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

le 8 Septembre 2022

Par : Junior Désiré ASSI
Titre : Apport de données télématiques dans la modélisation du risque géographique en assurance automobile.

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Sonia GUELOU
Anthony DERIEN
Signature :

Entreprise :
ADDACTIS FRANCE
Signature :

Membres présents du jury de l'EURIA :
Franck VERMET

Directeur de mémoire en entreprise :
Nabil RACHDI
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

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

Apport de données télématiques
Dans la modélisation du risque géographique
En assurance automobile

addactis
THE RISKTECH FOR INSURANCE

En partenariat avec

MICHELIN

Junior ASSI

Résumé

La prise en compte de l'aspect géographique du portefeuille de l'assureur est un facteur d'amélioration considérable de la qualité des modèles tarifaires en assurance IARD. Celle-ci passe par une modélisation spécifique visant à segmenter le risque grâce à divers critères géospatiaux. Ces dernières années, de nouveaux critères basés sur les habitudes de conduite ont vu le jour dans le domaine de l'assurance automobile. Ce sont des données dites télématiques.

L'objectif de ce mémoire est d'évaluer l'apport de ces données, agrégées à une maille géographique, sur l'amélioration de la significativité de la segmentation dans ce domaine. Pour atteindre cet objectif, deux méthodologies sont proposées. La première méthode est l'ajout d'informations géographiques dans un modèle classique d'assureur (GLM) en y intégrant directement une sélection de ces variables télématiques. La seconde méthode consiste à synthétiser par un algorithme de *Machine Learning*, ces informations géographiques en une unique variable appelée zonier et qui est ensuite intégrée dans le modèle GLM. Enfin, les différents impacts observés sur le modèle tarifaire suivant chacune de ces méthodes sont présentés.

Mots clefs: Assurance automobile, Tarification, Segmentation, Risque géographique, Télématique, GLM, Machine Learning, Lissage géospatial, Zonier.

Abstract

Considering the geographical aspect of the insurer's portfolio is a factor that considerably improves the quality of P&C insurance pricing models. This requires specific modeling to segment the risk using various geospatial criteria. In recent years, new criteria based on driving habits have emerged in the field of car insurance. These are the telematics data.

This thesis aims therefore at evaluating the contribution of these data, aggregated at a geographical level, on the improvement of the segmentation significance in this domain. In order to achieve this objective, two methodologies are proposed. The first method consists in adding geographical information in a classical insurance model (GLM) by directly integrating a selection of these telematics data. The second method consists in synthesizing by a Machine Learning algorithm, this geographical information into a single variable called risk zoning which is then integrated into the GLM model. Finally, the different impacts observed on the pricing model following each of these methods are presented.

Keywords: Car insurance, Pricing, Segmentation, Geographic risk, Telematics, GLM, Machine Learning, spatial smoothing, zoning.

Note de synthèse

Il s'agit de l'histoire d'un assureur automobile, à la recherche de nouveaux instruments lui permettant d'améliorer sa prise en compte du risque géographique dans ses modèles tarifaires. L'impact géographique étant un des facteurs les plus discriminants pour la sinistralité en assurance automobile, celui-ci espère, en améliorant la prise en compte de ce facteur, proposer des tarifs plus ajustés et plus compétitifs sur le marché.

Informé de la récente apparition d'une nouvelle typologie de données dites télématiques qui, renseignent les habitudes de conduite des conducteurs, cet assureur réfléchit à leur possible utilisation dans la modélisation de ce risque géographique. A cet effet, il se pose deux questions fondamentales :

- *La connaissance des habitudes de conduite des conducteurs dans les différentes zones du territoire peut-elle permettre de comprendre le risque sous-jacent à ces dernières ?*
- *Si oui, quels sont les réels apports techniques et opérationnels de ces nouvelles données dans la modélisation du risque géographique ?*

Pour répondre à ces interrogations, l'assureur décide de se procurer auprès de **MICHELIN**, leader mondial de la fabrication et de la commercialisation de Pneumatiques, une base de données contenant des variables télématiques fiablement collectées sur un échantillon représentatif de conducteurs français et agrégées à la maille communale.

Les données de cette base externe dénommée *Smart road data* peuvent être regroupées en quatre grandes catégories :

- **Les scores de comportement** : ce sont des notes sur la manière de freiner et sur l'allure de la conduite dans les différentes communes.
- **Les variables concernant l'usage du véhicule** : elles renseignent les kilomètres parcourus et les temps de trajet suivant différents contextes (météo, relief, urbanité,..).
- **Les variables de saisonnalité** : elles informent sur les niveaux du trafic automobile dans les communes selon différentes périodes de l'année.

- **Les variables cartographiques** : ce sont des variables sur les types de routes à l'intérieur des communes.

Au delà de l'aspect "clé en main" des données télématiques contenues dans la base *Smart road data*, un autre avantage de celles-ci est la **contextualisation** dont bénéficient quelques unes d'entre elles. Cette contextualisation permet, en effet, d'affiner la connaissance des habitudes de conduite des conducteurs sur les différentes zones du territoire.

Après une étude statistique approfondie de ces données externes, l'assureur les relie aux données internes dont il dispose par le biais de la clé de jointure géographique INSEE. Il est dès lors prêt à évaluer l'apport de ces données télématiques dans la modélisation du risque géographique.

Il décide de mener cette étude sur sa garantie Responsabilité Civile Matérielle (RCM) qui recouvre l'ensemble de son portefeuille d'assurés. Sa stratégie d'évaluation comporte deux volets :

- **un volet technique**
  Dans un premier temps, l'assureur souhaite analyser les gains statistiques découlant de l'utilisation des données télématiques suivant deux différentes approches de modélisation du risque géographique.
- **un volet opérationnel**
  Dans un second temps, il compte étudier les déformations de son tarif, qui résultent d'une modélisation du risque géographique basée sur ces variables télématiques.

Évaluations techniques

Modélisation du risque géographique par approche naïve

La première évaluation technique opérée par l'assureur se base sur une modélisation naïve du risque géographique. Dans cette approche, l'impact des comportements à l'intérieur des zones géographiques sur la fréquence de sinistres est modélisé par une combinaison des variables télématiques reflétant les habitudes de conduite dans ces zones. L'assureur injecte donc dans un **modèle GLM initial** (sans données géographiques) de fréquence de sinistres, une sélection de ses données télématiques externes et obtient un nouveau modèle de fréquence de sinistres appelé **modèle naïf**. En outre, dans l'optique d'exploiter davantage ce modèle naïf, ce dernier mène une étude d'interactions entre les variables de ce modèle.

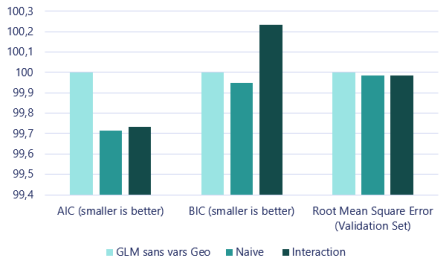

FIGURE 1 – *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle initial sans variables géographiques)*

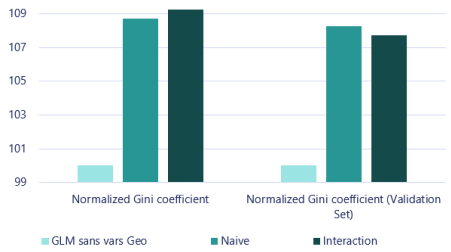

FIGURE 2 – *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle initial sans variables géographiques)*

Les graphiques ci-dessus mettent en lumière une amélioration générale des indicateurs de la qualité statistique du modèle initial (sans variables géographiques) lorsqu'on lui fournit des informations télématiques géospatialisées.

Il est possible d'observer qu'en présence d'informations télématiques, le modèle discrimine mieux la fréquence de sinistres. Cela s'observe via l'amélioration de plus de 7% de la valeur du Gini normalisé lors du passage du modèle initial aux nouveaux modèles naïf et naïf contenant une interaction.

Il ressort de cette première évaluation technique que la modélisation naïve du risque géographique par une combinaison de variables télématiques permet d'améliorer la qualité statistique et la précision du modèle de fréquence de sinistres. Les habitudes de conduite dans les différentes zones seraient donc effectivement des facteurs permettant de comprendre le risque géographique en assurance automobile.

Cependant, dans la pratique, la mise en production de cette approche peut être assez difficile pour l'assureur. En effet, rajouter plusieurs variables à sa structure tarifaire de base constitue une trop grande modification à gérer. Il décide donc de palier cette contrainte en synthétisant les informations géographiques apportées par les données télématiques externes en une unique variable appelée zonier.

Modélisation du risque géographique par zonier

La deuxième évaluation technique opérée par l'assureur se base sur une modélisation du risque géographique par zonier. Un zonier est une classification des différentes zones géographiques suivant les niveaux de risque qui leur sont rattachés. Dans cette approche, l'assureur s'attelle à construire deux zoniers à partir des résidus géographiques issus de la modélisation initiale (sans variable géographique) de la fréquence de sinistres.

Le premier zonier construit est un **zonier traditionnel**. Il est obtenu par un lissage géospatial par crédibilité de ces résidus. Ce zonier représente le zonier standard de l'as-

sureur. Son intégration dans le modèle initial de fréquence de sinistres permet d'obtenir le **modèle de référence**.

Le second zonier construit est un **zonier innovant**. Son obtention passe d'abord par une étape de modélisation statistique des résidus avec pour variables explicatives les données télématiques externes, puis par un lissage géospatial par crédibilité des prédictions de ce modèle. Ce zonier est dit innovant car sa construction combine à la fois une méthodologie moderne et l'intégration d'informations rarement utilisées pour l'établissement d'un zonier sur le marché de l'assurance automobile. L'intégration de ce zonier dans le modèle initial de fréquence de sinistres permet d'obtenir le **modèle innovant**.

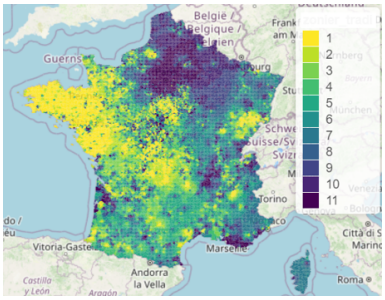

FIGURE 3 – *Zonier traditionnel*

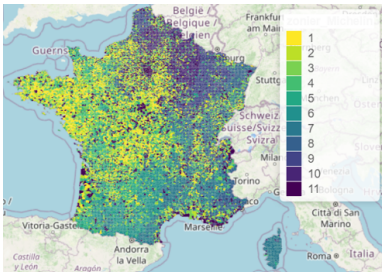

FIGURE 4 – *Zonier innovant*

Les zones de ces zoniers sont disposées par ordre croissant de niveau de risque. La zone 1 représente donc les communes les moins risquées pour l'assureur et la zone 11 renferme les communes les plus risquées pour ce dernier.

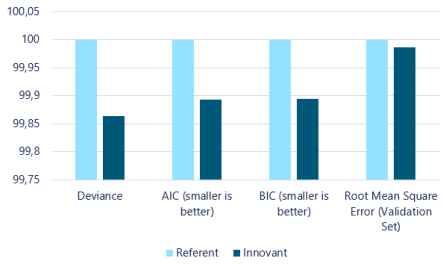

FIGURE 5 – *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle référent)*

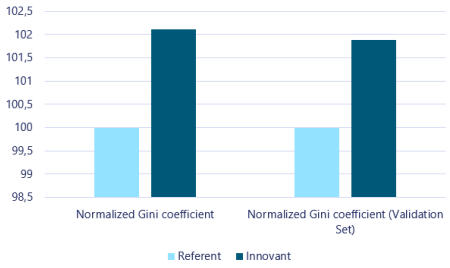

FIGURE 6 – *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle référent)*

Les figures ci-dessus semblent démontrer que le modèle innovant est statistiquement de meilleure qualité que le modèle référent. Cependant, les écarts de performances entre ces deux modèles sont relativement moins conséquents que ceux obtenus lors des comparaisons dans l'approche naïve. Ici, les indices de Gini évoluent au maximum de 2% en passant du modèle référent au modèle innovant. Ce constat est totalement objectif étant donné que dans cette nouvelle approche, les deux modèles comparés contiennent

une couche de segmentation géographique. Un trop grand écart de performances entre ces derniers aurait pu signifier la mauvaise construction d'un des zoniers au profit de l'autre.

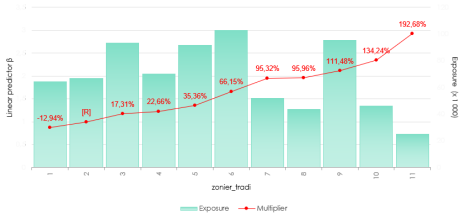

FIGURE 7 – *Courbe des facteurs multiplicatifs des modalités du zonier traditionnel*

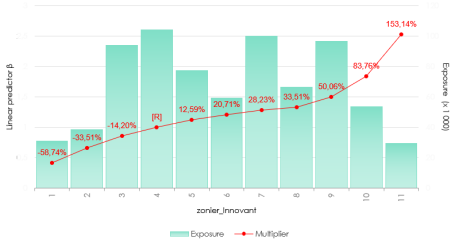

FIGURE 8 – *Courbe des facteurs multiplicatifs des modalités du zonier innovant*

Les zones de risque des deux zoniers analysés sont cohérentes avec la fréquence de sinistres. En effet, le passage d'une zone moins risquée à une zone plus risquée entraîne une variation positive entre les facteurs multiplicatifs de ces deux zones. Cela justifie la pente positive de ces deux zoniers. Ainsi, passer d'une zone moins risquée d'un de ces zoniers à une zone plus risquée du même zonier augmente la probabilité d'avoir un sinistre.

