**Mémoire présenté devant l'Université de Paris-Dauphine**
**pour l'obtention du Certificat d'Actuaire de Paris-Dauphine**
**et l'admission à l'Institut des Actuaires**

**le**

Par : Ralph BITAR
Titre : Etude et intégration de la mobilité dans la construction d'un modèle de tarification automobile avec zonier

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

---

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

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

*Entreprise :*
Nom : Finactys
Signature :

FINACTYS
853 Ch. SAINT MAYMES
06660 Antibes
contact@finactys.fr
SIRET : 839 954 948 00031

*Membres présents du jury du certificat d'Actuaire de Paris-Dauphine :*

*Directeur de Mémoire en entreprise :*
Nom : Maxime DELCAMBRE
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)***

*Secrétariat* :

*Bibliothèque* :

*Signature du responsable entreprise*

*Signature du candidat*

**Résumé**

L'assurance automobile est un marché très concurrentiel. Diverses méthodes sont mises en place afin d'offrir aux assurés la prime la plus adaptée à leur risque. Avec l'essor du *Big Data*, l'utilisation du *Machine Learning* est devenue un atout majeur pour la tarification.

Avec la prise en compte des caractéristiques propres à l'assuré telles que son âge ou sa catégorie socio-professionnelle, ces méthodes ont surtout permis l'intégration du facteur géographique : c'est le zonier.

Tarifer un produit d'assurance, c'est avant tout connaître le profil des assurés afin de prendre en compte leurs spécificités. Historiquement, l'AGPM assure les militaires, et aujourd'hui encore cette population est majoritaire dans le portefeuille. Les militaires se distinguent par une forte tendance migratoire : leur carrière les oblige à transiter régulièrement entre les différentes communes durant diverses missions. Ainsi, la mesure du risque d'une zone peut être biaisée par ces migrations.

L'objectif de ce mémoire est de mettre en évidence l'existence du phénomène de mobilité, et d'évaluer son impact à travers la construction de modèles de tarification intégrant ce facteur. L'enjeu est de modéliser ce phénomène à travers la sinistralité afin de donner à la compagnie les outils nécessaires dans ses prises de décisions stratégiques.

Un modèle GLM avec zonier a été construit afin de modéliser la sinistralité des assurés. Le modèle a été augmenté par l'ajout d'une variable mobilité indiquant si l'individu est mobile ou non. Puis, un modèle de régression logistique a été mis en place afin de prédire la probabilité qu'un assuré soit mobile. Sur la base des résidus du modèle logistique, un zonier a été construit, permettant de cibler les zones avec un fort taux de mobilité. Enfin, l'utilisation d'un GLMtree a permis de segmenter les assurés selon s'ils sont mobiles ou non, permettant de construire deux zoniers différents et de distinguer clairement les différences entre les deux populations. Cette méthode a notamment permis de construire un zonier qui permet d'indiquer les zones de provenance à risque, en agrégeant les résidus selon le lieu de départ des assurés mobiles.

*Mots-clés* : *Tarification automobile*, *Mobilité*, *Zonier*, *Machine Learning*, *Arbre de décision GLM (GLM-tree)*, *Biais de sélection.*

**Abstract**

Car insurance is a highly competitive market. Various methods are used to offer policyholders the premium best suited to their risk. With the rise of *Big Data*, the use of *Machine Learning* has become a major asset in underwriting.

In addition to taking into account policyholder characteristics such as age or socio-professional category, these methods have above all made it possible to integrate the geographical factor: the zoning.

Pricing an insurance product means first and foremost knowing the profile of policyholders, so as to take their specific characteristics into account. Historically, AGPM has insured military personnel, and today this population still accounts for the majority of the portfolio. Military personnel are characterized by a strong migratory tendency: their careers require them to move regularly between different communes during various missions. As a result, an area's risk measurement can be skewed by its migratory patterns.

The aim of this thesis is to highlight the existence of the mobility phenomenon, and to assess its impact through the construction of pricing models integrating this factor. The challenge is to model this phenomenon through claims experience, in order to provide the company with the tools it needs to make strategic decisions.

A GLM model with zoning was built to model the claims frequency of insured individuals. The model was enhanced by adding a mobility variable indicating whether the individual is mobile or not. Then, a logistic regression model was implemented to predict the probability that an insured person is mobile. Based on the residuals of the logistic model, a zoning was constructed, allowing the identification of areas with a high mobility rate. Finally, the use of a GLMtree allowed for segmenting the insured individuals based on whether they are mobile or not, enabling the construction of two distinct zonings and clearly distinguishing the differences between the two populations. This method, in particular, helped construct a zoning that highlights high-risk origin areas by aggregating the residuals based on the departure location of mobile insured individuals.

*Keywords* : *Car pricing*, *Mobility*, *Zoning*, *Machine Learning*, *Generalized Linear Model Trees (GLMtree)*, *Selection Biais*.

# Note de Synthèse

**Contexte de l'étude : la tarification**

Le secteur de l'assurance est caractérisé par l'inversion du cycle de production. En échange d'une certaine somme appelée la prime, l'assureur propose à l'assuré une couverture contre ses éventuels sinistres. Le montant de la prime est donc fixé avant la survenance du sinistre. L'enjeu pour l'assureur est d'estimer au mieux le montant de cette prime afin de pouvoir tenir son engagement envers l'assuré : c'est la tarification.

La prime payée par l'assuré est appelée la prime commerciale. Son montant correspond à la somme de plusieurs composantes, dont la prime pure, qui représente le risque de l'assuré et sert à couvrir le montant de ses dommages potentiels sur l'année. Son calcul repose sur le modèle coût-fréquence, c'est-à-dire que le montant de versement espéré par l'assureur pour un assuré peut s'exprimer comme le produit du montant moyen d'un sinistre (coût) multiplié par le nombre de sinistres espérés dans l'année (fréquence). Ce mémoire se focalise uniquement sur la modélisation de la fréquence de sinistres annuels pour un assuré donné, pour la garantie responsabilité civile automobile.

Formellement, la quantité qui est estimée est $\mathbf{E}[Y|X = x]$ où $Y$ est la variable aléatoire représentant le nombre de sinistres annuels pour un assuré $X$, dont les caractéristiques telles que l'âge ou la catégorie socioprofessionnelle sont représentées par le vecteur $x$. Cette quantité est estimée à l'aide d'un GLM (Modèle Linéaire Généralisé), c'est-à-dire qu'elle s'exprime, à une fonction près, comme une combinaison linéaire des caractéristiques de l'assuré :

$$\mathbf{E}[Y|X = x] = f(\theta_0 + \theta_1 x_1 + ... + \theta_p x_p).$$

Chaque coefficient représente donc l'impact (linéaire) d'une caractéristique de l'assuré dans l'estimation du risque. Cependant, deux assurés ayant les mêmes caractéristiques mais présents dans deux communes différentes ne sont pas nécessairement exposés aux mêmes risques. L'estimation peut donc être enrichie en intégrant le facteur géographique à travers la construction d'un zonier. Cette variable illustre le niveau de risque géographique d'un lieu donné et est construite sur la base de la perception du risque géographique du portefeuille.

Une fois le modèle sans le zonier construit, chaque individu possède à la fois un nombre de sinistres prédit et un nombre de sinistres observé (ces derniers sont ceux qui ont servi à estimer le modèle). L'écart entre ces deux quantités pour chaque individu est appelé le résidu. Puisque le risque géographique n'est pas pris en compte en amont par le modèle, une partie de l'existence de ces écarts est due à la non-prise en compte du facteur géographique. Ce sont donc ces résidus qui contiennent l'information géographique. La méthode choisie pour la construction du zonier est basée sur ces résidus et est présentée en Figure 1.

Les résidus du modèle sont agrégés à la maille INSEE, puis un modèle de Machine Learning est entraîné à l'aide de données en Open Data afin de prédire le risque des zones manquantes du portefeuille. Deux méthodes sont mises en compétition : les forêts aléatoires et le Gradient Boosting Machine. La performance des deux modèles est estimée par une méthode de validation croisée. Le lissage est inspiré par la méthode de crédibilité, puis les résidus lissés sont classés par un algorithme k-means.

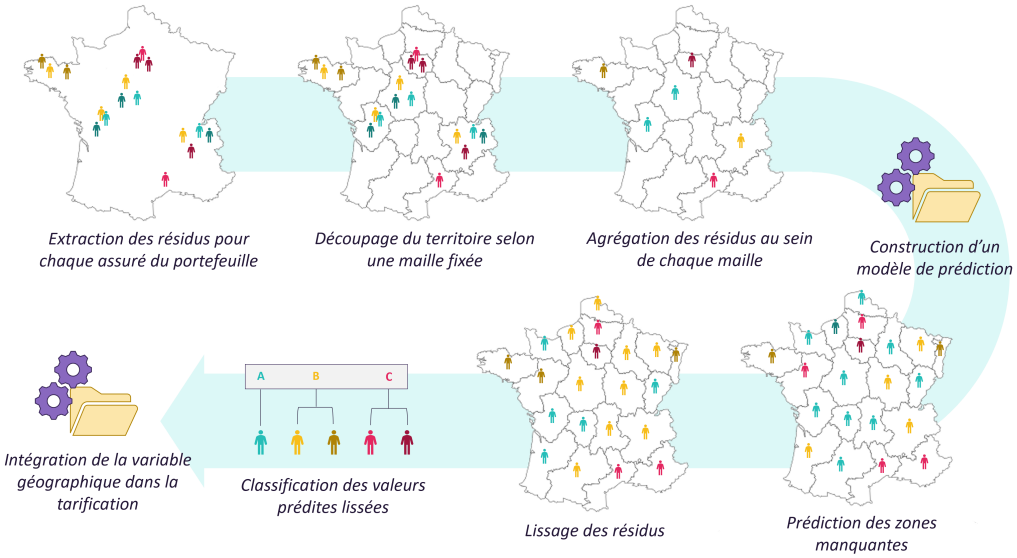



FIGURE 1 : Processus de construction de la variable zonier

**La problématique : un portefeuille mobile**

Le groupe AGPM a la particularité d'assurer une population majoritairement militaire. Au cours de leur carrière, ils sont amenés à effectuer des missions dans diverses régions du monde. C'est ce qui est appelé une "mutation". L'impact de ce phénomène est double. D'un côté, le comportement des assurés qui ont muté peut être affecté par le changement d'environnement. L'estimation des coefficients du modèle peut donc être biaisée par ce changement de comportement. De l'autre côté, les assurés qui ont muté ne perçoivent peut-être pas le risque géographique de la même manière que ceux qui n'ont pas muté, ce qui peut biaiser la construction du zonier. Pour rappel, ce sont les résidus qui portent l'information du risque géographique. Ainsi, le risque d'une zone sera une agrégation des résidus de tous les individus de la commune en question. Néanmoins, si parmi ces derniers se trouvent des "nouveaux entrants" qui viennent d'intégrer la commune, il est possible que leurs résidus soient biaisés par le fait d'avoir changé d'environnement. En effet, ces résidus contiennent à la fois l'écart dû à la zone géographique, mais aussi à l'effet de la mobilité. Ainsi, une agrégation simple des résidus des individus d'une zone sans distinguer les nouveaux entrants fausserait la mesure du risque géographique.

Plus largement, un assuré qui arrive dans un nouvel environnement pourrait être exposé à un risque d'accidentalité plus élevé, dû à une mauvaise connaissance des pièges de circulation, des endroits à éviter, ou du stress lié à la conduite dans une ville inconnue. L'objet de ce mémoire répond donc aux problématiques suivantes : le phénomène de mobilité a-t-il un impact (significatif) sur la mesure du risque automobile ? Comment le modéliser le cas échéant ?

**Présentation du portefeuille**

L'AGPM étant placée à Toulon, ce sont principalement des militaires de la marine qui sont assurés. Les bases de Brest et de Toulon sont les principaux points de défense maritime de la France. À l'échelle de la métropole, ce sont ces deux villes qui détiennent le flux migratoire le plus important. Cependant, d'autres flux existent, notamment le plus important qui est celui reliant la métropole à la Guyane (Figure 2).

La mobilité est un phénomène relativement présent dans le portefeuille. Chaque année, c'est en moyenne 7,44% des contrats renouvelés ou nouveaux entrants qui correspondent à des individus ayant changé de commune. De façon plus large, ce phénomène concerne 18,69% des polices de l'AGPM sur la période d'étude (2016 à 2022, sans considérer l'année 2020 marquée par la pandémie).

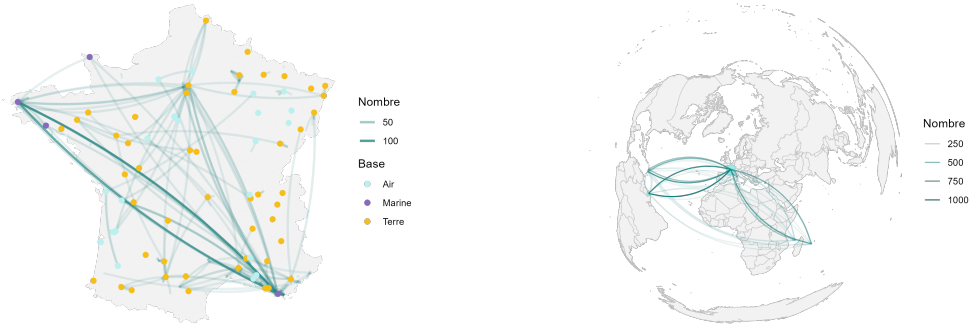



FIGURE 2 : Cartographie des migrations du portefeuille

**Démarche de résolution**

L'ensemble des modèles construits et utilisés afin de répondre à la problématique est représenté en Figure 3.

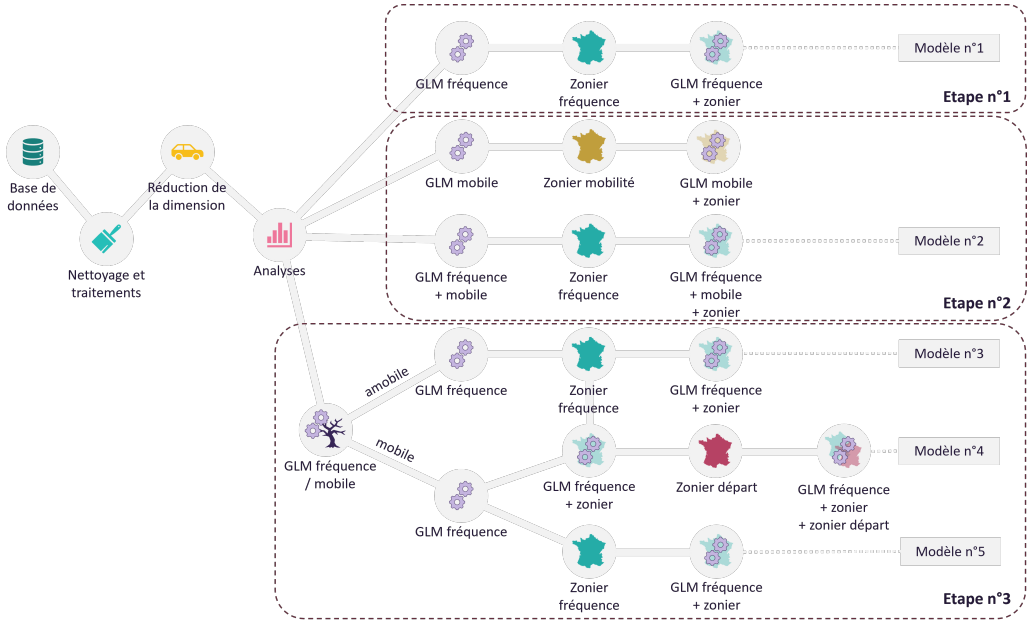
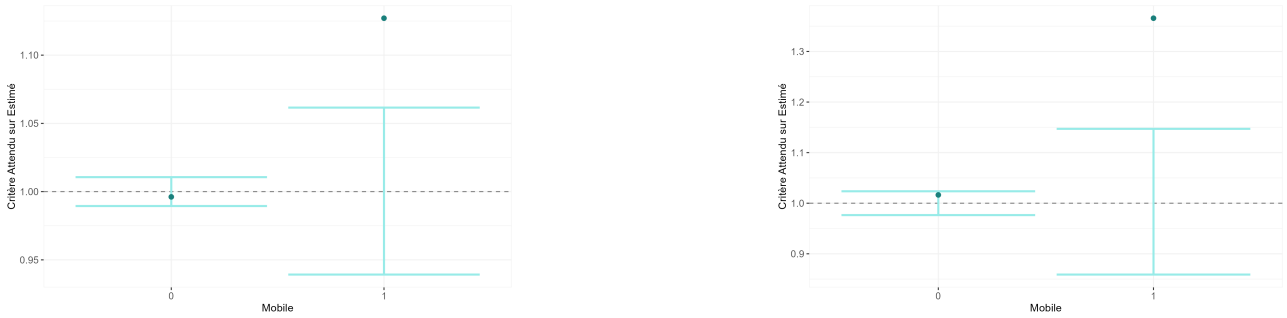



FIGURE 3 : Démarche de résolution générale

**Étape n°1 : construction d'un modèle de référence**

Après traitement de la base de données, la première étape consiste à construire un modèle qui ne prend pas en compte le phénomène de mutation afin d'obtenir un modèle de référence. C'est un GLM Poisson qui sera retenu. Puis, sur la base de ce modèle, il est possible de quantifier la significativité du phénomène de mutation dans l'estimation du risque. Pour cela, le critère « attendu sur estimé » (rapport entre la somme des valeurs cibles et la somme des valeurs estimées) a été calculé sur les deux populations (mobile et non mobile), que ce soit au test et à l'entraînement (Figure 4).

Dans les deux cas, il s'avère que le critère est significativement proche de 1 pour la population non mobile, mais significativement supérieur à 1 pour la population mobile (et donc que la somme des valeurs prédites est inférieure à la somme des valeurs cibles). Plus précisément, ce coefficient est supérieur à 1,35 au test, ce qui signifie que le risque réel des mobiles est en réalité 1,35 fois supérieur au risque estimé. Ce résultat signifie que




(a) Comparaison à l'entraînement

(b) Comparaison au test

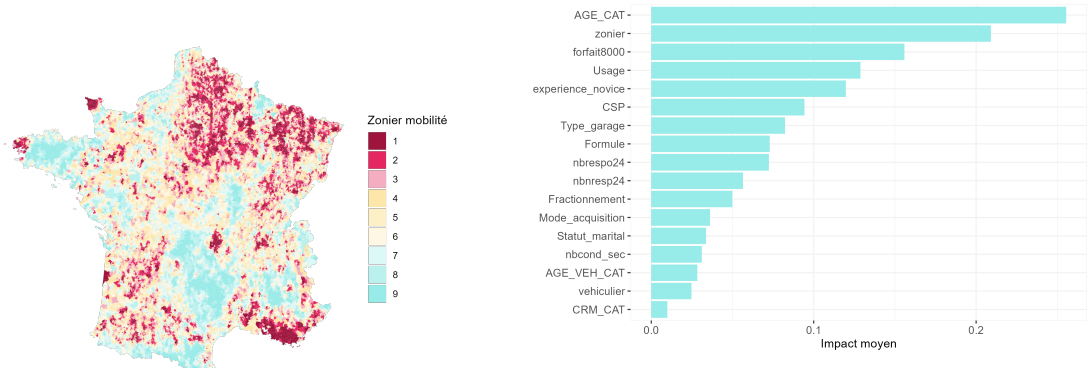
FIGURE 4 : Comparaison de la performance du modèle selon la mobilité

le modèle sous-estime clairement le risque de la population mobile, ou que ces derniers représentent un risque plus important que la population de base.

La mobilité est donc un phénomène qui, au global, aggrave la sinistralité des assurés. Cette hypothèse étant satisfaite, l'objectif des prochaines étapes est de prendre en compte ce phénomène et de le modéliser.

**Étape n°2 : intégration d'une variable mobilité et identification des assurés avec un fort taux de mobilité**

La prochaine étape est de reconstruire le modèle de tarification en y intégrant cette fois-ci la variable *mobile*, qui vaut 1 si l'individu est mobile, et 0 sinon. Cependant, cette caractéristique n'est pas fournie lors de la souscription. Cette variable a été construite sur la base de l'historique du portefeuille. Dans la mesure où cette variable est binaire, l'idée est de remplacer sa valeur par la probabilité d'être un individu mobile (qui est une valeur comprise entre 0 et 1). Un modèle de régression logistique a donc été mis en place. À l'aide des résidus du modèle, un zonier a été construit. Ce zonier permet de mettre en avant les zones avec un fort taux de mobilité, c'est-à-dire les zones où les individus mobiles sont les plus concentrés. La Figure 5 permet de visualiser le zonier en question et l'impact des coefficients dans l'estimation de la probabilité d'être mobile.




(a) Zonier mobilité

(b) Impact des coefficients sur la probabilité d'être mobile

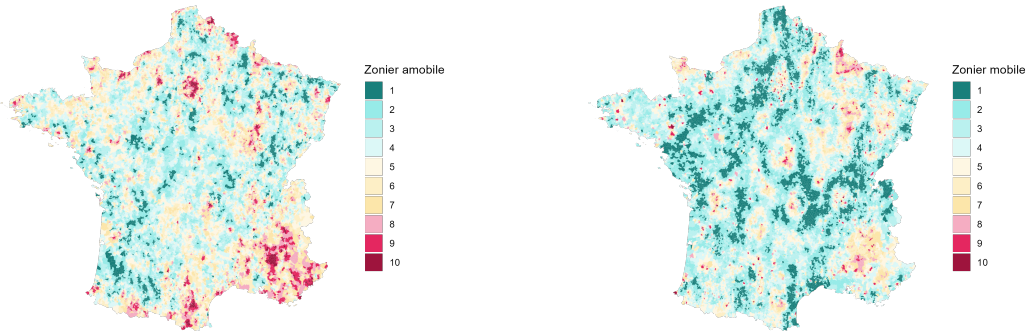
FIGURE 5 : Modélisation de la probabilité d'être mobile

Le zonier obtenu est une variable qualitative ordinale à 10 modalités : la valeur 1 (rouge) signifie que la zone a un fort taux de mobilité (contient beaucoup d'assurés mobiles), tandis que la valeur 10 (vert) signifie que la zone a un très faible taux de mobilité. Il s'avère que les zones les plus en proie à la mobilité sont les zones proches des bases de défense, notamment celles de la marine (Brest et Toulon). Ce zonier permet entre

autres de localiser le risque de mobilité. Par ailleurs, les variables du modèle de régression logistique ont été analysées. Une des variables les plus importantes est le forfait8000, qui est une option permettant à l'assuré de réduire sa prime s'il roule moins de 8000 km par an. De ce fait, il est plus probable qu'un assuré roulant plus de 8000 km par an soit mobile. Ces modèles sont donc des moyens d'anticiper de potentiels accidents dus à la mobilité en mettant en place un système de suivi, de prévention ou d'accompagnement des assurés.

**Étape n°3 : modélisation segmentée entre assurés mobiles et non mobiles, avec identification des zones de départ à risque**

Jusqu'à présent, les modèles construits appliquent les mêmes coefficients pour les deux types de populations. Cependant, il pourrait être intéressant d'estimer des coefficients différents selon les deux populations considérées. L'idée est d'appliquer un GLMtree : si l'estimation des coefficients est sensible à la population considérée, alors le modèle divise la base de données et construit deux modèles distincts, sinon, il renvoie un seul GLM. Le test de significativité renvoie une p-valeur inférieure à $0{,}01$, et deux GLM sont renvoyés. Ces modèles sont ensuite optimisés par sélection de variables AIC. Puis, sur la base des résidus, deux zoniers ont été construits (Figure 6).




(a) Zonier amobile (b) Zonier mobile

FIGURE 6 : Comparaison des zoniers amobile et mobile

Sur ces deux zoniers, une valeur à 1 signifie zone peu risquée, et une valeur à 10 signifie zone très risquée. Cependant, le risque porté par la modalité 1 chez les non-mobiles (amobiles) n'est pas équivalent à celui de la modalité 1 chez les mobiles (il n'y a pas correspondance des risques entre deux modalités identiques pour les deux zoniers). Le zonier des non-mobiles (amobiles) représente le risque géographique réel puisque ces derniers sont statiques. En revanche, le zonier construit sur la base des assurés mobiles est complètement différent. Par ailleurs, il ne semble pas intuitif dans la représentation du risque géographique : certaines zones au sud de la métropole ou en Île-de-France sont moins risquées qu'elles ne devraient l'être (en se référant au zonier des amobiles). En réalité, ce zonier représente un mélange de risques : pour une zone donnée, le risque est estimé sur la base d'assurés qui proviennent de plusieurs zones différentes. Le risque perçu est donc un mélange entre le risque réel de la zone et la somme des risques des zones de provenance.

Enfin, il peut être intéressant de repérer les zones de provenance à risque (c'est-à-dire identifier les zones dont la provenance aggrave la sinistralité). Pour ce faire, le zonier des non-mobiles est intégré au GLM(mobile) (puisque ce dernier représente le risque géographique réel). Les résidus du modèle sont extraits, et au lieu d'agréger les résidus au sein de la commune de présence des assurés au moment de l'accident, ces résidus sont agrégés dans la commune de provenance. On obtient alors un zonier qui indique, pour une zone donnée, le risque "d'en provenir" (Figure 7).

Une zone classée 10 (en rouge) signifie que les assurés mobiles provenant de cette zone sont très risqués, tandis que ceux provenant d'une zone 1 (en vert) le sont moins.

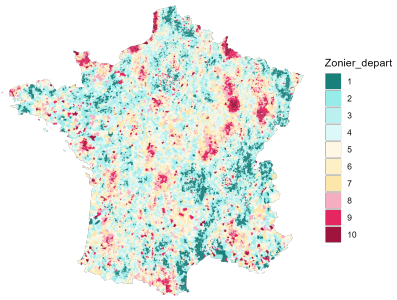



FIGURE 7 : Zonier de provenance

**Comparaison des performances**

La performance de l'ensemble des modèles de fréquence construits est analysée sur une base test selon deux critères : la MAE (Erreur Absolue Moyenne : somme de la différence en valeur absolue entre les valeurs cibles et les valeurs prédites), et la MSE (Erreur Quadratique Moyenne : identique à la MAE, mais la valeur absolue est remplacée par le carré de la différence). Les résultats sont présentés dans le tableau ci-dessous :

| Modèles | MAE | MSE | MAE (amobile) | MAE (mobile) | MSE (amobile) | MSE (mobile) |
|---|---|---|---|---|---|---|
| 1 | 0.04523 | 0.02275 | 0.04617 | 0.02761 | 0.02311 | 0.01602 |
| 2 | 0.04522 | 0.02276 | 0.04609 | 0.02904 | 0.02312 | 0.01594 |
| 3 et 5 | 0.04521 | 0.02461 | 0.04602 | 0.03008 | 0.02495 | 0.01821 |
| 3 et 4 | 0.04521 | 0.02460 | 0.04602 | 0.03005 | 0.02495 | 0.01816 |

TABLE 1 : Performances des modèles de fréquence

Les numéros des modèles correspondent à ceux présentés sur la Figure 3, ordonnés dans le tableau par complexité. Plus le modèle est complexe (modèles 3 et 5 ou 3 et 4), plus la MAE diminue alors que la MSE se dégrade. Or, la MSE accorde plus d'importance aux erreurs conséquentes, ce qui, dans ce contexte, est dû à une mauvaise détection des individus à risque. La complexification du modèle a donc permis de mieux détecter les individus peu risqués, au détriment des individus risqués. Ceci se justifie par le fait que la variable cible est zéro-inflatée : la population risquée est sous-représentée, et il est donc difficile de la détecter, notamment avec un modèle simple de type GLM.

Ces deux critères ont aussi été utilisés pour comparer la performance des modèles sur les deux types de populations séparément. Chez la population non mobile, la MAE s'améliore tandis que la MSE se dégrade ; en revanche, chez les mobiles, les deux critères se dégradent. Le phénomène de mobilité reste complexe, et potentiellement sous-représenté pour que le modèle produise des résultats significatifs.

**Conclusion**

Dans le cadre de ce mémoire, il a été montré que le phénomène de mobilité, tel que défini, a un impact sur la tarification (sous-estimation du risque des mobiles). Cependant, les modèles mis en place afin de modéliser ce phénomène ne permettent pas d'obtenir des résultats suffisamment significatifs pour fournir à la compagnie des outils utiles à la prise de décision stratégique.

Par ailleurs, la mobilité a été analysée uniquement du point de vue de la fréquence. Il serait toutefois pertinent d'analyser également ce phénomène du point de vue du coût moyen. Si la fréquence des sinistres chez les mobiles est plus élevée, il se pourrait que ce soient en réalité des sinistres à faible coût, ce qui compenserait la sursinistralité observée. L'impact du phénomène sur la prime, et plus largement sur la solvabilité, reste une piste d'ouverture à explorer.

# Synthesis note

**Context of the Study: Pricing**

The insurance sector is characterized by the reversal of the production cycle. In exchange for a certain sum called the premium, the insurer offers the insured coverage against potential claims. The amount of the premium is therefore set before the occurrence of the claim. The challenge for the insurer is to estimate the amount of this premium in order to fulfill its commitment to the insured: this is pricing.

The premium paid by the insured is called the commercial premium. Its amount corresponds to the sum of several components, including the pure premium, which represents the risk of the insured and covers the amount of potential damage over the year. Its calculation is based on the cost-frequency model, meaning that the expected payout for an insured individual can be expressed as the product of the average cost of a claim (cost) multiplied by the expected number of claims in the year (frequency). This thesis focuses solely on modeling the annual claim frequency for a given insured, for automobile liability coverage.

Formally, the quantity being estimated is $\mathbf{E}[Y|X = x]$ where $Y$ is the random variable representing the annual number of claims for an insured $X$, whose characteristics such as age and socioeconomic category are represented by the vector $x$. This quantity is estimated using a GLM (Generalized Linear Model), meaning it is expressed, with some function, as a linear combination of the characteristics of the insured

$$\mathbf{E}[Y|X = x] = f(\theta_0 + \theta_1 x_1 + ... + \theta_p x_p).$$

Each coefficient thus represents the (linear) impact of a characteristic of the insured in the risk estimation. However, two insured individuals with the same characteristics but located in two different municipalities may not necessarily be exposed to the same risks. The estimation can therefore be enhanced by incorporating the geographical factor through the construction of a zoning variable. This variable illustrates the geographical risk level of a given area and is built based on the geographical risk perception of the portfolio.

Once the model without the zoning variable is constructed, each individual has both a predicted number of claims and an observed number of claims (the latter are those used to estimate the model). The difference between these two quantities for each individual is called the residual. Since the geographical risk is not accounted for upfront by the model, part of the existence of these differences is due to the absence of the geographical factor. It is these residuals that contain the geographical information. The method chosen for constructing the zoning variable is based on these residuals and is shown in Figure 8.

The model residuals are aggregated at the INSEE level, and then a Machine Learning model is trained using Open Data to predict the risk in missing areas of the portfolio. Two methods are put to the test: Random Forests and Gradient Boosting Machines. The performance of both models is assessed using cross-validation. Smoothing is inspired by the credibility method, and the smoothed residuals are then classified using a k-means algorithm.

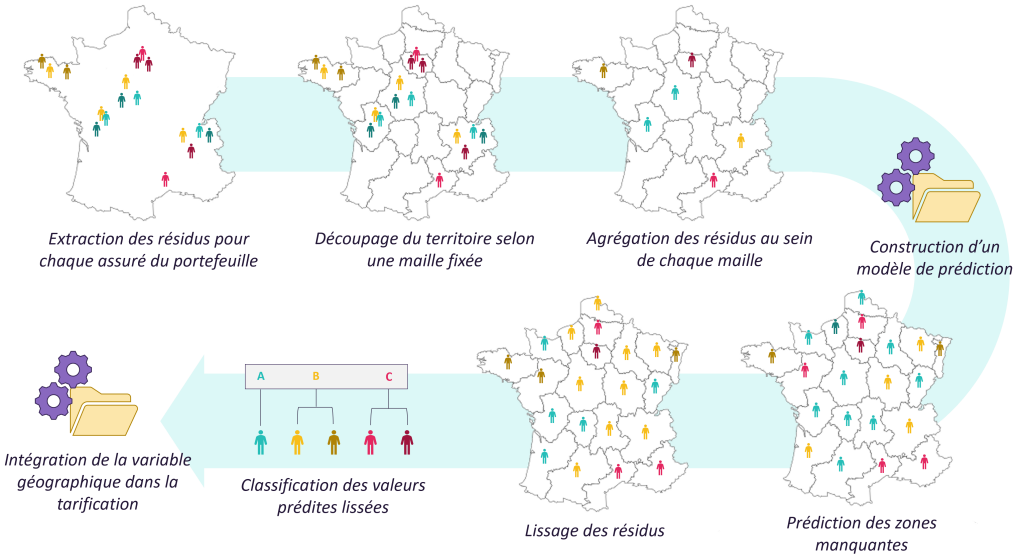



Figure 8: Process of constructing the zoning variable

**The Issue: A Mobile Portfolio**

The AGPM group is unique in that it insures a predominantly military population. Throughout their careers, military personnel are required to undertake missions in various regions around the world. This is referred to as a "posting." The impact of this phenomenon is twofold. On one hand, the behavior of the insured who have been posted may be affected by the change in environment. This could bias the estimation of the model coefficients. On the other hand, insured individuals who have been posted may perceive geographical risk differently than those who have not been posted, which could skew the construction of the zoning variable. As a reminder, it is the residuals that carry the information about geographical risk. Therefore, the risk of an area will be an aggregation of the residuals of all the individuals in the given municipality. However, if among these individuals there are "new entrants" who have just moved to the area, it is possible that their residuals are biased due to their change in environment. These residuals contain both the difference due to the geographical area and the effect of mobility. Thus, a simple aggregation of the residuals of individuals in an area without distinguishing new entrants would distort the measure of geographical risk.

More generally, an insured individual entering a new environment could be exposed to a higher risk of accidents due to unfamiliarity with traffic hazards, places to avoid, or stress related to driving in an unfamiliar city. The purpose of this thesis therefore addresses the following issues: does the mobility phenomenon have a (significant) impact on the measurement of automobile risk? If so, how can it be modeled?

**Presentation of the Portfolio**

Since AGPM is based in Toulon, it mainly insures military personnel from the navy. The naval bases of Brest and Toulon are France's main maritime defense points. On a national scale, these two cities have the highest migration flows. However, other flows exist, notably the largest one linking mainland France to French Guiana (Figure 9).

Mobility is a relatively common phenomenon in the portfolio. Each year, an average of 7.44% of renewed or newly entered contracts correspond to individuals who have changed municipalities. More broadly, this phenomenon concerns 18.69% of AGPM policies over the study period (2016 to 2022, excluding the year 2020 marked by the pandemic).

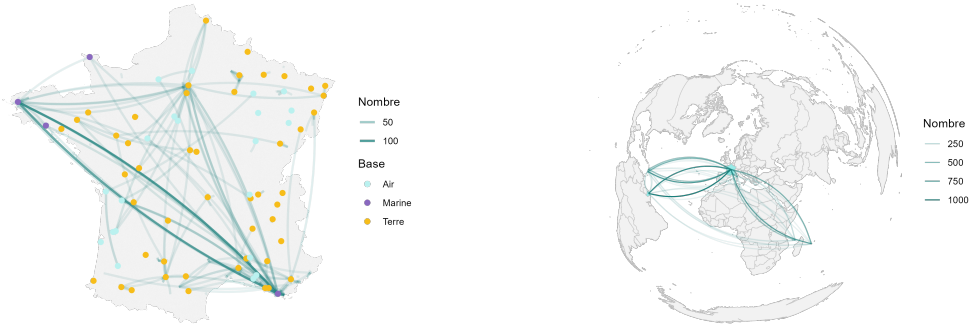



Figure 9: Migration mapping of the portfolio

**Problem Solving Approach**

The set of models constructed and used to address the issue is depicted in Figure 10.

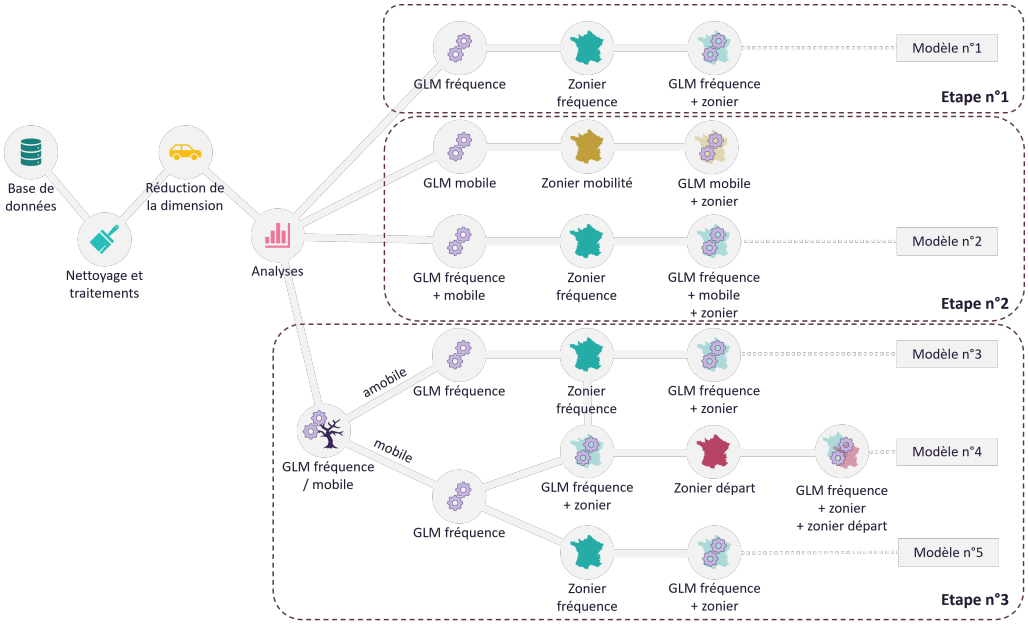
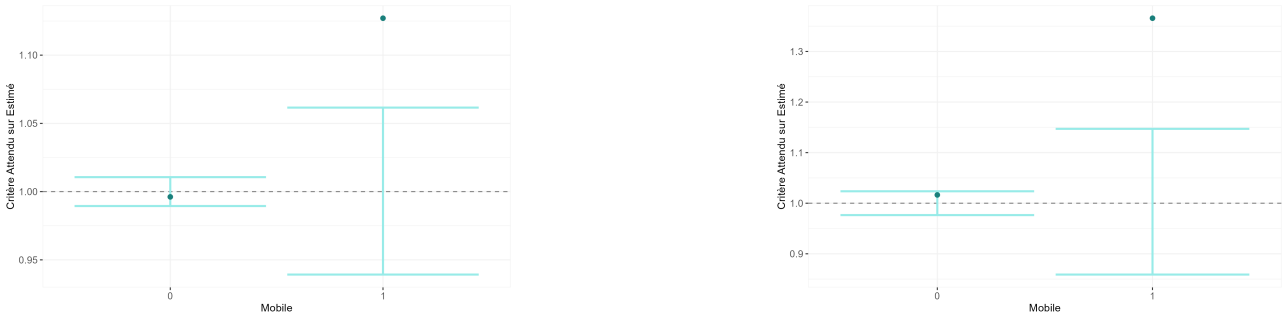



Figure 10: Resolution approach

**Step 1: Building a Reference Model**

After processing the database, the first step is to build a model that does not account for the posting phenomenon in order to obtain a reference model. A Poisson GLM will be selected. Based on this model, it is possible to quantify the significance of the posting phenomenon in risk estimation. To do this, the ”expected overestimated” criterion (the ratio of the sum of target values to the sum of predicted values) was calculated for both populations (mobile and non-mobile), for both testing and training (Figure 11).

