monthly, and even annual driving score.

This preprocessing consists in correcting the errors during the transmission of the data because it happens that the box presents operating faults which can cause a lack of data in the base. It also helps determine the driving score for each vehicle. The driving score is a variable, between

0 and 100, which captures an individual's risk profile based on their driving style. It is created from 4 behavioral sub-scores (score per event: acceleration, deceleration, cornering, and sudden movement) The score is constructed based on observed empirical claims and driving events recorded on the equipped fleet. A simplified representation of the scoring model is as follows:

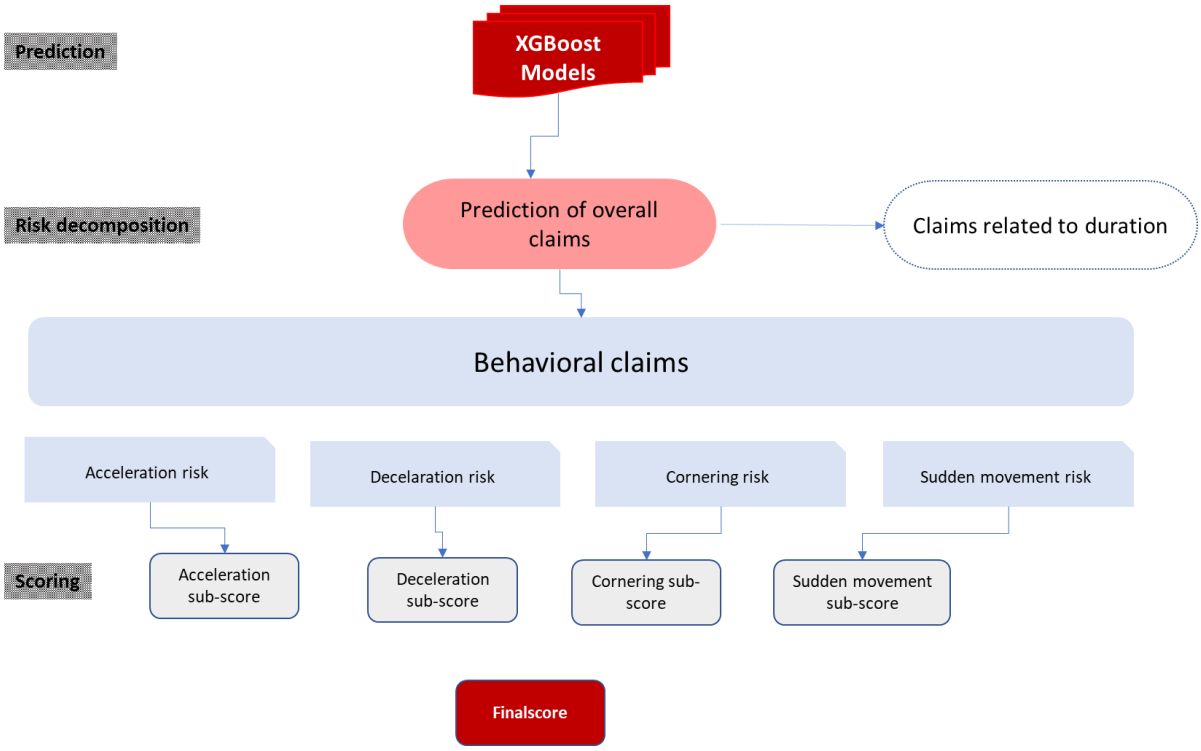


Figure 4 – Method of final driving score calculation

The prediction part is carried out according to an XGBoost model which begins by creating a first prediction model, giving each variable a coefficient of addition or reduction of the probability of having a claim depending on the value of the variable. We can notice that on the diagram above during the "risk breakdown" part, the impact of the duration on the loss ratio is set aside. The behavioral risk is then split into four sub-scores to finally produce an overall daily score, derived from a weighted average of the four sub-scores according to their weight in the total behavioral loss experience.

The following table presents the main explanatory variables constituting the modeling base:

Contract variables	**Telematics variables**
- Contract year - Vehicle age - Energy type - Business area - Zone* - Fleet size -Number of tax horses of the vehicle -Vehicle brand	-Annual distance -Annual duration -Number of annual driven days -Acceleration score -Deceleration score -Sudden movement score -Cornering score -Annual global score - Urban Driving Rate (TRU)

The realization of descriptive statistics of the base created allow us to identify the existing correlations between the telematic variables. For example, the number of days driven by a vehicle in the year is strongly correlated with the duration and distance traveled per year, which are themselves correlated with each other. Or a dependence between the overall driving score and the Urban Driving Rate (TRU, ratio of the distance traveled on urban roads to the overall distance): the lower the urban driving rate, the higher the overall driving score. high (in particular, we note that a TRU lower than 20, with strong chances, would correspond to an overall score higher than 80 and a TRU higher than 70 to an overall score lower than 30).

GLM Modeling

In this thesis, we modeled the claims frequency and the average cost according to the contract variables and according to the telematics variables to highlight the scope of the pandemic but also the contribution of the telematics variables in the determination of the pure premium on the MTPL guarantee for car fleet. The models studied are the following:

Models		Descriptions
Frequency	F1	Only contract variables (vehicle age, brand, segmentation car/van…)
	F1-bis	F1 model + « *Covid* » variable
	F2	Telematics variables (score, duration…) and some contract variables
	F2-bis	F2 model + « *Covid* » variable
Average cost	C1	Contract variables + « *Covid* » variable
	C2	Telematics variables + some contract variable + « *Covid* » variable

Table 7 – Main explanatory variables in the created database

By adding the “Covid” variable to F1 and F2 models, we create, respectively, F1-bis and F2-bis models. The objective is to make a choice between F1 and F1-bis but also between F2 and F2-bis. The choice is made according to several criteria:

- Suitability assessment criteria by AIC and BIC
- Measurement of the RMSE of each model
- Predictive power of each model in a test database created beforehand

In the telematics frequency model, all the variables are significant, and the driving score captures part of the information about the driver. In fact, in car fleets, we do not have any driver information in the database (lack of driver information).

Depending on the RMSE, the AIC and the BIC, the choices relate to F1-bis and F2-bis models:

	Learning RMSE	*Test RMSE*	*AIC*	*BIC*
F1	0,28106	0,26877	171 639	171 896
F1-bis	0,28087	0,26856	171 171	171 439
F2	0,28110	0,26834	170 744	170 022
F2-bis	0,28055	0,26793	169 794	169 084

Table 8 – Different frequency and severity models studied

These choices prove the positive effect of adding the "Covid" variable to both models.

Also, segmenting according to telematics variables constitutes a plausible hypothesis given the strong explanatory power of the latter on material claims. Indeed, by estimating the number of claims in the test base according to the F1-bis and F2-bis models, we obtain a negligible difference between the observed claims and the claims predicted by the model:

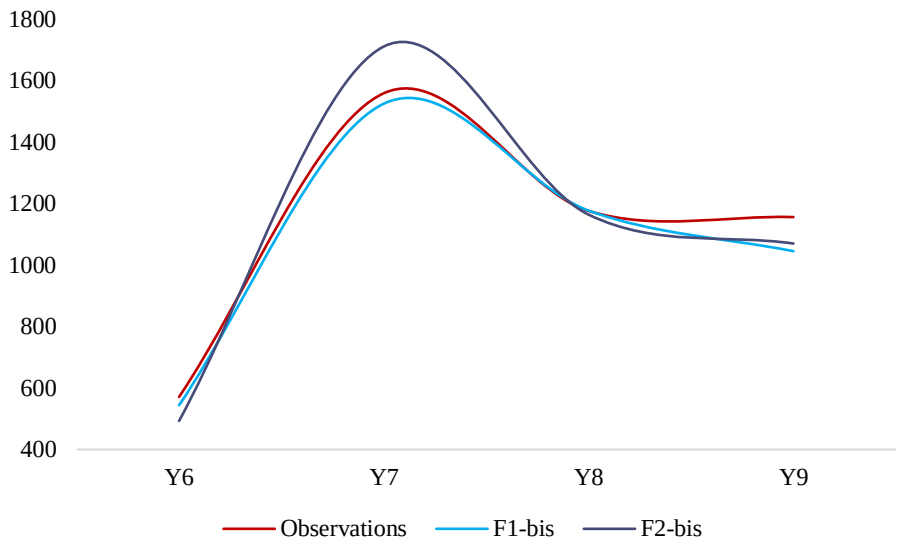


Figure 5 – Number of claims observed vs predicted by F1-bis F2-bis models

In average cost models, some modalities of some variables are not significant. The average cost estimated by the models reveals that the telematics model applied to the portfolio estimates more high costs than the contract model. The C2 model is therefore a model that overestimates the claims charge. Consequently, it can be considered a conservative model, which is an asset in insurance, even if the main objective is to stay as close as possible to the real cost of claims. Also, the driving score represents a leverage effect on the information contained in the model.

Looking now at the value of the pure premium, we notice a decline in the premium during the Covid years; the C1 model reveals a decrease of 23% while the C2 model presents a larger decrease, around 27%. According to the years and according to each model, we notice that the pure premium obtained on the telematics model is almost identical to the pure premium observed. The contract model also correctly estimates the pure premium. At the same time, we observe this decrease in the risk premium reported during the COVID years:

	Contrat	Télématique	Observée
Y6	215 €	216 €	215 €
Y7	200 €	203 €	203 €
Y8	150 €	148 €	148 €
Y9	150 €	152 €	156 €

Table 9 – Observed burning cost per year vs burning cost according to contract and telematics models

In view of these analyzes we were able to capture the impact of the COVID-19 crisis in the pure premium by flagging the years but also the contribution of telematics variables in it. We set up two models for frequency and two models for average cost. According to these models, those of telematics were retained in view of the power of prediction.

After this step, we have projected the contract (F1-bis) and telematics (F2-bis) models on the year Y10. Indeed, so far, we have trained and tested the different models built on the years Y6 to Y9, we will now try to determine the risk premium on the year Y10 (first 6 months: from November 2021 to April 2022) by testing two scenarios: first we look at what the burning cost will look like if we consider Y10 as a COVID year, then we look at what it would look like in the opposite case. We, therefore, obtain the following results according to the car/van segmentation:

		Contrat	Télématique	Observée
COVID=0	Car	174 €	169 €	127 €
	Van <2100kg	231 €	217 €	150 €
	Van >=2100kg	439 €	431 €	346 €
	PP Globale	191 €	185 €	140 €

		Contrat	Télématique	Observée
COVID=1	Car	136 €	123 €	127 €
	Van <2100kg	181 €	158 €	150 €
	Van >=2100kg	344 €	314 €	346 €
	PP Globale	150 €	135 €	140 €

Tableau 6 – Scenarios comparison for year Y10

According to the observed burning cost and the ones obtained with the contract and telematics models, we can see that the year Y10 looks more like a COVID year than a non-COVID year. Indeed, if the year Y10 were considered as a non-COVID year, this would mean that the two models do not estimate the burning cost well since they predict very high-risk premiums compared to the observed one.

Conclusion

This thesis shows that it is possible to predict the risk premium by modeling the claims frequency and the average cost over a given period. This modeling of the material claims frequency and average cost for MTPL cover was done using two approaches. We started with a model containing only contract variables, then to this model, we added the “COVID” variable to identify the overall effect of the pandemic on claims and severity. Subsequently, with a view to raising the effectiveness of the integration of telematic variables in the modeling, we focused on the study of models taking these variables into account.

On the modeling of the material frequency, the contract and telematics models both adjust the loss ratio well (reconciliation of the observed and estimated loss ratios). Nevertheless, it would be beneficial to deepen the telematics model since it contains crucial information, the only one

(in the case of our fleets), on the driver, which is the overall driving score. This variable perfectly reflects the driver's behavior. As for the modeling of the material average cost, we have seen that the telematics model overestimates the claims charge and that therefore it can be considered as a conservative model.

It is thus possible to imagine various ways of integrating telematics variables into pricing to build a pricing policy adapted to the desired offer. As we have seen, we can consider an increase in the rate for drivers with low scores and a decrease for high scores. Nevertheless, there is no question of penalizing bad drivers but of applying price reductions only to drivers who have obtained good scores.

In the last chapter, we focused on the correlation of the tariff (premium) with macro-economic effects. Given the scars left by COVID-19 and the fallout from war in Ukraine, long-term growth would be hard to envision, even for connected insurance. It would therefore be interesting to consider measures considering all the factors that could have an impact on the tariff. The adoption of a new convention would be a very good illustration of this.

Remerciements

Il y a deux ans, j'ai entrepris un voyage pour obtenir un diplôme de master d'actuariat. Ce mémoire marque la fin dudit voyage. La route pour y arriver fut généralement brillante et pleine de réalisations, parfois sombre et pleine d'obstacles, mais toujours très passionnante et valait vraiment le coup. Et comme c'est généralement le cas, aucune de ces réalisations n'a été gagnée seule et aucun de ces obstacles n'a été surmonté seule. Je veux donc prendre un peu de temps pour remercier tous ceux qui m'ont rejoint durant ce voyage et m'ont ainsi aidée à franchir la ligne d'arrivée.

Avant de penser à atteindre la ligne d'arrivée, on doit d'abord se rendre au point de départ.

Je voudrais dans un premier temps remercier, toute l'équipe Actuariat Dommages International (ADI), pour m'avoir assistée et partagée des connaissances considérables. Je tiens à leur témoigner toute ma reconnaissance, pour l'expérience enrichissante et pleine d'intérêt qu'elle m'a fait vivre au sein de la Direction Technique et Protection (TPR).

Merci à Thibault VAN EVERBROECK, responsable de l'équipe DOM, qui par bienveillance a su m'assister du début jusqu'à la fin, se souciant à tout instant de mon bien-être. Ce qui n'a cessé d'accroître ma motivation. J'ai aussi beaucoup appris de lui étant donné son art de l'organisation et ses qualités de manager.

Mes remerciements s'adressent vivement à Paul SCHMITT-BAILER, qui à travers son rôle de tuteur, m'a beaucoup apprise sur les défis à relever dans chaque type de mission. Il n'hésite jamais à me partager ses connaissances, son savoir-faire et son expertise dans ce domaine, tout en m'accordant sa confiance et une large indépendance dans l'exécution de tâches valorisantes.

Merci aussi à tous mes collègues : Nicoleta NOKO, Tadeusz DESLANDES, Manon PIEREN, Fabrice GADAUD, Valentina MAURO et Giacomo ALIBARDI qui ont été toujours présents en cas de besoin et m'ont énormément appris lors de cette expérience.

Je souhaite également remercier Olivier LOPEZ, directeur de l'ISUP, pour m'avoir permis d'intégrer l'école, pour son accompagnement et aussi ses conseils.

Enfin, un grand merci à TOUTE ma famille pour leur soutien inconditionnel. J'ai beaucoup de chance d'avoir grandi dans une famille chaleureuse qui apprécie la compagnie de l'autre. Les plus grands éloges iront très certainement à mes personnes favorites ; mes parents. Les mots ne peuvent décrire à quel point je suis reconnaissante pour tout que vous avez fait pour moi durant toute ma vie. MERCI BEAUCOUP !

Table des matières

INTRODUCTION

Dans un environnement en constante mutation, les nouvelles mobilités prennent de plus en plus de place dans les déplacements quotidiens de la population. Les différentes périodes de confinement qu'a vécues le monde en 2020 et 2021 ont également eu un impact important sur la mobilité des usagers, qui ont été nombreux à ne plus se servir de leur voiture quotidiennement.

Les mois qui ont suivi la première vague de la COVID-19 ont découvert des défis que peu auraient pu prévoir et moins encore auraient pu se préparer. Les impacts de la pandémie se sont répercutés sur de nombreuses branches de l'assurance, mais l'une des conséquences les plus directes à l'échelle mondiale a été visible dans l'assurance automobile qui a enregistré un recul de sinistralité d'environ -20% en 2020.

Aussi, la sinistralité des portefeuilles de flottes automobiles s'est vue considérablement réduite. Une flotte automobile correspond à l'ensemble des véhicules que possède une entreprise. Le marché français de l'assurance flotte automobile est chiffré à 2,4 Md€ et concerne 4,6 M de véhicules en 2020.

Aujourd'hui, malgré le retour des conducteurs sur les routes, les répercussions de la pandémie sur leurs habitudes de conduite, sur l'utilisation des véhicules ainsi que sur la couverture d'assurance se poursuivront probablement à l'avenir. Depuis l'avènement de la télématique mobile, il existe maintenant de multiples variations dans les modèles d'assurance à l'usage. La télématique est un terme générique qui désigne, pour les véhicules, des systèmes de communication embarqués permettant de produire, d'envoyer, de recevoir et de stocker de l'information par des moyens de télécommunication et des systèmes de navigation satellitaire. L'intérêt pour les modèles d'assurance connectés a fortement augmenté depuis le début de la pandémie en Europe :

- L'assurance au kilomètre : également connue sous le nom d'assurance « Pay as you drive » (pour « Payez comme vous conduisez »).
- L'assurance à l'usage classique : basée essentiellement sur le comportement de conduite.
- L'assurance qui récompense : également basée sur le comportement, avec la différence que le score de conduite n'est pas directement lié à la prime.

L'intérêt des conducteurs pour la télématique n'a jamais été aussi élevé et ne cesse de croître, la crise de la COVID-19 a fortement participé à la découverte des nombreux avantages de la télématique aux conducteurs.

Le but de ce mémoire est donc de capter l'impact de la pandémie croisé avec la contribution des paramètres télématiques dans les flottes automobiles. Pour ce faire, nous allons modéliser la fréquence et le coût moyen des sinistres sur la garantie responsabilité civile dans la flotte ALD Italie en se basant sur les modèles linéaires généralisées (GLM).

Ainsi, dans la première partie, nous allons expliquer plus en détail le contexte ainsi que les différentes données qui vont être utilisées pour la modélisation. Dans la seconde partie, nous

présenterons la tarification automobile et la théorie sur les GLM avant d'effectuer une analyse détaillée des données à notre disposition. Dans la troisième partie, dans le but de déterminer la prime de risque, nous étudierons l'impact de la pandémie en exposant les différents modèles mis en place pour modéliser la fréquence et le coût moyen des sinistres matériels. Pour finir, nous essayerons de concrétiser notre étude en effectuant des conjectures sur la prime pure avant d'exhiber la corrélation de cette analyse avec les effets macro-économiques.

PARTIE I

CONTEXTE GLOBAL DE L'ETUDE

Chapitre 1 : PRESENTATION DU GROUPE ET CONTEXTE DE L'ETUDE

1.1. Le Groupe Société Générale

La Société Générale, créée au XIXème siècle, est aujourd'hui l'une des principales banques françaises et l'un des tous premiers groupes européens de services financiers. Fort de ses 155 ans d'histoire et présent presque partout dans le monde, le Groupe compte 133 000 collaborateurs qui accompagnent au quotidien plus de 30 millions de clients.

1.1.1. Sogecap

SOGECAP est la compagnie d'assurance vie et de capitalisation du groupe Société Générale. Elle est la société mère des entités composant le métier Assurances (assurance de personnes et assurance de dommages) du groupe Société Générale en France et à l'International. Elle est au cœur de la stratégie de développement du Groupe, en synergie avec l'ensemble de ses métiers de banque de détail, banque privée et services financiers.

Présente en France avec Sogecap (entité qui s'occupe de la partie vie), Antarius, Sogessur et Oradea Vie, et dans 9 pays à l'étranger, SOGECAP propose une gamme complète de produits et services pour répondre aux besoins des clients particuliers, professionnels et entreprises en Assurance-vie Epargne, Epargne retraite et Particuliers Protection.

Société Générale Assurances s'appuie sur la performance de son modèle de bancassurance intégré alliant l'efficacité du digital et l'expertise du conseiller pour poursuivre sa dynamique de croissance et de gains de parts de marché.

Une banque assurance intégrée

L'intégration de l'assurance au sein de la banque est une source de satisfaction et de fidélisation des clients du groupe Société Générale, et de revenus et de synergies pour les réseaux. Le modèle de Société Générale Assurances s'appuie sur une très grande fréquence de contacts avec les clients, au travers des réseaux d'agence des Banques de détail et des Banques privées, des conseillers des centres de Relation Client, via les sites internet et les applications bancaires.

Société Générale Assurances met ses expertises au service des réseaux de distribution du Groupe, pour faire évoluer en continu ses gammes de produits et de services et être aux côtés des clients dans les moments qui comptent.

Partenariats

Pour développer son activité partenariale en assurance vie épargne, Société Générale Assurances peut s'appuyer sur deux franchises phares : Oradea Vie en France, et Sogelife au Luxembourg, qui répondent aux attentes d'une clientèle exigeante avec des offres innovantes distribuées via des plates-formes de conseillers en gestion de patrimoine indépendants (CGPI) de premier plan et de banques privées.

En protection, Société Générale Assurances dispose déjà de partenariats solides et poursuit sa dynamique de développement tant en prévoyance qu'en assurances dommages avec des partenaires internes et externes au groupe Société Générale, notamment en assurance des flottes automobiles.

En 2021, SOGECAP présente les chiffres clés suivants :

Chiffre d'affaires	15,8 Md€
Encours	136 Md€
Résultat	363 M€
Clients	14 M
Contrats gérés	23 M
Présence	9 Pays
Collaborateurs	2 900

Figure 6 – Chiffres clés sur l'année 2021 de Sogecap

1.1.2. Sogessur

Sogessur est la compagnie d'assurance dommages du groupe Société Générale Assurances. Elle constitue avec Sogecap, la ligne Métier Assurances du groupe Société Générale. Créée en 1996, elle a la charge de la conception et de la gestion des produits d'assurance dommages (automobile, habitation, garantie des accidents de la vie, assurance des moyens de paiement, protection juridique, assurance scolaire) distribués par les réseaux bancaires de la Société Générale. En forte croissance, Sogessur accompagne également depuis 2011 le développement de l'assurance dommages à l'international (Italie, Allemagne, Pologne, Roumanie, Russie, République Tchèque…).

1996	Création de Sogessur
1997	Commercialisation des contrats d'assurance auto et habitation
2002	Commercialisation de la garantie accidents de la vie (GAV)
2004	Commercialisation de l'assurance scolaire
2005	Commercialisation de la protection juridique et de l'assurance habitation étudiant
2008	Souscription du millionième contrat
2011	Gestion des contrats d'assurances des moyens de paiement

Figure 7 – Evolution de l'entité Sogessur depuis sa création jusqu'en 2011

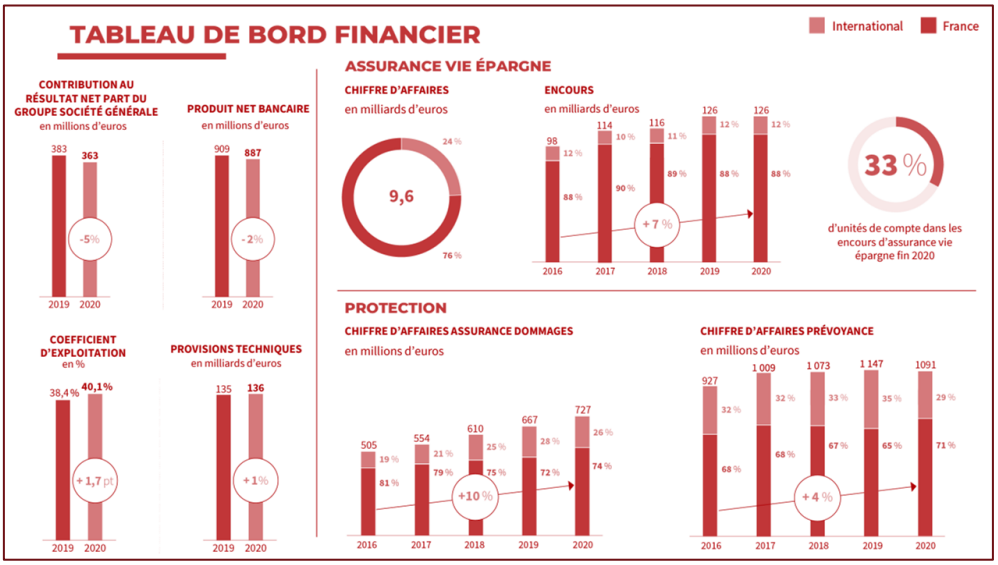


Figure 8 – Tableau de bord financier de Sogecap 2019-2020

1.1.3. TDP/TPR

La Direction Technique Dommages et Prévoyance (TDP), devenue depuis janvier 2022 la Direction Technique Protection (TPR) est au sein de la Direction Développement des Partenariats et Entreprises (DPE) et est constituée de 6 pôles :

- Le pôle Technique Prévoyance et Santé Collective (PSC),
- Le pôle Technique Dommage (DOM),
- Le pôle Produits et Contrats (PCO),
- Le pôle Réassurance Modélisation de la rentabilité (RMO),
- Le pôle Support Technique et Infocentre (STI),
- Le pôle Pilotage des Projets Transverses (PJT)

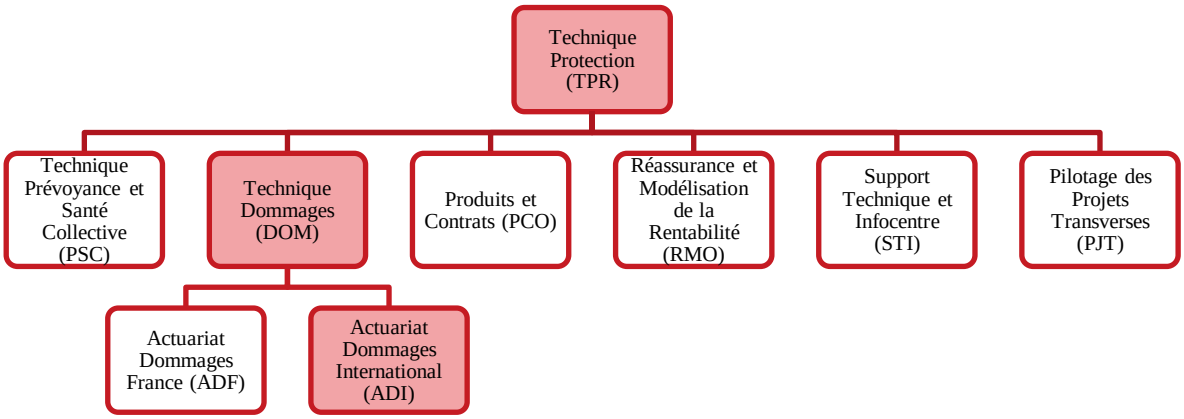


Figure 9 – Composition de la direction Technique et Protection (TPR)

1.1.4. ADI

Le pôle DOM comprend deux équipes : l'équipe ADF (Actuariat Dommage France) et l'équipe ADI (Actuariat Dommage International). Le secteur ADI couvre 5 pays chez Sogessur. Ce dernier compte 4 succursales non-vies (L'Italie, la Pologne, l'Allemagne et la Roumanie) et une filiale non-vie (la République Tchèque).

La Russie constituait aussi une filiale non-vie. Néanmoins, depuis le début des événements en Ukraine, le Groupe a rapidement conclu qu'il faisait face à un bouleversement géopolitique changeant radicalement et durablement les perspectives de collaboration avec la Russie.

Ainsi pour gérer au mieux les premières conséquences de la crise, Société Générale a décidé de cesser ses activités de banque et d'assurance en Russie et de signer un accord en vue de céder la totalité de sa participation dans Rosbank (une banque universelle russe) ainsi que ses filiales russes d'assurance à Interros Capital (l'une des plus grandes sociétés d'investissement en Russie et le précédent actionnaire de Rosbank).

Cet accord, qui a été conclu au terme de plusieurs semaines de travail intensif, permettait au Groupe de se retirer de manière effective et ordonnée de la Russie.

Concernant ses projets d'implantation, Société Générale Assurance a lancé une compagnie Takaful au Maroc, filiale détenue à 100% par La Marocaine Vie. La filiale participative propose, comme offre de départ, une couverture takaful destinée aux personnes bénéficiaires de financement Mourabaha (transaction entre le vendeur ~le client~ et l'acheteur ~la banque islamique~), suivie de la commercialisation de produits multirisque particulier et entreprise.

La société prévoit d'élargir sa gamme de produits notamment en assurance dommages. Le pôle DOM est sollicité à cet égard sur le développement d'un produit MRH (Multirisque Habitation) pour 2023.

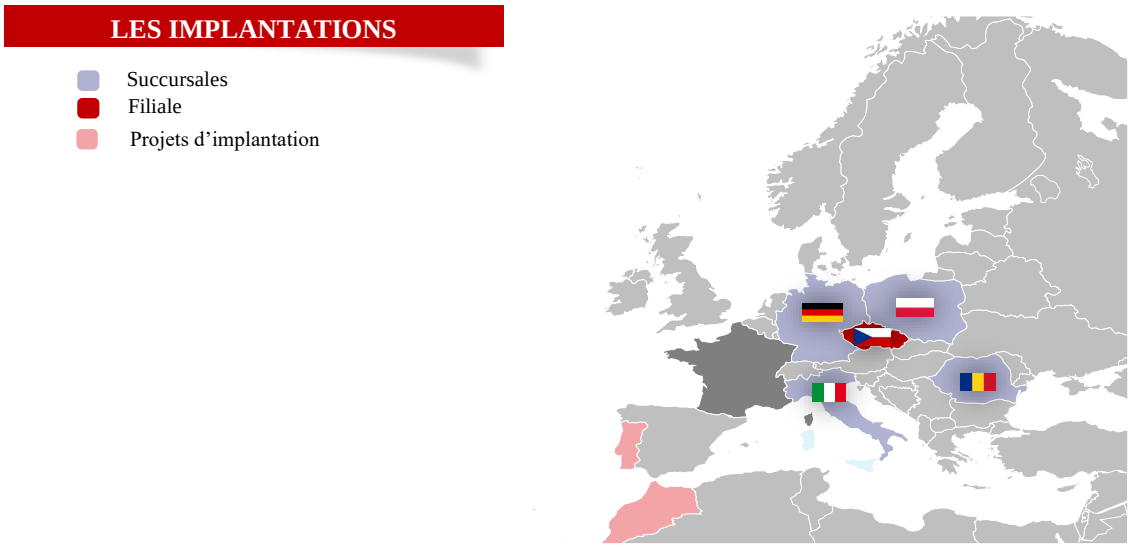


Figure 10 – Les différents pays couverts par l'équipe ADI

Le rôle de ADI pour les filiales, est d'effectuer la supervision technique ainsi qu'une revue des produits commercialisés et construits sur place, par les équipes d'actuariat locales. Cependant

pour les succursales, ADI est responsable de toutes les études actuarielles et de toute la partie provisionnement et tarification.

Le travail principal d'ADI est donc de réaliser des études actuarielles d'analyses de performances et de clôtures, mais aussi d'effectuer la réponse aux appels d'offres et la construction de produits avec les partenaires externes ou intra-groupe à l'étranger.

Pour répondre à ces multiples sollicitations venant des succursales et pouvoir superviser les filiales, l'équipe ADI comprend 6 actuaires à temps plein. Les missions chez ADI sont multiples : Tout d'abord, chaque mois, l'équipe réalise la clôture des comptes pour les produits présents dans toutes les succursales avec le calcul des provisions et IBNR ainsi qu'une projection des résultats attendus. Ces clôtures sont réalisées mensuellement pour mieux piloter les produits et avoir une meilleure vue d'ensemble chaque mois de l'évolution des business et des écarts par rapport aux budgets estimés.

Ensuite, l'équipe est énormément sollicitée sur les appels d'offres où il faut construire le tarif et donc mener des études statistiques et prendre l'information sur le marché dans le pays concerné pour pouvoir construire des produits avec des prix attractifs mais rentables. L'équipe est aussi en charge des études tarifaires et de rentabilité pour les renouvellements des produits existants.

Pour réaliser ces différentes missions au quotidien, l'Actuariat Dommage International a de nombreux interlocuteurs avec qui elle collabore et travaille. Les premières personnes en lien avec l'équipe sont les équipes des succursales qui la sollicitent pour les appels d'offres, fournissent les données pour les clôtures et attendent les retours de provisions estimées ainsi que les résultats consolidés.
Le deuxième interlocuteur de l'équipe est le service Produits et Contrats (PCO) situé lui aussi dans la Direction Technique de Sogessur. La collaboration entre les deux services permet à Sogessur de répondre aux appels d'offres avec des produits bien structurés qui répondent aux attentes du client.
Un autre interlocuteur sans qui l'équipe actuariat dommage international ne pourrait travailler est le service de Support Technique et Infocentre (STI) : c'est l'équipe qui met en forme et à disposition la majorité des bases de données nécessaires à l'actuariat pour pouvoir faire le suivi des produits.
Enfin, le dernier interlocuteur majeur est l'équipe Réassurance et Modélisation de la Rentabilité (RMO). Cette équipe réalise les budgets et les business plans de chaque produit lancé par ADI et s'occupe aussi des traités de réassurance.

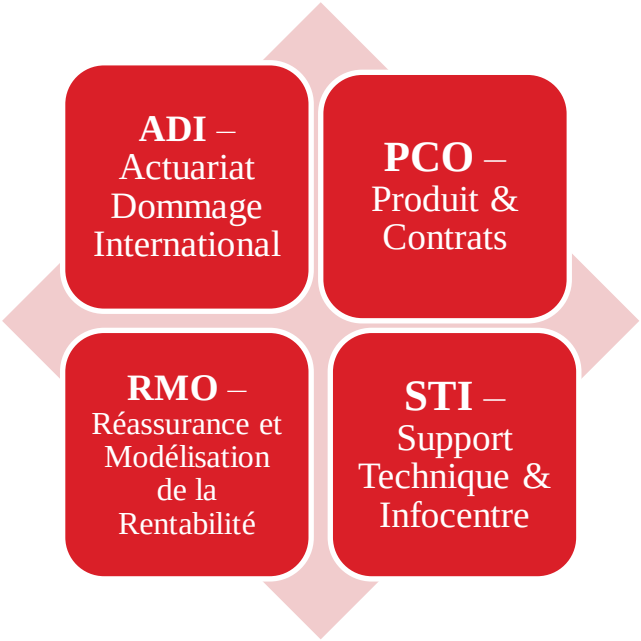


Figure 11 – Les différents interlocuteurs d'ADI

1.1.5. ALD automotive (LLD)

ALD Automotive est une filiale du groupe Société générale créée en 1998 et spécialisée dans la location longue durée (LLD) et la gestion de parc automobile :

- Location longue durée : l'entreprise cliente confie au loueur le financement et la gestion de son parc automobile en contrepartie d'un loyer mensuel ;
- Gestion de flotte (fleet management) : l'entreprise cliente reste propriétaire de son parc automobile sans en supporter la gestion quotidienne, prise en charge par le gestionnaire.

En 2014, le Groupe Société Générale Assurances s'est lancé dans l'expérience de la télématique avec les voitures connectées. Depuis cette date, ALD et SGA travaillent en étroite collaboration sur les données télématiques issues des véhicules équipés dans la flotte ALD italienne.

1.2. Projet télématique : évolution et enjeux

La Télématique représente l'ensemble des techniques qui combinent les moyens de l'informatique avec ceux des télécommunications. Elle vise à récolter des informations (des données) sur les véhicules dans le but d'apporter de nouveaux services et une valeur ajoutée pour le client et le conducteur.

La notion « télématique » remonte aux années 1960 mais la mise en place de cette technologie auprès du grand public a mis du temps à voir le jour. L'une des principales fonctionnalités de la télématique est la technologie du GPS. Grâce à cette dernière et aux différentes mesures prises par l'Europe pour développer la télématique et améliorer la sécurité routière, la télématique est devenue un outil du quotidien bien que peu connu.