	Zonier traditionnel	zonier innovant
Multiplier 1	0,8706	0,4126
Multiplier 11	2,9268	2,5314
Exposition 1	64425	30870
Exposition 11	25207	29420
spread standard	**70,25%**	**83,70%**
spread normalisé	**27,49%**	**79,77%**

FIGURE 9 – *Tableau récapitulatif des comparaisons de spread des zoniers traditionnel et innovant*

D'après ce tableau de synthèse, quelque soit la formule de calcul utilisée, le *spread* du zonier innovant surpasse celui du zonier traditionnel. Cela signifie qu'en terme de discrimination du risque, le zonier innovant est nettement meilleur que le zonier traditionnel.

Toutes ces analyses présentées se rejoignent une nouvelle fois sur la significativité de l'apport des données télématiques dans la modélisation du risque géographique en assurance automobile.

Avant de passer à l'étape d'évaluation opérationnelle, l'assureur, assez satisfait de ces premiers résultats techniques, décide d'investiguer la possibilité d'accroître ces performances statistiques en associant aux données télématiques un tout autre type de données qui a fait ses preuves dans la modélisation du risque géographique depuis quelques années : les *Open data*.

Les *Open data* sont des données numériques, provenant généralement de structures publiques dont, l'accès, l'usage, la modification et la rediffusion sont librement ouverts à

tous les usagers. Elles sont de divers ordres (socio-économique, emploi, réseaux routiers, météo,...).

A l'issu de cette courte investigation menée de façon connexe à l'évaluation de l'apport des données télématiques dans la modélisation du risque géographique, l'assureur a pu découvrir que l'alliance *Open data* - données télématiques est très prometteuse. Elle mérite donc d'être mieux étudiée de manière à profiter des avantages qu'offrent ces deux types de données très complémentaires.

Évaluations opérationnelles

Déformations de la segmentation géographique de l'assureur

Ici, il est supposé que l'assureur utilise présentement pour sa modélisation du risque géographique le zonier traditionnel déjà présenté.

Après avoir apprécié statistiquement l'apport du zonier innovant dans son modèle de fréquence de sinistres, l'attention de l'assureur s'est portée sur la déformation de ces zones de risque lors du passage de son zonier traditionnel actuel au nouveau zonier innovant.

Afin de quantifier la migration d'une commune entre ces deux zoniers, celui-ci utilise un indicateur appelé *Switch* qui se base sur la différence des numéros des zones dans lesquelles se trouve cette commune dans chacun de ces zoniers.

$$Switch = zonier_{traditionnel} - zonier_{innovant}$$

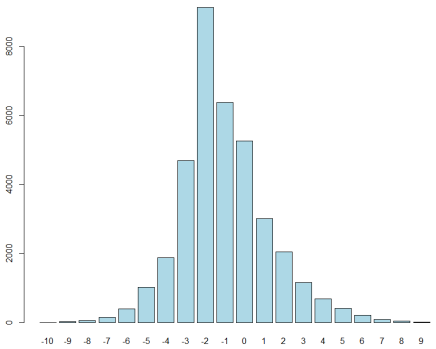

FIGURE 10 – *Distribution du switch entre les zoniers traditionnel et innovant*

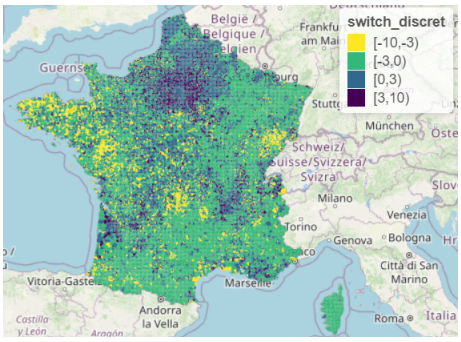

FIGURE 11 – *Cartographie du switch entre les zoniers traditionnel et innovant*

La figure 10 montre que la répartition du *Switch* entre les zoniers traditionnel et innovant est concentrée entre -3 et +2 avec un pic en -2. La déformation apportée

par le zonier innovant n'est donc ni trop élevée ni trop faible. Le mode de cette série statistique discrète valant -2, cela signifie que sur la plupart des communes, le zonier innovant suggère un sur-classement du risque de deux classes par rapport à la classe proposée par le zonier traditionnel.

D'après la cartographie (Figure 11), les classes de *Switch* extrêmes se localisent principalement sur la Bretagne (sur-classement extrême) et sur l'Île-de-France (déclassement extrême). Dans la majeure partie du reste de la France, le *Switch* est assez modéré, variant entre -3 et 2.

Cette dernière cartographie laisse l'assureur un peu perplexe. En effet, il est vrai que le *Switch* semble assez modéré sur une grande partie du territoire. Néanmoins les quelques parties touchées par les *Switch* extrêmes, très faibles en nombre, semblent représenter une part non négligeable de l'exposition de celui-ci, eu égard à leur localisation.

Cela motive l'assureur à évaluer l'impact de ces différents *Switch* sur sa prime pure afin d'achever de se convaincre de leur réelle importance.

Impact des migrations de zones sur la prime pure

L'assureur procède au calcul du delta entre les primes pures innovantes [1] et les primes pures de référence [2] :

$$Delta = \frac{Prime_{inno}}{Prime_{ref}}$$

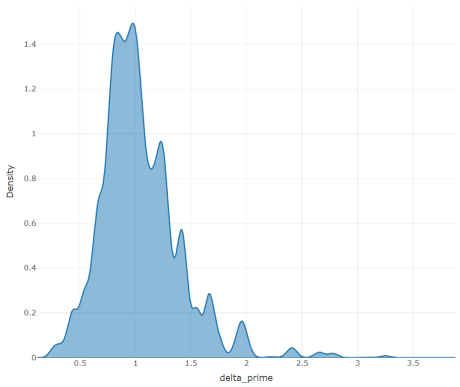

FIGURE 12 – *Distribution du delta entre les primes innovantes et les primes de référence*

Le delta varie entre 0.25 et 3.75 (bornes extrêmes très faiblement représentées). Il se concentre dans l'intervalle [0.5 ;1.5] et sa valeur moyenne est de 1. Cette concentration du delta entre 0.5 et 1.5 montre qu'en moyenne, la minoration (respectivement majoration) maximale suggérée par la structure innovante revient à diviser(respectivement

1. primes pures estimées par la structure tarifaire contenant le zonier innovant
2. primes pures estimées par la structure tarifaire contenant le zonier traditionnel

multiplier) le tarif de référence par 2, ce qui paraît acceptable.

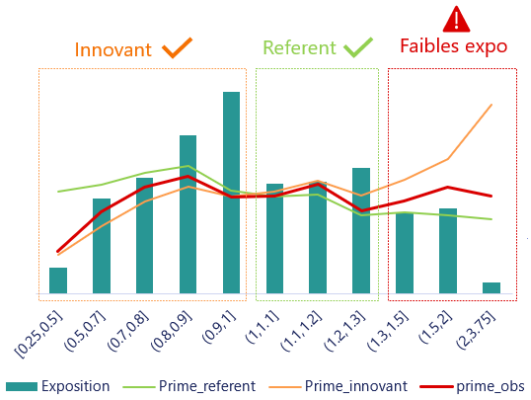

FIGURE 13 – *Comparaison des primes innovantes et de référence relativement aux burning cost*

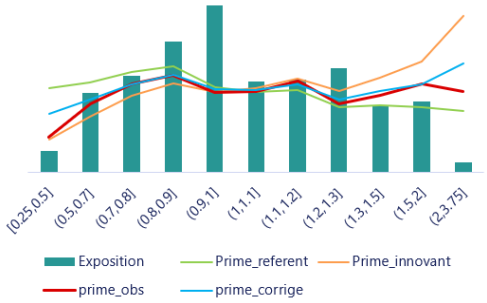

FIGURE 14 – *Représentation de la prime pure corrigée*

L'analyse de ces delta a permis à l'assureur d'obtenir une structure de primes pures plus adaptée. Ainsi, l'objectif initial de ce dernier, qui était d'ajuster son tarif en améliorant sa segmentation du risque géographique a donc bien été atteint.

En définitive, les données télématiques sont donc effectivement un moyen efficace pour comprendre le risque géographique en assurance automobile afin de parfaire sa modélisation.

Executive summary

This is the story of a car insurer looking for new tools to improve the way it takes geographic risk into account in its pricing models. Geographic impact is one of the most discriminating factors in car insurance, this insurer hopes to improve its ability to take this into account the market with more adjusted and competitive rates.

Having been made aware of the recent appearance of a new style of data called telematics, which provides information on the driving habits of drivers, this insurer is considering their possible use in its modeling of this geographic risk.
He is getting two fundamental questions :

- *Can knowledge of the driving habits of drivers in different areas of the territory provide insight into the risk underlying these ?*
- *If so, what are the real technical and operational contributions of these new data in geographic risk modeling ?*

To answer these interrogations, the insurer decides to obtain from **Michelin**, world leader in the manufacture and the marketing of tires, a data base containing telematic variables reliably collected on a representative sample of French drivers and aggregated at the communal level.

The data of this external database named Smart road data can be grouped in four main categories :

- **Behavioral scores** : These are notes on how to brake and how to drive in different municipalities.
- **Car usage data** : they provide information on the kilometers traveled and the travel time according to different contexts (weather, relief, urbanity, etc.).
- **Seasonality data** : They provide information on traffic levels in the municipalities at different times of the year.
- **Cartographic data** : these are variables on the types of roads within the municipalities.

Beyond the "turnkey" aspect of the telematic data contained in the Smart Road Data database, another advantage is the contextualization of some of them. This contextualization makes it possible to refine the knowledge of the driving habits of drivers in different areas of the territory.

After an in-depth statistical study of these external data, the insurer links them to the internal data through the INSEE geographic join key. He is then ready to evaluate the contribution of these telematic data in the modeling of the geographic risk.

He decided to conduct this study on its Material Liability cover, which covers its entire portfolio of insureds. Its evaluation strategy has two components :

- **a technical component**
  First, the insurer wishes to analyze the statistical gains resulting from the use of telematics data according to two different approaches to modeling geographic risk.
- **an operational component**
  In a second step, he intends to study the distortions in his technical premium that result from geographic risk modeling based on these telematic variables.

Technical evaluations

Geographic risk modeling using a naive approach

The first technical assessment made by the insurer is based on a naive modeling of geographic risk. In this approach, the impact of behavior within geographic areas on the claims frequency is modeled by a combination of telematic variables reflecting driving habits in those areas. The insurer therefore injects a selection of its external telematics data into an **initial GLM** claims frequency model (without geographical data) and obtains a new claims frequency model called a **naive model**. In addition, in order to further exploit this naive model, he conducts an interaction study between the variables of this model.

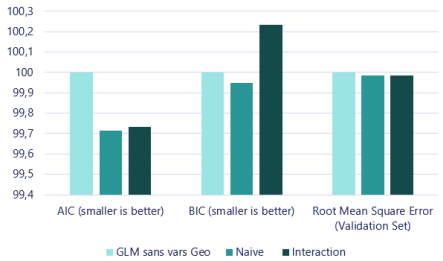

FIGURE 15 – *Comparison of the quality indicators of the models (base 100 with the values of the indicators of the initial model without geographical data as reference)*

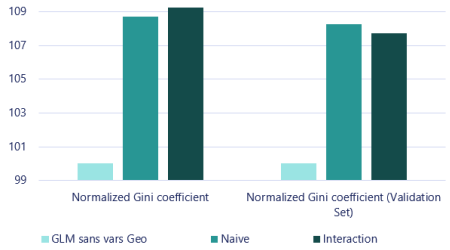

FIGURE 16 – *Comparison of standardized Gini indices (base 100 with the values of the indicators of the initial model without geographical data as reference)*

The graphs above highlight a general improvement in the statistical quality indicators of the initial model (without geographic data) when provided with geospatialized telematics information.

For instance, it is possible to observe that in the presence of telematic information, the model better discriminates the claims frequency. This can be observed through the improvement of more than 7% of the value of the normalized Gini when passing from the initial model to the new naive and naive containing an interaction models.

This first technical evaluation shows that naive modeling of geographic risk using a combination of telematics variables improves the statistical quality and accuracy of the claims frequency model. Driving habits in different areas would therefore indeed be factors in understanding geographic risk in car insurance.

However, in practice, the implementation of this approach can be quite difficult for the insurer. Indeed, adding several variables to its basic pricing structure is too much to manage. The insurer therefore decides to overcome this constraint by synthesizing the geographic information provided by the external telematics data into a single variable called risk zoning.

Geographic risk modeling using a risk zoning

The second technical evaluation performed by the insurer is based on a geographic risk model by risk zoning. A risk zoning is a classification of the different geographical areas according to the levels of risk attached to them. In this approach, the insurer build two risk zonings from the geographic residuals resulting from the initial modeling (without geographic data) of the claims frequency.

The first one built is a **traditional risk zoning**. It is obtained by a geospatial smoothing of these residuals. This risk zoning represents the standard of any insurer. Its integration in the initial claims frequency model allows him to obtain the **benchmark**

model.

The second one built is an **innovative risk zoning**. It is obtained first by a statistical modeling of the residuals with external telematic data as explanatory variables, and then by a geospatial smoothing of this model predictions. This risk zoning is said to be innovative because its construction combines both a modern methodology and the integration of information rarely used for the establishment of a risk zoning in the car insurance market. The integration of this risk zoning in the initial claims frequency model allows us to obtain the **innovative model**.