In both cases, it turns out that the criterion is significantly close to 1 for the non-mobile population, but significantly greater than 1 for the mobile population (meaning that the sum of the predicted values is lower than the sum of the target values). More specifically, this coefficient is greater than 1.35 in the test, which means that the actual risk for the mobile population is actually 1.35 times higher than the estimated risk. This result indicates that the model clearly underestimates the risk for the mobile population, or that they represent a greater risk than the base population.




(a) Comparison during training

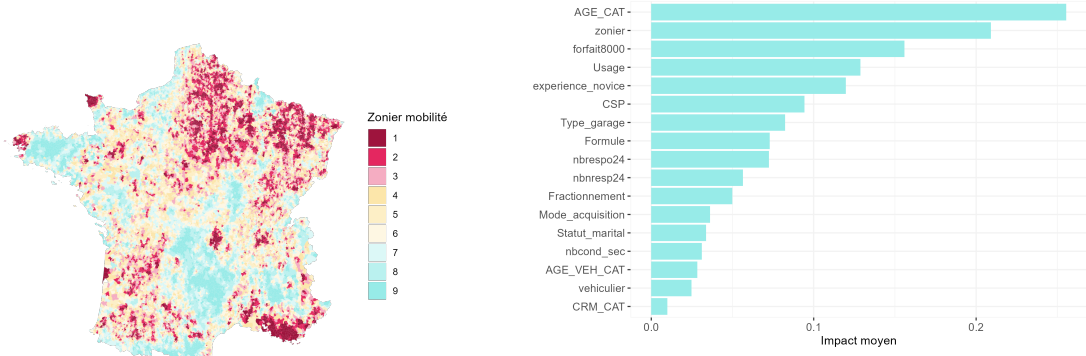
(b) Comparison during testing

Figure 11: Comparison of model performance by mobility status

Mobility is therefore a phenomenon that, overall, increases the claims frequency for insured individuals. This hypothesis being satisfied, the next steps aim to account for and model this phenomenon.

**Step 2: Incorporating a Mobility Variable and Identifying Insured Individuals with High Mobility**

The next step is to rebuild the pricing model by incorporating a mobility variable, which is set to 1 if the individual is mobile, and 0 otherwise. However, this characteristic is not provided at the time of subscription. This variable was constructed based on the historical data of the portfolio. Since this variable is binary, the idea is to replace its value with the probability of being a mobile individual (a value between 0 and 1). A logistic regression model was therefore implemented. Using the residuals of the model, a zoning variable was constructed. This zoning variable highlights areas with a high mobility rate, i.e., areas where mobile individuals are most concentrated. Figure 13 allows visualizing the zonier in question and the impact of the coefficients on the estimation of the probability of being mobile.




(a) Mobility zoning variable

(b) Impact of coefficients on the probability of being mobile

Figure 12

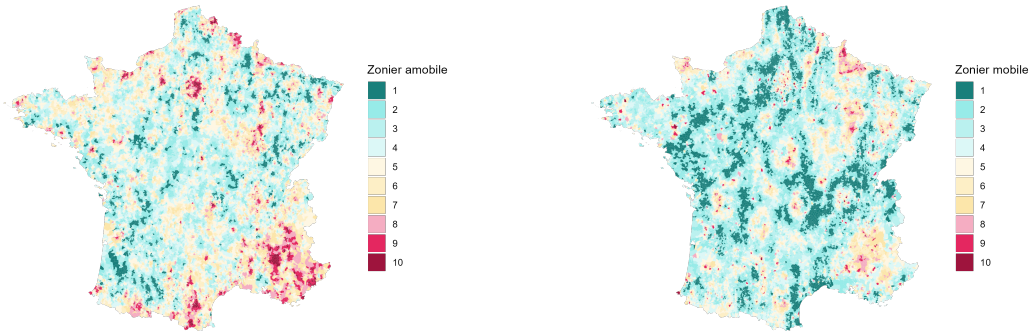
Figure 13: Modeling the probability of being mobile

It turns out that the areas most affected by mobility are those near defense bases, particularly those of the navy (Brest and Toulon). This zoning variable helps to locate mobility risk. Furthermore, the variables from the logistic regression model were analyzed. One of the most significant variables is the ”forfait8000,” an option that allows the insured to reduce their premium if they drive less than 8000km per year. Consequently, it is more likely that an insured individual driving more than 8000km per year is mobile. These models are therefore ways to anticipate potential accidents due to mobility by implementing a monitoring, prevention, or support system

for the insured.

**Step 3: Segmented Modeling for Mobile and Non-Mobile Insured Individuals, with Identification of High-Risk Departure Zones**

Until now, the models have applied the same coefficients for both populations. However, it might be interesting to estimate different coefficients for the two populations. The idea is to apply a GLMtree: if the estimation of coefficients is sensitive to the population considered, the model splits the database and constructs two distinct models; otherwise, it returns a single GLM. The significance test returns a p-value less than 0.01, and two GLMs are returned. These models are optimized by variable selection using AIC. Then, based on the residuals, two zoniers were constructed (Figure 14).




(a) Non-mobile zoning variable

(b) Mobile zoning variable

Figure 14: Comparison of non-mobile and mobile zoning variables

The non-mobile zoning variable is similar to the first model's zoning variable (which makes sense since they are the majority). This zoning variable represents the real geographical risk since they are static. On the other hand, the zoning variable based on mobile insured individuals is quite different. Moreover, it doesn't seem intuitive in the representation of geographical risk: some areas in the southern part of the country or in Île-de-France are less risky than they should be (when referring to the non-mobile zoning variable). In reality, this zoning variable represents a mixture of risks: for a given area, the risk is estimated based on insured individuals from several different areas. The perceived risk is therefore a mix of the real risk of the area and the sum of risks from the areas of origin.

Finally, it may be interesting to identify high-risk origin zones (i.e., to identify areas where the origin worsens the claims rate). To do this, the non-mobile zoner is integrated into the GLM(mobile) (since the latter represents the actual geographical risk). The residuals of the model are extracted, and instead of aggregating the residuals within the municipality where the insured person resides at the time of the accident, these residuals are aggregated in the municipality of origin. This results in a zoner that indicates, for a given area, the risk of 'originating from there' (Figure 15).

**Performance Comparison**

The performance of all the constructed frequency models is analyzed on a test dataset using two criteria: MAE (the sum of the absolute differences between target and predicted values) and MSE (similar to MAE but with the squared differences). The results are presented in the table below:

The more complex the model, the lower the MAE, but the MSE deteriorates. However, the MSE places more importance on significant errors, which in this context is due to poor detection of high-risk individuals. The complexity of the model has thus allowed for better detection of low-risk individuals at the expense of high-risk individuals. This can be explained by the fact that the target variable is zero-inflated: the risky population

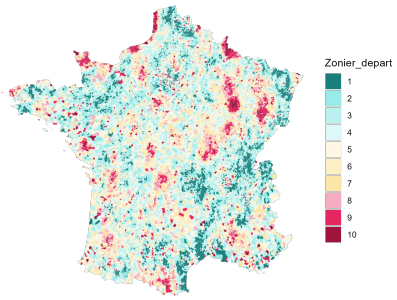



Figure 15: Origin zoning

| Modèles | MAE | MSE | MAE (amobile) | MAE (mobile) | MSE(amobile) | MSE(mobile) |
|---|---|---|---|---|---|---|
| 1 | 0.04523234 | 0.02275856 | 0.04617525 | 0.02761145 | 0.02311878 | 0.01602687 |
| 2 | 0.04522914 | 0.02276068 | 0.04609515 | 0.02904554 | 0.02312535 | 0.01594575 |
| 3 et 5 | 0.04521208 | 0.02461193 | 0.04602132 | 0.03008929 | 0.02495445 | 0.01821112 |
| 3 et 4 | 0.04521051 | 0.02460953 | 0.04602132 | 0.03005825 | 0.02495445 | 0.01816386 |

Table 2: Performance of frequency models

is underrepresented, making them difficult to detect, particularly with a simple model like a GLM.

These two criteria were also used to compare the performance of the models on the two populations separately. For the non-mobile population, the MAE improves while the MSE worsens, but for the mobile population, both criteria worsen. The mobility phenomenon remains complex, and potentially underrepresented for the model to yield significant results. However, the study allowed for identifying risks and providing the company with the necessary tools for its development.

**Conclusion**

This thesis has shown that the phenomenon of mobility, as defined, has an impact on pricing (underestimation of mobile risk). However, the models set up to model the phenomenon do not provide meaningful results, unless they can provide the company with the tools it needs to make strategic decisions.

In addition, mobility has been analyzed solely from the point of view of frequency. However, it would also be relevant to analyze the phenomenon from the point of view of average cost. If the frequency of mobile claims is higher, it could be that these are actually low-cost claims, which would compensate for the observed excess claims rate. The impact of this phenomenon on premiums and, more broadly, on solvency remains an avenue worth exploring.

# Remerciements

La rédaction de ce mémoire est pour moi la fin d'un beau parcours riche en expérience et surtout plein de belles rencontres. J'ai eu la chance d'échanger avec des personnes qui ont marqué mon esprit et m'ont permis de devenir la personne que je suis aujourd'hui.

Tout d'abord, un grand merci à mes tuteurs de stage, **Maxime DELCAMBRE** et **Eléa LAMOUROUX**, pour votre soutien et votre encadrement. Je remercie particulièrement **Léonie LE BASTARD** pour toute la connaissance et l'expérience qu'elle m'a apportée. Je remercie aussi toute l'équipe FINACTYS : **Charles BODDELE**, **Amélie CAUSSE**, **Hugo CIVEL**, **Nicolas LEROUX**, **Damien LOUREIRO**, et **Lucas THOUVENOT**. Chacun d'entre vous est pour moi une grande source d'inspiration, et c'est un plaisir et un honneur de faire partie de cette famille. Un grand merci à vous !

Je remercie aussi **Guillaume GONNET** et **Nidhal JOUINI** avec qui j'ai eu le plaisir de collaborer dans le cadre de ce projet de fin d'études. Merci à vous et à toute l'équipe de l'AGPM.

Une partie de ma réussite je la dois à mes professeurs qui m'ont transmis leur savoir. Je remercie le directeur du master Actuariat **Quentin GUIBERT**, mon tuteur académique **André GRONDIN**, et tous les professeurs intervevants de l'université de Paris Dauphine. Je remercie également **Anthonny REVEILLAC** de m'avoir permis d'intégrer le master de Dauphine. Je remercie tous mes anciens professeurs pour ces quatres années d'études passées à vos côtés. Je tiens à mentionner ma tutrice académique INSA **Béatrice LAURENT** et mon ancien professeur **Mélisande ALBERT** pour leur pédagogie et la rigeur qu'elles m'ont transmises.

***Enfin, et pour moi le plus important, merci à ma famille et à mes amis qui m'ont soutenu inconditionnellement toute ma vie. C'est à vous que je dédie ma réussite, et la rédaction de ce mémoire.***

# Table des matières

# Introduction

Le secteur de l'assurance est caractérisé par l'inversion du cycle de production. En échange d'une certaine somme appelée la prime, l'assureur propose à l'assuré une couverture contre ses éventuels sinistres. Le montant de la prime est donc fixé avant la survenance du sinistre. L'enjeu pour l'assureur est d'estimer au mieux le montant de cette prime afin de pouvoir tenir son engagement envers l'assuré : c'est la tarification.

Le développement du *Machine Learning* a permis de construire des modèles de tarification plus adaptés aux spécificités de l'assuré, en tenant compte de ses caractéristiques telles que son âge ou sa catégorie socioprofessionnelle. L'essor du *Big Data* a permis à l'assureur de stocker et d'utiliser ces informations afin de s'adapter et de faire face à la concurrence. Par ailleurs, l'émergence de l'*Open Data* a permis à l'assureur d'enrichir ces modèles à l'aide d'informations fournies en *Open Source*. Cette avancée lui a notamment permis d'intégrer le risque géographique dans les modèles de tarification, appelé zonier.

Le portefeuille d'un assureur est généralement très diversifié. Cependant, il peut se spécialiser selon la politique de la compagnie. Un portefeuille spécialisé signifie que certains profils sont présents en majorité. C'est le cas du groupe AGPM (Association Générale de Prévoyance Militaire). Initialement créé en 1951 pour les militaires français durant la guerre d'Indochine, ce n'est qu'en 1998 que la compagnie s'est ouverte au grand public. Encore aujourd'hui, son portefeuille est marqué d'une forte empreinte militaire.

La spécialisation d'un portefeuille peut faire émerger un biais de sélection important. Le biais de sélection apparaît lorsque la population d'un échantillon donné n'est pas représentative de la population générale. S'agissant ici des militaires, il peut être pertinent de prendre en compte les spécificités de cette population afin que les offres proposées soient plus adaptées.

La particularité de cette population qui est au cœur de ce mémoire est la mobilité. Au cours de leur carrière, les militaires sont amenés à effectuer des missions dans diverses régions du monde. C'est ce qui est appelé une "mutation". L'impact de ce phénomène est double. D'un côté, le comportement des assurés qui ont muté peut être affecté par le changement d'environnement, ce qui peut biaiser la mesure de leur risque individuel par rapport à ceux qui n'ont pas muté. De l'autre côté, les assurés qui ont muté ne perçoivent peut-être pas le risque géographique de la même manière que ceux qui n'ont pas muté, ce qui peut biaiser la construction du zonier. Bien que ce phénomène soit inspiré du comportement militaire, il peut se généraliser à tous les portefeuilles lorsque l'étude s'y prête.

Le sujet de ce mémoire porte donc sur la modélisation du phénomène de mobilité à travers la tarification automobile. L'objectif est d'étudier ce phénomène grâce à une modélisation par *Machine Learning* à travers la construction de modèles de tarification. L'intérêt est aussi de fournir à la compagnie les outils et les références nécessaires pour l'aider dans son pilotage stratégique.

L'écrit se structure en quatre chapitres. Le premier 1 définit le cadre du mémoire et aborde les différentes notions qui seront utilisées par la suite. Il présente le contexte assurantiel de l'étude, les modèles de tarification classiques, et la problématique.

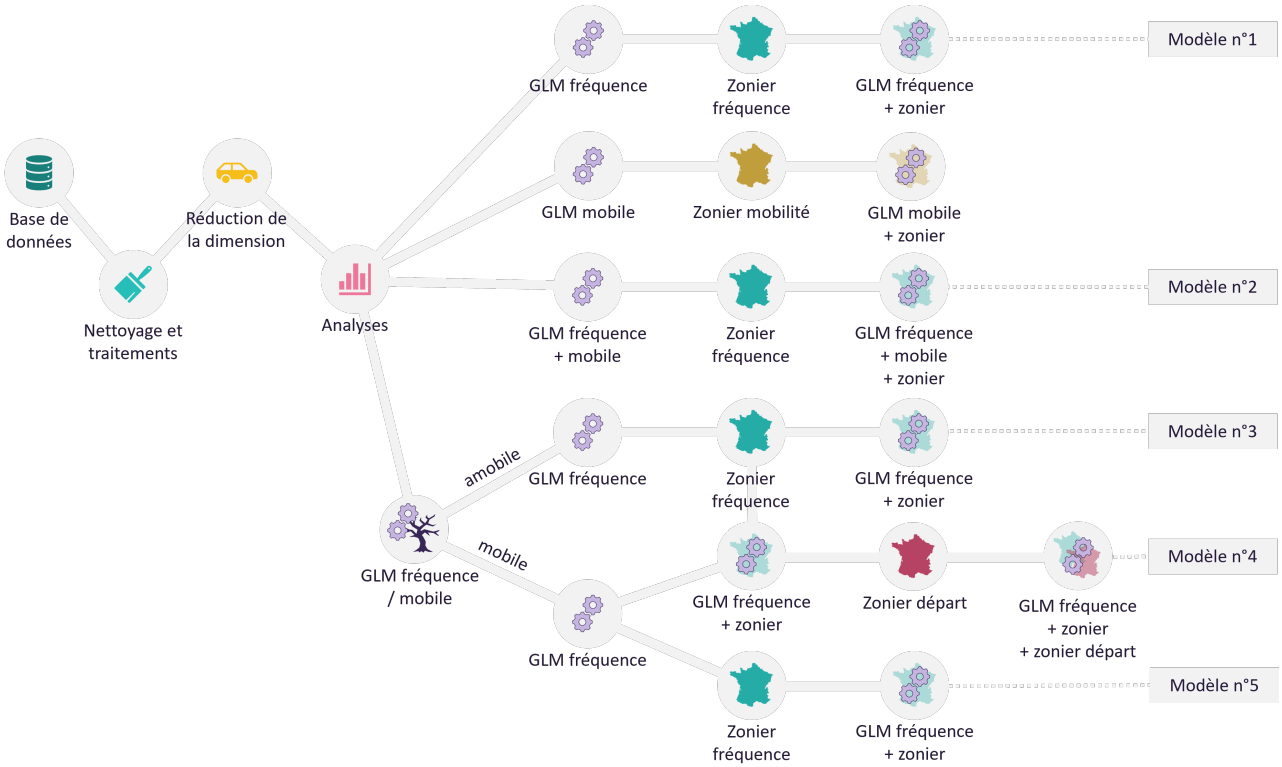
Le deuxième chapitre 2 propose un premier modèle de tarification qui ne prend pas en compte le phénomène de mobilité. La construction d'un tel modèle permet d'avoir une référence pour pouvoir évaluer les autres

modèles qui seront mis en place. C'est donc dans ce chapitre qu'est construit le premier zonier.

C'est dans le troisième chapitre 3 que le phénomène de mobilité sera directement intégré dans le modèle. Par ailleurs, une modélisation du risque de mobilité (probabilité qu'un assuré soit mobile) permettra d'obtenir un zonier appelé "zonier mobilité". Ce zonier permettra d'identifier les zones avec un fort taux de mobilité.

Enfin, le dernier chapitre 4 présente des modèles plus complexes afin de comprendre le phénomène. Une première segmentation entre les assurés mobiles et non mobiles est faite grâce à l'application d'un GLMtree. Deux modèles sont obtenus, et donc deux zoniers. Puis, un zonier dit "de provenance" est construit sur la base des assurés mobiles. Enfin, leur risque sera modélisé à travers les trois zoniers construits : le zonier des non mobiles, le zonier mobilité, et le zonier de provenance. Cette étude permettra d'interpréter la façon dont la mobilité impacte la perception du risque géographique des assurés mobiles.

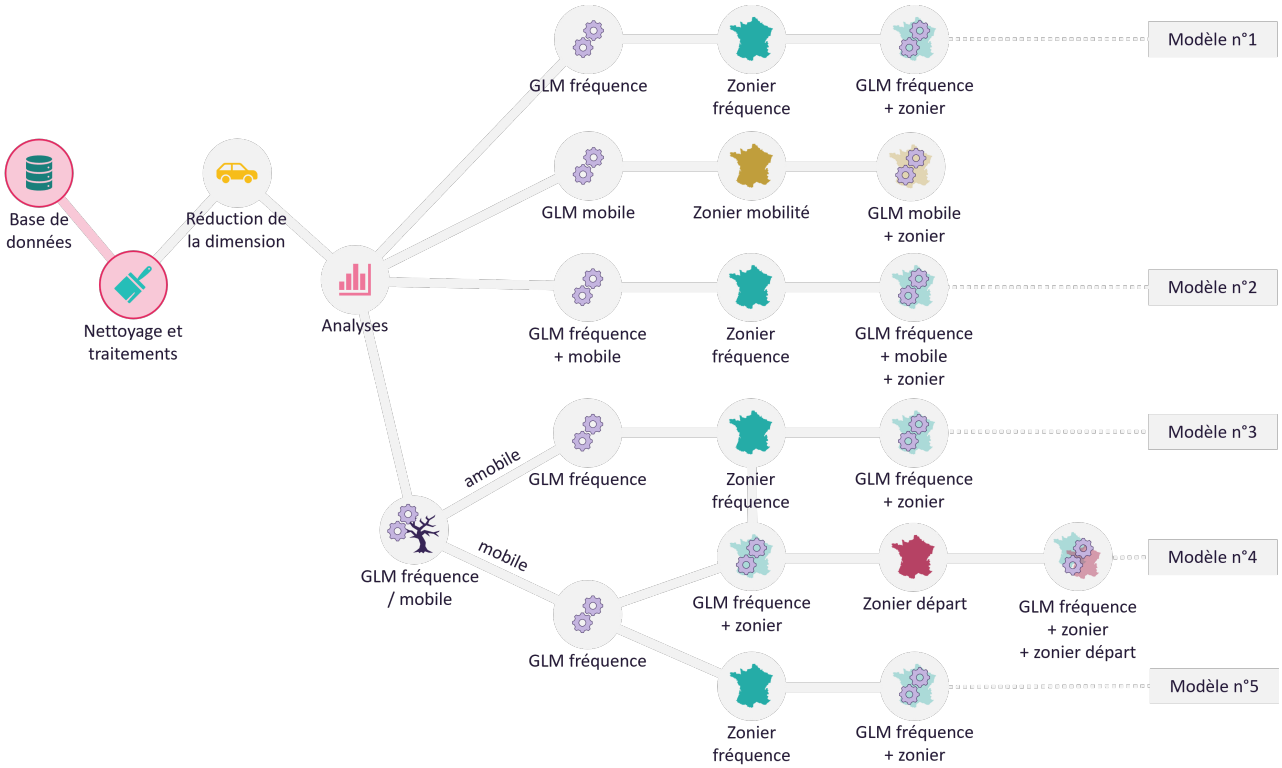
L'ensemble de ces étapes est représenté par le diagramme de suivi ci-dessous. Ce diagramme sera repris à chaque chapitre afin de positionner le point d'avancement de l'étude.



# Chapitre 1

# Tarification automobile chez les militaires : le cas de la mobilité

*L'objectif de ce chapitre est de dessiner progressivement le sujet de ce mémoire et sa problématique. La première partie introduit brièvement le secteur de l'assurance avant de se focaliser sur l'assurance automobile. La deuxième partie rappelle les grands principes de la tarification à travers le Machine Learning et la construction du zonier. La troisième partie sera dédiée à la présentation de la problématique et à la démarche de résolution.*



## 1.1 Introduction à l'assurance automobile

### 1.1.1 Présentation de l'activité d'assurance

#### Les organismes du secteur assurantiel

Un contrat d'assurance est un accord signé entre deux parties : l'assureur et l'assuré. En échange d'une somme versée par l'assuré, l'assureur s'engage à réaliser une prestation financière au profit d'un bénéficiaire en cas de réalisation d'un risque. La somme versée par l'assuré est appelée la prime. Trois organismes peuvent exercer l'activité d'assurance :

- **Les sociétés d'assurances** : régies par le CODE DES ASSURANCES (1930), elles se présentent généralement sous la forme de sociétés anonymes dont l'objectif est de faire du profit afin de redistribuer les bénéfices aux actionnaires.
- **Les mutuelles** : régies par le CODE DE LA MUTUALITÉ (1945) et le CODE DE LA SÉCURITÉ SOCIALE (1956), ce sont des sociétés de personnes à but non lucratif.
- **Les institutions de prévoyance** : régies par le CODE DE LA SÉCURITÉ SOCIALE (1956), elles couvrent principalement les risques de prévoyance et de santé.

#### Une activité réglementée

Tous ces organismes sont soumis au contrôle de l'AUTORITÉ DE CONTRÔLE PRUDENTIEL ET DE RÉSOLUTION (ACPR) (2010). Cette dernière est chargée de la supervision du secteur bancaire et de l'assurance. Elle veille notamment à ce que les compagnies soient en mesure de tenir leurs engagements envers les assurés à tout moment[1].

#### Une organisation par type de risques

L'activité d'assurance se divise généralement en deux grandes familles : l'assurance-vie et l'assurance non-vie (ou assurance dommage). Une partie importante de l'assurance non-vie est appelée l'assurance IARD (Incendie, Accidents et Risques Divers). Cette dernière regroupe l'ensemble des contrats d'assurance qui permettent à des particuliers, des entreprises ou d'autres entités de s'assurer contre des dommages qui ne relèvent pas de la vie humaine. Ce sont principalement les dommages matériels qui sont couverts, le reste étant dédié à l'assurance de personnes (maladie, épargne, capitalisation, etc.).

L'article R321-1 du CODE DES ASSURANCES (1930) définit plusieurs branches d'exercice dont l'agrément administratif est accordé par l'ACPR. L'assurance automobile correspond principalement aux branches 3 et 10.

### 1.1.2 Focus sur l'assurance automobile

#### La Responsabilité Civile automobile

En France, toute personne a l'obligation de réparer financièrement les dommages matériels et immatériels causés à autrui. C'est le principe juridique de la Responsabilité Civile (RC), défini par l'article 1240 du CODE CIVIL (1804). Ce principe s'applique notamment dans le cas des accidents automobiles. La garantie Responsabilité Civile automobile a pour but de protéger financièrement les conducteurs de véhicules et les tiers en cas de dommages causés par la conduite. En France, elle est obligatoire pour tous les propriétaires de véhicules. Selon l'article L211-1 du CODE DES ASSURANCES (1930) :

[1] Se référer aux articles L612-1 et L321-1 du Code des assurances.

*"Toute personne physique ou toute personne morale autre que l'État, dont la responsabilité civile peut être engagée en raison de dommages subis par des tiers résultant d'atteintes aux personnes ou aux biens dans la réalisation desquels un véhicule est impliqué, doit, pour faire circuler celui-ci, être couverte par une assurance garantissant cette responsabilité, dans les conditions fixées par décret en Conseil d'État."*

**Le Coefficient de Réduction-Majoration**

Une autre spécificité importante de l'assurance automobile est le coefficient de réduction-majoration (CRM), défini par l'article A121-1 du CODE DES ASSURANCES (1930), aussi appelé système bonus-malus. Il s'agit d'un coefficient qui va minorer ou majorer la prime d'assurance à chaque échéance annuelle en fonction du comportement de l'assuré. Sa prime sera réduite s'il n'a eu aucun sinistre durant l'année, et augmentée en fonction du nombre de sinistres ayant impliqué sa responsabilité. En notant $c_n$ le CRM d'un assuré à l'année $n$, le CRM de l'assuré pour l'année $n+1$ s'écrit :

$$c_{n+1} = \begin{cases} c_0 = 1 \\ c_{n+1} = \min\left(\max\left(1.25^k \cdot 1.125^p \cdot 0.95^{\mathbb{1}_{\{k+p=0\}}} \cdot c_n,\ 0.5\right),\ 3.5\right) \end{cases}$$

où $k$ est le nombre de sinistres responsables et $p$ le nombre de sinistres partiellement responsables. La première année, le CRM est de 1. Si l'assuré n'a subi aucun accident durant l'année, son CRM diminue de $5\,\%$. Sinon, il augmente de $25\,\%$ pour chaque sinistre responsable, et de $12{,}5\,\%$ pour chaque sinistre partiellement responsable. Enfin, ce coefficient est toujours compris entre $50\,\%$ et $350\,\%$.

**Les différentes garanties et formules**

Si la garantie RC automobile est obligatoire, d'autres garanties facultatives peuvent être proposées :

- **La garantie Vol** : couvre les dommages liés à une tentative de vol du véhicule (portière forcée, carreaux brisés, etc.).
- **La garantie Incendie** : couvre les dommages causés par un incendie ou une explosion.
- **La garantie Bris de glace** : couvre les dommages causés au pare-brise, aux vitres latérales, et aux lunettes arrière.
- **La garantie Dommage Tout Accident** : couvre tout dommage subi par le véhicule, y compris ceux dont l'assuré est responsable.

Ces garanties sont ensuite regroupées au sein de divers produits (ou formules). Ce sont ces produits qui seront proposés à l'assuré. Il a généralement le choix entre trois produits de souscription :

- **Formule Tiers** : niveau de protection minimal exigé (RC automobile).
- **Formule Tiers plus** : niveau de protection plus large que la formule Tiers, mais excluant les dommages dont l'assuré est lui-même responsable.
- **Formule Tout risque** : niveau de protection maximal. L'assuré est couvert de tous les risques, responsables ou pas.

La Table 1.1 résume l'ensemble de ces garanties et de ces formules. Toutes les garanties citées ne sont pas exhaustives. D'autres offres peuvent être proposées selon la politique de la compagnie. L'objet de ce mémoire

| Garantie \ Formule | Tiers | Tiers plus | Tout risque |
|---|---|---|---|
| RC automobile | ✓ | ✓ | ✓ |
| Vol | | ✓ | ✓ |
| Incendie | | ✓ | ✓ |
| Bris de glace | | ✓ | ✓ |
| Dommage tout accident | | | ✓ |

TABLE 1.1 : Garanties et formules de souscription classiques en assurance automobile

portera uniquement sur la **garantie Responsabilité Civile automobile**.

**Sinistralité des compagnies françaises**

Le rapport d'analyse 2022 de la FÉDÉRATION FRANÇAISE DE L'ASSURANCE (FFA) (2022) fournit des chiffres sur l'évolution de la sinistralité (en fréquence annuelle) des compagnies françaises. La Figure 1.1 est une reconstitution graphique de ces données par garantie.

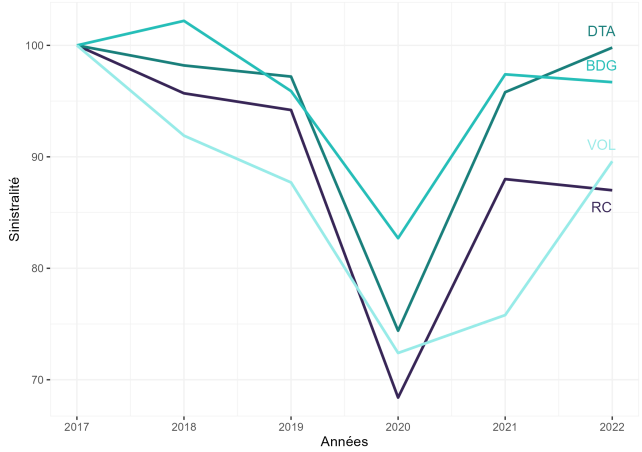



FIGURE 1.1 : Sinistralité (fréquence base 100 depuis 2017) des compagnies françaises par garantie automobile

Quelque soit la garantie, les compagnies françaises connaissent une tendance générale quasi à la baisse de la fréquence des sinistres[2]. Cette évolution est surtout marquée par le pic de sinistralité en 2020 dû à la pandémie du COVID, et plus précisément au confinement. Plusieurs facteurs peuvent potentiellement expliquer cette baisse de sinistralité, tels que l'entrée en vigueur du DÉCRET N°2018-487 (2018) sur la limitation de vitesse à 80 km/h.

Une attention sera donc portée sur l'évolution de la sinistralité dans le portefeuille. Si la tendance reflète

[2]Cependant, cela ne signifie pas que les coûts ont eux aussi diminué. Il s'avère qu'avec l'inflation post-COVID, la baisse de la fréquence des sinistres ne suffit pas à compenser l'augmentation des coûts.

celle observée à l'échelle macroscopique, il peut s'avérer pertinent de prendre en compte l'année calendaire dans le modèle de tarification.[3]

### 1.1.3 Présentation du portefeuille

#### Structure du jeu de données initial

Le jeu de données fourni est une table où chaque police est représentée sur une ou plusieurs lignes. Une ligne correspond à l'observation d'une police depuis le début de sa date d'observation jusqu'au minimum entre la date de fin d'observation et l'année comptable. Une colonne représente une caractéristique de l'assuré associé à la police. La structure chronologique d'une police dans la table est représentée sur la Figure 1.2.

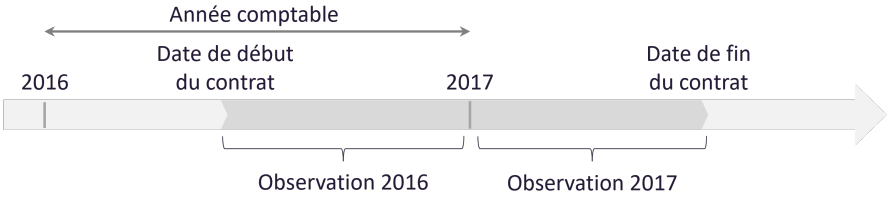



FIGURE 1.2 : Suivi d'une police dans la base de données

Lorsqu'un client souscrit[4] ou renouvelle son contrat en cours d'année, deux lignes seront présentes dans la base pour ce même individu. La première correspond à la période bornée par le début de la date d'observation (ou renouvellement) et la fin de l'année comptable. La seconde correspond à la période bornée par le début de l'année comptable suivante et la date de fin d'observation.

Les assurés ne sont donc pas nécessairement observés sur toute une année dans la base de données. Or, la prime d'assurance est construite sur l'historique du portefeuille de façon à couvrir les assurés sur une année entière. Il est donc primordial de prendre en compte la durée d'observation d'une police pour l'estimation des paramètres du modèle. De ce fait, un assuré ayant été observé sur une année entière a plus de poids dans la construction du modèle qu'un assuré ayant été observé sur la moitié de l'année. Cette durée d'observation est appelée l'exposition.

#### Choix des variables tarifaires

Une base de données a été récupérée. La dimension de la base est de plus de 4 millions de lignes pour 90 colonnes (soit 90 variables). Parmi les 90 variables extraites, 45 ont été sélectionnées pour l'étude[5]. Certaines variables sont des dates qui servent uniquement à formater et nettoyer la base. Les variables qui constituent des critères de tarification sont les suivantes. Voici une liste non exhaustive des variables utilisées dans le modèle de tarification :

- Usage (Qualitative à 5 modalités) : Utilisation du véhicule (*Promenade, Tous déplacements, ...*).
- CRM (Quantitative) : Coefficient de Réduction Majoration.
- nbcond_sec (Quantitative) : Nombre de conducteurs secondaires.

[3] Cette pratique permet, par exemple, pour la prédiction du coût moyen, de prendre en compte l'effet de l'inflation.
[4] C'est généralement le terme "affaires nouvelles" qui est utilisé pour les nouvelles polices du portefeuille.
[5] Certaines variables ne seront plus conservées après traitement et analyse

- AGE_COND (Quantitative) : Âge du conducteur principal.
- AGE_PER (Quantitative) : Âge du permis.
- CSP (Qualitative à 7 modalités) : Catégorie Socioprofessionnelle (*Militaire, Retraité, ...*).
- Groupe (Quantitative) : Groupe du véhicule (référence SRA).
- Classe (Qualitative à 16 modalités) : Classe du véhicule (référence SRA).
- AGE_VEH (Quantitative) : Âge du véhicule.
- Energie (Qualitative à 4 modalités) : Énergie d'alimentation du véhicule (*Essence, Électrique, ...*).
- Marque (Qualitative à 18 modalités) : Marque du véhicule.
- Valeur_veh (Qualitative à 18 modalités) : Valeur du véhicule.
- Vitesse (Qualitative à 18 modalités) : Vitesse maximale du véhicule.

**Traitements effectués**

Afin de préparer les données à l'apprentissage, les traitements suivants ont été effectués :

- **Traitement des expositions** : il a été vérifié que l'exposition calculée correspond bien à la période de segmentation de l'assuré. Cette exposition doit, en théorie, être inférieure ou égale à 1 et strictement supérieure à 0. Certaines lignes ont des expositions nulles, ce qui a permis de supprimer 20.23% des lignes du jeu de données.
- **Suppression des doublons** : lors de l'extraction, certains doublons ont pu apparaître. Cependant, seules une dizaine de lignes ont été supprimées.
- **Traitement des sinistres aux bords** : dans la base de données, la date de fin de segmentation correspond à la date de début de segmentation de l'observation suivante[6]. Ainsi, un sinistre survenant à la date de fin de segmentation apparaît en double. Il convient donc de supprimer ces doublons.
- **Traitement des anomalies** : certaines anomalies ont été repérées, telles que des coefficients de réduction-majoration inférieurs à 0.5. Les lignes correspondantes ont été supprimées, soit une centaine de lignes. Les variables `Valeur_veh` et `Vitesse` présentent de nombreuses valeurs aberrantes, notamment des 0. L'analyse des corrélations (Chapitre 2) montrera que ces variables sont fortement corrélées à d'autres et ne seront donc pas conservées.
- **Valeurs manquantes** : certaines variables contiennent des valeurs manquantes. Elles ont été simplement remplacées par 0.

## 1.2 Principe de tarification

La prime est le montant que l'assuré paie à l'assureur afin d'être couvert contre ses éventuels sinistres. La prime payée est appelée la prime commerciale. Son montant correspond à la somme de trois composantes :

[6] sauf si la date de fin de segmentation est au 31/12. Dans ces cas, la date de début de segmentation est bien le 01/01.

(i) La prime pure : elle représente le risque de l'assuré et sert à couvrir le montant de ses dommages potentiels sur l'année.

(ii) Les chargements d'acquisition et de gestion : ils servent à couvrir les frais de la compagnie (gestion des contrats, infrastructures, etc.).

(iii) Le chargement de sécurité : il représente une marge permettant à l'assureur de se couvrir contre le risque de souscription, c'est-à-dire qu'il permet de couvrir une éventuelle mauvaise tarification ou une hausse imprévue de la sinistralité.

Ce mémoire se focalise sur la prime pure, c'est-à-dire sur la modélisation de la sinistralité de l'assuré. Le calcul de cette prime repose sur le modèle coût-fréquence. Plus précisément, c'est le modèle de fréquence qui sera considéré ici.

### 1.2.1 Modèle coût-fréquence : entre segmentation et mutualisation

**Le modèle coût-fréquence**

Soit un assuré dont les caractéristiques sont représentées par un vecteur $x$. Ces caractéristiques peuvent être son âge, la classe de son véhicule, ou encore sa catégorie socioprofessionnelle. Pour $p \in \mathbb{N}^*$, $x \in \mathbb{R}^p$. Soit $N_x \in \mathbb{N}^*$ son nombre de sinistres annuel et $C_x^1, \dots, C_x^{N_x} > 0$ les coûts de chacun de ses $N_x$ sinistres. Soit $S_x$ sa charge totale de sinistres annuelle. Avec la convention $\sum_{n=1}^{0} C_x^n = 0$, les quantités définies précédemment sont liées par la formule suivante :

$$S_x = \sum_{n=1}^{N_x} C_x^n.$$

Soit $\pi_x^*$ la prime pure de l'assuré. Comme elle représente son risque, elle doit correspondre au mieux à sa charge totale de sinistres annuelle, afin de lui affecter la prime la plus "juste". Généralement, la prime pure répond au problème d'optimisation quadratique suivant :

$$\pi_x^* \in \arg\min_{\pi_x \in \mathbb{R}} \{\mathbb{E}[(S_x - \pi_x)^2]\}.$$

La résolution de ce problème[7] conduit à affecter à l'assuré la prime pure $\pi_x^*$ telle que $\pi_x^* = \mathbb{E}[S_x]$. Le calcul de cette quantité s'effectue ensuite sur la base de deux hypothèses importantes :

(i) Les charges de sinistres unitaires sont indépendantes et identiquement distribuées.

(ii) La charge de sinistres est indépendante du nombre de sinistres survenus.

Mathématiquement, cela revient à écrire $(C_x^n)_{n \in \mathbb{N}^*} \overset{\text{iid}}{\sim} C_x$ et $C_x \perp\!\!\!\perp N_x$. Sous ces conditions, la prime pure se décompose alors comme le produit de deux espérances[8] :

$$\pi_x^* = \mathbb{E}[S_x] = \mathbb{E}[N_x]\mathbb{E}[C_x].$$

[7] Démonstration en Annexe A.1.

[8] Démonstration en Annexe A.2.

C'est le modèle coût-fréquence, qui doit son nom à cette décomposition de la prime pure comme le produit de la fréquence[9] annuelle de sinistres $\mathbb{E}[N_x]$ et la charge de sinistres unitaire $\mathbb{E}[C_x]$. Il est alors possible de modéliser séparément la fréquence et la charge.