Le premier point d'entrée de la télématique dans l'automobile s'est fait par l'intermédiaire des flottes automobiles. Il était nécessaire pour les entreprises d'avoir une vue globale de leurs flottes et une estimation précise des kilomètres parcourus, de l'essence consommée et du coût

total de la flotte à un moment donné. C'est donc pour répondre à ces besoins que les différents acteurs de la télécommunication et de l'informatique ont développés des technologies capables de suivre en temps réel ces différentes informations d'un parc automobile. La vitesse à laquelle la télématique a évolué depuis sa création dans les années 1960 est tout simplement impressionnante :

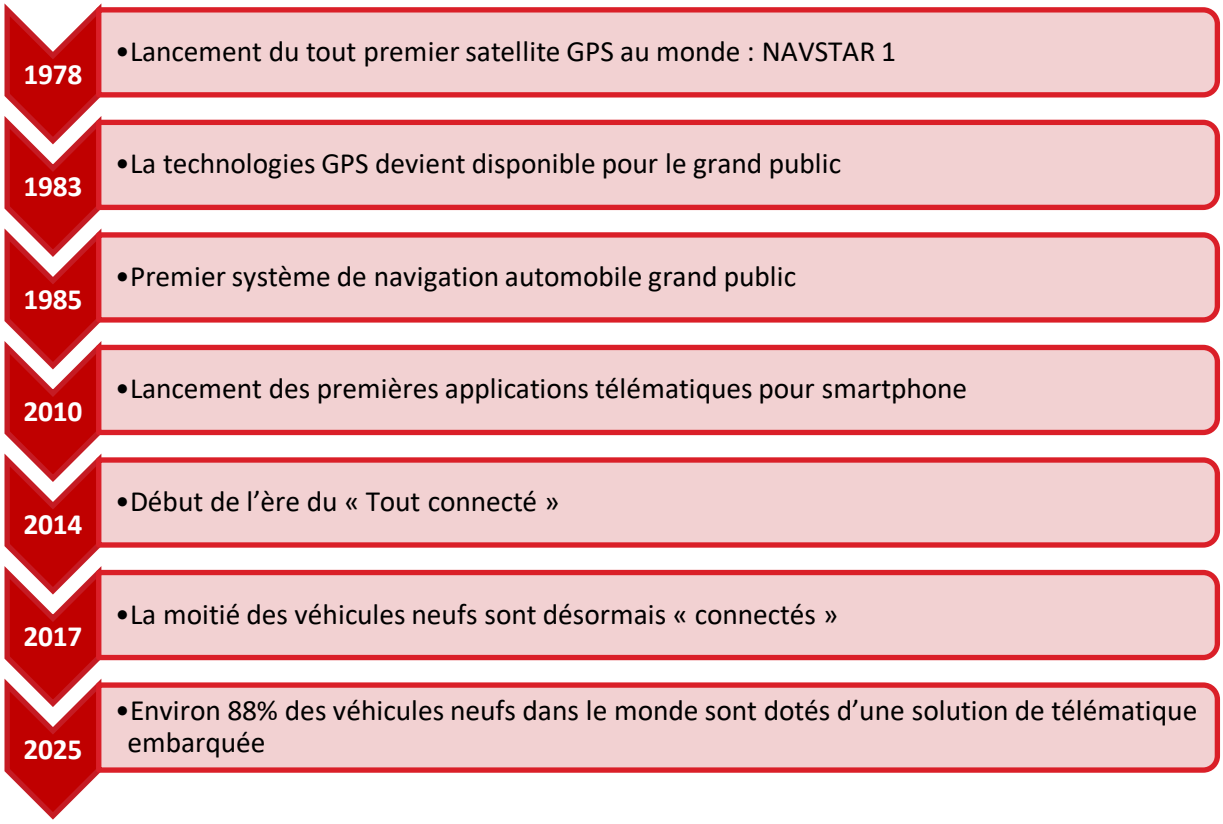


Figure 12 – Evolution de la télématique depuis 1978

1.2.1. Télématique et politique

Les développements technologiques récents décrits ci-dessus ont (et continueront d'avoir) des impacts sur les structures et les normes politiques. Les politiques démocratiques de masse peuvent être affectées par des changements dans l'environnement de l'information domestique dans trois domaines critiques : le contrôle des sources d'information fournies par les nouvelles technologies ; le contenu, la qualité et l'étendue des informations fournies.

Dans le secteur automobile, la télématique vise à récolter des informations sur les véhicules dans le but d'apporter de nouveaux services et une valeur ajoutée pour le client et conducteur. Les gouvernements et Etats ont mis en place plusieurs politiques publiques pour développer les technologies de prévention et d'aide à la conduite dans le but de réduire le nombre d'accidents de la route. L'assurance est l'un des secteurs où la télématique peut permettre une évolution majeure dans l'appréciation du risque et la prévention des usagers de la route. Le nerf de la guerre dans cette nouvelle technologie et offre d'assurance est la donnée de l'assuré. Le problème étant que cette ressource est de plus en plus surveillée ; les entreprises ne peuvent plus stocker tous les types de données. De plus, les assurés sont de plus en plus sensibles aux données que les entreprises ou personnes tierces possèdent à leur égard. Ainsi, ces nouveaux

produits sont dépendants d'une ressource qui n'est pas en libre-service mais qui en plus devient de moins en moins accessible.

D'ailleurs, un texte de loi publié au Journal officiel de l'Union européenne en 2019 impose aux constructeurs automobiles, à partir de juillet 2022, d'équiper les nouveaux véhicules d'un boîtier qui enregistre la vitesse, le freinage et plusieurs paramètres de sécurité en cas de choc. Les voitures d'occasion devront y passer en 2024. Ces dispositifs embarqués seront en mesure de collecter de nombreuses données de navigation comme la consommation, les freinages, les trajectoires ou les accélérations. Reliés par GPS, ils seront capables de ralentir automatiquement le véhicule en cas de dépassement de la vitesse autorisée.

Selon le texte, « les données qu'ils sont capables d'enregistrer peuvent être communiquées aux autorités nationales, sur la base du droit de l'Union ou d'un droit national, pour les seuls besoins de l'étude et de l'analyse des accidents ». En revanche, « les enregistreurs de données d'événement devraient fonctionner suivant un système en circuit fermé » ce qui pourrait se traduire par : les données ne pourront pas être transmises, en direct, au constructeur ou aux assurances.

Le texte est aussi clair sur l'identification du propriétaire, « un enregistreur de données d'événement n'est pas capable d'enregistrer et de mémoriser les quatre derniers chiffres de la partie "désignation du véhicule" du numéro d'identification du véhicule, ni aucune autre information qui pourrait permettre l'identification individuelle du véhicule concerné, de son propriétaire ou de son détenteur ».

Pour finir, en ce qui concerne la désactivation de la boîte noire dans le véhicule, au même titre que l'airbag qui assure la sécurité du véhicule (à l'exception de celui du passager avant pour permettre l'installation d'un siège bébé si besoin), ce dispositif ne pourra pas être supprimé à l'intérieur de la voiture. L'alinéa b de l'article 6 précise que les dispositifs d'enregistrement « ne peuvent pas être désactivés ».

1.2.2. Télématique et RGPD

Le Règlement Général sur la Protection des Données (RGPD) a été adopté le 14 avril 2016 au Parlement Européen et promulgué le 27 avril 2016 au journal officiel de l'Union Européenne. Il est entré en vigueur le 25 mai 2018. Il concerne la protection des personnes physiques à l'égard du traitement de leurs données personnelles, ainsi que la libre circulation de celles-ci. Le RGPD est l'extension de la loi européenne de protection des données personnelles à toutes les entreprises étrangères traitant de la donnée concernant les citoyens européens. Ce règlement s'applique à tous les états membres de l'Union Européenne, fournissant ainsi un cadre juridique unifié pour l'ensemble de l'Union Européenne.

Une « donnée personnelle » est « toute information se rapportant à une personne physique identifiée ou identifiable ». Une personne peut être identifiée :

- directement (nom, prénom)
- ou indirectement (par un identifiant (n° client), un numéro (de téléphone), une donnée biométrique, plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale, mais aussi la voix ou l'image).

Le règlement renforce les droits des utilisateurs sur l'utilisation de la donnée personnelle par les entreprises. Il définit notamment la notion de « consentement » : les utilisateurs doivent être informés de l'usage qui sera effectué sur leurs données et ils devront donner leur accord ou s'y opposer. De plus, les données devront être « portables » : une personne peut récupérer les données sous une forme facilement réutilisable et peut les transférer à un tiers. Ceci permet de redonner de la maîtrise des données aux utilisateurs. Enfin, la dernière disposition clé est une meilleure application du « droit à l'oubli » pour les citoyens : c'est le fait de pouvoir exiger la suppression de tout ou une partie de ses données personnelles.

Dans le domaine des flottes et de la télématique, le RGPD n'a pas vraiment bouleversé les pratiques. Avant, la Commission nationale de l'informatique et des libertés (CNIL) réglementait déjà les usages de la télématique dans les flottes : elle laissait aux conducteurs la possibilité de se mettre en mode privé ou elle n'ouvrait pas aux administrateurs la possibilité de visualiser les données anciennes de plus de deux mois. » Mais le RGPD a remis l'accent sur des règles légales, à l'image de l'interdiction de contrôler les vitesses des conducteurs par l'entreprise.

Le « profilage » est certainement le risque le plus élevé pour le traitement des données télématiques. Il désigne notamment le fait « d'utiliser des données à caractère personnel pour évaluer certains aspects personnels relatifs à une personne physique, notamment pour analyser ou prédire des éléments concernant le rendement au travail, la situation économique, la santé, les intérêts, la fiabilité, le comportement, la localisation ou les déplacements de cette personne physique ». Dans le cadre de la télématique, récupérer la séquence des positions GPS du véhicule, tout en ayant accès à des informations comme la date et l'heure de chacune de ces positions GPS, est un cas typique de « profilage » puisqu'il est alors possible d'obtenir la localisation et le déplacement du véhicule.

Tout ceci conduit à un besoin de normalisation pour assurer la cyber sécurité des données mais aussi une interopérabilité entre les systèmes pour que tous les acteurs puissent se comprendre et parler le même langage.

1.3. Le marché de l'assurance flotte automobile en France

Selon le rapport de la FFA (Fédération Française de l'Assurance) au 31/12/2020, relativement au marché de l'assurance flotte auto en France, le parc des véhicules assurés par un contrat flotte est estimé à un peu plus de 4,6 millions de véhicules, en hausse de 5,1 %. Il représente 8,3 % de l'ensemble du parc des véhicules assurés.

	Parc (En milliers)	Variations 2020/2019	Cotisations (En M€)	Variations 2020/2019
Véhicules assurés en mono contrats	51 130	1,30%	21 070	2,80%
Véhicules assurés en contrats flottes	4 610	5,10%	2 412	4,80%

Tableau 10 – Chiffres du parc flotte automobile en France en 2020 (Source : FFA)

Après une année de hausse très soutenue (+7,5 % en 2019), l'année 2020 qui a connu une crise inédite en raison de la pandémie de la COVID-19, se caractérise par une baisse des immatriculations des flottes automobiles neuves de -23,1 %.

Avec un chiffre d'affaires de 2,4 milliards d'euros, les contrats flottes d'entreprises à usage professionnel représentent 10,3 % de l'ensemble des cotisations d'assurance automobile et 4,0 % de l'ensemble des assurances de biens et responsabilité.

La législation française rend obligatoire la souscription de l'assurance de responsabilité civile par le propriétaire d'un véhicule automobile s'il souhaite le mettre en circulation. La garantie Responsabilité civile (RC) représente le minimum obligatoire auquel se doit de souscrire tout automobiliste pour rester dans la légalité. C'est également la moins chère, mais son niveau de couverture est le plus bas, puisqu'elle couvre uniquement les dégâts matériels et corporels que l'assuré est susceptible de causer à autrui avec son véhicule. Cela implique que l'assureur prend à sa charge l'indemnisation des dommages matériels ou corporels de la victime. Il ne permet pas de rembourser le conducteur de ses propres dommages mais il permettra de rembourser ses passagers en cas de sinistre.

Parmi les garanties facultatives que le conducteur peut souscrire, il est possible de citer la garantie dommages au véhicule. Cette garantie complémentaire permet au propriétaire du véhicule d'obtenir une indemnisation des dommages subis par le véhicule, en cas de sinistre où le conducteur est responsable. Le montant de l'indemnisation est en général évalué par un expert indépendant.

En 2020, excepté pour la garantie Vol, l'évolution des fréquences est favorable. Cette évolution est la conséquence directe de la baisse du trafic routier. La baisse la plus marquée s'observe sur la garantie Dommages au véhicule (- 25,5 % par rapport à 2019). En revanche, les coûts moyens sont toujours orientés à la hausse.

	Fréquence 2020		**Coût moyen 2020**	
	Niveau (‰)	**Variation vs 2019**	**Niveau (€)**	**Variation vs 2019**
RC corporels	4,1 ‰	-23,60%	*N.D.*	*N.D.*
RC matériels	58,7 ‰	-25,00%	1 240 €	2,50%
Vol	4,2 ‰	2,70%	5 550 €	12,90%
Bris de glaces	58,8 ‰	-16,70%	600 €	8,90%
Dommages tous accidents	49,7 ‰	-25,50%	2 370 €	7,40%

Tableau 11 – Fréquence et coût moyen par garantie de l'année 2020 comparée à 2019

N.D. : Non disponible

Le ratio sinistre à primes de l'ensemble des contrats flottes s'améliore et représente 74 % des cotisations, en recul de 12 points de pourcentage par rapport à 2019.

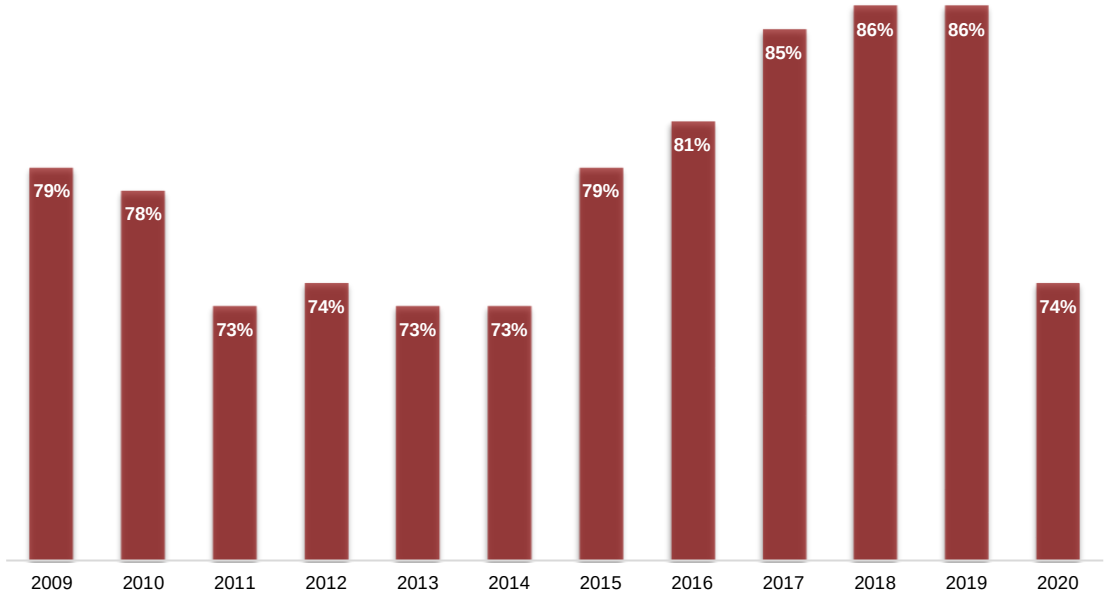


Figure 13 – Evolution du ratio sinistres à primes de 2009 à 2020 sur l'ensemble des contrats flottes

Chapitre 2 : TARIFICATION AUTOMOBILE

2.1. Définition

Lorsqu'un client souhaite souscrire à un nouveau contrat d'assurance, l'assureur doit déterminer la prime que l'assuré devra payer en fonction de certaines caractéristiques. Cette prime reflète le coût d'un assuré au cours d'une certaine période ; elle doit être à la fois suffisamment élevée pour que la compagnie d'assurance puisse prendre en charge un éventuel sinistre subi par l'assuré, dont le coût peut être élevé, et à la fois suffisamment basse pour que l'assuré accepte de signer le contrat et qu'il ne parte pas dans une autre compagnie d'assurance.

Tarifer un risque automobile, c'est donc apprécier la probabilité que l'assuré et le véhicule à qui sont accordés une garantie soient impliqués dans un accident de la circulation. Pour déterminer cette probabilité, l'assureur se base sur des études statistiques menées sur la fréquence et le coût des accidents.

C'est aux actuaires de déterminer les différents paramètres à prendre en considération pour tarifer un risque donné, en partant de l'analyse globale de l'ensemble des données statistiques relatives à la circulation automobile. Ce travail fondamental permet à chaque assureur de déterminer un tarif indispensable à l'équilibre de ses résultats dans la branche « automobile ».

2.2. Les différentes structures de tarification

Les systèmes de tarification combinent l'affectation d'un assuré à une grille de classification a priori, selon des caractéristiques objectives observables sans coût, avec un mécanisme de modulation de la prime selon l'expérience au volant. L'importance relative des deux éléments de tarification est variable selon les pays.

En Europe, la tarification selon l'expérience est très développée suivant des mécanismes souvent communs à l'ensemble des assureurs, notamment du fait de la réglementation, et se superpose à une classification a priori qui peut aller d'une structure très sommaire à une structure intégrant de nombreux critères comme en France. Avec l'évolution vers une plus grande liberté tarifaire, on constate une tendance à la diversification des structures de tarification a priori ; il existe une grande diversité des produits d'assurance, qui associent souvent assurance responsabilité civile et assurance dommages en un même contrat, les tarifs pouvant varier selon les assureurs à la fois par une classification a priori, une tarification en fonction de l'expérience.

2.3. Les types de garanties en assurance automobile

En assurance automobile, on distingue les garanties suivantes :

- La garantie Responsabilité civile : c'est le minimum obligatoire auquel se doit de souscrire tout automobiliste pour rester dans la légalité. C'est également la moins chère, mais son niveau de couverture est le plus bas, puisqu'elle couvre uniquement les dégâts matériels et corporels que l'assuré est susceptible de causer à autrui avec son véhicule.

- La garantie dommages au véhicule : Les garanties dommages sont des garanties facultatives. Elles permettent, en cas d'accident responsable, d'obtenir une indemnisation des dommages

subis par son propre véhicule. Le montant de cette indemnisation est déterminé par l'examen du véhicule par un expert indépendant.

Les garanties dommages sont donc utiles, notamment lorsque le véhicule assuré est neuf ou a peu roulé. On distingue deux niveaux de garanties dommages :

*La garantie dommage collision permet à un assuré de recevoir une indemnisation en cas d'accident responsable avec un tiers identifiable. Il peut s'agir d'un autre véhicule (motorisé ou non), d'un piéton ou d'un animal dont il est possible d'identifier le propriétaire. L'indemnisation reçue par l'assuré est proportionnelle aux dommages subis par son véhicule.

*La garantie dommages accidents permet à un assuré de recevoir une indemnisation en cas d'accident responsable, même en l'absence de tiers. Elle est donc beaucoup plus large que la garantie dommages collisions et offre, par exemple, une couverture si le véhicule subit des dommages alors qu'il se trouve en stationnement.

Il faut noter que la garantie dommages est généralement assortie d'une franchise. Il faut donc veiller à ce que le niveau de cette franchise corresponde bien au niveau d'usage du véhicule assuré. Il existe 3 types de franchise :

-La franchise simple : si cette franchise est fixée à x€, l'assuré est indemnisé quand le coût de son sinistre dépasse x€,

-La franchise intégrale : cette franchise est considérée comme étant un seuil. Si elle est fixée à x€, seuls les sinistres ayant un coût supérieur à x€ sont remboursés ;

-La franchise agrégée : si cette franchise est fixée à x€, tant que le coût total de tous les sinistres ne dépasse pas x€, aucun sinistre n'est remboursé.

*La garantie corporelle du conducteur : le plus souvent présente dans les formules d'assurance tous risques, la garantie corporelle du conducteur est la seule protection pouvant prendre en charge les dommages corporels subis par le conducteur assuré, indépendamment de sa responsabilité dans l'accident. Selon les conditions prévues par le contrat, la garantie du conducteur couvre ce dernier à bord de la voiture. La garantie corporelle du conducteur prend en charge les frais médicaux, le préjudice subi et éventuellement la perte des revenus liée à un arrêt de travail. Il faut noter que les coûts liés aux sinistres corporels qui, bien qu'ils soient moins fréquents, peuvent engendrer des coûts très importants.

*La garantie Vol et Incendie : cette garantie indemnise le propriétaire du véhicule lorsque celui-ci est endommagé ou détruit par le feu, ou encore lorsqu'il est volé (ou même lorsqu'il y a tentative de vol, en effet le véhicule peut subir des dégâts à la suite d'une tentative de vol). L'indemnité varie en fonction des conditions prévues par l'assurance, elle peut se baser sur la valeur de la voiture au moment du sinistre ou sur une valeur conventionnelle indiquée dans le contrat.

En plus de l'historique, les caractéristiques du contrat influent sur le montant de la prime (montant des franchises, des conservations, des plafonds). Enfin, le montant des différents frais liés au contrat est pris en compte (frais de gestion des sinistres, d'administration des contrats et d'acquisition principalement).

2.4. Eléments du tarif d'un produit d'assurance automobile

2.4.1. La prime pure

La prime d'assurance payée par le souscripteur du contrat se base en premier lieu sur une « prime pure » calculée par l'assureur. La prime pure (appelée aussi « prime d'équilibre », « prime de risque » ou « prime technique ») correspond au montant estimé de la charge des sinistres. Elle est déterminée par référence à l'historique statistique sur les exercices précédents de la fréquence des sinistres et de leur coût moyen.

Le calcul de la prime pure nécessite la détermination de deux ratios :

- Fréquence

$$Fréquence = \frac{Nombre\ de\ sinistres}{l'exposition}$$

Le numérateur représente le nombre de sinistres garanties. L'exposition d'un assuré au sein d'un portefeuille correspond à la durée durant laquelle celui-ci a été observé dans le portefeuille. En général, cette exposition s'exprime en années-police (1 année police correspond donc à l'observation d'un assuré dans le portefeuille pendant une année entière).

En fonction des dates de début et de fin d'étude, et des dates de début et de fin de contrat, l'exposition se calcule selon la formule suivante :

$$Expo = \max(0, \frac{\min(dt_{etude}^{fin}, dt_{contrat}^{fin}) - \max(dt_{etude}^{deb}, dt_{etude}^{deb})}{366}$$

On divise par *366* car la période d'étude est une année bissextile. Si l'étude était plus générale, on prendrait plutôt *365,25*.

-Sévérité (ou coût moyen)

$$Sévérité = \frac{Coûts\ Totaux}{Nombre\ de\ sinistres}$$

La sévérité peut être calculée en utilisant le rapport des coûts des sinistres payés sur le nombre de sinistres fermées ou le rapport des coûts des sinistres déclarés sur le nombre de sinistres déclarés. Les dépenses reliées aux sinistres peuvent être inclues /exclues de ce calcul.

La prime pure représente dès lors le coût moyen payé par chaque unité d'exposition et est calculé de cette façon :

$$Prime\ Pure = \frac{Coûts\ Totaux}{l'exposition}$$

$$= \frac{Coûts\ Totaux}{l'exposition} * \frac{\text{Nombre de Sinistres}}{\text{Nombre de Sinistres}}$$

$$\boxed{Prime\ Pure = Fréquence * Sévérité}$$

Modéliser la prime pure revient donc à effectuer deux modélisations : une sur la fréquence des sinistres et une autre sur le coût moyen des sinistres. La fréquence est une mesure du taux auquel les réclamations surviennent pour le risque spécifique. La sévérité est une mesure du coût moyen des sinistres pour le risque spécifique.

2.4.2. Les autres éléments du tarif

En plus de la prime pure, d'autres éléments viennent s'ajouter pour définir le tarif final d'un contrat d'assurance :

-un chargement de sécurité, c'est à dire une majoration lui permettant de faire face à la volatilité naturelle des sinistres (il peut s'agir d'une hausse soudaine et non prévue de la sinistralité au cours de l'année)

-des frais de fonctionnement, réunissant les coûts d'acquisition (commissions à verser aux intermédiaires, marketing et publicité…) et les frais de gestion des différents services de l'entreprise.

La prime totale est la somme effectivement payée par le souscripteur. Elle ajoute à la prime nette d'éventuels frais accessoires et les taxes imposées par la législation.

-Les frais accessoires : également dénommés « compléments de prime » ou « frais de police », ils rémunèrent les frais de gestion inhérents au seul contrat. Ils sont perçus lors de son émission ou à celle d'un avenant, ainsi qu'à l'occasion de chaque échéance.

-Les taxes fiscales et contributions : à la prime nette et aux frais accessoires s'ajoutent les taxes imposées par la loi

D'autres éléments sont également pris en compte dans la détermination du tarif final. Il s'agit de :

-La réassurance : les compagnies d'assurance cèdent, dans certains cas, une partie du risque assuranciel à une autre compagnie.

-La marge : elle correspond à la rentabilité du contrat pour l'assureur. Cette partie est celle qui suscite le plus de négociations.

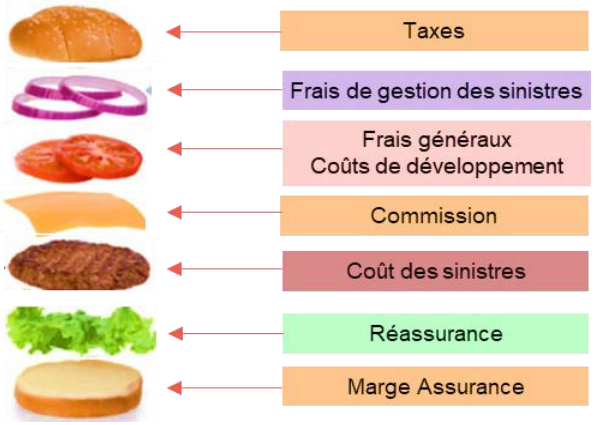


Figure 14 – Les autres éléments du tarif

2.5. Techniques de tarification : segmentation et mutualisation, les deux faces d'une même pièce

La segmentation est une technique permettant à un assureur de classer les risques selon certains critères pour établir son tarif et /ou déterminer les modalités des garanties offertes. Tandis que le principe de mutualisation des risques, qui est au cœur de l'activité d'assurance, consiste à répartir le coût de la réalisation d'un sinistre entre les membres d'un groupe soumis potentiellement au même risque. Les assureurs collectent, chaque année, les primes de tous les assurés. Ces primes vont être utilisées pour dédommager ceux qui auront subi un sinistre. Les primes d'assurance automobile de tous les conducteurs seront donc utilisées pour dédommager les accidentés.

2.6. Types de tarification : particuliers vs flottes

Durant la vie d'un contrat d'assurance automobile : une première tarification a priori effectuée au début du contrat, puis à chaque renouvellement, une tarification a posteriori qui vient modifier le tarif fixé au début du contrat.

Classiquement, pour les particuliers (contrats mono-véhicule), la tarification a priori s'effectue à partir de critères relatifs à l'assuré et son véhicule mais également le type de contrat souscrit (formules, plafonds, franchises). Ici il s'agit de comment utiliser les caractéristiques d'un assuré pour déterminer sa prime.

Pour des contrats « flotte », le montant de la cotisation dépendra davantage de la sinistralité antérieure de l'ensemble de la flotte, l'usage des véhicules, la zone de circulation ainsi que le type de véhicule. L'assureur étudie la sinistralité antérieure sur une période donnée (plusieurs années, généralement entre 3 et 5 ans), à savoir le nombre de sinistres survenus, le montant de ces derniers, ainsi que la garantie ayant été impactée, et en particulier les coûts liés aux sinistres corporels qui, bien qu'ils soient moins fréquents, peuvent engendrer des coûts très importants. En plus de l'historique, les caractéristiques du contrat influent sur le montant de la prime (montant des franchises, des conservations — franchises correspondant au montant annuel cumulé à charge de l'entreprise —, des plafonds). Enfin, le montant des différents frais liés au contrat est pris en compte (frais de gestion des sinistres, d'administration des contrats et d'acquisition principalement).

PARTIE II

THEORIE DES GLM ET PRESENTATION DES DONNEES

Chapitre 3 : THEORIE SUR LES MODELES LINEAIRES GENERALISES (GLM)

Dans cette section, nous présenterons le cadre théorique autour des GLM.

3.1. Introduction

La plupart des modèles de tarification en assurance automobile se basent sur des modèles linéaires généralisés (GLM). L'objectif des GLM est de modéliser la relation existante entre une variable réponse (ou variable à expliquer) et une ou plusieurs variables explicatives.

Avant d'exposer la théorie sur les GLM, il paraît nécessaire de comprendre le modèle linéaire gaussien. Les modèles linéaires gaussiens ont longtemps été utilisés pour modéliser la fréquence et le coût moyen. Cependant, ils ne sont pas adaptés à la réalité, puisque la variable à modéliser, c'est-à-dire la variable réponse, n'est pas nécessairement gaussienne.

Ainsi, le modèle linéaire généralisé a été créé afin d'étendre le modèle linéaire aux variables non gaussiennes et plus précisément aux variables dont la loi fait partie de la famille exponentielle.

3.2. Contexte

Le principe de la régression est de modéliser $E[Y|X]$ comme une fonction g des variables explicatives X, soit $E[Y \mid X] = g(X)$. La variable Il s'écrit alors $Y = g(X) + \varepsilon$ où ε est un bruit aléatoire. Il représente l'écart entre Y et son espérance conditionnelle, soit l'erreur que l'on commet lorsque l'on remplace Y par son espérance conditionnelle.

Supposons que nous disposons d'un échantillon de taille n de $(p + 1) - uplets\ (X, Y)$, le but est donc de retrouver la fonction g. Le modèle le plus simple, le modèle linéaire gaussien, suppose que g est linéaire et que le bruit est gaussien.

Nous pouvons observer deux problèmes :

- La forme linéaire de g peut être trop restrictive ;
- Le cadre gaussien (du bruit) n'est peut-être pas adapté aux données.

Le but des modèles linéaires généralisés est de relâcher ces deux restrictions.

Avantages des GLM :

- Ils permettent de conserver la simplicité des modèles linéaires tout en autorisant une forme plus générale ;
- Les coefficients du modèle sont estimés par maximisation d'une vraisemblance qui provient d'une famille exponentielle de lois (qui peuvent ne pas être gaussiennes).

Inconvénients des GLM

- La procédure d'estimation n'est efficace que si la vraie loi conditionnelle appartient à cette famille exponentielle ;

- Il y a une liberté sur le choix de la forme de $E\,[Y|X]$ à travers la fonction de lien. Mais, ce choix est souvent imposé par un certain choix canonique correspondant à la famille exponentielle choisie ;
- Utiliser une famille exponentielle impose de ne pas avoir de valeurs extrêmes.

3.2.1. Le modèle linéaire gaussien

Le but des modèles linéaires (LM) (tels que les modèles linéaires généralisés) est d'exprimer la relation entre une variable de réponse observée, Y (par exemple : fréquence de sinistralité), et un certain nombre de covariables (également appelées variables prédictives), X. Les deux modèles considèrent les observations Y_i, comme étant des réalisations de la variable aléatoire Y.

Les modèles linéaires conceptualisent Y comme la somme de sa moyenne μ et d'une variable aléatoire, ε : $Y = \mu + \varepsilon$

Le modèle suppose que :

3. La valeur attendue de Y , μ , peut être écrite comme une combinaison linéaire des covariables, X.
4. Le terme d'erreur, ε , est normalement distribué avec une moyenne nulle et une variance σ^2

Le modèle linéaire cherche à exprimer l'élément observé Y comme une combinaison linéaire d'une sélection de variables prédictives, plus une variable aléatoire $\varepsilon \sim N(0, \sigma^2)$:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon$$

- $(X_1, X_2, \dots X_n)$ sont les variables prédictives, également appelées co-variables.
- $(\beta_1, \beta_2, \dots \beta_n)$ sont les paramètres du modèle à estimer.
- Ce modèle suppose donc que la variable d'observation Y est normalement distribuée avec une moyenne $\beta_0 + \sum_{i=1}^{n} \beta_i X_i$ et une variance σ^2.

Nous introduisons les notations matricielles suivantes :

$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ est le vecteur colonne de $\mathbb{R}^n$ dont les composantes correspondent aux valeurs observées pour la variable réponse ;

$X_{pi} = \begin{pmatrix} X_{p1} \\ X_{p2} \\ \vdots \\ X_{pn} \end{pmatrix}$ sont les vecteurs colonnes des p co-variables avec des composantes égales aux valeurs observées ;

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ est le vecteur colonne des $p+1$ paramètres, et $\varepsilon = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}$ (l'estimation des erreurs) le vecteur des n résidus.

Pour simplifier davantage la notation, les vecteurs X_i peuvent être agrégés en une seule matrice X. Cette matrice est appelée la matrice de conception et est définie comme suit :

$$X = \begin{pmatrix} 1 & X_{11} & X_{21} & & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & & X_{pn} \end{pmatrix}$$

Le système d'équations prend alors la forme : $Y = \beta X + \varepsilon$

Le but de la modélisation linéaire est de trouver les β qui permettent de minimiser les résidus. Quand des β satisfaisants sont déterminés, il est possible de prendre de nouvelles observations, correspondant à $(1, x_{n+k}^1, \dots, x_{n+k}^p)$ avec k un entier naturel, et de lui appliquer le modèle afin de trouver un Y_{n+k} estimé. Cette modélisation est soumise à trois hypothèses :

- Les résidus sont indépendants.
- Les résidus suivent une loi Normale de moyenne nulle et de variance résiduelle.
- Les résidus sont homogènes.

Ces hypothèses sont simples mais contraignantes. Dans la modélisation de la sinistralité automobile, le nombre de sinistres est une variable de comptage qui appartient à $\mathbb{N}$et suit une distribution de Poisson. La variance du nombre de sinistres est le paramètre de la loi de Poisson. Ce paramètre augmente donc avec le nombre de sinistres, ce qui implique une augmentation de la variance. Ceci permet de démontrer que l'hypothèse d'homogénéité des résidus n'est pas envisageable dans notre cas.

De plus, le fait d'avoir des résidus qui suivent une loi Normale implique la possibilité d'avoir des prédictions négatives alors qu'il est impossible d'avoir un nombre de sinistres négatif.

Les hypothèses du modèle linéaire empêchent de modéliser le nombre de sinistres tel qu'on le souhaiterait. L'utilisation de modèles linéaires généralisés permettra donc de résoudre ces problèmes.

Les GLM consistent en une large gamme de modèles qui incluent les modèles linéaires comme cas particulier. Ici, les restrictions des modèles linéaires (hypothèses de normalité, de variance constante et d'additivité des effets) sont supprimées et la variable de réponse est supposée plutôt appartenir à une famille exponentielle.

3.2.2. Famille exponentielle

La famille exponentielle est une famille dont la densité de probabilité est une loi à 2 paramètres définie comme suit : $f_{\theta,\varphi}(y) = exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right\}$

où $a(\varphi), b(\theta)\, et\, c(y,\varphi)$ sont des fonctions spécifiques. Les conditions imposées à ces fonctions sont les suivantes :

3. a est positif et continu sur $\mathbb{R}$
4. b est une fonction définie sur $\mathbb{R}$, deux fois dérivable par la dérivée seconde une fonction positive (en particulier $b(\theta)$ est une fonction convexe)
5. c est indépendant du paramètre θ
6. θ est un paramètre lié à la moyenne.
7. φ est un paramètre d'échelle lié à la variance.

Ces trois fonctions sont liées par le simple fait que f doit être une fonction de densité de probabilité. Le paramètre θ est appelé paramètre canonique et φ est le paramètre d'échelle. Les distributions familières appartenant à la famille exponentielle sont les distributions Normale, Poisson et Gamma qui présentent un grand intérêt pour les applications d'assurance.