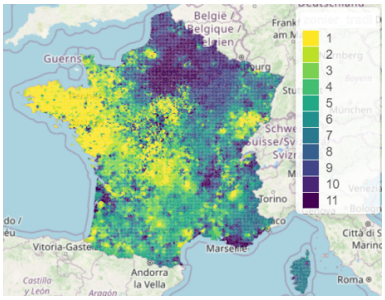

FIGURE 17 – *Traditional risk zoning*

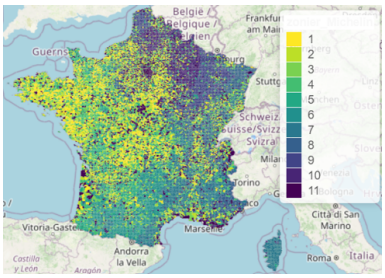

FIGURE 18 – *Innovative risk zoning*

The modalities of these risk zonings are arranged in ascending order of risk level. Modality 1 represents the least risky municipalities for the insurer and modality 11 contains the riskiest municipalities for the insurer.

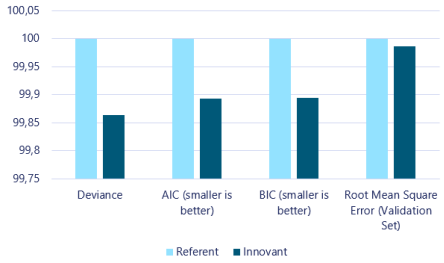

FIGURE 19 – *Comparison of the quality indicators of the models (base 100 with the values of the indicators of the benchmark model as reference)*

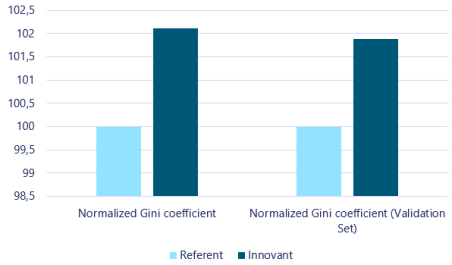

FIGURE 20 – *Comparison of normalized Gini indexes (base 100 with the values of the indicators of the benchmark model as reference)*

The figures above seem to show that the innovative model is statistically better than the benchmark model. However, the differences in performance between these two models are relatively smaller than those obtained in the naive approach. Here, the Gini indices evolve by a maximum of 2% when moving from the benchmark model to the innovative model. This finding is completely objective since in this new approach, both models compared contain a geographic segmentation layer. Too great a difference in

performance between them could have meant that one of the risk zonings was poorly built in favor of the other.

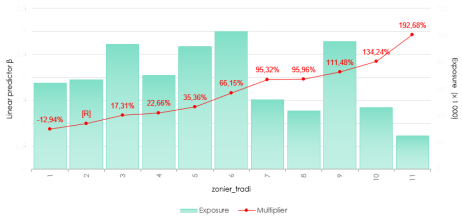

FIGURE 21 – *Curve of multiplicative factors of the modalities of the traditional risk zoning*

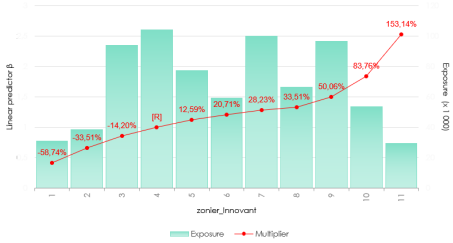

FIGURE 22 – *Curve of multiplicative factors of the modalities of the innovative risk zoning*

The modalities of the both risk zonings analyzed are consistent with the claims frequency. Indeed, the transition from a less risky modality to a more risky modality leads to a positive variation between the multiplicative factors of these two modalities. This justifies the positive spread of these two risk zonings. Thus, moving from a less risky area in one of these risk zoning to a more risky area in the same risk zoning increases the probability of having a claim.

	Zonier traditionnel	zonier innovant
Multiplier 1	0,8706	0,4126
Multiplier 11	2,9268	2,5314
Exposition 1	64425	30870
Exposition 11	25207	29420
spread standard	**70,25%**	**83,70%**
spread normalisé	**27,49%**	**79,77%**

FIGURE 23 – *Summary table of spread comparisons of traditional and innovative risk zonings*

According to this summary table, regardless of the calculation formula used, the innovative risk zoning outperforms the traditional risk zoning. This means that in terms of risk discrimination, the innovative risk zoning is clearly better than the traditional risk zoning.

All of the analyses presented once again agree on the significance of the contribution of telematics data in modeling geographic risk in car insurance.

Before moving on to the operational evaluation stage, the insurer, quite satisfied with these initial technical results, decided to investigate the possibility of increasing these statistical performances by associating with telematic data another type of data which has proved its worth in the geographic risk modelling over the last few years : the Open data.

The Open data are digital data, generally coming from public structures, whose access, use, modification and redistribution are freely open to all users. They are of various kinds (socio-economic, employment, road networks, weather,...).

At the end of this short investigation conducted in connection with the evaluation of the contribution of telematics data in geographic risk modeling, the insurer was able to discover that the alliance between telematics data and open data is very promising. It deserves to be better studied in order to take advantage of the benefits offered by these two very complementary types of data.

Operational evaluations

Distortions in the insurer's geographic segmentation

Here, it is assumed that the insurer is currently using the traditional risk zoning already presented for its geographic risk modeling.

After having statistically assessed the contribution of the innovative risk zoning in its claims frequency model, the insurer's attention was focused on the distortion of its geographic segmentation when switching from its current traditional risk zoning to the new innovative risk zoning.

In order to quantify the migration of a municipality between these two risk zonings, he uses an indicator called *Switch* which is based on the difference of the numbers of the risk zoning modality in which this municipality is located within each of these risk zoning.

$$Switch = risk\ zoning_{traditionnal} - risk\ zoning_{innovative}$$

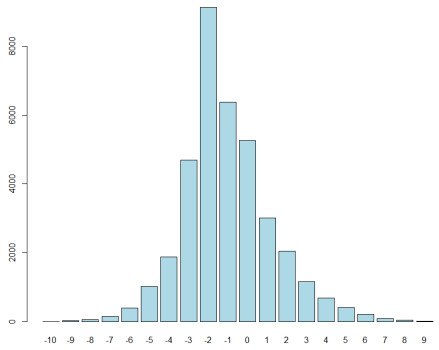

FIGURE 24 – *Switch distribution between traditional and innovative risk zonings*

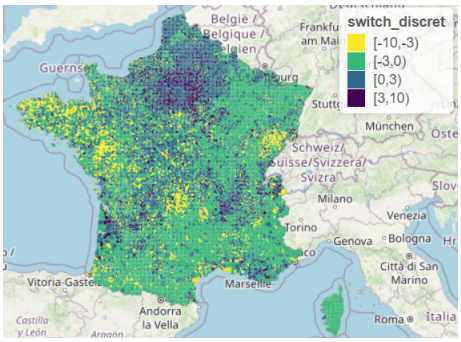

FIGURE 25 – *Cartography of switch between traditional and innovative risk zonings*

The figure 24 shows that the distribution of the *Switch* between the traditional and innovative risk zonings is concentrated between -3 and +2 with a spike at -2. The distortion contributed by the innovative risk zoning is therefore neither too high nor too low. The mode of this discrete statistical series being -2, this means that on most of the

municipalities, the innovative risk zoning suggests an over-classification of the risk by two classes compared to the class proposed by the traditional risk zoning.

According to the cartography (Figure 25), the extreme *Switch* classes are located mainly on Bretagne (extreme over-classification) and on Île-de-France (extreme downgrading). In most of the rest of France, the *Switch* is quite moderate, varying between -3 and 2.

This last cartography leaves the insurer a little perplexed. Indeed, it is true that the *Switch* seems to be quite moderate over a large part of the territory. Nevertheless, the few parts affected by the extreme *Switch* are certainly very weak in number, but they seem to represent a not negligible part of the exposure of the insurer, considering their localization.

It motivates him to evaluate the impact of these different *Switch* on his technical premium in order to convince himself of their real importance.

Impact of *Switch* on the technical premium

The insurer calculates the delta between the innovative technical premiums [3] and the benchmark technical premiums [4] :

$$Delta = \frac{Premium_{inno}}{Premium_{benchk}}$$

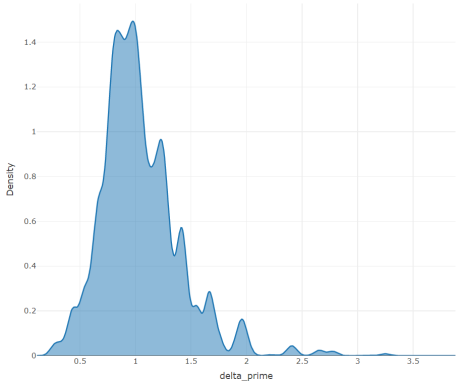

FIGURE 26 – *Delta distribution between innovative and benchmark technical premiums*

The delta varies between 0.25 and 3.75 (extreme limits very weakly represented). It is concentrated in the interval [0.5 ;1.5] and its average value is 1. This concentration of the delta between 0.5 and 1.5 shows that, on average, the maximum reduction (respectively increase) suggested by the innovative structure amounts to dividing (respectively

3. technical premiums estimated by the pricing structure contening the innovative risk zoning
4. technical premiums estimated by the pricing structure contening the traditionnal risk zoning

multiplying) the benchmark premium by 2, which seems acceptable.

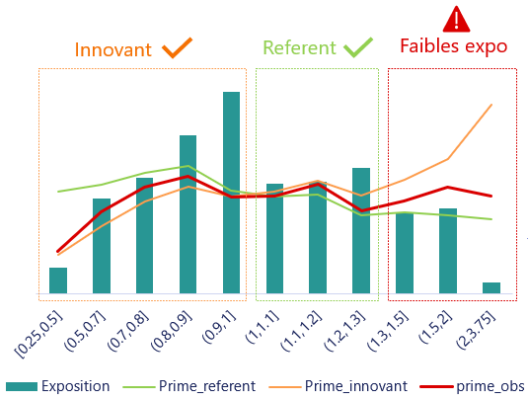

FIGURE 27 – *Comparison of innovative and benchmark premiums relative to burning costs*

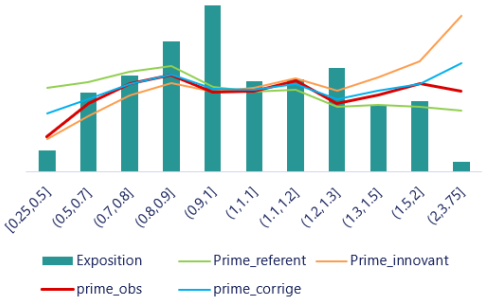

FIGURE 28 – *Representation of technical premium corrected*

The analysis of these delta allowed the insurer to obtain a more appropriate technical premium structure. Thus, the insurer's initial objective, which was to adjust its rate by improving its geographic risk segmentation, was indeed achieved.

Ultimately, therefore, telematics data is indeed an effective way to understand geographic risk in car insurance in order to perfect its modeling.

Remerciements

« *La préparation d'un mémoire n'est pas un sprint mais un marathon [...]* »
Tel était le premier conseil que me donnait mon tuteur en entreprise au début de mes travaux. Il a raison.

Je tiens en ces quelques lignes à remercier toutes ces personnes qui, de près ou de loin, m'ont aidé, épaulé et conseillé, tout comme le public qui au bord du parcours encourage les marathoniens jusqu'à la ligne d'arrivée.

Je voudrais, tout d'abord, adresser mes sincères remerciements à Nabil RACHDI, mon tuteur en entreprise, pour la proposition de ce sujet très passionnant, pour son encadrement et pour tous ses conseils très avisés dont j'ai pu bénéficier.

Un grand merci également à toute l'équipe P&C Pricing & Analytics d'Addactis France pour les belles expériences partagées. De façon particulière, je tiens à exprimer toute ma gratitude aux consultants Victoria DELAVAUD et Cédric DENNIEL pour les conseils à la fois techniques et rédactionnels qu'ils m'ont donnés.

Mes remerciements vont aussi aux membres de l'équipe DDI du groupe Michelin, pour la confiance qu'ils m'ont accordée sur ce projet.

Je voudrais remercier aussi, l'ensemble du corps enseignant de l'Euro Institut d'Actuariat (EURIA) de Brest pour la qualité de la formation dispensée et reçue pendant ces trois années d'études supérieures. J'adresse particulièrement mes remerciements à mon tuteur académique Franck VERMET, pour sa disponibilité et ses précieux conseils pendant nos séances de travail en rapport avec le mémoire.

Enfin, je tiens à exprimer mon infinie reconnaissance à ma famille et à mes amis qui, tout au long de ce parcours, n'ont cessé de croire en moi en m'encourageant. Leur soutien a été ma plus grande motivation.