Si $N_x$ est la variable aléatoire représentant le nombre de sinistres annuels d'un assuré ayant des caractéristiques $x$, alors le résultat obtenu se généralise pour toutes les caractéristiques $X \in \mathbb{R}^p$ présentes dans le portefeuille, en notant $N_x$ comme étant $N|X = x$. Le même raisonnement s'applique à la charge unitaire. En conclusion, l'expression de la prime pure d'un assuré dans le portefeuille ayant des caractéristiques $x$ est :

$$\pi_x^* = \mathbb{E}[N|X = x]\mathbb{E}[C|X = x].$$

La méthode généralement employée par les assureurs pour estimer ces quantités est l'utilisation des *Modèles Linéaires Généralisés* (GLM).

**Segmentation et mutualisation**

La prime pure déterminée précédemment se présente sous la forme d'un produit d'espérances conditionnelles. Si les assurés ne sont pas discriminés selon leurs caractéristiques, alors l'ensemble du portefeuille constitue un seul groupe, et tous les assurés se verront affecter la même prime pure, notée $\pi^* = \mathbb{E}[N]\mathbb{E}[C]$. Il est alors possible d'estimer chacune de ces quantités en utilisant la *Loi (faible) des grands nombres* :

*Soient* $X_1, \dots, X_n \overset{\text{iid}}{\sim} X$ *tels que* $\mathbb{E}[X] < +\infty$. *Alors la moyenne empirique de l'échantillon converge en probabilité vers l'espérance de la loi :*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\mathbb{P}} \mathbb{E}[X].$$

C'est le principe de *mutualisation* : les risques de tous les assurés sont mutualisés afin de se compenser. La prime payée est donc une moyenne des risques présents dans le portefeuille. Le problème d'une mutualisation parfaite est que certains assurés, appelés les "bons risques", se retrouvent à payer une prime plus élevée que celle qui correspond réellement à leur risque. Inversement, les "mauvais risques" payent une prime inférieure à celle qui représente leur risque réel. En raison de la compétitivité des compagnies d'assurances, les bons risques seront attirés par des offres plus avantageuses d'autres compagnies. La compagnie n'est alors plus en mesure de compenser ses mauvais risques par les bons risques.

Afin de remédier à ce phénomène, il est nécessaire de segmenter le portefeuille en plusieurs classes de risques, ou groupes homogènes, et d'opérer ainsi la mutualisation au sein de chaque groupe. C'est le principe de *segmentation*.

La segmentation apparaît dans la formule de la prime pure sous forme d'une espérance conditionnée à un groupe d'appartenance, noté ici $X$. Cependant, une segmentation parfaite, où chaque assuré est l'unique individu de son groupe, n'est pas souhaitée. En effet, un faible nombre de représentants au sein d'un groupe ne permet pas de rendre l'estimation du risque suffisamment robuste.

### 1.2.2 Utilisation du Machine Learning en assurance

Le *Machine Learning* est un domaine vaste, toujours en pleine évolution et faisant l'objet de nombreuses recherches scientifiques. Seuls les concepts élémentaires seront abordés ici. Le lecteur intéressé est invité à se

[9] En réalité, parler de fréquence est un abus de langage. La quantité $\mathbb{E}[N_x]$ n'est pas une fréquence, mais un nombre de sinistres espéré sur une période de temps fixée à 1 an.

référer à JAMES et al. (2013) pour approfondir le sujet.

Pour rappel, la quantité à estimer est $\mathbb{E}[N|X = x]$. Cette quantité sera modélisée par un *Modèle Linéaire Généralisé* (GLM). Ce modèle fait partie des méthodes d'apprentissage statistique, et plus largement des méthodes de *Machine Learning*. Il est donc important de rappeler les fondements de ces méthodes et leur cadre d'application.

**Définition d'un modèle de Machine Learning**

Soit $Y$ la variable à prédire et $X$ l'ensemble des facteurs explicatifs qui servent à la prédiction. Soit $(Y, X)$ un couple de variables aléatoires de loi jointe inconnue $\mathcal{P}$ sur $\mathcal{Y} \times \mathcal{X}$, où $\mathcal{Y} \subset \mathbb{R}$ et $\mathcal{X} \subset \mathbb{R}^p$.

Un modèle est une fonction $f : \mathcal{X} \to \mathcal{Y}$, appelée prédicteur, qui lie $Y$ à $X$. La qualité de la prédiction (ou la performance) est mesurée par son erreur de généralisation, qui correspond à l'écart moyen entre $Y$ et $f(X)$. Cet écart est quantifié par une fonction $l : Y, f(X) \mapsto l(Y, f(X))$ appelée fonction coût. Ainsi, l'erreur de généralisation d'un modèle $f$ s'écrit :

$$R_{\mathcal{P}}(f) = \mathbb{E}_{(Y,X)\sim\mathcal{P}}[l(Y, f(X))].$$

Il est alors possible de définir le meilleur modèle $f^*$, appelé prédicteur optimal (ou oracle), comme la fonction qui minimise l'erreur de généralisation sur l'ensemble[10] $\mathcal{F}$ de tous les prédicteurs possibles :

$$f^* \in \arg\min_{f\in\mathcal{F}}\{R_{\mathcal{P}}(f)\}.$$

Dans le cas où $Y$ est une variable numérique, la fonction coût généralement utilisée est la fonction de perte quadratique $l : Y, f(X) \mapsto (Y - f(X))^2$. Il est alors possible de montrer[11] que le prédicteur optimal est la fonction :

$$f^* : x \mapsto \mathbb{E}[Y|X = x].$$

Il est supposé qu'il existe une fonction $f^*$ telle que $Y = f^*(X) + \epsilon$, de sorte que l'écart moyen entre $Y$ et $f^*(x)$ soit minimal, où $\epsilon$ représente l'ensemble des facteurs non observés ou qui viennent détériorer la mesure de $Y$. La Figure 1.3 permet d'illustrer ce résultat géométriquement.

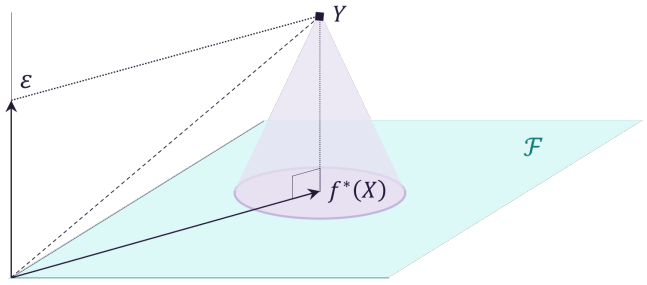



FIGURE 1.3 : Illustration du prédicteur optimal comme projection orthogonale

[10] $\mathcal{F} = \{f : \mathbb{E}[f(X)^2] < +\infty\}$
[11] Démonstration en Annexe A.3.

Si les variables aléatoires sont des vecteurs appartenant à un espace vectoriel (hilbertien), alors $f^*(X)$ est la projection de $Y$ sur le plan de l'image de $X$ par $f$. De ce fait, $f^*$ est la meilleure approximation de $Y$ sachant l'information $X$ disponible.

En conclusion, l'intérêt du *Machine Learning* est de construire une fonction $f$ qui se rapproche au mieux d'une espérance conditionnelle. Dans le cadre actuariel, $Y$ peut jouer le rôle du nombre de sinistres annuels ou bien celui de la charge de sinistres unitaire. Ici, il s'agira du nombre de sinistres.

Dans le cas où $Y$ est une variable qualitative à $K$ classes, la fonction coût à minimiser est généralement la fonction $l : y, x \mapsto -\sum_{k=1}^{K} \mathbb{1}_{y=k} \log(f_k(x))$ où $f_k(x) = \mathbb{P}(Y = k|X = x)$. Elle est appelée la fonction de cross-entropy (inverse de la log-vraisemblance). Dès lors, il est possible de montrer que le prédicteur optimal est la fonction $f^*$ telle que

$$\mathbb{P}(Y = f^*(x)|X = x) = \max_{y \in \mathcal{Y}} \{\mathbb{P}(Y = y|X = x)\}.$$

Cette fonction $f^*$ est par ailleurs appelée la règle de Bayes. En somme, pour un individu présentant des caractéristiques $x$, la probabilité d'appartenance à chacune des classes de $Y$ est calculée et la règle de prédiction optimale consiste à affecter à l'individu la classe avec la plus grande probabilité d'appartenance.

**Construction d'un modèle et évaluation**

Les résultats précédents justifient l'intérêt du *Machine Learning* en assurance : l'objectif est de construire des modèles qui estiment la valeur d'une espérance conditionnelle.

Cependant, l'expression de $f^*$ n'est pas toujours connue, à moins de se restreindre à certaines hypothèses sur la loi de $Y|X$. Il s'agit de modèles paramétriques[12]. C'est le cas des *Modèles Linéaires Généralisés*. Seulement, l'intérêt du *Machine Learning* est de pouvoir construire des modèles qui ne se fondent pas nécessairement sur une loi statistique. Il s'agit alors de modèles non-paramétriques.

Soit $f$ le modèle à estimer, et $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ un échantillon suivant la distribution $\mathcal{P}$. À partir de cet échantillon, le modèle $f$ est estimé en minimisant le risque empirique (ou erreur d'entraînement), qui est une estimation de l'erreur de généralisation théorique :

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i)).$$

D'après la Loi des Grands Nombres, plus la taille de l'échantillon augmente, plus le risque empirique approche l'erreur de généralisation théorique.

Néanmoins, le but d'un modèle de *Machine Learning* n'est pas uniquement l'estimation du modèle. C'est aussi et surtout la prédiction. Un modèle de *Machine Learning* s'évalue donc sur deux critères : la qualité de l'estimation et la capacité de prédiction. Il y a donc deux étapes importantes lors de la construction d'un modèle : l'apprentissage et l'évaluation. La méthodologie est la suivante :

1. Division de l'échantillon récolté $\mathcal{D}_n$ en un échantillon d'apprentissage $\mathcal{D}_n^{\text{app}}$ et un échantillon de test $\mathcal{D}_n^{\text{test}}$. Il faut bien veiller à ce que les échantillons d'apprentissage et de test soient disjoints. L'algorithme ne doit en aucun cas être construit sur la base d'observations qu'il sera amené à prédire.
2. Apprentissage de l'algorithme à l'aide de l'échantillon d'apprentissage $\mathcal{D}_n^{\text{app}}$ en minimisant le risque empirique $R_n$. Il est alors possible d'évaluer la qualité de l'estimation du modèle.

[12] Les modèles paramétriques définissent généralement des fonctions dont la forme est fixe, avec des coefficients à estimer. Une hypothèse est faite sur la loi de distribution du jeu de données. Les modèles non-paramétriques ne font aucune hypothèse quant à la distribution des données ou l'expression de la fonction. Ils sont construits en se basant sur la structure des données à disposition.

3. Évaluation de la capacité de prédiction à l'aide de l'échantillon de test $\mathcal{D}_n^{\text{test}}$. L'algorithme est appliqué sur l'échantillon de test et les sorties sont comparées avec les valeurs attendues.

Afin d'évaluer le modèle, que ce soit à l'entraînement ou au test, plusieurs critères peuvent être choisis. Seuls les deux critères suivants seront utilisés :

- **La Mean Absolute Error (MAE)** : elle mesure l'écart moyen en valeur absolue entre les valeurs cibles et les valeurs prédites. La formule est donnée par

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|.$$

- **Mean Square Error (MSE)** : elle mesure l'écart moyen au carré. Cette formule donne plus de poids aux mauvaises prédictions dans l'évaluation de l'erreur. La formule est donnée par

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

D'autres critères existent tels que la RMSE (Root Mean Square Error), la RMSLE (Root Mean Squared Log Error) ou encore la MAPE (Mean Absolute Percentage Error). La RMSE sera aussi utilisée et correspond simplement à la racine carrée de la MSE.

**Le dilemme biais-variance**

Afin d'illustrer le concept de biais et de variance en *Machine Learning*, un exemple basé sur le modèle polynomial est utilisé. L'hypothèse faite est

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + ... + \theta_p X^p + \epsilon.$$

C'est un modèle paramétrique où les coefficients $\theta_i$ sont à estimer à partir de l'échantillon d'apprentissage. Le degré du polynôme est donné par la valeur de $p$. En entraînant le modèle pour différents $p$, les résultats obtenus sont ceux présentés sur la Figure 1.4.

Plus le degré du polynôme augmente, plus le modèle explique bien les données qui lui servent d'entraînement. Le modèle explique parfaitement les données d'entraînement lorsque le degré du polynôme est égal à la taille de l'échantillon. Ce dernier cas correspond simplement à une interpolation. Néanmoins, lorsque le modèle est appliqué pour de nouvelles observations, des performances plus nuancées sont constatées.

Lorsque le modèle est trop simple (degré faible), naturellement, il n'arrivera pas à performer face à de nouvelles données. Il s'agit d'un cas de sous-apprentissage. À l'inverse, lorsque le modèle est trop complexe, il s'accommode si bien aux données qui lui ont servi d'entraînement qu'il perd en capacité de généralisation. Dans ce cas, il s'agit de surapprentissage. Cette idée est illustrée sur la Figure 1.5. Il est donc possible de distinguer deux sources d'erreurs qui empêchent les modèles de *Machine Learning* de se généraliser au-delà de l'échantillon d'apprentissage :

(i) Le biais : les choix sur les hypothèses peuvent parfois être trop simplistes (modèle peu complexe ou structure pas adaptée) et éloignent donc le modèle du modèle optimal (sous-apprentissage).

(ii) La variance : à vouloir être trop complexe, le modèle apprend les erreurs induites par le bruit et devient sensible aux plus petites fluctuations (sur-apprentissage). Le modèle devient alors très sensible au jeu d'apprentissage, c'est-à-dire qu'une faible variation de la base d'apprentissage modifie très fortement le modèle.

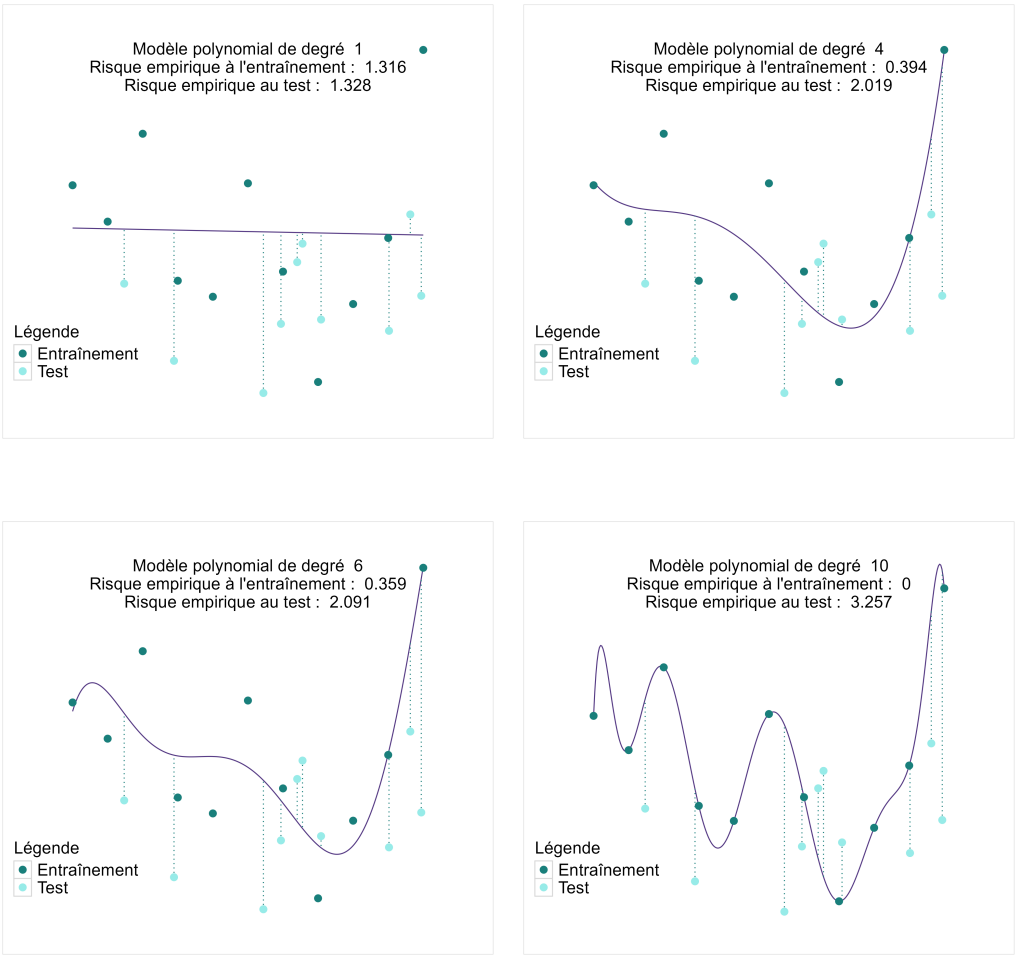



FIGURE 1.4 : Application d'un modèle polynomial pour différents degrés sur une base fictive

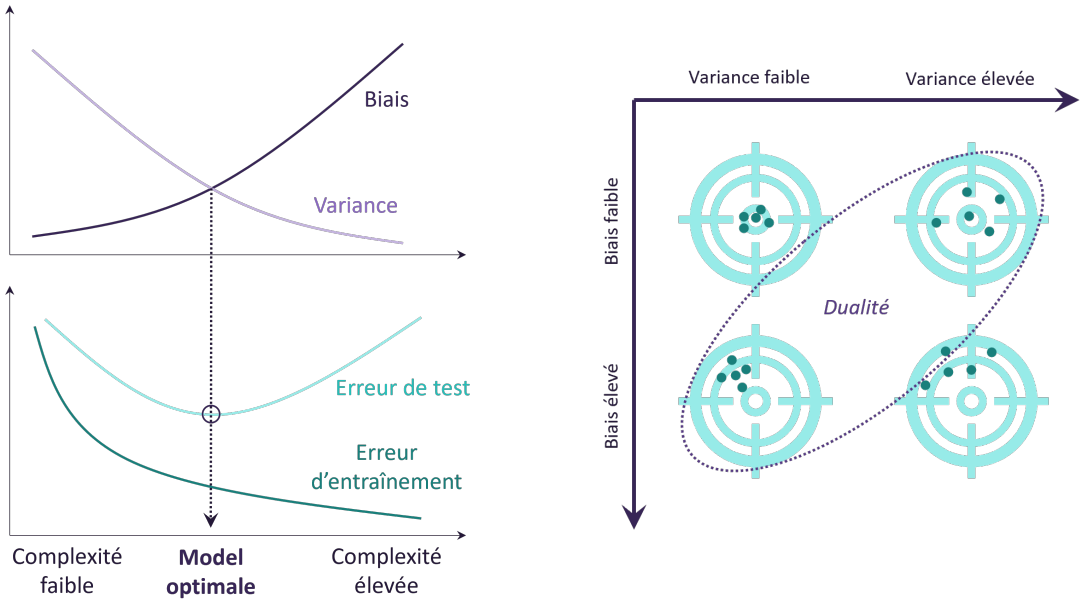



FIGURE 1.5 : Illustration du dilemme biais variance

Ce dilemme biais-variance se formalise mathématiquement à travers la relation suivante

$$R_{\mathcal{P}}(\hat{f}_F) - R_{\mathcal{P}}(f^*) = \underbrace{R_{\mathcal{P}}(\hat{f}_F) - \inf_{f \in F}\{R_{\mathcal{P}}(f)\}}_{\text{Erreur d'estimation (variance)}} + \underbrace{\inf_{f \in F}\{R_{\mathcal{P}}(f)\} - R_{\mathcal{P}}(f^*)}_{\text{Erreur d'approximation (biais)}}.$$

Le biais est une mesure de la justesse du modèle tandis que la variance est une mesure de la sensibilité du modèle aux données d'entraînement.

**Sélection des modèles**

Soit une collection de modèles $\mathcal{C} = \{F_1, ..., F_J\}$ et leur prédicteur respectif $f_{F_1}, ..., f_{F_J}$. La question est de savoir comment déterminer le meilleur modèle $F^*$.
En ayant estimé les prédicteurs $f_{F_1}, ..., f_{F_J}$ par $\hat{f}_{F_1}, ..., \hat{f}_{F_J}$, le meilleur modèle $\hat{F}^*$ est celui qui vérifie

$$\hat{F}^* \in \arg\min_{F \in \mathcal{C}} \left\{ R_n(\hat{f}_F) + \text{pen}(F) \right\}.$$

Un terme de pénalisation a été ajouté à l'estimation du risque empirique. La définition explicite du terme de pénalisation fait encore l'objet de diverses recherches. En outre, ce terme est lié à la complexité de l'algorithme (par exemple, au degré du polynôme dans l'exemple précédent). En minimisant uniquement le risque empirique, le critère tend à choisir le modèle ayant le biais le plus faible, c'est-à-dire le modèle le plus complexe. D'un autre côté, en minimisant le terme de pénalité, le critère tend à choisir le modèle le moins complexe, c'est-à-dire celui dont la variance est la plus faible. L'ajout d'un terme de pénalité permet donc de faire le compromis entre le biais et la variance. Il existe principalement deux critères qui permettent d'évaluer les modèles selon ce principe.

- Le critère AIC (AKAIKE (1974)) : la formule est donnée par

  $$\text{AIC} = -2\ln(\mathcal{L}) + 2p,$$

  où la quantité $\mathcal{L}$ est la vraisemblance du modèle, $n$ est la taille de l'échantillon, et $p$ la complexité du modèle liée au nombre de paramètres estimés. La quantité $-2\ln(\mathcal{L})$ est appelée la déviance du modèle. Lorsque les paramètres sont estimés, la vraisemblance du modèle est maximisée, ce qui revient à minimiser la déviance. Elle fait donc office de risque de généralisation.

- Le critère BIC (SCHWARZ (1978)) : la formule est donnée par

  $$\text{BIC} = -2\ln(\mathcal{L}) + \log(n)p.$$

  La différence du critère BIC par rapport au critère AIC est que le coefficient 2 est remplacé par le logarithme de la taille de l'échantillon pour le terme de pénalité. Ainsi, le critère BIC tend à pénaliser les modèles complexes plus sévèrement que le critère AIC à mesure que la taille de l'échantillon augmente.

En revanche, il est parfois difficile de définir la vraisemblance d'un modèle, notamment dans le cas non-paramétrique. Une façon empirique d'évaluer et de comparer des modèles est d'appliquer une validation croisée.

**La validation croisée**

La qualité de l'estimation d'un modèle peut être sensible à l'échantillon d'apprentissage choisi. Avec une autre division du jeu de données en un échantillon d'apprentissage et de test, les résultats ne seraient pas les mêmes. La validation croisée (KOHAVI (1995)) permet d'évaluer la performance des modèles selon l'échantillon d'apprentissage considéré, et donc leur robustesse. Les étapes de la validation croisée, dite *K-fold cross validation*, sont les suivantes :

1. Division de la base de données en $K$ blocs disjoints.
2. Exécution de $K$ processus indépendants :
   (a) Pour chacun de ces processus, apprentissage du modèle sur $K-1$ blocs.
   (b) Test sur le bloc qui n'a pas servi à l'apprentissage.

Finalement, $K$ estimations différentes de l'erreur (à l'entraînement et au test) sont obtenues pour un modèle donné. Puis, l'analyse de la distribution de l'erreur de test permet de visualiser la capacité de généralisation du modèle.

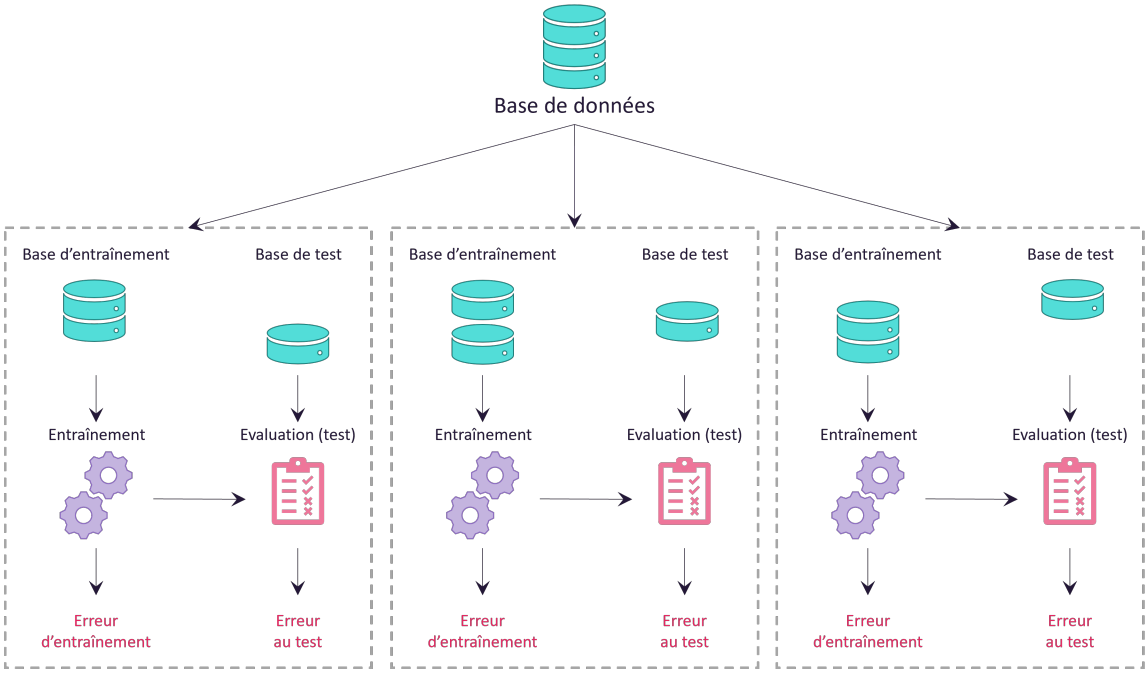



FIGURE 1.6 : Processus de la validation croisée

Il existe plusieurs types de validation croisée selon la façon dont la division de la base initiale est faite. Généralement, la méthode la plus utilisée est la *K-fold*. C'est l'exemple de la Figure 1.6. Elle consiste à diviser la base initiale en $K$ paquets de même taille (appelés *fold*) et d'opérer à chaque fois l'apprentissage sur $K-1$ paquets puis de tester sur le paquet restant.

Un autre type de validation croisée est le *Leave-One-Out*. Il s'agit simplement de mettre de côté à chaque fois un seul individu au lieu d'un ensemble d'individus. Si l'objectif est de tester la robustesse du modèle dans une vision temporelle, une *Time Series Cross-Validation* peut être appliquée. Il s'agit d'une validation croisée dont le découpage respecte la chronologie des données.

### 1.2.3 Les modèles linéaires généralisés

Les *Modèles linéaires généralisés* (MCCULLAGH et NELDER (1989)), notés GLM, sont les modèles de base en tarification. La littérature regorge de documents à ce sujet, ils sont surtout utilisés pour leur interprétabilité.

**Définition du *Modèle Linéaire Généralisé***

Rappelons que l'objectif en *Machine Learning* est d'estimer $f^*$ où $f^*(x) = \mathbb{E}[Y|X = x]$. Les GLM représentent une famille de modèles paramétriques où $Y$ est supposé appartenir à une famille exponentielle dont les paramètres dépendent des facteurs explicatifs $X$ appelés les régresseurs.

*Soit $Y$ une variable aléatoire unidimensionnelle. La variable $Y$ appartient à une famille exponentielle si la densité de la loi de $Y$ s'écrit*

$$f_Y(y, \omega, \phi) = \exp\left(\frac{y\omega - b(\omega)}{\gamma(\phi)} - c(y, \phi)\right),$$

*où $\omega$ est appelé le paramètre naturel de la distribution et $\phi$ le paramètre de dispersion.*
Formellement, l'écriture la plus adaptée serait $Y|X = x \sim \mathcal{F}_{Exponentielle}(\omega(x), \phi(x))$.

Dans le cas où $Y$ appartient à une famille exponentielle, il est possible de montrer que $\mathbb{E}[Y|X = x] = b'(\omega(x))$ et $\mathbb{V}[Y|X = x] = b''(\omega)\gamma(\phi(x))$.

Remarquons que le paramètre naturel peut alors s'écrire comme une fonction de l'espérance de $Y|X = x$

$$\omega(x) = b'^{-1}(\mathbb{E}[Y|X = x]) = g(\mathbb{E}[Y|X = x]),$$

où la fonction $g : x \mapsto b'^{-1}(x)$ qui associe l'espérance de $Y|X = x$ au paramètre naturel $\omega$ est appelée la fonction de lien. La fonction $g$ se doit d'être bijective sur l'espace de définition afin de pouvoir être inversible. Finalement, les modèles linéaires généralisés sont construits à partir de trois composantes :

(i) une composante aléatoire $Y$ appartenant à une famille exponentielle ;

(ii) une composante fixe $X$ qui sont les régresseurs du modèle (ou facteurs explicatifs) ;

(iii) une fonction lien $g$ liant l'espérance de $Y|X$ et la composante fixe $X$.

La Table 1.2 répertorie quelques distributions appartenant à la famille exponentielle et les liaisons entre les différentes composantes.

| Distribution | $\omega$ | $\gamma(\phi)$ | $g(x)$ | $\mathbb{E}[Y\|X = x]$ | $\mathbb{V}[Y\|X = x]$ |
|---|---|---|---|---|---|
| Gaussienne $\mathcal{N}(\mu, \sigma^2)$ | $\mu$ | $\phi = \sigma^2$ | $x$ | $\mu = \omega(x)$ | $\sigma^2$ |
| Bernouilli $\mathcal{B}(p)$ | $\ln\left(\frac{p}{1-p}\right)$ | 1 | $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$ | $p = \frac{e^{\omega(x)}}{1+e^{\omega(x)}}$ | $p(1-p)$ |
| Poisson $\mathcal{P}(\lambda)$ | $\ln(\lambda)$ | 1 | $\ln(x)$ | $\lambda = e^{\omega(x)}$ | $\lambda$ |
| Gamma $\mathcal{G}(\mu, \nu)$ | $-\frac{1}{\mu}$ | $\frac{1}{\nu}$ | $-\frac{1}{x}$ | $\mu = -\frac{1}{\omega(x)}$ | $\frac{\mu^2}{\nu}$ |
| Inverse Gamma $\mathcal{IG}(\mu, \sigma^2)$ | $-\frac{1}{2\mu^2}$ | $\sigma^2$ | $-\frac{1}{2x^2}$ | $\mu = \frac{1}{\sqrt{-2\omega(x)}}$ | $\mu^3\sigma^2$ |

TABLE 1.2 : Relations entre les différentes composantes d'un GLM

En réalité, n'importe quelle fonction lien peut être utilisée pour une loi donnée, sous certaines conditions de régularité avec la variable cible. Néanmoins, il est préférable d'utiliser les fonctions de lien canonique qui possèdent des propriétés mathématiques intéressantes.

**Les GLM : une extension du modèle linéaire**

Dans le cadre du modèle linéaire, l'hypothèse faite sur la variable cible est

$$Y = X\theta + \epsilon,$$

où $\epsilon \sim \mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}^p$, ce qui implique que $Y \sim \mathcal{N}(X\theta, \sigma^2)$. Il advient alors que $\mathbb{E}[Y|X = x] = x\theta$. Par identification, c'est le paramètre naturel de la distribution qui s'exprime comme une combinaison linéaire des régresseurs, c'est-à-dire

$$\omega(x) = \langle x, \theta \rangle.$$

L'idée des GLM est donc d'étendre la relation entre le paramètre naturel, exprimé comme une combinaison linéaire des régresseurs, et l'espérance de la loi, à d'autres distributions que la gaussienne (à savoir celles appartenant à la famille exponentielle). Cette généralisation est faite à travers la fonction lien.

**Estimation des paramètres**

Les GLM sont des modèles paramétriques, c'est-à-dire que l'apprentissage se fait en estimant les paramètres du modèle, à savoir ici $\theta$. Cette estimation se fait par maximum de vraisemblance : la log-vraisemblance du modèle est à maximiser.
Soit un échantillon de taille $n$. Pour tout individu $i = 1, ..., n$,

$$\begin{aligned}
&Y_i|X = x_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{F}_{Exponentielle}(\omega(x_i), \phi(x_i)),\\
&g(\mathbb{E}[Y_i|X = x_i]) = \omega(x_i)\\
&\omega(x_i) = \theta_0 + \theta_1 x_{i,1} + \cdots + \theta_p x_{i,p}.
\end{aligned}$$

Dans la mesure où les observations sont indépendantes, et en notant $(y_1, x_1), \ldots, (y_n, x_n)$ les $n$ observations, la vraisemblance du modèle est la fonction $\mathcal{L}$ telle que

$$\mathcal{L}(y_1, \ldots, y_n; \theta) \overset{\perp\!\!\!\perp}{=} \prod_{i=1}^{n} f_{Y_i}(y_i, \omega_i, \phi_i; \theta).$$

La log-vraisemblance est la fonction $l$ telle que

$$\begin{aligned}
l(y_1, \ldots, y_n; \theta) &:= \ln(\mathcal{L}(y_1, \ldots, y_n; \theta))\\
&= \sum_{i=1}^{n} \ln(f_{Y_i}(y_i, \omega_i, \phi_i; \theta))\\
&:= \sum_{i=1}^{n} \left( \frac{y_i \omega(x_i; \theta) - b(\omega(x_i; \theta))}{\gamma(\phi(x_i))} - c(y_i, \phi(x_i)) \right).
\end{aligned}$$

L'objectif est donc de trouver $\hat{\theta}_{MV}$, l'estimateur de maximum de vraisemblance de $\theta$, tel que

$$\hat{\theta}_{MV} \in \arg\max_{\theta \in \mathbb{R}^{p+1}} \{l(Y_1, \ldots, Y_n; \theta)\}.$$

En notant le score

$$S(Y_1, \ldots, Y_n; \theta) := \left( \frac{\partial l}{\partial \theta_0}(Y_1, \ldots, Y_n; \theta), \ldots, \frac{\partial l}{\partial \theta_p}(Y_1, \ldots, Y_n; \theta) \right)^T,$$

l'estimateur de maximum de vraisemblance vérifie

$$S(Y_1, \ldots, Y_n; \hat{\theta}_{MV}) = 0_{(p+1)}.$$

Cependant, il n'existe pas dans le cas général de formule fermée[13] pour l'estimation de $\theta$. Ainsi, l'estimation repose très souvent sur des algorithmes d'optimisation, de type Iterative Reweighted Least Squares (GREEN (1984)). L'estimation de $\theta$ par une formule fermée fait l'objet de diverses recherches. À titre d'exemple, BROUSTE et al. (2020) proposent une formule fermée pour l'estimation de $\theta$ dans des cas spécifiques.

**Cas de l'offset**

L'offset est un terme ajouté à l'expression du paramètre naturel et dont le coefficient multiplicateur vaut toujours 1 :

$$\omega(x) = \langle x, \theta \rangle + \text{offset}.$$

Pour estimer la fréquence annuelle de sinistres d'un assuré, le nombre de sinistres survenus pour cet assuré est pris en compte sur une année. Dans les faits, les assurés ne sont pas toujours observés sur une année complète. Le temps d'observation d'un assuré est appelé l'exposition, notée $e$, et vaut 1 pour une observation d'une année (unité annuelle). Plus formellement, le calcul de l'exposition se fait de la façon suivante :

$$\text{exposition} = \frac{\text{nombre de jours entre la date de début et de fin d'observation}}{\text{nombre de jours dans l'année}}.$$

Si $N$ est le nombre de sinistres survenus sur une année, cette variable est généralisable en définissant $N_e$ comme le nombre de sinistres survenus sur une période $e$. La variable $N$ est alors un cas particulier où $e = 1$.

Pour un assuré $X$ observé sur un temps d'exposition $e$, le nombre de sinistres survenus peut être vu comme un processus de Poisson homogène tel que

$$N_e|X \sim \mathcal{P}(e\lambda(X)),$$

où $\lambda$ est l'intensité du processus homogène. C'est la variable $N_e|X$ qui, dans les faits, est observée. L'objectif est de calculer $\mathbb{E}[N|X]$. Or, la variable $N$ est le nombre de sinistres survenus sur une exposition d'un an, c'est-à-dire dans le cas où $e = 1$. Dans le cas d'un processus de Poisson homogène, les espérances de $N_e|X$ et $N|X$ sont liées par la relation

$$\mathbb{E}\left[\frac{N_e}{e}|X\right] = \lambda(X) = \mathbb{E}[N|X].$$

L'intensité du processus est la fréquence annuelle de sinistres. Dans la mesure où le nombre de sinistres suit une loi de Poisson, la fonction de lien logarithmique est appliquée à l'espérance :

$$\begin{aligned}
&\ln(\mathbb{E}[N|X = x]) = \langle x, \theta \rangle \\
&\Leftrightarrow \ln\left(\mathbb{E}\left[\frac{N_e}{e}|X = x\right]\right) = \langle x, \theta \rangle \\
&\Leftrightarrow \ln(\mathbb{E}[N_e|X = x]) = \langle x, \theta \rangle + \ln(e).
\end{aligned}$$

Il suffit alors d'appliquer un GLM Poisson à la variable observée $N_e$ en rajoutant le logarithme de l'exposition $e$ en offset pour normaliser la mesure de la fréquence.

De façon plus générale, il est possible d'exprimer la fréquence annuelle de sinistres d'un assuré comme le rapport du nombre de sinistres observés sur l'exposition :

$$\text{fréquence sinistre annuelle} = \frac{\text{nombre de sinistres sur la période d'observation}}{\text{exposition}}.$$

[13] Le cas où $b(x) = x$ est un cas où la solution est explicite. Ce cas correspond par ailleurs au cas gaussien, c'est-à-dire au modèle linéaire simple.

**Résidus du modèle**

Une fois le modèle entraîné, une prédiction $\hat{\mu}_i$ est obtenue pour chaque assuré $i$ telle que

$$\hat{\mu}_i = \hat{\mathbb{E}}[Y_i|X = x_i] = g^{-1}(\hat{\omega}(x_i)).$$

Le résidu de cet individu est l'écart entre la valeur observée et sa prédiction. Le résidu (additif) d'un individu $i$ est défini comme :

$$\hat{\epsilon}_i = Y_i - \hat{\mu}_i.$$

D'autres types de résidus peuvent être définis, comme les résidus de Pearson, les résidus de déviance, ou encore les résidus d'Anscombe. Ces résidus sont présentés dans la Table 1.3.

| Résidus | Formule | Description |
|---|---|---|
| Brut | $Y_i - \hat{\mu}_i$ | Écart entre la valeur cible et la valeur prédite. |
| Pearson | $\frac{Y_i - \hat{\mu}_i}{\sqrt{\mathbb{V}_{\hat{\mu}_i}(Y_i)}}$ | Normalisation des résidus bruts par l'estimation de la variance de la loi. Ils sont utilisés dans les cas où les résidus n'ont pas la même variance. |
| Déviance | $d_i \operatorname{sgn}(Y_i - \hat{\mu}_i)$ | Les résidus correspondent à la contribution de l'observation $i$ à la déviance du modèle. |
| Anscombe | $\frac{2}{3}\frac{Y_i^{2/3} - \hat{Y}_i^{2/3}}{\hat{Y}_i^{1/6}}$ | Les valeurs cibles et prédites sont transformées de façon à rendre les résidus gaussiens. |

TABLE 1.3 : Définition des différents résidus d'un modèle

Généralement, ce sont les résidus d'Anscombe qui sont utilisés pour la construction de zonier. Enfin, dans le cas où $Y|X$ suit une loi de Poisson, les réponses prédites ajustées $\hat{Y}_i$ peuvent être calculées ainsi :

$$\hat{Y}_i \in \arg\max_{k \in \mathbb{N}} \left\{ e^{-\hat{\mu}_i} \frac{\hat{\mu}_i^k}{k!} \right\}.$$

Dans ce cas, $\hat{Y}_i$ correspond à la valeur $k$ maximisant la probabilité estimée $\hat{\mathbb{P}}(Y = k|X = x_i)$.