Le tableau ci-dessous résume les distributions utiles qui sont membres de la famille exponentielle (1) :

	θ	φ	$a(\varphi)$	$b(\theta)$	$c(y,\varphi)$
Normale(μ,σ^2)	μ	σ^2	$\frac{\varphi}{\theta}=\frac{\sigma^2}{\mu}$	$\frac{\sigma^2}{2}$	$-\frac{1}{2}\left\{\frac{y^2}{\sigma^2}+\ln(2\pi\sigma^2)\right\}$
Poisson(μ)	$\ln(\mu)$	1	$\frac{\varphi}{\theta}=\frac{1}{\ln(\mu)}$	$\exp(\theta)$	$-\ln(y)$
Gamma(μ,α)	$-\frac{1}{\mu}$	$\frac{1}{\alpha}$	$\frac{\varphi}{\theta}=-\frac{1}{\alpha\mu}$	$-\ln(-\theta)$	$\alpha\ln(\alpha y)-\ln(y)-\ln(\Gamma(\alpha))$

Tableau 12 – Distributions de quelques familles exponentielles

Le choix standard pour $\alpha(\varphi)$ est $\frac{\varphi}{\theta}$.

Un membre de la famille exponentielle a les deux propriétés suivantes :

- La distribution est complétement spécifique en termes de moyenne et de variance.
- La variance de Y_i est fonction de sa moyenne.

Cette seconde propriété est soulignée en exprimant la variance comme suit : $Var(Y_i)=\varphi V(\mu_i)$ (2) où $V(x)$, appelée fonction de variance, est une fonction spécifiée et le paramètre φ met à l'échelle la variance.

Pour chaque observation Y_i , nous supposons une distribution définie comme dans (1). Ainsi chaque observation a un paramètre canonique $\theta_i$différent mais le paramètre d'échelle φ est le même pour toutes les observations. On suppose en outre que les fonctions $a(\varphi), b(\theta)\ et\ c(y,\varphi)$,sont les mêmes pour tout i. Ainsi, chaque observation provient de la même classe au sein de la famille exponentielle, mais faire varier θ correspond à faire varier la moyenne de chaque observation.

De plus, les paramètres $\theta_i$et φ enveloppent les informations de moyenne et de variance sur Y_i. On peut montrer que pour cette famille de distributions :

$\mu_i=E(Y_i)=b'(\theta_i)$ (3)

$Var(Y_i)=b''(\theta_i).a(\varphi)$ (4)

Où b' est la dérivée de b par rapport à θ_i. Cf démonstration page 119*(Annexe A).*

(3) montre que le paramètre canonique est essentiellement équivalent à la moyenne. (4) peut être interprété comme établissant que la variance est une fonction de la moyenne multipliée par un paramètre d'échelle $a(\varphi)$. Ceci est conforme à la relation (2) que nous avons déjà vue plus haut.

3.3. Modèle linéaire généralisé – cadre théorique

Dans les modèles linéaires, la variable de réponse est supposée suivre une distribution normale, avec une variance constante pour chaque observation. Dans les GLM, ces limitations sont supprimées. Au lieu de cela, la variable réponse est supposée appartenir à une distribution de famille exponentielle. Par conséquent, la variance est autorisée à varier avec la moyenne. Nous pouvons esquisser les hypothèses suivantes pour les GLM :

- Composante aléatoire : chaque composante de Y est indépendante et appartient à l'une des distributions de famille exponentielle.
- Composante systématique : les p co-variables sont combinées pour donner le "prédicteur linéaire" η: $\eta = \boldsymbol{X}.\beta$
- Fonction de lien : La relation entre la moyenne et les composantes systématiques est spécifié via une fonction de lien, g, qui est différentiable et monotone telle que : $E(Y) = \mu = g^{-1}(\eta)$ (5)

Cette formulation équivaut à dire que : $g(\mu) = \boldsymbol{X}.\beta$

Le choix de la fonction de lien est donc très important selon le type de variable réponse que l'on veut modéliser. Le tableau suivant présente la forme de modèle typique utilisée dans tarif assurance :

Y	Fréquence	Coût moyen
Fonction de lien	$\ln(x)$	$1/x$
Résidus	$Poisson$	$Gamma$

Tableau 13 – Fonctions de lien pour les modélisations GLM de fréquence et de coût moyen

Une fois la forme du modèle (variable de réponse, covariables, fonction de lien et structure d'erreur) définies, il faut estimer les composantes de β. Pour ce faire, on maximise la fonction de vraisemblance (*likelihood function*). Par définition, cette méthode cherche à trouver les paramètres qui, lorsque appliqués à la forme supposée du modèle, produisent les données observées avec la probabilité la plus élevée.

La vraisemblance est définie comme étant le produit des probabilités d'observer chaque valeur de la variable Y. Pour les distributions continues telles que les distributions normale et gamma, la fonction de densité de probabilité est utilisée à la place de la probabilité. Il est courant de considérer le logarithme de la vraisemblance (*log-likelihood*) puisqu'il s'agit d'une somme d'observations plutôt que d'un produit, cela donne des calculs plus gérables (et tout maximum de la vraisemblance est également un maximum de la log-vraisemblance).

Ci-dessous, une illustration du processus de résolution :

D'après la relation (3) : $\theta_i = b'^{-1}(\mu_i) = h(\mu_i)$

De plus, d'après (5) : $\mu_i = g^{-1}(\beta_0 + \sum_{k=1}^{p} \beta_k . X_{ik})$.

De là nous pouvons écrire que $\theta_i = h(g^{-1}(\beta_0 + \sum_{k=1}^{p} \beta_k . X_{ik}))$.

Ecrivons la fonction du maximum de vraisemblance :

$$l(y_1, \dots, y_n, \theta_1, \dots, \theta_n) = \prod_{i=1}^{n} f(y_i, \theta_i)$$

Ainsi,

$$l(y_1, \dots, y_n, \beta_1, \dots, \beta_n) = \prod_{i=1}^{n} f\left(y_i, h(g^{-1}(\beta_0 + \sum_{k=1}^{p} \beta_k . X_{ik}))\right)$$

Nous pouvons alors estimer les paramètres β_i en résolvant le système de p équations suivantes :

$$\frac{\partial \ln l(y_1, \dots, y_n, \beta_1, \dots, \beta_n)}{\partial \beta_k} = 0$$

Nous obtenons donc un ensemble d'estimations de paramètres pour le modèle $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$.

En pratique, le grand nombre d'observations auxquelles nous sommes confrontés signifie que la résolution du système d'équations se fait rarement à l'aide de l'algèbre linéaire. Au lieu de cela, des techniques numériques (et en particulier des algorithmes multidimensionnels de Newton-Raphson) sont utilisées.

Le terme d'offset :

Une caractéristique importante qui sera utile dans notre modèle GLM revisité est l'introduction d'un paramètre de décalage. Dans certains cas, l'effet d'une variable explicative est connu. Dans ce cas, plutôt que d'estimer les paramètres β relatifs à cette variable, il convient d'inclure des informations sur cette variable dans le modèle en tant qu'effet connu. Ceci peut être réalisé en introduisant un « terme d'offset » ξ dans la définition du prédicteur linéaire η : $\eta = \boldsymbol{X}.\beta + \xi$

Cependant, nous obtiendrons : $E(Y) = \mu = g^{-1}(\eta) = g^{-1}(\boldsymbol{X}.\beta + \xi)$

Par exemple, l'offset peut être utilisé lors de l'ajustement d'un GLM sur le nombre de sinistres (au lieu de la fréquence de sinistralité). Étant donné que le nombre de sinistres est proportionnel à la mesure de l'exposition ; nous pouvons définir le terme d'offset pour qu'il soit égal au logarithme de l'exposition de chaque observation. Cela se traduira par un modèle de fréquence multiplié par l'exposition.

3.3.1. Qualité de l'ajustement

Un enjeu clé lors de l'exécution d'un modèle à l'aide de GLM est de sélectionner le bon sous-ensemble de covariables. C'est une phase importante de l'analyse et nécessite un bon sens de la modélisation de la part l'analyste ; le but est de faire un compromis pour choisir le meilleur modèle :

- qui est la meilleure explication (ou ajustement) de la variable de réponse Y.
- qui a le moins de covariables possible pour être une méthode de prédiction robuste.

En effet, plus nous introduirons de covariables, plus la modélisation de la variable réponse sera précise. Cependant nous serons confrontés à un risque de surajustement. En réalité ce n'est pas possible pour un modèle de décrire parfaitement la variable de réponse compte tenu de l'expérience de Y qui inclut une erreur aléatoire, également appelée bruit. Le surajustement se produit lorsqu'un modèle décrit une erreur ou bruit aléatoire au lieu de la relation sous-jacente. Cela se produit généralement lorsqu'un modèle est excessivement complexe, comme par exemple, avoir trop de paramètres par rapport au nombre d'observations. Un modèle qui a été sur-ajusté aura généralement de mauvaises performances prédictives. On dit alors que le modèle n'est pas robuste.

Afin d'estimer la qualité du modèle, le modélisateur utilisera différentes statistiques, en particulier la Déviance et le Khi-carré de Pearson.

Déviance :

Une déviance est une mesure de la différence entre les valeurs ajustées et les observations. La définition de la déviance est la suivante :

$$D = 2\phi \sum_{i=1}^{n} [ln\, l(y_i, y_i) - \ln l(y_i, \hat{\mu}_l)]$$

ϕ est le paramètre d'échelle.

$l(y_i, y_i)$ est la vraisemblance "saturée" qui correspond à la vraisemblance dans le cas où les paramètres seraient exactement égaux aux observations ; cela implique que le nombre de paramètres est égal au nombre d'observations, et la vraisemblance saturée est la vraisemblance maximale réalisable.

$l(y_i, \hat{\mu}_l)$ est la vraisemblance avec la moyenne estimée.

Par conséquent, la déviance représente la différence entre la vraisemblance maximale réalisable et la vraisemblance de notre modèle. Plus le modèle est bon, plus la valeur de la déviance est faible.

Khi-carré de Pearson :

Le Khi-carré de Pearson χ^2 est une autre statistique qui évalue la qualité globale de l'ajustement du modèle. Il est défini comme suit :

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i, \hat{\mu}_i)^2}{var(\hat{\mu}_i)}$$

Ces statistiques indiquent à quel point le modèle s'adapte à l'expérience, mais elles ne tiennent pas compte du nombre de paramètres utilisés. C'est pourquoi d'autres statistiques ont été créées telles que les statistiques d'AIC (*Akaike Information Criterion*) et de BIC (*Bayesian Information Criterion*). La logique sous-jacente est que l'inclusion d'un paramètre supplémentaire devrait être pénalisé ; ce qui est logique car nous visons à sélectionner le meilleur modèle qui donne le meilleur ajustement avec le moins de variables possible.

AIC :

La formule de l'AIC est : $AIC = -2 \ln l\,(y_i, \hat{\mu}_l) + 2p$

avec p le nombre de paramètres estimés dans le modèle. Nous mentionnerons également une forme alternative de l'AIC, qui est l'AIC corrigé donné par : $AICc = -2 \ln l\,(y_i, \hat{\mu}_l) + 2p\frac{n}{n-p-1}$

où n est le nombre total d'observations utilisées. L'AICc est généralement utilisé pour les échantillons où le nombre d'observations n est faible ou le nombre de paramètres p est grand ; par conséquent, l'AICc pénalisera davantage l'inclusion d'une covariable supplémentaire.

BIC :

Le BIC est une mesure similaire à l'AIC et est définie comme suit : $BIC = -2 \ln l\,(y_i, \hat{\mu}_l) + p.\log(n)$

Ces statistiques sont utilisées pour comparer les modèles entre eux, par conséquent, plus la statistique est faible, meilleur est le modèle. Ceci est utilisé lorsqu'il s'agit de sélectionner un sous-ensemble de covariables en examinant le modèle incluant et excluant la variable.

3.3.2. Sélection de variables

Une fois les premières modélisations réalisées, l'objectif est de fiabiliser au mieux le modèle et de l'améliorer autant que possible en sélectionnant les variables adéquates. L'optimisation du modèle se déroule en deux étapes :

- la création de modalités pour chaque variable afin de rendre le modèle le plus efficace possible
- la sélection des variables afin de conserver uniquement les variables ajoutant de l'information.

Les variables utilisées peuvent être objectivées de multiples façons : la variable localisation, par exemple, peut être une ville précise, un département ou une région.

Si la variable est trop précise, elle ne revêt que très peu d'informations : si, par exemple, un seul contrat est associé à la ville de Lyon et qu'un sinistre est recensé sur ce contrat, alors la ville de Lyon apparaîtra à tort. Ce biais représente un risque pour le modèle. Au contraire, une segmentation trop large ne permettra pas d'identifier efficacement les différentes zones de risque.

L'optimisation des variables est donc à la fois primordiale et délicate. Par ailleurs, ces variables ne doivent être ni redondantes, ni vectrices d'erreurs tout en conservant le plus d'informations possible.

Une variable est non-significative quand elle présente pour le modèle une $p - value > 5\%$ sur toutes ces modalités, c'est donc une variable qui n'apporte aucune information.

Afin de déterminer si l'inclusion d'une nouvelle covariable est significative ou non, nous pouvons effectuer différents tests :

- la statistique du Wald Chi-Square pour chaque modalité. Cette méthode consiste à tester l'hypothèse nulle. Si le test rejette l'hypothèse, nous pouvons utiliser le modèle le plus complexe incluant la variable. La statistique du Wald Chi-Square se définit de la manière suivante : $Wald\ ChiSquare = \frac{Valeur\ estimée(\beta)}{Erreur\ type^2}$

  Cet indicateur donne une mesure de la quantité d'informations apportée par chaque modalité dans le modèle. Plus le Wald Chi Square est élevé plus la modalité est estimée avec précision.
- Estimation de l'erreur standard des paramètres : l'estimation de l'erreur standard des paramètres $\hat{\beta}$ est possible en raison de la nature asymptotique du maximum de vraisemblance. Sous certaines hypothèses de régularité, l'estimation du maximum de vraisemblance est asymptotiquement normalement distribuée. De plus la moyenne de la distribution est nulle et sa variance égale à l'inverse de la matrice de l'information Fisher. Nous pouvons écrire : $\sqrt{n}(\hat{\beta} - \beta) \longrightarrow N(0, \hat{I}^{-1})$

  Où $\hat{I}$ est l'estimateur de la matrice de l'information de Fisher : $\hat{I}_{j,k} = E\left[\frac{\partial \ln l}{\partial \ln l}.\frac{\partial \ln l}{\partial \hat{\beta}_k}\right]$

  On obtient donc l'intervalle de confiance à 95% pour le paramètre $\hat{\beta}_i$

  $$[\hat{\beta}_i - 1.96 \times \sqrt{\hat{I}_{i,i}^{-1}}\ ; \hat{\beta}_i - 1.96 \times \sqrt{\hat{I}_{i,i}^{-1}}]$$

  Cet intervalle est également appelé *intervalle de confiance de Wald*. Plus l'intervalle autour de la courbe estimée est étroit, plus les paramètres sont précis et meilleurs est le modèle. Un test courant consiste à tracer une ligne horizontale à partir de la base de niveau ; si la ligne peut traverser l'intervalle de confiance, nous pouvons décider de ne pas inclure la variable étant donné que l'incertitude autour des paramètres estimés est trop élevée.
- Jugement : Bien que l'utilisation des statistiques soit importante pour la sélection d'une variable, il est essentiel de faire preuve de jugement et de bon sens. Par conséquent, il est important de vérifier également si le modèle est logique.
- Cohérence dans le temps : ce test consiste à vérifier l'évolution des paramètres lorsqu'ils interagissent avec une variable temporelle (ex : année d'accident). Si le facteur affiche une tendance constante dans le temps, nous pouvons être plus sûrs que la tendance est prédictive de l'avenir.

Afin d'identifier la meilleure combinaison de variables, il existe des processus de sélection de variables qui créent des modèles GLM de façon répétée :

GLM Forward : cette méthode, connue sous le nom de « procédure ascendante », construit une première modélisation avec, en plus de la constante, une unique variable explicative. La variable choisie par le modèle est celle qui implique le plus petit AIC. En d'autres termes, on construit un premier modèle avec la variable qui apporte le plus d'informations. Dans un deuxième temps, on construit un deuxième modèle avec la première variable et la seconde variable qui fait baisser l'AIC du modèle. On répète cette action jusqu'à ce qu'il ne reste plus que des variables capables d'augmenter l'AIC. Ces variables ne seront pas retenues car elles dégradent la significativité du modèle.

GLM Backward : cette méthode est connue sous le nom de « procédure descendante ». Elle est considérée comme étant la procédure opposée à la précédente. Elle utilise un premier GLM avec toutes les variables, puis supprime la variable créant le plus d'erreurs. On voit ainsi l'AIC

diminuer. La procédure réitère cette boucle tant qu'il existe une variable génératrice d'erreurs, c'est-à-dire une variable qui fait augmenter l'AIC quand elle est présente. A la fin de cette procédure, on se retrouve avec un modèle rendu aussi précis que possible.

3.3.3. Validation du modèle

La validation du modèle est une étape importante pour les GLM. Cette validation consiste à vérifier si le modèle est approprié, c'est-à-dire qu'il a de bonnes pouvoir explicatif et qu'il est prédictif. Elle vérifie donc qu'il n'y a pas de surapprentissage. Pour ce faire, le modélisateur ou l'analyste utilisera différents tests parmi lesquels :

L'analyse des Résidus :

Diverses mesures des résidus peuvent être dérivées pour montrer, pour chaque observation, comment la valeur ajustée diffère de l'observation réelle. En pratique, nous utiliserons les résidus de Déviance et Pearson, qui sont liés aux statistiques de Déviance et de Khi-carré de Pearson. Nous donnerons une définition conceptuelle ci-dessous :

Le résidu de déviance r_i^d d'une observation i représente la contribution de l'observation à la Déviance, on peut donc écrire :

$$D = \sum_{i=1}^{n} r_i^d$$

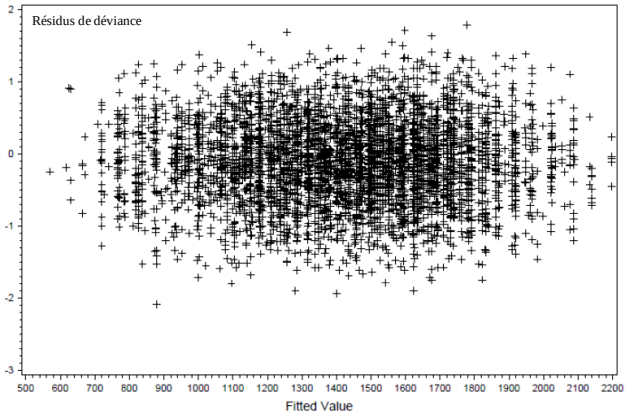
De même, le résidu de Pearson r_i^p d'une observation i représente la contribution de l'observation à la statistique de Pearson, on peut donc écrire :

$$\chi^2 = \sum_{i=1}^{n} r_i^p$$

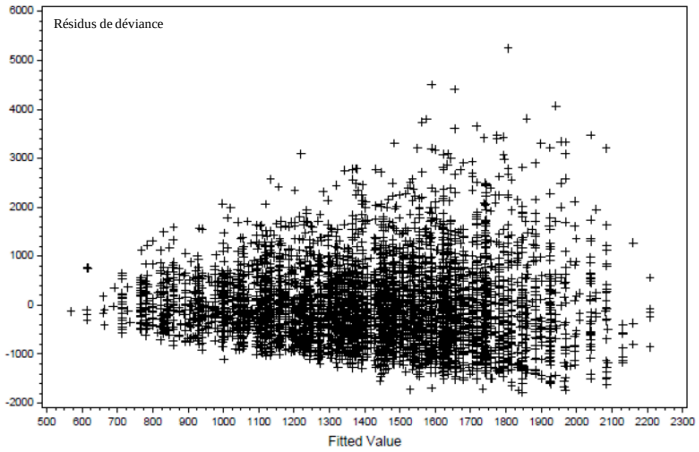
$r_i^d \; et \; r_i^p$ ont le même signe que le résidu brut $(y_i - \hat{\mu}_i)$. Le test consiste donc à observer le nuage de points des résidus par rapport à la valeur estimée $\hat{\mu}_i$ en abscisse.

Si la fonction d'erreur supposée est appropriée, nous nous attendrions à ce que les résidus soient distribués symétriquement autour de l'axe nul, sans tendance évidente du nuage de points. Nous supposerons alors que le terme d'erreur reflète uniquement le bruit.

Par exemple, dans le nuage de points ci-dessous qui montre la déviance résiduelle d'un hypothétique modèle : de gauche à droite du graphique, la moyenne générale et la variabilité des résidus sont raisonnablement constantes, ce qui suggère que la fonction de variance supposée est appropriée :



Contrairement, dans le graphique ci-dessous, on observe que la variabilité augmente avec la valeur ajustée. Cela indique qu'une fonction d'erreur inappropriée a peut-être été sélectionnée et que la variance des observations augmente avec les valeurs ajustées dans une plus grande mesure qu'on ne le supposait. Nous ne devrions pas valider ce modèle.



La base test :

Une méthode couramment utilisée consiste à comparer la prédiction du modèle sur un échantillon retenu. En fait, étant donné que les estimations du modèle sont dérivées de l'échantillon de données ; nous nous attendrions à ce que la prédiction soit proche de l'expérience. En utilisant un ensemble de données de test, nous pouvons vérifier le pouvoir de prédiction du modèle ou vérifier qu'il ne présence pas de surapprentissage.

On se permet donc de diviser les données en deux sous-ensembles avant la modélisation :

3. Données d'apprentissage ou données de modélisation, pour construire le modèle (en général elles comprennent 80% à 90% des données totales).
4. Données de test, le but consistera à comparer les valeurs prédites du modèle sur le jeu de données de test en comparaison avec les observations réelles. Les données de test sont généralement sélectionnées au hasard dans l'échantillon.

   Une autre option consiste à retester le modèle sur les données du futur trimestre ; dans ce cas nul besoin de diviser la base initiale, les données de test seront les données d'une période future.

En générale pour la méthode de la base test, on détermine le MSE (*Mean Squared Error ou erreur quadratique moyenne*) qui est une mesure souvent utilisée pour sa simplicité et son efficacité. Elle peut, en effet, s'appliquer à tous les modèles prédictifs :

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

En pratique, on cherche à minimiser cette mesure.

A la place du *MSE*, on peut utiliser le *RMSE* (Root Mean Square Error). Comme son nom l'indique, c'est la racine carrée du *MSE* et nécessite lui aussi la pratique d'une validation croisée de type « apprentissage et test », c'est à dire que l'on divise la base de données en deux échantillons.

Chapitre 4 : PRESENTATION ET ANALYSE PRELIMINAIRE DES DONNEES

4.1. Données assurantielles

Cette étude utilisera les données issues de la flotte automobile ALD Italie. C'est une grande flotte automobile comprenant plus de 120 000 véhicules assurés par Sogessur sur la garantie responsabilité civile.

Pour les données assurancielles, nous travaillerons essentiellement avec trois bases de données : une base contrat historisée, et deux bases sinistres historisées.

4.1.1. La base contrats

La base contrat est la base de données comportant toutes les informations des assurés et des véhicules assurés. Elle est fournie par STI (Support Technique et Infocentre), ce qui fait d'elle une base très exploitable, bien renseignée qui ne présente pas d'anomalies particulières. C'est une base historisée c'est-à-dire qu'elle contient tous les détails des contrats depuis la première année.

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
Nombre de véhicules	108408	82154	90483	107832	138716	138718	156469	138753	127455	127670
Exposition	106057	82865	82401	101797	130138	135580	159787	144715	130056	52917

Tableau 14 – Exposition et nombre de véhicules dans la base contrats

Les périodes de Y1 à Y10 constituent les années contrat qui commencent toujours en novembre et finissent et en octobre. Le contrat dure donc 1 an et est renouvelable tous les ans. Nous avons la première année (Y1) qui va de novembre 2012 à octobre 2013, ainsi de suite jusqu'à l'année Y10 allant de novembre 2021 à octobre 2022. A noter que les chiffres sur l'année en cours (Y10) vont de novembre 2021 à fin mars 2022.

Dans la base contrat, on trouve entre autres les variables suivantes :

Intitulé de la variable	Descriptif
COIMMAT	Le numéro de la plaque d'immatriculation
COMARK	La marque du véhicule
COMODEL	Le modèle du véhicule
COTYPVEH	Le type du véhicule à savoir si c'est un car (voiture) ou un van (voiture au-delà de 2100kg)
CONBCVF	Nombre de chevaux fiscaux
COPOIDS	Poids du véhicule
COENERGIE	Type d'énergie ou de carburant
COPUISS	Puissance du véhicule
CONBPLACE	Nombre de place du véhicule
COCYLINDREE	Cylindrée du véhicule
COBLACKBOX	Modèle boîtier télématique
CODTIMMAT	Date d'immatriculation du véhicule
COPROVIMMAT	Province d'immatriculation

COKMCONTRAT	Nombre de kilomètres au contrat
CODUREECONTRAT	Durée du contrat : durée de location incluse dans le contrat de leasing
COCONSMAG	Evaluation du client
ATECO_CODE	Code ATECO à la maille deux chiffres permettant d'identifier le secteur d'activité
COTAILLEFLO	Taille de la flotte
COPROVI	Province du contrat
COZONAERO	Zone aéroportuaire : indique si le véhicule se trouve dans une zone aéroportuaire
COMTFRCH	Franchise

Tableau 15 – Les variables présentes dans la base contrats

4.1.2. La base sinistres

Une base sinistre ou base de réclamations, est une base où sont recensées toutes les informations relatives aux accidents déclarés par les assurés. Dans le tableau qui suit, nous avons le nombre de sinistres garantis depuis Y1 ainsi que le nombre de sinistres liés aux voitures équipées :

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
Nombre de Sinistres Garantie	13 168	8 714	9 183	10 235	12 675	13 437	15 473	9 497	8 729	6 589
Sinistres télématiques	-	-	-	1 196	4 462	6 102	9 082	6 946	7 013	5 423

Tableau 16 – Sinistres garanties dans la flotte ALD Italie depuis l'année Y1

***Y10 sur 8 mois : de novembre 2021 à juillet 2022.*

Comme la base contrat, la base sinistre est aussi une base historisée, fournie par STI et mise à jour de façon mensuelle.

Dans la base sinistre, on trouve entre autres les variables suivantes :

Intitulé de la variable	Descriptif
NSIN	Numéro du sinistre
SILIMMAT	Numéro de plaque d'immatriculation
SILDTSURV	Date de survenance du sinistre
SILDTCLOT	Date de clôture du sinistre
SILDTREOUV	Date de réouverture
SILETAT	Etat ou statut du sinistre (ouvert, fermé avec paiement, sans suite)
SILCAUSE	Code de la cause du sinistre
SILRESP	Niveau de responsabilité : totale, partielle ou pas de responsable
SILTYPE	Type de sinistre : matériel, corporel ou mixte
SILPROD	Produit : type de garantie
SILFRAUD	Variable oui ou non, indiquant s'il y a fraude ou pas
SILLIEUSIN	Lieu ou localisation du sinistre
SILPROVINCE	Province de survenance du sinistre
SILPAYS	Pays de survenance du sinistre
SILFRCH	Franchise (oui ou non)

SIGMTPROVOUV	Montant des réserves à l'ouverture
SIGMTPROV	Montant des réserves
SIGMTPROVEXP	Montant des réserves sur les frais d'expertise
SIGMTPAIE	Montant des paiements
SIGMTPAIEEXP	Montant des frais d'expertise
SIGMTPREV	Montant des recours prévus
SIGMTREC	Montant des paiements des recours

Tableau 17 – Les variables dans la base sinistres

Dans cette étude, nous travaillons uniquement sur les sinistres liés à la garantie responsabilité civile : nous ne recenserons que les sinistres où l'assuré est responsable ou partiellement responsable. Grâce à la base sinistre, nous pouvons observer la fréquence matérielle.

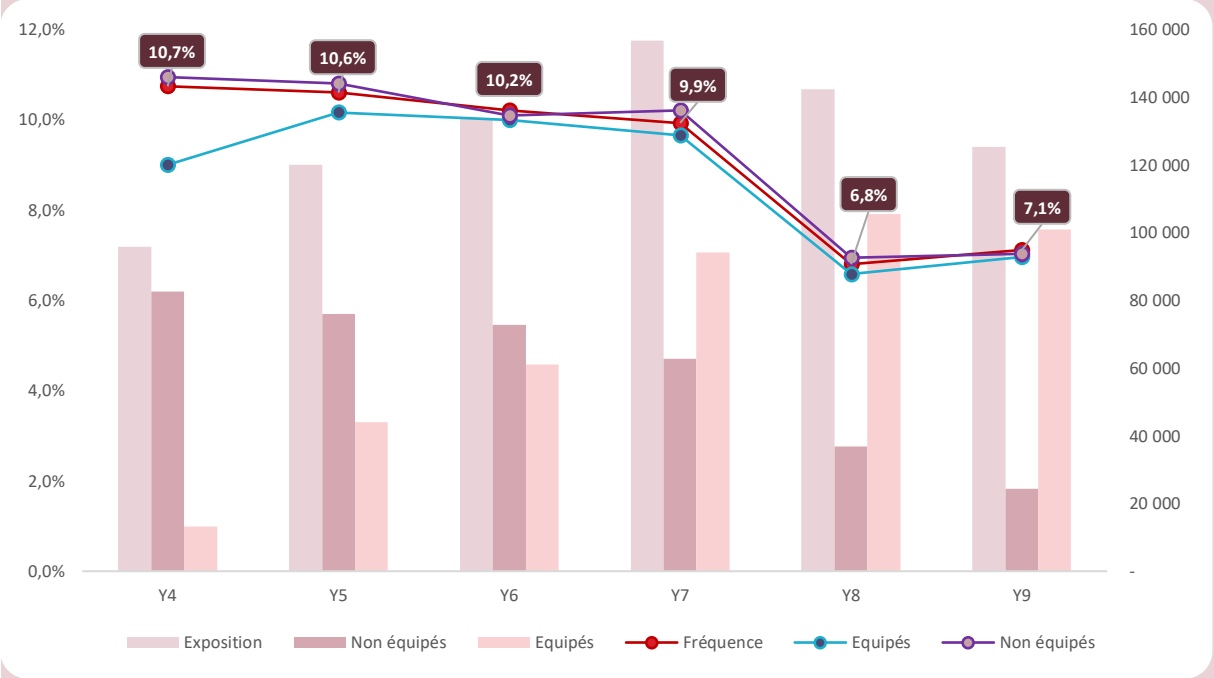


Figure 15 – Evolution de l'exposition et de la sinistralité en fonction des années et des types de véhicules (équipés ou non équipés)

En analysant ce graphique, nous voyons que, la flotte dans sa globalité est composée à environ 80% de voitures équipés et donc 20% de véhicules non-équipés. Partant de l'année Y1 à Y9 nous remarquons une baisse des non-équipés au profit des équipés en termes d'exposition. En ce qui concerne la fréquence totale, nous observons une baisse depuis l'année Y1. Entre l'année qui précède la crise sanitaire (Y7) et l'année COVID (Y8), nous avons une baisse de la fréquence de 3 points (elle passe de 9,9% à 6,8%). Pour l'année qui suit (Y9) nous avons une légère augmentation : fréquence à 7,1%.

En faisant un split de le fréquence entre équipés et non-équipés, la courbe de la fréquence pour les non-équipés présente la même tendance que la fréquence totale mais avec un léger décalage : la fréquence des véhicules équipés est plus élevée. En ce qui concerne les véhicules équipés, ils sont moins dangereux que les non-équipés avec une fréquence 6,95% en année Y9.

4.2. Données télématiques

4.2.1. La base télématique

Les données contenues dans cette base sont un produit direct d'ADI. Ces données proviennent d'un prétraitement préalable fait sur des données brutes à la maille journalière mais aussi d'une construction d'un score de conduite journalier, mensuel et même annuel.

4.2.2. Prétraitement des données brutes

Les données brutes sur lesquelles nous effectuons un prétraitement portent sur 3 bases. Ces bases sont au format csv et nous viennent d'OCTO (compagnie qui met en place les boîtiers télématiques dans les véhicules et fournisseur des données brutes télématiques). Il s'agit de :

• La base des trajets détaillés (Trip_D) qui renseigne l'ensemble des informations à la maille GPS associées à l'ensemble des trajets télématiques observés. Les informations notamment contenues sont l'identifiant du véhicule, l'identifiant du trajet, l'horodatage du point GPS, les coordonnées du point GPS, la qualité du signal, la direction de la boussole, la province, la vitesse et la distance parcourue.

• La base des « crashs » (Crash_S) qui est composée de l'ensemble des informations permettant de géolocaliser un « crash », dont notamment la position et l'horodatage du « crash », et donc de le rattacher à un trajet. Un « crash » se déclenche lors de la survenance d'un événement de conduite dépassant le seuil de déclenchement du boîtier.

• La base des données comportementales (Trip_B) qui comporte les informations relatives à un événement, avec notamment l'identifiant du véhicule, le type d'événement, l'horodatage de l'événement, la position et la qualité du point GPS liées à l'événement.

Nom de la base	Production	Réception	Description
Trip_D	Hebdo	Hebdo	Données détaillées de chaque trajets.
Crash_S	Quotidien	Quotidien	Données agrégées des crashs détectés.
Trip_B	Hebdo	Hebdo	Données comportementales détaillées.

Tableau 18 – Les données télématique à la réception

Une phase de prétraitement de la donnée est nécessaire, car se trouvent dans les bases de nombreuses anomalies du fait d'une mauvaise qualité de transmission de la donnée au niveau du boîtier ou d'un prétraitement automatique inapproprié des données par le fournisseur.

En effet, il peut arriver que le boîtier ou le récepteur présente des défauts de fonctionnement. Cela peut causer un manque de données dans la base. La qualité de la transmission étant plus faible, les informations transférées peuvent alors être bruitées ou de variance aléatoire, ce qui n'est pas souhaitable. Un dysfonctionnement de l'instrument de mesure inclus dans le boîtier peut aussi provoquer des erreurs de relevé et transmettre des données avec une forte incertitude. Enfin, le fournisseur peut, lui aussi, transmettre des données qu'il a lui-même prétraité automatiquement avant de les fournir, et peut ne pas être exempts de tout reproche. Ce genre d'erreurs survient notamment lorsque le fournisseur change sa méthode de calcul sans prévenir.

La première phase du prétraitement des données brutes consiste à un nettoyage de la base. Cela vise principalement à supprimer les doublons, les valeurs manquantes et toutes les aberrations

générées par le système. A la deuxième étape, les données sont agrégées par trajet avec les données comportementales. Et pour finir, ces données sont assemblées par jour et rattachées aux informations contrats et sinistres. A la fin de cette étape, nous disposons d'une base contenant, pour chaque jour et chaque véhicule ayant circulé, les kilomètres effectués, le temps de conduite, le nombre d'événements de chaque type ainsi que les informations contrats et sinistres.

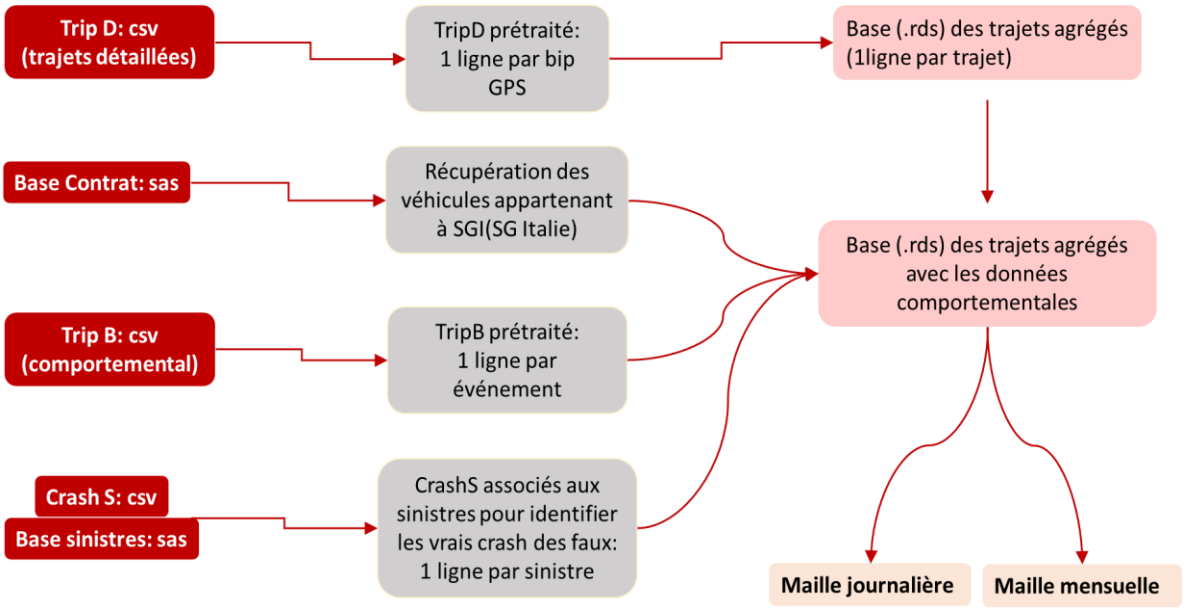


Figure 16 – Les étapes du prétraitement télématique

4.2.3. Création du score de conduite

Le score de conduite est une variable, comprise entre 0 et 100, qui capte le profil de risque d'un individu en fonction de sa manière de conduire.