Table des matières

Résumé i

Note de synthèse v

Remerciements xxi

Glossaire 1

Introduction 3

1 Contexte et enjeux 5
- 1.1 Assurance automobile . . . 6
  - 1.1.1 Généralités . . . 6
  - 1.1.2 Garanties en assurance automobile . . . 6
  - 1.1.3 Quelques chiffres clés du marché en 2020 . . . 7
- 1.2 Risque géographique en assurance automobile . . . 9
  - 1.2.1 Segmentation : définition, importance et limites . . . 10
  - 1.2.2 Segmentation par des critères géospatiaux : Une vision géographique du risque en assurance automobile . . . 12
- 1.3 A la découverte de la télématique . . . 13
  - 1.3.1 Définition de la télématique . . . 13
  - 1.3.2 Télématique embarquée dans le milieu de l'automobile . . . 13
  - 1.3.3 RGPD et utilisation éthique des données télématiques . . . 14
  - 1.3.4 Quand la télématique s'invite en assurance automobile . . . 15
- 1.4 Enjeux du mémoire . . . 16

2 Descriptions statistiques des sources de données de l'étude 19
- 2.1 Présentation de la base interne : données d'un assureur . . . 20
  - 2.1.1 Description des variables de la base interne . . . 20
  - 2.1.2 Analyse temporelle du portefeuille . . . 21
- 2.2 Présentation de la base externe : *Smart Road Data* . . . 23
  - 2.2.1 Les scores de comportement . . . 24
  - 2.2.2 Les variables concernant l'usage du véhicule . . . 25
  - 2.2.3 Les variables de saisonnalité . . . 26

2.2.4 Les variables cartographiques . . . 26
2.3 Analyse statistique de la base *Smart Road Data* . . . 27
2.3.1 Étude des corrélations entre les variables . . . 27
2.3.2 ACP et réduction de dimensions . . . 30
2.4 *Feature Engineering* et sélection non supervisée . . . 33
2.4.1 *Feature Engineering* . . . 33
2.4.2 Sélection non supervisée des variables de la base *Smart Road Data* 37
2.5 Fusion des deux sources de données . . . 38

3 Modélisation classique de la prime pure 41
3.1 Modèle collectif . . . 42
3.2 Théorie des Modèles Linéaires Généralisés (GLM) . . . 43
3.3 Sélection supervisée de variables et qualité de modèle . . . 44
3.3.1 Méthodes de Sélection supervisée de variables . . . 44
3.3.2 Indicateurs de qualité d'un modèle . . . 46
3.4 Application sur le portefeuille d'assurance de l'étude . . . 50
3.4.1 Modélisation de la survenance de sinistres . . . 51
3.4.2 Modélisation des coûts moyens . . . 56
3.4.3 Modélisation du burning cost . . . 57

4 Modélisation du risque géographique 1 : approche naïve 59
4.1 Présentation théorique de l'approche naïve . . . 60
4.2 Sélection supervisée et analyses des variables . . . 60
4.2.1 Sélection supervisée des variables . . . 60
4.2.2 Analyses des variables retenues par sélection supervisée . . . 63
4.3 Étude des interactions entre les variables sélectionnées . . . 68
4.3.1 Théorie des arbres de décision CART . . . 68
4.3.2 Application du CART pour la détection d'interactions . . . 70
4.3.3 Test de significativité et sélection des interactions . . . 71
4.4 Première évaluation de l'apport des données télématiques contextualisées 73
4.4.1 Comparaisons des performances statistiques des modèles : initial et naïfs . . . 73
4.4.2 Comparaison visuelle des prédictions des modèles : initial et naïf . 74

5 Modélisation du risque géographique 2 : Zonier 77
5.1 Présentation des étapes de la construction d'un zonier . . . 78
5.2 Calcul et agrégation géographique des résidus . . . 78
5.3 Méthodes de traitements du signal géographique . . . 82
5.3.1 Théorie du lissage géospatial par crédibilité . . . 82
5.3.2 Théorie des forêts aléatoires . . . 86
5.3.3 Méthodes de clustering . . . 89
5.4 Application de ces méthodes pour la construction de zoniers . . . 90
5.4.1 Construction d'un zonier traditionnel . . . 90
5.4.2 Construction d'un zonier innovant . . . 93

5.5 Deuxième évaluation de l'apport des données télématiques contextualisées 96
5.5.1 Comparaisons statistiques des performances des modèles : référent et innovant . . . 96
5.5.2 Comparaison du spread des zoniers : traditionnel et innovant . . . 97
5.6 Open data et Smart road data, une union envisageable ? . . . 99
5.6.1 Comparaisons statistiques des performances des modèles : Open data et hybride . . . 100
5.6.2 Comparaison du spread des zoniers : Open data et hybride . . . 101

6 Évaluation actuarielle : Quels impacts sur la prime de l'assureur ? 103
6.1 Analyse des migrations entre les zoniers . . . 104
6.1.1 Switch entre deux zoniers . . . 104
6.1.2 Application : calcul et évaluation du switch entre les zoniers traditionnel et innovant . . . 105
6.2 Impact des changements de zones sur la prime pure de l'assureur . . . 107
6.2.1 Delta entre deux primes . . . 108
6.2.2 Application : calcul et évaluation du delta entre les primes innovantes et les primes de référence . . . 108

Conclusion 113

A Tables descriptives des variables 117

B Splines 119

Bibliographie 120

Table des figures 127

Glossaire

Big data = Quantité massive de données à disposition

CNIL = Commission Nationale de l'Informatique et des Libertés

Feature Engineering = Retraitement des variables

GPS = Global Positioning System

IARD = Icendies Accidents et Risques Divers

INSEE = Institut National de la Statistique et des Etudes Economiques

IRIS = Ilots Regroupés pour l'Information Statistique

OBD = On Board Diagnosis

PAYD = Pay As You Drive

PHYD = Pay How You Drive

RCM = Responsabilité Civile Matérielle

RGPD = Règlement Général sur la Protection de Données

Scoring = Système d'évaluation par le biais de scores ou de notes

UBI = Usage Based Insurance

Introduction

Big data, intelligence artificielle, robot autonome, voiture connectée,...
Ce champ lexical qui hier fut l'apanage du domaine de la science-fiction, est aujourd'hui une isotopie récurrente martelant la connotation réaliste de ces expressions. L'avancée du numérique est donc bien réelle et est impulsée par son unique moteur : **la donnée**.

Cette digitalisation progressive de la société, pilotée par la donnée, impacte profondément tous les pans du tissu socio-économique. Le secteur de l'assurance, en particulier celui de l'assurance automobile dont l'activité dépend inéluctablement de l'exploitation des données est un des domaines les plus touchés par ce boom technologique. En effet, l'émergence de nouvelles mobilités, l'utilisation de voitures beaucoup plus sophistiquées engendrant de nouvelles habitudes de conduite sont entre autres des impacts pouvant influencer la sinistralité sur cette branche de l'assurance.

Face à cette situation, une veille technologique s'impose à tous les acteurs de ce marché qui souhaitent rester compétitifs tout en assurant la pérennité de leur activité. Il leur incombe de constamment rechercher de nouvelles méthodes et de nouvelles sources d'informations afin d'affiner continuellement la connaissance du risque qu'ils souhaitent assurer.

C'est donc dans cette quête d'outils innovants, permettant de mieux comprendre le risque en assurance automobile, que s'inscrivent les travaux de ce mémoire. La problématique à étudier est l'évaluation de la pertinence et de la significativité d'une modélisation du risque géographique basée sur de nouvelles données révélatrices des habitudes de conduite appelées **données télématiques**. En d'autres termes :

Quel est l'apport des données télématiques dans la modélisation du risque géographique en assurance automobile ?

La réponse à cette question nécessite de rapprocher les données de la sinistralité d'un assureur à des variables télématiques agrégées à une maille géographique. Ici, la sinistralité analysée est celle d'une garantie Responsabilité Civile Matérielle (RCM) et les données télématiques utilisées sont agrégées à la maille communale. La démarche adoptée pour répondre à cette interrogation est résumée sur la figure ci-après.


Présentation du contexte et des enjeux & Descriptions des données
Modélisation tarifaire sans données géographiques
Première modélisation du risque géographique : approche naïve
Deuxième modélisation du risque géographique : Zonier
Evaluation des impacts de l'utilisation de la télématique sur la prime pure

Chapitre 1

Contexte et enjeux du mémoire

« Pour bâtir haut, il faut creuser profond. »

Proverbe mongole

L'objectif de ce mémoire est d'évaluer l'apport de données télématiques dans la modélisation du risque géographique en assurance automobile. Ce premier chapitre pose les bases nécessaires à la compréhension du contexte et des enjeux de l'étude menée. Il couvrira les thèmes majeurs du sujet à savoir l'assurance automobile, le risque géographique en assurance automobile et la télématique.

1.1 Assurance automobile

1.1.1 Généralités

L'assurance automobile est une branche de l'assurance de biens encore appelée assurance Incendies Accidents et Risques Divers (*IARD*). Cette dernière fait, elle même, partie de la famille des assurances dites Non-vie. L'assurance automobile couvre les dommages matériels et corporels subis ou causés involontairement par un assuré à autrui avec son véhicule.

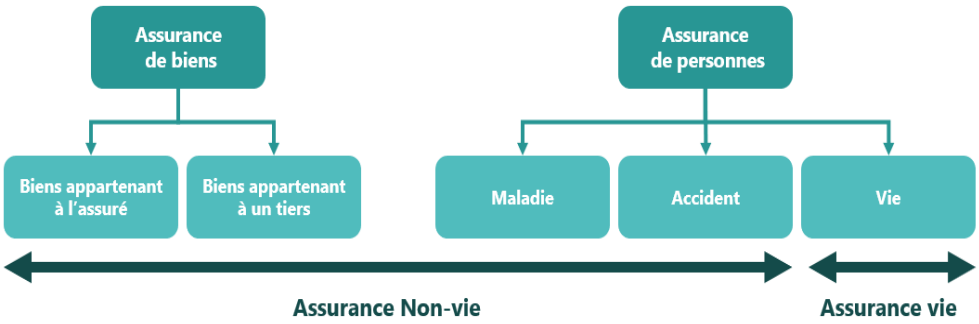

FIGURE 1.1 – *Les familles d'assurance*

En assurance automobile, les contrats ont une durée d'un an avec la plupart du temps une clause de tacite reconduction. Depuis 2015, avec la mise en application de la loi Hamon, les assurés ont le droit de résilier leur contrat à n'importe quelle période et sans obligation de justification auprès de leur assureur, à partir du premier anniversaire de leur adhésion.
A l'intérieur d'un contrat d'assurance automobile peuvent figurer trois grandes catégories de garanties : les garanties de responsabilité civile, les garanties de dommages aux biens et les garanties complémentaires.

1.1.2 Garanties en assurance automobile

a) **Les garanties de responsabilité civile** :

> *« Tout fait quelconque de l'homme, qui cause à autrui un dommage, oblige celui par la faute duquel il est arrivé, à le réparer. »* Article 1240 du code civil français.

Les garanties de responsabilité civile se fondent sur l'obligation de réparer un dommage **causé** à autrui. Elles sont de ce fait obligatoires depuis 1958 dans un contrat d'assurance automobile. Lorsque les dégâts sont causés sur le véhicule d'autrui, c'est la garantie de responsabilité civile matérielle qui intervient. Dans le cas où les dommages sont physiquement subis par l'autre partie, c'est la garantie

de responsabilité civile corporelle qui agit. Une personne qui ne possède que les garanties de responsabilité civile dans son contrat d'assurance automobile est dite assurée « au tiers ».

b) **Les garanties de dommages aux biens** :

Il existe plusieurs garanties de dommages aux biens. Elles couvrent les préjudices matériels **subis** par l'assuré. Elles sont facultatives, mais très fréquemment associées aux garanties de responsabilités civiles dans les contrats. Parmi celles-ci se trouvent par exemple :

- **La garantie dommages tous accidents** : L'assureur s'engage à couvrir tous les sinistres subis par le véhicule de l'assuré, lorsqu'ils résultent d'une collision avec un corps fixe ou mobile.
- **La garantie bris de glace** : L'assureur garantit les dommages liés aux parties vitrées du véhicule de l'assuré (pare-brises, vitres latérales,...).
- **La garantie vol** : Elle intervient en cas de vol du véhicule ou de pièces du véhicule ou d'effets personnels de l'assuré présents dans son véhicule. Si le véhicule est retrouvé dans le mois suivant le vol, l'assureur prend en charge les frais de remise en état dans la limite de sa valeur avant sinistre. Sinon une indemnité contractuelle est due. Les conditions tarifaires sont fonction des mesures de protection mises en place par l'assuré.
- **La garantie incendie** : L'assureur endosse les frais dus à l'incendie du véhicule, que cet incendie se soit déclenché à l'intérieur ou à l'extérieur de celui-ci. Les incendies provoqués par tous types de cigarettes sont exclus du contrat.

c) **Les garanties complémentaires** :

En complément de toutes ces garanties déjà mentionnées, l'assureur peut en proposer bien d'autres, afin, d'améliorer sa couverture du risque. Elles peuvent être de divers ordres et ne sont pas toujours étroitement liées au véhicule. En guise d'exemple, sont citées :

- **La garantie catastrophes naturelles** : L'assureur garantit les dégâts engendrés par des catastrophes naturelles sous réserve de parution au journal officiel de l'arrêté interministériel constatant l'état de catastrophe naturelle.
- **La garantie protection juridique** : L'assureur couvre les frais de défense de l'assuré devant les tribunaux.

1.1.3 Quelques chiffres clés du marché en 2020

La crise sanitaire engendrée par la pandémie de la Covid-19 a impacté les activités des différents secteurs économiques. Concernant le secteur de l'automobile, les mesures de confinements imposées par le gouvernements français ont occasionné des baisses du trafic routier comme le montre la figure 1.2.