**Analyse de performance**

Dans la mesure où les GLM sont des modèles paramétriques, ils définissent une vraisemblance et, par conséquent, un AIC et un BIC. Ces critères peuvent être utilisés pour sélectionner les variables du modèle en proposant divers modèles GLM expliqués par différentes variables.

Une méthode plus empirique consiste à utiliser les critères de performance tels que la MSE présentée en section 1.2.2. Cependant, ces critères, tels qu'ils sont présentés, ne prennent pas en compte le poids des individus. En effet, tous les individus n'ont pas nécessairement le même poids. Par exemple, dans le cas de la prédiction de la fréquence de sinistres, tous les assurés ne sont pas observés sur des durées similaires. Afin de prendre en compte l'exposition d'un individu, la MSE peut être reformulée comme suit :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\frac{1}{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\omega_i(Y_i - \hat{Y}_i)^2,$$

où $\omega_i$ est le poids de l'individu $i$. Le critère peut donc être ajusté en fonction du poids des individus utilisés pour l'apprentissage.

Par ailleurs, il est possible de montrer qu'un GLM Poisson qui prédit $N_e$ avec le logarithme de l'exposition en offset, et une fonction de lien logarithmique, est numériquement équivalent à un modèle qui prédit la fréquence $\frac{N_e}{e}$ avec l'exposition de l'individu en poids. Le GLM Poisson peut donc être comparé à d'autres méthodes de *Machine Learning* non nécessairement paramétriques.

### 1.2.4 Apport du zonier et de l'Open Data

#### Importance du zonier

Un facteur important pouvant être pris en compte dans la tarification est la position géographique de l'assuré. Toutes choses égales par ailleurs, un assuré vivant dans une zone où le taux de sinistralité est élevé (en ville, par exemple) n'est pas exposé aux mêmes risques qu'un assuré vivant dans une zone avec un taux de sinistralité plus faible (en campagne, par exemple). Aussi, avec l'essor de l'*Open Data*, la prise en compte du facteur géographique est devenue un atout majeur dans la tarification. L'utilisation de données externes permet de mesurer le risque des zones qui ne sont pas exposées dans le portefeuille. Cet avantage permet d'affecter une prime à un nouvel assuré vivant dans une zone qui jusque-là n'était pas représentée dans le portefeuille.

#### Les données utilisées

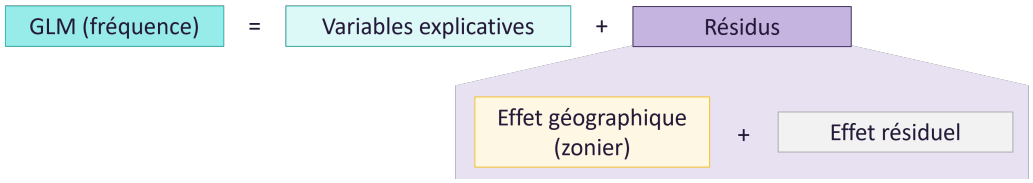
La plupart des données utilisées pour construire le zonier sont les mêmes que celles utilisées pour l'ancien zonier déjà construit. Cependant, dans le cadre de ce mémoire, des données liées à la position des bases de défense en France ont été collectées et utilisées. En somme, les données utilisées sont les suivantes :

- **Bases de données annuelles des accidents corporels de la circulation routière** (de L'INTÉRIEUR (2023)) : ces données contiennent l'ensemble des accidents automobiles ayant eu lieu sur le territoire français entre 2005 et 2022. Seules les années de 2016 à 2022 (avec exclusion de 2020) seront utilisées. Ces données permettent d'obtenir une première estimation du risque automobile dans chaque zone.
- **Bases de correspondance des codes INSEE et des codes postaux** (OPENDATASOFT (2016)) : cette base a été récoltée puisqu'elle permet d'obtenir des informations supplémentaires telles que le type de zone (urbaine, couronne, etc.) ou encore le type de sol (montagne, plateau, etc.).
- **Bases de dossier complet** (INSEE (2024)) : cette base fournie par l'INSEE permet d'obtenir diverses informations sur la démographie de chaque commune (âge moyen, pourcentage d'hommes et de femmes, taux de déménagement, etc.).
- **GitHub contenant les tracés des entités géographiques et administratives françaises suivantes au format GeoJSON** (GREGOIREDAVID (2018)) : ce dossier permet de tracer la carte de France (avec les départements d'Outre-Mer) sur le logiciel R Studio.
- **Localisation des bases de défense** (des ARMÉES (2014)) : ces données contiennent une image permettant de localiser les bases de défense en France. Le sujet étant en lien avec l'armée, les données sont difficilement accessibles en *Open Source*. La position des bases étant représentée sur une image, il a fallu créer "à la main" une base de données qui répertorie leur position. Cette dernière restera uniquement à disposition du groupe AGPM.

**Méthode de construction considérée**

La construction d'un zonier repose sur l'exploitation des résidus du modèle GLM. L'hypothèse fondamentale est la suivante :

*Une partie de l'existence des résidus est due à la non-prise en compte du facteur géographique.*




Les résidus contiennent donc une information géographique. L'objectif est de synthétiser cette information sous la forme d'une variable "zonier". Pour construire cette variable, l'approche suivante est adoptée :

1. **Extraction des résidus pour chaque assuré du portefeuille** : Un premier modèle de tarification est construit à partir d'un GLM. Sur la base de l'hypothèse fondamentale précitée, les résidus du modèle sont extraits. Plusieurs types de résidus peuvent être choisis (Pearson, déviance, Ancombes, etc.).

2. **Découpage du territoire selon une maille géographique** : Une fois les résidus extraits, une maille géographique est définie. Dans ce mémoire, la maille utilisée est le code INSEE (à l'échelle de la commune). L'ensemble du territoire sera donc découpé selon les différents codes INSEE, et le risque géographique sera mesuré pour chaque maille. Pour cela, les résidus de tous les individus sont agrégés au sein de chaque code INSEE selon une métrique choisie.

3. **Agrégation des résidus au sein de chaque maille** : L'agrégation des résidus peut se faire en calculant, par exemple, la moyenne des résidus au sein de chaque code INSEE. L'objectif est d'obtenir une base de données où chaque ligne représente un code INSEE, la variable cible est le résidu agrégé, et les variables explicatives sont celles collectées en *Open Source*.

4. **Construction d'un modèle de prédiction à l'aide de l'*Open Data*** : Étant donné que tous les assurés du portefeuille ne sont pas présents dans chaque code INSEE, un modèle de prédiction des résidus agrégés est mis en place en utilisant les données collectées.

5. **Prédiction des zones manquantes** : Grâce à l'algorithme de prédiction construit, le risque pour chaque code INSEE, qu'il soit présent ou non dans le portefeuille, est estimé. L'exposition de toutes les communes est normalisée à 1.

6. **Lissage des résidus** : Théoriquement, deux codes INSEE géographiquement proches doivent avoir des résidus agrégés similaires. Toutefois, quelques discordances peuvent être observées dans le dégradé géographique du risque. Cette étape consiste donc à lisser les résidus géographiques de manière à ce que deux codes INSEE voisins présentent des risques similaires. Le lissage effectué ici s'inspire de la théorie de la crédibilité.

7. **Clustering des valeurs prédites lissées** : Jusqu'à ce point, la mesure du risque a été effectuée à partir des résidus, qui sont des variables quantitatives. L'objectif est de créer une variable zonier qualitative, où chaque modalité correspond à un risque spécifique. Les résidus lissés sont donc regroupés en plusieurs classes qui formeront les modalités de la variable zonier.

8. **Intégration de la variable zonier dans le modèle de tarification** : Enfin, le modèle de tarification est reconstruit en prenant cette fois en compte l'effet géographique, grâce à la variable zonier construite.

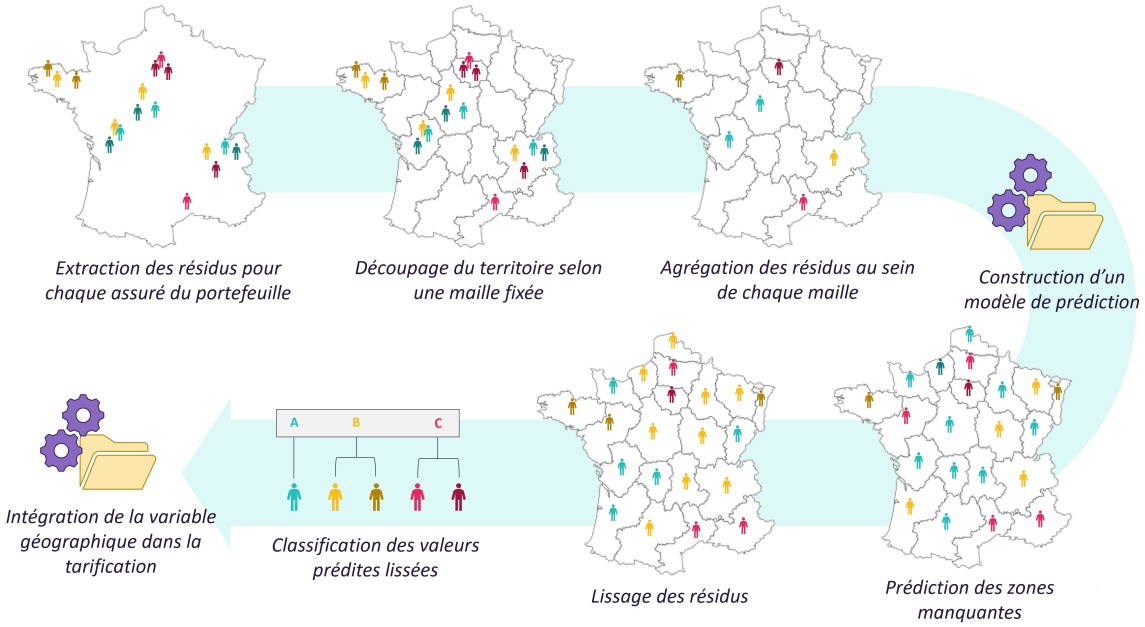



FIGURE 1.7 : Processus de construction de la variable zonier

Toutes ces étapes sont illustrées dans la Figure 1.7. Bien que des modèles comme le Krigeage (CRESSIE (1990)) soient largement utilisés, la méthode de construction proposée a été choisie pour les raisons suivantes :

- Le portefeuille étant très spécifique, l'analyse (Chapitre 2) montre que les assurés sont principalement situés dans des zones proches des bases de défense. Des "îlots" de densité sont observés, laissant de larges zones désertes. Ce phénomène rend difficile l'application d'un lissage classique. En général, le lissage utilisé est basé sur la crédibilité, prenant en compte l'exposition d'une zone. Cependant, ici, de vastes zones n'ont aucune exposition. Il a donc été décidé de réaliser un lissage après avoir obtenu les risques pour toutes les zones (par *Machine Learning*), afin de ramener l'exposition de toutes les communes à 1.
- La machine utilisée n'a pas une grande capacité de mémoire. Des méthodes lourdes en calculs (en termes de temps et de stockage) ne peuvent pas être appliquées, notamment si l'on doit considérer la distance entre plusieurs villes simultanément. C'est pourquoi la méthode de construction du zonier se présente sous la forme d'une succession d'étapes simples et intermédiaires. Cependant, cette contrainte a permis de proposer une méthode de construction innovante, offrant l'avantage de pouvoir contrôler chaque partie du processus de création.

## 1.3 Présentation de la problématique

### 1.3.1 Un portefeuille atypique composé de militaires

#### Le cadre de vie militaire

L'AGPM est une compagnie qui a la particularité d'assurer de nombreux militaires. Leurs modes de vie sont souvent marqués par des missions qui les obligent à s'installer dans des régions éloignées de leur domicile. L'AGPM étant basée à Toulon, une grande partie de ses assurés sont des militaires de la marine.

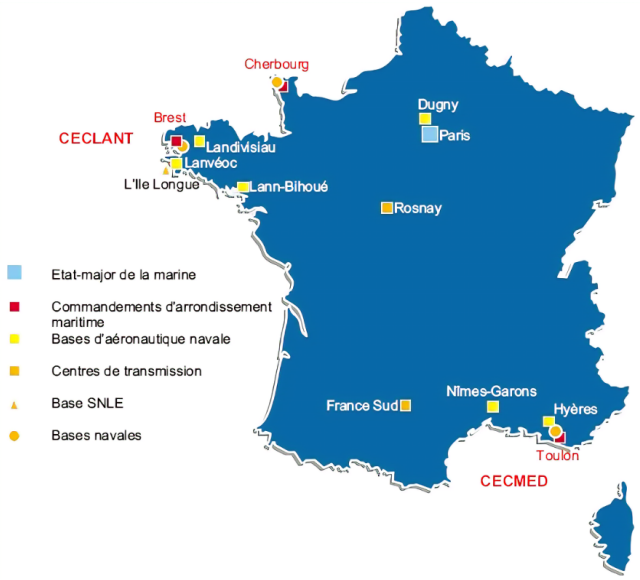



FIGURE 1.8 : Cartographie des principales bases militaires de la marine

La Figure 1.8 présente les principales bases maritimes en Métropole. Les bases de Brest et Toulon représentent les points de défense maritime majeurs de la France. Il est donc attendu qu'il y ait un réseau de transition important entre ces deux villes. Cependant, ces bases ne présentent pas nécessairement le même risque géographique. On pourrait supposer, toutes choses égales par ailleurs, que les conducteurs toulonnais soient de moins bons conducteurs que les conducteurs brestois. Des assurés toulonnais s'installant à Brest pourraient ainsi augmenter la sinistralité de la zone en apportant avec eux leur expérience de mauvais conducteurs. Inversement, un flux d'assurés brestois vers Toulon pourrait conduire à une sous-estimation du risque à Toulon. Une surtarification à Brest et une sous-tarification à Toulon pourraient donc être observées.

Plus généralement, un assuré qui arrive dans un nouvel environnement pourrait être exposé à un risque d'accidentalité plus élevé, en raison d'une mauvaise connaissance des pièges de circulation, des zones à éviter ou du stress lié à la conduite dans une ville inconnue.

**Un portefeuille présentant des "îlots" d'exposition sur le territoire**

Une autre problématique liée à la singularité de ce portefeuille est la localisation des assurés. L'analyse (Chapitre 2) montrera que la majorité du portefeuille est concentrée dans des zones proches des bases militaires. Cette répartition géographique pourrait introduire un biais : le biais de sélection. Ce biais survient lorsque la population étudiée n'est pas représentative de la population globale. Étant principalement composée de militaires, et donc localisée près des bases de défense, la distribution géographique des résidus pourrait ne pas refléter l'ensemble du territoire.

D'une part, une forte hétérogénéité dans la répartition des individus génère des "îlots" d'expositions très distants les uns des autres. Cela rend la notion de "voisins proches" moins pertinente et pourrait rendre difficile l'application d'un lissage sur des zones peu exposées. C'est la raison pour laquelle le lissage est appliqué après la prédiction des résultats par l'algorithme, et non sur les données observées en amont, car l'algorithme garantit que l'exposition de toutes les communes est normalisée à 1.

D'autre part, bien que les bases militaires soient souvent situées dans des zones où le taux de sinistralité automobile est élevé, ce facteur pourrait être sous-estimé par les algorithmes d'apprentissage, car les communes utilisées pour l'entraînement du modèle pourraient ne pas être suffisamment discriminantes par rapport à ce critère, étant toutes similaires. En conséquence, si les résidus couvrent des zones géographiquement proches mais similaires, la performance des algorithmes de prédiction pourrait être réduite. Pour pallier ce problème,

une variable indiquant la distance d'une commune à la base de défense la plus proche a été intégrée. Cela permet de capter les variations de risque mesurées sur des zones proches, même si le portefeuille présente une exposition géographiquement concentrée.

### 1.3.2 La mobilité, un phénomène impactant la mesure du risque à deux niveaux

La méthode de construction du zonier repose sur les résidus du modèle, qui portent l'information géographique. Ainsi, le risque d'une zone est une agrégation des résidus des assurés de cette commune. Toutefois, si parmi ces assurés se trouvent des "nouveaux entrants" ayant récemment intégré la commune, il est possible que leurs résidus soient biaisés par le fait d'avoir changé d'environnement. Le phénomène de mobilité est donc un facteur exogène qui impacte la tarification à deux niveaux : le risque individuel et le zonier.

**Impact théorique de la mobilité sur la construction du zonier**

Les résidus extraits ne portent pas uniquement l'effet géographique. La Figure 1.9 illustre l'impact théorique de la mobilité sur la construction du zonier.

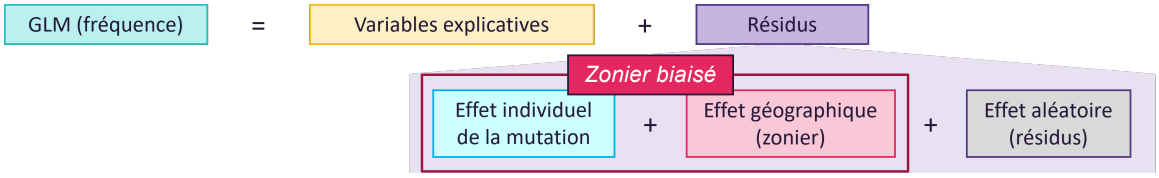



FIGURE 1.9 : Impact théorique de la mobilité sur la construction du zonier

Ainsi, en agrégeant les résidus des individus d'une zone sans distinguer les nouveaux entrants, la mesure du risque géographique peut être faussée, surtout si les biais apportés par ces entrants ne se compensent pas.

**Impact théorique de la mobilité sur le risque individuel**

Il suffirait a priori d'ajouter une variable "mutation" dans le modèle de base pour prendre en compte ce phénomène. Ce modèle servira de référence pour comparer les performances des modèles plus complexes (Chapitre 4) qui seront développés. Toutefois, intégrer directement cette variable pourrait s'avérer insuffisant. Le GLM estime des paramètres qui s'appliquent à tous les individus, mais dans le cas de la population militaire, qui possède des caractéristiques spécifiques, certains paramètres pourraient être inutiles pour estimer leur risque et pourraient donc être implicitement biaisés. Pour cette raison, d'autres modèles plus complexes seront appliqués afin d'enrichir l'étude.

### 1.3.3 Formalisation du phénomène de mutation

La variable *mutation* est définie de la manière suivante : les données sont ordonnées chronologiquement par police, et si le code INSEE d'une police pour une observation donnée est différent de celui de l'observation précédente, la variable prend la valeur 1 ; sinon, elle prend la valeur 0.

La Figure 1.10 illustre le processus de construction de la variable mutation. Cette construction est toutefois limitée par l'aspect temporel, car l'exposition n'est pas prise en compte dans la création de cette variable. Un assuré ayant muté avec une date de début d'observation en décembre aura une exposition de mobilité d'un mois, tandis qu'un assuré dans le même cas, mais avec une date de début en avril, aura une exposition de mobilité de neuf mois. Cependant, un assuré ayant muté en décembre et ayant un sinistre en janvier aurait probablement un sinistre lié à sa mutation, mais ce sinistre sera comptabilisé comme un sinistre sans effet de mobilité dans le cadre de la structure en année comptable.

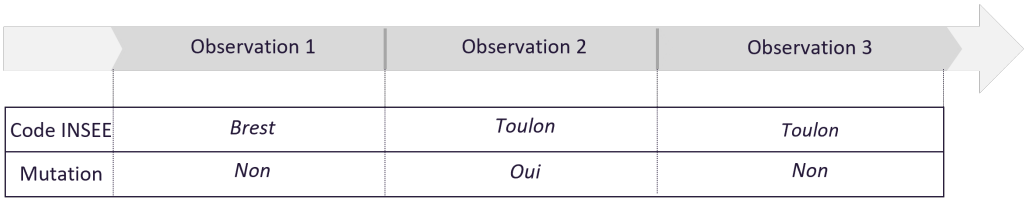



FIGURE 1.10 : Schéma de construction de la variable mutation

**Conclusion**

L'objectif de ce chapitre est de clarifier le contexte de l'étude. Il s'agit d'une tarification automobile qui se confronte à un portefeuille atypique, composé de militaires. Le mode de vie militaire, rythmé par des mutations, implique des déplacements fréquents à travers le territoire. En théorie, ce phénomène a un double impact : sur le risque individuel et sur le zonier. Ce mémoire a donc pour objectifs :

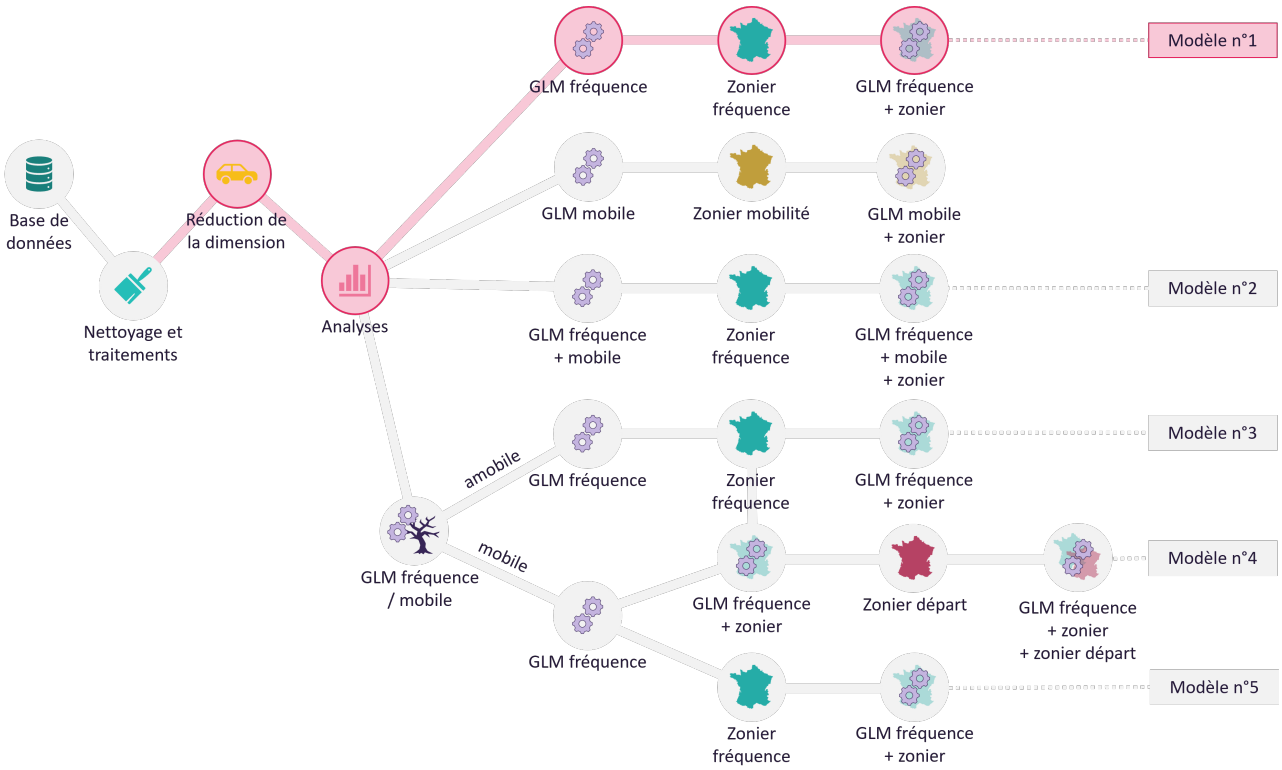
- d'étudier le profil des assurés soumis à la mobilité, et de prédire leur comportement afin d'assister la compagnie dans ses prises de décisions stratégiques ;
- de mettre en évidence l'existence et l'impact de la mobilité sur la mesure du risque, que ce soit à travers le modèle ou le zonier ;
- de proposer une modélisation et des méthodes adaptées à la spécificité du portefeuille.

Le prochain chapitre sera consacré à la construction d'un modèle de fréquence simple, qui ne prend pas en compte la mobilité. Ce chapitre présentera également les premiers résultats de la méthode de construction du zonier.

# Chapitre 2

# Construction d'un premier modèle de fréquence

*L'objectif de ce chapitre est de construire un premier modèle de fréquence. Ce modèle servira de référence pour la modélisation et l'analyse de l'impact de la mobilité. Dans un premier temps, les données sont analysées et préparées pour l'apprentissage, ce qui constitue la phase de preprocessing. La deuxième partie est dédiée à la mise en place d'un modèle de fréquence sans prendre en compte le risque géographique. Enfin, dans la troisième partie, le zonier sera construit à partir des résidus et intégré au modèle.*



## 2.1 Analyse et preprocessing des données

Cette section vise à explorer le jeu de données disponible, en comprendre la structure et anticiper les éventuelles difficultés pouvant surgir lors de l'apprentissage.

### 2.1.1 Analyse descriptive

#### Analyse univariée

L'analyse univariée consiste à examiner chaque variable individuellement et à évaluer son influence sur la sinistralité des assurés. Pour des raisons de concision, l'étude se concentre principalement sur l'évolution de la sinistralité dans le portefeuille, ainsi que sur l'âge et la catégorie socioprofessionnelle des assurés, qui représentent en grande partie la structure démographique du portefeuille. Les autres variables sont détaillées en Annexe B.

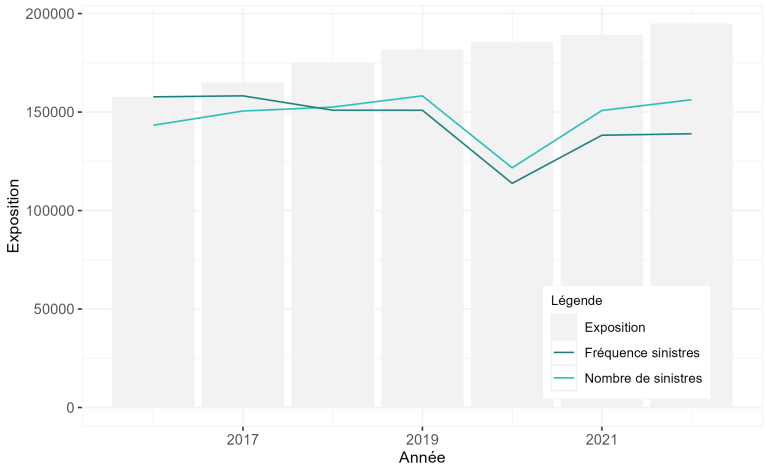
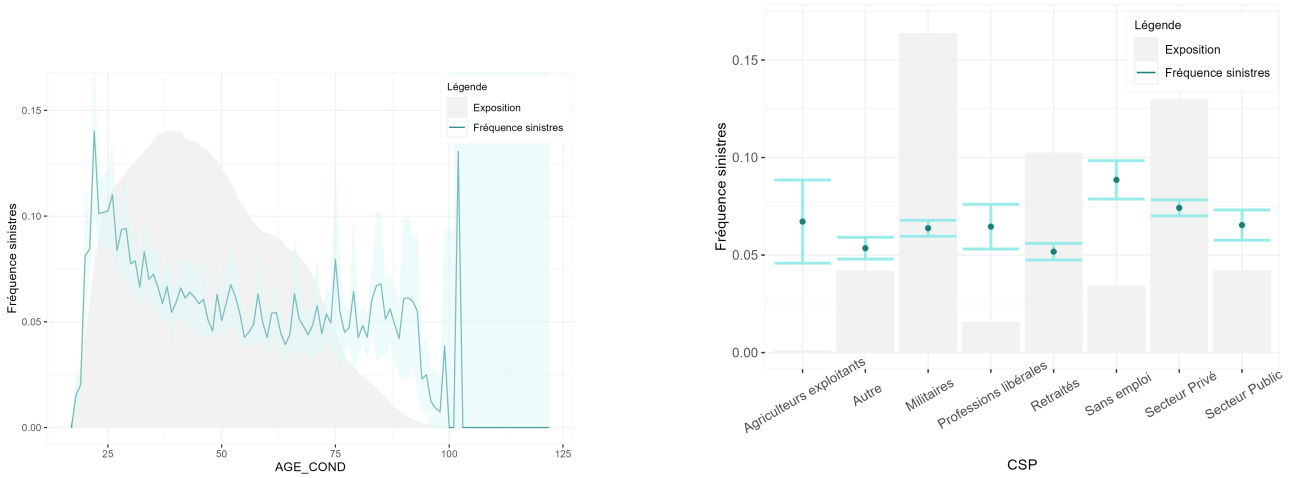



FIGURE 2.1 : Évolution de la sinistralité du portefeuille

La première analyse porte sur l'évolution de la sinistralité. Les résultats, présentés dans la Figure 2.1, montrent que bien que le nombre de sinistres augmente avec l'exposition des assurés, la fréquence des sinistres semble diminuer. Il existe une décroissance progressive de la sinistralité, marquée par une chute notable en 2020, due à la pandémie de COVID-19. Cette tendance est représentative de l'évolution de la sinistralité automobile des compagnies françaises au niveau national, comme le montre la Figure 1.1. L'année calendaire est donc une variable importante qui sera intégrée dans le modèle.

L'analyse se poursuit avec les variables d'âge et de catégorie socioprofessionnelle, dont les résultats sont présentés dans la Figure 2.2. Sur toute la période observée, la compagnie a principalement assuré des conducteurs dont l'âge moyen est de 47 ans. La majorité des assurés se situe dans la tranche d'âge 30-50 ans, avec une décroissance progressive vers les âges plus avancés. Le portefeuille est relativement jeune, bien qu'une proportion notable de retraités (probablement d'anciens militaires) soit également présente. Du point de vue de la sinistralité, le phénomène de la "*bosse des accidents*" se manifeste, c'est-à-dire une augmentation de la fréquence des sinistres chez les jeunes conducteurs (entre 17 et 25 ans). Cette sinistralité diminue ensuite avec l'âge, puis se stabilise à partir de 50 ans, bien qu'une certaine volatilité persiste. Une instabilité particulièrement élevée est observée pour les conducteurs très âgés (plus de 90 ans), en raison de leur faible représentativité dans le portefeuille.

Cependant, il est important de noter que la hausse de la sinistralité chez les jeunes conducteurs peut être due




(a) Analyse de l'âge

(b) Analyse de la catégorie socioprofessionnelle

FIGURE 2.2 : Structure démographique du portefeuille

à d'autres facteurs. Par exemple, il est possible que l'âge ne soit pas le principal facteur influent, mais plutôt le statut d'emploi, avec un risque accru chez les individus sans emploi, qui sont souvent jeunes. Cela donne l'illusion que les jeunes conducteurs sont davantage à risque. De telles hypothèses nécessitent des analyses plus approfondies, notamment une étude des interactions entre les différentes variables.

L'AGPM étant une compagnie d'assurance spécialisée dans les profils militaires, l'analyse révèle que cette catégorie d'assurés représente la majorité du portefeuille. Cependant, cette majorité reste relative, car d'autres catégories, comme les salariés du secteur privé ou les retraités, sont également fortement représentées. En termes de sinistralité, les assurés sans emploi se distinguent par un taux de sinistres plus élevé que les autres groupes. Il est également à noter que la sinistralité des militaires semble se situer dans la moyenne.

**Analyse cartographique**

Dans la mesure où un des enjeux de ce mémoire est la construction d'un zonier, il est intéressant d'analyser les données d'un point de vue géographique en les géocodant. Il est ainsi possible de visualiser la répartition des assurés sur l'ensemble du territoire. Le résultat est présenté sur la Figure 2.3.

La répartition des assurés est assez concentrée et non uniformément répartie sur toute la métropole. Des regroupements se distinguent, par exemple, au sud-est en Côte d'Azur, sur la côte Atlantique en Gironde et à Brest, ou encore vers les grandes villes comme Paris et Lyon. Le risque est donc très localisé. Afin de quantifier l'hétérogénéité de la répartition des assurés, le coefficient de Gini (CERIANI et VERME (2012)) est calculé sur l'exposition de toutes les communes.

Le coefficient de Gini est un indicateur utilisé en économie qui permet de mesurer la répartition des richesses au sein d'une population dans un territoire donné. Un coefficient de Gini proche de 0 signifie que la répartition de la richesse est homogène, tandis qu'un coefficient à 1 signifie que cette répartition n'est pas du tout égalitaire. Ici, la "richesse" en question est l'exposition et la "population" est l'ensemble des communes. L'application donne un coefficient de Gini égal à 0.87, ce qui signifie qu'il y a bien une forte hétérogénéité dans la répartition du portefeuille sur le territoire.

Aussi, puisque le portefeuille est principalement constitué de militaires, il est intéressant d'étudier la répartition des assurés en fonction de la proximité à une base de défense.

La Figure 2.4 permet d'étudier cette corrélation entre l'exposition d'une commune et la proximité à une

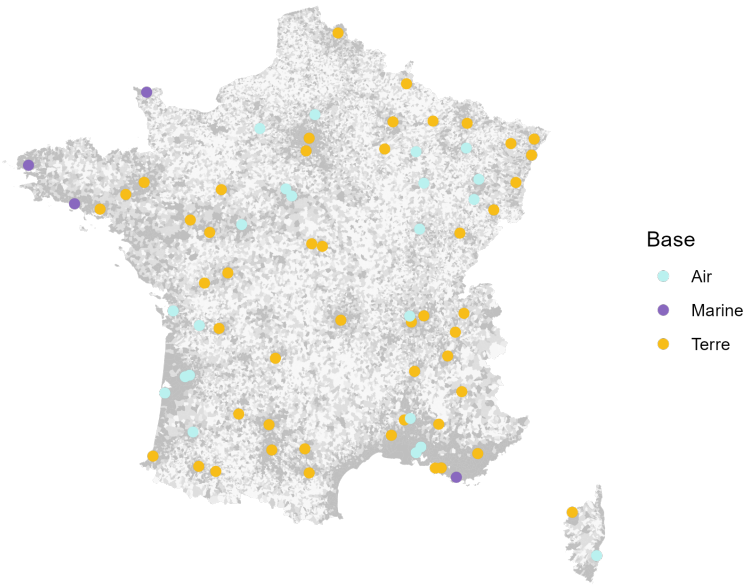
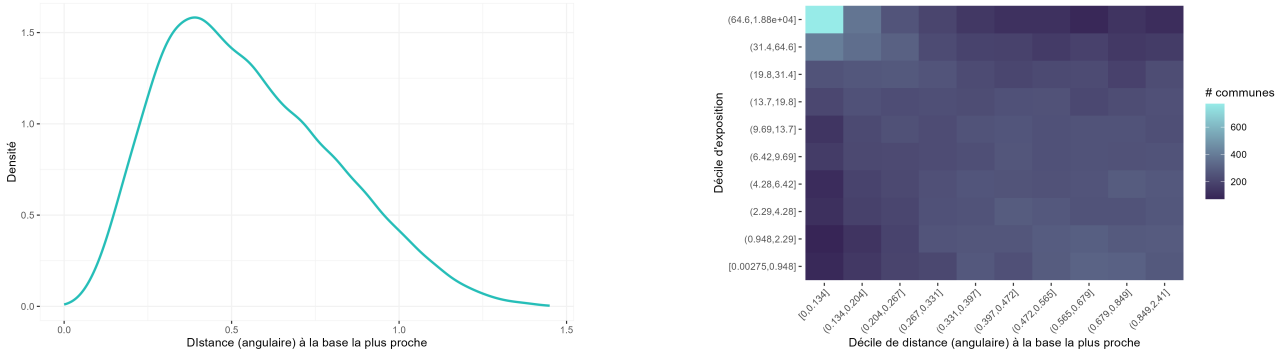



FIGURE 2.3 : Exposition du portefeuille sur le territoire




(a) Distribution pour les communes non exposées

(b) Distribution des communes exposées

FIGURE 2.4 : Corrélation entre l'exposition d'une commune et la distance à la base la plus proche

base de défense. La Figure 2.4a montre qu'aucune commune non exposée[1] n'est située à proximité d'une base de défense, mais elles sont presque toujours éloignées d'une distance (angulaire) d'au moins 0.5°. En ce qui concerne les communes exposées, la Figure 2.4b montre bien que les communes les plus exposées sont toujours situées très proches des bases de défense. Les distances et les expositions sont divisées en déciles, c'est-à-dire que chaque segment selon une variable contient 10% des communes. Quantitativement, le croisement entre le décile d'exposition le plus élevé et le décile de distance le plus faible (distance proche de 0) représente 37% de l'exposition totale du portefeuille. Plus largement, les quatre cadrants en haut à gauche représentent 49% de l'exposition totale, soit près de la moitié de l'exposition du portefeuille.

L'exposition géographique du portefeuille est donc bien concentrée au niveau des communes proches des bases de défense. La variable représentant la distance d'une commune à la base de défense la plus proche

[1] C'est-à-dire où aucun assuré n'a été présent dans la zone

sera intégrée au modèle de prédiction du zonier afin de soutenir l'algorithme dans la détection du risque géographique et faciliter le lissage.

**Analyse multivariée**

Une analyse est désormais portée sur les liens entre les différentes variables. Pour rappel, le jeu de données contient deux types de variables : qualitatives et quantitatives. Pour un couple de variables choisi, selon leur type respectif, différents critères sont utilisés pour évaluer l'intensité de leur liaison. Selon les cas de figure, le package *greybox* permet de calculer les quantités suivantes :

- Dans le cas de deux variables quantitatives $X$ et $Y$, le critère de liaison est le coefficient de corrélation de Pearson
$$\frac{Cov(x,y)}{\sigma_x \sigma_y},$$
où, en notant $x_1, ..., x_n$ (respectivement $y_1, ..., y_n$) les $n$ observations de la variable $X$ (resp. $Y$) :
  - $Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n \bar{y}_n$ est la covariance observée entre les variables $X$ et $Y$,
  - $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ est la moyenne des observations de $X$,
  - $\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2}$ est l'écart-type des observations de $X$.
- Dans le cas de deux variables qualitatives $X$ et $Y$, le critère de liaison est le V de Cramer. Soit $m_1, ..., m_J$ (resp. $l_1, ..., l_K$) les $J$ modalités (resp. $K$) de la variable $X$ (resp. $Y$). Le V de Cramer est défini de la façon suivante
$$\sqrt{\frac{\frac{\chi^2}{n}}{\min(J-1, K-1)}},$$
  - $\chi^2 = \frac{1}{n}\sum_{j,k} \frac{(n_{j,k} - n_{j\cdot} n_{\cdot k})^2}{n_{j\cdot} n_{\cdot k}}$ est la statistique du khi-deux,
  - $n_{j,k} = \sum_{i=1}^{n} \mathbb{1}_{x_i = m_j} \mathbb{1}_{y_i = l_k}$ est le nombre d'individus prenant à la fois la modalité $j$ de la variable $X$ et la modalité $k$ de la variable $Y$,
  - $n_{j\cdot}$ (resp. $n_{\cdot k}$) est le nombre d'individus prenant la modalité $j$ de la variable $X$ (resp. $k$ de la variable $Y$).
  - Dans le cas mixte (une variable quantitative et une variable qualitative), c'est le coefficient de corrélation multiple qui est utilisé. Un modèle linéaire ANOVA est mis en place où
$$Y = \theta_0 + \theta_1 \mathbb{1}_{X=x^1} + \cdots + \theta_m \mathbb{1}_{X=x^m} + \epsilon_i,$$
où $x^1, ..., x^m$ sont les $m$ modalités de la variable $X$, et $\epsilon_i$ les termes d'erreur (indépendants et identiquement distribués de loi normale centrée et de variance constante). Le coefficient de corrélation multiple correspond simplement à la racine du coefficient de détermination $R^2$ du modèle, où
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$
avec $\hat{Y}_i$ les valeurs prédites, et $\bar{Y}$ la moyenne de $Y$ dans l'échantillon.

Toutes ces métriques prennent des valeurs entre $-1$ et $1$. Une valeur de $0$ signifie que les deux variables sont indépendantes ou décorrélées, tandis qu'une valeur proche de $1$ ou $-1$ signifie que les deux variables sont liées (ou corrélées). C'est finalement une table de croisement donnant l'intensité de liaison entre chaque couple de variables qui est obtenue.