Le but initial est de créer 4 sous scores comportementaux (par événement) que l'on combinera pour déterminer le score de conduite final. Le score sera construit sur la base de la sinistralité empirique observée et des événements de conduites relevés sur la flotte équipée. Une représentation simplifiée du modèle de scoring est représentée ci-dessous :

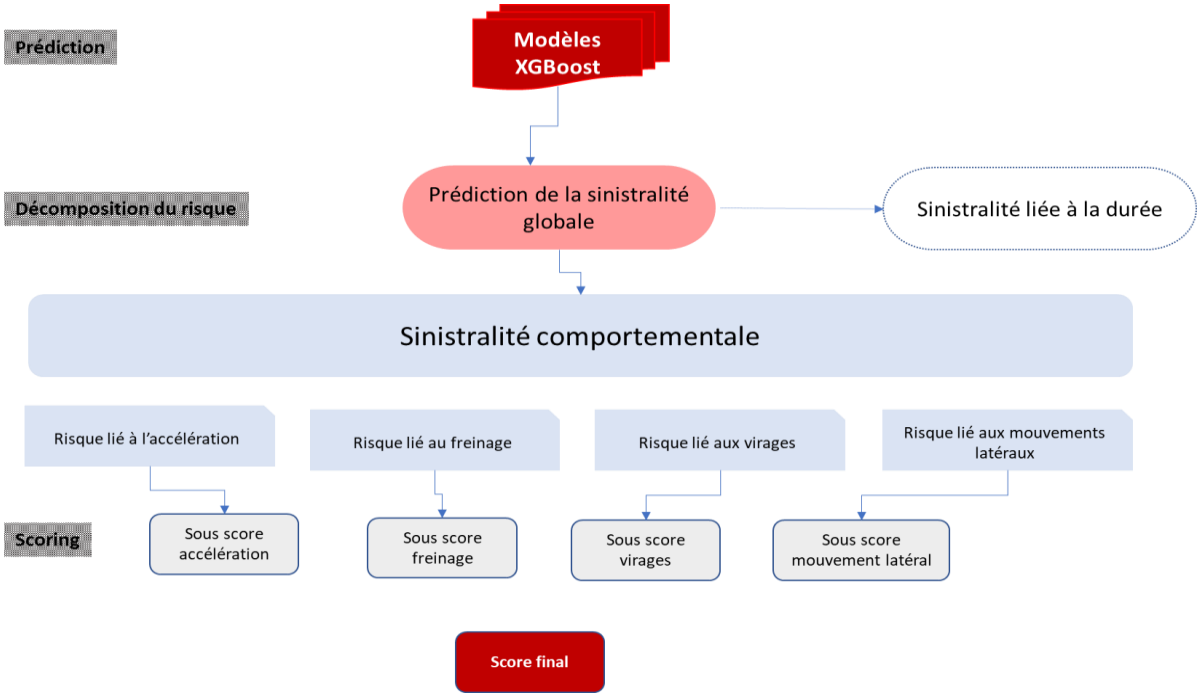


Figure 17 – Procédé de détermination du score de conduite final

Pour la partie prédiction, il a été choisi de se servir d'un modèle XGBoost (Cf. page 121 *Annexe B*). C'est un modèle de Machine Learning basé sur l'apprentissage d'ensemble séquentiel et les arbres de décision. Il fonctionne par récurrence. En effet, il commence par créer un premier modèle de prédiction, donnant à chaque variable un coefficient d'ajout ou de diminution de la probabilité d'avoir un sinistre en fonction de la valeur de la variable. On peut remarquer que sur le schéma ci-dessus lors de la partie « décomposition du risque », nous prendrons uniquement la partie explicative comportementale du modèle et ainsi nous mettrons de côté l'impact de la durée sur la sinistralité. Le risque comportemental sera ensuite scindé en quatre sous scores pour enfin produire un score global journalière, issue d'une moyenne pondérée des quatre sous scores en fonction de leur poids dans la sinistralité comportementale totale.

Si nous voulons ensuite obtenir un score de conduite mensuel ou annuel, il suffit de faire une moyenne des scores du véhicule sur la période désirée, pondérée par la distance conduite chaque jour.

4.3. Statistiques descriptives

4.3.1. Evolution de la distance globale suivant les années

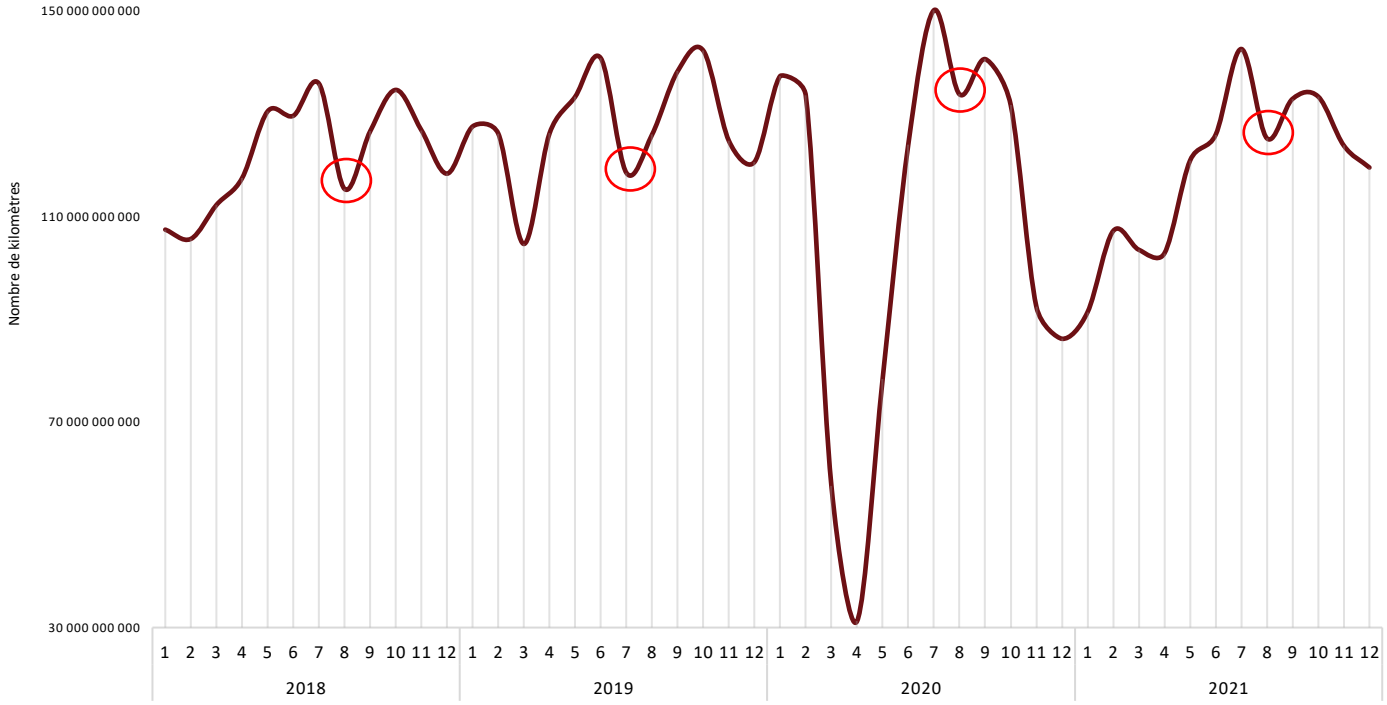


Figure 18 – Evolution du kilométrage total mensuel dans la flotte télématique

En analysant ce graphique, nous remarquons d'abord un pic vers le bas au mois d'août de chaque année. En effet, pendant le mois d'août les Italiens sont en vacances car les températures chaudes ne permettent que difficilement de travailler, c'est le mois préféré des Italiens pour partir en congés. Les villes comme Pise, Venise ou Milan ne contiennent souvent que des touristes étrangers et la circulation y est plus facile que pendant les autres mois de l'année.

Le plus grand pic vers le bas est celui qu'on observe au mois d'avril 2020 en pleine crise sanitaire. Cela s'explique par le confinement qu'a vécu la population sur la période février – mars et qui sera prolongé par la suite jusqu'u mois de mai2020. Pendant cette période, les déplacements étaient strictement limités dans l'ensemble du pays expliquant ainsi cette baisse énorme de la distance parcourue par les véhicules.

Un nouveau confinement va s'installer vers la fin de l'année afin de limiter la propagation du virus pendant les fêtes. Cela explique ainsi le pic vers le bas observé au mois de décembre 2020.

4.3.2. Evolution du score de conduite par année et par génération

Sur les graphiques suivants, nous avons pour chaque année une représentation de la proportion du nombre de véhicules par tranche de score.

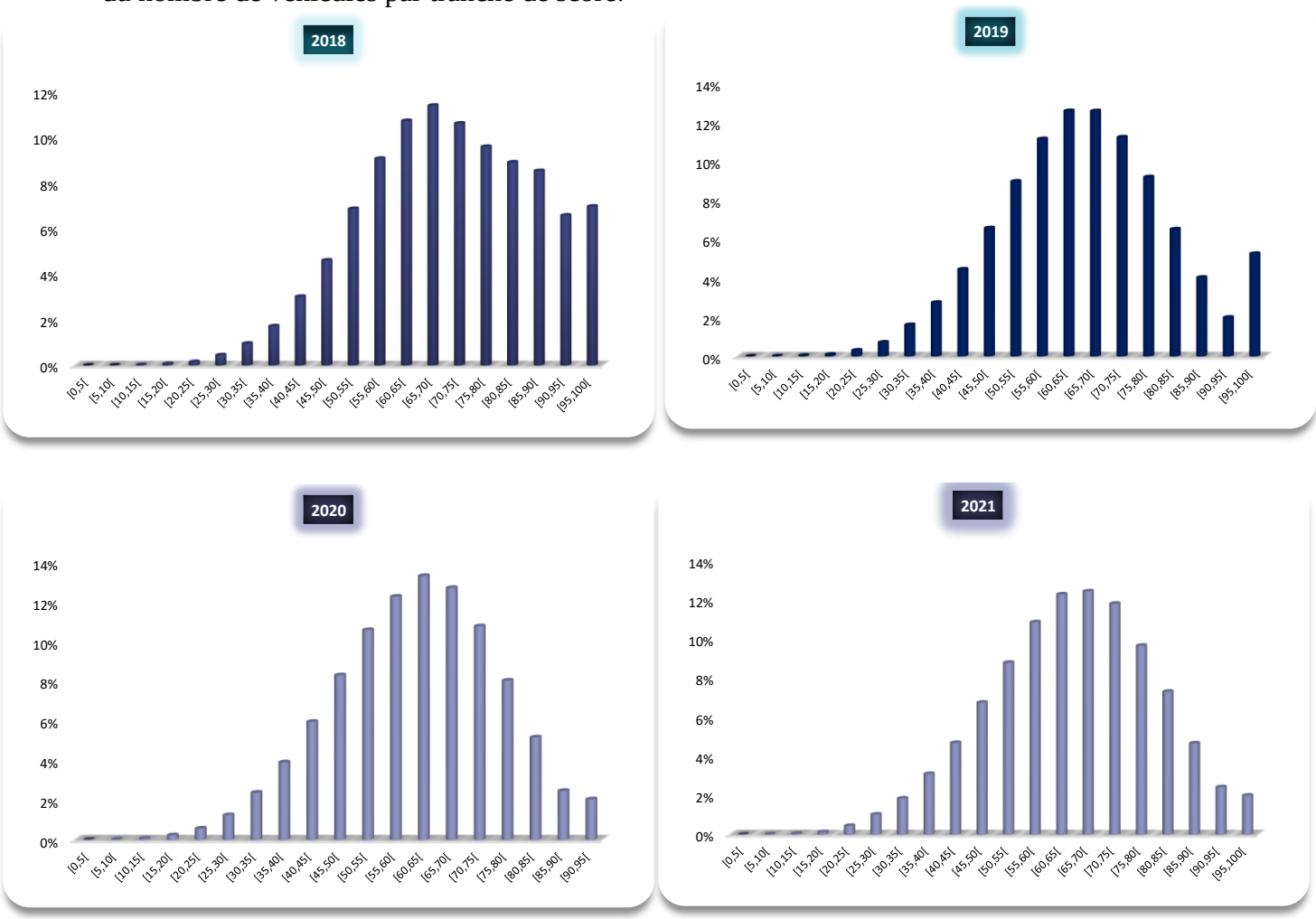


Figure 19 – Evolution du score annuel de conduite (par tranche) par année

Sur l'année 2018, la population la plus représentée est celle qui a un score compris entre 65 et 70. Pour cette même année, 85% des véhicules de la flotte télématique ont un score supérieur à 50. Cela engendre donc une baisse de la population avec un mauvais score au profit de celle avec un bon score de conduite. Les très bons scores, c'est-à-dire ceux supérieure à 80, sont aussi très bien représentés ; environ 30% de la flotte télématique.

Les années qui suivent (2019, 2020 et 2021), quant à elles, montrent une forte baisse des très bons scores : il y'a environ 17% des véhicules qui ont un score compris entre 80 et 100 sur ces années. Selon une étude faite par l'Observatoire de l'Assurance (IoT, Insurance Observatory), cette baisse de la portion des très bons scores est due à une baisse de la vigilance sur leurs habitudes de conduite. En effet, les conducteurs souscrits à une assurance télématique basé sur un score de conduite, sont très attentifs à leurs comportements sur le volant au début, parce que voulant bénéficier à une prime moins chère. Néanmoins, ils se lassent au fil du temps et ne font plus très attention à leur habitudes de conduite. Pour optimiser ce système et inciter les conducteurs à maintenir les bonnes habitudes, L'IoT met en avant comme solution de proposer

aux conducteurs des récompenses pour les motiver encore plus. Par exemple, un café, un smoothie ou un bon de stationnement gratuit (une récompense par semaine) pour chaque tranche de 100 kilomètres consécutifs sans événement.

En ce qui concerne la fraction des scores compris entre 50 et 80, elle a augmenté au fil des années. Elle est passée de 58% (en 2018) à 68% (en 2020). Cela s'explique essentiellement par les mesures restrictives et les confinements observés en 2020 qui ont engendré une baisse de la distance de conduite et donc une évolution du score.

Si nous nous intéressons maintenant à une vision par génération (celles des années Y5 àY9), nous pouvons représenter l'évolution du score moyen par génération dans le portefeuille télématique :

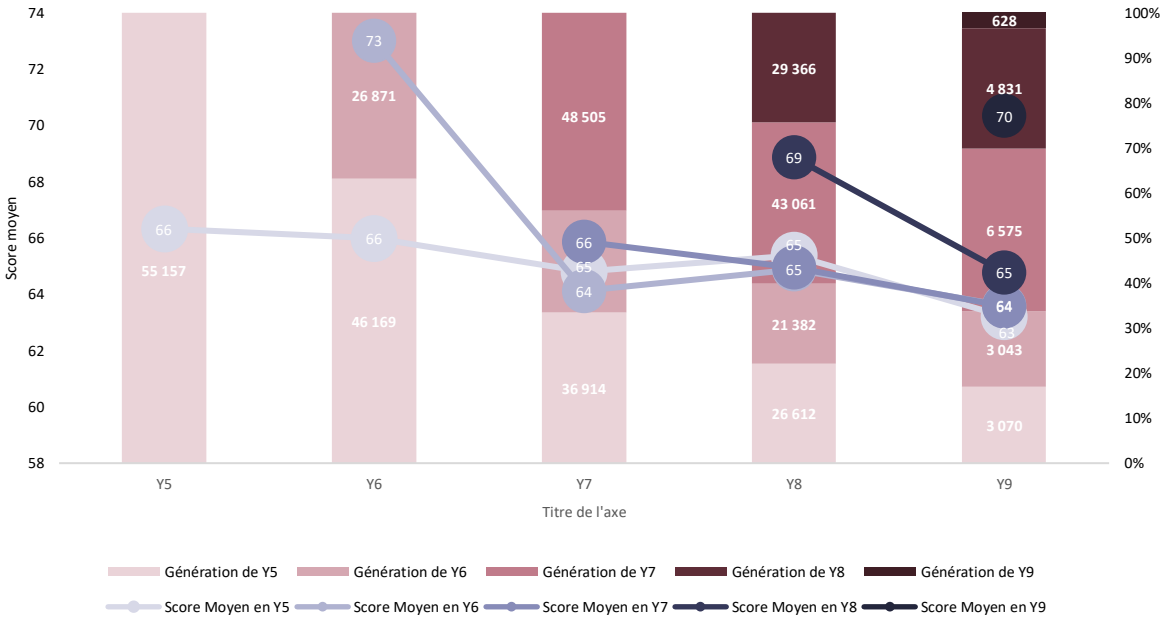


Figure 20 – Evolution du score moyen par génération

Nous voyons que chaque nouvelle génération a un meilleur score la première année et revient à la normale en année N+1. Cela s'explique toujours par une baisse de la vigilance des conducteurs sur leurs habitudes.

4.3.3. Statistiques globales

Dans l'optique d'utiliser de nous familiariser à la base télématique et des variables qu'elle contient, nous effectuerons quelques statistiques descriptives. La base qui sera utilisée pour la réalisation de ces statistiques descriptives est la base scorée annuelle.

La base scorée annuelle a été créée à partir des bases scorées mensuelles des années (contrat) Y6 à Y9. Dans cette base, une ligne représente une plaque d'immatriculation (unique par année) accompagnée des données télématiques qui vont avec (distance et durée totale, distance et durée par type de route et par partie de la journée, nombre de jours roulés dans l'année, les scores par type d'événement et le score global).

A cette base on ajoute le base contrat historisée pour récupérer le type de véhicule dont il s'agit. Cela nous permettra de faire un split entre les voitures et les fourgonnettes.

				Distances en kilomètres							Temps en heures					
Années	véhicules	Nb jours moyen	Km moyen	M	U	O	D	N	W	Heures moyen	M	U	O	D	N	W
Y6	69 184	216	20 168	41%	25%	34%	57%	10%	34%	587	16%	52%	32%	58%	9%	33%
Y7	100 214	207	15 704	42%	31%	28%	60%	11%	29%	460	16%	60%	24%	60%	9%	30%
Y8	103 482	197	12 857	42%	36%	22%	60%	10%	30%	380	16%	67%	17%	61%	9%	31%
Y9	100 803	231	14 820	42%	36%	22%	62%	9%	30%	441	16%	67%	17%	63%	7%	30%
CAR																
Y6	58 782	217	19 339	42%	25%	32%	55%	10%	35%	540	17%	53%	30%	55%	10%	35%
Y7	88 859	208	15 206	42%	31%	27%	59%	10%	30%	434	17%	60%	23%	59%	9%	32%
Y8	94 274	199	12 521	43%	36%	21%	60%	10%	30%	363	17%	67%	16%	60%	9%	32%
Y9	91 836	233	14 389	43%	36%	21%	61%	8%	30%	422	17%	67%	16%	62%	7%	31%
VAN																
Y6	10 402	211	24 857	34%	25%	41%	63%	8%	29%	852	12%	50%	38%	68%	7%	25%
Y7	11 355	199	19 598	36%	30%	34%	63%	13%	24%	659	13%	57%	30%	66%	10%	23%
Y8	9 208	176	16 297	38%	36%	26%	64%	13%	23%	561	14%	67%	20%	68%	9%	22%
Y9	8 967	208	19 242	39%	35%	25%	65%	12%	23%	638	15%	65%	20%	68%	9%	23%

M= Motorways	D= Day(journée)
U= Urbanways	N= Night(soirée)
O= Other roads type	W= Working Hours

Tableau 19 – Statistiques sur les taux de roulages dans l'année en distance et en durée dans la flotte télématique

Dans ce tableau, nous avons pour chaque année, le nombre de véhicules, le nombre de jours moyen roulé dans l'année (jours) la distance (respectivement la durée) moyenne roulée dans l'année en km (respectivement en heures) (moy km respectivement moy h), la distance ainsi que la durée moyenne par type de route et par partie de la journée. Nous avons ensuite reconduit ces résultats le faisant le split voitures standards et commerciaux.

D'après le tableau, en comparant l'année 6 à l'année 7, on voit que l'année Y7 comporte beaucoup plus de véhicule que l'année Y6 (100 214 vs 69 184) néanmoins, on compte en moyenne plus de jours roulés sur l'année 6. Cela se fait ainsi sentir sur la distance mais aussi sur la durée moyenne parcourue.

En termes de type de routes et parties de la journée :

- Distance : les véhicules roulent beaucoup plus sur les autoroutes et en journée
- Durée : les véhicules passent beaucoup plus de temps sur les routes départementales et en journée.

L'année Y8 (du 01/11/2019 au 31/10/2020) sera donc beaucoup plus impactée par la COVID-19. Pour cette année, autre que le nombre de véhicules qui augmente, tous les autres chiffres sont en baisse par rapport à Y6 et Y7. Et comparée à Y9, l'année Y8 présente les mêmes tendances que l'année Y9.

Suivant le split voitures/fourgonnettes, faut savoir qu'environ 85% des véhicules de la flotte télématique ALD Italie sont de type car. Nous observons donc les mêmes résultats que précédemment si nous nous focalisons que sur les voitures. Par ailleurs, même si les voitures sont bien plus nombreuses que les fourgonnettes, les distances ainsi que les durées parcourues par ces derniers dépassent de loin les trajets effectués par les voitures et cela sur n'importe quel type de route et n'importe quelle partie de la journée.

Pour se rapprocher beaucoup plus de la réalité, nous pondérons les résultats précédents par le nombre moyen de jours roulés dans l'année. Ce que nous résumons dans le tableau suivant :

				Distances en km/jour							Temps en min/jour					
Années	**véhicules**	**Nb jours moyen**	**km/jour**	**M**	**U**	**O**	**D**	**N**	**W**	**h/jour**	**M**	**U**	**O**	**D**	**N**	**W**
Y6	69 184	216	93	38	24	32	53	9	31	2,7	26	85	52	95	15	53
Y7	100 214	207	76	32	23	21	46	8	22	2,2	21	79	32	81	12	40
Y8	103 482	197	65	28	24	14	39	7	19	1,9	19	78	19	71	10	35
Y9	100 803	231	64	27	23	14	40	6	19	1,9	19	77	19	72	8	34
CAR																
Y6	58 782	217	89	38	23	29	49	9	31	2,5	25	79	45	82	15	52
Y7	88 859	208	73	31	22	20	43	7	22	2,1	21	75	29	74	11	40
Y8	94 274	199	63	27	23	13	38	6	19	1,8	18	74	18	65	9	35
Y9	91 836	233	62	26	22	13	38	5	19	1,8	18	73	18	67	8	34
VAN																
Y6	10 402	211	118	40	29	48	74	10	34	4	29	121	92	165	17	60
Y7	11 355	199	98	36	30	33	62	13	24	3,3	26	114	59	132	21	46
Y8	9 208	176	93	35	34	24	59	12	21	3,2	26	128	38	131	18	43
Y9	8 967	208	93	36	33	24	60	11	21	3,1	27	120	37	125	17	42

Tableau 20 – Statistiques sur les taux de roulages dans la journée en distance et en durée dans la flotte télématique

Dans ce tableau, contrairement au précédent, on voit que les années Y5 et Y6 présentent des résultats très proches en termes de distance et de durée. Ensuite l'année Y7 est en baisse par rapport aux deux années précédentes. Enfin les année Y8 et Y9 qui présentent aussi les mêmes tendances sont en forte baisse par rapport aux trois premières années.

Nous retenons que dans la flotte télématique, en se basant que sur les deux dernières années, une voiture effectue en moyenne 63km/jour et 1,9 heures/jour et une fourgonnette en fait 93km/jour et 3,1 heures/jour.

Le graphique qui suit montre la répartition des véhicules dans la flotte télématique ALD Italie par année et suivant leur type :

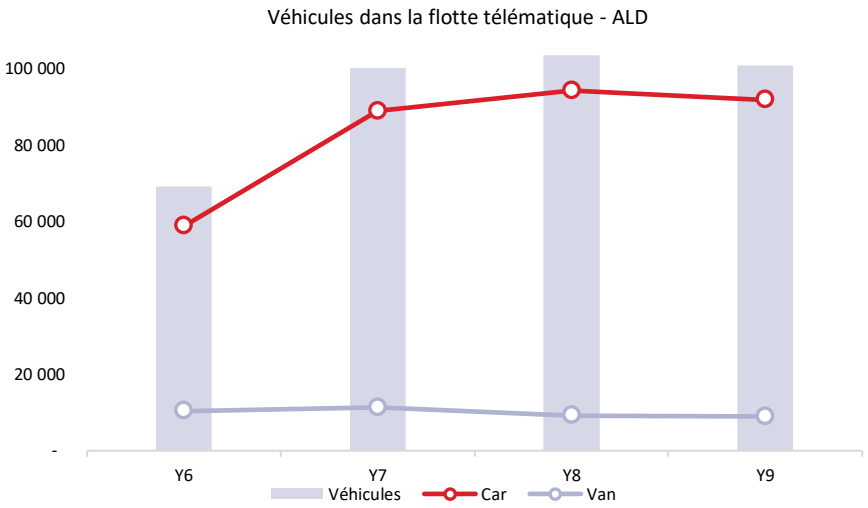


Figure 21 – Répartition des véhicules dans la flotte télématiques ALD Italie par type (voitures standards/commerciales)

Dans la flotte télématique, plus de 88% des véhicules sont de type « car » (voitures standards). Les 12% restant est constitué par les « van » (voitures commerciales). En effet, on retrouve cette même répartition des véhicule dans la flotte ALD Italie dans sa globalité.

4.3.4. Statistiques sur la distance

A partir des données globales de la base scorée par année, nous avons établis des tranches de distance de sorte à avoir des modalités plus ou moins équilibrées. Nous mettons sous forme graphique la répartition de ces tranches de distances en fonction du pourcentage du nombre de véhicules dans la flotte par année.

Pour chaque graphique, nous mettrons à coté la distribution de la fonction de répartition (pourcentage cumulé de véhicules dans la flotte).

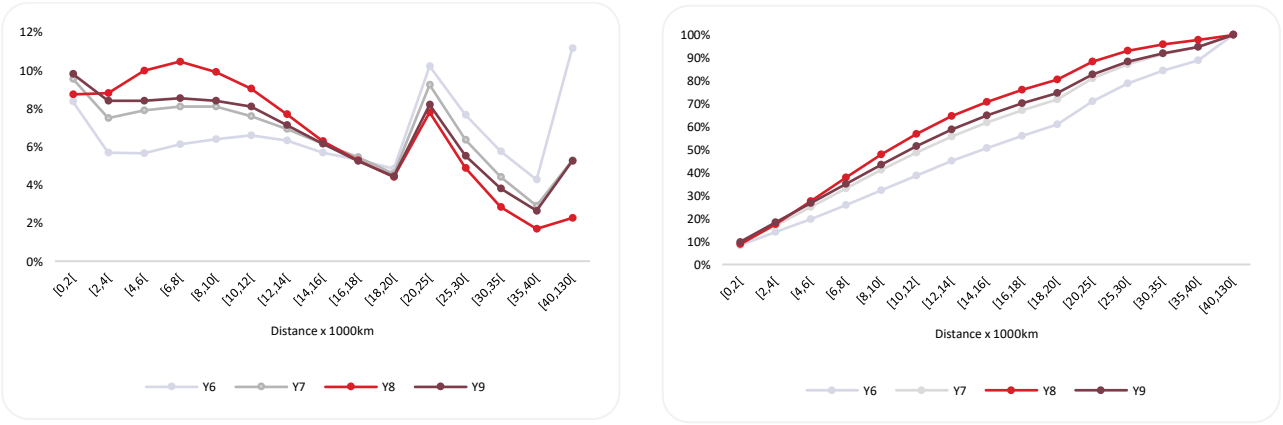


Figure 22 – Répartition des tranches de kilomètres dans la flotte télématique

Sur ces graphiques, en partant de l'année Y6 à l'année Y9, nous observons une augmentation de la fraction des véhicules effectuant de petites distances. Nous voyons qu'environ 60% des véhicules en Y9 roulent moins de 10 000km dans l'année.

Inversement, nous remarquons une diminution de la fraction des véhicules effectuant de grandes distances. De Y6 à Y9, le pourcentage des véhicules faisant plus de 20 000km a fortement baissé.

Sur ces graphes, la rupture observée sur la tranche [20,25[est due au pas de la tranche (5 000km de différence).

En faisant une pondération de la distance par le nombre de jours moyen roulés dans l'année, nous pouvons avoir une vision plus détaillée des résultats précédents comme le montrent les graphiques suivants :

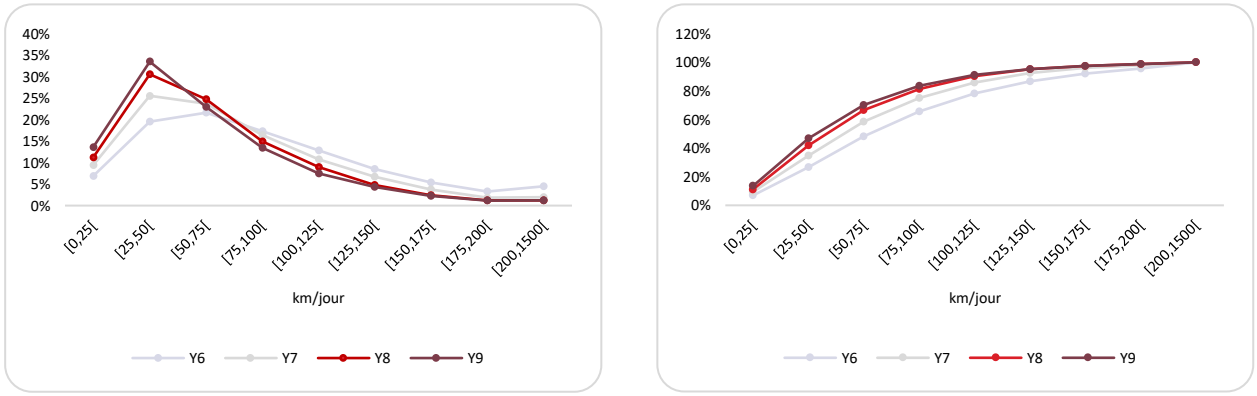


Figure 23 – Répartition des tranches de kilomètres par jour dans la flotte télématique

D'après ces deux graphes, nous observons une augmentation de la portion des véhicules faisant moins de 75km/jour (en Y9 environ 70% des véhicules) et une diminution de la fraction des véhicules faisant plus de 100km/jour.

Split car/van :

En faisant une analyse différentiée selon le type de véhicule, nous constatons que les résultats obtenus précédemment sur l'analyse de la distance globale sont les mêmes pour les voitures, néanmoins pour les fourgonnettes, même si nous avons les mêmes tendances suivant les années, nous voyons que nous avons plus de 66% des van qui roulent moins de 20 000km dans l'année.

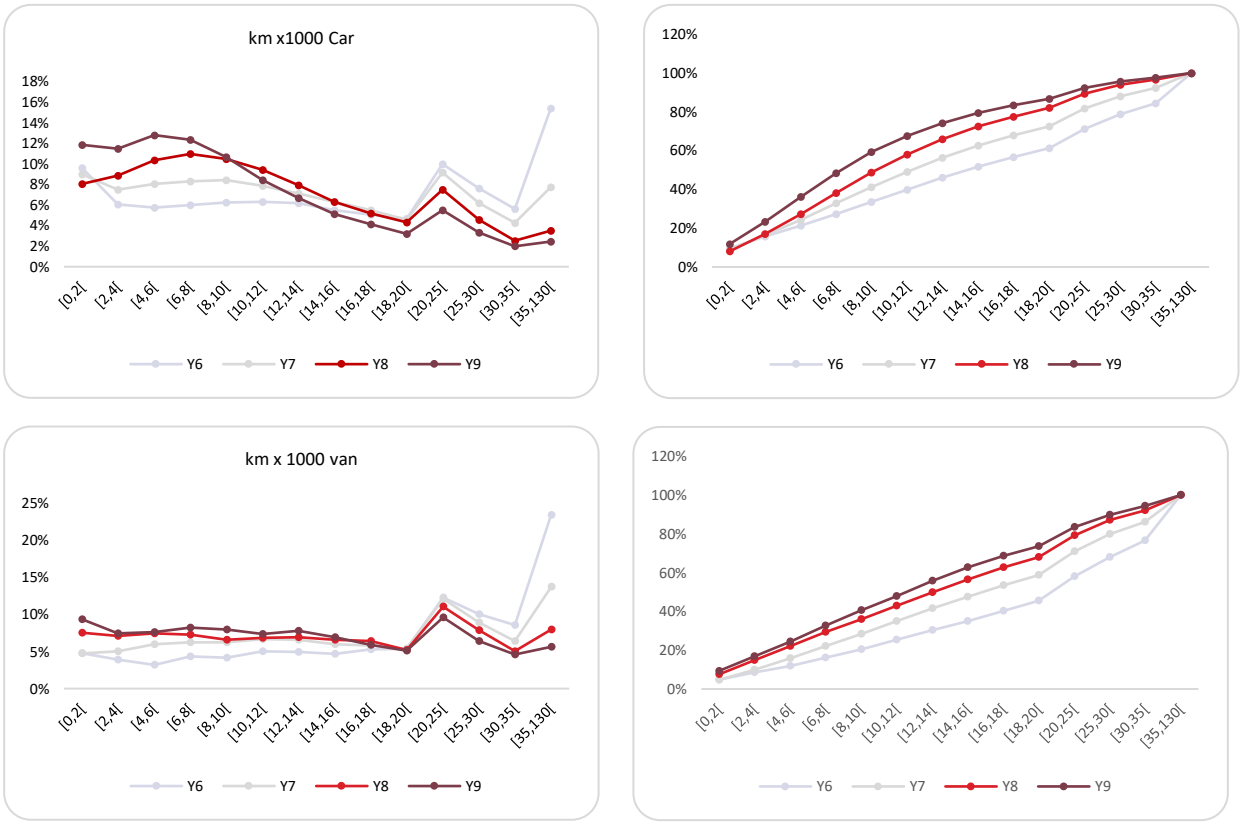


Figure 24 – Répartition des tranches de kilomètres suivant les types de véhicules dans la flotte télématique

4.3.5. Statistiques sur la durée

Si nous passons à l'analyse du temps passé sur les routes par année, nous observons les mêmes tendances que sur les distances. Les variables « durée de conduite » et « distance de conduite » sont fortement corrélées.