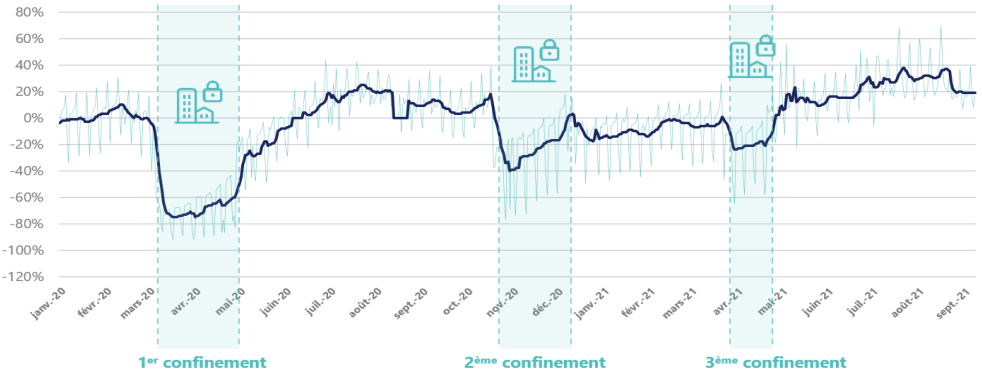

FIGURE 1.2 – *Évolution du trafic routier en France entre Janvier 2020 et Septembre 2021*
Source de la donnée : Cerema - Indicateur du trafic routier

De cet impact direct a découlé des répercutions indirectes sur le marché de l'assurance automobile. Cette section présentera donc les variations de différents indicateurs sur ce marché dans le contexte atypique de l'année 2020 (*France Assureurs* cf.[3]).

a) **Le risque assuré** :
Au cours de l'année 2020, le parc de véhicules assurés de première catégorie [1] (hors flottes) a augmenté de +1,2% succédant à une hausse de +1,1% en 2019. Cette croissance du parc assuré reflétant l'intérêt des particuliers dans la protection de leurs véhicules, peut être l'une des causes du vieillissement du parc automobile français. En effet, en France, l'âge moyen des véhicules est passé de 9 à 10,6 ans entre 2017 et 2020 [2]. Cela signifie que les particuliers utilisent beaucoup plus longtemps leurs véhicules ou qu'ils préfèrent acheter des véhicules d'occasion. Dans les deux cas, cela implique que les voitures fonctionnent sur de plus longues durées et sont donc mieux protégées.

b) **Les cotisations** :
Les cotisations sur le marché de l'assurance automobile français conservent leur tendance à la hausse. Après une croissance de 3,1% en 2019 elles augmentent encore de 3% en 2020. La prime moyenne globale d'un véhicule de première catégorie assuré s'établit à 423 euros en 2020, progressant ainsi de 1,3% par rapport à 2019. Grâce à ces chiffres, cette branche conserve son statut d'incontournable, représentant 39% de la famille des assurances de biens.

c) **La sinistralité** :
L'année 2020 a vu le niveau le plus bas du ratio combiné comptable net de réassurance jamais observé depuis 1995 sur le marché de l'assurance automobile. Le ratio combiné est le rapport de la somme des charges de sinistres et des frais de gestion décaissés par le montant total de primes encaissées. Ce ratio s'établit à 94,7% en 2020, symbolisant une amélioration de 7,3 points par rapport à 2019.

1. Les Véhicules de première catégorie sont des voitures particulières ou des véhicules utilitaires.
2. https ://www.bfmtv.com/auto/le-parc-automobile-francais-vieillit_AN-202002140008.html

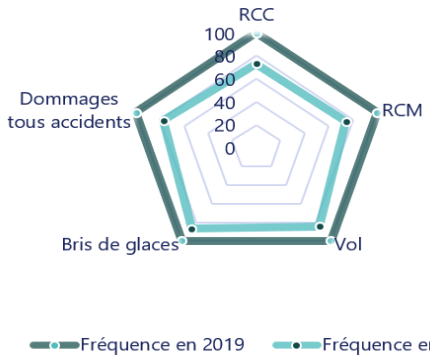

FIGURE 1.3 – *Niveaux de fréquence de sinistres par garanties et suivant l'année (Base 100 2019)*

FIGURE 1.4 – *Niveaux de coûts moyens par garanties et suivant l'année (Base 100 2019)*

La figure 1.3 montre une baisse des fréquences de sinistres toutes garanties confondues entre 2019 et 2020. Cette baisse de sinistralité en 2020 peut s'expliquer par les chutes de trafic routier occasionnées par les périodes de confinement sur cette année (voir Figure 1.2). Inversement, La figure 1.4 montre une augmentation des coûts moyens sur le marché de l'assurance automobile suivant toutes les garanties entre 2019 et 2020. Cette hausse des coûts endossés par les assureurs est due entre autres à l'augmentation constante des prix des pièces détachées des voitures.

En somme, malgré le contexte social particulier qui a marqué l'année 2020, l'assurance automobile continue sa progression. Cette progression est sans aucun doute le fruit de la concurrence accrue entre les acteurs de ce marché. Ces derniers s'attelant à toujours proposer des services de meilleures qualités, basés sur des méthodes toujours plus innovantes.

1.2 Risque géographique en assurance automobile

Le risque est la possibilité de survenance d'un évènement préjudiciable. Lorsqu'une personne (physique ou morale) se fait assurer, elle cède le risque inhérent au motif assuré à son assureur. On parle alors de transfert de risque de la cédante : la personne ayant souscrit au contrat d'assurance, vers le preneur de risque : l'assureur. En vue de maintenir la rentabilité de son activité, ce dernier se doit d'encadrer par différentes méthodes cette prise de risque. Et l'une des méthodes les plus utilisées pour y parvenir est la segmentation.

1.2.1 Segmentation : définition, importance et limites

Définition de la segmentation

La segmentation consiste pour un assureur à partitionner son portefeuille d'assurés en classes de risques homogènes. Cela revient à regrouper les individus présentant des profils de risque proches. Au sein de chaque classe, les assurés paient pour le risque moyen des individus qui la composent. Dans l'optique de segmenter son portefeuille, l'assureur s'appuie sur différents critères. Les critères classiques utilisés proviennent des informations collectées à la souscription du contrat. En assurance automobile, l'âge du conducteur peut par exemple être un critère significatif de segmentation du risque. En effet, il est courant d'observer une fréquence de sinistres plus élevée chez les conducteurs dont l'âge est supérieur à 65 ans que chez ceux ayant un âge compris entre 40 et 65 ans. De ce fait, le risque moyen diffère entre ces deux classes d'âges et cette différence doit se ressentir dans les tarifs proposés.

Importance de la segmentation

La segmentation permet à l'assureur de contrôler un facteur important dans ses prises de risques à savoir l'antisélection. L'antisélection est la résultante d'une asymétrie d'informations entre la cédante et le preneur de risque lors de la souscription du contrat. Supposons par exemple deux assureurs commercialisant des contrats d'assurance automobile :

- l'assureur A décide de ne pas segmenter son portefeuille,
- l'assureur B quant à lui, décide de segmenter son portefeuille suivant l'âge du conducteur.

Supposons que les assureurs A et B proposent des contrats ayant les mêmes garanties.

	Assurés de 40 à 65 ans	Assurés de +65 ans
Assureur A	50	50
Assureur B	30	70

FIGURE 1.5 – *Montants de primes proposés par chacun des assureurs*

	Assurés de 40 à 65 ans	Assurés de +65 ans
Montants de sinistres	20	60

FIGURE 1.6 – *Sinistralité observée au cours de l'exercice*

L'assureur A, proposant la même prime sans distinction de la classe d'âge, fait des profits sur les conducteurs dont l'âge est compris entre 40 et 65 ans (bons risques) qu'il utilisera pour couvrir les pertes endossées sur les conducteurs âgés de plus de 65 ans (mauvais risques). Ainsi, dans ce type de fonctionnement, les « bons risques » paient pour les « mauvais risques ».

L'assureur B, en décidant de segmenter son portefeuille suivant l'âge du conducteur, adapte son tarif au niveau de risque sur les différentes classes d'âges. En effet, les mauvais risques paient une prime plus élevée du fait de leur forte probabilité d'avoir un sinistre au cours de l'exercice, et les bons risques paient une moindre prime.

En fin d'exercice, des migrations seront observées sur ce marché concurrentiel simulé. Ces migrations sont les conséquences de la prise en compte de la segmentation dans la construction des tarifs des assureurs. En effet, les « bons risques » décideront de se faire assurer par l'assureur B leur proposant des primes plus faibles, et les « mauvais risques » iront se faire assurer chez l'assureur A afin de payer moins chers. Cette situation est dommageable pour la rentabilité de l'assureur A. Ce dernier, en décidant de ne pas segmenter le risque sur son portefeuille, se prive d'informations importantes et se retrouve à assurer en majorité des « mauvais risques ».

Limites de la segmentation

Le précédent exemple, certes minimaliste, prouve la nécessité pour un assureur de segmenter le risque de manière homogène sur son portefeuille. Toutefois, la segmentation présente certaines limites à ne pas franchir.

- **Des contrats d'assurance unisexe** :
  « La prise en compte du sexe de l'assuré en tant que facteur de risque dans les contrats d'assurance constitue une discrimination. » Extrait de La *Gender Directive* (2004/113/CE)

  Le 21 décembre 2012, entrait en vigueur un arrêt de la Cour de justice de l'Union Européenne qui interdit la distinction homme-femme dans les tarifs d'assurance (*Actuariel* cf.[1]). De ce fait, toute utilisation de critères basés sur le sexe de l'assuré dans une optique de segmentation du risque est prohibée à tous les assureurs en vue de conserver l'égalité des genres sur le marché de l'assurance.

- **D'un tarif mutualisé à un tarif individualisé** :
  Lorsque la segmentation est poussée à l'extrême, atteignant des niveaux de finesse très excessifs, le tarif proposé tend à devenir individuel, brisant le principe de mutualisation du risque (*F. PLANCHET* cf.[16]). Cette hyper segmentation est susceptible de causer un problème d'éthique. En effet, certains assurés ayant subi ou causé des accidents de la circulation seront contraints à renoncer à s'assurer du fait des montants trop élevés de leur prime.

Il convient donc pour les assureurs de rester pragmatiques dans la segmentation du risque sur leur portefeuille, en n'utilisant que des critères à la fois statistiquement pertinents et légalement convenables.

1.2.2 Segmentation par des critères géospatiaux : Une vision géographique du risque en assurance automobile

En assurance IARD, ou plus précisément en assurance automobile, l'un des aspects impactant le plus la sinistralité est l'aspect géographique du portefeuille. En effet, la sinistralité n'est pas uniforme sur tout le territoire. Elle évolue suivant la zone de résidence, de stationnement ou de circulation et cela pour diverses raisons. Ces raisons peuvent être entre autres : la densité de la population, le niveau de vie, le taux de chômage, la pluviométrie et le niveau du trafic routier.

Prenons pour exemple, deux jeunes étudiants de 21 ans ayant les mêmes véhicules, l'un habitant à Marseille et l'autre à Plougastel. Un assureur qui n'axerait la segmentation pour une garantie de responsabilité civile que sur la classe d'âge, le type de véhicule et la catégorie socio-professionnelle, proposerait le même tarif à ces deux profils. Pourtant, compte tenu de la différence de densité du trafic routier à Marseille et à Plougastel, il est clair que ces deux profils ont des niveaux de risque très différents. Il serait donc préjudiciable pour cet assureur de ne pas tenir compte de ce facteur susmentionné dans sa modélisation tarifaire.

Il existe une multitude de critères géospatiaux pouvant être utilisés selon les besoins des garanties à tarifer. Ces critères apportent des informations à différentes échelles de zones géographiques appelées mailles géographiques. Les mailles les plus communes sont : la région, le département et la maille communale encore appelée maille INSEE [3]. Il en existe bien d'autres encore plus fines.

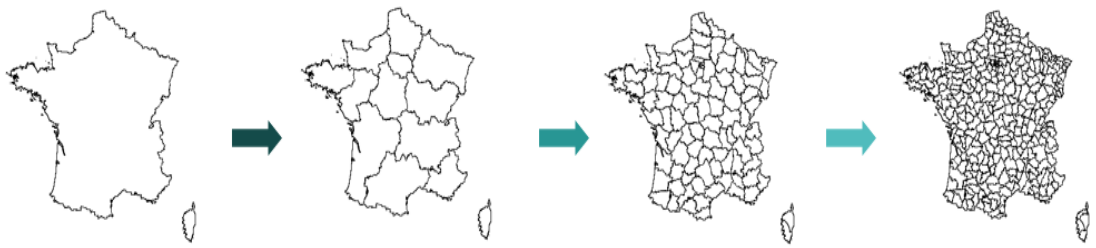

FIGURE 1.7 – *Exemples de mailles géographiques*

L'assureur ne pouvant utiliser n'importe quelle variable géographique dans sa modélisation, il se doit de n'en sélectionner que celles qui sont les plus significatives et qui améliorent la qualité de son modèle. Le traitement et l'intégration de ces informations

3. Institut National de la Statistique et des Etudes Economiques, service public français chargé de la production et de l'analyse des différentes données statistiques concernant les collectivités, la géographie, les populations et les entreprises.

géographiques dans la structure de segmentation du risque est communément appelée la « modélisation du risque géographique ».

1.3 A la découverte de la télématique

1.3.1 Définition de la télématique

Selon le dictionnaire Larousse, la télématique est « l'ensemble des techniques et des services qui associent les télécommunications et l'informatique ». La télécommunication désigne le partage d'informations par le biais de moyens technologiques et l'informatique est la science afférente à la gestion, au traitement et à l'analyse de l'information via des ordinateurs. La télématique consiste donc à échanger des informations entre des systèmes et des ordinateurs grâce à des appareils connectés. Le mot télématique a été officiellement introduit dans la littérature pour la première fois en décembre 1977, après la publication du rapport Nora-Minc. Ce rapport résume une étude visant à faire progresser la réflexion sur les moyens de conduire l'informatisation de la société.
La télématique se retrouve aujourd'hui dans différents domaines comme la santé et le bâtiment. Toutefois, le domaine le plus avancé sur le sujet reste celui de l'automobile.