Cette table de croisement peut être interprétée de façon géométrique. En prenant l'intensité de liaison en valeur absolue, une liaison de 0 signifie "indépendant" et 1 signifie "lié". En transformant par la fonction $x \mapsto 1 - |x|$, les valeurs à 0 signifient cette fois-ci "lié" et 1 "indépendant". En fait, les notions de "lié" et "indépendant" peuvent être traduites par la notion de distance (ou d'éloignement). Deux variables sont proches au sens de la distance (distance proche de 0) lorsqu'elles sont liées. À l'inverse, elles sont éloignées (proches de 1) si elles sont indépendantes.

Il est alors possible d'appliquer une *Multidimensional Scaling* (MDS) (KRUSKAL (1978)) qui est une méthode permettant d'obtenir la position de points dans un plan cartésien à partir d'une table de distance entre chaque point. Les points sont ici les variables tarifaires. C'est la fonction *cmdscale* du package *stats* qui est utilisée pour appliquer la MDS.

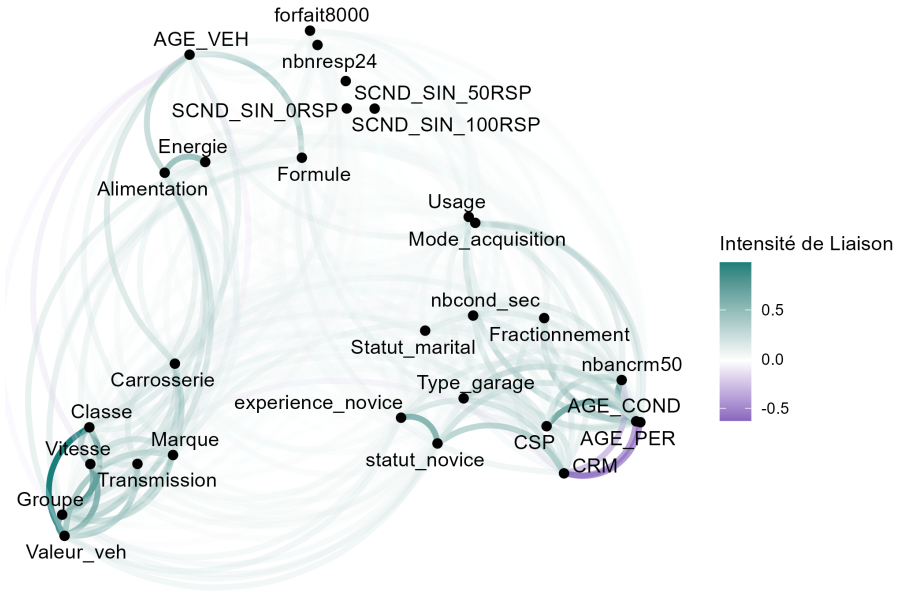



FIGURE 2.5 : Réseau de liaison entre les variables

La Figure 2.5 est le résultat de cette méthode appliquée aux données. C'est un réseau de connexions où l'intensité des arêtes dépend de l'intensité de la liaison entre les variables constituant les sommets. En analysant le réseau obtenu, trois groupes se distinguent : celui de gauche qui représente l'ensemble des caractéristiques du véhicule, celui de droite qui représente les caractéristiques du conducteur assuré, et celui du dessus qui représente l'historique de sinistralité. L'âge du permis et l'âge du conducteur sont très corrélés. Il serait donc inutile de conserver les deux variables pour le GLM puisque les deux variables sont linéairement liées. Seul l'âge du conducteur sera conservé.

Par ailleurs, les variables relatives aux véhicules, en plus d'être pour la plupart qualitatives avec beaucoup de modalités, sont aussi fortement liées. Il serait intéressant de les regrouper en une seule variable. Cette idée sera développée dans la suite à travers le véhiculier.

### 2.1.2 Préprocessing

L'étape de préprocessing consiste à formater le jeu d'apprentissage (et de test). L'objectif est le suivant : dans un souci d'interprétation et de segmentation des polices, toutes les variables doivent être rendues qualitatives. Par ailleurs, les variables véhiculaire sont pour la plupart qualitatives et contiennent beaucoup de

modalités. Une variable qui contient beaucoup de modalités peut entraîner un surapprentissage de l'algorithme. S'il y a peu de représentants pour une modalité donnée, l'estimation ne sera pas suffisamment robuste et l'algorithme s'ajustera trop aux individus.

#### Bases d'entraînement et base de test

La base d'entraînement correspond aux observations allant de 2016 à 2021, tandis que la base de test correspond aux observations de 2022.

#### Catégorisation des variables quantitatives

La prochaine étape consiste à catégoriser les variables quantitatives. Pour ce faire, l'approche de l'Algorithme 1 sera adoptée.

**Algorithme 1** Catégorisation d'une variable quantitative

Appliquer le modèle : *gam(nb_sin* $\sim$ *s(variable) + offset(log(expos)))*
Récupérer la fonction spline du modèle
Discrétiser cette spline par un algorithme CART
▷ *Chaque feuille terminale est une modalité de la variable catégorisée et la profondeur de l'arbre détermine le nombre de modalités. Le nombre de modalités est ensuite optimisé.*
**pour** un nombre de modalités $m$ **:**
Élaguer l'arbre de sorte à obtenir $m$ modalités
Appliquer le modèle : *glm(nb_sin* $\sim$ *variable catégorisée + offset(log(expos)))*
Calculer le BIC du modèle
**fin pour**
Choisir le nombre de modalités optimal en élaguant l'arbre de sorte à minimiser le BIC

Les modèles linéaires supposent une relation de linéarité entre les variables explicatives et la valeur cible (ou une transformation de cette valeur). Néanmoins, l'effet des variables étudiées n'obéit pas nécessairement à cette hypothèse en pratique.

Catégoriser une variable peut être vu comme appliquer une fonction de discrétisation à la variable dans l'expression du modèle. Exprimer une espérance par une combinaison de variables transformées, c'est exactement le principe des *Modèles Additifs Généralisés* (GAM) (HASTIE et TIBSHIRANI (1987)). Ces modèles se présentent ainsi

$$g(\mathbb{E}[Y|X = x]) = \theta_0 + f_1(x_1) + \cdots + f_p(x_p),$$

où les fonctions $f_i$ sont des fonctions dites *splines de lissage* (c'est généralement le terme de *smoothing splines* qui est employé dans la littérature). Ces fonctions s'expriment comme une combinaison linéaire de fonctions de bases (appelées *basis functions*)

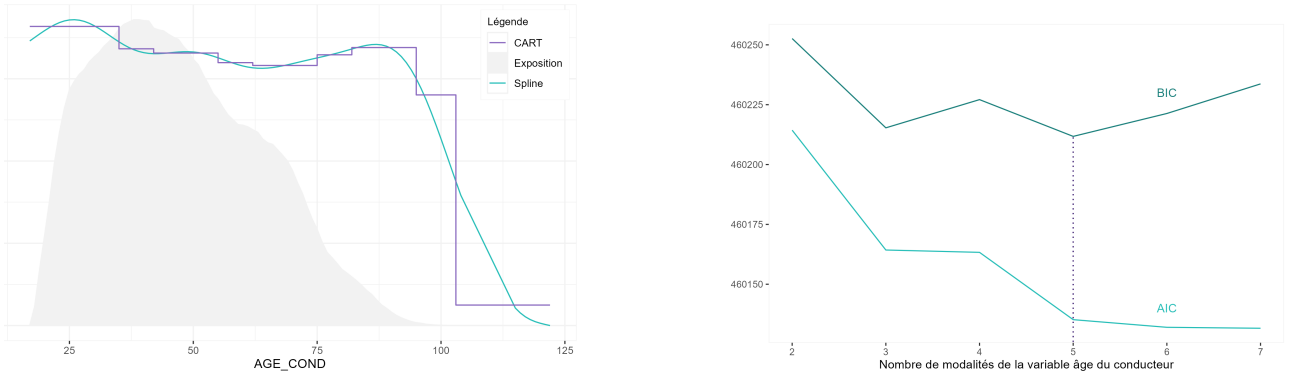
$$f_i(x) = \sum_{j=1}^{k} \beta_{i,j} b_{i,j}(x).$$

La commande *gam* de R utilise des splines dites en *plaque minces (thin plate regression spline)* dont la complexité augmente avec $j$. Par défaut, la complexité maximale est $k = 10$.

La spline obtenue pour la variable âge du conducteur[2] est présentée sur la Figure 2.6a en vert. Cette spline a été discrétisée à l'aide d'un arbre de régression (BREIMAN et al. (1984)) en pondérant chaque âge par son

[2]Le modèle GAM a été appliqué individuellement sur chaque variable numérique à la fois. Il aurait été plus judicieux d'appliquer directement un GAM sur toutes les variables en même temps, en appliquant la spline uniquement sur les variables quantitatives. Cela

exposition dans le portefeuille. Le résultat de cette discrétisation est représenté par la courbe en escalier en violet. Ce découpage représente l'arbre de complexité maximale. Chaque plateau représente un découpage de la variable âge du conducteur. La prochaine étape est de déterminer le nombre de modalités optimal, c'est-à-dire à quelle profondeur il faut élaguer l'arbre. À chaque niveau d'élagage, une variable catégorisée de l'âge du conducteur est obtenue. Pour chaque nombre de modalités différent, un GLM Poisson est appliqué sur cette variable et le nombre de modalités optimal sera choisi comme étant celui qui minimise le BIC du modèle.




(a) Catégorisation de la spline

(b) Détermination du nombre de modalités optimal

FIGURE 2.6 : Application de la méthode de catégorisation sur l'âge du conducteur

Toujours dans l'exemple de l'âge du conducteur, la Figure 2.6b montre que le plus petit BIC est obtenu pour un nombre de modalités égal à 5. C'est donc ce partitionnement qui sera appliqué pour la variable. Cette méthode sera appliquée à toutes les variables quantitatives.

**Réduction de la dimension : construction d'un véhiculier technique**

Un véhiculier est une variable tarifaire qui permet de segmenter les assurés selon les caractéristiques techniques de leur véhicule. Classiquement, la méthode de construction d'un véhiculier est similaire à celle du zonier. Elle consiste à extraire les résidus d'un modèle qui ne prend pas en compte les variables relatives au véhicule, puis à agréger et modéliser les risques à l'aide de l'*Open Data*. Une segmentation des véhicules est déjà fournie par l'organisme SRA (SOciété de Réparation Automobiles) à travers les variables Groupe et Classe. Cependant, l'intérêt du véhiculier est de permettre à l'assureur d'adapter cette segmentation SRA à son propre portefeuille ou encore de lisser les tarifs afin de favoriser la mutualisation.

La méthode généralement utilisée afin de construire le véhiculier est similaire à celle du zonier, c'est-à-dire qu'elle se base sur les résidus du modèle. En revanche, cette approche ne convient pas dans le cadre de ce mémoire. En effet, cette méthode suppose que l'information relative au véhicule est contenue dans les résidus du modèle. Or, les résidus contiennent également l'information géographique, et potentiellement encore l'information liée aux mutations. Les résidus qui seront extraits seront donc biaisés par de nombreux facteurs cumulés. Seul un véhiculier "technique" sera construit, c'est-à-dire une variable qui regroupe les caractéristiques techniques similaires relatives aux véhicules des assurés. L'objectif est donc de synthétiser l'information répartie selon toutes les variables véhiculaire en une seule. Néanmoins, selon la méthode de regroupement choisie, il peut y avoir une perte d'information plus ou moins importante. En effet, le véhiculier construit est une variable dont les modalités sont moins nombreuses que toutes celles présentes à l'origine. En contrepartie de cette perte d'information, son intérêt est justifié pour les raisons suivantes :

aurait permis de capter les effets marginaux de toutes les variables par rapport aux autres. Cependant, la mémoire de la machine à disposition ne permet pas une telle application.

– **Indépendance des variables** : les variables d'un GLM doivent être linéairement indépendantes afin d'améliorer la qualité de l'estimation des paramètres. L'indépendance permet de pallier au problème de singularité de la matrice construite lors de la minimisation du risque empirique dans l'algorithme de Newton-Raphson.

– **Principe de parcimonie** : il est toujours préférable de construire un modèle expliquant un phénomène complexe avec le moins de variables explicatives possible. Cette condition facilite, entre autres, l'interprétation du modèle.

– **Mutualisation** : en regroupant des variables corrélées (ou liées), la mutualisation dans le portefeuille est favorisée.

– **Surajustement** : dans le cas où des modalités sont représentées par un faible nombre d'individus, l'estimation du coefficient n'est pas suffisamment robuste et l'algorithme perd en capacité de généralisation.

La première étape est de constituer une base représentative des véhicules dans le portefeuille. À partir du jeu de données initial, on agrège en un seul tableau l'ensemble des combinaisons possibles des caractéristiques des véhicules en sommant l'exposition et le nombre de sinistres. La Figure 2.7 illustre le passage de la maille police à la maille véhicule. Cela permet de réduire la dimension des données dans un premier temps.

| Police | Marque | Classe | # sinistres | Exposition |
|---|---|---|---|---|
| 1 | CITROEN | M | 0 | 0,75 |
| 2 | CITROEN | M | 1 | 0,45 |
| 3 | CITROEN | P | 1 | 0,67 |
| 4 | MERCEDES | C | 0 | 0,90 |
| 5 | MERCEDES | C | 2 | 0,83 |

| Marque | Classe | # sinistres | Exposition |
|---|---|---|---|
| CITROEN | M | 1 | 1,2 |
| CITROEN | M | 1 | 0,67 |
| MERCEDES | C | 2 | 1;73 |

FIGURE 2.7 : Construction de la base véhiculier

Un algorithme CART est construit afin de regrouper les variables véhiculier : à partir des caractéristiques du véhicule, on prédit la fréquence sinistre associée, en intégrant l'exposition en poids. Puis, chaque feuille terminale devient une modalité de la variable véhiculier.

En résumé, les 7 variables relatives au véhicule ont été regroupées en une seule variable à $14$ modalités, soit $14$ modalités à estimer pour le GLM contre $66$ avec toutes les variables véhiculaire.

## 2.2 Mise en place d'un prédicteur tarifaire

**Distribution de la variable cible**

La variable cible correspond au nombre de sinistres comptabilisés sur une période d'exposition donnée pour un assuré. Sa distribution est donnée sur la Figure 2.8.

Il faut bien noter que tester directement l'adéquation de $Y$ à une loi de Poisson ne donne aucune indication quant à la pertinence du GLM Poisson, car l'hypothèse faite sur la distribution n'est pas sur $Y$ mais sur $Y|X$. Une distribution qui suppose que $Y|X$ suit une loi de Poisson n'implique pas que $Y$ suit une loi de Poisson[3]. Le graphique permet uniquement de confirmer que la variable cible est bien une variable de comptage, et qu'il pourrait être pertinent de supposer que $Y|X$ suit une loi de Poisson.

[3]La loi dépend aussi de la distribution de $X$, qui est trop complexe pour être intuitivement comprise.

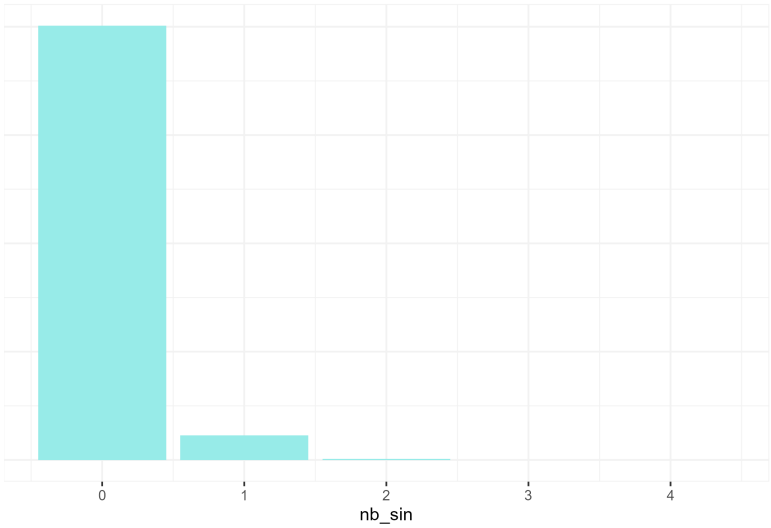



FIGURE 2.8 : Distribution de la variable cible

Une autre information importante peut être tirée du graphique : il y a une explosion d'individus dont le nombre de sinistres est égal à 0. Il est fort probable que la distribution de $Y$ soit une distribution dite zéro-inflatée. Il sera donc difficile pour le modèle de détecter des individus qui auront au moins un sinistre, puisque ces derniers sont noyés parmi ceux qui n'en ont jamais eu. C'est un cas classique de sous-représentation d'une classe d'individus dans le jeu de données. Des alternatives existent afin de contourner ou prendre en compte ce phénomène :

- **Sous-échantillonnage** : constituer un nouveau jeu de données à partir du jeu de données initial en tirant aléatoirement des individus (avec ou sans remise) sous la contrainte que la répartition des individus dans chacune des classes soit similaire.
- **Suréchantillonnage** : appliquer une méthode (par exemple la méthode SMOTE) de sorte à "générer" des nouveaux individus similaires à ceux appartenant à la classe minoritaire, et rééquilibrer ainsi les classes.
- **Changer la distribution de la variable cible** : modéliser par une Binomiale négative d'ordre 1, une Binomiale négative d'ordre 2, ou encore une loi géométrique.
- **Modélisation séparée** : appliquer d'abord un modèle binomial afin de prédire la probabilité d'avoir au moins un sinistre, puis modéliser les valeurs strictement positives par une loi zéro tronquée, comme la loi Poisson zéro-tronquée ou la loi binomiale négative zéro tronquée. Les commandes *family = pospoisson* et *family = posnegbinomial* du package *VGAM* permettent de modéliser respectivement ces deux lois.
- **Mélange de lois** : utiliser des modèles type ZIP (Zero Inflated Poisson), ZINB (Zero Inflated Negative Binomial) ou Hurdle (du package *pcsl*).

Afin de rester synthétique, le GLM Poisson sera seulement confronté à la binomiale négative. Le lecteur intéressé trouvera plus d'informations sur la modélisation des données zéro-inflatées dans FENG (2021).

### 2.2.1 Application d'un GLM Poisson

#### Sélection des variables

Afin de rendre compte de l'importance d'une variable sur la qualité du modèle, une méthode de sélection de variables itératives dite AIC backward est appliquée. À partir du modèle initial, un nouveau modèle est ajusté mais en enlevant seulement une variable. Ce procédé est répété pour toutes les variables une à une. Les AIC de tous ces modèles sont comparés avec celui du modèle initial qui contient toutes les variables. Le modèle qui a le plus petit AIC est conservé. Ce modèle devient le nouveau modèle de référence. Le processus est réitéré jusqu'à ce qu'aucun modèle ne donne un AIC qui soit inférieur au modèle de référence. Le dernier modèle obtenu est celui qui sera conservé.

Appliqué au jeu de données, le résultat de l'AIC backward est la suppression de la variable nbancrm50 qui est le nombre d'années avec un CRM à 50%, passant d'un AIC à 381 045,6 à 381 042,5. D'autres variantes sont applicables comme l'AIC forward qui cette fois-ci part du modèle nul[4] et rajoute une variable au fur et à mesure de sorte à minimiser l'AIC. Pour des raisons de temps de calcul, seule la méthode backward sera appliquée. Une dernière variante est l'AIC "step both" qui combine les méthodes backward et forward en choisissant à chaque itération de rajouter ou supprimer une variable.

#### Tests d'adéquation et sur/sous-dispertion

Une fois le modèle mis en place, il est possible de calculer les valeurs ajustées. La valeur prédite par le modèle est une estimation du paramètre de la loi de Poisson, et non un nombre de sinistres entier. La valeur ajustée du modèle pour un individu $i$ est définie comme l'entier $\hat{Y}_i$ tel que

$$\hat{Y}_i \in \arg\max_{k\in\mathbb{N}} \Big\{ e^{-\hat{\mu}_i} \frac{\hat{\mu}_i^k}{k!} \Big\}.$$

En appliquant cette formule aux résultats, toutes les valeurs ajustées sont à $0$. Dans la mesure où le jeu de données a une structure zéro-inflation, le modèle préfèrera toujours estimer une fréquence de sinistres qui soit proche de $0$ afin de minimiser l'erreur.

Un autre problème induit par l'explosion de zéros est la surdispertion. Le modèle de Poisson suppose que l'espérance est égale à la variance. Néanmoins, face à un jeu de données zéro-inflation, cette hypothèse n'est pas toujours respectée. Il est possible de tester la surdispertion du modèle à l'aide de la fonction *dispertiontest* du package *AER*. La p-valeur obtenue est de l'ordre de $10^{-16}$ ce qui signifie qu'il y a de la surdispertion. Elle est d'ailleurs évaluée à 1.07, sachant qu'une dispersion "normale" est de 1. Le lecteur intéressé trouvera plus d'informations sur la surdispersion, et plus largement sur les modèles de comptage dans CAMERON et TRIVEDI (2013).

Une façon de remédier à ce problème est d'appliquer une autre distribution au GLM telle que la Binomiale négative afin de relâcher l'hypothèse d'égalité entre la variance et l'espérance.

### 2.2.2 GLM Poisson vs GLM Binomiale négative

Soit $X$ une variable aléatoire suivant une loi de Poisson, alors sa variance est égale à son espérance. En revanche, si $X$ suit une loi Binomiale négative notée $\mathcal{BN}(r,p)$, sa variance est

$$\mathbb{V}[X] = \mathbb{E}[X]\Big(1 + \frac{1}{r^2}\mathbb{E}[X]\Big).$$

[4] Modèle qui ne dépend que de l'intercept.

Selon la valeur de $r$ qui est un paramètre de la loi en question, un ajustement supplémentaire sur la variance est possible.

| Modèle | AIC | BIC |
|---|---|---|
| Poisson | 381042.5 | 381889.5 |
| Binomiale négative | 380117 | 380975.8 |

TABLE 2.1 : Résultats sur les GLM Poisson et Binomiale négative

La Table 2.1 donne les résultats pour les deux modèles appliqués au portefeuille. Il semblerait que le modèle binomial négatif soit meilleur. Néanmoins, les critères AIC et BIC ne sont applicables que sur les jeux d'entraînement et ne donnent qu'un a priori sur la performance des algorithmes. Afin de rendre compte du pouvoir prédictif du modèle, une validation croisée Monte Carlo est appliquée.

Une validation croisée Monte Carlo consiste à effectuer une validation croisée sur tous les modèles à disposition, ce qui permet de comparer la distribution des erreurs. Les itérations sont exécutées en parallèle à l'aide des packages *parallel* et *doParallel*. La machine à disposition contient 4 cœurs. Ce sont 10 itérations qui sont exécutées sur 3 cœurs[5] en même temps, ce qui signifie que l'exécution de la boucle est trois fois plus rapide que s'il n'y avait pas de parallélisme.

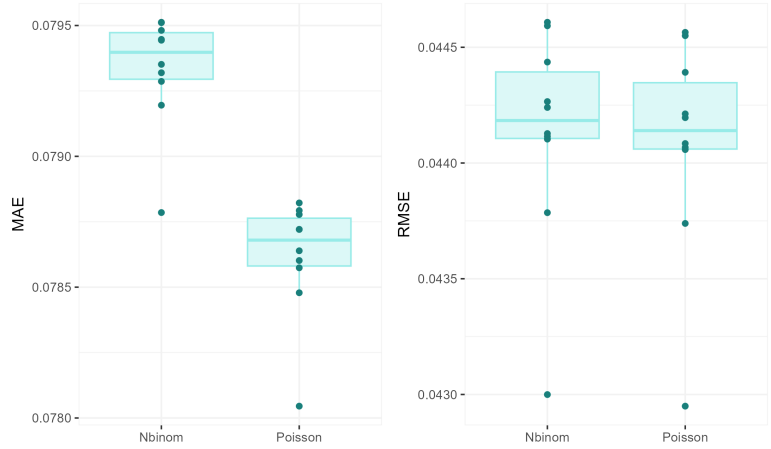



FIGURE 2.9 : Comparaison des performances du GLM Poisson et GLM Binomiale négatif

Les résultats de la validation croisée Monte Carlo sont présentés en Figure 2.9. Du point de vue de la MAE, le modèle Poisson est clairement meilleur que le modèle Binomial négatif. Au niveau de la RMSE, il est légèrement meilleur.

En réalité, comparer l'AIC (ou le BIC) des deux modèles permet de mesurer uniquement la qualité de l'apprentissage mais pas celui du pouvoir prédictif, car les deux modèles comparés ont exactement la même complexité, puisqu'ils ont le même nombre de variables. En effet, le terme de pénalisation pour les deux modèles

[5] Un cœur reste disponible afin de pouvoir effectuer des calculs annexes en parallèle sur la machine. Ainsi, trois cœurs sont exécutés au service de la machine virtuelle et il reste 1 à disposition.

est exactement le même. La différence est donc au niveau de la déviance. Si l'AIC du modèle Binomial négatif est inférieur à celui du modèle Poisson, c'est uniquement parce qu'il a une déviance plus faible, et qu'il est donc mieux ajusté aux données d'apprentissage. En effet, les modèles en question sont estimés en minimisant l'inverse de la log-vraisemblance, c'est-à-dire en minimisant la déviance.

## 2.3 Construction du zonier

La dernière partie de ce deuxième chapitre porte sur la construction du zonier et son implémentation dans la tarification. La méthode de construction est rappelée en Section 1.2.4. Les résultats seront ici présentés et analysés.

### 2.3.1 Modèle de prédiction

**Agrégation des résidus à la maille INSEE**

Soit $r_i$ le résidu pour un individu $i$ donné. Soit $z$ une commune quelconque et $I_z$ l'ensemble des polices appartenant à cette commune. Les résidus sont agrégés pour chaque code INSEE de la façon suivante

$$r_z = \frac{\sum_{i \in I_z} r_i}{\sum_{i \in I_z} e_i},$$

où $e_i$ est l'exposition de la police $i$. Par ailleurs, chaque commune est munie d'un poids qui est la somme des expositions de ses polices. En effet, plus l'exposition d'une commune dans le portefeuille est grande, plus le calcul de son résidu agrégé est en théorie robuste. Ce sont ces résidus agrégés qui seront prédits.

Pour rappel, plusieurs résidus ont été introduits : additifs, multiplicatifs, Pearson, et Anscombe. Bien que certains d'entre eux présentent des propriétés intéressantes, il est en principe complexe de déterminer à l'avance lequel permettra d'obtenir le meilleur résultat. La distribution de ces résidus est présentée sur la Figure 2.10.

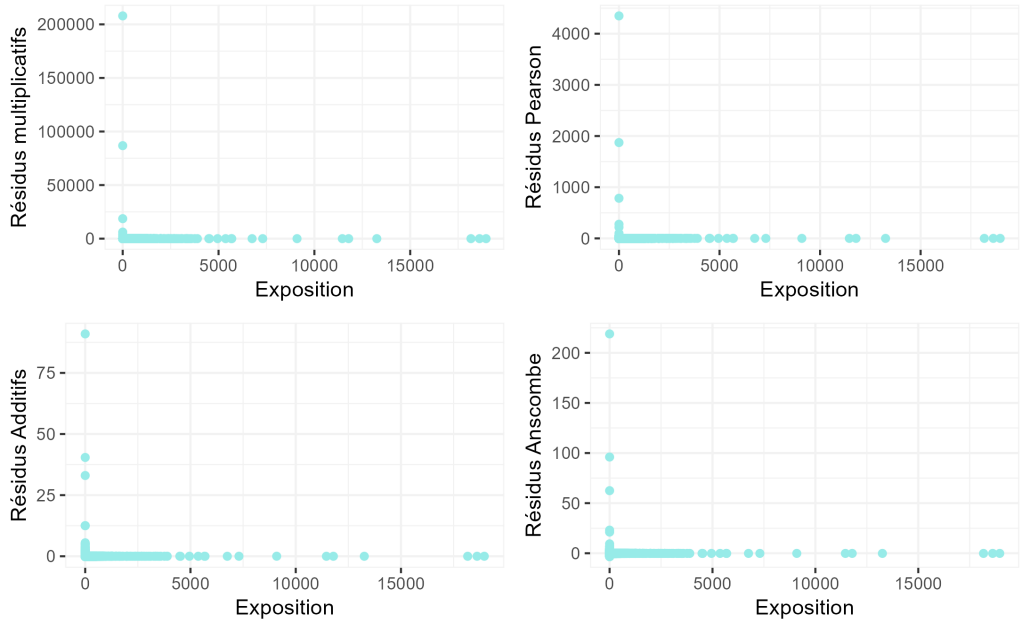



FIGURE 2.10 : Distribution de la variable cible

Quoi qu'il en soit, la distribution est assez atypique. Certaines communes se démarquent nettement des autres en termes de valeurs brutes. Les communes avec des résidus qui explosent et une faible exposition peuvent s'expliquer ainsi : un seul individu du portefeuille est présent dans une commune et a eu un seul sinistre sur une durée d'observation très courte. Ces outliers sont des cas typiques de biais de représentation : un seul individu a été observé sur une période restreinte et a eu au moins un sinistre, faisant ainsi exploser la fréquence.

Ce phénomène est principalement dû à la concentration de l'exposition du portefeuille dans certaines communes proches des bases de défense. Quoi qu'il en soit, la forme de la distribution est la même quel que soit le type de résidus considéré. Seuls les résidus additifs seront considérés.

#### Construction d'un algorithme de prédiction

Chaque commune possède donc un résidu agrégé et plusieurs variables explicatives. Un algorithme de prédiction peut donc être construit. Deux modèles seront confrontés : les *Forêts aléatoires* et le *Gradient Boosting Machine* (GBM). Ces deux algorithmes opposent deux concepts en *Machine Learning* : le *bagging* et le *boosting*. Si le *bagging* correspond à une moyenne de modèles indépendants, le *boosting* est une combinaison linéaire de modèles qui s'adaptent les uns aux autres.

#### Modèle *Random Forest (bagging)*

L'idée du *bagging* est illustrée dans l'Algorithme 2. La démarche est la suivante : diviser le jeu d'entraînement en plusieurs échantillons qui serviront chacun à entraîner indépendamment un sous-modèle. La prédiction finale correspond à la moyenne de la prédiction de tous les sous-modèles obtenus. Dans le cas d'une classification, c'est simplement la classe majoritaire.

**Algorithme 2** Bagging

**Entrée** Échantillon d'apprentissage
**pour** b = 1 à B **:**
Obtenir un échantillon bootstrap $z_b$ de la base d'apprentissage
Estimer le modèle $\hat{f}_{z_b}$ à l'aide de l'échantillon bootstrap
**fin pour**
Pour un individu $x_0$, la prédiction correspond à la moyenne $\hat{f}_B = \frac{1}{B}\sum_{i=1}^{B} \hat{f}_{z_b}(x_0)$ des sous-modèles estimés. Dans le cas d'une classification, c'est la classe majoritaire.

En ce qui concerne la division de l'échantillon d'apprentissage, il n'est pas toujours concevable de diviser le jeu d'apprentissage en plusieurs échantillons indépendants, et notamment si la taille du jeu de données n'est pas suffisante pour rendre l'estimation robuste. L'échantillon construit est donc un échantillon bootstrap qui est une pioche aléatoire avec remise des individus de la base. Certains individus peuvent être présents plusieurs fois au sein d'un même échantillon.

Soit $\rho(x_0) = \text{Corr}(\hat{f}_{z_b}(x_0), \hat{f}_{z_{b'}}(x_0))$ la corrélation pour un individu $x_0$ entre le modèle estimé par $z_b$ et celui estimé par $z_{b'}$. Soit $\text{V}(x_0) = \mathbb{V}[\hat{f}_{z_b}(x_0)]$ la variance du modèle $z_b$. Ces quantités seront supposées identiques quel que soit le couple $(b, b')$ considéré. La variance du modèle agrégé dans le cas de la régression est

$$
\begin{aligned}
\mathbb{V}[\hat{f}_B] := \mathbb{V}[\frac{1}{B}\sum_{i=1}^{B}\hat{f}_{z_b}(x_0)] &= \frac{1}{B^2}\mathbb{V}[\sum_{i=1}^{B}\hat{f}_{z_b}(x_0)] \\
&= \frac{1}{B^2}\sum_{i=1}^{B}\mathbb{V}[\hat{f}_{z_b}(x_0)] + \frac{2}{B^2}\sum_{1\leq i<j\leq B}\text{Cov}(\hat{f}_{z_b}(x_0), \hat{f}_{z_{b'}}(x_0)) \\
&= \frac{1}{B}V(x_0) + \frac{2}{B^2}\sum_{1\leq i<j\leq B}\rho(x_0)V(x_0) \\
&= \frac{1}{B}V(x_0) + \frac{B-1}{B}\rho(x_0)V(x_0) \\
&= \rho(x_0)V(x_0) + V(x_0)\frac{1-\rho(x_0)}{B} \underset{B\to+\infty}{\longrightarrow} \rho(x_0)V(x_0) \leq V(x_0).
\end{aligned}
$$

L'intérêt du bagging donc est de diminuer au maximum la variance de la prédiction à l'aide de sous-modèles qui soient le plus indépendants possible. Ainsi, BREIMAN (2001) propose alors la démarche suivante : appliquer le bagging dans le cas où les sous-modèles sont des arbres de décision (CART). De plus, à chaque itération dans la construction d'un arbre, la variable de partitionnement choisie sera la plus pertinente non pas parmi toutes les variables à disposition, mais parmi un nombre de variables piochées aléatoirement (en général, le nombre de variables piochées est égal à la racine du nombre total de variables). Cette intégration de l'aléa dans le modèle permet de renforcer l'indépendance des sous-modèles entre eux. C'est ce qui donne son nom aux *Random Forest* (Forêts aléatoires). La Figure 2.11 illustre le principe de construction de l'algorithme *Random Forest*.

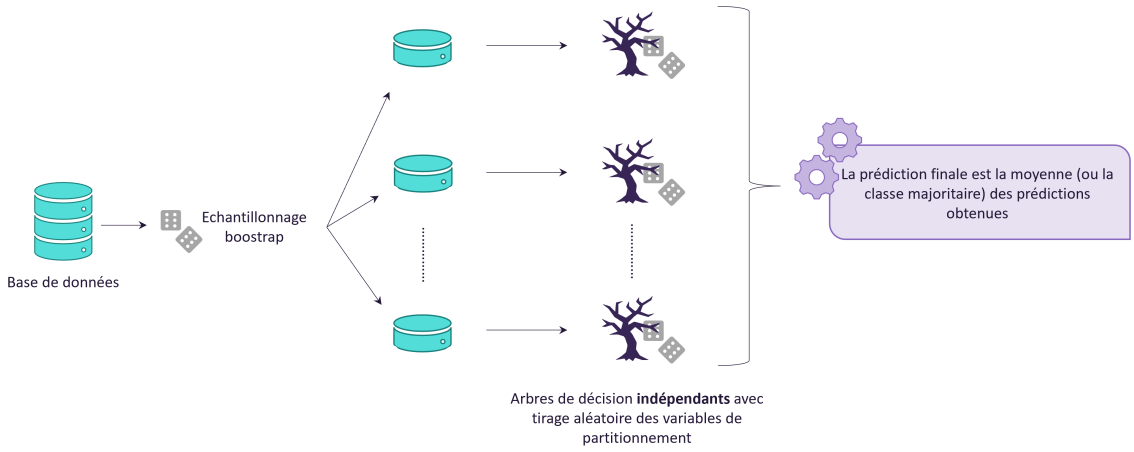



FIGURE 2.11 : Illustration de l'algorithme *Random Forest*

**Modèle *Gradient Boosting Machine (boosting)***

Il existe plusieurs variantes du *boosting*. Celle qui sera utilisée ici est le *Gradient Boosting Machine*, aussi dit GBM. Contrairement au *Random Forest* qui est une moyenne de sous-modèles indépendants, le modèle final obtenu est une combinaison linéaire de sous-modèles (arbres de décisions) construits de manière séquentielle. À chaque itération, le nouveau sous-modèle estimé est une version améliorée du précédent, construit en prédisant les résidus du modèle précédent. La Figure 2.12 illustre le principe de construction de l'algorithme.

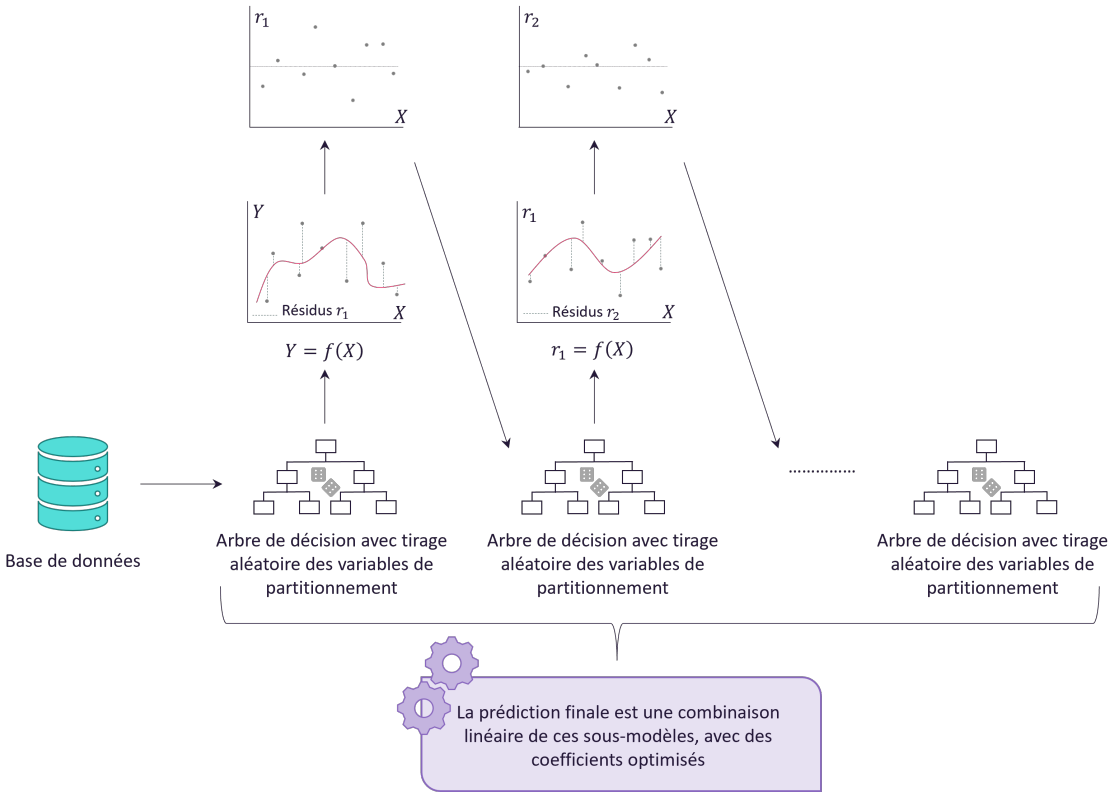



FIGURE 2.12 : Illustration de l'algorithme GBM

À l'origine, ces modèles ont été introduits par FREUND et SCHAPIRE (1997) dans le cas de la classification binaire. Par la suite, ces modèles se sont étendus à la classification multiple et à la régression. Formellement, le modèle final obtenu est

$$\hat{f}_M(x) = \sum_{i=1}^{M} \beta_m b(x, \gamma_m),$$

où $M$ est le nombre de sous-modèles, $(\beta_m)_{1 \leq m \leq M}$ sont les coefficients de pondération des sous-modèles, et la fonction $x, \gamma \mapsto b(x, \gamma)$ est un sous-modèle dont la structure dépend de $\gamma$. Ainsi, pour une fonction perte $l$, l'objectif est de résoudre le problème suivant

$$\min_{\{\beta_m, \gamma_m\}_{m=1}^{M}} \sum_{i=1}^{n} l\Big(Y_i, \sum_{i=1}^{M} \beta_m b(X_i, \gamma_m)\Big).$$

Dans le cas de la régression, c'est généralement la fonction de perte quadratique qui est utilisée. Le problème étant bien trop compliqué à résoudre, une procédure itérative est mise en place afin d'approcher la solution.