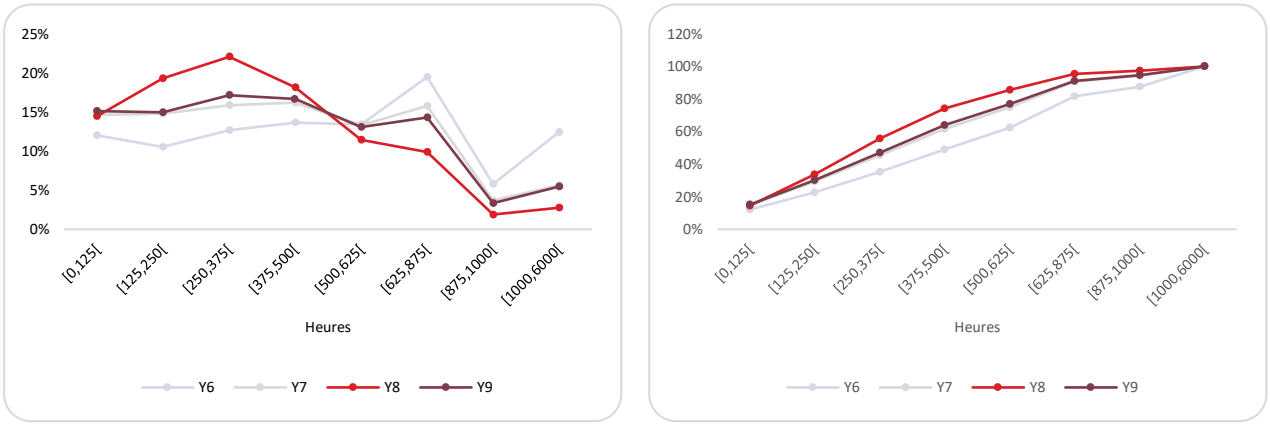


Figure 25 – Répartition des tranches d'heures dans la flotte télématique

Globalement, nous avons en Y9 environ 64% de la flotte qui roulent moins de 375 heures dans l'année.

PARTIE III

ETUDES DES MODELES ET PROJECTIONS

Chapitre 5 : ANALYSE DES IMPACTS DE LA COVID-19

5.1. Introduction – chiffres clés

La flambée de la pandémie et la succession de confinements plus ou moins sévères au cours de l'année 2020 ont eu un impact particulier sur l'assurance responsabilité civile automobile.

D'une part, selon le rapport (2020) de l'Ania (Association Nationale des compagnies d'assurance), les primes ont diminué (de près de 6 %). D'autre part, la fréquence des sinistres, elle aussi, a connu une baisse considérable à la suite du changement significatif de la mobilité, en raison des mesures restrictives.

Le taux de rétention a fortement chuté en mars, mais est revenu aux niveaux prépandémiques en mai 2020, alors que les mesures de restriction se sont assouplies et que les initiatives de réduction des primes ont eu un impact sur les comportements de conduite. Il en résulte, malgré une baisse drastique des profits de placement, une amélioration du résultat du compte technique de cette branche, qui s'établit à 1,5 milliard d'euros (Tout véhiculés concernés).

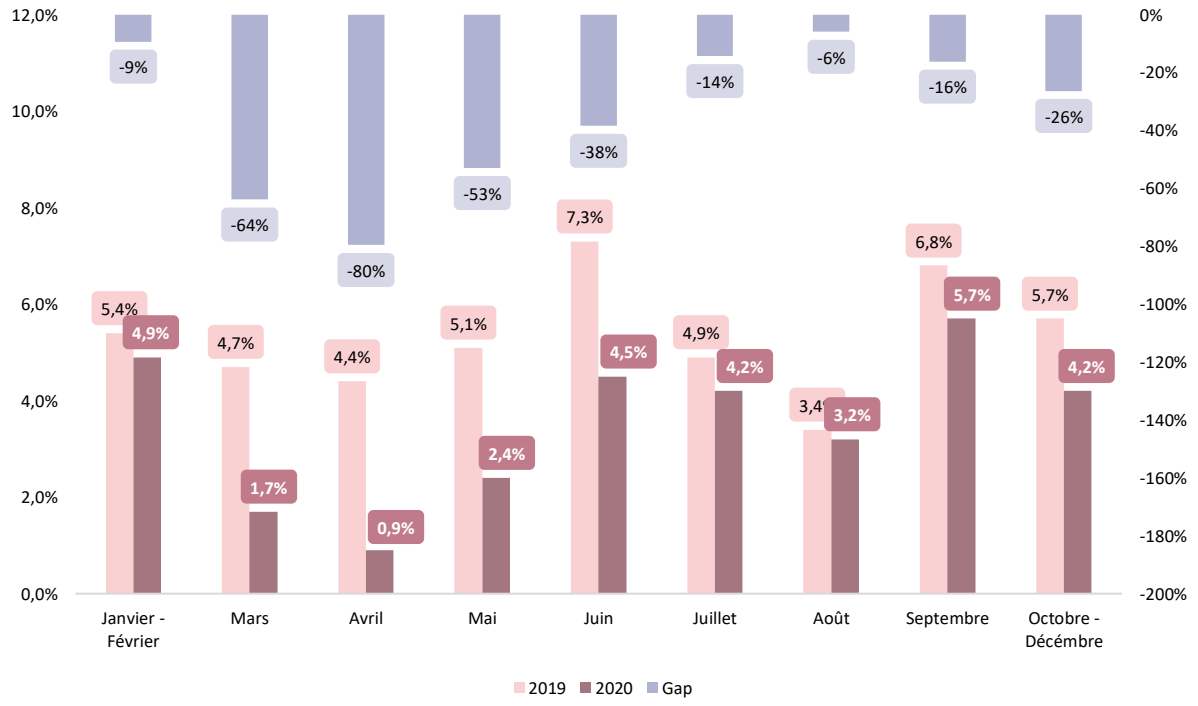


Figure 26 – Fréquences des sinistres sur la garantie RC en 2019 et 2020. Source : ASTIN (Actuarial Studies in Non-life Insurance)

Le graphique ci-dessous montre l'évolution de la fréquence de sinistralité en 2019 et en 2020 ainsi que les changements annuels observés. Cela nous permet donc de relever les gaps observés entre ces deux années. Nous pouvons nettement voir la baisse significative de la fréquence sur la période mars – mai qui représente la première phase de confinement.

Aussi, comme le montre le graphique suivant, le coût moyen relatif des sinistres a considérablement augmenté (+13 %) par rapport à l'année précédente. Cela s'explique par l'inflation, mais aussi par la baisse du nombre de petits sinistres (sinistres à coûts bas).

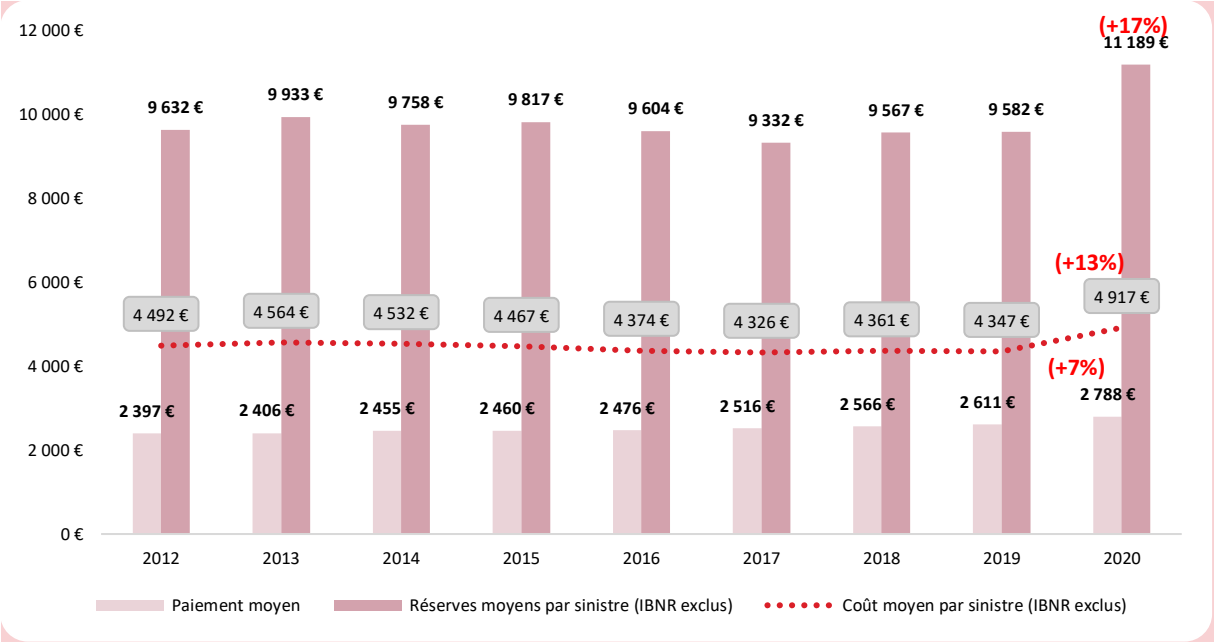


Figure 27 – Coûts moyens des sinistres sur la garantie RC en 2019 et 2020. Source : ASTIN

Cette augmentation atténue en partie les effets bénéfiques de la baisse de la fréquence de sinistralité.

La flotte ALD Italie suit la même tendance observée sur le marché quant à l'évolution de sa fréquence de sinistralité.

5.2. Sinistralité – graphiques (ALD Italie)

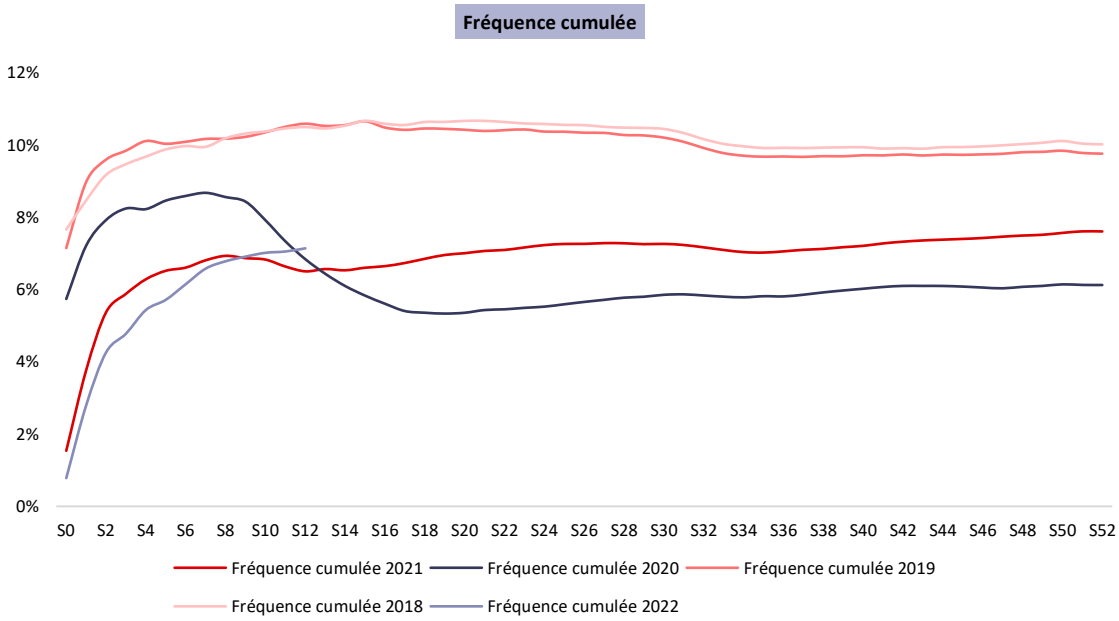


Figure 28 – Evolution des cumuls de fréquences matérielles dans la flotte ALD Italie de 2018 à 2022

Les années 2018 et 2019 présentent les mêmes tendances en termes d'évolution et de niveau de la fréquence. La fréquence en 2021 se situe entre la période pré-COVID et l'année 2020 impactée par la pandémie. En ce qui concerne l'année en cours, la courbe suit la même tendance que 2021. La fréquence en 2022 est légèrement plus faible que celle de 2021 sur les 8 première semaines. La remontée de la fréquence à partir de la 8ème semaine de 2022, malgré le télétravail, s'explique par la situation du trafic qui reprend petit à petit mais qui n'a pas retrouvé son état d'avant-crise.

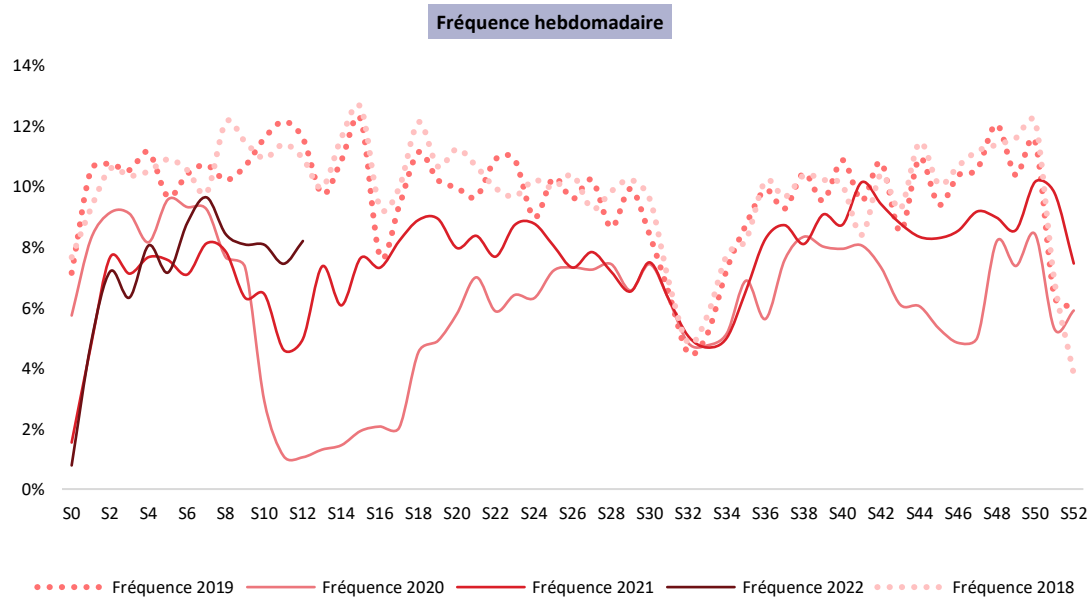


Figure 29 – Evolution de la fréquence matérielle dans la flotte ALD Italie suivant les semaines de l'année (de 2018 à 2022)

Les pics de fréquence en 2018 et 2019 s'observent principalement pendant les 16 et 51ème semaine de l'année. En 2020 on observe une baisse de la fréquence à partir de la 9ème semaine (début mars). Pic à 9,7% (en février). En 2021, nous avons une fréquence plus faible qu'en 2019 et plus haute qu'en 2020, avec un pic à 10,8% en octobre. Pour l'année 2022, nous tirons la même conclusion que sur le graphique des fréquences cumulées.

Sur le graphe suivant, nous allons essayer de mettre en évidence une corrélation entre et kilomètres parcourus en nous concentrant sur les années 2019 – 2021 :

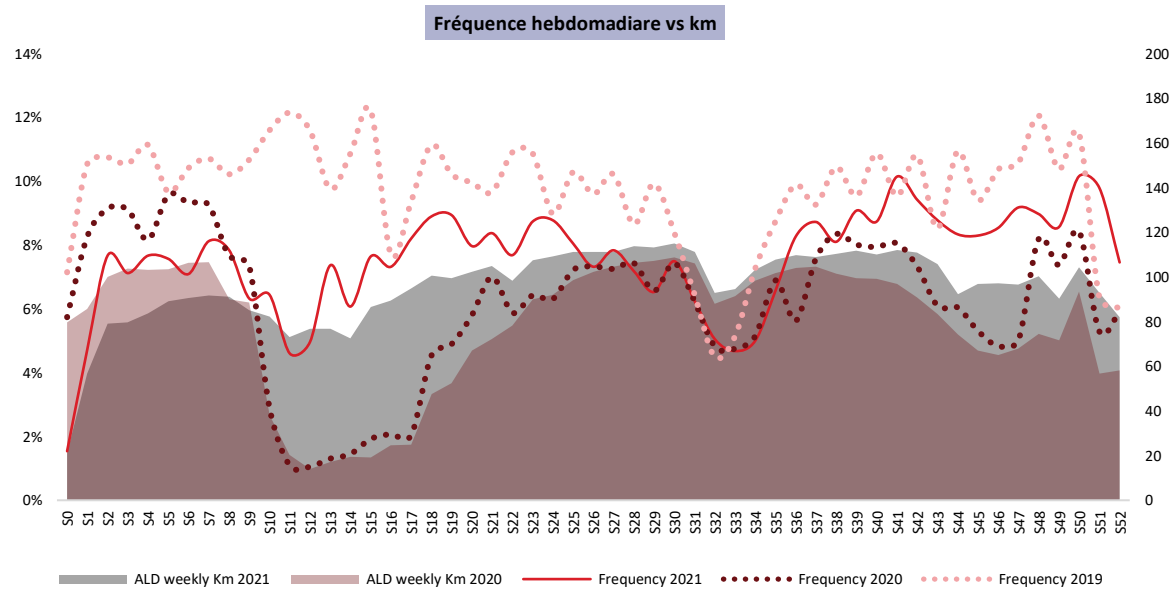


Figure 30 – Evolution de la fréquence matérielle et des kilomètres hebdomadaires de 2019 à 2021

La fréquence en 2019 est plus élevée que celles observées en 2020 et en 2021. Avant la 8ème semaine de l'année, les distances observées en 2021 sont plus faibles que celles observées en 2020 et après c'est la situation inverse qui se produit. Cela s'explique toujours par le confinement sur la période avril-mai 2020. Sur l'année 2020, la courbe de la fréquence et celle des distances suivent une même tendance en termes d'évolution. Nous verrons cela plus en détail par la suite.

Chapitre 6 : MODELISATION GLM DE LA FREQUENCE MATERIELLE

Dans cette partie, l'objectif est d'évaluer l'effet de la COVID ainsi que l'apport des variables télématiques dans la modélisation GLM de la fréquence matérielle pour la garantie RC. Pour le modèle de fréquence, la variable cible est le nombre de sinistres. Nous utilisons l'exposition comme variable de décalage (offset) pour ajuster le nombre de sinistres puisque l'objectif est de modéliser le nombre de sinistres attendu sur une période d'exposition d'un an.

Dans un premier temps, nous quantifierons la contribution des variables télématiques dans le modèle GLM classique (modèle avec juste des variables contrat) à la suite de la mise-à-jour des modalités des variables. Ensuite, nous tenterons de prouver cet apport télématique ou du moins de comparer le modèle retenu avec le modèle classique. Puis, dans la perspective d'évaluer la portée de la COVID, nous créerons une nouvelle variable « *COVID* » que nous intégrerons dans chaque modèle. Pour finir nous effectuerons des prédictions sur ces quatre modèles pour valider et retenir le modèle qui prédit le mieux la sinistralité, en d'autres termes le modèle qui est plus proche des observations.

Modèles	Descriptions
F1	Prend en compte que les variables contrat (âge véhicule, marque, segmentation car-van…)
F1-bis	Modèle F1 + variable « *Covid* »
F2	Prend en compte les variables télématiques (score, distance…) plus quelques variables contrat
F2-bis	Modèle F2 + variable « *Covid* »

Tableau 21 – Les différents modèles mis en place pour la modélisation de la fréquence matérielle

6.1. Rappel sur les données utilisées

Dans le but d'étudier les modèles décrits ci-dessus, nous utiliserons :

- La base contrat
- Les bases sinistres
- La base télématique scorée par année

Nous nous concentrerons sur les années de 6 à 9 (novembre 2017 à octobre 2021).

6.2. Quelques statistiques

6.2.1. Exposition et fréquence dans la base télématique

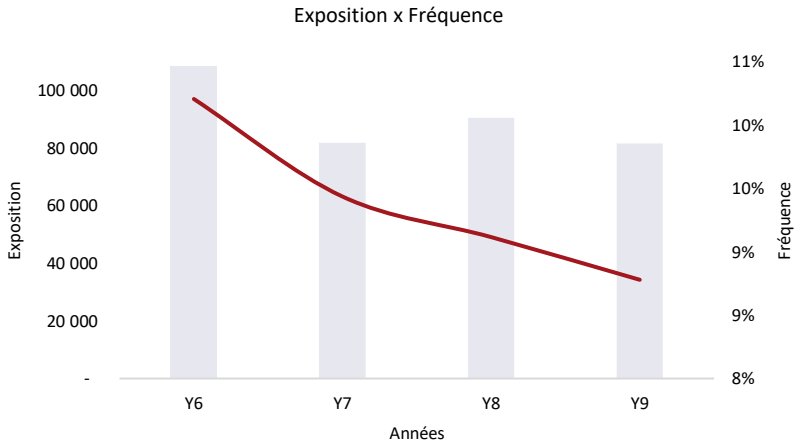


Figure 31 – Evolution de l'exposition et de la fréquence de sinistralité matérielle dans la flotte télématique de Y6 àY9

Partant de l'année 6 à l'année 9, nous remarquons une baisse de la fréquence de sinistralité dans la flotte télématique. Comme dans toute la flotte, cette baisse s'explique par les mesures restrictives durant les différentes confinements et donc par une réduction de l'activité des automobilistes sur les routes.

6.2.2. Fréquence vs score global

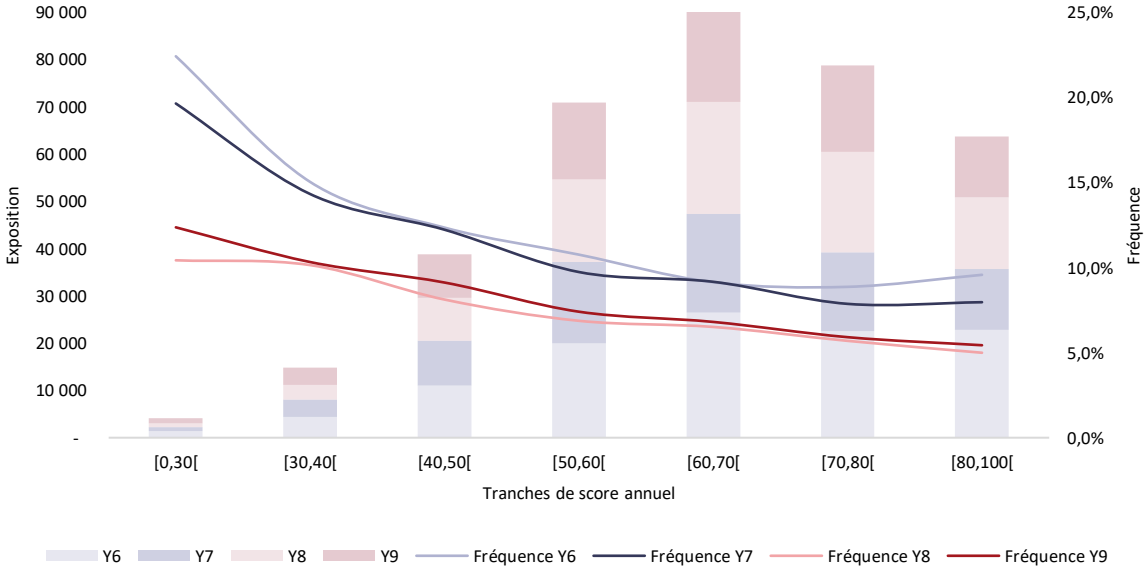


Figure 32 – Evolution de la fréquence matérielle en fonction su score global de conduite dans l'année

Ici, Le premier constat que nous faisons ressort de la baisse de la fréquence de sinistralité suivant l'augmentation du score. Ce qui est tout à fait légitime : plus on a un bon score de conduite moins on est dangereux. Les tendances de sinistralité sur les années non-COVID (Y6 et Y7) sont quasi les mêmes. Les fréquences sur les années COVID (Y8 et Y9) sont aussi très

proches, même si la fréquence sur l'année Y9 est légèrement supérieure à celle de l'année Y8 (conséquence de la reprise progressive des activités des conducteurs).

6.2.3. Fréquence vs distance

Durant la crise sanitaire, à la suite des confinements, le taux de congestion (indicateur pour mesurer l'ampleur des embouteillages) a fortement baissé. En effet selon une étude menée par *TomTom* (éditeur de logiciels de planification d'itinéraires et fabricant de systèmes de navigation GPS mobiles ou embarqué dans certains véhicules) en janvier 2021, auprès de 600 millions de véhicules dans le monde et notamment en France, il ressort que la congestion routière a chuté. Nous avons comme exemple une baisse de 14 % en 2020 dans l'Hexagone et même de 21 % en moyenne aux heures de pointe, comparé à 2019. Cette baisse drastique du trafic est une conséquence des restrictions de travail (télétravail) et de voyage. Cela a engendrée une décrue des déplacements effectués. Pour la sécurité routière, cela a eu un avantage : moins de trajets signifie moins d'accidents.

Le graphique ci-dessous montre l'évolution de la fréquence de sinistralité matérielle en fonction de la distance effectuée :

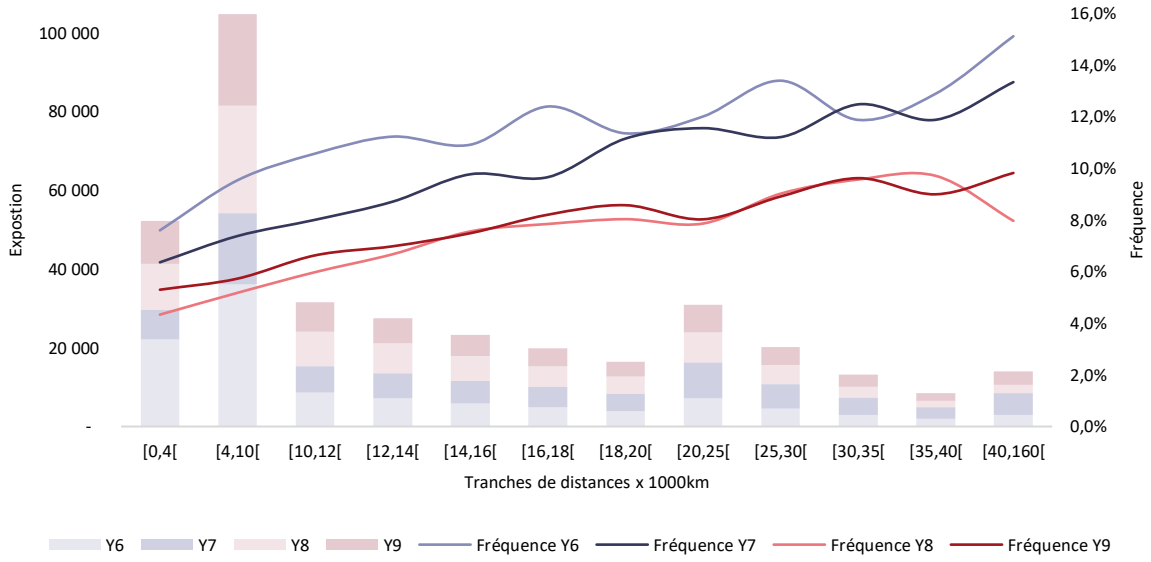


Figure 33 – Evolution de la fréquence matérielle en fonction des kilomètres effectués par véhicule dans l'année

N.B : sur ce graphique, en abscisse nous avons les tranches de distance x 1000km

Nous retrouvons la même conclusion, quand nous mettons la distance à la place du score de conduite, même si pour la distance : suivant les années, plus la distance augmente, plus la fréquence augmente. Sur l'année Y8, la baisse de la fréquence sur les grandes distances (supérieur à 40 000km à l'année) est due à une faible exposition pour cette tranche de population. En effet, avec la crise sanitaire, les trajets en voiture ont été considérablement réduits en raison du confinement, entraînant une diminution considérable des trajets longs.

6.2.4. Fréquence vs score urban

La variable « *score urban* » est le rapport entre la distance globale et la distance sur les routes urbaines (en ville), c'est donc le taux de roulage sur les routes urbaines. Elle est comprise entre 0 et 100. Dans toute la suite nous la qualifierons par TRU (taux de roulage urbain).

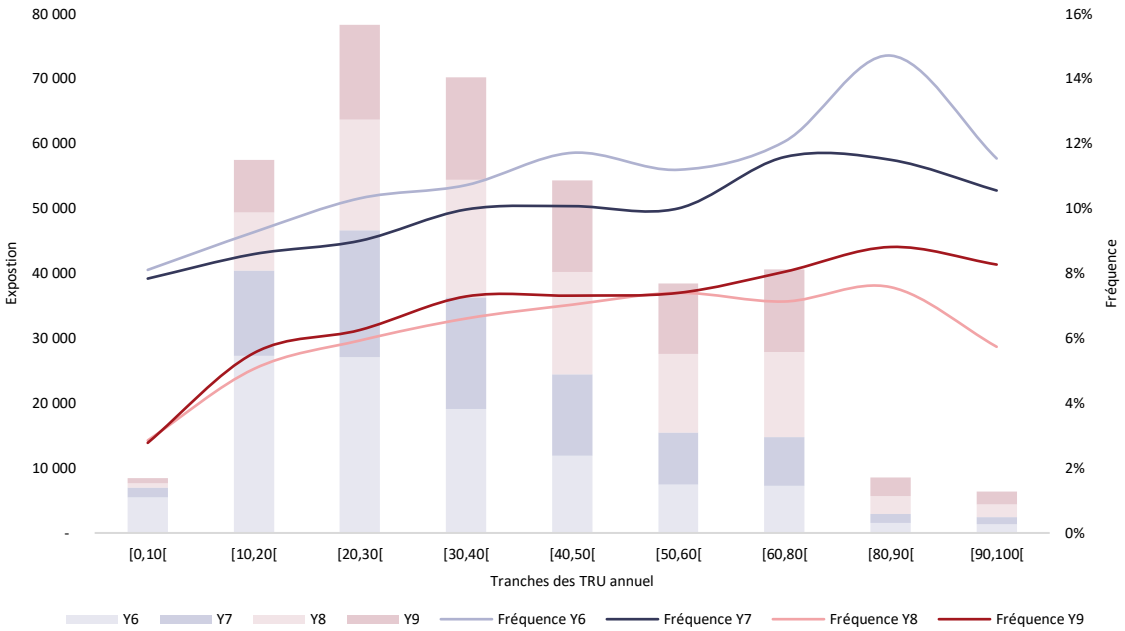


Figure 34 – Evolution de la fréquence matérielle en fonction du taux de roulage urbain (TRU) annuel

En remplaçant la distance globale par le TRU, cette même tendance à la baisse de la sinistralité par année se manifeste. Pour la fraction de population qui a un TRU entre 90 et 100, la baisse de la fréquence s'explique par une faible exposition et du nombre de sinistres.

6.2.5. Fréquence vs taux de roulage non-urbain (TRNU)

Nous observons la situation inverse avec les TRNU (taux de roulage sur les routes non urbaines). Le TRNU est le rapport de la distance effectuée sur les autoroutes plus la distance effectuée sur les autres types de routes et la distance globale. Ainsi plus le TRNU est faible, plus la fréquence est élevée. Ceci en comparaison avec le graphique précédent permet de dire que les routes urbaines sont plus dangereuses comparées aux autres types des routes.

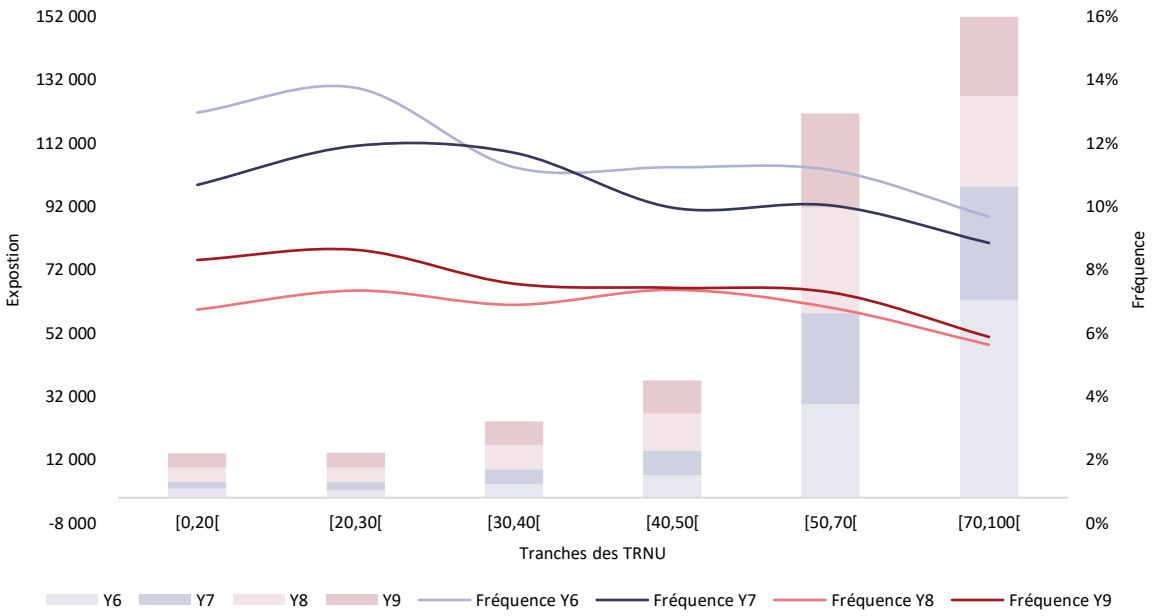


Figure 35 – Evolution de la fréquence matérielle en fonction du taux de roulage urbain (TRU) annuel

Après cette analyse détaillée de quelque variables télématiques, nous pouvons maintenant passer à l'étude des différents modèles.

6.3. Mode opératoire

Nous divisons la base initiale (base contenant les informations contrats, sinistres et télématiques) en deux bases : une base apprentissage qui contient 80% des observations et une base test qui contient les 20% restant. Ainsi, nous entraînons d'abord nos quatre modèles sur la base apprentissage ensuite nous mesurerons leur pouvoir de prédiction sur la base test.

6.4. Premières modélisations les modèles F1 et F1-bis : mesure de l'impact de la variable « covid » sur la modélisation classique

Pour la mise en place des premiers modèles, nous présenterons d'abord les variables que nous jugeons nécessaires à intégrer dans l'étude, ensuite nous analyserons les résultats obtenus pour en déduire si la variable « COVID » induit un effet positif ou négatif dans le modèle classique.

Les variables que l'on retrouve dans ces premiers modèles sont :

- L'âge du véhicule
- La segmentation (voiture standard, fourgonnette légère, lourde)
- Le type d'énergie
- Le code ATECO (secteur d'activité)
- La zone
- La taille de la flotte
- Le nombre de chevaux fiscaux du véhicule (*NBCVF*)
- La marque du véhicule.

Pour le moment, nous n'utilisons pas la variable « année » dans nos modèles puisqu'elle servira à créer la variable « COVID » par la suite. Aussi nous nous baserons sur cette variable « année » pour effectuer les analyses prédictives.