1.3.2 Télématique embarquée dans le milieu de l'automobile

Dans le milieu de l'automobile, la télématique s'effectue généralement [4] par le biais d'un boîtier appelé *On Board Diagnosis* (OBD). C'est un système embarqué qui contient principalement un récepteur GPS, complété très souvent par d'autres dispositifs comme un accéléromètre, un modem et une interface du moteur. Ce boîtier extrait plusieurs informations générées par le véhicule comme la position GPS, la vitesse, la "force G" c'est à dire toutes les mesures relatives à l'accélération et les états de différents composants du véhicule. Ces données sont ensuite transférées vers des systèmes de stockage.

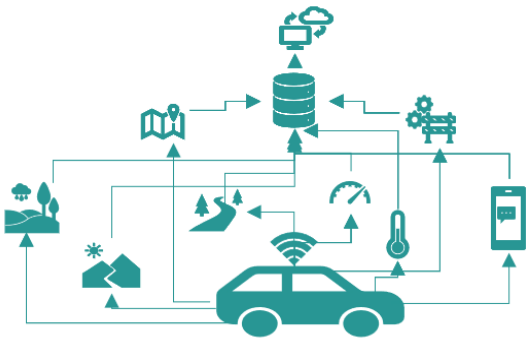

FIGURE 1.8 – *Illustration shématique du fonctionnement de la télématique automobile*

4. Il existe de nouveaux types de systèmes télématiques soit basés sur le smartphone du conducteur ou soit nativement connectés au véhicule.

Cette grande variété d'informations collectées puis traitées, devient par la suite une base de données dans laquelle se trouvent par exemple des variables sur les différentes positions GPS occupées par le véhicule dans le temps, les distances et les temps de trajet parcourus, les vitesses du véhicule, les temps de freinage et de ralenti, la consommation de carburant, les pannes du véhicule, la tension de la batterie...

Ces informations peuvent être transmises en temps réel au conducteur par le biais d'un logiciel sous forme d'indicateurs ou de services d'aide à la conduite (analyse de la conduite en temps réel, prévention sur le comportement au volant, suggestion d'itinéraire, indicateurs sur l'état du moteur et d'autres composantes du véhicule, indicateurs sur la consommation journalière de carburant ...). En outre, elles peuvent faire l'objet d'analyses statistiques à diverses finalités comme le pilotage d'un business de flottes automobiles ou encore la conception de produits et de services en assurance automobile.

1.3.3 RGPD et utilisation éthique des données télématiques

La protection des personnes physiques à l'égard du traitement de leurs données personnelles est un sujet majeur à l'ère du *Big data*[5] et de la digitalisation. Le 14 avril 2016, le parlement européen a donc adopté le Règlement Général sur la Protection de Données (RGPD) afin d'aider les citoyens européens à protéger leur identité numérique. En France la structure chargée de veiller à l'application du RGPD est la Commission Nationale de l'Informatique et des Libertés (CNIL).

D'après l'article 4 de cette réglementation, la donnée à caractère personnelle se définit comme « *toute information se rapportant à une personne physique identifiée ou identifiable,[...]* ». Une personne physique est dite identifiable si elle peut être identifiée directement ou indirectement, par le biais d'identifiants propres à son identité physique, physiologique, génétique, psychique, économique, socio-culturelle, géographique ou encore numérique.

D'une part, la mise en place de cette réglementation renforce les droits des citoyens quant à l'utilisation de leurs données personnelles par les entreprises sur trois points.

- **Le consentement** : Tout entreprise souhaitant utiliser les données personnelles d'un particulier pour diverses raisons, est dans l'obligation de lui demander son accord avant de débuter l'exploitation.
- **La portabilité** : Les données doivent conserver un aspect portable de sorte à ce que les particuliers puissent à tout instant les transférer à d'autres acteurs du marché.
- **Le droit à l'oubli** : Le particulier a le pouvoir d'exiger à tout moment la suppression d'une partie ou de l'entièreté de ses données personnelles exploitées.

5. Quantité massive de données à disposition

D'autre part, elle durcit les obligations des entreprises dans leur exploitation des données personnelles sous le signe de la transparence, du contrôle et de la responsabilisation. Cela implique pour les entreprises d'utiliser le moins de données personnelles possible dans la gestion de leur service et de veiller à maintenir une anonymisation et une sécurité rigoureuse de celles-ci. En cas de manquement à ces devoirs, les entreprises peuvent encourir des sanctions allant du simple blâme à l'arrêt de l'exploitation des données et à une amende.

Les données télématiques recueillies dans les véhicules constituent des données à caractère personnel puisqu'elles sont révélatrices des habitudes de conduite et de déplacement des conducteurs. Ainsi leur utilisation est aussi soumise au RGPD. Toute entreprise (constructeurs automobile, gestionnaires de flottes, fournisseurs de service, assureurs, ...) souhaitant exploiter ces types de données se doit donc d'observer toutes les mesures succinctement présentées plus haut.

1.3.4 Quand la télématique s'invite en assurance automobile

L'essor de la télématique dans le milieu de l'automobile a favorisé la création d'un nouveau type de produit d'assurance appelé « assurance télématique ». L'assurance télématique est une des sous-branches de la famille des *Usage-Based Insurance* (UBI) en assurance automobile. Un contrat d'assurance UBI est un contrat dont la prime dépend, en plus des critères traditionnels, du niveau d'utilisation que l'assuré fait de son véhicule. Le marché des assurances UBI se décline sous trois différentes formes :

— **L'assurance fondée sur l'auto-déclaration (*self-reporting based*)** :
La prime payée est calculée en fonction du kilométrage annuel déclaré par l'assuré. Une remise est appliquée au tarif si la distance parcourue est inférieure à un seuil fixé. Ce type de contrat est moins répandu sur le marché car il implique pour l'assureur de se fier uniquement aux dires de l'assuré.

— **L'assurance au kilomètre ou Pay-As-You-Drive (PAYD)** :
Dans le cadre d'un contrat d'assurance PAYD, les données sur le kilométrage parcouru sont calculées à partir d'un odomètre. Un odomètre est un dispositif qui ne fait que compter le nombre de kilomètres parcourus par un véhicule. Ainsi, l'assureur peut vérifier en fin d'exercice le nombre de kilomètres parcourus déclarés par l'assuré avant d'ajuster son tarif.

— **L'assurance fondée sur le comportement ou Pay-How-You-Drive (PHYD)** :
L'assurance PHYD est la branche à laquelle appartient l'assurance télématique. Dans ce type de contrat, un boîtier OBD est placé sur le véhicule de l'assuré de sorte à ce que l'assureur récolte en temps réel les données nécessaires à ses modifications tarifaires. Ici le calcul de la prime se base non seulement sur le kilométrage parcouru par l'assuré, mais aussi sur de nouvelles variables de conduite enregistrées par le boîtier comme les comportements au volant, les allures de conduite ou de

freinage...

Le PHYD est aujourd'hui la forme la plus complète des assurances basées sur le niveau d'utilisation que l'assuré fait de son véhicule. Néanmoins, il est aussi le plus difficile à mettre en production.

En effet, dans ce format de produit, les données télématiques sont directement reliées aux assurés puisqu'elles sont collectées sur leurs véhicules. Elles font donc partie des **données internes** de l'assureur. Cette manière de fonctionner est entravée par plusieurs facteurs tels que : les coûts élevés des dispositifs OBD et de leur installation sur les véhicules de chacun des assurés, la complexité de l'exploitation en temps réel des données collectées, la réticence des individus à souscrire à ce nouveau type de produit d'assurance soit par peur de l'utilisation abusive de ces données sensibles soit à cause de préjugés considérant ce nouveau produit comme un effet de mode qui finalement ne modifie pas vraiment la prime à payer...

En outre, les données télématiques classiquement utilisées dans les contrats d'assurance PHYD sont en général démunies d'un aspect qui peut dans certains cas se révéler très important à savoir : **l'aspect contextuel**. En effet, la mise en contexte des données télématiques permet d'affiner les information apportées par celles-ci et par conséquent d'améliorer la connaissance générale des habitudes de conduite des individus.

1.4 Enjeux du mémoire

Ce mémoire s'inscrit dans le cadre de la mise en place d'un partenariat entre le cabinet de conseil en actuariat ADDACTIS France et le groupe leader mondial de la fabrication et de la commercialisation de pneumatique MICHELIN. Le but de ce partenariat est de développer une offre de service visant à aider les acteurs du marché de l'assurance automobile dans leur appréhension du risque.

Dans ce mémoire, les travaux portent sur l'évaluation de l'apport de données télématiques dans la modélisation du risque géographique lors de la tarification des produits d'assurance automobile. L'objectif final est donc d'évaluer l'impact des habitudes de conduite à l'intérieur des différentes zones géographiques sur le risque en assurance automobile et sur l'amélioration de la qualité des modèles dans ce domaine.

Cette évaluation s'effectuera sur la garantie responsabilité civile d'un assureur français. Tout au long des travaux, interagissent les données internes de l'assureur et les données externes fournies par MICHELIN. Ici, l'usage des données télématiques dans la modélisation actuarielle diffère de celui qui est fait classiquement en assurance. En effet, non seulement les données télématiques sont utilisées en tant que données externes puisqu'elles sont reliées aux informations des assurés par le biais d'une clé de jointure

géographique, mais aussi, elles ont l'avantage pour la majeur partie d'entre elles d'être **contextualisées** de sorte à étudier l'évolution du risque suivant différentes situations.

Chapitre 2

Descriptions statistiques des sources de données de l'étude

« Au commencement était...la donnée ? »

Parodie du premier verset du livre de Jean

Dans l'optique d'évaluer l'apport de la télématique dans la modélisation du risque géographique en assurance automobile, deux types de sources de données sont utilisées : des données internes et des données externes. Les données internes sont l'ensemble des informations recueillies par l'assureur au moment de la souscription du contrat. Les données externes, quant à elles, sont des variables exogènes au processus de souscription, mais reliables aux données internes par le biais d'une ou plusieurs clés de jointure.

Dans le cadre de ce mémoire, la donnée interne est le portefeuille d'un assureur français et la principale base de données externes est la base ***Smart Road Data***. Cette dernière contient à la fois des variables télématiques et des données sur les types de routes à la maille commune.

L'objectif de ce chapitre est de présenter ces bases : interne puis externe. Cette présentation s'articulera autour de la description, des analyses statistiques et des retraitements effectués sur celles-ci.

2.1 Présentation de la base interne : données d'un assureur

La base interne utilisée pour les travaux de ce mémoire est le portefeuille d'un assureur français sur trois années d'exercice 2017-2019.
L'étude sur ce portefeuille se focalisera sur les sinistres attritionnels de la garantie responsabilité civile matérielle (RCM). Un sinistre est dit attritionnel si le coût qu'il engendre n'est pas considéré comme grave. D'une part, le choix de la garantie responsabilité civile est dû au fait qu'elle regroupe tous les assurés du portefeuille puisqu'elle est obligatoire. D'autre part, la restriction aux sinistres attritionnels permet de concentrer les travaux sur le cœur du sujet à savoir l'évaluation de l'apport des données télématiques dans la modélisation du risque géographique en assurance automobile.

2.1.1 Description des variables de la base interne

Les variables de la base interne proviennent des questionnaires de souscription de l'assureur. Elles peuvent être regroupées en trois grandes catégories :

- **Les variables concernant le profil de l'assuré** : Ce sont les caractéristiques en lien direct avec l'assuré tels que son âge, sa catégorie socio-professionnelle, le nombre d'années d'ancienneté de son permis de conduire...

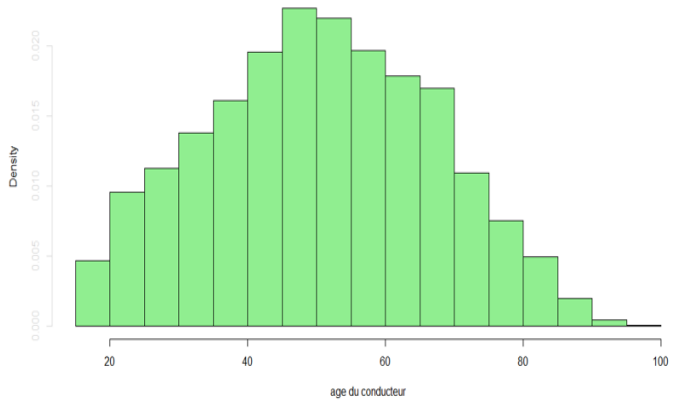

FIGURE 2.1 – *Histogramme des âges des conducteurs sur le portefeuille*

La figure 2.1 montre qu'une grande partie des conducteurs a un âge compris entre 40 et 60 ans. Toutefois, la répartition des âges semble assez hétérogène. Cette hétérogénéité est positive pour l'assureur puisqu'elle implique une connaissance du risque en automobile sur pratiquement toutes les classes d'âges.

- **Les variables concernant le véhicule de l'assuré** : Ce sont les spécificités du véhicule de l'assuré. Elles sont entre autres l'âge du véhicule, la marque, le modèle, la puissance ou le nombre de places assises à l'intérieur de celui-ci.