FRIEDMAN (2001) propose alors un algorithme appelé MART[6] qui par la suite sera généralisé sous le nom de *Gradient Boosting Models* (GBM). Ces algorithmes peuvent se baser sur différentes fonctions coût selon la nature de la variable cible. Les sous-modèles sont des arbres CART, mais la particularité de ces algorithmes est que leur optimisation est fondée sur le gradient de la fonction coût (d'où son nom). La méthode de construction est présentée dans l'Algorithme 3

**Algorithme 3** Gradient Boosting Machine pour la régression

Initialisation : $\hat{f}_0(x) = \arg\min_{\gamma \in \mathbb{R}} \sum_{i=1}^{n} l(Y_i, \gamma)$
**pour** $m = 1$ *à* $M$ **:**
  Calculer : $r_{i,m} = -\left[\frac{\partial l(Y_i, f(X_i))}{\partial f(X_i)}\right]_{f=\hat{f}_{m-1}}; i = 1, \ldots, n$
  Ajuster un arbre de régression $\delta_m$ sur $(X_i, r_{i,m})_{i=1,\ldots,n}$
  ▷ *Les feuilles terminales seront notées* $(R_{j,m}, j = 1, \ldots, J_m)$.
  Calculer : $\gamma_{j,m} = \arg\min_{\gamma \in \mathbb{R}} \sum_{X_i \in R_{j,m}} l(Y_i, \hat{f}_{m-1}(X_i) + \gamma)$
  Mise à jour : $\hat{f}_m(x) = \hat{f}_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{j,m} \mathbb{1}_{x \in R_{j,m}}$
**fin pour**

Le paramètre $\nu$ est un paramètre qui détermine la contribution de chaque arbre dans la prédiction finale. Trois hyperparamètres sont à calibrer : le nombre d'arbres $M$ (ou de sous-modèles), le paramètre $\nu$, et l'expression de la fonction coût. Ces modèles sont calibrés par validation croisée. La commande *gbm* du package *gbm* réalise la procédure. Les meilleurs résultats ont été obtenus avec une fonction coût "Laplace", c'est-à-dire la différence en valeur absolue de la variable cible et la valeur prédite.

**Résultats des algorithmes de prédiction**

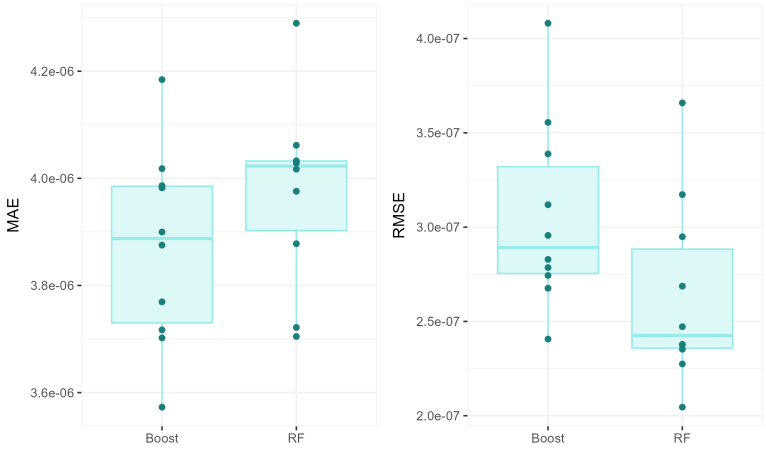



FIGURE 2.13 : Résultats de la validation croisée MC au test

Afin de confronter les deux modèles de prédiction, une validation croisée Monte Carlo est mise en place. Les résultats sont obtenus à la Figure 2.13. En ce qui concerne la MAE, de meilleurs résultats sont obtenus avec l'algorithme de *boosting*. Ces résultats semblent cohérents puisque le modèle a été calibré avec une fonction

[6] Multiple Additive Regression Trees

coût de Laplace, ce qui correspond aussi à l'expression de la MAE. À l'inverse, c'est le *Random Forest* qui performe le mieux du point de vue de la RMSE. Cependant, quel que soit le critère choisi, la distribution des erreurs pour le *Random Forest* semble moins volatile que celle du GBM. Le modèle *Random Forest* sera donc conservé.

**Remarque sur l'intégration de la variable de distance à une base de défense**

Une des sorties de l'algorithme *Random Forest* est l'importance des variables. Pour rappel, pour chaque arbre construit, des variables de partitionnement sont sélectionnées aléatoirement parmi l'ensemble des variables disponibles. L'importance d'une variable représente en quelque sorte le nombre de fois que cette variable a été choisie comme variable de partitionnement, chaque fois qu'elle faisait partie d'une pioche. Une variable importante signifie donc que chaque fois qu'elle est présente dans une pioche, elle sera souvent sollicitée pour le partitionnement. C'est donc une variable avec un fort pouvoir discriminant. L'importance des variables en sortie est affichée sur la Figure 2.14.

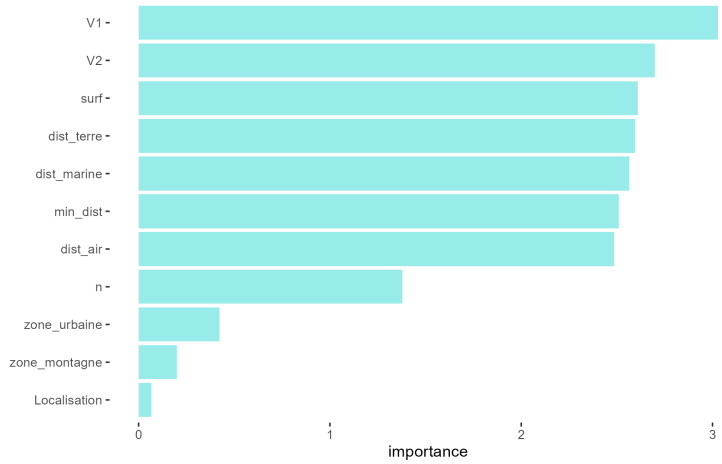



FIGURE 2.14 : Importance des variables par la méthode RF

Les variables qui représentent la distance à la base la plus proche (marine pour dist_marine, terrestre pour dist_terre, aérienne pour dist_air, et la plus proche des trois pour min_dist) sont des variables très sollicitées par l'algorithme. Les variables[7] les plus discriminantes sont les variables démographiques V1 et V2. L'intégration des variables de distance est donc pertinente du point de vue de l'algorithme, et lui permet de prendre en compte une forme de lissage implicite.

### 2.3.2 Prédiction, Lissage, et classification

**Corrélation géographique des prédictions**

L'étude est désormais focalisée sur les résidus prédits et toutes les communes ont désormais le même poids. Ces nouveaux résidus sont cartographiés sur la Figure 2.15.

[7] La variable n représente le taux de sinistralité de la zone.

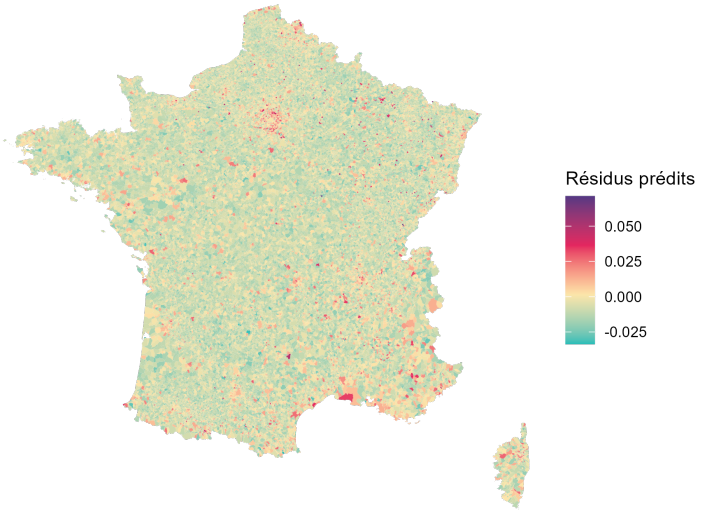



FIGURE 2.15 : Cartographie des résidus prédits

Les prédictions obtenues semblent avoir lissé la distribution des résidus initiaux. Puis, afin d'apprécier la dépendance géographique de ces résidus, un variogramme est construit. Le variogramme est une fonction qui permet d'analyser la corrélation géographique d'une certaine grandeur $Z(x)$ qui dépend de la position géographique $x$. Formellement, elle se définit de la façon suivante

$$\gamma(h) = \frac{1}{2}\mathbb{E}_{|x-y|=h}[(Z(x) - Z(y))^2].$$

Néanmoins, du fait de la discrétisation des données, il est impossible de définir une fonction pour tout $h > 0$. Une façon de faire est de définir un seuil $\delta > 0$ et de calculer le variogramme empirique défini par

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{i,j \in n(h)} (Z(x_i) - Z(x_j))^2,$$

où $n(h)$ est le nombre de paires de points dont la distance est comprise entre $h - \delta h$ et $h + \delta h$.

Cependant, appliquer cette formule sur le jeu de données requiert une trop grande quantité de mémoire, notamment pour le calcul des distances. En effet, pour environ 36 000 communes, il sera nécessaire de calculer une matrice de distance intercommunale de taille 36 000×36 000, soit de l'ordre de $10^9$ composantes matricielles. Afin de contourner ce problème, une logique de voisinage sera adoptée pour la distance. La distance $h$ est simplement remplacée par le $v^{\text{ième}}$ plus proche voisin. Le variogramme empirique devient

$$\hat{\gamma}(v) = \frac{1}{2n} \sum_{i=1}^{n} (Z(x_i) - Z(x_{ppv(i,k)}))^2,$$

où $ppv(i, v)$ est l'indice du $v^{\text{ième}}$ plus proche voisin de $i$ et $n$ le nombre de communes.

A l'aide de la fonction *get.knn* du package *FNN*, il est possible de récupérer directement pour toutes les communes les $k$ plus proches voisins. Dans cette étude, la corrélation est étudiée pour un nombre maximal de 100 voisins. Les résultats sont illustrés à la Figure 2.16.

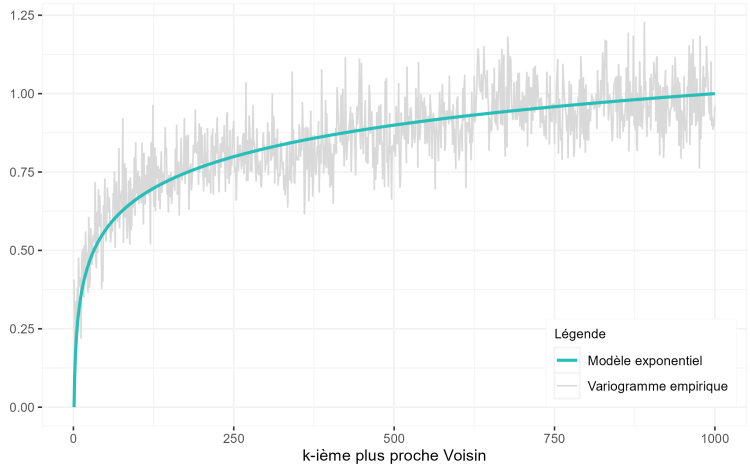



FIGURE 2.16 : Variogramme empirique des résidus

Le variogramme est caractérisé par trois composantes :

- **Le palier** : c'est le niveau maximal atteint par la fonction. Généralement, lorsqu'un phénomène de dépendance géographique est étudié, cette dépendance s'estompe à mesure que la distance entre les points s'éloigne. Cet effet se traduit par un plateau atteint à la limite du variogramme.
- **La portée** : c'est la distance (ici le nombre de voisins) à partir de laquelle le palier est atteint (approximativement). En réalité, le palier n'est jamais vraiment atteint. La portée est une distance à partir de laquelle la courbure de la fonction est suffisamment faible pour considérer qu'il s'agit d'un plateau.
- **L'effet pépite** : en théorie, à mesure que la distance entre les points diminue, la différence de valeurs entre les points doit tendre vers 0. Cependant, face à des données expérimentales, ce n'est pas toujours le cas. D'une part, il n'est pas possible d'obtenir des informations sur la corrélation entre des points dont la distance est inférieure à la distance minimale observée sur le jeu de données. D'autre part, s'il existe une corrélation géographique globale, il existera une forme de bruit à l'échelle locale. L'intensité de ce bruit est traduite par l'amplitude de l'effet pépite.

Selon la forme du graphique obtenu, plusieurs modèles permettent d'estimer le variogramme. Ici, il sera modélisé par le modèle exponentiel défini ainsi

$$V(x) = C(1 - e^{-\frac{x}{a}}),$$

où $C$ représente le palier, et $a$ la portée. Ces paramètres sont estimés par $\hat{C}$ et $\hat{a}$ de sorte que

$$\hat{C}, \hat{a} \in \arg\min \frac{1}{100} \sum_{v=1}^{100} (\hat{\gamma}(v) - V(v))^2.$$

L'algorithme BFGS (Méthode de Broyden-Fletcher-Goldfarb-Shanno) est utilisé pour optimiser ces paramètres, avec la fonction *optim* de R.

L'algorithme BFGS est une procédure itérative qui permet de résoudre des problèmes d'optimisation à l'aide d'une descente de gradient. L'idée de la méthode est d'utiliser une approximation de la matrice Hessienne. La procédure est présentée dans l'Algorithme 4. La condition de convergence peut être une condition du type $|\nabla f(x_k)| < \epsilon$ avec $\epsilon > 0$ un nombre petit.

**Algorithme 4** Optimisation BFGS

**Entrée** $f, \nabla f, x_0$

Initialisation de la matrice Hessienne $B_0 = I$ (matrice identité) et $k = 0$.

**tant que** Condition de convergence non atteinte **:**

Calculer $p_k$ tel que $B_k p_k = -\nabla f(x_k)$

Calcul du pas optimal $\alpha_k$ par recherche linéaire

Calculer $x_{k+1} = x_k + \alpha_k p_k = x_k + s_k$

Calculer $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$

Mettre à jour la matrice Hessienne $B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}$

**fin tant que**

### Lissage par crédibilité

L'objectif du lissage est d'obtenir un variogramme qui soit plus lisse, c'est-à-dire que la courbe soit moins volatile. Le lissage appliqué est inspiré du lissage par crédibilité[8]. Soit $r_z$ le résidu de la commune $z$ et $r_z^*$ son résidu crédibilisé, alors

$$r_z^* = \alpha r_z + (1-\alpha)\frac{1}{\sum_{v \in ppv(z,k)} \frac{1}{d(z,v)^n}} \sum_{v \in ppv(z,k)} \frac{1}{d(z,v)^n} r_v,$$

où $\alpha$ est le facteur de crédibilité, $ppv(z,k)$ est l'ensemble des $k$ plus proches voisins de la commune $z$, $d(z,v)$ est la distance entre la commune $z$ et le voisin $v$, et $n$ un paramètre qui ajuste le poids donné à la distance entre les communes. Il y a trois paramètres dans ce modèle : $\alpha$, $n$, et le nombre maximal de voisins $k$. Ce dernier sera fixé comme étant la portée calculée précédement, soit environ 243.

L'enjeu est maintenant de savoir quel paramètre choisir pour $\alpha$ et $n$. En théorie, plus $\alpha$ se rapproche de 0, plus le lissage est important. Le variogramme se rapprochera d'une fonction qui croît linéairement vers le palier puis se stabilise. Par ailleurs, si le degré $n$ est trop grand, le lissage d'une commune sera surtout impacté par la valeur de son voisin direct. Le couple $(\alpha, n)$ est choisi de sorte à favoriser le lissage sans pour autant déformer la courbure du variogramme théorique estimé. En notant $\hat{\gamma}^*$ le variogramme des résidus lissés, les estimations $\hat{\alpha}$ et $\hat{n}$ sont calculées de sorte que

$$\hat{\alpha}, \hat{n} \in \arg\min \frac{1}{100} \sum_{v=1}^{100} (\hat{\gamma}^*(v, \alpha, n) - V(v))^2.$$

La méthode employée est un *gridsearch* sur un ensemble prédéfini de valeurs possibles pour $\alpha$ et $n$. Le meilleur résultat est obtenu pour un $\alpha$ à $0,5$ et un $n$ à 2.

Les résultats du lissage sont présentés en Figure 2.17. Les résidus obtenus sont bien plus lisses et le dégradé obtenu est plus fondu sur la carte. Désormais, les résidus contiennent bien l'information géographique en tenant compte du voisinage. Aussi, le variogramme obtenu (en violet) est moins volatile et se rapproche tout autant du modèle de corrélation exponentiel estimé. La dernière étape est de classifier ces résidus. La méthode *kmeans* (MACQUEEN (1967)) est appliquée. Le nombre de classes choisi est de 10 afin de respecter la politique de l'AGPM. Le zonier obtenu est présenté sur la Figure 2.18.

[8] Généralement, les méthodes de lissage par crédibilité prennent en compte l'exposition des assurés dans la commune. Cependant, puisque les assurés ne sont pas uniformément répartis sur le territoire, il a été choisi d'appliquer le lissage sur les valeurs prédites qui ont toutes une exposition de 1.

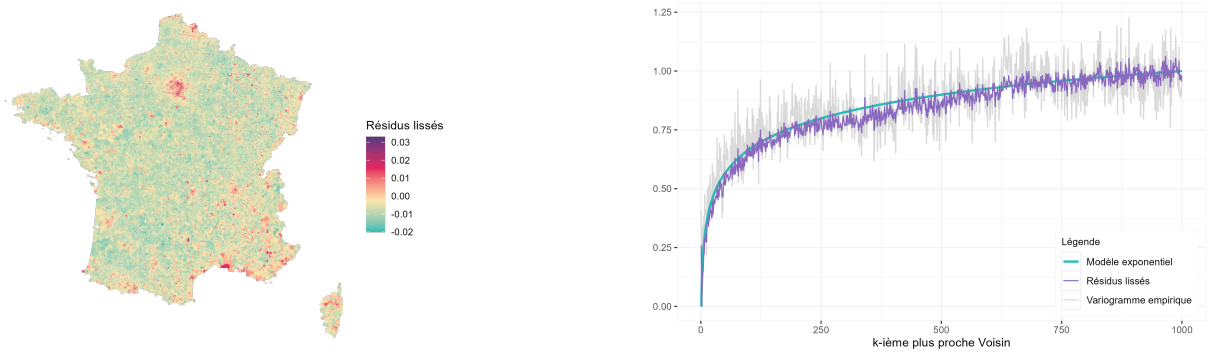



FIGURE 2.17 : Résultat de la méthode de lissage sur les résidus

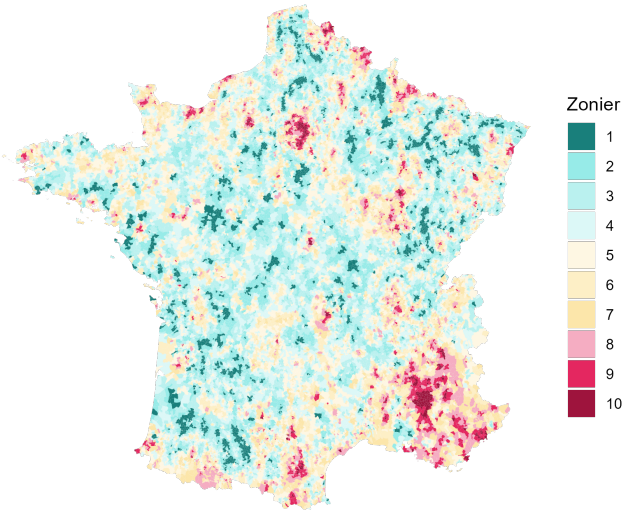



FIGURE 2.18 : Cartographie du zonier

### 2.3.3 Intégration du zonier dans la tarification

#### Performances

Après intégration du zonier comme nouvelle variable tarifaire, les performances des modèles Poisson et Binomiale négatif sont comparées sur la base test (2022). Les résultats sont présentés sur la Table 2.2.

| Modèle | MAE | MSE |
|---|---|---|
| GLM sans variable géographique | 0,04525288 | 0,02283927 |
| GLM avec ancien zonier | 0,04523357 | 0,02282251 |
| GLM avec nouveau zonier | 0,04523234 | 0,02275856 |

TABLE 2.2 : Performance sur la bases test du modèle fréquence

Bien que la différence soit très légère, il y a bien une amélioration du modèle par intégration du nouveau zonier. Aussi, il semblerait que le nouveau zonier construit permette d'obtenir un modèle plus performant qu'avec l'ancien zonier. Cette progression peut être expliquée par l'ajout de nouvelles données, notamment sur la localisation des bases de défenses.

**Interprétation des variables**

Le but est d'interpréter les 75 coefficients estimés en sortie du GLM Poisson. Chaque modalité de chacune des variables est devenue une variable à part entière du point de vu du GLM. C'est la dummification[9]. A titre d'exemple, pour un GLM appliqué sur une variable qualitative $x$ à $m$ modalités $x_1, \ldots, x_m$, le modèle s'exprime de la façon suivante

$$g(\mathbb{E}[Y|X=x]) = \theta_0 + \theta_1 \mathbb{1}_{\{x=x_1\}} + \ldots + \theta_m \mathbb{1}_{\{x=x_m\}}.$$

Pour $p$ variables qualitatives $x^1, \cdots, x^p$ où chaque variable $j$ détient $m_j$ modalités, il est possible d'écrire la formule reliant l'espérance aux paramètres de la façon suivante

$$g(\mathbb{E}[Y|X=x]) = \theta_0 + \begin{pmatrix} \theta_{1,1}\mathbb{1}_{\{x^1=x_1^1\}} \\ \vdots \\ \theta_{1,m_1}\mathbb{1}_{\{x^1=x_{m_1-1}^1\}} \end{pmatrix} + \cdots + \begin{pmatrix} \theta_{p,1}\mathbb{1}_{\{x^p=x_1^p\}} \\ \vdots \\ \theta_{p,m_p}\mathbb{1}_{\{x^p=x_{m_p-1}^p\}} \end{pmatrix}.$$

La lecture de cette expression est très simple : pour chaque variable, la ligne choisie correspondant à la modalité représentant l'individu et seul un coefficient est considéré (puisque l'indicatrice vaut 0 ou 1). Cependant, il y a toujours une modalité en moins pour chaque variable qui sera captée en intercept.

Soient deux individus $x_1$ et $x_2$ identiques sur toutes les variables sauf une seule notée $x^j$. Les modalités respectives pour cette variable seront notées $\theta_{j,k}$ et $\theta_{j,k'}$. Dans le cas d'un GLM Poisson avec le lien logarithmique, le rapport de leur estimation devient

$$\frac{\mathbb{E}[Y|X=x_1]}{\mathbb{E}[Y|X=x_2]} = e^{\theta_{j,k}-\theta_{j,k'}}.$$

Sur ce principe, l'impact marginal d'une variable sur la tarification peut être quantifié en calculant la moyenne de la différence en valeur absolue de chacun des couples de coefficients possibles. Pour une variable $\theta_j$, l'indice de l'effet marginal est calculé ainsi

$$\frac{1}{C_{m_j-1}^2 + m_j - 1}\left( \sum_{1\leq k<k'\leq m_j-1} |\theta_{j,k} - \theta_{j,k'}| + \sum_{m=1}^{m_j-1} |\theta_{j,m}| \right),$$

où $C_{m_j-1}^2$ est le nombre de combinaisons de 2 modalités qu'il est possible de faire avec la variable $j$. La somme de droite est rajoutée afin de prendre en compte les cas où la modalité prise par l'un des deux individus est en intercept.

Une fois la méthode appliquée au jeu de données, les résultats obtenus sont ceux de la Figure 2.19. C'est la variable Usage qui impacte le plus la tarification, c'est-à-dire que toutes choses égales par ailleurs, si deux individus diffèrent pour la variable usage, l'impact sera en moyenne plus important qu'une différence sur une autre variable.

[9]Méthode qui consiste à transformer une variable qualitative de sorte à ce que chacune de ses modalités devienne une variable quantitative prenant la valeur 0 ou 1. Notons que deux modalités ne peuvent pas être prises en même temps.

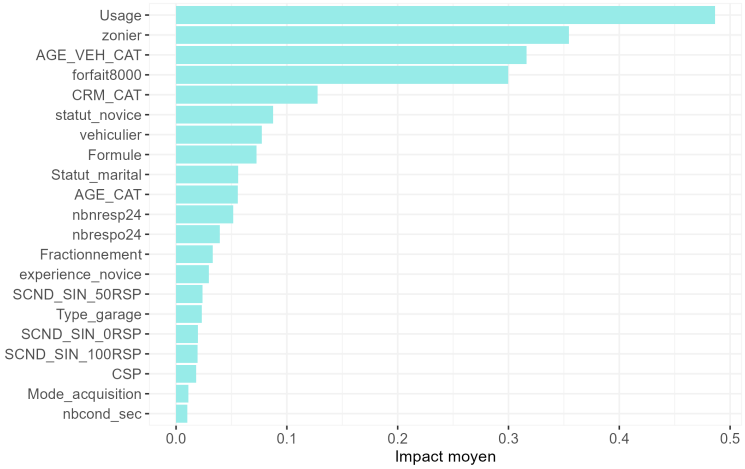



FIGURE 2.19 : Impact des coefficients dans la tarification

**Conclusion**

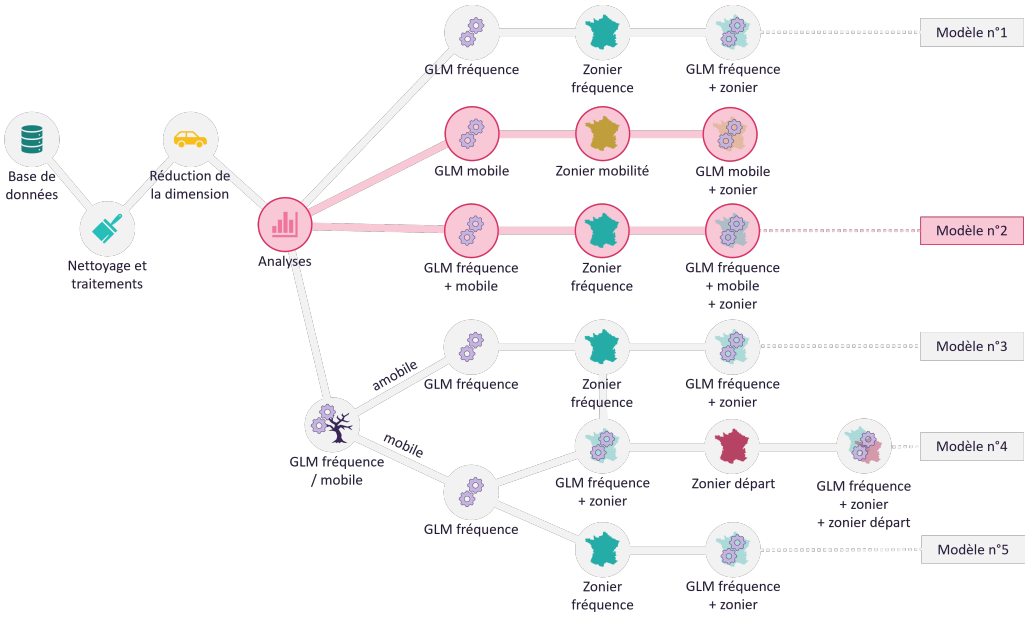
L'objectif de ce chapitre est de mettre en place un premier modèle de tarification qui ne prend pas en compte la mobilité. La prochaine étape est de modéliser ce phénomène de deux façons :

- mettre en place d'un modèle pour prédire la probabilité qu'une police soit mobile. Ce modèle permettra de contruire un zonier "mobilité" afin de repérer les zones ayant un fort taux de mobilité.
- intégrer la variable mutation dans le modèle de tarification.

# Chapitre 3

# Modélisation et intégration de la mobilité dans la tarification

*Jusqu'à présent, c'est un modèle de tarification ne prenant pas en compte le phénomène de mobilité qui a été construit. Pour rappel, l'objectif est de modéliser et d'étudier l'impact de ce phénomène à travers le modèle de fréquence, tout en proposant une tarification qui soit plus juste et équitable. Une façon simple de procéder est d'intégrer une variable mutation au GLM construit. Cette approche sera par la suite comparée à d'autres modèles plus complexes (chapitre 4).*



## 3.1 Analyse descriptive du phénomène

Deux approches seront adoptées pour l'analyse descriptive : une étude des individus mobiles entre eux, et une étude comparative entre les individus mobiles et ceux qui seront dit "amobiles"[1].

### 3.1.1 Statistiques et cartographie

#### La mobilité en quelques chiffres

Pour rappel, l'étude est portée sur une période de 2016 à 2022, sans prise en compte de l'année 2020 liée au COVID. Sur cette période, la mobilité est un phénomène relativement présent dans le portefeuille. Chaque année, c'est en moyenne 7,44% des contrats renouvelés, ou nouveaux entrants, qui correspondent à des individus ayant changé de commune. Un pic de mobilité est observé en 2018 avec 8,2% des polices qui sont mobiles. De façon plus large, ce phénomène concerne 18,69% des polices de l'AGPM sur cette période.

En ce qui concerne le nombre de mutations par individu, la majorité des individus mobiles, soit 77,6%, a muté une seule fois sur la période d'observation. Cependant, il faut bien noter que l'intérêt n'est pas porté sur le nombre de mutations mais sur les critères qui permettent de prédire la probabilité qu'une police soit mobile ou pas.

#### Transitions intra-métropole

Le phénomène de mobilité est caractérisé par les différentes transitions des assurés entre les communes. Les transitions qui ont lieu au sein de la Métropole sont représentées sur la Figure 3.1. Elles sont illustrées par des segments incurvés dans le sens trigonométrique.[2]

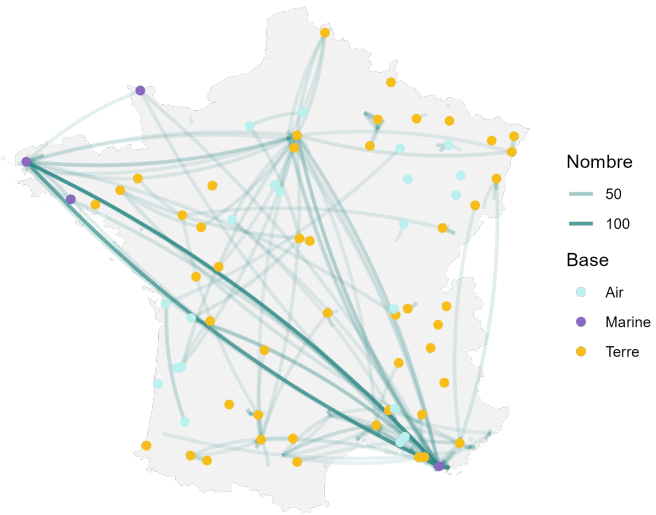



FIGURE 3.1 : TOP 500 des transitions intra-Métropole

Le flux le plus important est bien le réseau Brest-Toulon. Dans la mesure où l'AGPM est basée à Toulon, et

[1] Ceux qui ne sont pas mobiles.

[2] Pour deux points géographiques, la lecture des segments de transitions se fait dans le sens inverse des aiguilles d'une montre.

que la plupart des assurés sont des militaires de la marine, il est logique de retrouver ce flux. Le triangle entre Brest, Paris, et Toulon représente un réseau important. Néanmoins, d'autres flux sont tout aussi présents. Il existe aussi une forte mobilité entre les différentes communes du sud-est. De façon plus large, il est intéressant de noter que la plupart des transitions se font d'une base à l'autre, c'est-à-dire que dans la majorité des cas, les bases de défense sont les points de départ ou d'arrivé des flux transitoires. Enfin, dans la plupart des cas, les transitions semblent symétriques dans les deux sens, c'est-à-dire qu'il y a presque autant de transition dans un sens que dans l'autre entre deux communes.

#### Transitions inter-métropole

Une autre particularité du phénomène de mobilité dans le portfeuille est l'importance des transitions entre la Métropole et les autres départements d'Outre-Mer. Ces transitions sont représentées sur la Figure 3.2.

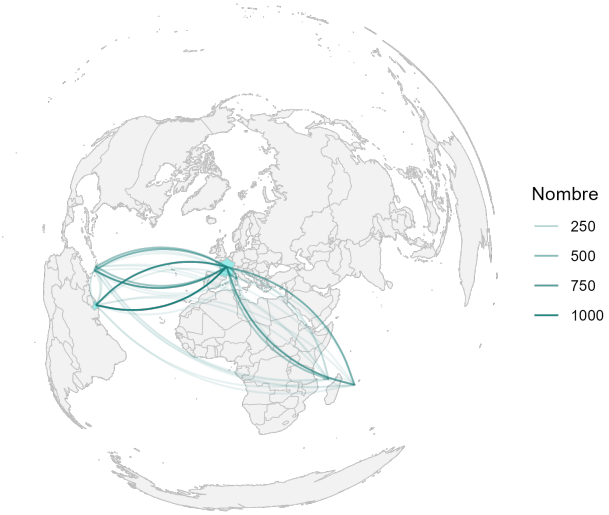



FIGURE 3.2 : Transition inter-Métropole

Géographiquement, ces transitions relient la métropole à trois points majeurs : les antilles avec la Martinique et la Guadeloupe, la Guyane, et le sud-est de l'Afrique avec Mayotte et l'île de la Réunion. Aussi, les transitions entre la Guyane et la métropole représentent les flux les plus importants du portefeuille, quelque soit le sens considéré. Ces transitions sont intéressantes car l'impact du changement d'envrionnement entre ces zones est en théorie plus important qu'une transition sein de la Métropole. Une attention sera donc portée sur ces flux à grande échelle.

### 3.1.2 Population mobile vs population amobile

#### Comparaison démographique

Les deux populations peuvent être comparées sur des critères démographiques. Encore une fois, l'étude est uniquement focalisée sur l'âge et la catégorie socioprofessionnelle.

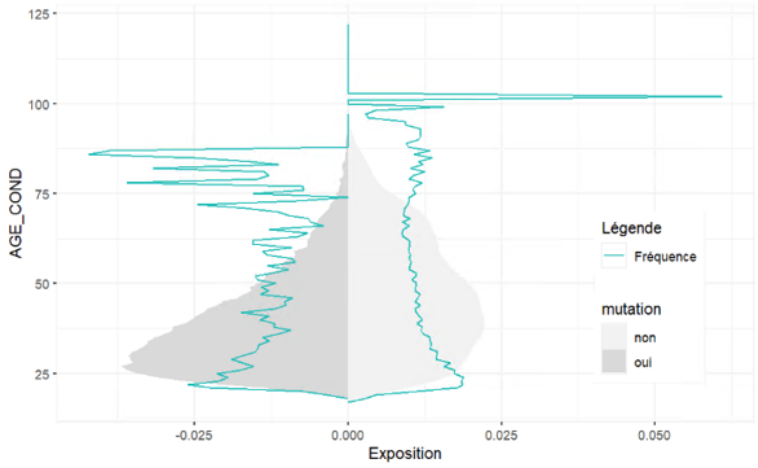



FIGURE 3.3 : Représentation par âge

La Figure 3.3 permet de comparer les deux populations du point de vue de l'âge[3]. L'exposition affichée n'est pas absolue mais relative, c'est-à-dire que seule la forme de la distribution de l'exposition peut être comparée. Ainsi, la population mobile semble être représentée en majorité par des jeunes avec un pic vers 26 ans, tandis que la population amobile est représentée par un spectre plus large avec un pic vers 42 ans.

La fréquence de sinistres est représentée en absolue. Ainsi, en terme de sinistralité, les deux populations sont presque identiques, avec une sinistralité légèrement plus importante chez les jeunes mobiles. La volatilité aux âges élevés ne permet pas de définir une comparaison fiable puisque les mobiles âgés sont très peu présents dans le portefeuille.

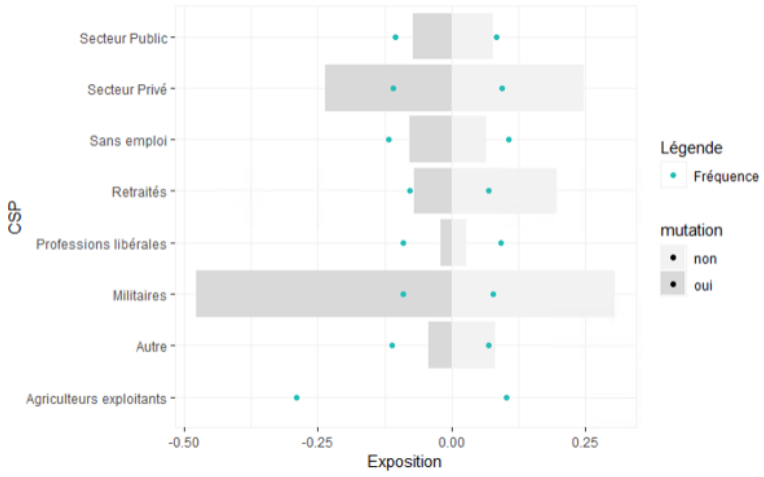



FIGURE 3.4 : Représentation par CSP

En ce qui concerne la catégorie socioprofessionelle (Figure 3.4), les deux populations sont principalement composées de militaires, avec une part plus importante chez les mobiles. Ce résultat reste cohérent avec la

[3]Cette figure est inspirée de la pyramide des âges, graphique souvent utilisé dans les études démographiques. Usuellement, elle permet d'analyser la répartition par âge et par sexe d'une population à un instant donné. Cependant, la distinction ici ne se fait pas selon le sexe mais selon si la population est mobile ou pas.

structure du portefeuille. Aussi, il est intéressant de remarquer que le secteur privé représente aussi une part importante de la population mobile. En ce qui concerne la sinistralité, les deux populations semblent très proches. Une différence est notable pour les individus "Agriculteurs exploitants" mais leur faible exposition là encore ne permet pas de porter une analyse qui soit fiable.

Si le portefeuille de l'AGPM est spécialisé dans le profil militaire, l'analyse montre que la population mobile est encore plus spécifique : il s'agit en grande partie d'une population militaire jeune. Ainsi, ces critères pourraient ne plus être suffisament discriminants du point de vue de la sinistralité chez les mobiles.

**Comparaison de la sinistralité**

Enfin, il est possible de comparer les deux populations du point de vue de la sinistralité. Bien que les mobiles représentent 3% des sinistres survenus, la Figure 3.5 permet de montrer que la fréquence de sinistres chez les mobiles est supérieure à celle des amobiles.

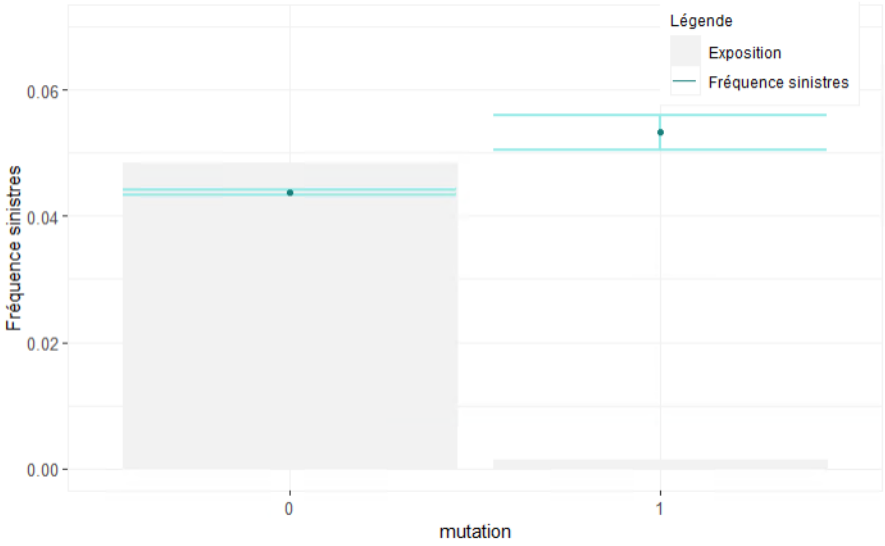



FIGURE 3.5 : Comparaison de la fréquence sinistre globale

Cependant, la comparaison démographique a montré que la population mobile est surtout représentée par des jeunes conducteurs et des militaires. Si ces deux facteurs sont généralement des cas qui aggravent la sinistralité, il est donc primordiale de s'assurer que cette sursinistralité chez les mobiles n'est pas uniquement dû à sa structure démographique dont les facteurs sont déjà pris en compte dans le modèle de référence. Il serait donc inutile de discriminer les assurés selon s'ils sont mobiles ou pas.