Appliquons maintenant le modèle F1 sur la base apprentissage, avec la procédure GENMOD sur SAS nous obtenons la sortie suivante :

Analyse des paramètres estimés du maximum de vraisemblance du modèle F1										
Paramètres		DDL	Estimation	Erreur	IC de Wald à95%		Khi-2 de Wald	Pr > khi-2	Relativy	
Intercept		1	-2,672	0,0189	-2,7091	-2,6349	19883,8	<,0001	6,9%	
Age véhicule	+3	1	-0,1812	0,0371	-0,2539	-0,1085	23,89	<,0001	0,83	6,3%
Age véhicule	0	1	-0,0982	0,0168	-0,131	-0,0653	34,35	<,0001	0,91	6,9%
Age véhicule	2	1	-0,112	0,017	-0,1453	-0,0787	43,42	<,0001	0,89	6,8%
Age véhicule	3	1	-0,1907	0,0225	-0,2349	-0,1466	71,67	<,0001	0,83	6,3%
Age véhicule	1	0	0	0	0	0 ,		,	1,00	7,6%
Segmentation	1) Van <2100kg	1	0,3746	0,0325	0,3109	0,4383	132,81	<,0001	1,45	11,0%
Segmentation	2) Van >=2100kg	1	0,7528	0,0196	0,7144	0,7912	1476,8	<,0001	2,12	16,1%
Segmentation	3) Car	0	0	0	0	0 ,		,		
NRJ	Elec/Hybrid	1	-0,3139	0,0292	-0,3711	-0,2567	115,63	<,0001	0,73	5,5%
NRJ	Other	0	0	0	0	0 ,		,		
ATECO	1)BANQUE/ASSU	1	0,0201	0,0519	-0,0816	0,1218	0,15	0,6986	1,02	7,7%
ATECO	2)Commercial	1	0,0473	0,0146	0,0186	0,076	10,43	0,0012	1,05	8,0%
ATECO	3)PHARMA/Loc	1	0,1142	0,0245	0,0662	0,1622	21,77	<,0001	1,12	8,5%
ATECO	4)Transport	1	0,482	0,0291	0,4249	0,5391	273,69	<,0001	1,62	12,3%
ATECO	5)Autre	0	0	0	0	0 ,		,		
ZONE	1)	1	0,4049	0,0228	0,3601	0,4496	314,36	<,0001	1,50	11,4%
ZONE	5)	1	-0,1988	0,026	-0,2498	-0,1479	58,47	<,0001	0,82	6,2%
ZONE	6)	1	0,1235	0,0149	0,0943	0,1527	68,75	<,0001	1,13	8,6%
ZONE	2-3-4)	0	0	0	0	0 ,		,		
Taille flotte	GRANDE FLOTTE	1	-0,1038	0,0218	-0,1465	-0,0611	22,71	<,0001	0,90	6,8%
Taille flotte	MOYENNE FLOTTE	1	-0,0478	0,0176	-0,0822	-0,0133	7,39	0,0065	0,95	7,2%
Taille flotte	PETITE FLOTTE	0	0	0	0	0 ,		,		
NBCVF	Hyper Puissant	1	0,1458	0,0156	0,1153	0,1763	87,85	<,0001	1,16	8,8%
NBCVF	Puissant	1	-0,1057	0,0231	-0,1509	-0,0604	20,95	<,0001	0,90	6,8%
NBCVF	Très Puissant	0	0	0	0	0 ,		,		
Marque	BM/Autres	1	0,056	0,0193	0,0182	0,0938	8,45	0,0037	1,06	8,0%
Marque	Ford/Autres	1	0,0109	0,019	-0,0264	0,0481	0,33	0,5671	1,01	7,7%
Marque	Honda/Autres	1	0,2489	0,0301	0,19	0,3078	68,58	<,0001	1,28	9,7%
Marque	Nissan/Autres	1	0,091	0,0244	0,0432	0,1387	13,94	0,0002	1,10	8,3%
Marque	Toyota/Renault/Autres	1	0,1231	0,026	0,0721	0,1741	22,39	<,0001	1,13	8,6%
Marque	Hyundai/Volzwagen/Cheveloret/Opel/Autres	0	0	0	0	0 ,		,		

Tableau 22 – Sortie SAS de la modélisation GLM de la fréquence matérielle du modèle F1

Dans cette sortie :

- La première colonne liste les différentes variables explicatives (Paramètres).
- La deuxième colonne reprend les modalités de chacune des variables explicatives, il s'agit de l'ensemble des classes de risque.
- La quatrième colonne donne l'estimation du paramètre pour chacune des classes.
- La cinquième colonne donne l'erreur standard sur l'estimation du paramètre, les deux colonnes suivantes montrent les intervalles de confiance de Wald à 95%.
- *Pr > khi 2* est la valeur du test du Khi-deux de significativité des paramètres (et donc des modalités).

L'intercept ou la constante caractérise la classe d'échelle. Il désigne la classe de risque ayant la plus grande exposition parmi toutes les combinaisons. Dans notre cas il s'agit des véhicules âgés d'un (1) an de type *car* dont le type d'énergie et le code *ATECO* sont autre, roulant dans les zones 2-3-4), appartenant à la classe des petites flottes, qui sont très puissants en termes de nombre de chevaux fiscaux et dont la marque se trouve dans la liste « Hyundai/Volzwagen/Cheveloret/Opel/Autres ».

L'interprétation de toute autre classe de risque caractérisée par chaque paramètre du modèle doit se faire relativement à cette classe de référence.

Le tableau suivant fournit les informations globales du modèle :

Critères d'évaluation de l'adéquation du modèle F1

Critère	DDL	Valeur	Valeur/DDL
Ecart	340 000	126 105	0,3761
Déviance normalisée	340 000	126 105	0,3761
Khi2 de Pearson	340 000	381 745	1,1385
Pearson normalisé X2	340 000	381 745	1,1385
Log-vraisemblance		-84 280	
Log-vraisemblance complète		-85 795	
AIC		171 639	
AICC		171 639	
BIC		171 896	

Tableau 23 – Critères d'évaluation de l'adéquation du modèle F1

Il montre que l'AIC du modèle est de *171 639* et que le BIC est de *171 896* avec un nombre de degré de liberté à *340 000*.

Dans l'ensemble, le modèle est significatif. Selon les modalités des variables, nous observons les résultats suivants :

6.4.1. « Segmentation » et « Age véhicule »

La variable « âge véhicule » représente exactement l'âge du véhicule depuis sa date d'immatriculation et la variable « Segmentation » est le morcellement selon le type et poids du véhicule. Dans le modèle F1, ces variables profilent plus de poids selon le Khi2 de Wald. De plus nous avons des relativités assez significatives :

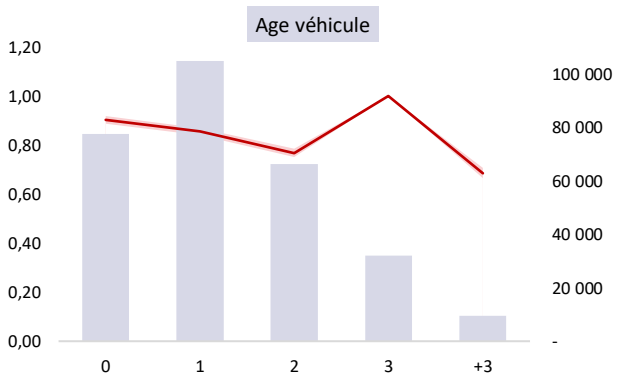


Figure 36 – Exposition et relativité en fonction de l'âge du véhicule

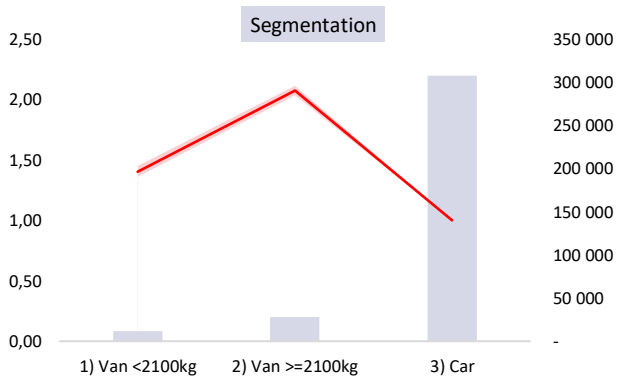


Figure 37 – Exposition et relativité en fonction de la segmentation (car/van)

Nous voyons qu'en termes de fréquence, les véhicule qui ont 3 ans ont une sinistralité assez marquée comparés aux autres. Nous constatons le même résultat pour les voitures commerciales lourdes (van lourds).

6.4.2. Les autres variables

Pour les autres variables, nous obtenons aussi en général assez de significativité suivant leurs modalités :

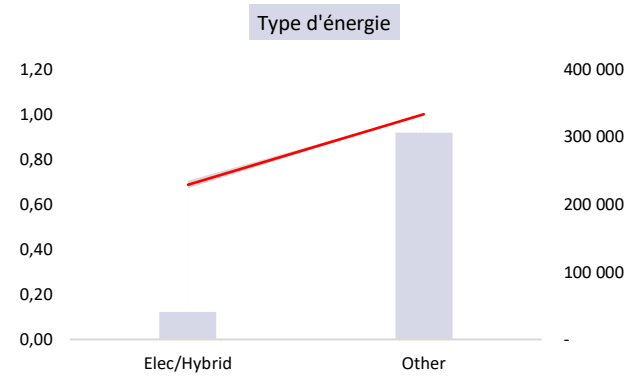


Figure 39 – Exposition et relativité en fonction du type d'énergie

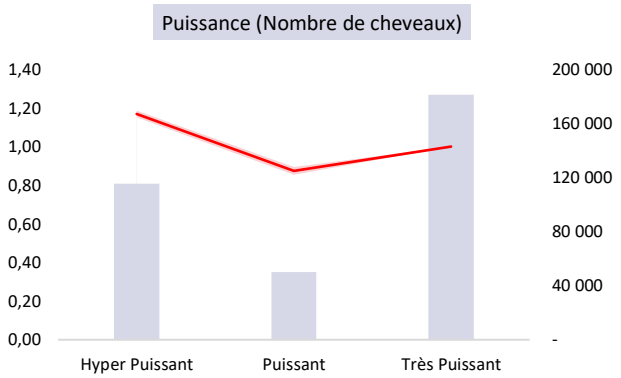


Figure 38 – Exposition et relativité en fonction de la puissance (nombre de chevaux fiscal)

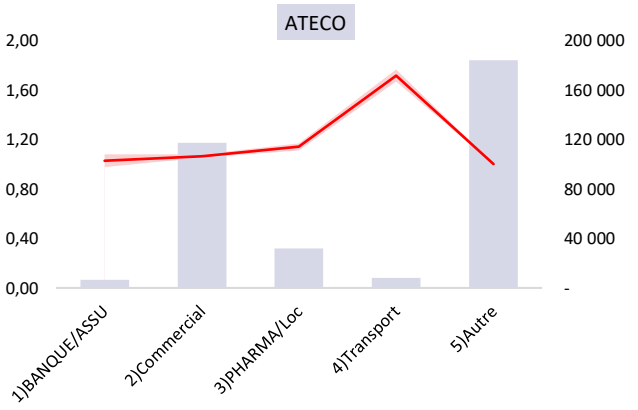


Figure 40 – Exposition et relativité en fonction du secteur d'activité (code ATECO)

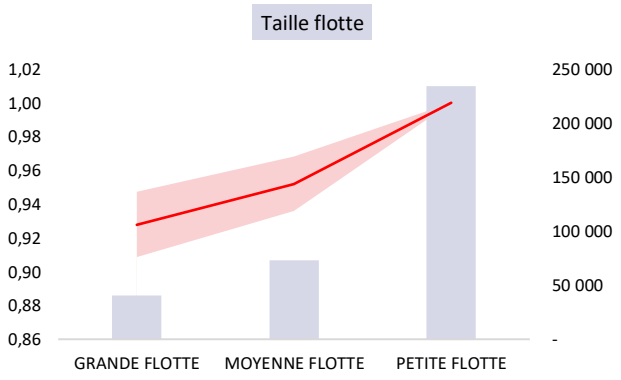


Figure 42 – Exposition et relativité en fonction de la taille de la flotte

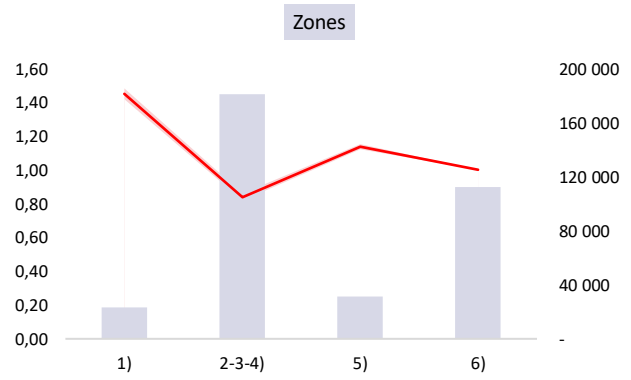


Figure 43 – Exposition et relativité en fonction des zones

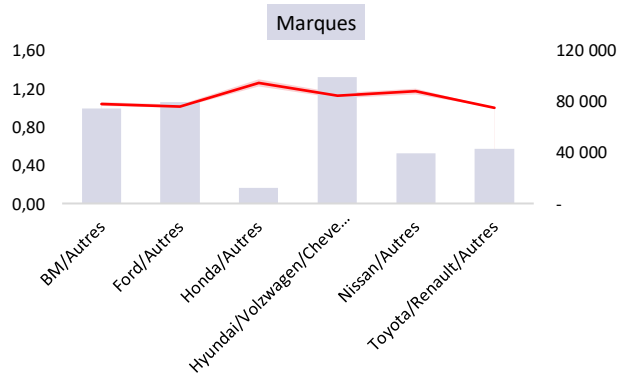


Figure 41 – Exposition et relativité en fonction des marques de véhicules

Pour mesurer l'importance de la variable *COVID* dans le modèle F1, nous décidons de créer la variable en flaguant les années, comme le montre la figure suivante :

COVID → 0	Y6, Y7
COVID → 1	Y8, Y9

Nous ajoutons donc cette variable au modèle F1 pour créer le modèle F1-bis. L'objectif ici est de faire un choix entre les modèles F1 (modèle sans la variable COVID) et F1-bis (modèle avec la variable COVID). Comme il n'existe pas de critère universel permettant de définir la notion de meilleur modèle, nous utiliserons comme critère la capacité de prédiction du modèle.

Conformément aux critères d'évaluation de l'adéquation, nous obtenons donc les résultats suivants pour les modèles *F1* et *F1-bis* :

	Critères d'évaluation de l'adéquation	
F1	AIC	171 639
	BIC	171 896
F1-bis	AIC	171 171
	BIC	171 439

Tableau 24 – Comparaison de critères d'évaluation de l'adéquation des modèles F1 et F1-bis

Selon l'AIC et le BIC, le choix porte sur le modèle F1-bis, ce qui confirme l'hypothèse selon laquelle la variable « COVID » a un impact positif sur le pouvoir explicatif du modèle.

Selon le tableau ci-dessous :

		Modèle F1			Modèle F1-bis		
Paramètres		Khi-2 de Wald	Pr > khi-2	Relativy	Khi-2 de Wald	Pr > khi-2	Relativy
Intercept		19883,8	<,0001	6,9%	17695,4	<,0001	7,6%
Age véhicule	+3 ans	23,89	<,0001	0,83	0,01	0,9142	1,00
Age véhicule	0	34,35	<,0001	0,91	31,46	<,0001	0,91
Age véhicule	2	43,42	<,0001	0,89	13,35	0,0003	0,94
Age véhicule	3	71,67	<,0001	0,83	6,34	0,0118	0,94
Age véhicule	1	,	,	1,00	,	,	1,00
Segmentation	1) Van <2100kg	132,81	<,0001	1,45	103,79	<,0001	1,39
Segmentation	2) Van >=2100kg	1476,8	<,0001	2,12	1348,19	<,0001	2,06
Segmentation	3) Car	,	,		,	,	
NRJ	Elec/Hybrid	115,63	<,0001	0,73	41,46	<,0001	0,83
NRJ	Other	,	,		,	,	
ATECO	1)BANQUE/ASSU	0,15	0,6986	1,02	0,11	0,7435	1,02
ATECO	2)Commercial	10,43	0,0012	1,05	5,66	0,0173	1,04
ATECO	3)PHARMA/Loc	21,77	<,0001	1,12	9,2	0,0024	1,08
ATECO	4)Transport	273,69	<,0001	1,62	213,03	<,0001	1,53
ATECO	5)Autre	,	,		,	,	
ZONE	1)	314,36	<,0001	1,50	310,02	<,0001	1,49
ZONE	5)	58,47	<,0001	0,82	62,24	<,0001	0,81
ZONE	6)	68,75	<,0001	1,13	65,88	<,0001	1,13
ZONE	2-3-4)	,	,		,	,	
Taille flotte	GRANDE FLOTTE	22,71	<,0001	0,90	30,04	<,0001	0,89
Taille flotte	MOYENNE FLOTTE	7,39	0,0065	0,95	5,6	0,018	0,96
Taille flotte	PETITE FLOTTE	,	,		,	,	
NBCVF	Hyper Puissant	87,85	<,0001	1,16	92,46	<,0001	1,16
NBCVF	Puissant	20,95	<,0001	0,90	10,64	0,0011	0,93
NBCVF	Très Puissant	,	,		,	,	
Marque	BM/Autres	8,45	0,0037	1,06	15,36	<,0001	1,08
Marque	Ford/Autres	0,33	0,5671	1,01	4,96	0,0259	1,04
Marque	Honda/Autres	68,58	<,0001	1,28	64,84	<,0001	1,27
Marque	Nissan/Autres	13,94	0,0002	1,10	16,58	<,0001	1,10
Marque	Toyota/Renault/Autres	22,39	<,0001	1,13	17,65	<,0001	1,12
Marque	Hyundai/Volzwagen/Cheveloret/Opel/Autres	,	,		,	,	
COVID	1				461,05	<,0001	0,73
COVID	0				,	,	

Tableau 25 – Comparaison des sorties SAS des modélisations GLM de la fréquence matérielle selon les modèles F1 et F1-bis

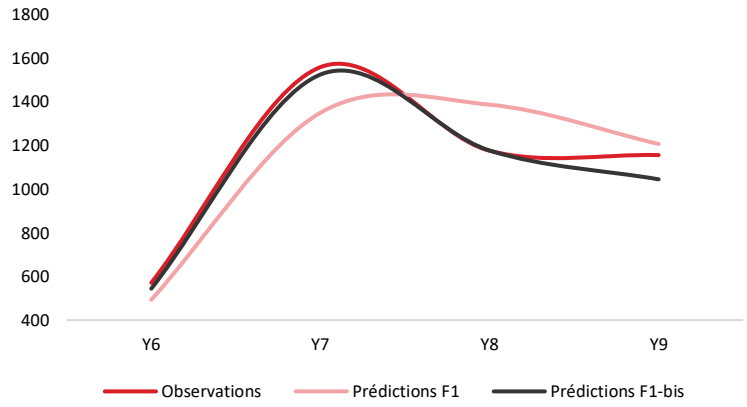
Nous pouvons déceler directement cette baisse de la fréquence engendrée par la COVID. En effet, on visualise l'effet lié à la pandémie : en Y8 et Y9, les véhicules ont 10% moins de probabilité d'avoir un sinistre comparé à Y6 et Y7 (ici la baisse de la fréquence en fonction de la COVD n'est que 10% parce que nous avons combiné les années 2020 et 2021 aussi nous sommes uniquement sur la garantie RC).

De surcroît des indicateurs AIC et BIC, nous mesurons aussi le pouvoir de prédiction de chaque modèle. Pour ce faire, nous allons les appliquer chacun sur la base test afin de relever le nombre de sinistres observés et prédits dans cette base :

		Observations	Prédiction	MSE	
F1	Y6	571	493	0,08300	
	Y7	1559	1350	0,08543	
	Y8	1176	1387	0,06127	MSE Global
	Y9	1156	1207	0,06641	0,07224
F1-bis	Y6	571	544	0,0829	
	Y7	1559	1525	0,0853	
	Y8	1176	1177	0,0611	MSE Global
	Y9	1156	1045	0,0664	0,07212

Tableau 26 – Comparaison des sinistres prédites et observées selon les modèles F1 et F1-bis

En comparant le nombre de sinistres prédits par les modèles, nous voyons que sur les années non-COVID (Y6 et Y7) le modèle F1 prédit moins de sinistres que le modèles F1-bis, et sur les années COVID (Y8 et Y9), il prédit plus de sinistres comme l'illustre le graphique suivant :

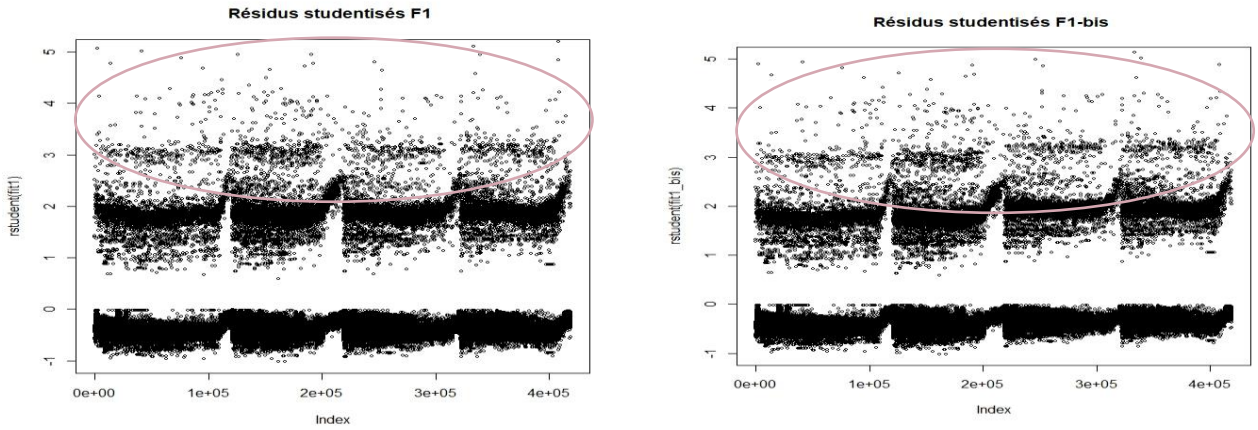


	RMSE apprentissage	*RMSE* test
F1	0,28106	0,26877
F1-bis	0,28087	0,26856

Figure 44 – Comparaison des sinistres prédites et observées selon les modèles F1 etF1-bis

Sur toutes les sinistres observés dans la base test, le modèle F1-bis présente des prédictions très proches des observations contrairement au modèle F1. Aussi, en évaluant la moyenne des erreurs standards (*MSE*) globale nous remarquons que le modèle F1-bis est mieux puisque son MSE est légèrement plus faible.

Ainsi, le modèle F1-bis donne un RMSE plus faible, pour confirmer que c'est le modèle à choisir, appuyons notre hypothèse par une analyse de résidus des deux modèles :



La représentation des résidus studentisés des deux modèles donne presque la même tendance. Néanmoins sur le haut des graphiques (dans les ellipses rouges), nous avons plus de points dans le premier graphique. Cela montre que dans le modèle F1-bis, nous avons des moyennes d'erreur plus faibles, même-si nous sommes en présence de plusieurs points atypiques.

Ainsi, au vu de tous ces résultats nous portons notre choix sur le modèle *F1-bis*.

6.5. Utilisation des informations issue de la télématique afin d'expliquer l'évolution des indicateurs techniques sur les années atypiques :

6.5.1. Corrélation entre les variables télématiques

Pour repérer les éventuelles dépendances entre deux variables quantitatives, on peut se baser sur la corrélation ρ de *Pearson.* Dans le cas où la corrélation est trop importante, il sera nécessaire de retirer certaines variables du modèle à mettre en place afin de ne pas le biaiser.

La corrélation de Pearson permet de déterminer si deux variables quantitatives sont dépendantes. Deux variables sont dites fortement dépendantes l'une de l'autre lorsque la valeur absolue de leur corrélation est proche de 1.

6.5.1.1. Théorie du ρ de pearson :

Soient X et Y deux variables quantitatives prenant respectivement les valeurs $(x_i)_{i=1\dots n}$ et $(y_i)_{i=1\dots n}$.On obtient ρ $(coefficient\ de\ corrélation\ de\ Pearson)$ avec la formule suivante :

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

On calcule l'estimateur $\hat{\rho}_{XY}$ sans biais de ρ_{XY} :

$$\hat{\rho}_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

Avec $\sigma_X\ et\ \sigma_Y$ les corrélations empiriques respectives de X et Y.

$$\hat{\sigma}_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ Et } \hat{\sigma}_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Et $\hat{\sigma}_{XY}$ la covariance empirique de X et Y :

$$\hat{\sigma}_{XY} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$$

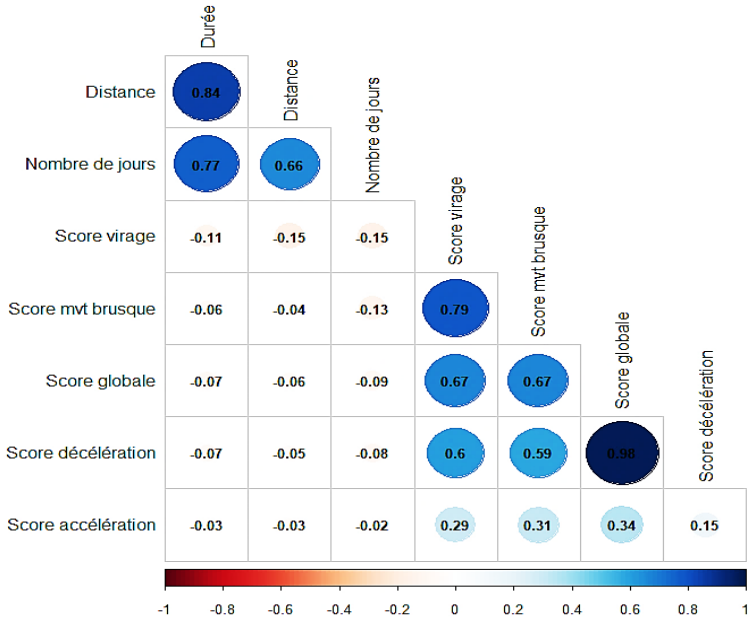


Figure 45 –Graphiques de corrélation sur les variables télématiques

Le graphique ci-dessus montre la corrélation existante entre certaines variables télématiques. Sur ce graphique, le « nombre de jours » désigne le nombre de jours roulés par un véhicule dans l'année. Nous voyons que cette variable est fortement corrélée à la durée et à la distance parcourues par année, qui sont elles-mêmes corrélées entre elles.

Par ailleurs, nous remarquons que les scores par événements (accélération, décélération, virage et mouvement latéral brusque) sont aussi fortement corrélés entre eux. Ainsi, au vu de cette analyse, la construction d'un modèle GLM sur les variables télématiques se basera essentiellement sur la distance et le score global de conduite.

6.5.1.2. Autres analyses de relation entre les variables

Une heatmap, ou carte de chaleur, est une représentation graphique de données. Elle fonctionne sur le principe d'un nuancier de couleurs allant généralement du bleu (plus représenté) au rouge (moins représenté). Ce procédé permet de donner à des données un aspect visuel plus facile à comprendre et à interpréter.

Pour établir une certaine relation entre des variables qualitatives, nous pouvons passer par les cartes de chaleurs. Ces dernières sont intuitives à lire et peuvent donner un aperçu des relations entre des VI (variables indépendantes) et des VD (variables dépendantes) qui sont parfois difficiles à capter dans d'autres méthodes. Ces cartes servent à effectuer des analyses croisées pour certaines variables comme nous le verrons.

6.5.1.2.1. Variable COVID vs taux de roulage urbain (TRU)

COVID	Tranches de TRU [0,20[	[20,30[	[30,40[	[40,50[	[50,70[	[70,100]	Total
0	25%	24%	19%	13%	13%	6%	100%
1	11%	18%	19%	17%	22%	11%	100%

Tableau 27 – Proportion des véhicules selon les taux de roulage urbain avant et après Covid

D'après ce premier tableau réalisé sur l'étude de la corrélation entre les années Covid et les tranches de TRU, nous pouvons voir que sur les années hors COVID, 49% des véhicules ont un TRU inférieur à 30 : ce sont les véhicules qui effectuent moins de 30% de leur distance totale sur l'année sur les routes urbaines. Si nous regardons ceux qui ont un TRU supérieur à 50, c'est-à-dire effectuant plus 50% de leur distance totale les routes urbaines, alors ils représentent 19%.

Sur les années Y8 et Y9, 29% des véhicules ont un TRU inférieur à 30 et 33% ont un TRU supérieur à 50. Cela signifie donc que la proportion des véhicules roulant plus sur les routes urbaines a augmenté et celle des véhicules roulant moins sur les routes urbaines a baissé pendant la crise. En effet, pour leurs déplacements en période de pandémie, beaucoup d'automobilistes privilégient la voiture et fuient la promiscuité des transports en commun :il y'a moins de risque de transmission du virus.

6.5.1.2.2. Variable score global vs TRU

En détaillant l'analyse précédente selon les tranches de score global, nous obtenons les résultats suivants :

		Tranches de TRU						
	COVID =0	[0,20[	[20,30[	[30,40[	[40,50[	[50,70[	[70,100]	Total
Score global	[0,30[	10%	15%	18%	18%	24%	15%	100%
	[30,40[	12%	18%	20%	17%	22%	11%	100%
	[40,50[	16%	22%	21%	16%	18%	7%	100%
	[50,60[	20%	25%	21%	15%	15%	5%	100%
	[60,70[	25%	26%	20%	12%	12%	5%	100%
	[70,80[	29%	25%	18%	12%	11%	5%	100%
	[80,100[	34%	22%	16%	10%	11%	8%	100%

		Tranches de TRU						
	COVID = 1	[0,20[	[20,30[	[30,40[	[40,50[	[50,70[	[70,100]	Total
Score global	[0,30[	4%	6%	11%	16%	33%	30%	100%
	[30,40[	4%	11%	15%	18%	32%	21%	100%
	[40,50[	6%	14%	18%	18%	28%	15%	100%
	[50,60[	8%	17%	20%	19%	25%	11%	100%
	[60,70[	11%	20%	21%	17%	21%	9%	100%
	[70,80[	14%	21%	20%	17%	19%	9%	100%
	[80,100[	18%	19%	18%	15%	18%	12%	100%

Tableau 28 – Proportion des véhicules selon les taux de roulage urbain et le score global annuel avant et après Covid

D'après ces tableaux, sur la totalité des observations, la population la plus représentée est celle qui a un TRU inférieur à 30 pendant les années non-COVID et pour les années COVID, c'est celle qui présente un TRU supérieur à 50.

Nous pouvons aussi relever une corrélation négative qui existerait entre le score de conduite et le TRU : plus le taux de roulage urbain est bas, plus le score global de conduite est élevé (en particulier, on note qu'un TRU inférieur à 20, avec de fortes chances, correspondrait à un score global supérieur à 80 et un TRU supérieur à 70 à un score global inférieur à 30).

Après l'analyses des corrélations entre les variables télématiques, nous pouvons maintenant confronter les modèles F2 et F2-bis.

6.5.2. Comparaison entre F2 et F2-bis

Nous rappelons les paramètres des deux modèles :

Modèles	Descriptions
F2	Prend en compte les variables télématiques (score, distance…) plus quelques variables contrat
F2-bis	Modèle F2 + variable « *Covid* »

Dans ces deux modèles, les variables contrat que nous intégrons sont pour le moment les seules variable prises en compte dans le tarif ALD Italie. Il s'agit de :

- La segmentation (voiture standard, van léger et van lourd)
- Et du type d'énergie (Electrique/autre).

Selon les mêmes dispositifs de comparaison des modèles F1 et F1-bis, nous confrontons le modèle F2 au modèle F2-bis.

Les variables que l'on retrouve dans ces premiers modèles sont :

- Le score global de conduite
- La distance effectuée dans l'année (en milliers de kilomètres)
- La segmentation (voiture standard, van léger et van lourd)
- Le type d'énergie
- Le taux de roulage urbain (TRU)
- La variable COVID (juste dans le modèle F2-bis)

En appliquant les deux modèles sur la bases apprentissage, nous obtenons les résultats suivants :

		Modèle F2			Modèle F2-bis		
Paramètres		Khi-2 de Wald	Pr > khi-2	Relativy	Khi-2 de Wald	Pr > khi-2	Relativy
Intercept		16661,6	<,0001	5,6%	14907,3	<,0001	6,3%
Score global	[0,30[	290,55	<,0001	2,15	257,37	<,0001	2,06
Score global	[30,40[	202,85	<,0001	1,52	172,2	<,0001	1,47
Score global	[40,50[	188,2	<,0001	1,35	164,34	<,0001	1,32
Score global	[50,60[	31,32	<,0001	1,11	25,36	<,0001	1,10
Score global	[70,80[	37,41	<,0001	0,88	30,54	<,0001	0,90
Score global	[80,100[	66,42	<,0001	0,84	68,3	<,0001	0,84
Score global	[60,70[	,	,		,	,	
Distance x 1000km	[0,4[	48,21	<,0001	0,85	70,82	<,0001	0,82
Distance x 1000km	[10,12[	16,17	<,0001	1,11	24,36	<,0001	1,14
Distance x 1000km	[12,14[	35,43	<,0001	1,17	49,67	<,0001	1,21
Distance x 1000km	[14,16[	71,09	<,0001	1,27	92,3	<,0001	1,31
Distance x 1000km	[16,18[	95,71	<,0001	1,33	122,33	<,0001	1,38
Distance x 1000km	[18,20[	96,14	<,0001	1,36	124,95	<,0001	1,42
Distance x 1000km	[20,25[	189,16	<,0001	1,40	243,14	<,0001	1,47
Distance x 1000km	[25,30[	190,5	<,0001	1,48	246,6	<,0001	1,56
Distance x 1000km	[30,35[	196,44	<,0001	1,59	244,35	<,0001	1,67
Distance x 1000km	[35,40[	158,28	<,0001	1,64	199,62	<,0001	1,74
Distance x 1000km	[40,160[	306,94	<,0001	1,74	374,26	<,0001	1,85
Distance x 1000km	[4,10[	,	,		,	,	
Segmentation	1) Van <2100kg	64,78	<,0001	1,29	37,23	<,0001	1,21
Segmentation	2) Van >=2100kg	3212,71	<,0001	2,65	2944,51	<,0001	2,55
Segmentation	3) Car	,	,		,	,	
NRJ	Elec/Hybrid	85,62	<,0001	0,78	24,41	<,0001	0,88
NRJ	Other	,	,		,	,	
TRU	[0,20[	34,39	<,0001	0,88	64,99	<,0001	0,84
TRU	[30,40[	22,92	<,0001	1,10	43,2	<,0001	1,14
TRU	[40,50[	41,68	<,0001	1,15	93,41	<,0001	1,24
TRU	[50,70[	75,96	<,0001	1,21	181,11	<,0001	1,34
TRU	[70,100[	140,21	<,0001	1,39	277,03	<,0001	1,60
TRU	[20,30[	,	,		,	,	
COVID	1				929,29	<,0001	0,65
COVID	0				,	,	

Tableau 29 – Comparaison des sorties SAS des modélisations GLM de la fréquence matérielle selon les modèles F2 et F2-bis

La sortie ci-dessous émet, pour le modèle F2, qu'une voiture standard avec un score global compris entre 60 et 70, effectuant une distance entre 4000km et 10000km à l'année, dont le type d'énergie est qualifié de « autre » et avec un score urbain compris entre 20 et 30, présente une fréquence de sinistralité de 5,6%. Cependant, pour le modèle F2-bis, cette même classe de référence, laisse paraître une fréquence de sinistralité de 6,3% sur les années non-Covid (Y6 et Y7) et de 4,1% (0,65x6,3%) sur les années Covid (Y8 et Y9).

Ainsi, nous décelons de cette analyse un apport et impact importants de la variable « COVID ». Cet impact se traduit par une baisse capitale de la fréquence de sinistralité due à la baisse du nombre global de kilomètres parcourus depuis le début de la crise (exemple en France le nombre global de kilomètres parcourus a connu une variation -17% en 2020 par rapport à 2019 selon le SDES 2021 -Service des Données et Etudes Statistiques-). En effet, les confinements et déconfinements ont fortement influencé les comportements des déplacements des usagers.

En termes d'adéquation, le modèle F2-bis fournit AIC et un BIC plus faibles que le modèle F2.

	Critères d'évaluation de l'adéquation	
F2	AIC	170 744
	BIC	170 022
F2-bis	AIC	169 794
	BIC	169 084

Tableau 30 – Comparaison des AIC et BIC des modèles F1 et F1-bis

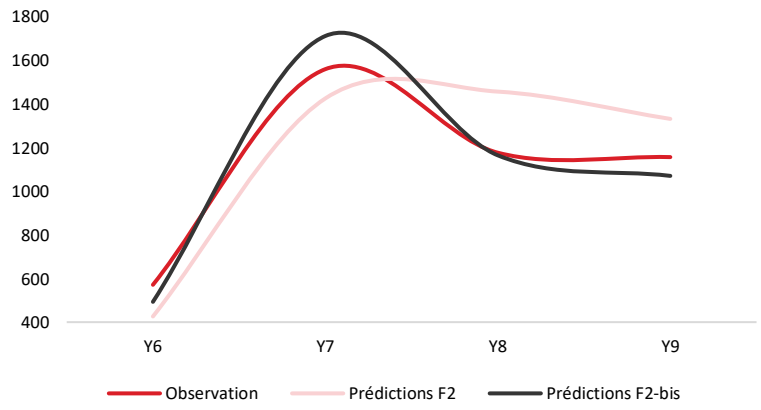
Aussi nous retrouvons cette même corrélation qui a été décelée entre les variables score global et TRU d'après les analyses de relation entre les variables réalisées plus haut ; nous constatons que les véhicules standards avec un score global supérieur à 80 présentent le même coefficient multiplicateur que ceux avec un TRU inférieur à 20 et ce coefficient est évalué à 0,85 (ce qui signifie qu'une voiture avec un score global de conduite supérieur 80 ou un TRU inférieur à 20 comporte une probabilité de 15% moins d'avoir un sinistre).