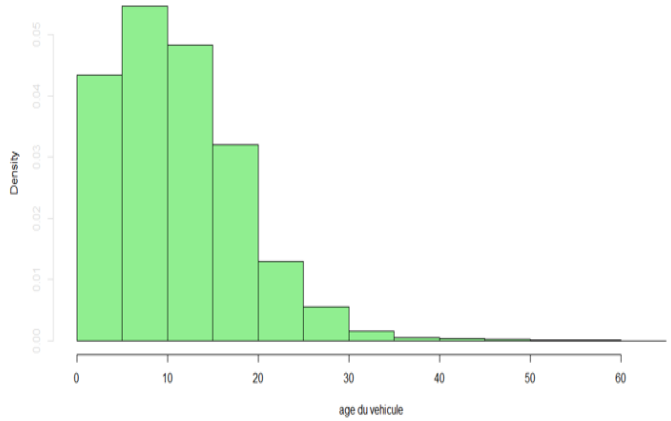

FIGURE 2.2 – *Histogramme des âges des véhicules sur le portefeuille*

La figure 2.2 montre que la distribution des âges de véhicules sur le portefeuille est assez asymétrique. Cette asymétrie résulte du fait que la majeur partie de ces véhicules a un âge compris entre 0 et 30 ans et une infirme partie de ces véhicules a un âge supérieur à 30 ans. Entre 2017 et 2020, l'âge moyen des véhicules du parc automobile français avoisinait 10 ans. Cette tendance nationale se retrouve bien sur le portefeuille de l'assureur.

- **Les variables géographiques** : elles renseignent sur le lieu de résidence de l'assuré. Ce sont des variables telles que le code département et le code IRIS. Le code IRIS est un découpage du territoire plus fin que le code INSEE. les IRIS couvrent généralement entre 1800 et 5000 habitants et leurs limites s'appuient sur les coupures du tissu urbain.

2.1.2 Analyse temporelle du portefeuille

L'analyse temporelle d'un portefeuille consiste à en observer l'évolution au fil du temps suivant des indicateurs de tarification tels que l'exposition, la fréquence de sinistres et les coûts moyens. Cela permet à l'assureur d'ajuster sa politique tarifaire au fil du temps et de piloter ses prises de risque en vue de conserver sa rentabilité et d'anticiper de potentielles dérives.

Définitions

Exposition :
L'exposition d'une police est le ratio de la durée réelle pendant laquelle elle a été couverte par la durée définie à la signature du contrat. Par exemple, pour les contrats d'assurance automobile (durée d'un an), la formule de l'exposition peut s'écrire :

$$Exposition = \frac{Nombre\ de\ mois\ de\ couverture}{12}$$

Fréquence de sinistres :
C'est la probabilité de survenance d'un sinistre pendant la durée de vie du contrat. Elle se calcule de la manière suivante :

$$Fréquence = \frac{Nombre\ de\ sinistres}{Exposition}$$

Coût moyen de sinistres :
C'est le montant décaissé en moyenne par l'assureur afin de couvrir un sinistre. Il s'obtient en faisant :

$$Coût\ moyen = \frac{Coût\ de\ sinistres}{Nombre\ de\ sinistres}$$

Analyses Temporelles

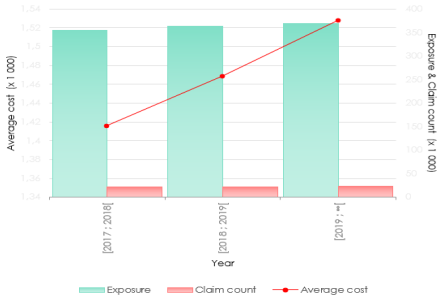

FIGURE 2.3 – *Évolution de l'exposition et des coûts moyens de la garantie RCM sur le portefeuille suivant l'année (2017 à 2019)*

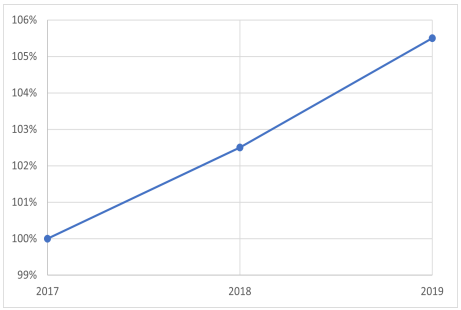

FIGURE 2.4 – *Évolution des coûts moyens de la garantie RCM sur le marché de l'assurance automobile en base 100 de 2017*
source : Données clés de l'assurance française en 2019 FFA

Sur la figure 2.3, les barres vertes représentent les totaux d'exposition sur les différentes années d'exercice. Au cours de ces trois années le total d'exposition a évolué en moyenne d'environ 4% . Cela dénote une quasi stabilité temporelle des prises de risque de l'assureur.

Sur la même figure, la courbe de couleur rouge présente l'évolution des coûts moyens endossés par l'assureur suivant chacune des années au titre de la garantie responsabilité civile (RCM). La pente positive de cette courbe signifie l'accroissement de ces coûts moyens au fil des années. La courbe de couleur bleue de la figure 2.4 montre que l'évolution des coûts moyens de la garantie RCM sur le marché de l'assurance automobile français suit ces mêmes tendances à la hausse. Ainsi, L'évolution des coûts moyens de l'assureur respectent les tendances du marché. Ces accroissements des coûts moyens de sinistres sur le marché peuvent s'expliquer entre autres par l'augmentation annuelle des prix des pièces de rechange des véhicules.

Cependant, d'après les figures 2.5 et 2.6 la fréquence de sinistres sur le portefeuille de l'assureur semble suivre une tendance différente de celle du marché sur l'année 2019.

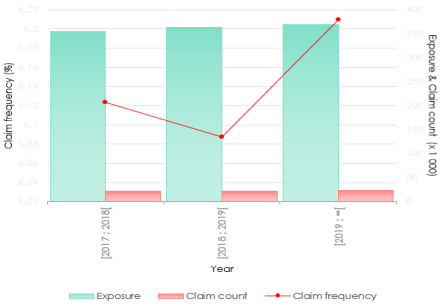

FIGURE 2.5 – *Évolution de l'exposition et des fréquences de sinistres de la garantie RCM sur le portefeuille suivant l'année (2017 à 2019)*

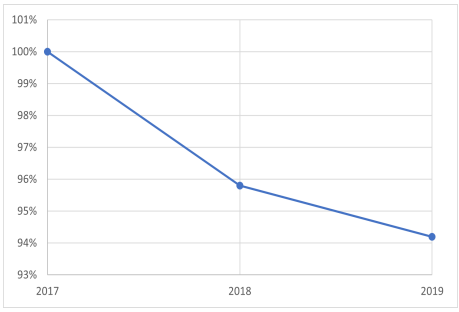

FIGURE 2.6 – *Évolution des fréquences de sinistres de la garantie RCM sur le marché de l'assurance automobile en base 100 de 2017 (source : Données clés de l'assurance française en 2019 FFA)*

Cela peut être la conséquence de différentes causes. Lors d'une refonte de son système tarifaire, l'assureur se doit de rechercher ces causes et d'y prêter une grande attention.

2.2 Présentation de la base externe : *Smart Road Data*

Les données de la base *Smart Road Data* font partie d'une offre de service du groupe **MICHELIN**, leader de la fabrication et de la commercialisation de pneumatiques dans le monde. C'est un tableau de **30 993 lignes** représentant chacune une commune et **58 colonnes** dont les noms de ces communes, leurs codes de département et des données agrégées sur celles-ci. Ces dernières colonnes sont toutes de type continue et peuvent être regroupées en 4 grandes catégories :

- Les scores de comportement ;
- Les variables concernant l'usage du véhicule ;
- Les variables de saisonnalité ;
- Les variables cartographiques.

Ces données sont récoltées sur une communauté de conducteurs créée par le groupe et portant le nom de *Better Driving Community*. Cette communauté contient plus de 200 000 conducteurs ayant différents types de véhicules et parcourant plus de 1.5 milliards de kilomètres par an. La base fournit des informations sur environ 89% des communes du territoire Français. Cependant, il peut arriver que dans certaines variables, il y ait des valeurs manquantes sur quelques communes présentes dans cette base. Cela représente pour la majeur partie des variables concernées moins de 1% des communes à l'exception des variables de saisonnalité. Ces dernières contiennent toutes environ 15% de valeurs manquantes.
Au cours de l'étude, lorsque le besoin d'avoir une information exhaustive sur le territoire s'est fait sentir, des stratégies ont été mises en place afin d'imputer des valeurs aux communes dont les informations n'étaient pas disponibles.

2.2.1 Les scores de comportement

Ces variables émanent du *scoring* de la conduite des automobilistes. L'idée est de connaître les différents comportements qu'ils adoptent au volant selon la commune par laquelle ils passent. Les scores à disposition se focalisent sur la manière de freiner (intensité, fréquence, distance, freinage brusque) et l'allure de la conduite selon le type de routes (lignes droites ou virages). Leurs interprétations respectent la logique des systèmes de notation classique. En effet, plus un score est élevé, meilleure est la manière de conduire. Ces variables sont organisées en trois niveaux : (voir Figure 2.7)

FIGURE 2.7 – *Les différents niveaux de scores de comportement*

Au premier niveau se trouvent les scores S2. Ce sont les scores les plus détaillés selon le phénomène qu'ils sont censés modéliser (manière de freiner / allure de conduite). Par exemple la variable *S2_anticipation_brakingIntensity* évalue spécifiquement l'intensité de freinage des conducteurs sur les différentes communes. Au second niveau, se situent les scores S3. Ce sont des combinaisons de scores S2. Ils synthétisent toutes les informations apportées par ces derniers. Au troisième et dernier niveau se positionne le score global, qui est une combinaison des scores S3. Il représente donc une note générale attribuée aux conducteurs sur une commune. Toutefois, l'analyse de cette dernière variable peut être difficile. Cela est dû au fait qu'elle est la combinaison de variables qui elles mêmes sont le fruit de combinaisons d'autres variables.
Enfin, pour tous ces scores de comportement, il convient de souligner que ni les formules permettant de les construire ni les méthodes utilisées pour les agréger n'ont été fournies. Le tableau décrivant l'ensemble des scores est consultable en Annexe A.

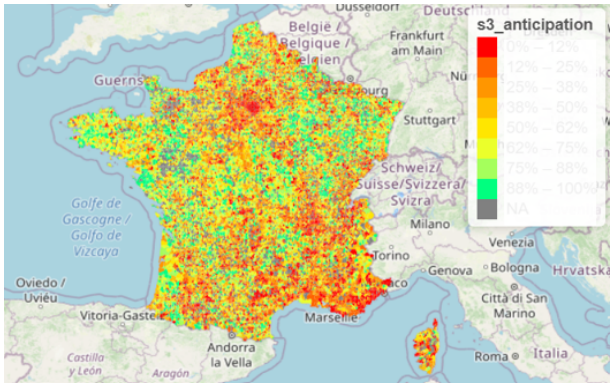

FIGURE 2.8 – *Cartographie du score S3_anticipation*

La variable *S3_anticipation* représente la note générale attribuée aux conducteurs suivant leur façon de freiner sur une commune donnée. Le gradient de couleurs passe du rouge correspondant aux zones ayant des scores très faibles, au vert reflétant celles ayant des scores très élevés. Il apparaît par exemple que dans le bassin parisien, le massif central et les Alpes, la note de freinage est majoritairement faible. Cela dénote une moins bonne manière de freiner de la part des conducteurs dans ces différents secteurs. En outre, la cartographie met en lumière le manque de données sur quelques communes apparaissant en gris.

2.2.2 Les variables concernant l'usage du véhicule

Elles renseignent principalement sur les kilomètres parcourus et les durées totales de trajets (en secondes) dans les communes selon différents contextes :

- général,
- météo (temps sec, temps humide),
- relief (plaine, vallée, montagne),
- urbanité (route urbaine, route extra urbaine, autoroute),
- courbure des routes (grand virage, virage moyen, virage serré, ligne droite).

Le tableau présentant l'ensemble de ces variables est consultable en Annexe A.

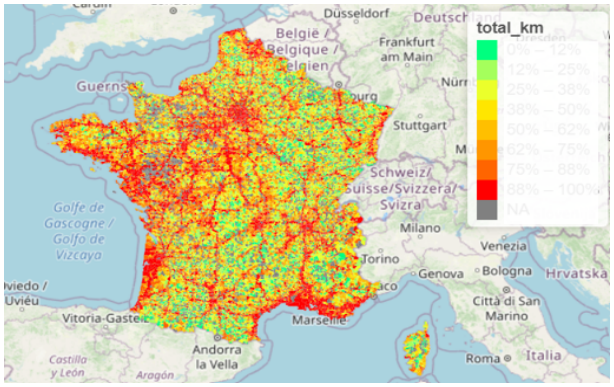

FIGURE 2.9 – *Cartographie du total de kilomètres parcourus à l'intérieur des communes*

Sur la cartographie précédente (Figure 2.9), le gradient de couleur passe du vert, se rapportant aux communes ayant un faible total de kilomètres parcourus, au rouge représentant les communes ayant un total de kilomètres parcourus très élevé. Il est par exemple observable dans des zones comme l'île de France, Lyon, la côte méditerranéenne ou encore le littoral atlantique, que le compteur kilométrique est au plus haut. Il en est de même pour les communes traversées par des autoroutes telles que l'A7 ou encore l'A6.

2.2.3 Les variables de saisonnalité

Elles sont au nombre de douze (12), représentant chacune la différence (en pourcentage) entre la distance moyenne parcourue sur une année et celle parcourue au cours d'un mois sur une commune donnée. Elles permettent d'évaluer le niveau du trafic automobile dans les différentes zones géographiques selon les différentes périodes de l'année.