Une façon de mettre en avant l'impact réel de la mobilité est d'analyser le ratio "Attendu sur Estimé" (rapport A sur E) des deux populations. Ce dernier correspond au rapport de la somme des valeurs cibles sur la somme des valeurs prédites par le modèle pour une population $\mathcal{A}$ donnée

$$\text{Ratio A/E} = \frac{\sum_{i\in\mathcal{A}} Y_i}{\sum_{i\in\mathcal{A}} \hat{\mu}_i}.$$

Ce ratio s'interprète de la façon suivante :

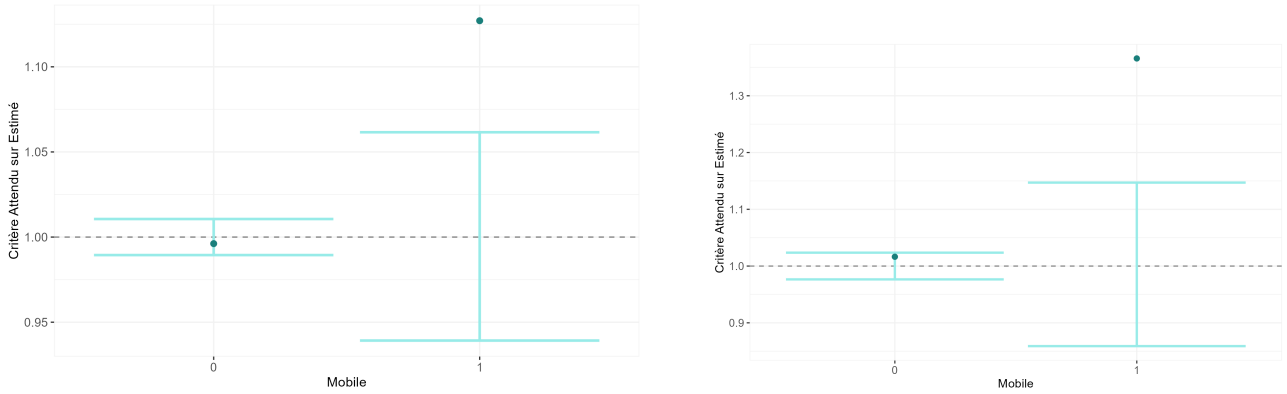
- Si le ratio est égal à 1, le modèle renvoie des prédictions proches de la valeur cible.
- Si le ratio est supérieur à 1, le modèle a tendance à sous-estimer la sinistralité des individus.

– Si le ratio est inférieur à 1, le modèle a tendance à surrestimer la sinistralité des individus.

La thèse de LE BASTARD (2024) fournit un intervalle de confiance pour ce critère. En notant $A_{\mathcal{A}}$ la somme des valeurs cibles attendues et $E_{\mathcal{A}}$ la somme des valeurs cibles estimées pour une population $\mathcal{A}$ donnée, l'hypothèse nulle testée est $H_0 : A_{\mathcal{A}} \sim \mathcal{P}(E_{\mathcal{A}})$. Sous cette hypothèse, l'inégalité suivante est vérifiée

$$\mathbb{P}\Big(\frac{q_{\alpha/2}(E_{\mathcal{A}})}{E_{\mathcal{A}}} \leq \frac{A_{\mathcal{A}}}{E_{\mathcal{A}}} \leq \frac{q_{1-\alpha/2}(E_{\mathcal{A}})}{E_{\mathcal{A}}}\Big) \geq 1-\alpha,$$

où $q_\gamma(E_{\mathcal{A}})$ est le $\gamma$ quantile d'une loi de Poisson de moyenne $E_{\mathcal{A}}$, et $\alpha$ est le seuil de confiance usuel de $5\%$.




(a) Comparaison à l'entraînement

(b) Comparaison au test

FIGURE 3.6 : Comparaison de la performance du modèle selon la mobilité

La Figure 3.6 montre que ce ratio est proche de 1 pour les amobiles mais significativement supérieur à 1 chez les mobiles, que ce soit à l'entraînement ou au test. Le risque des assurés mobiles est donc significativement sous-estimé par le modèle. Ainsi, cette hausse de la sinistralité chez les mobiles n'est pas uniquement dû à un effet de l'âge ou de la catégorie socioprofessionnelle, mais il y a bien un facteur sous-jacent propre au phénomène qui n'est pas capté par le modèle. Par ailleurs, l'estimation du risque des amobiles ne semble pas être significativement impacté par la présence des mobiles dans le modèle. En effet, les amobiles sont majoritaires dans le portefeuille, et ont donc une inertie plus forte. C'est pour cette raison que l'intervalle de confiance de cette population est plus précise que celle des mobiles dans les deux cas.

## 3.2 Modèles de prédiction

L'objectif est désormais de mettre en place un modèle permettant de prédire la probabilité qu'un assuré souscripteur, ou renouvelant son contrat, soit un mobile. Il s'agit de mettre en place une regression logistique afin de construire un zonier en se basant sur les résidus du modèle.

### 3.2.1 Quelques rappels théoriques

#### La régression logistique

La variable cible est une variable à deux états : 0 pour amobile et 1 pour mobile. Cette variable suit une loi de Bernouilli dont le paramètre dépend des caractéristiques de l'assuré. Aussi, la loi de Bernouilli fait partie de la famille des lois exponentielles présentées en Section 1.2.3. Il est donc possible d'appliquer un GLM : c'est la régression logistique.

Ce modèle doit son nom à la fonction de lien qui lui est appliquée : la fonction logistique. La réciproque de cette fonction est la fonction $x \mapsto \frac{e^x}{1+e^x}$. Elle est représentée sur la Figure 3.7.

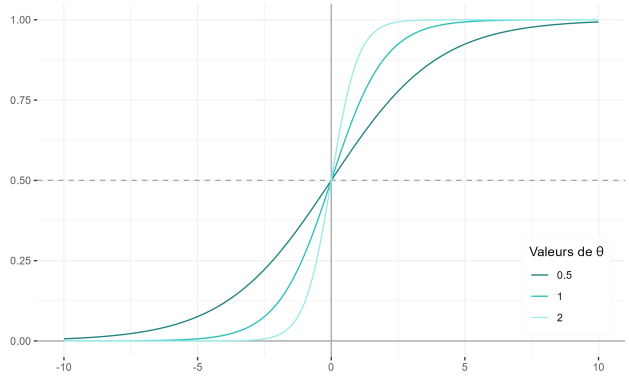



FIGURE 3.7 : Fonction logistique

Cette fonction est à valeur dans $]0;1[$. L'avantage est qu'elle permet de modéliser des valeurs de probabilités, en l'occurence ici la probabilité d'être un mobile ou pas. En effet, l'espérance d'une loi de Bernouilli est égale à la probabilité que la variable aléatoire prenne l'état positif, ici mobile, d'où l'intérêt d'estimer la fréquence par une valeur comprise entre 0 et 1. L'espérance de ce modèle s'exprime donc de la façon suivante

$$\mathbb{E}[Y|X=x] = \frac{e^{\omega(x)}}{1+e^{\omega(x)}},$$

où $x$ est un vecteur contenant l'ensemble des caractéristiques de l'assuré, et $\omega(x)$ est le paramère naturel.

**Interprétation du modèle par odds ratio**

La méthode généralement employée pour interpréter la régression logistique est le calcul des odds ratio. L'odds ratio se définit comme un rapport de "chances". Les "chances" pour qu'un individu soit classé en 1 peut être défini par

$$\text{odds}(x) = \frac{\pi(x)}{1-\pi(x)},$$

où $\pi(x) = \mathbb{P}(Y=1|X=x) = \mathbb{E}[Y|X=x]$. Par exemple, si $\pi(x) = \frac{1}{4}$, l'odds vaut $\frac{1}{3}$, ce qui signifie qu'il y a "une chance sur trois" pour que l'individu soit classé en 1. Puis, l'odds ratio est simplement le rapport de chance entre deux individus $x$ et $\tilde{x}$ et

$$\text{OR}(x,\tilde{x}) = \frac{\text{odds}(x)}{\text{odds}(\tilde{x})}.$$

Le calcul du odds ratio permet de mesurer l'impact d'une variable sur la prédiction. En effet, il est possible de montrer que pour deux individus qui diffèrent sur une seule variable $j$, l'odds ratio s'écrit

$$\text{OR}(x,\tilde{x}) = e^{\theta_j(x_j-\tilde{x}_j)}.$$

Dans la mesure où les variables sont rendues qualitatives, l'odds ratio se calcule par l'exponentiel de la différence des coefficients des modalités, toutes choses égales par ailleurs. Ainsi, la prédiction est impactée par la différence des coefficients pour chaque combinaison de modalités d'une même variable. Il est alors possible d'appliquer le critère d'interprétation utilisé en Section 2.3.3.

**Critère de performance**

Plusieurs critères permettent d'évaluer les modèles dont la prédiction est binaire. Pour évaluer la performance de la régression logistique, des critères dérivés de la matrice de confuction seront utilisés : le rappel et la spécificité. Il existe d'autres mesures telles que la précision ou la F-mesure qui est une moyenne harmonique du Rappel et de la Précision.

La matrice de confusion est un tableau à double entrées qui permet de répertorier la classe prédite des indivus selon leur classe réelle. C'est à partir de cette table que sont calculés les autres critères.

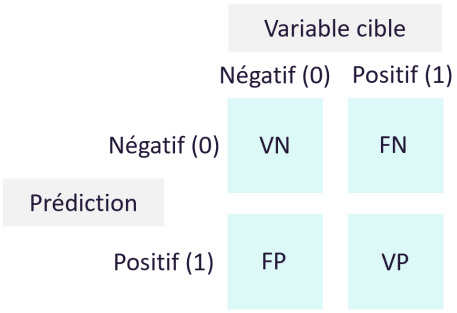



FIGURE 3.8 : Evaluation des modèles binaires

La Figure 3.8 illustre les 4 cas de figures possibles :

- Les individus réellement négatifs sont prédits négatifs (bonne prédiction), ils sont appelés les "vrais négatifs" et notés VN.
- Les individus réellement négatifs sont prédits positifs (mauvaise prédiction), ils sont appelés les "faux positifs" et notés FP.
- Les individus réellement positifs sont prédits négatifs (mauvaise prédiction), ils sont appelés les "faux négatifs" et notés FN.
- Les individus réellement positifs sont prédits positifs (bonne prédiction), ils sont appelés les "vrais positifs" et notés VP.

A partir de là, il est possible de définir les critères de performances suivants :

- **Le rappel (ou sensibilité)** : c'est le taux de vrais positifs, c'est-à-dire la proportion de positifs correctements prédits
$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}}.$$
- **La spécificité** : c'est le taux de vrais négatifs, c'est-à-dire la capacité du modèle à détecter les individus négatifs (inversement au rappel)
$$\text{Spécificité} = \frac{\text{VN}}{\text{VN} + \text{FP}}.$$
Aussi, l'antispécificité est la valeur qui correspond à 1 moins la Spécificité. C'est le taux de négatifs réels classés positifs.

La régression logistique permet de renvoyer une valeur de probabilité. Notament, lorsque le paramètre naturel est nul, la valeur renvoyée est de 0,5. C'est le seuil généralement fixé pour calculer les valeurs ajustées.

Dans le cas de la régression logistique, en notant $\hat{\mu}_i$ la prédiction de l'individu $i$, la valeur ajustée $\hat{Y}_i$ de cet individu est

$$\hat{Y}_i = \mathbb{1}_{\{\hat{\mu}_i > 0.5\}},$$

c'est-à-dire que la valeur renvoyée est 1 si $\omega(x) > 0$. Cependant, ce seuil de 0,5 pourrait arbitrairement être modifié afin d'améliorer la performance du modèle.

La régression logistique est utilisée pour classer les individus de façon binaire 0 ou 1, et la classe 1 est généralement la classe d'intérêt et souvent sous représentée. C'est le cas ici puisqu'il s'agit des individus mobiles. Il est généralement important de bien prédire les individus appartenant à cette classe. L'objectif est donc que le modèle prédise correctement par 1 les individus qui le sont effectivement, mais aussi qu'il ne prédise pas par 1 les individus qui ne le sont pas. il y a donc une dualité entre sensibilité et antispécificité. Ainsi, il est possible de représenter pour chaque seuil la valeur de la sensibilité et de l'antispécificité renvoyée par le modèle : c'est la courbe ROC.

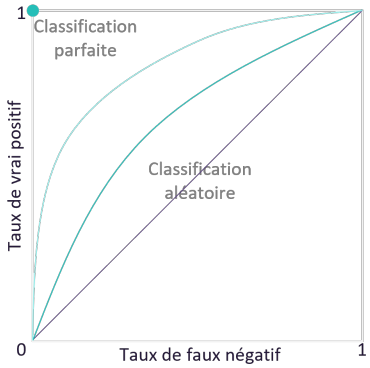



FIGURE 3.9 : Représentation de différentes courbes ROC

La Figure 3.9 est une représentation de différentes coubres ROC selon la performance du modèle. L'objectif est d'atteindre le point de coordonnées (0,1). Une courbe ROC telle que celle de la droite linéaire signifie que le modèle ne fait pas mieux qu'un tirage aléatoire. Plus la courbe se rapproche du cadrant haut gauche, plus le modèle est performant.

Enfin, afin de quantifier cette performance du modèle, il est possible de calculer l'aire sous la courbe ROC, aussi appelée AUC (Area Under the Curve). Un AUC égale à 0,5 signifie que le modèle ne fait pas mieux qu'un tirage aléatoire. Plus l'AUC est proche de 1, plus le modèle est performant.

### 3.2.2 Application au portefeuille et construction du zonier mobilité

#### Construction d'un zonier mobilité

Une régression logistique a été mise en place. Les résidus additifs ont été extraits. Afin d'améliorer les performances du modèle, un zonier est construit sur la base de ces résidus. Cependant, il est possible de se demander quel sens donner à ce zonier ?

L'étude révèle un lien important entre la cartographie des résidus et le taux de mobilité d'une commune. le taux de mobilité d'une commune $z$, noté $\tau_z$, est défini comme le rapport de la somme des expositions des polices mobiles sur la somme des expositions des assurés de la commune

$$\tau_z = \frac{\sum_{\{i \in z\} \cap \{\text{mobile}\}} e_i}{\sum_{\{i \in z\}} e_i}.$$

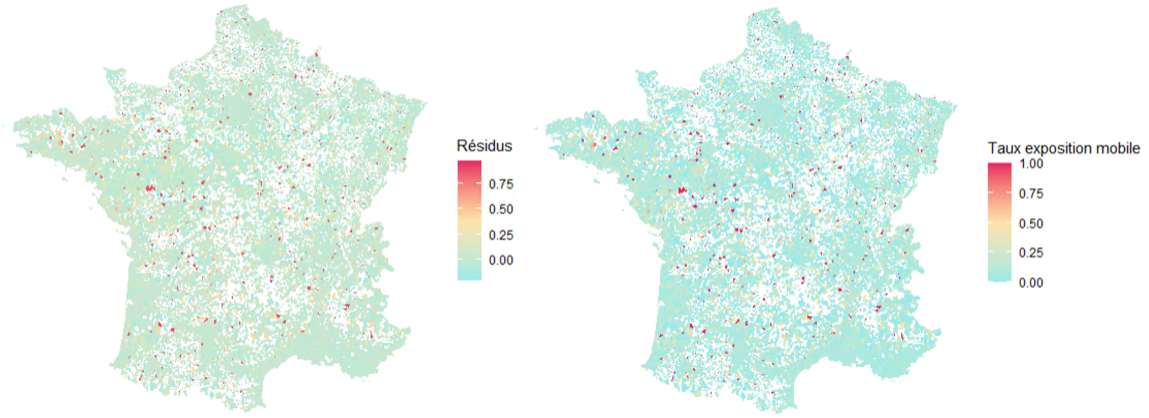



FIGURE 3.10 : Interprétation du zonier mobile

En affichant côte à côte cette carte avec celle des résidus (Figure 3.10), ces deux cartes sont similaires en terme de niveau de valeurs. Les deux cartes ne sont pas à la même échelle, mais les communes avec des valeurs élevées dans un cas le sont aussi dans l'autre. Afin de quantifier cette similitude entre les deux cartes, il est intéressant de comparer les rangs. Pour chaque carte, les communes sont ordonnées selon leur valeur et sont ensuite affectées d'un rang (de celle qui a la valeur la plus élevée à la plus faible). Il est ensuite possible de mettre en place un test apparié permettant de comparer la similitude entre les rangs d'une même commune dans les deux cartes. De cette façon est testée la similitude entre les deux cartes sans considérer l'échelle de valeurs.

Le test mis en place est un test apparié de Wilcoxon. L'hypothèse nulle testée est $H_0$ : "*Les communes conservent leur rang dans les deux cas de figures*". La p_valeur renvoyée est de 0,946 $>$ 0,05, ce qui signifie que l'hypothèse nulle $H_0$ n'est pas rejetée au seuil de $5\%$. Ainsi, les deux cartes sont similaires.le zonier construit peut donc être interpété comme étant le taux de mobilité d'une commune, c'est-à-dire la part de mobiles présents dans la zone.

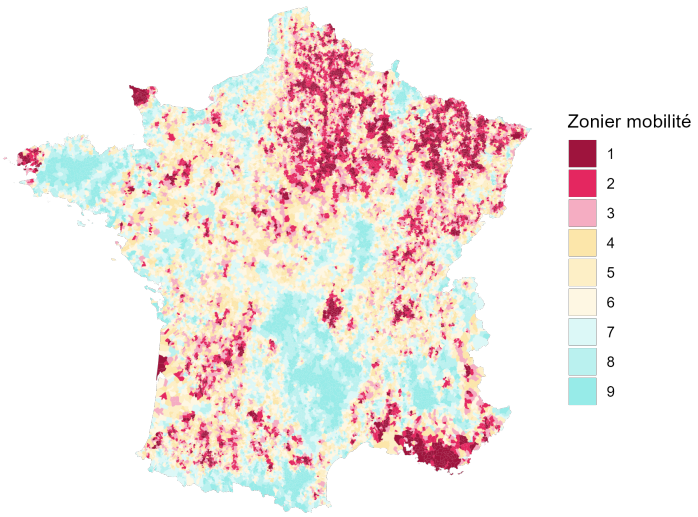



FIGURE 3.11 : Résultat du zonier mobilité

Après prédiction, lissage et classification, c'est le zonier de la Figure 3.11 qui est obtenu. Un zonier mobilité de 1 signie que la zone présente un fort taux de mobilité. A l'inverse, un zonier mobilité de 10 signie que

la zone ne présente pas (ou peu) de polices mobiles. Des points migratoires important sont repérés comme par exemple Brest, Toulon, Lyon, Bordeaux, ou encore Cherbourg. Ce sont des zones où la mobilité est importante. Puisque la mobilité est un facteur aggravant la sinistralité, La construction de ce zonier permet donc de cibler des zones à risque potentielle.

#### Sélection des variables

Une fois le zonier intégré, les variables importantes sont sélectionnées grâce à une méthode de sélection backward basée sur l'AIC. Les variables supprimées sont les trois variables qui ressassent le nombre de sinistres responsables, non responsables, et partiellement responsables sur les 24 derniers mois, et le nombre d'années d'ancienneté avec un CRM à 50%.

#### Interprétation

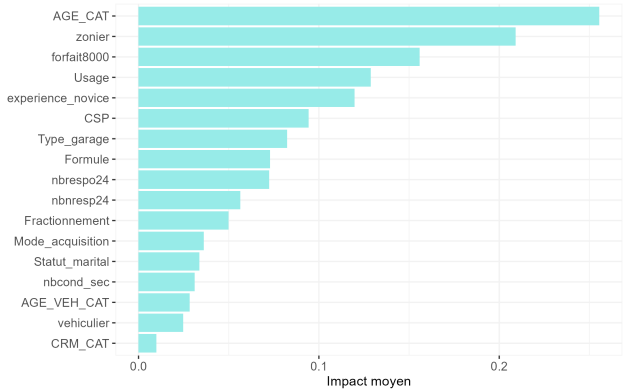



FIGURE 3.12 : Impact des coefficients sur la probabilité d'être mobile

La Figure 3.12 permet de visualiser l'impact des variables sur la probabilité qu'une police donnée soit mobile. La variable la plus discriminante est l'âge du conducteur. Pour rappel, l'analyse descriptive a montré que la population mobile est surtout représentée par des jeunes conducteurs. Il est donc cohérent d'estimer qu'un assuré jeune ait plus de chance d'être mobile qu'un assuré plus âgé.

Une autre variable importante, classée en troisième position juste après le zonier, est le forfait8000. Cette variable prend deux modalités selon si l'assuré a souscrit au forfait ou pas. C'est une offre propre à l'AGPM. C'est une option qui permet à l'assuré de réduire son tarif si la distance annuelle parcourue par son véhicule est inférieure à 8000 km. Il est donc cohérent que cette variable soit significative pour la prédiction, car une distance de parcours élevée est naturellement liée à une forte mobilité de l'assuré.

#### Performance

La courbe ROC du modèle est affichée sur la Figure 3.13. Par ailleurs, un modèle *gbm* a été construit afin de challenger la régression logistique.

Un détail important est que la courbe ROC est définit pour un seuil compris dans l'interval $]0, 0.42[$. Le modèle de régression logistique ne prédit aucun profil avec une probabilité de mobilité supérieure à $0.42$. Cette difficulté de l'algortihme à clairement distinguer les mobiles des amobiles peut être due au fait que les mobiles soient sous-représentés dans le jeu de données.

Les deux courbes ROC sont assez proches et aucun modèle ne se distingue de l'autre. L'AUC est de $65\%$.

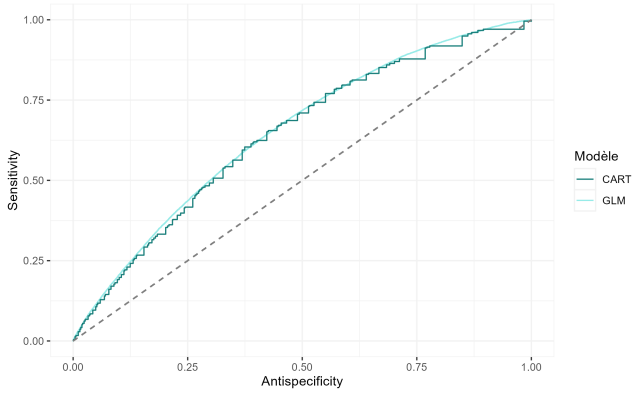



FIGURE 3.13 : Courbes ROC

La performance du modèle est donc moyenne et est encore perfectible. Bien que certains modèles ne seront pas mis en production, la modélisation permet à la compagnie de comprendre la structure de son portefeuille, et de cibler les zones où la mobilité est importante.

### 3.2.3 Prédiction de la fréquence de sinistres avec intégration de la mobilité

#### Interprétation

La variable ”mutation” a directement été intégrée au modèle de référence. La Figure 3.14 montre que c’est la cinquième variable la plus importante, après le forfait8000. Cette variable est donc relativement importante dans la prédiction.

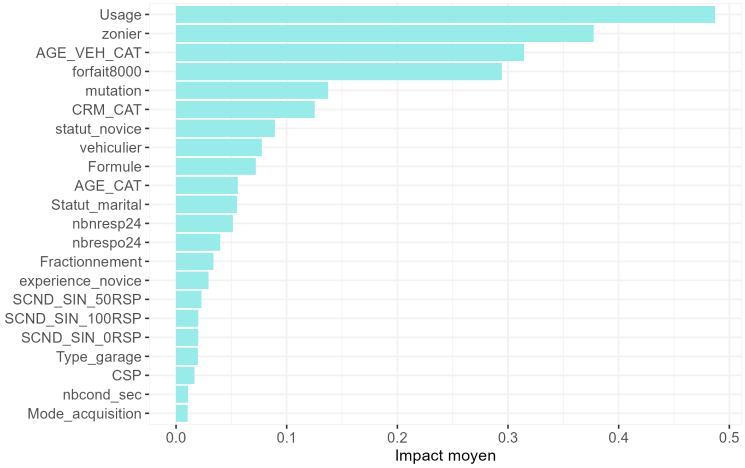



FIGURE 3.14 : Impact des coefficients (modèle avec mutation)

**Analyse des performances**

La dernière étape est de comparer les performances du modèle en l'appliquant sur la base test. Cette performance sera comparée à celle du modèle de base[4].

| Modèle | MAE | MSE |
|---|---|---|
| GLM (1) | 0,04523234 | 0,02275856 |
| GLM + mutation (2) | 0,04522914 | 0,02276068 |

TABLE 3.1 : Résultat sur la base test

La Table 3.1 montre que l'intégration de la mobilité améliore la MAE mais pas la MSE. Cependant, ce changement est très peu significatif. Pour un coût de sinistre moyen de l'ordre de 1000 euros, la différence est au centime près.

| Modèle | MAE (amobile) | MAE (mobile) | MSE (amobile) | MSE (mobile) |
|---|---|---|---|---|
| GLM (1) | 0,04617525 | 0,02761145 | 0,02311878 | 0,01602687 |
| GLM + mutation (2) | 0,04609515 | 0,02904554 | 0,02312535 | 0,01594575 |

TABLE 3.2 : Distinction des résultats selon mobile et amobile

Il serait aussi intéressant d'analyser la performance du modèle en distinguant les résultats sur la population mobile et la population amobile. Ces résultats sont affichés sur la Table 3.2. Ainsi, l'intégration de la mobilité permet d'améliorer la MAE des amobiles et la MSE des mobiles. La MSE est un critère qui donne plus de poids que la MAE aux erreurs importantes. Or, les modèles prédisent dans les deux cas une espérance très faible[5]. Les grandes erreurs sont donc dues aux individus ayant eu un sinsitre. L'amélioration de la MSE chez les mobiles signifie donc que le modèle avec la variable mutation capte mieux la sinsitralité des mobiles que le modèle de base, peut-être au détriment des amobiles. Cependant, le modèle avec la variable mutation reste plus performant.

**Conclusion**

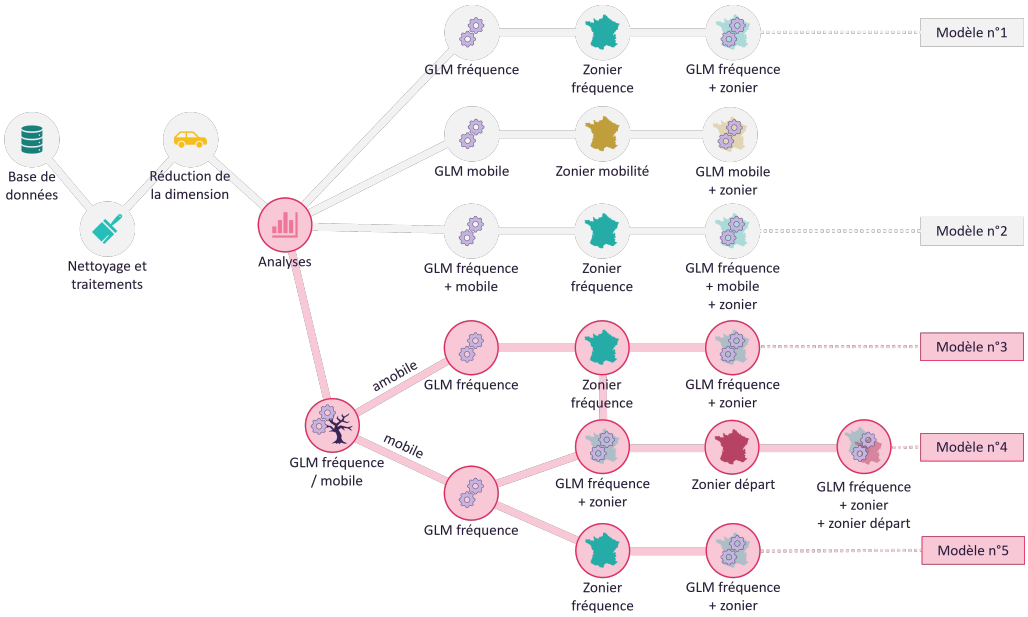
L'objectif de ce chapitre est de modéliser le phénomène de mobilité à travers un zonier, puis d'intégrer ce phénomène dans le modèle de référence. Cette dernière étape montre que l'intégration de ce phénomène n'améliore pas considérablement la performance du modèle. Cette faible progression peut être dûe à une sous représentation de cette population dans le portefeuille. Néanmoins la modélisation a permis de comprendre le phénomène en distinguant les facteurs favorables à la mobilités et les communes les plus impactées par ce phénomène.

[4]Les modèles sont numérotés afin de pouvoir les retrouver sur le schéma de suivi présenté au début du chapitre.
[5]Puisque pour rappel les donnée sont zéro-inflatées.

# Chapitre 4

# Modélisation segmentée : une interprétation enrichie de la mobilité

*Il a été montré que le modèle de fréquence initial sous-estime l'impact des mobiles. Sa prise en compte a permis d'obtenir une modélisation dont le pouvoir prédictif reste quasi inchangé. A défaut de ne pas obtenir des performances qui soient significativements plus intéressantes, l'intérêt de l'étude réside avant tout dans la modélisation et la compréhension du phénomène. Jusqu'à présent, le modèle mis en place applique les mêmes coefficients sur tous les individus. Cependant, puisque les mobiles possèdent une structure démographique atypique, il pourrait être judicieux de modéliser séparément cette population du reste du portefeuille. Ce choix est notamment appuyé par l'application du GLMtree.*



## 4.1 Intérêt du GLMtree

### 4.1.1 Fondement de la démarche

**Le GLM, un modèle dont les coefficents sont indirectements biaisés ?**

Lors de la contruction du GLM, l'ensemble des paramètres sont estimés "en même temps", de sorte à minimiser la déviance du modèle. Pour une modalité donnée, c'est le même paramètre qui est appliqué à l'ensemble du portfeuille. Ainsi, la non prise en compte du facteur mobilité peut biaiser l'estimation des paramètres et déteriorer la prédiction pour les deux populations mobile et amobile.

Afin d'extraire l'effet de ce phénomène sur les paramètres, la variable mobile a été directement intégrée dans le GLM (Chapitre 3). Cependant, cette variable particulière semble concerner une certaine catégorie de risque. L'analyse descriptive montre que ce phénomène touche principalement les jeunes conducteurs militaires.

Ainsi, dans la mesure où les individus mobiles sont principalement des jeunes conducteurs, il est possible que l'âge n'a pas nécessairement un effet significatif dans l'estimation de leur risque. Cependant, le coefficient de l'âge est estimé pour tous les individus, mobiles et amobiles. Alors que l'effet de l'âge ne devrait peut-être pas être pris en compte pour l'estimation du risque des mobiles, la mutualisation de ce coefficient pour les deux types de population oblige sa prise en compte. Ainsi, cette variable pourrait être indirectement biaisée.

C'est généralement le cas des variables qui interragissent entre elles. L'effet de l'interraction peut être pris en compte dans le GLM en intégrant le coefficient de l'âge, le coefficient de mobilité, et un coefficient correspondant au croisement de l'âge et de la mobilité. Ainsi, les coefficients initiaux sont corrigés des effets croisés. Cependant, la mobilité pourrait indirectement impacter d'autres variables. Prendre en compte l'interraction de la mobilité pour toutes les variables serait trop complexe et illisible en terme d'interpétation. Une façon de procéder est de simplement séparer les deux populations et de construire deux modèles distincts.

**Justification mathématique de la segmentation mobile/amobile**

L'intérêt est d'obtenir des modèles plus adaptés et interprétables selon les cas. Ici, la discrimination faite porte sur le caractère mobile de l'assuré. La faisabilité de cette méthode est appuyée par la formule de l'espérance totale

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, M]\mathbb{P}(M|X) + \mathbb{E}[Y|X, \bar{M}]\mathbb{P}(\bar{M}|X),$$

où $M$ est une variable aléatoire indiquant si l'assuré est mobile($M$) ou pas ($\bar{M}$), $Y$ est le nombre de sinistres, et $X$ l'ensemble des caractéristiques propres à l'assuré. Cependant, la variable $M$ est une variable binaire modélisée par une loi de Bernouilli (Chapitre 3). L'expression de l'espérance devient

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, M]\mathbb{E}[M|X] + \mathbb{E}[Y|X, \bar{M}]\mathbb{E}[M|X].$$

Ainsi intervient le modèle de mobilité contruit au chapitre précédent. Néanmoins, il reste à se demander s'il est réellement pertinent de segmenter le jeu de données et de construire deux modèles distincts. En autre, l'objectif est de quantifier le biais potentiel de la mobilité sur tous les autres paramètres.

### 4.1.2 Construction d'un GLMtree

**Principe général**

Le GLMtree (ZEILEIS et al., 2008) fait partie de la famille des méthodes de partitionnement, c'est-à-dire des méthodes dont la constuction est basée sur une division progressive des données. Cette méthode permet de construire des GLM selon une segmentation pertinente du portefeuille.

A titre d'exemple, l'algorithme CART est une méthode partitionnement. Rappelons que son principe repose sur la division répétée des données en sous-groupes, en choisissant à chaque étape la variable qui permet de séparer les indivius de sorte à améliorer la prédiction de la variable cible. Chaque division est effectuée en maximisant un critère de pureté comme le critère de Gini ou l'entropie pour la classification. Une division correspond à un noeud. Un noeud conduit à deux feuilles. Chaque feuille contient donc une partition du jeu de données. La valeur prédite par une feuille est la moyenne des individus qui la compose[1] et qui ont servi à sa construction[2]. Une feuille devient à son tour un noeud et le processus est répété jusqu'à ce qu'un certain nombre de conditions soit atteint, comme la profondeur maximale de l'arbre ou un minimum d'individus dans une feuille. Les dernières feuilles sont appelées les feuilles terminales.

Les méthodes de partitionnement sont généralisées à travers l'algorithme MOB*(Model-Based recursive partitioning)*. Le concept reprend celui de l'algorithme CART, mais au lieu de prédire à chaque feuille par la moyenne (ou la classe majoritaire), c'est un modèle qui est mis en place. Dans le cas où c'est un GLM, il s'agit d'un GLMtree.

Il y a donc deux types de variables : celles qui servent à partitionner le jeu de données (et donc de construire l'arborescence), et celles qui servent à construire le modèle d'une feuille. Ici, la variable qui servira de partitionnement est la variable mobile, et les variables qui serviront à construire les GLM sont les autres variables tarifaires. Dans le cas des arbres de décision, le partitionnement est réalisé selon des critères de pureté (où d'homogénéïté). L'intérêt des GLMtree est de généraliser ce critère par un test d'instabilité des coefficients. Entre autre, ce test sera interprété comme une quantification du biais indirect de la mobilité sur les paramètres.

**Principe théorique**

Cette section reprend les travaux de ZEILEIS et al. (2008). L'objectif de n'est pas d'exposer une théorie exhaustive des GLMtree mais de présenter uniquement les éléments pertinents pour la modélisation.

Soient $Y$ la variable cible, $X^1, ..., X^p$ les $p$ variables explicatives, $Z^1, ..., Z^k$ les $k$ variables qui servent à partitionner le jeu de données, $\theta$ un vecteur contenant l'ensemble des paramètres du modèle, et $\Psi$ la fonction objectif[3]. Le principe de l'algorithme est le suivant :

**Algorithme 5** Contruction du GLMtree (version simplifiée)

**Entrée** Variables cibles, variables explicatives, variables de partitionnement
**tant que** *Critère d'arrêt non vérifié* **:**
*Ajustement du modèle (GLM) sur toutes les variables*
*Test d'instabilité des paramètres selon l'ensemble des variables de partitionnement*
*Partitionner le jeu de données selon la variable associée à la plus grande instabilité*
*(Ce partitionnement est effectué si l'instabilité mesurée est significative)*
**fin tant que**

A chaque noeud, l'algorithme mesure l'instabilité de l'estimation des paramètres pour chaque variable de partitionnement. C'est cette instabilité qui sera interprétée comme le biais de la mobilité sur les variables, puisque la seule variable de partitionnement utilisée est la variable mobile. Dans le cas où cette instabilité est captée, la sortie de l'algorithme sera un modèle de Poisson pour chaque catégorie d'assuré (mobile et amobile). En somme, l'application de cette méthode conduit à deux résultats possibles : ou bien la mobilité n'a pas

[1] Dand le cas de la régression. Sinon, dans le cas de la classification, la classe prédite est la classe majoritaire des individus qui composent à la feuille et qui ont servi à sa construction.

[2] C'est-à-dire les individus de la base d'entraînement.

[3] Les paramètres du modèle sont estimés de sorte à optimiser cette fonction

d'impact sur les paramètres et le modèle renvoie un seul GLM, ou bien l'impact est significatif et le modèle renvoie deux GLM.

L'estimation des paramètres passe par l'optimisation d'une fonction coût, notée ici $\Psi$[4]. Optimiser cette fonction signifie trouver la valeur de $\theta$ tel que

$$\sum_{i=1}^{n} \psi(Y_i, \theta) = 0,$$

où

$$\psi(Y, \theta) = \frac{\partial \Psi}{\partial \theta}(Y, \theta),$$

et $n$ est le nombre d'individus dans la base. Il est important de noter que l'écriture $\frac{\partial \Psi}{\partial \theta}(Y, \theta)$ est abusive[5] et ne renvoie pas un scalaire mais un vecteur où chaque composante est la dérivée par rapport à un paramètre $\theta_j$ du vecteur $\theta$. Ainsi, afin de mesurer l'instabilité des paramètres selon la variable de partitionnement, l'idée est de mesurer la fluctuation de $\psi(Y, \theta)$ autour de 0 selon cette variable. Cette fluctuation est quantifiée par le processus de fluctuation empirique

$$W_j(t) = \hat{J}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \hat{\psi}_{\sigma(Z_{ij})},$$

pour $t \in [0, 1]$, $\sigma(Z_{ij})$ est le rang de la $i$-ième observation de la variable $Z_j$ dans la base, $\hat{\psi}_{\sigma(Z_{ij})}$ est l'estimation du coût associé à l'individu $\sigma(Z_{ij})$, et $\hat{J}$ une estimation de la matrice de covariance des coûts

$$\hat{J} = \frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, \hat{\theta}) \psi(Y_i, \hat{\theta})^T.$$

Sous l'hypothèse nulle de stabilité des paramètres, ce processus de fluctuation empirique converge vers un mouvement brownien. Il devient alors possible d'appliquer un test statistique en capturant la fluctuation du processus par application d'une certaine fonction $\lambda$. Le lecteur intéressé trouvera des écrits plus approfondis sur la nature du test dans ZEILEIS et HORNIK (2007).

## 4.2 Double modélisation : mobile vs amobile

### 4.2.1 Résultats du GLMtree

La p-valeur liée au test de fluctuation est inférieur à 0.001 (et donc inférieur à 0.05), ce qui signifie que l'hypothèse de stabilité des paramètres est rejetée. Le modèle renvoie donc deux GLM, un pour chaque catégorie d'assuré (mobile et amobile).

La prochaine étape est d'extraire les résidus des deux modèles afin de construire deux zoniers et de comprendre ainsi la perception géographique du risque selon les deux populations. Le modèle des mobiles pourra par la suite être enrichi par d'autres hypothèses tels que la localisation de provenance (Métropole, Martinique, Mayotte, ...).