Les prédictions sur la base test des deux modèles fournissent les résultats suivants :

		Observations	Prédiction	MSE	
F2	Y6	571	426	0,0834	
	Y7	1559	1425	0,0849	
	Y8	1176	1456	0,0612	MSE Global
	Y9	1156	1331	0,0661	0,072009
F2-bis	Y6	571	493	0,0829	
	Y7	1559	1711	0,0848	
	Y8	1176	1164	0,0609	MSE Global
	Y9	1156	1070	0,0659	0,071785

Tableau 31 – Comparaison des sinistres prédites et observés selon les modèles F2 et F2-bis

Comme dans l'analyse faite pour les modèles de contrats, nous voyons que le modèle F2 prédit moins de sinistres que le modèles F2-bis sur les années non-COVID (Y6 et Y7), et plus de sinistres sur les années COVID (Y8 et Y9) comme le montre le graphique suivant :



	RMSE apprentissage	*RMSE* test
F2	0,28110	0,26834
F2-bis	0,28055	0,26793

Figure 46 – Comparaison des sinistres prédites et observés selon les modèles F2 et F2-bis

Nous voyons ainsi que les sinistres prédits par modèle F2-bis se rapprochent particulièrement aux sinistres théoriques. Nous retenons non seulement l'effet du COVID dans ce modèle mais aussi et surtout la contribution des variables télématiques.

En effet, segmenter selon les variables télématiques représente une hypothèse plausible étant donné le fort pouvoir explicatif de ces dernières sur la sinistralité matérielle. Comme nous venons de le voir, en estimant les nombres des sinistres dans la base test selon le modèle télématique (F2-bis), nous obtenons une différence négligeable entre la sinistralité réelle et la sinistralité prédite par le modèle.

Pour démontrer l'apport télématique dans la modélisation GLM de la fréquence matérielle, nous allons opposer les modèles F1-bis (modèle contrat) et F2-bis (modèle télématique). Pour ce faire, nous créons depuis la base initiale, 1000 échantillons aléatoires égales (dans chaque échantillon, nous avons 5% des observations de la base initiale stratifiées selon l'année et la segmentation type de véhicule), ensuite, nous déterminons la fréquence estimée par les modèles ainsi que la fréquence observée pour chaque observations dans les 1000 échantillons. Par la suite, nous déterminons la moyenne des fréquences estimées par chaque modèle dans chaque échantillon. Nous constituons dès lors une base de 1000 observations et pour chaque observation, sont connues les fréquences de sinistralité selon les deux modèles ainsi que la fréquence observée.

L'objectif étant de comparer pour la nouvelle base (de 1000 observations) la distributions des différences de probabilités de sinistralités entre le modèle contrat (F1-bis) et l'observé et entre le modèle télématique (F2-bis) et l'observé. Ce procédé permettra de tester non seulement la stabilité des modèles, mais aussi d'analyser de plus près leur différence.

Nous obtenons alors la figure ci-dessous :

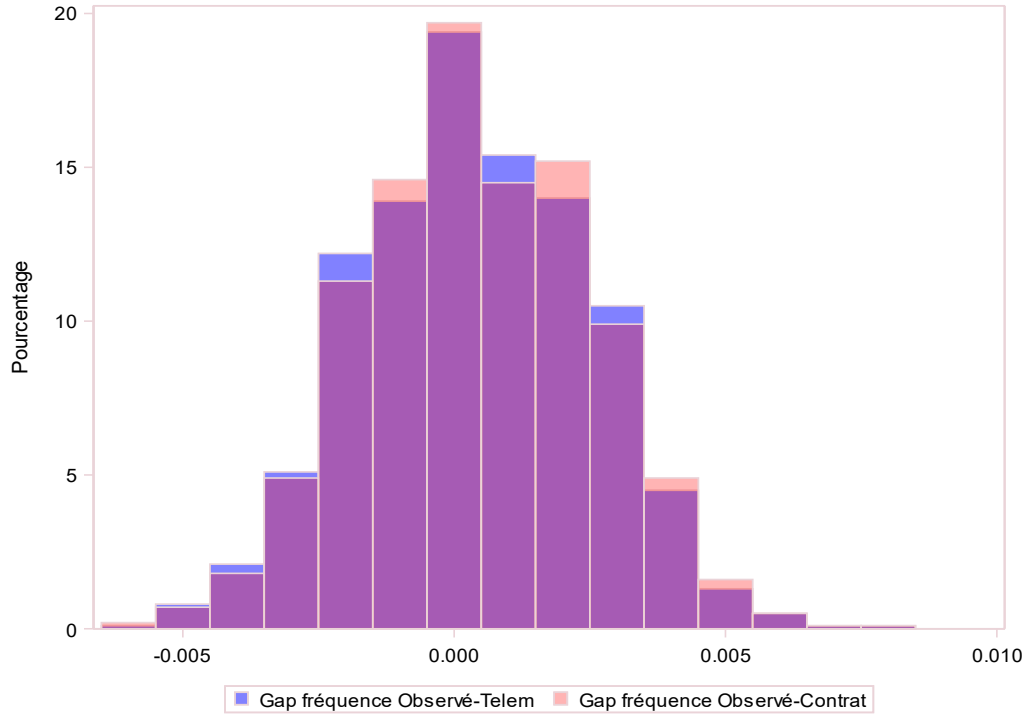


Figure 47 – Histogrammes des différences sur les probabilités de sinistralité entre l'observé et les modèles F1-bis et F2-bis

Lorsque nous comparons les deux distributions, nous remarquons que les faibles estimations de probabilités de sinistres sont plus présentes avec le modèle télématique qu'avec le modèle contrat. Aussi comme vu plus haut, nous avons une estimation du nombre total de sinistres présents dans la base test par le modèle F2-bis à 4 438 contre 4 462 réellement observés. Cette même estimation avec le modèle F1-bis donne un nombre de sinistres égale à 4 297.

Par conséquent, nous avons un modèle télématique qui révèle une diminution de la sinistralité observée par rapport à celle attendue mais aussi qui ajuste bien cette sinistralité (rapprochement avec les données observées). Il s'avère dès lors bénéfique de considérer ce modèle d'autant plus qu'il contient une information cruciale, la seule d'ailleurs, sur le conducteur. Cette information n'est autre que le score global de conduite qui reflète parfaitement son comportement au volant.

Il serait donc plus judicieux de porter notre choix sur le modèle F2-bis.

La connaissance unique de la fréquence de sinistralité ne nous permet pas de calculer une prime pure. La détermination de cette dernière nécessite la connaissance du coût moyen. En effet cette prime s'obtient en multipliant les coefficients des modèles de fréquence et de coût. La prime pure représente le montant moyen que l'assureur est prêt à accepter pour garantir le risque.

Chapitre 7 : MODELISATION GLM DU COÛT MOYEN MATERIEL

Nous modéliserons le coût moyen avec les mêmes variables présentes dans les deux modèles de fréquence retenus. Pour ce faire, examinons d'abord la significativité globale de ces variables dans le modèle de sévérité selon le critère de l'AIC :

Synthèse des sélections

Etape	Effet saisi	AIC	p-value
0	Intercept	492 567	.
1	Distance	492 263	<,0001
2	ATECO	492 107	<,0001
3	Taille flotte	491 955	<,0001
4	Score global	491 862	<,0001
5	Age véhicule	491 804	<,0001
6	TRU	491 747	<,0001
7	COVID	491 685	<,0001
8	ZONE	491 650	<,0001
9	Marque	491 617	<,0001
10	Segmentation	491 600	<,0001
11	NRJ	491 593	0,0027
12	NBCVF	491 591*	0,0421

*Valeur optimal de l'AIC

Le tableau ci-dessous résume la significativité des variables si nous les ajoutons à la modélisation du coût moyen. Nous voyons ainsi que toutes les variables pourrons être intégrées dans les modèle même si nous observons une p-value plus grande pour les variables NRJ (type d'énergie) et NBCVF (nombre de chevaux fiscaux).

Nous étudierons donc le coût moyen matériel sur la base de deux modèles : un modèle C1 avec les variables contrat et un modèle C2 avec les variables télématiques. La modélisation GLM du coût moyen se fait en utilisant la loi Gamma avec une fonction lien logarithme. Pour modéliser ce coût, nous travaillons uniquement sur la base sinistres auquel nous ajoutons les informations télématiques et contrats.

7.1. Coût moyen selon le modèle contrat C1

Les variables que l'on retrouve dans ce premier modèle sont :

- L'âge du véhicule
- La segmentation (voiture standard, fourgonnette légère, lourde)
- Le type d'énergie
- Le code ATECO (secteur d'activité)
- La zone
- La taille de la flotte
- Le nombre de chevaux fiscaux du véhicule (NBCVF)
- La marque du véhicule.

- La variable COVID

La sorite SAS du modèle fournit les résultats suivants :

Analyse des paramètres estimés du maximum de vraisemblance du modèle C1									
Paramètre		DDL	Estimation	Erreur alle	de confiance de Wald		Khi-2 de Wald	Pr > khi-2	Realtivy
Intercept		1	7,6137	0,016	7,5823	7,645	226496	<,0001	2 026 €
Segmentation	1) Van <2100kg	1	-0,0775	0,0271	-0,1307	-0,0243	8,15	0,0043	0,93
Segmentation	2) Van >=2100kg	1	0,0185	0,0178	-0,0164	0,0535	1,08	0,299	1,02
Segmentation	3) Car	0	0	0	0	0	,	,	1,00
NRJ	Elec/Hybrid	1	-0,0723	0,024	-0,1193	-0,0253	9,07	0,0026	0,93
NRJ	Other	0	0	0	0	0	,	,	1,00
Taille flotte	GRANDE FLOTTE	1	-0,1663	0,018	-0,2016	-0,131	85,31	<,0001	0,85
Taille flotte	MOYENNE FLOTTE	1	-0,141	0,0146	-0,1695	-0,1124	93,36	<,0001	0,87
Taille flotte	PETITE FLOTTE	0	0	0	0	0	,	,	1,00
ZONE	1)	1	0,096	0,0201	0,0566	0,1354	22,78	<,0001	1,10
ZONE	5)	1	0,0535	0,0212	0,012	0,095	6,37	0,0116	1,05
ZONE	6)	1	0,0227	0,0123	-0,0013	0,0468	3,43	0,0641	1,02
ZONE	2-3-4)	0	0	0	0	0	,	,	1,00
ATECO	1)BANQUE/ASSU	1	-0,0484	0,0426	-0,1318	0,035	1,29	0,2557	0,95
ATECO	2)Commercial	1	0,1047	0,012	0,0812	0,1282	76,18	<,0001	1,11
ATECO	3)PHARMA/Loc	1	0,2081	0,0201	0,1688	0,2474	107,64	<,0001	1,23
ATECO	4)Transport	1	0,0929	0,0286	0,0369	0,1489	10,57	0,0012	1,10
ATECO	5)Autre	0	0	0	0	0	,	,	1,00
Marque	BM/Autres	1	-0,0656	0,0161	-0,0971	-0,034	16,56	<,0001	0,94
Marque	Ford/Autres	1	-0,0011	0,0156	-0,0316	0,0294	0	0,9448	1,00
Marque	Honda/Autres	1	0,0364	0,0281	-0,0186	0,0914	1,69	0,1941	1,04
Marque	Nissan/Autres	1	-0,0801	0,0201	-0,1195	-0,0407	15,88	<,0001	0,92
Marque	Toyota/Renault/Autres	1	-0,0599	0,0212	-0,1015	-0,0183	7,98	0,0047	0,94
Marque	Hyundai/Volzwagen/Cheveloret/Opel/Autres	0	0	0	0	0	,	,	1,00
NBCVF	Hyper Puissant	1	-0,0206	0,0132	-0,0465	0,0054	2,41	0,1205	0,98
NBCVF	Puissant	1	0,0701	0,019	0,0328	0,1074	13,55	0,0002	1,07
NBCVF	Très Puissant	0	0	0	0	0	,	,	1,00
Age véhicule	3	1	0,2024	0,0319	0,1398	0,2649	40,23	<,0001	1,22
Age véhicule	0	1	0,049	0,014	0,0216	0,0763	12,3	0,0005	1,05
Age véhicule	2	1	0,0274	0,0143	-0,0006	0,0555	3,68	0,0552	1,03
Age véhicule	3	1	0,0654	0,0191	0,0279	0,1029	11,68	0,0006	1,07
Age véhicule	1	0	0	0	0	0	,	,	1,00
COVID	1	1	0,0674	0,0119	0,0441	0,0907	32,13	<,0001	1,07
COVID	0	0	0	0	0	0	,	,	1

Figure 48 – Analyse des paramètres estimés du maximum de vraisemblance du modèle C1

Ici, d'après la classe référente, le coût moyen est estimé à 2 026€. Comme évoqué en introduction, la portée de la pandémie sur le coût moyen n'a pas été négligeable. En effet, Il a connu une variation positive comparée aux années précédant la crise. Cette augmentation du coût moyen est vérifiée par le modèle C1 selon lequel le coût moyen a augmenté de 7% comparé à la classe référente sur les années COVID.

Suivant ce modèle, en considérant la segmentation selon le type de véhicule, nous avons une sévérité qui est plus élevée pour les van lourds que pour les voitures standards et selon l'âge du véhicule, plus il augmente, plus la sévérité croît. Cela s'explique par la hausse du prix des réparations automobiles. Les coûts augmentent d'une année à l'autre et c'est le prix des pièces détachées qui est principalement concerné par cette hausse. Nous pouvons prendre l'exemple d'un rétroviseur, si son coût tournait autour de 190€ il y a 10 ans, il pourrait coûter aujourd'hui entre 500 et 600€.

7.2. Coût moyen selon le modèle télématique C2

Les variables que l'on retrouve dans ce deuxième modèle sont :

- Le score global de conduite annuel
- La distance de conduite annuelle en millier de km
- La segmentation (voiture standard, fourgonnette légère, lourde)
- Le type d'énergie
- Le taux de roulage urbain (TRU)
- La variable COVID

La sorite SAS du modèle fournit les résultats suivants :

Analyse des paramètres estimés du maximum de vraisemblance du modèle C2									
Paramètre		DDL	Estimation	Erreur	confiance de Wald à95%		Khi-2 de Wal	Pr > khi-2	Relativy
Intercept		1	7,7138	0,0178	7,6789	7,7486	188470	<,0001	2 239 €
Score global	[0,30[	1	0,1942	0,0378	0,1201	0,2684	26,38	<,0001	1,21
Score global	[30,40[	1	0,1107	0,0243	0,063	0,1584	20,69	<,0001	1,12
Score global	[40,50[	1	0,0852	0,0182	0,0495	0,1209	21,9	<,0001	1,09
Score global	[50,60[	1	0,0344	0,016	0,003	0,0659	4,61	0,0318	1,03
Score global	[70,80[	1	-0,0584	0,0165	-0,0908	-0,026	12,49	0,0004	0,94
Score global	[80,100[	1	-0,099	0,0181	-0,1345	-0,0636	29,94	<,0001	0,91
Score global	[60,70[	0	0	0	0	0	,	,	1,00
Distance x 1000km	[0,4[	1	0,2161	0,0192	0,1784	0,2537	126,71	<,0001	1,24
Distance x 1000km	[12,20[	1	-0,0873	0,0142	-0,1152	-0,0595	37,68	<,0001	0,92
Distance x 1000km	[20,25[	1	-0,1456	0,0195	-0,184	-0,1073	55,52	<,0001	0,86
Distance x 1000km	[25,40[	1	-0,1879	0,0174	-0,222	-0,1538	116,65	<,0001	0,83
Distance x 1000km	[40,160[	1	-0,2104	0,0256	-0,2606	-0,1603	67,65	<,0001	0,81
Distance x 1000km	[4,12[	0	0	0	0	0	,	,	1,00
Segmentation	) Van <2100k	1	-0,0075	0,0266	-0,0597	0,0447	0,08	0,7772	0,99
Segmentation	Van >=2100l	1	0,1123	0,0155	0,0818	0,1427	52,28	<,0001	1,12
Segmentation	3) Car	0	0	0	0	0	,	,	1,00
NRJ	Elec/Hybrid	1	-0,0919	0,0214	-0,1338	-0,0499	18,43	<,0001	0,91
NRJ	Other	0	0	0	0	0	,	,	1,00
TRU	[0,20[	1	0,0173	0,0178	-0,0175	0,052	0,94	0,3311	1,02
TRU	[30,40[	1	-0,0485	0,0169	-0,0816	-0,0155	8,27	0,004	0,95
TRU	[40,50[	1	-0,0642	0,0183	-0,1	-0,0285	12,39	0,0004	0,94
TRU	[50,70[	1	-0,0805	0,0181	-0,116	-0,0451	19,83	<,0001	0,92
TRU	[70,100[	1	-0,1539	0,023	-0,1989	-0,1088	44,83	<,0001	0,86
TRU	[20,30[	0	0	0	0	0	,	,	1,00
COVID	1	1	0,1089	0,0116	0,0861	0,1316	88,02	<,0001	1,12
COVID	0	0	0	0	0	0	,	,	1,00

Figure 49 – Analyse des paramètres estimés du maximum de vraisemblance du modèle C1

Suivant le modèle C2, sur une année non-COVID, une voiture standard de type d'énergie « autre » ayant un score de conduite compris entre 60 et 70 et un score urbain compris entre 20 et 30, effectuant entre 4 000 et 12 000 km à l'année porte un coût moyen à 2 239€. Conformément à ce modèle, le coût moyen diminue quand le score de conduite, la distance et le TRU (taux de roulage urbain) augmentent.

Les graphiques ci-dessous montrent l'évolution de la sévérité en fonction de la distance, du score et du TRU :

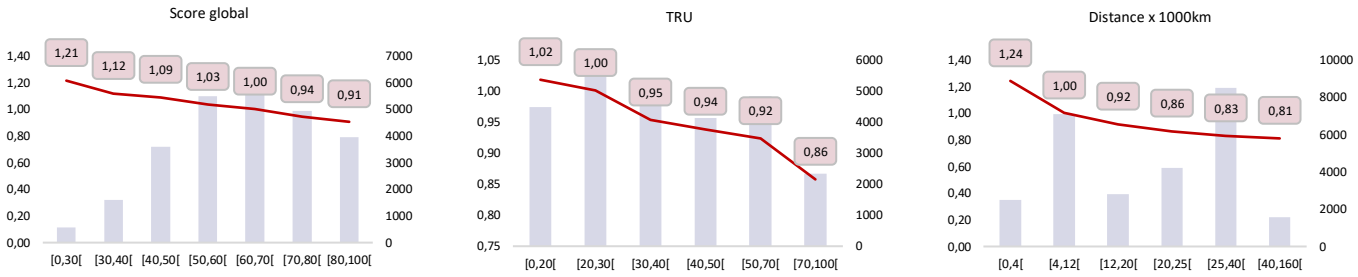


Figure 50 – Coefficients des variables télématiques (score global, TRU et distance) dans la modélisation du coût moyen.

Au vu de la distance effectuée à l'année et du taux de roulage urbain, la baisse du coût moyen sur certaines tranches pourrait s'expliquer par la présence de pleins de petits sinistres (sinistres à coût bas) qui absorbent les coûts des gros sinistres.

7.3. Comparaison des modèles C1 et C2

En mettant à la balance les deux modèles, suivant le critère de l'AIC et du BIC, le choix porte sur le modèle télématique encore une fois.

	C1	***C2***
AIC	492 454	492 395
BIC	492 665	492 577

De plus si nous traçons les histogrammes de la distribution des différences entre coût moyen observé et coût moyen estimé par les modèles, nous relevons que le modèle télématique appliqué au portefeuille estime plus de coûts élevés que le modèle contrat. Le modèle C2 est donc un modèle qui surestime la charge. Par conséquent il peut être considéré comme un modèle prudent, ce qui est un atout en assurance, même si l'objectif principal est de rester au plus près du coût réel des sinistres.

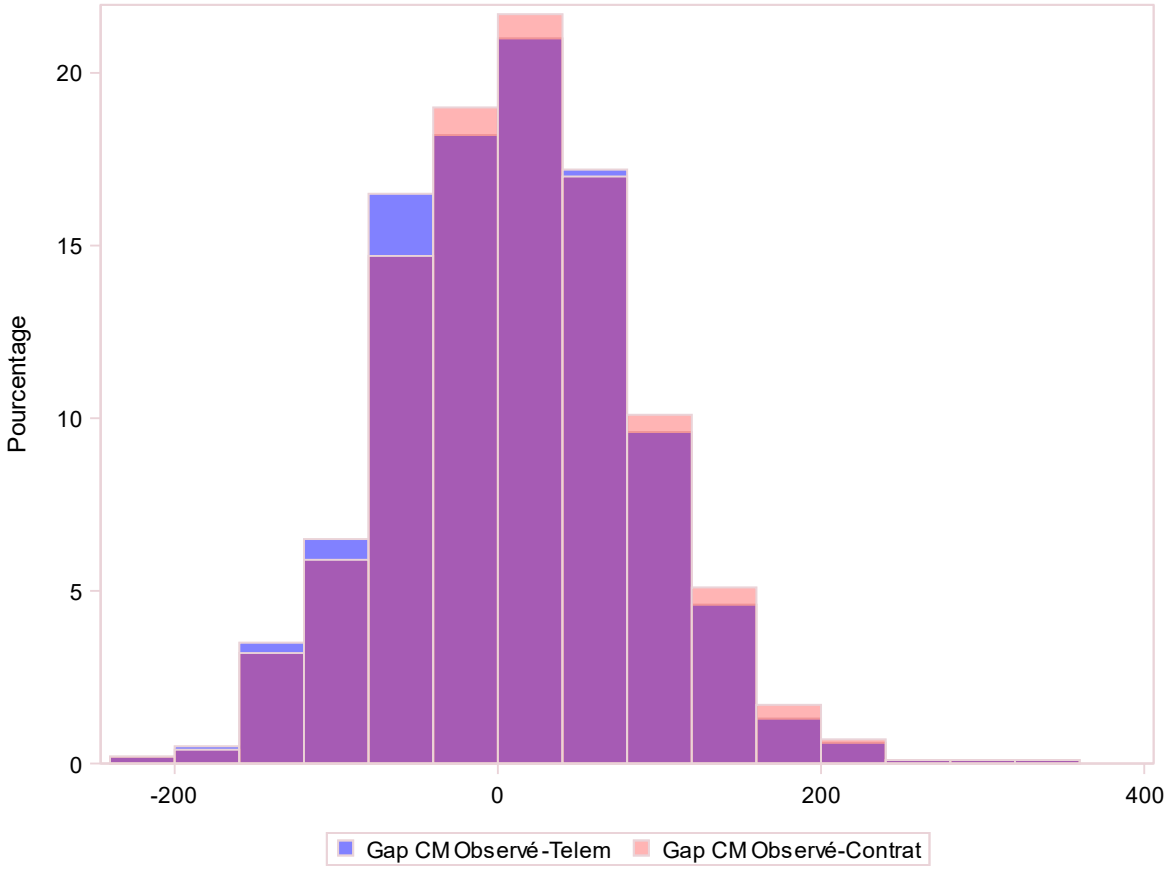


Figure 51 – Histogrammes de la distributions des différences de coûts moyens entre l'observée et les modèles C1 et C2

Compte tenu de ces résultats, nous portons notre choix porte sur le modèle télématique (C2).

Etant donné que nous avons modéliser et la fréquence de sinistralité et le coût moyen, nous pouvons maintenant passer à la détermination de la prime de risque.

Chapitre 8 : DETERMINATION DE LA PRIME PURE

8.1. La prime pure F1-bis x C1

Paramètres		Coefficients PP	Prime Pure
Constante		153,7 €	
Age véhicule	0	0,96	147 €
Age véhicule	1	1,00	154 €
Age véhicule	2	0,97	148 €
Age véhicule	3	1,01	155 €
Age véhicule	+3	1,22	187 €
Segmentation	3) Car	1,00	154 €
Segmentation	1) Van <2100kg	1,29	198 €
Segmentation	2) Van >=2100kg	2,10	322 €
NRJ	Elec/Hybrid	0,77	118 €
NRJ	Other	1,00	154 €
ATECO	1)BANQUE/ASSU	0,97	149 €
ATECO	2)Commercial	1,15	177 €
ATECO	3)PHARMA/Loc	1,33	204 €
ATECO	4)Transport	1,68	259 €
ATECO	5)Autre	1,00	154 €
ZONE	1)	1,65	253 €
ZONE	5)	0,86	132 €
ZONE	6)	1,15	177 €
ZONE	2-3-4)	1,00	154 €
Taille flotte	GRANDE FLOTTE	0,75	116 €
Taille flotte	MOYENNE FLOTTE	0,83	128 €
Taille flotte	PETITE FLOTTE	1,00	154 €
NBCVF	Hyper Puissant	1,14	175 €
NBCVF	Puissant	0,99	153 €
NBCVF	Très Puissant	1,00	154 €
Marque	BM/Autres	1,01	155 €
Marque	Ford/Autres	1,04	160 €
Marque	Honda/Autres	1,32	203 €
Marque	Nissan/Autres	1,02	157 €
Marque	Toyota/Renault/Autres	1,05	162 €
Marque	Hyundai/Volzwagen/Cheveloret/Opel/Autres	1,00	154 €
COVID	1	0,78	120 €
COVID	0	1,00	154 €

Tableau 32 – Coefficient de la prime pure obtenue avec le modèle de fréquence contra (F1-bis) et le modèle de coût moyen contrat (C1)

Si nous multiplions les coefficients obtenus dans le modèle de fréquence contrat par ceux obtenus dans le modèle de coût moyen, nous obtenons une prime de risque à 153,7€ pour un véhicule avec les modalités de référence. Sur le modèle contrat, nous voyons que la prime pure est passée de 154€ à 120€ pendant les années COVID. Ce modèle laisse paraître donc une baisse de la prime pure d'environ 23%.

8.2. La prime pure F2-bis x C2

Conformément au modèle télématique sur la fréquence et sur la sévérité, la prime pure obtenu pour l'intercept est de 141,6€. Nous voyons que sur toutes les variables du modèle sauf sur le score de conduite, la prime pure agit selon une corrélation positive (plus les modalités de ces variables augmentent, plus la prime augmente). En effet pour le score global de conduite : plus on est dangereux (mauvais conducteur) plus la prime de risque est élevée.

Figure 52 – Evolution de la prime pure suivant le modèle télématique en fonction du score de conduite, de la distance, du TRU et de la segmentation par type de véhicule

Au moment de l'établissement du tarif, certains coefficients pour certaines variables seront lissés car ces variables accusent parfois des effets non-linéaires qui peuvent être malvenus puisqu'il s'agit de variables ordonnées. C'est le cas ici de la variable « distance ».

Si nous nous intéressons maintenant à la valeur de la prime pure suivant les années et suivant chaque modèle, nous remarquons que la prime pure obtenue sur le modèle télématique est presque identique à la prime pure observée. Le modèle contrat quant à lui estime bien aussi la prime pure. En même temps, nous remarquons la baisse de la prime de risque pendant les années COVID :

	Contrat	Télématique	Observée
Y6	215 €	216 €	215 €
Y7	200 €	203 €	203 €
Y8	150 €	148 €	148 €
Y9	150 €	152 €	156 €

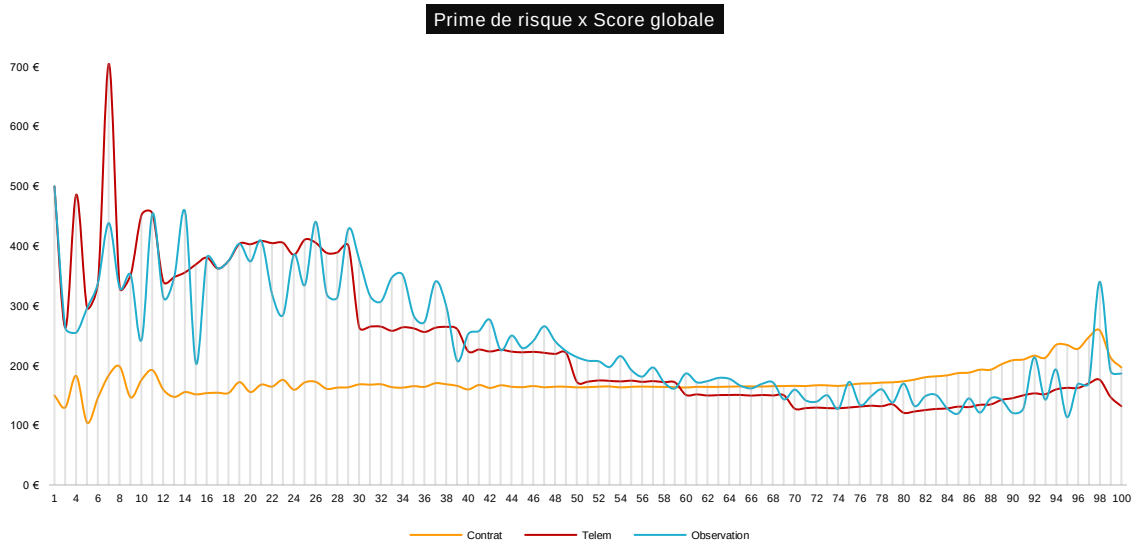
Tableau 33 – Prime de risque par année selon les modèles contrat et télématique et selon l'observée

Bonne conduite → baisse de prime :

Dans ce paragraphe nous nous intéressons à l'évolution de la prime pure en fonction du score de conduite selon les deux modèles. Ici nous mettons en phase les trois primes de risque :

- La prime pure selon le modèle contrat,
- La prime pure selon le modèle télématique
- Et la prime pure observée (coût observé/exposition) ajustée un peu sur les scores entre 0 et 30

Nous obtenons le graphique suivant :



Les tendances générales que nous observons selon ce graphique sont :

- Sur modèle télématique ; plus le score augmente, plus la prime baisse. Nous relevons cette même tendance sur la courbe de la prime réelle
- Sur le modèle contrat, nous relevons une tendance assez constante autour de 170€

Il apparaît dès lors une très bonne segmentation du portefeuille avec le modèle télématique pour le score de conduite.

8.3. Conclusion de la partie

Au vu des études réalisées dans cette partie nous avons pu capter l'impact de la crise de la COVID-19 dans la prime pure en flaguant les années mais aussi l'apport des variables télématiques dans celle-ci. Nous avons mis en place deux modèles pour la fréquence et deux modèles pour le coût moyen. Selon ces modèles, ceux de la télématique ont été retenus au vu du pouvoir de prédiction.

Dans le modèle de fréquence matérielle télématique, toutes les variables sont significatives et le score de conduite permet de capter une partie de l'information sur le conducteur dont nous ne disposons pas dans la base de données (absence d'informations conducteurs dans la base). Le nombre de sinistres que prédit ce modèle paraît très proche du nombre de sinistres réellement observé.

Dans le modèle de coût moyen matériel télématique, certaines modalités de certaines variables n'étaient pas significatives. Aussi, le score de conduite contribue dans l'apport d'informations pour ce modèle.

A ce stade, nous avons donc démontré l'importance des variables télématiques mais aussi relevé l'effet de la pandémie. Nous résumons les résultats obtenus dans le tableau suivant :

Paramètres		Coefficients CM	Coefficients Freq	Coefficients PP	Prime Pure
Constante		2 239 €	6,3%	141,6 €	
Score global	[0,30[	1,21	2,06	2,50	354 €
Score global	[30,40[	1,12	1,47	1,65	233 €
Score global	[40,50[	1,09	1,32	1,44	204 €
Score global	[50,60[	1,03	1,10	1,14	162 €
Score global	[60,70[	1,00	1,00	1,00	142 €
Score global	[70,80[	0,94	0,90	0,84	120 €
Score global	[80,100[	0,91	0,84	0,76	107 €
Distance x 1000km	[0,4[	1,24	0,82	1,01	144 €
Distance x 1000km	[4,10[	1,00	1,00	1,00	142 €
Distance x 1000km	[10,12[	1,00	1,14	1,14	161 €
Distance x 1000km	[12,14[	0,92	1,21	1,11	157 €
Distance x 1000km	[14,16[	0,92	1,31	1,20	170 €
Distance x 1000km	[16,18[	0,92	1,38	1,26	179 €
Distance x 1000km	[18,20[	0,92	1,42	1,30	184 €
Distance x 1000km	[20,25[	0,86	1,47	1,27	179 €
Distance x 1000km	[25,30[	0,83	1,56	1,29	183 €
Distance x 1000km	[30,35[	0,83	1,67	1,39	196 €
Distance x 1000km	[35,40[	0,83	1,74	1,44	204 €
Distance x 1000km	[40,160[	0,81	1,85	1,50	212 €
Segmentation	3) Car	1,00	1,00	1,00	142 €
Segmentation	1) Van <2100kg	0,99	1,21	1,21	171 €
Segmentation	2) Van >=2100kg	1,12	2,55	2,86	404 €
NRJ	Elec/Hybrid	0,91	0,88	0,80	113 €
NRJ	Other	1,00	1,00	1,00	142 €
TRU	[0,20[	1,02	0,84	0,86	121 €
TRU	[20,30[	1,00	1,00	1,00	142 €
TRU	[30,40[	0,95	1,14	1,09	154 €
TRU	[40,50[	0,94	1,24	1,16	164 €
TRU	[50,70[	0,92	1,34	1,24	175 €
TRU	[70,100[	0,86	1,60	1,37	194 €
COVID	1	1,12	0,65	0,73	103 €
COVID	0	1,00	1,00	1,00	142 €

Chapitre 9 : PROJECTIONS ET CORRELATIONS AVEC LES EFFETS MACRO-ECONOMIQUES

Jusqu'ici, nous avons entraîné et testé les différents modèles construits sur les années Y6 à Y9. Dans cette partie, nous allons essayer dans un premier temps d'indiquer la prime de risque sur l'année Y10 (6 premiers mois : de novembre 2021 à avril 2022) selon qu'elle soit une année COVID ou non-COVID. Et dans un deuxième temps, nous analyserons, de façon générale, les effets macro-économiques sur l'évolution de la prime pure.

9.1. Projections sur l'année Y10

Pour réaliser nos projections sur l'année 10, nous allons tester deux scénarios : d'abord nous regardons ce que donnerait la prime si on considère Y10 comme une année COVID puis on regarde ce que cela donnerait dans le cas contraire. Nous obtenons dès lors les chiffres suivants :

		Contrat	Télématique	Observée
COVID=0	Car	174 €	169 €	127 €
	Van <2100kg	231 €	217 €	150 €
	Van >=2100kg	439 €	431 €	346 €
	PP Globale	191 €	185 €	140 €

		Contrat	Télématique	Observée
COVID=1	Car	136 €	123 €	127 €
	Van <2100kg	181 €	158 €	150 €
	Van >=2100kg	344 €	314 €	346 €
	PP Globale	150 €	135 €	140 €

Selon la prime pure observée et les primes pures obtenues avec les modèles contrat et télématique,, nous pouvons voir que l'année Y10 ressemble plus à une année COVID qu'à une année non-COVID. En effet, si l'année Y10 était considérée comme une année non-COVID, cela signifierait que les deux modèles n'estiment pas bien la prime pure puisqu'ils prédisent des primes de risques très élevées comparées à l'observé. Cela étant dit, la pandémie se poursuit toujours sur l'année Y10 (novembre 2021 à octobre 2022) ; nous pouvons ainsi dire que l'hypothèse prônant que l'année Y10 serait une année COVID est justifiable.

Même si nous ne sommes plus à l'époque des confinements et des couvre-feux, le télétravail se poursuit toujours, favorisant ainsi la réduction des distances effectuées par les conducteurs sur les routes. Aussi, la hausse du carburant observée ces derniers temps, aura certainement un impact sur les primes d'assurance automobile.

9.2. Corrélations avec les effets macro-économiques

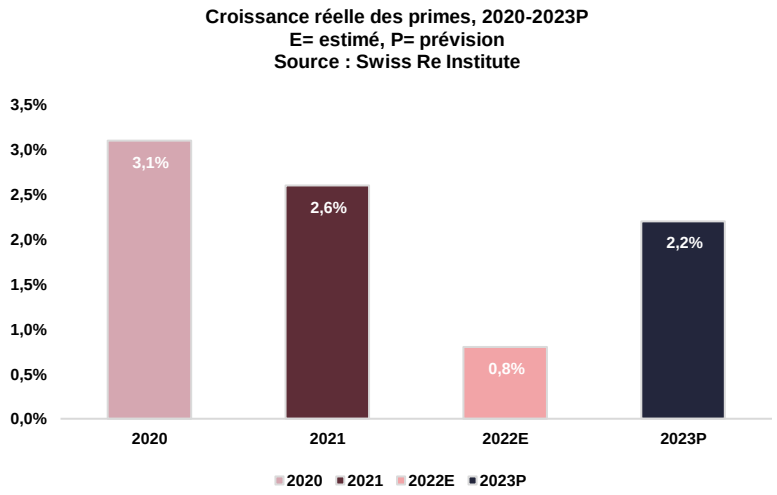
Avec la guerre en Ukraine, il est facile de perdre de vue le fait que la récession mondiale la plus sévère a frappé il y a à peine deux ans, lorsque la COVID-19 s'est emparé du monde. Les blessures économiques de la pandémie, qui mettront des années à cicatriser, avaient déjà réduit les perspectives de croissance durable avant la guerre. L'invasion russe en cours a aggravé un environnement de croissance déjà ralenti. Avec autant de défis, un changement structurel est nécessaire pour renforcer la résilience.

9.2.1. COVID et inflation ou… stagflation

La pandémie a contribué à créer une inflation structurellement plus élevée au cours de la prochaine décennie ; l'économie mondiale ralentit fortement et l'inflation atteint des sommets. Les banques centrales augmentent les taux d'intérêt, ciblant la stabilité des prix plutôt que la croissance économique. Selon le rapport « sigma 4/2022 sur l'assurance mondiale », cela contribuera à éviter la stagflation à la manière des années 1970. En effet, les craintes de stagflation (combinaison d'une inflation élevée depuis plusieurs décennies et d'un ralentissement de la croissance économique) refont surface.

Pour les assureurs, l'impact principal de l'inflation se traduira par la hausse des coûts des sinistres, et cela, davantage en assurance non-vie. En effet, les hausses de prix les pièces automobiles dépassent celles de l'économie au sens large. Des vents contraires à la rentabilité, mais aussi des vents favorables en raison d'un nouveau durcissement des taux dans le secteur de l'IARD seront à envisager, les secteurs d'activité automobile et responsabilité civile seront sans doute les plus touchés.

Une forte croissance nominale de 6,1 % des primes totales (non-vie et vie) en 2022 est envisageable. En termes réels, cela se traduit par une croissance quasi stable (+0,4 %).



En non-vie, l'inflation des valeurs d'exposition et le durcissement des taux pourrait stimuler la croissance des primes mondiales.

L'impact négatif se fera aussi sentir à travers la hausse des prix de l'énergie, en particulier en Europe.

9.2.2. Prix du carburant et taux de congestion

Selon l'Insee, les prix à la consommation du gaz, des carburants et dans une moindre mesure de l'électricité ont fortement augmenté entre décembre 2020 et octobre 2021, de l'ordre de 41 %, 21 % et 3 % respectivement. Le prix des carburants se sont rapprochés des niveaux atteints à l'automne 2018 et ceux du gaz les ont dépassés. La hausse des prix depuis le début de l'année a conduit en octobre 2021 à un surcroît des dépenses mensuelles d'énergie d'un peu plus de 40 € en moyenne par ménage, dont 20 € pour les carburants. Néanmoins, fin 2020, les prix des carburants étaient relativement bas, du fait de la crise sanitaire, mais les retombées de la guerre en Ukraine en décrivent une tout autre trajectoire.

Le prix du carburant semble être un facteur contributif à la congestion du trafic. En ce sens, la hausse des prix du carburant engendre une baisse de la sinistralité, et cela, pour deux principales raisons. D'abord, pour économiser de l'essence, les conducteurs ont tendance à « lever le pied ». Ensuite, toujours par souci d'économie, certains conducteurs se verront contraints de reporter certains trajets. Or, qui dit baisse de la vitesse et réduction des déplacements, dit chute du nombre d'accidents et de sinistres. Ce même phénomène a été observé à l'occasion des confinements successifs : en 2020, engendrant une baisse de la mortalité sur les routes de plus de 20%.

La hausse des prix du carburant a aussi un effet sur le taux de congestion. La congestion provoque une consommation supplémentaire sur la quantité de carburant mais aussi une augmentation de la durée d'un trajet. Une voiture circulant dans un trafic embouteillé aura tendance à consommer plus de carburant qu'une voiture circulant à des vitesses constantes et raisonnables. Dans le but d'éviter des factures d'essence très élevées, certains automobilistes favorisent d'autres moyens de transport (transports en commun, deux roues, trottinettes…), ou même le télétravail. Cela aura pour impact une baisse du taux de congestion et donc moins de bouchons sur les routes.

Néanmoins la voiture reste majoritaire, dans l'ensemble des déplacements et peu importe la distance parcourue. Selon l'Insee (sur une analyse de la mobilité avant et pendant la pandémie), dans l'ensemble, 74% des trajets domicile-travail sont réalisés en voiture, quand seulement 16% le sont en transport en commun et 8,5% en mode doux (marche à pied et vélo).

Pour caractériser plus finement l'usage de la voiture, il est intéressant de regarder l'évolution des parts modales en fonction de l'éloignement du lieu de travail. Il faut noter que deux tiers des actifs effectuent plus de 5 km pour se rendre à leur travail. Et sans surprise, plus les actifs vivent loin de leur lieu de travail, plus ils plébiscitent la voiture, avec par exemple 80% des trajets ayant une distance supérieure à 5 km qui sont effectués en voiture.

Si la voiture règne sur les déplacements longs, on pourra s'étonner qu'elle reste prédominante sur les petites distances : 60% des déplacements de moins de 5 km sont effectués en voiture, à comparer avec 23% pour les modes doux et 15% en transport en commun, le reste des déplacements étant faits en deux-roues. Nous pouvons dès lors nous intéresser à comment évolue la prime pure suivant les distances effectuées à l'année.

Le graphique qui suit montre l'évolution de la prime pure suivant des tranches de distance (en milliers de kilomètres) effectuée à l'année :

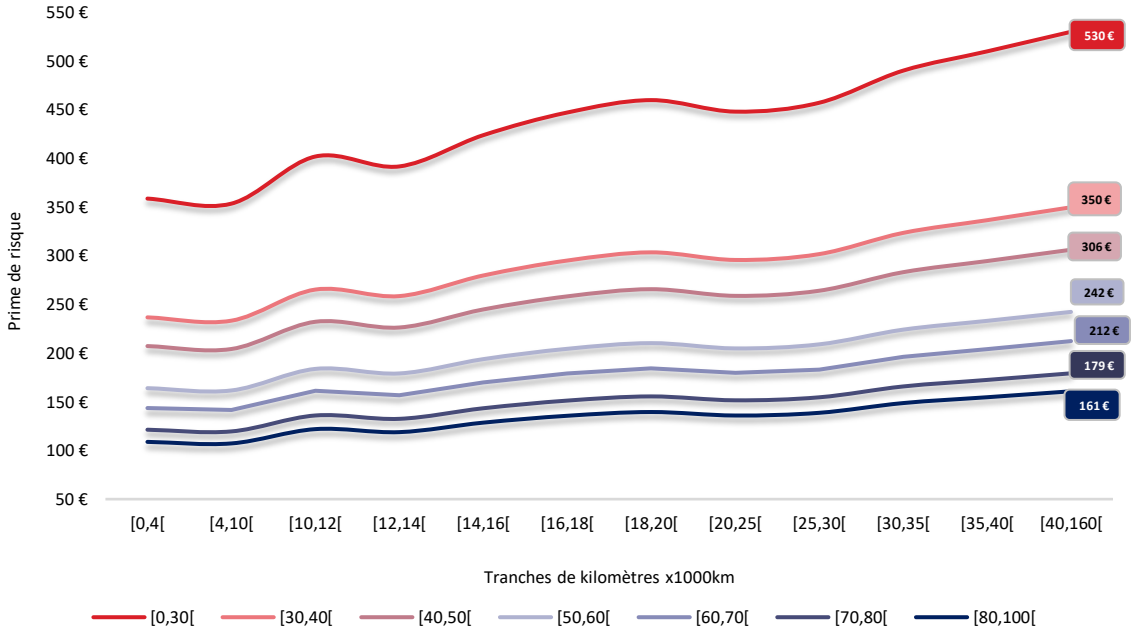


Figure 53 – Prime de risque selon la distance réalisée à l'année et le score global de conduite annuel

Sur ce graphique nous voyons que la prime de risque augmente quand la distance effectuée à l'année augmente et quand le score annuel de conduite baisse. Par exemple pour un véhicule effectuant plus de 40 000km à l'année se verrait avec une prime de risque à 530€ si son score était inférieur à 30. Pour ce même véhicule, sa prime de risque baisserait jusqu'à 161€ si son score annuel de conduite était supérieur à 80. Cela représente donc un gap significatif selon le score.

9.2.3. Profitabilité post-COVID ?

Les assureurs automobiles ont bien performé au début de la pandémie de COVID-19, car les confinements ont entraîné moins de voitures sur les routes et donc moins de sinistres. Mais la forte inflation de cette année a entraîné une forte augmentation du coût des réparations automobiles. Le marché de l'assurance automobile a connu des niveaux significatifs d'inflation des sinistres au premier semestre, principalement en raison de la hausse des prix des voitures d'occasion, de l'allongement des délais de réparation et de l'inflation du coût des pièces de rechange.

Dans cette foulée, deux scénarios peuvent être projetés :

→*Dans le meilleur des cas* : le scénario de croissance économique négative pour l'industrie. Il peut être défini comme étant un déclin économique prolongé et des comportements de conduite plus conservateurs. Les complications du ralentissement économique entraîneraient une longue récession. Après la crise (si toute fois la pandémie prend fin), de nouvelles normes « comportementales » entraîneraient moins de voyages et contribueraient à une vision plus prudente de la vie. L'impact net serait une continuation de la tendance favorable de la fréquence des sinistres après le pic à la baisse du premier semestre 2020. Le coût moyen des sinistres se modérerait conformément aux comportements plus conservateurs.

→*Dans le pire des cas* : le scénario de contraction économique et de changements de comportement. Le résultat serait de nouvelles normes comportementales qui compromettent les capacités de maintien de l'ordre et d'application du prix de la pandémie. La tendance favorable de la fréquence des sinistres depuis des décennies s'inverserait après la pointe à la baisse du premier semestre 2020. Le coût moyen des sinistres poursuivrait sa trajectoire ascendante.

Dans les deux cas : la pression réglementaire pourrait faire baisser davantage les tarifs ou forcer l'extension de la couverture, aggravant ainsi la dégradation des performances du ratio combiné. Les assureurs peuvent faire face à une pression sociale, en plus de la pression réglementaire, pour rembourser ou réduire les primes pendant cette période, phénomène déjà rencontré au début de la crise sanitaire.

Parmi les scénarios ci-dessus, le second est, bien sûr, moins probable, mais, pour l'analyse des risques, il est essentiel de les anticiper tous. Il faut ainsi envisager des mesures, notamment en augmentant les prix et en déployant une nouvelle capacité de tarification.

9.3. Analyse de perspectives

La crise renforcera l'élan déjà en cours pour les automobilistes, les agents et les consommateurs qui se tournent vers les canaux numériques. L'automobile personnelle a été en première ligne de la numérisation pour le secteur de l'assurance, mais la pandémie souligne la nécessité de mettre davantage l'accent sur les canaux numériques.

9.3.1. Adoption accrue de la télématique

Avec la perspective de pandémies récurrentes à l'avenir, la télématique répond directement au besoin des consommateurs de payer des primes moins élevées lorsque les véhicules sont moins utilisés pendant les périodes de confinement. Les consommateurs rechercheront des produits qui « s'endorment » pendant les périodes de non-utilisation, ce qui maintiendrait la conformité aux réglementations d'assurance de l'État tout en permettant une protection d'assurance qui reflète de manière plus appropriée une utilisation réduite. La télématique éliminerait les conjectures et l'ambiguïté, permettant une tarification basée sur l'utilisation qui est plus équitable pour le consommateur et plus précise pour la compagnie d'assurance.

9.3.2. Adoption du télétravail

Selon des études, le mode de travail à distance s'est avéré efficace. Adopter ce mode de travail irait dans un sens positif pour l'assurance auto puisque cela diminuerait l'utilisation de la voiture pour les trajets domicile-travail impactant positivement la fréquence des accidents. Cette crise est un signal d'alarme pour repenser le modèle d'exploitation et construire une infrastructure plus efficace et résiliente. Le voyage vers le modèle de l'état futur pourrait s'accélérer, étant donné la reconnaissance que les épidémies ultérieures à la COVID-19 ou d'autres pandémies pourraient avoir un impact similaire et soudain à l'avenir.

9.3.3. Changement des habitudes des conducteurs

La pandémie pourrait changer l'attitude des consommateurs à l'égard des ressources partagées. Les individus peuvent moins compter sur les transports en commun et le co-voiturage pour éviter le risque d'infection par les modes de transport fortement utilisés. Le déploiement des véhicules autonomes pourrait ralentir, compte tenu de leur interdépendance avec les flottes sans

conducteur. À l'inverse, les perspectives des flottes sans conducteur pourraient devenir plus attrayantes, notamment en tant qu'alternative aux transports en commun.

CONCLUSION

L'intérêt des consommateurs pour la télématique n'a jamais été aussi élevé et ne cesse de croître. La crise de la COVID a fait découvrir les avantages de la télématique à de nombreux conducteurs. L'appétit pour les modèles d'assurance connectés a fortement augmenté depuis le début de la pandémie.

L'idée originale derrière le traitement des données télématiques dans le secteur de l'assurance est de suivre l'utilisation d'un véhicule pour tarifer l'assurance en fonction du comportement de conduite de l'assuré. Aujourd'hui, il permet aux compagnies d'assurance de récompenser ceux qui optent pour une conduite prudente ou des programmes basés sur la distance, les comportements à risque, le score...

Dans ce mémoire, nous avons essayé de modéliser la fréquence et le coût moyen des sinistres matériels pour la garantie RC suivant deux approches. Nous sommes partis d'un modèle contenant uniquement des variables contrats, puis à ce modèle, nous avons ajouté la variable « COVID » afin de relever l'effet global de la pandémie sur la sinistralité et la sévérité. Par la suite, dans l'optique de relever l'efficacité de l'intégration des variables télématiques dans la modélisation, nous nous sommes penchés sur l'étude des modèles prenant en compte ces variables.

Sur la modélisation de la fréquence matérielle, les modèles contrat et télématique ajustent bien la sinistralité (rapprochement des sinistralités observée et estimée). Néanmoins Il serait bénéfique de considérer le modèle télématique puisqu'il contient une information capitale, la seule d'ailleurs (dans le cas de nos flottes), sur le conducteur qu'est le score global de conduite. Cette variable reflète parfaitement le comportement conducteur au volant. Quant à la modélisation du coût moyen matériel, nous avons vu que le modèle télématique surestime la charge et que par conséquent il peut être considéré comme un modèle prudent.

Ainsi, il est possible d'imaginer diverses manières d'intégrer les variables télématiques dans la tarification afin de construire une politique tarifaire adaptée à l'offre souhaitée. Comme nous l'avons vu, nous pouvons envisager une hausse du tarif pour les conducteurs avec des scores bas et une baisse pour les bons scores. Cependant la question n'est pas de pénaliser les mauvais conducteurs mais d'appliquer des abattements tarifaires uniquement aux conducteurs ayant obtenu de bons scores.

La technologie télématique permet d'accéder à une toute nouvelles catégories d'informations sur le comportement des assurés. Ces données comportementales sont très utiles dans le cadre des flottes, pour comprendre le profil de risque d'un titulaire de police. La télématique permet dès lors de mettre en place de nouveaux modèles commerciaux plus robustes en mettant en évidence les paramètres les plus remarquables telles que le score de conduite, la distance de conduite ou le taux de roulage sur les routes urbaines.

Dans le dernier chapitre, nous avons mis l'accent sur les la corrélation du tarif (de la prime) avec les effets macro-économiques. Compte tenu des cicatrices laissées par la COVID-19 et des retombées de la guerre en Ukraine, une croissance à long terme serait difficile à envisager, même pour l'assurance connectée. Il faut donc envisager des mesures, notamment en augmentant les prix et en déployant de nouvelles capacités de tarification pour rétablir les

marges. Il est clair que l'inflation des sinistres s'accélère à un rythme considérable. Mais, même avec la hausse des prix, il faut s’attendre à ce que les marges se détériorent considérablement.

Table des figures

Liste des tableaux

Références

[1] M. DENUIT et A. CHARPENTIER (2005). Mathématiques de l'assurance non-vie, tome II Tarification et Provisionnement.

[2] A. CHARPENTIER (2010). Statistique de l'assurance. URL : https://cel.archives-ouvertes.fr/cel-00550583/document (pages 5-60).

[3] A. CHARPENTIER (2011). Segmentation et mutualisation, les deux faces d'une même pièce.

[4] GALEA (2019) Tarification automobile à l'aide de modèles de machine learning et apport des données télématiques.

[5] F. CASANOVA AIZPUN et al. (2022). Global insurance premium volumes to reach new high in 2022. URL : https://www.swissre.com/institute/research/sigma-research/sigma-2022-04.html

[6] T. CAILLOL et C. GIRAUD (2017). Bancassureurs : les nouveaux champions de l'assurance ? URL : https://www.insurancespeaker-wavestone.com/2017/01/bancassureurs-champions-assurance/

[7] E. DURAND (2016). Les bancassureurs sortent du lot en termes de croissance et de rentabilité (Facts & Figures). URL : https://www.argusdelassurance.com/acteurs/les-bancassureurs-sortent-du-lot-en-termes-de-croissance-et-de-rentabilite-facts-figures.110658

[8] Rapport 2020–2021ANIA* https://www.ania.it/documents/35135/0/Italian+Insurance-2021+WEB2.pdf/78fd5027-9f19-b262-679b-be1c84dec1ab?version=1.0&t=1642695880855

[9] ASTIN (2021) Pandemic effects on Italian insurance market

[10] Rapport FFA (2020) étude sur le marché des flottes automobiles en 2020

[11] Magazine Flottes Automobiles (2022), assurance des flottes 2022 : la hausse encore et toujours, https://www.flotauto.com/assurance-flottes-2022-hausse-20220614.html

[12] C. CHOW. Utilisation des données télématiques pour l'analyse de la sinistralité automobile. URL : https://www.institutdesactuaires.com/sedocumenter/memoire-d-actuariat-38?id=bb61369ed7b8629dd91a707f2422a157

[13] T. DESLANDES. Construction d'une échelle Bonus-Malus à l'aide du score de conduite [Mémoire d'actuariat consulté en interne].
URL : https://www.institutdesactuaires.com/se-documenter/memoires-d-actuariat-38?id=36f79bfebfc104b134798e0147766f8c

[14] SAS SUPPORT. Sas Entreprise Guide https://support.sas.com/en/software/enterprise-guide-support.html#documentation

ANNEXE

Annexe A : MODELES LINEAIRES GENERALISES (GLM)

Démonstration de la relation de la moyenne et de la variance pour une distribution appartenant à une famille exponentielle :

1. Soit Y une variable de distribution appartenant à une famille exponentielle. La fonction moment de Yest définie par :

$$M_Y(t) = E(e^{tY}) = \int e^{ty+\frac{y\theta-b(\theta)}{a(\theta)}-c(y,\theta)}dy$$

$$= e^{-\frac{b\theta}{a(\theta)}}\int e^{ty+\frac{y\theta}{a(\theta)}-c(y,\theta)}dy$$

$$= \frac{e^{-\frac{b\theta}{a(\varphi)}}}{e^{-\frac{b(a(\varphi)t+\theta)}{a(\varphi)}}}\int e^{\frac{a(\varphi)ty+\theta y}{a(\varphi)}-c(y,\theta)-\frac{b(a(\varphi)t+\theta)}{a(\varphi)}}dy$$

$$= \frac{e^{-\frac{b\theta}{a(\varphi)}}}{e^{-\frac{b(a(\varphi)t+\theta)}{a(\varphi)}}}\int e^{\frac{y(a(\varphi)t+\theta)-b(a(\varphi)t+\theta)}{a(\varphi)}-c(y,\theta)}dy$$

$$= \frac{e^{-\frac{b\theta}{a(\varphi)}}}{e^{-\frac{b(a(\varphi)t+\theta)}{a(\varphi)}}}$$

$$= e^{-\frac{b(\theta)+b(a(\varphi)t+\theta)}{a(\varphi)}}$$

Ainsi la moyenne de Y peut être obtenu comme suit : $E(Y) = M'_Y(t)|_{t=0}$ où M' est la dérivée de Mpar rapport à t.

$$E(Y) = \left(e^{-\frac{b(\theta)+b(a(\varphi)t+\theta)}{a(\varphi)}}\right)'|_{t=0}$$

$$= e^{-\frac{b(\theta)+b(a(\varphi)t+\theta)}{a(\varphi)}}\left(\frac{b\prime(a(\varphi)t+\theta)a(\varphi)}{a(\varphi)}\right)|_{t=0}$$

$$= e^{\frac{-b(\theta)+b(\theta)}{a(\varphi)}}b'(\theta)$$

$$= b'(\theta)$$

2. On détermine la variance de Y par :

$$Var(Y) = (M"_Y(t) - M'_Y(t)^2)|_{t=0}$$

D'après ce qui précède, nous avons $M'_Y(t)^2|_{t=0} = \left(e^{-\frac{b(\theta)+b(a(\varphi)t+\theta)}{a(\varphi)}}\right)^{'2}|_{t=0} = (b'(\theta))^2$

$$M"_Y(t) = \left(e^{-\frac{b(\theta)+b(a(\varphi)t+\theta)}{a(\varphi)}}\right)" \Big|_{t=0}$$

$$= e^{-\frac{b(\theta)+b(a(\varphi)t+\theta)}{a(\varphi)}} \left\{\left(b'(a(\varphi)t+\theta)\right)^2 + b"(a(\varphi)t+\theta)a(\varphi)\right\}\Big|_{t=0}$$

$$= e^{\frac{-b(\theta)+b(\theta)}{a(\varphi)}} \left\{\left(b'(\theta)\right)^2 + b"(\theta)a(\varphi)\right\}$$

$$= b'(\theta)^2 + b"(\theta)a(\varphi)$$

Ainsi nous obtenons la variance de Y:

$$Var(Y) = b'(\theta)^2 + b"(\theta)a(\varphi) - b'(\theta)^2$$

$$Var(Y) = b"(\theta)a(\varphi)$$

Dans le logiciel SAS Enterprise Guide, nous avons utilisé la procédure GENMOD pour tourner nos modèles GLM. La procédure GENMOD ajuste les modèles linéaires généralisés, tels que définis par Nelder et Wedderburn (1972). La classe des modèles linéaires généralisés est une extension des modèles linéaires traditionnels qui permet à la moyenne d'une population de dépendre d'un prédicteur linéaire via une fonction de lien non linéaire et permet à la distribution de probabilité de réponse d'être n'importe quel membre d'une famille exponentielle de distributions. De nombreux modèles statistiques largement utilisés sont des modèles linéaires généralisés. Ceux-ci incluent des modèles linéaires classiques avec des erreurs normales, des modèles logistiques et probit pour les données binaires et des modèles log-linéaires pour les données multinomiales. De nombreux autres modèles statistiques utiles peuvent être formulés sous forme de modèles linéaires généralisés en sélectionnant une fonction de lien et une distribution de probabilité de réponse appropriées.

Pour illustrer cela, nous pouvons considérer que nous voulons modéliser la fréquence pour une base donnée. Soient donc Y le nombre de sinistres, Xi les variables explicatives, Mij les modalité j de chaque variable Xi et Mk, k<j la modalité référence de chaque Xi. On appelle modalité de référence, la modalité la plus représentée i.e. la modalité avec plus de données. La procédure se présente donc comme suit :

```
proc genmod data=Base;
            class X1(ref="M12") X2(ref="M24")…;
            model Y = X1 X2 …
          / dist   = poisson  link   = log offset = v1;
          output    out = Base_bis
                    pred = Pred
                    resraw = Resraw
                    reschi = Reschi
                    resdev = Resdev
                    stdreschi = Stdreschi
                    stdresdev = Stdresdev
                    reslik = Reslik ;
       store out = work.Modele_freq;
run;
```

Ici l'offset est la variable d'écart. En générale c'est le logarithme de l'exposition. L'output permet d'obtenir une base comportant les valeurs prédites et les valeurs des différents types de résidus. Le store out permet de pouvoir stocker notre modèle et de l'utiliser après sur une base test par exemple :

```
Proc PLM restore=Modele_freq;
score data=Base_test
out = Base_test / ilink ;
run;
```

C'est la procédure PLM qui permet de calculer les valeurs prédites d'un modèle sur une autre base.

Annexe B : PRINCIPE DU GRADIENT BOOSTING (XGBOOST)

1. Le boosting

Dans l'apprentissage automatique, le boosting est un méta-algorithme d'ensemble pour réduire principalement les biais, ainsi que la variance dans l'apprentissage supervisé, et une famille d'algorithmes d'apprentissage automatique qui convertissent les apprenants faibles en apprenants forts.

Le boosting est une méthode utilisée pour réduire les erreurs dans l'analyse prédictive de données. Les données utilisées sont des données étiquetées pour entraîner des modèles de machine learning à faire des prédictions sur des données non étiquetées. Un modèle de machine learning unique peut commettre des erreurs de prédiction selon la précision du jeu de données d'entraînement. Le boosting s'efforce de résoudre ce problème en entraînant successivement plusieurs modèles afin d'améliorer la précision du modèle global.

C'est une méthode d'agrégation développée par *Freund* et *Schapire* (1996) qui repose sur des stratégies adaptatives (adaboost pour adaptative boosting). Il cherche à optimiser l'affectation des poids en fonction des prévisions. Il crée ainsi des classifieurs faibles h_i de façon à obtenir le classifieur H tel que : $H = signe(\sum \alpha_t h_t)$.

le boosting construit des arbres en série, c'est-à-dire que chaque arbre généré (sauf le premier) a accès à son prédécesseur, ou plus précisément à l'erreur de son prédécesseur. Le nouvel arbre construit aura pour but de se concentrer sur les lacunes de son prédécesseur désormais dévoilées, en donnant plus de poids aux données mal prédites.

2. L'eXtreme Gradient Boosting (XGBoost)

Le Gradient Boosting est une technique d'apprentissage automatique pour les problèmes de régression et de classification, qui produit un modèle de prédiction sous la forme d'un ensemble de modèles de prédiction faibles, généralement des arbres de décision.

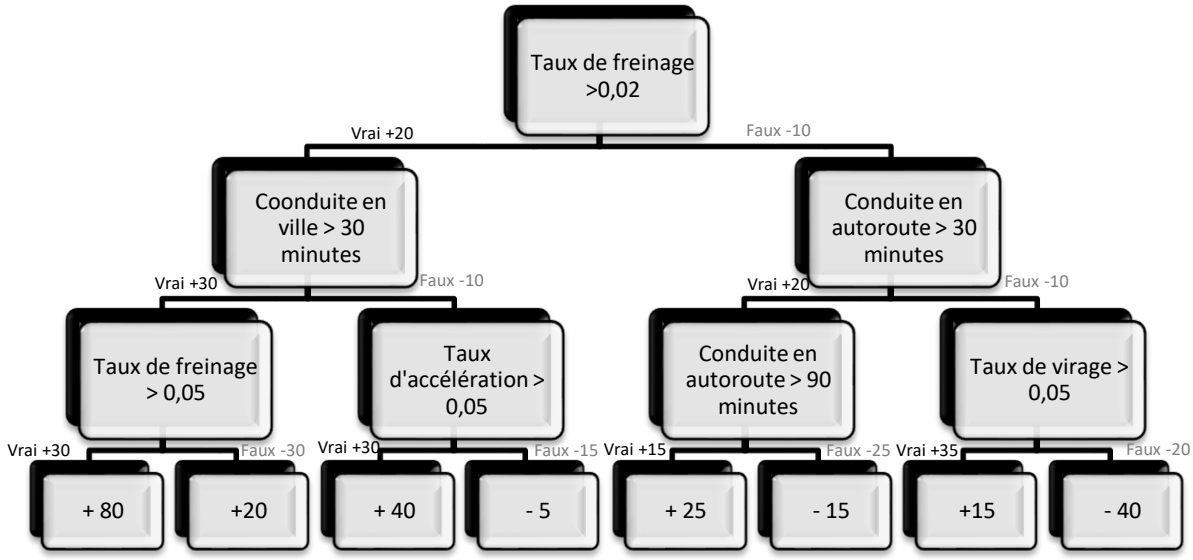
Il construit le modèle par étapes comme le font d'autres méthodes de boosting, et il les généralise en permettant l'optimisation d'une fonction de perte différentiable arbitraire.

XGBoost est l'une des implémentations du concept Gradient Boosting, mais ce qui rend XGBoost unique, c'est qu'il utilise "une formalisation de modèle plus régularisée pour contrôler

le surajustement, ce qui lui donne de meilleures performances", selon l'auteur de l'algorithme, *Tianqi Chen*.

3. Exemple d'arbre XGBoost dans la détermination du score de conduite

Dans la création du score de conduite, le modèle XGBoost, combine plusieurs algorithmes afin d'obtenir un résultat unique issu de tous les autres. Il travaille par récurrence : il commence par créer un premier modèle de prédiction, donnant à chaque variable un coefficient d'ajout ou de diminution de la probabilité d'avoir un sinistre en fonction de la valeur de la variable. On peut le représenter sous la forme d'une arborescence :



Le modèle fournie également une mesure de la significativité pour chaque individu. Ensuite un second modèle est réalisé en prenant en compte les bonnes prédictions du premier. Le troisième modèle fonctionne de manière similaire et conserve les bonnes prédictions des deux premiers modèles. Il est ainsi de meilleure qualité. Ce schéma se reproduit en boucle à de nombreuses reprises si bien que le modèle obtenu à la fin de l'algorithme est le meilleur possible.

Après avoir appliqué ce modèle, il en découle une prédiction de la sinistralité chaque jour, pour chaque véhicule, grâce aux variables explicatives et leurs modalités.

Ce modèle de Machine Learning est très efficace pour faire des prédictions issues de variables non linéaires (par exemple des variables ayant un cap sur la valeur du maximum). Il présente cependant un défaut : l'effet "boîte noire". Cet effet provient du fait qu'il prend les arguments et ressort une prédiction sans expliquer ses différentes actions, ce qui peut le rendre difficile à utiliser. Pour pallier ce défaut, nous avons modifié l'algorithme afin de récupérer l'impact de chaque variable dans chacun des modèles qu'il crée.

Pour cela, nous évaluons l'impact de chaque variable prédictive dans chaque modèle créé lors de l'algorithme XGBoost. Puis nous moyennons ces impacts afin d'obtenir l'impact moyen de chaque variable. L'impact moyen obtenu est l'impact marginal moyen de la variable. Ces impacts marginaux permettent de créer des sous-scores à chacune des quatre variables comportementales : accélération, décélération, mouvement latéral et vitesse angulaire. Ils sont créés grâce à la formule :

$$Score_{ij} = 100 \times \left(1 - \frac{\min\left(\max\left(I_m, q_{5\%}(I_m)\right), q_{95\%}(I_m)\right) - q_{5\%}(I_m)}{q_{95\%}(I_m) - q_{5\%}(I_m)}\right)$$

Avec :

- I_m= $ipmact_{ij}^{event}$
- i = véhicule i
- j = journée j
- Event = événement intérêt
- $q_{\%}(x)$= quantile d'ordre I de la variable x

Annexe C : LES DETAILS SUR LES VARIABLES EXPLICATIVES

1. Age du véhicule en nombre d'années

Age véhicule	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
+3	29700	6.07	29700	6.07
0	136898	27.96	166598	34.03
1	148428	30.32	315026	64.35
2	112482	22.98	427508	87.33
3	62037	12.67	489545	100.00

2. Segmentation dépendant du type et du poids des véhicules

Segmentation	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
1) Van <2100kg	16200	3.31	16200	3.31
2) Van >=2100kg	38739	7.91	54939	11.22
3) Car	434606	88.78	489545	100.00

3. Type d'énergie du véhicule

Type d'énergie	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
Elec/Hybrid	60328	12.32	60328	12.32
Other	429217	87.68	489545	100.00

4. Code Ateco (par secteur d'activité)

ATECO	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
1)BANQUE/ASSU	9260	1.89	9260	1.89
2)Commercial	166242	33.96	175502	35.85
3)PHARMA/Loc	45324	9.26	220826	45.11
4)Transport	11296	2.31	232122	47.42
5)Autre	257423	52.58	489545	100.00

5. Zones

ZONE	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
1)	33161	6.77	33161	6.77
2-3-4)	254210	51.93	287371	58.70
5)	44088	9.01	331459	67.71
6)	158086	32.29	489545	100.00

6. Nombre de chevaux fiscaux (<12, >19)

Nb chevaux fiscaux	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
Hyper Puissant	163859	33.47	163859	33.47
Puissant	71806	14.67	235665	48.14
Très Puissant	253880	51.86	489545	100.00

7. Variable Covid

COVID	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
0	289510	59.14	289510	59.14
1	200035	40.86	489545	100.00

8. Taille de la flotte (<2, >4)

Taille flotte	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
GRANDE FLOTTE	56547	11.55	56547	11.55
MOYENNE FLOTTE	102790	21.00	159337	32.55
PETITE FLOTTE	330208	67.45	489545	100.00

9. Marque des véhicules

Marques	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
BM/Autres	103962	21.24	103962	21.24
Ford/Autres	114955	23.48	218917	44.72
Honda/Autres	17014	3.48	235931	48.19
Hyundai/Volzwagen/Cheveloret/Opel/Autres	138633	28.32	374564	76.51
Nissan/Autres	54840	11.20	429404	87.71
Toyota/Renault/Autres	60141	12.29	489545	100.00

10. Tranches de score globale

Score global	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
[0,30[	6618	1.35	6618	1.35
[30,40[	21513	4.39	28131	5.75
[40,50[	52080	10.64	80211	16.38
[50,60[	92624	18.92	172835	35.31
[60,70[	119511	24.41	292346	59.72
[70,80[	104967	21.44	397313	81.16
[80,100[	92232	18.84	489545	100.00

11. Tranches de distance x 1000 km

Distance	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
[0,4[	139167	28.43	139167	28.43
[10,12[	34698	7.09	173865	35.52
[12,14[	29524	6.03	203389	41.55
[14,16[	24643	5.03	228032	46.58
[16,18[	20835	4.26	248867	50.84
[18,20[	17103	3.49	265970	54.33
[20,25[	31974	6.53	297944	60.86
[25,30[	20708	4.23	318652	65.09
[30,35[	13485	2.75	332137	67.85
[35,40[	8646	1.77	340783	69.61
[4,10[	134569	27.49	475352	97.10
[40,160[	14193	2.90	489545	100.00

12. Tranche des TRU (taux de roulage urbain)

TRU	Nombre	Pourcentage	Cumulé	Pourcentage cumulé
[0,20[	86193	17.61	86193	17.61
[20,30[	99855	20.40	186048	38.00
[30,40[	91441	18.68	277489	56.68
[40,50[	72603	14.83	350092	71.51
[50,70[	90199	18.43	440291	89.94
[70,100[	49254	10.06	489545	100.00