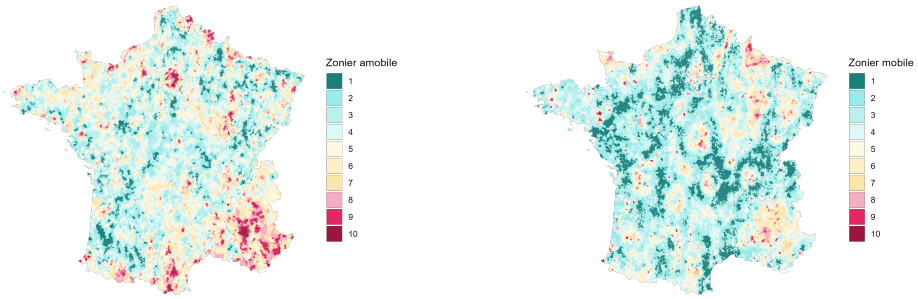
[4]Cette notation est choisie afin de rester cohérent avec l'article de référence.

[5]Mathématiquement, c'est le gradient de la fonction $\Psi$.

### 4.2.2 Confrontation des deux modèles

#### Zonier mobile et amobile

L'extraction des résidus de chacun des deux modèles permet de construire deux zoniers, chacun adapté à la population modélisée. Ces deux zoniers sont représentés sur la Figure 4.1.




(a) Zonier amobile (b) Zonier mobile

FIGURE 4.1 : Comparaison des zoniers amobile et mobile

Il est impotrant de noter que les modalités des deux zoniers ne correspondent pas, c'est-à-dire qu'une commune évaluée zonier 1 pour amobile ne représente pas le même risque qu'une commune évaluée zonier 1 pour mobile. La comparaison ne peut être faite que par contraste entre les communes.

Les deux zoniers construits sont loin d'être similaires. Or, le risque géographique est un risque fixe. Il devrait, en théorie, être le même quelque soit la population. Cette différence montre bien l'impact de la mobilité sur la perception du risque géographique des mobiles. Cependant, si le zonier des amobiles représente le risque géographique réel, il est important d'expliquer la différence portée par le zonier des mobiles.

#### Sélection des variables

Une méthode de sélection AIC backward est appliquée sur le GLM des mobiles. Comparé au modèle des amobiles, beaucoup de variables ont été supprimées. Les variables conservées sont affichées sur la Figure 4.2. Très peu de variables sont représentées pour les mobiles du fait de leur profil atypique.

Il est intéressant de noter que les variables représentant l'âge de l'assuré et sa catégorie socioprofessionelle ont été supprimées. En effet, la population est très spécifique : ce sont des jeunes militaires. Ces deux critères ne sont donc pas suffisament discriminants du point de vue de la sinistralité.

#### Comparaison des coefficents

Après sélection des variables, il est possible de comparer la valeur des coefficients des deux modèles (Figure 4.2). Les fluctuations importantes sont surtout présentes pour les variables Usage, Statut_marital et pour l'intercept. La valeur de l'intercept pour les mobiles est supérieure à celle des amobiles, ce qui appuie l'idée que les mobiles représentent un risque intrincèque plus important que les amobiles. Cependant, en prenant par exemple la variable Staut_marital, la plupart des modalités pour les mobiles sont inférieurs à celles des amobiles, ce qui montre que cette variable aggrave plus la sinistralité des amobiles.

Cependant, il est important de distinguer deux interprétations pour une variable donnée : la variation de la valeur des coefficients de chaque modalité pour un modèle (mobile ou amobile), et la différence des valeurs entre les deux modèles. Le premier permet de mesurer l'impact du changement de modalité dans la tarification, tandis que la seconde permet de comparer la contribution globale de la variable dans la mesure du risque entre les deux modèles.

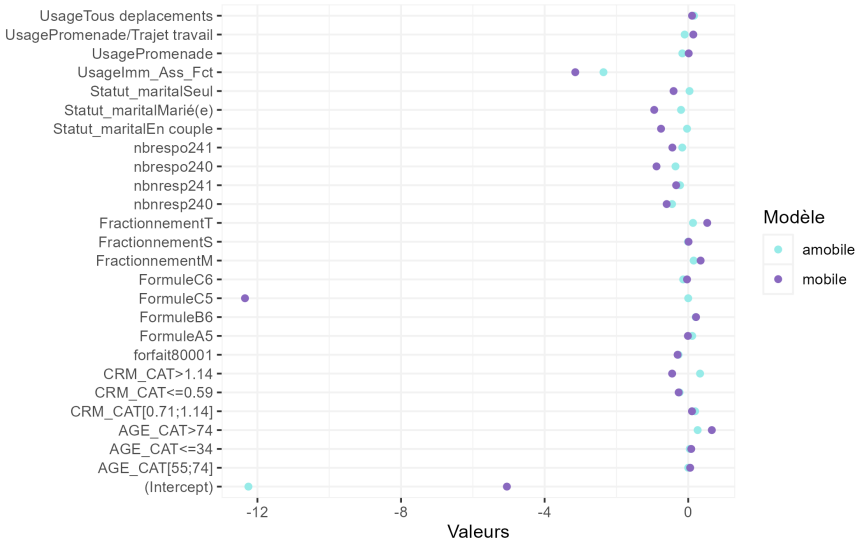



FIGURE 4.2 : Comparaison des coefficients pour le modèle mobile et amobile

Ainsi, pour la variable Statut_marital, les modalités ont des coefficients plus importants dans le modèle amobile, ce qui signifie que cette variable contribue plus à augmenter le risque globale des amobiles que les mobiles, mais la volatilité des coefficients pour les mobiles semble plus importante, ce qui signifie que changer de modalité pour cette variable impacte plus la tarification des mobiles que des amobiles.

### 4.2.3 Analyse des performances

| Modèles | MAE | MSE | MAE (amobile) | MAE (mobile) | MSE(amobile) | MSE(mobile) |
|---|---|---|---|---|---|---|
| 1 | 0.04523234 | 0.02275856 | 0.04617525 | 0.02761145 | 0.02311878 | 0.01602687 |
| 2 | 0.04522914 | 0.02276068 | 0.04609515 | 0.02904554 | 0.02312535 | 0.01594575 |
| 3 et 5 | 0.04521208 | 0.02461193 | 0.04602132 | 0.03008929 | 0.02495445 | 0.01821112 |

TABLE 4.1 : Analyse des performances sur la base de test

La Table 4.1 résume les performances des modèles sur la base de test[6]. Le modèle segmenté est celui qui donne les meilleurs résultats en terme de MAE et de MSE. Seulement, ce résultat est à nuancé. C'est la modélisation des amobiles qui est réellement améliorée (MAE et MSE), au détriment des mobiles qui est déteriorée. Seulement, puisque les amobiles sont majoritaires, on peut expliquer ce gain de performance par le fait que l'estimation de leur risque soit plus robuste. Néanmoins, bien que les mobiles soient minoritaires, cette progression de la performance globale confirme bien le biais indirect des mobiles sur l'estimation des paramètres.

Finalement, quelque soit le modèle utilisé, les différences de performances ne sont pas importants. Bien que l'impact ait été détecté de manière significative, les résultats obtenus peuvent s'expliquer par la faible volumétrie des mobiles. Cette faible volumétrie représente une certaine limite de la modélisation.

[6]Pour rappel les modèles sont numérotés afin de pouvoir les retrouver sur le schéma de suivi présenté au début du chapitre.

### 4.2.4 Construction d'un zonier de provenance géographique

Enfin, une dernière modélisation pouvant être appliquée est la construction d'un zonier de provenance géographique (Figure 4.3). En considérant que le zonier des amobiles représente le risque géographique réel, il est possible de l'intégrer au modèle des mobiles. Les résidus peuvent ensuite être extraits, et sur la base de ces derniers, il est possible de construire un zonier de provenance géographique.

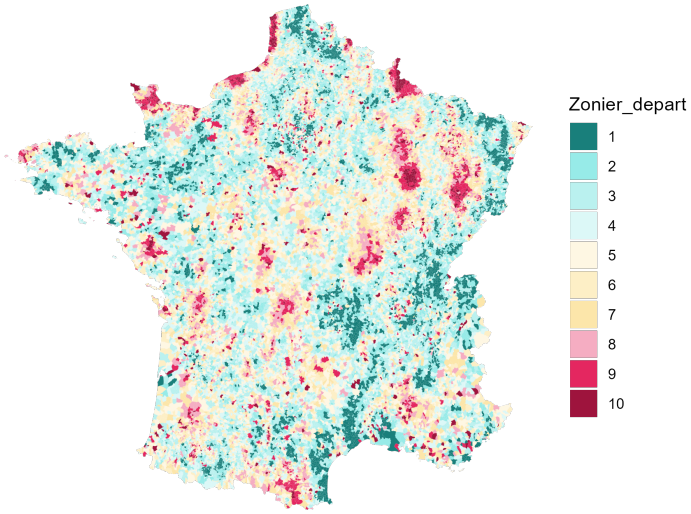



FIGURE 4.3 : Résultat du zonier départ

Initialement, pour construire un zonier, les résidus sont agrégés à la maille de la commune actuelle. Ici, les résidus extraits seront agrégés à la maille de la commune passée. Le zonier construit représente donc le risque de provenance géographique. Une modalité élevée pour une commune signifie que provenir de cette commune est plus risquée qu'une autre. Ce zonier de provenance sera appelé le zonier départ.

### 4.2.5 Analyse des performances

| Modèles | MAE | MSE | MAE (amobile) | MAE (mobile) | MSE(amobile) | MSE(mobile) |
|---|---|---|---|---|---|---|
| 1 | 0.04523234 | 0.02275856 | 0.04617525 | 0.02761145 | 0.02311878 | 0.01602687 |
| 2 | 0.04522914 | 0.02276068 | 0.04609515 | 0.02904554 | 0.02312535 | 0.01594575 |
| 3 et 5 | 0.04521208 | 0.02461193 | 0.04602132 | 0.03008929 | 0.02495445 | 0.01821112 |
| 3 et 4 | 0.04521051 | 0.02460953 | 0.04602132 | 0.03005825 | 0.02495445 | 0.01816386 |

TABLE 4.2 : Titre du tableau

Enfin, le modèle avec zonier de départ a été intégré au GLM des mobiles, et les performances ont une nouvelle fois été comparées. Cette dernière modélisation permet d'obtenir des résultats qui sont, une nouvelle fois , sensiblement meilleures aux autres modèles mis en place. La performance du GLM amobile n'a pas changé puisque le zonier de départ ne concerne que les mobiles. Cependant, pour le GLM des mobiles, en remplaçant le zonier construit avec le zonier des amobiles (risque géographique réel) et le zonier de départ, de meilleures performances sont obtenues (pour la MAE et pour la MSE). Le zonier des mobiles ne représente pas le risque géographique réel, et la comlplexité de la mobilité induit un zonier biaisé pour les mobiles.

## 4.3 Limites et prise de recul sur la modélisation

Les résultats de l'analyse descriptive ont montré une fréquence de sinistres significativement plus importante chez les individus mobiles que chez les amobiles. Par ailleurs, l'enjeu était de comprendre si ce phénomène n'était pas simplement dû à une démographique atypique qui favoriserait ce phénomène. En effet, il s'avère que les assurés mobiles sont en général plus jeunes que les autres. En ayant extrait les effets de toutes les variables tarifaires grâce au *Modèle Linéaire Généralisé*, le test du rapport Attendu sur Estimé montre qu'il existe bien un effet de la mobilité sur la fréquence de sinistres, bien que les autres facteurs tels que l'âge soient déjà pris en compte. C'est ce test qui fonde l'enjeu de ce mémoire. Cependant, la prise de recul sur les données est importante afin de comprendre les faits observés et ne pas tomber dans le **biais de conformité**[7].

Il existe deux formes de biais : celui des données vers l'utilisateurs, et celui de l'algorithme utilisé. Le biais de l'algorithme est un sujet qui a été abordé au Chapitre 1, notament à travers le dilemme biais-variance. Dans cette partie, ce n'est pas ce type de biais qui esttraité. Ce sont les biais présents dans les données qui sont étudiés.

### 4.3.1 Biais de sélection

Pour rappel, le biais de sélection (ou biais de représentativité) apparaît lorsque l'échantillon étudié n'est pas représentatif de la population. En l'occurence, dans le cas de l'AGPM, la population assurée est principalement militaire. Leur mode de vie singulier mérite une étude adaptée. C'est ce qui justifie l'étude de cette population sous l'angle de la mobilité.

#### Un portefeuille singulier

Outre le phénomène de mobilité, le caractère singulier du portefeuille se révèle implictement à travers le zonier. En effet, la répartition de l'exposition sur le territoire engendre des "îlots" dus à une concentration des assurés dans certaines zones, notamment des lieux proches de bases de défenses. Plusieurs zones sont alors désertes et leurs risques géographiques pourraît être difficiles à prédire.

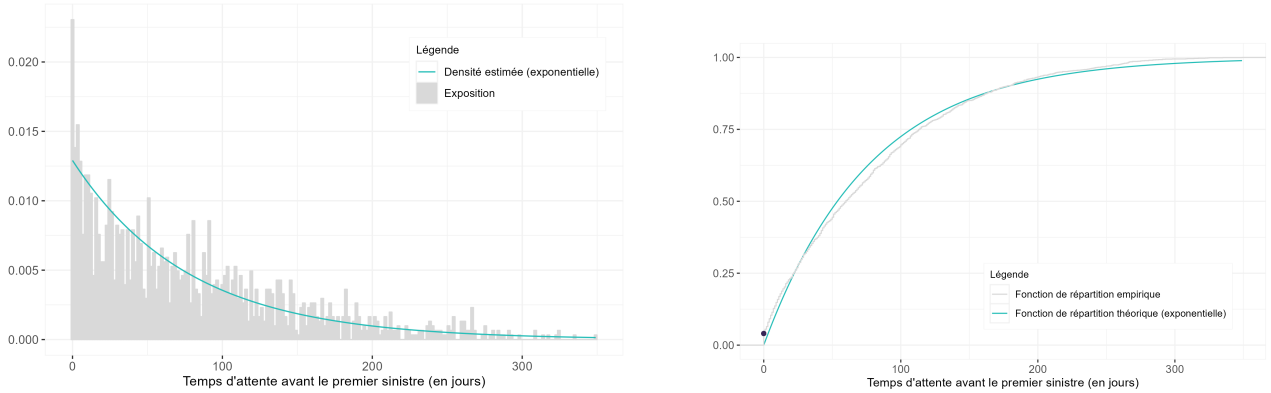
#### Déclaration d'un sinistre chez les mobiles

La variable mobile a été construite sur la base du suivi des assurés. Un assuré a été considéré mobile si au moment de la souscription ou du renouvellement du contrat, le code INSEE diffère de celui de la segmentation précédente. Cependant, il se pourrait qu'en réalité, les individus qui apparaîssent mobiles à ce moment là l'était déjà sans l'avoir déclaré. Ainsi, lorsqu'ils ont un sinistre, ils le déclarent, et c'est au moment de la déclaration du sinistre que leur dossier se met à jour avec le changement de code INSEE.

Prenons un assuré qui déménage en 2017. Supposons qu'il ne déclare pas ce déménagement. Puis, ayant eu un sinistre en 2019, soit 2 ans plus tard, le sinistre est déclaré avec une mise à jour de son code INSEE. Alors qu'il était mobile en 2017, cette déclaration tardive oblige à le considérer mobile en 2019 avec un sinistre qui en réalité au lieu moment où il n'était plus mobile. Cet exemple illustre pourquoi la sinistralité des mobiles pourraient potentiellement être surrestimée à tord.

Un moyen d'estimer l'impact de ce phénomène est d'analyser le temps avant la survenance du premier sinistre dans l'année chez la population mobile. Afin de modéliser un phénomène de temps d'attente, il est possible d'estimer la distribution par une loi exponentielle. Le but n'est pas de réaliser une étude poussée, mais

[7]Biais selon lequel les éléments qui confirment une hypothèse souhaitée sont préférés ou favorisés, en faisant (souvent) abstraction des facteurs pouvant l'infirmer.




(a) Distribution

(b) Fonction de répartition

FIGURE 4.4 : Analyse du temps d'attente avant le premier sinistre chez les mobiles

d'avoir une idée sur le comportement des assurés. L'estimation du paramètre de la loi conduit a un temps d'attente moyen d'environ 77 jours. Cependant, ce qui importe réellement, c'est le pic en 0. Un pic en 0 signifie que le sinistre survient le jour de la souscription. Cependant, ce pic ne représente qu'environ 4% des sinistres. L'effet du biais décrit précédement est potentiellement peu présent dans le portefeuille.

#### Phénomène de Censure à droite

Le phénomène de censure à droite est surtout étudier en démographique pour la construction de tables de mortalité. Il apparaît lorsque le phénomène étudié a lieu avant la date de début d'observation. Puisque les observations débutent en 2016, la proportion de mobiles pour cette année est surement sous-estimée, du fait de l'absence d'information sur les années antérieures.

Un autre phénomène évoque au Chapitre 1 est l'exposition des individus mobiles tronquée par l'année compatble. Pour reprendre l'exemple, un assuré qui aurait muté et dont la date de début d'observation est en Décembre aurait une exposition de mobilité d'un mois, tandis qu'un assuré dans le même cas avec une date de début d'observation en avril aurait une exposition de mobilité de neuf mois. Pourtant, l'assuré ayant muté en Décembre, la survenance d'un sinistre en janvier serait très probablement lié à la mutation du mois passé. Cependant, la structure en année comptable comptabilisera ce sinistre de janvier comme un sinistre sans effet de la mobilité.

### 4.3.2 Bias de suffisance

#### Vision fréquentielle vs vision temporelle

La vision adoptée tout le long de ce mémoire est une approche fréquentiste (ou distributive). L'aspect temporelle n'a jamais été prise en compte. Pourtant, un assuré qui aurait muté l'année précédente aurait en théorie très peu de chances d'être mobile durant l'année d'observation. L'intégration de l'aspect temporelle pourrait contribuer à améliorer le modèle de prédiction des profils mobiles. Cependant, étudier ce phénomène d'un point de vue temporelle ferait l'object d'une étude très poussée. A titre d'exemple, une analyse descriptive simple montre que le nombre de mobilité est nettement plus important en été que durant le reste de l'année (Figure 4.5). Cette analyse peut aider la compagnie dans la mise en place d'une campagne de prévention destinée aux assurés avec une mobilité élevée durant cette période.

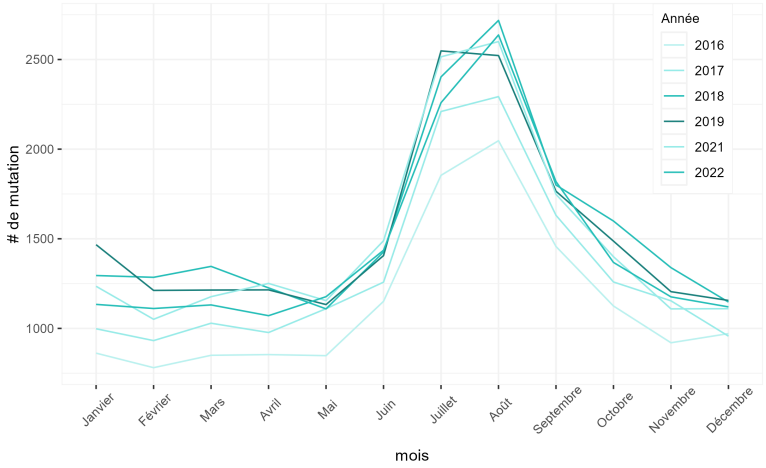



FIGURE 4.5 : Evolution du nombre de mutation

**Manque d'exhaustivité dans les variables tarifaires**

Les performances des modèles pour la modélisation du zonier mobile ont atteint les 63%. Cependant, c'est une performance qui reste moyenne et cela peut s'expliquer par un manque de variables explicatifs. Pour rappel, seuls 4 variables ont été utilisées. De plus, il a été montré que le traitement du phénomène de mobilité a permis d'améliorer la mesure du risque des amobiles, mais pas des mobiles. Ces résultats appuient le manque d'explicabilité des variables utilisées pour la modélisation de la sinistralité des mobiles.

**Un jeu de données peu volumétrique**

Bien que la mobilité soit importante en terme de sinistralité, sa volumétrie dans le portefeuille n'est pas assez conséquente. Elle ne concerne que 20% des polices, seuls quelques milliers de mutations sont comptabilisées, face à une centaine de milliers de profils différents.

**Impact non significatif sur la prime**

La mobilité a été analyser ici uniquement du point de vue de la fréquence. Cependant, il serait pertinent d'analyser aussi le phénomène du point de vue du coût moyen. Si la fréquence des sinistres chez les mobiles est plus importantes, il se pourraît que ce soit en réalité des sinistres à faibles coûts, ce qui compenserait la sursinistralité observée. L'impact du phénomène sur la prime et plus largement sur la solvabilité reste une piste d'ouverture à explorer.

### 4.3.3 Limites opérationnelles et éthique

**Automatisation du rescencement des mobiles**

La variable mutation a été construite sur la base de l'historique du portfeuille. Cette variable pourrait être construite de façon plus pertinente grâce à un processus de rescencement automatisé de la compagnie. Elle serait donc directement fournie à la souscription ou au renouvellement du contrat.

**Ethique et discrimination**

Reste à se demander s'il est éthiquement correcte de discriminer les assurés selon leur région de provenance. Toutes choses égales par ailleurs, deux assurés dans une même commune pourraient se retrouver avec des tarifs différents, sous prétexte qu'ils ne proviennent pas de la même région. En réalité, cette problématique est contournée par l'intégration de la probabilité d'être mobile. Au lieu d'interroger directement l'assuré au mo-

ment de la souscription (ou du renouvellement du contrat), une probabilité d'être mobile sera affectée pour tous les profils similaires. Tout de même, la prise en compte de cette variable comme facteur discriminant reste une question importante. L'actuaire doit pouvoir proposer des modèles de tarification justes, avec une segmentation adaptée lui permettant de faire face à la concurrence, tout en respectant une déontologie professionnelle.

**Conclusion**

L'objectif de ce chapitre est d'approfondir l'étude de la mobilité. L'utilisation du GLMtree a permis de segmenter les deux populations en les modélisant séparément. Cette double modélisation a ensuite permis de quantifier l'impact de la mobilité sur la tarification mais aussi sur le zonier : la mobilité impact de la zonier, et la zone de provenance est un facteur important. Cependant, la modélisation est confrontée à certaines limites qui ont été soulevée, ce qui permet d'amorcer de nouvelles poursuites d'études.
Enfin, l'intérêt de toutes ces modélisation est avant tout d'assister la compagnie dans ses prises de décisions stratégiques, et surtout de l'aider à comprendre la structure de son portefeuille et le comportement de ses assurés.

# Conclusion

L'objet de ce mémoire répond à un besoin client spécifique : comment modéliser et mesurer l'impact de la mobilité dans le portefeuille ?

La première étape était de construire un modèle de référence ne prenant pas en compte le phénomène de mobilité. Puis, ce modèle a été augmenté à l'aide d'une variable mobilité indiquant si l'individu est mobile ou pas, ce qui est permis d'améliorer très légèrement la MAE mais pas la MSE. Puis, une modélisation plus complexe a été mis en place à l'aide d'un GLMtree : modéliser le risque différemment selon les deux populations considérés. Là encore, une meilleur MAE est observée à défaut d'une dégradation de la MSE. Ce schéma de performance s'explique notamment par le fait que les données à disposition soient zéros-inflatées : la population à risque est sous-représentée, et bien que la fréquence de sinistres chez les mobiles soit nettement plus importante, l'échantillon représentatif de cette population n'est pas assez important en terme de volume. Le modèle de tarification peine donc à détecter correctement le risque des mobiles, et s'améliore donc dans la détection du « non risque ». Enfin, afin d'approfondir la compréhension du phénomène, un modèle de régression logistique a été construit afin de détecter les assurés mobiles mais aussi les zones avec un fort taux de mobilité ; et le GLMtree a permis de construire un zonier qui permet de détecter les zones de provenance à risque.

Cependant, la modélisation envisagée peut présenter certaines limites. Tout d'abord, le phénomène de mobilité n'a pas du tout été abordé d'un point de vue temporelle. Il peut-être fort probable qu'un individu ayant récemment muté est plus ou moins de chance de muter une nouvelle fois dans un avenir proche. Une autre limite importante réside dans la construction de la variable mobile : sur la base de l'historique du portefeuille, un individu a été considéré mobile si son code INSEE pour une observation donnée est différente de la précédente. Cependant, une « observation » du point de vue du portefeuille est bornée selon l'année comptable. A titre d'exemple, un individu ayant muté en novembre aura la mention mobile jusqu'au 31 décembre, puis redeviendra non mobile l'année d'après. En revanche, si un sinistre survient pour cet individu en janvier, soit à peine 2 mois après mutation, ce sinistre ne sera pas catégorisé comme mobile (alors qu'il est très probable que ce sinistre soit survenu à cause de la mobilité). La mobilité pourrait donc être un phénomène sous-estimé dans le portefeuille.

Enfin, les études réalisées dans ce mémoire offrent diverses pistes d'études intéressantes : étude de la valeur client du portefeuille, introduction des variables météorologiques dans le zonier, analyse de la mobilité dans une dimension temporelle, ou encore l'utilisation de méthodes d'interprétabilité des modèles de *Machine Learning* (autre que les valeurs de Shapley) afin d'approfondir la compréhension de la mobilité.

# Bibliographie


AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19.6, p. 716-723.

AUTORITÉ DE CONTRÔLE PRUDENTIEL ET DE RÉSOLUTION (ACPR) (2010). URL : https://acpr.banque-france.fr/page-sommaire/textes-de-reference.

BREIMAN, L. (2001). Random forests. *Machine learning* 45, p. 5-32.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. et STONE, C. J. (1984). Classification and Regression Trees. *Wadsworth International Group*.

BROUSTE, A., BOULESTEIX, A. L. et GIRAUD, C. (2020). A closed-form solution for the estimation of parameters in generalized linear models: a new perspective. *Statistical Modelling* 20.3, p. 238-256.

CAMERON, A. C. et TRIVEDI, P. K. (2013). Regression Analysis of Count Data. 2nd. Cambridge University Press.

CERIANI, L. et VERME, P. (2012). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality* 10, p. 421-443.

CODE CIVIL (1804). URL : https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006070721/2023-01-01.

CODE DE LA MUTUALITÉ (1945). URL : https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006074067/2022-03-29/.

CODE DE LA SÉCURITÉ SOCIALE (1956). Code de la sécurité sociale. URL : \url{https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006073189/2023-03-05}.

CODE DES ASSURANCES (1930). URL : https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006073984/2022-03-21.

CRESSIE, N. (1990). The Origins of Kriging. *Mathematical Geology* 22.3, p. 239-252.

De L'INTÉRIEUR, M. (2023). Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2022. URL : https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2022/.

Des ARMÉES, M. (2014). Bases de défense. URL : https://www.data.gouv.fr/fr/datasets/bases-de-defense-564168/.

DÉCRET N°2018-487 (2018). URL : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037076517.

FENG, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distribution and Applications* 8.8.

FREUND, Y. et SCHAPIRE, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55.1, p. 119-139.

FRIEDMAN, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29.5, p. 1189-1232.

FÉDÉRATION FRANÇAISE DE L'ASSURANCE (FFA) (2022). Les données clés de l'assurance française en 2022. URL : https://www.franceassureurs.fr/nos-chiffres-cles/donnees-globales/assurance-francaise-donnees-cles-2022/.

GREEN, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust Applications. *Journal of the Royal Statistical Society: Series B (Methodological)* 46.2, p. 149-164.

GREGOIREDAVID (2018). Contours des régions, départements, arrondissements, cantons et communes de France (métropole et départements d'outre-mer) au format GeoJSON. URL : https://github.com/gregoiredavid/france-geojson.

HASTIE, T. et TIBSHIRANI, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association* 82.398, p. 371-386.

INSEE (2024). Base du dossier complet. URL : https://www.insee.fr/fr/statistiques/5359146#consulter.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. et al. (2013). An introduction to statistical learning. T. 112. Springer.

KOHAVI, R (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Morgan Kaufman Publishing*.

KRUSKAL, J. B. (1978). Multidimensional scaling. *Murry Hill*.

LE BASTARD, L. (2024). Clustering of pathologies: application to Long-Term Care Insurance.

MACQUEEN, J (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*.

MCCULLAGH, P. et NELDER, J. A. (1989). Generalized Linear Models. 2nd. London : Chapman & Hall.

OPENDATASOFT (2016). Correspondance Code INSEE - Code Postal 2013. URL : https://public.opendatasoft.com/explore/dataset/correspondance-code-insee-code-postal/table/.

SCHWARZ, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* 6.2, p. 461-464.

ZEILEIS, A. et HORNIK, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica* 61.4, p. 488-508.

ZEILEIS, A., HOTHORN, T. et HORNIK, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 17.2, p. 492-514.

# Annexe A

# Démonstrations des principes théoriques

## A.1 Problème de la prime pure

L'objectif est de résoudre le problème suivant

$$\pi_x^* \in \arg\min_{\pi_x \in \mathbb{R}} \{\mathbb{E}[(S_x - \pi_x)^2]\},$$

où $S_x$ est une variable aléatoire réelle positive. Il s'agit d'un problème d'optimisation convexe sans contrainte. En effet, en développant le carré et par linéarité de l'espérance, le problème se réécrit

$$\pi_x^* \in \arg\min_{\pi_x \in \mathbb{R}} \{\mathbb{E}[S_x^2] - 2\pi_x \mathbb{E}[S_x] + \pi_x^2\},$$

qui est une fonction polynomiale de degré 2 en $\pi_x$ et dont le coefficient du degré maximal est positif. Notons que bien que $S_x$ soit une variable aléatoire, il n'influe en rien la résolution de ce problème puisqu'on manipule l'espérence de cette variable et l'espérence est un objet déterministe. La fonction à minimiser est donc convexe, la solution de ce problème existe et est unique.

La dérivée de la fonction est

$$\frac{\partial \mathbb{E}[(S_x - \pi_x)^2]}{\partial \pi_x} = \mathbb{E}\Big[\frac{\partial (S_x - \pi_x)^2}{\partial \pi_x}\Big] = \mathbb{E}[-2(S_x - \pi_x)] = -2\mathbb{E}[S_x] + 2\pi_x.$$

La solution $\pi_x^*$ vérifie la condition d'optimalité :

$$\mathbb{E}[-2(S_x - \pi_x^*)] = 0 \Leftrightarrow \pi_x^* = \mathbb{E}[S_x].$$

*-Fin de la démonstration-*

## A.2 Modèle coût-fréquence

Soit $S_x$ une variable aléatoire réelle telle que

$$S_x = \sum_{n=1}^{N_x} C_x^n,$$

où $(C_x^n)_{n\in\mathbb{N}^*} \overset{\text{iid}}{\sim} C_x$ et $C_x \perp\!\!\!\perp N_x$. Alors

$$\begin{aligned}
\mathbb{E}[S_x] &:= \mathbb{E}\Big[\sum_{n=1}^{N_x} C_x^n\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\sum_{n=1}^{N_x} C_x^n | N_x\Big]\Big] \\
&= \mathbb{E}\Big[\sum_{n=1}^{N_x} \mathbb{E}\Big[C_x^n | N_x\Big]\Big] \text{ par mesurabilité,} \\
&= \mathbb{E}\Big[\sum_{n=1}^{N_x} \mathbb{E}[C_x^n]\Big] \text{ par indépendance,} \\
&= \mathbb{E}[N_x \mathbb{E}[C_x]] \text{ car les distributions sont iid,} \\
&= \mathbb{E}[N_x]\mathbb{E}[C_x].
\end{aligned}$$

*-Fin de la démonstration-*

## A.3 Modèle de machine learning optimal

L'objectif est de résoudre le problème

$$f^* \in \arg\min_{f\in\mathcal{F}} \{\mathbb{E}_{(Y,X)\sim\mathcal{P}}[l(Y, f(X))]\},$$

où $\mathcal{F} = \{f : \mathbb{E}[f(X)^2] < +\infty\}$ est l'ensemble des prédicteurs possibles. $\forall f \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{E}[(Y - f(X))^2] &= \mathbb{E}[(Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - f(X))^2] \\
&= \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] - 2\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - f(X))] + \mathbb{E}[(\mathbb{E}[Y|X] - f(X))^2].
\end{aligned}$$

La quantité $\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - f(X))]$ se simplifie de la façon suivante

$$\begin{aligned}
&\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - f(X))] \\
&= \mathbb{E}[Y\mathbb{E}[Y|X] - Yf(X) - \mathbb{E}[Y|X]^2 + \mathbb{E}[Y|X]f(X)] \\
&= \mathbb{E}\Big[\mathbb{E}[Y\mathbb{E}[Y|X] - Yf(X) - \mathbb{E}[Y|X]^2 + \mathbb{E}[Y|X]f(X)|X]\Big].
\end{aligned}$$

Or $\mathbb{E}[Y|X]$ et $f(X)$ sont $X$-mesurables, d'où

$$\begin{aligned}
&\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - f(X))] \\
&= \mathbb{E}\Big[\mathbb{E}[Y|X]^2 - f(X)\mathbb{E}[Y|X] - \mathbb{E}[Y|X]^2 + \mathbb{E}[Y|X]f(X)\Big] \\
&= 0.
\end{aligned}$$

Finalement, la quantité que l'on cherche à minimiser s'écrit

$$\mathbb{E}[(Y - f(X))^2] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - f(X))^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y|X])^2]$$

car $\mathbb{E}[(\mathbb{E}[Y|X] - f(X))^2] \geq 0$.
En conclusion, $\forall f \in \mathcal{F}$, $\mathbb{E}[(Y - f(X))^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y|X])^2]$ d'où la solution

$$f^* : x \mapsto \mathbb{E}[Y|X = x].$$

*-Fin de la démonstration-*

## A.4 Equivalence entre GLM Poisson et modèle de régression

On se place dans le cas d'un GLM Poisson avec la fonction de lien logarithmique. On suppose donc que $Y_i|X = x_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{P}(\lambda(x_i))$ et que $\ln(\lambda(X_i)) = X_i\theta + \ln(e_i)$. Par indépendance des individus, la vraisemblance du modèle s'écrit

$$\mathcal{L}(Y_1, ..., Y_n; \theta) \overset{\perp\!\!\!\perp}{=} \prod_{i=1}^{n} e^{-\lambda(X_i)} \frac{\lambda(X_i)^{Y_i}}{Y_i!}.$$

La log-vraisemblance est la fonction $l$ s'écrit donc

$$l(Y_1, ..., Y_n; \theta) = \sum_{i=1}^{n} -\lambda(X_i) + Y_i \ln(\lambda(X_i)) - \ln(Y_i!).$$

En utilisant la fonction de lien logarithme, on a

$$l(Y_1, ..., Y_n; \theta) = \sum_{i=1}^{n} -e_i e^{X_i\theta} + Y_i X_i \theta - \ln(Y_i!).$$

Le paramètre $\theta$ est estimé en maximisant cette quantité, c'est-à-dire en anulant chaque composante du score $S(Y_1, ..., Y_n; \theta)$. Le $j$-ième élément de ce vecteur est

$$\frac{\partial l}{\partial \theta_j}(Y_1, ..., Y_n; \theta) = \sum_{i=1}^{n} -e_i X_i^j e^{X_i\theta} + Y_i X_j.$$

Supposons maintenant que l'on remplace $Y_i$ par $Z_i = \frac{Y_i}{e_i}$ et qu'on mette en application un GLM Poisson avec la fonction de lien logarithmique tel que $Z_i|X = x_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{P}(\lambda(x_i))$ et $\ln(\lambda(X_i)) = X_i\theta$ avec $e_i$ en poid. La log-vraisemblance s'écrit

$$l(Z_1, ..., Z_n; \theta) = \sum_{i=1}^{n} -e_i\lambda(X_i) + e_i Z_i \ln(\lambda(X_i)) - e_i \ln(Z_i!).$$

En utilisant la fonction de lien logarithmique, on calcule le $j$-ième élément du score

$$\frac{\partial l}{\partial \theta_j}(Z_1, ..., Z_n; \theta) = \sum_{i=1}^{n} -e_i X_i^j e^{X_i\theta} + e_i Z_i X_j = \sum_{i=1}^{n} -e_i X_i^j e^{X_i\theta} + e_i \frac{Y_i}{e_i} X_j = \frac{\partial l}{\partial \theta_j}(Y_1, ..., Y_n; \theta).$$

Dans les deux cas, on cherche donc à estimer $\theta$ en annulant la même quantité. Les deux modèles sont donc équivalents - *Fin de la démonstration* -.

# Annexe B

# Analyses descriptives complémentaires

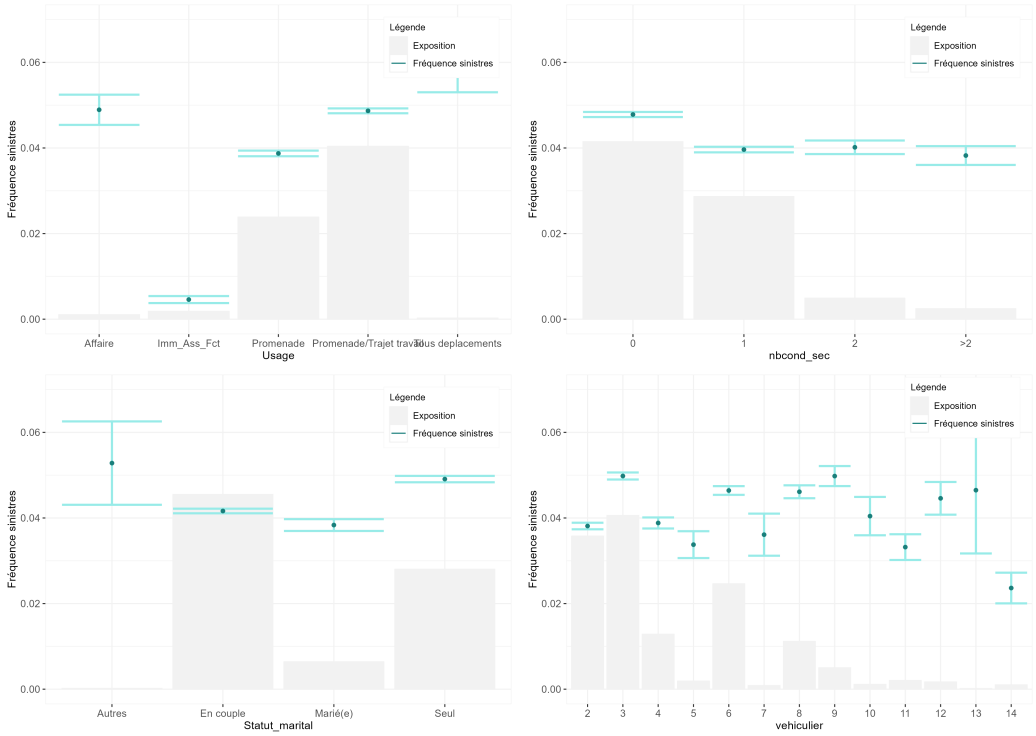



FIGURE B.1 : Statistiques descriptives pour quelques variables tarifaires

La Figure B.1 présente une statistique descriptive pour les autres variables. A titre d'exemple, il est possible de constater au niveau du nombre de conducteurs secondaires (nbcond_sec) que la fréquence de sinitres est plus importante lorsqu'il n'y a pas de conducteurs secondaires. La fréquence de sinistres semble aussi plus importante chez les conducteurs "célibataires". En ce qui concerne le véhiculier, des regroupements en 14 classes ont été effectués par un algorithme CART, et la plupart des assurés possèdent des véhicules catégorisés classe 2 ou 3, avec une plus forte sinistralité chez ces derniers.