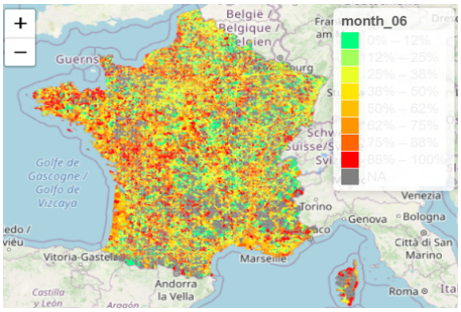

FIGURE 2.10 – *Cartographie du niveau de trafic en Juin*

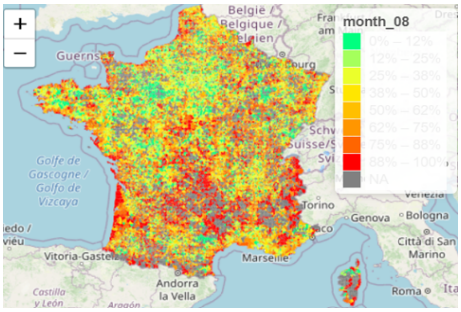

FIGURE 2.11 – *Cartographie du niveau de trafic en Août*

Les deux cartes ci-dessus permettent d'apprécier les niveaux de trafic par rapport à la moyenne annuelle sur les différentes zones en été. Il est par exemple notable que le trafic a tendance à s'intensifier dans le sud de la France durant cette saison. L'une des raisons pouvant expliquer ce phénomène est le tourisme pendant les grandes vacances.

2.2.4 Les variables cartographiques

Cette dernière catégorie regroupe des variables non télématiques. Elles proviennent de la cartographie des routes à l'intérieur des communes. Elles renseignent sur les différents types de route à l'intérieur de ces dernières suivant deux aspects : l'urbanité et la courbure. Ces colonnes sont conservées pour notre étude bien que n'étant pas des données télématiques. En effet, compléter les signaux télématiques de ces variables sur les routes permet de gagner en information sur le risque géographique en assurance automobile. La suite des travaux éclairera quant à la significativité de ces informations complémentaires.
Le tableau présentant ces variables est disponible en Annexe A.

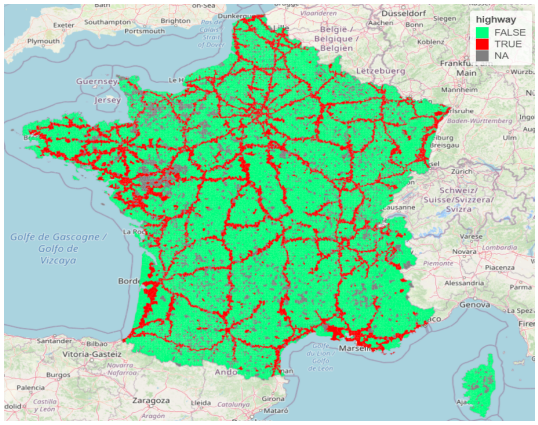

FIGURE 2.12 – *Cartographie des communes traversées par des autoroutes*

La variable *Highway* donne la distance totale d'autoroutes dans les communes. Elle vaudra donc 0 pour les communes n'étant pas traversées par celles-ci. Ainsi, afin de la représenter sur une carte, elle a dû être transformée en une variable binaire avec pour modalités :

- FALSE si elle vaut 0, ce qui signifie que la commune n'est pas traversée par des autoroutes,
- TRUE sinon.

Le résultat est observable sur la cartographie ci-dessus (Figure 2.12).

2.3 Analyse statistique de la base *Smart Road Data*

Pour rappel, dans toute cette section, les variables analysées sont des variables continues. Dans chacune des sous sections, le formalisme mathématique est d'abord présenté (*P. AILLIOT* cf.[2]) puis est suivi d'une application sur la base *Smart Road Data*.

2.3.1 Étude des corrélations entre les variables

L'étude des corrélations permet de résumer les relations de dépendances existantes entre les variables prises deux à deux. Le fait de conserver deux variables fortement corrélées dans une étude peut constituer une redondance inutile d'information. Il faut donc détecter dès le début ces répétitions afin de les diminuer voir les supprimer par diverses méthodes.

Considérons une base X contenant n lignes ($n > 1$) représentant chacune un individu et p colonnes ($p > 3$) représentant chacune une variable. Le coefficient de corrélation empirique $\sigma_{j,k}$ entre deux variables j et k de la base X s'obtient par la formule :

$$\sigma_{j,k} = cor(X_{.,j}, X_{.,k}) = \frac{v_{j,k}}{s_j s_k}$$

Avec $v_{j,k}$ étant la covariance empirique des variables j et k et s'écrivant :

$$v_{j,k} = cov(X_{.,j},X_{.,k}) = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k)$$

(où $\bar{x}_j$ et $\bar{x}_k$ sont les moyennes respectives des colonnes j et k),

Et avec $s_j = \sqrt{v_{j,j}}$ et $s_k = \sqrt{v_{k,k}}$ les écart-types respectifs des colonnes j et k.

Le coefficient $\sigma_{j,k}$ varie entre -1 et 1. Il se rapproche de -1 lorsque les deux variables considérées sont fortement corrélées négativement, ce qui signifie que la croissance de l'une correspond à la décroissance de l'autre. Inversement, il avoisine 1 si les deux variables sont fortement corrélées positivement, c'est à dire qu'elles évoluent simultanément dans le même sens.

Application : Dans la pratique on utilise un graphique appelé corrélogramme afin de lire plus facilement les niveaux de corrélation entre les variables. L'illustration suivante présente le corrélogramme des données de la base *Smart Road Data*.

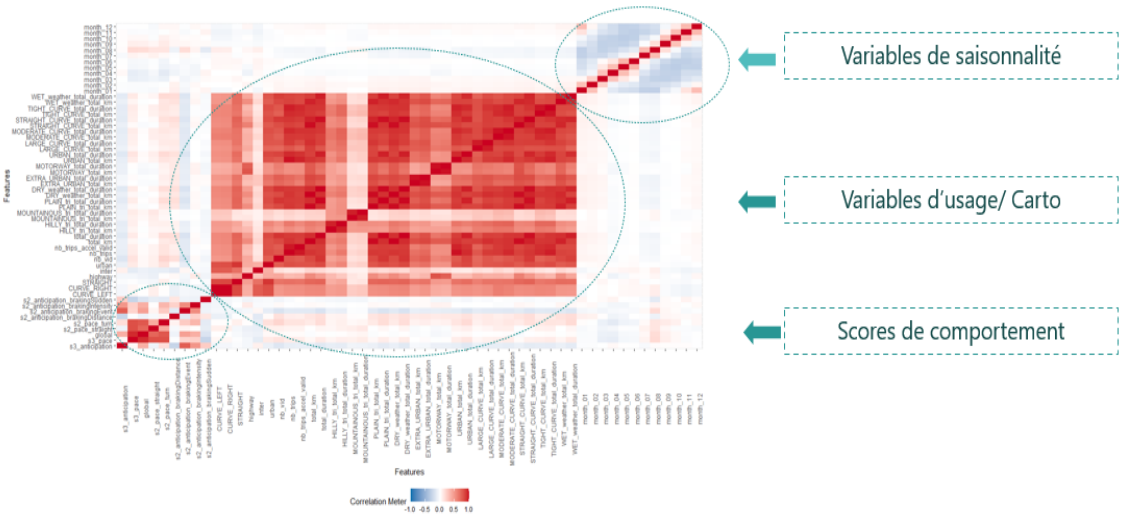

FIGURE 2.13 – *Corrélogramme des données de la base Smart Road Data*

Le corrélogramme révèle trois blocs distincts de variables ayant des corrélations prononcées. Un zoom sur ces trois différents blocs permet de mieux les évaluer et de mieux les comprendre.

- **Zoom sur le bloc des variables de saisonnalité :**

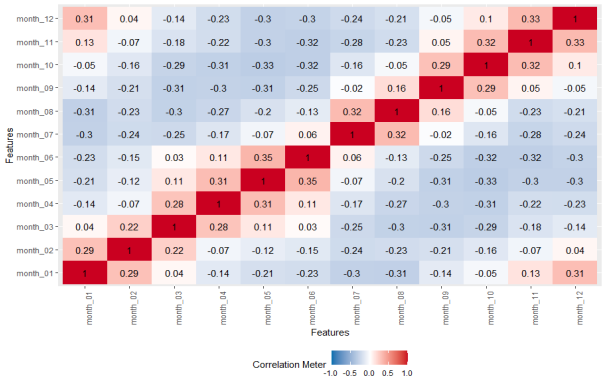

FIGURE 2.14 – *Corrélogramme des 12 variables de saisonnalité*

Ce corrélogramme montre que le niveau de trafic sur un mois quelconque de l'année est corrélé positivement à celui du mois le précédant et à celui du mois le succédant. Cependant il est corrélé négativement aux niveaux de trafic des mois plus éloignés. Le mois de janvier (*month_01*) est une bonne illustration de cette observation. D'après les corrélations présentées, le trafic évolue pareillement en janvier qu'en décembre ou encore en février. Mais dès le mois de mars, les sens des corrélations tendent à s'inverser. Cela est dû à l'effet des saisons : hiver, printemps, été, automne. Au cours des mois d'une même saison, les tendances de trafic sont pratiquement les mêmes, tandis qu'elles changent d'une saison à une autre.

- **Zoom sur le bloc des variables d'usage et de cartographie :**

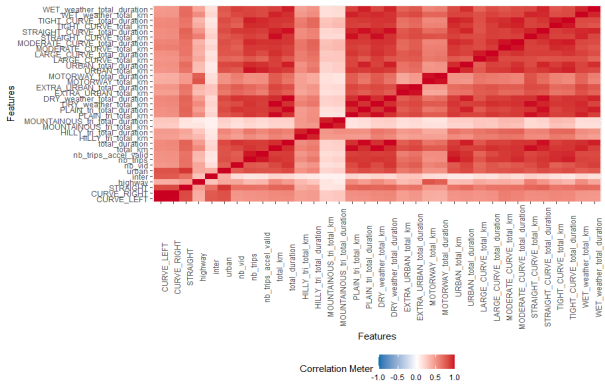

FIGURE 2.15 – *Corrélogramme des variables d'usage et de cartographie*

Dans ce bloc, les variables sont toutes corrélées positivement. Cela s'explique par trois raisons principales. D'abord, les kilomètres totaux parcourus et les durées totales des trajets évoluent dans le même sens, puisqu'en général plus on parcourt de kilomètres et plus le temps de trajet augmente. Ensuite, les informations sur les différents types de route (virages, lignes droites, routes urbaines,...) à l'intérieur des communes en terme de distances sont assez similaires. En effet, plus il y a

de routes en général dans une commune et plus on y trouvera différents types de route. Enfin, les niveaux de kilomètres parcourus et les durées totales de trajets sont généralement élevés dans les communes ayant de nombreuses routes.

- **Zoom sur le bloc des scores de comportement :**

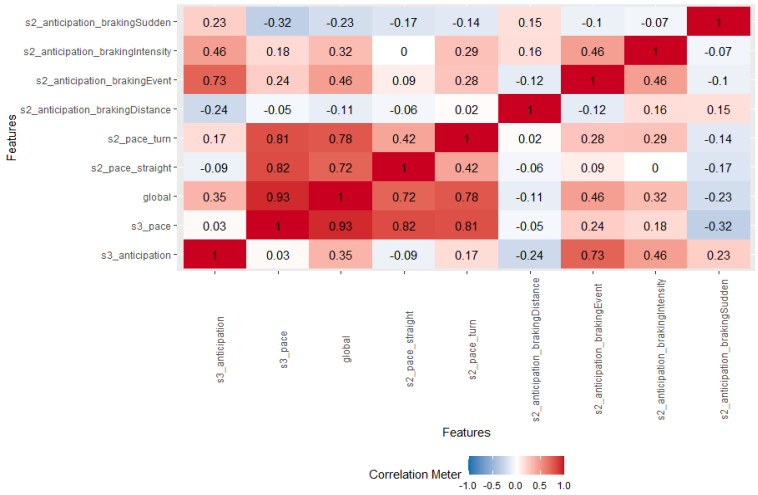

FIGURE 2.16 – *Corrélogramme des scores de comportement*

Dans ce bloc, les corrélations s'observent par niveaux. En effet, les scores S2 sont corrélés à leurs scores S3 correspondant et les scores S3 sont corrélés à la note globale.

2.3.2 ACP et réduction de dimensions

Considérons toujours notre base X contenant n individus ($n > 1$) et p variables ($p > 3$). Le but de l'Analyse en Composantes Principales (ACP) est de fournir une représentation graphique du nuage des individus de la base X sur un espace de dimension réduite q ($q < p$) où q vaut généralement 2 ou 3 afin de l'étudier.

L'ACP permet de transformer les variables initiales en de nouvelles variables appelées composantes principales en s'appuyant sur les relations de dépendance existant entre elles. Ces nouvelles variables sont decorrélées et ordonnées de sorte à ce que les premières contiennent la plus grande partie de l'information contenue dans la base de départ.

Ces composantes principales représentent les coordonnées des individus sur les axes factoriels. C'est à dire les directions de l'espace expliquant au mieux la variance des données initiales.
Algébriquement parlant, les composantes principales sont les vecteurs propres de la matrice de covariance de ces données initiales. La valeur propre λ_k associée à la composante principale k ($k = 1,...,p$) représente le pourcentage de variance expliquée par celle-ci. La

somme de ces valeurs propres permet donc d'obtenir une mesure de la dispersion totale du nuage de points appelée inertie totale I_T :

$$I_T = \sum_{k=1}^{p} \lambda_k$$

Dès lors, pour connaître la part d'inertie expliquée par les m premières composantes principales ($m < p$), il suffit de calculer le ratio :

$$\frac{\sum_{k=1}^{m} \lambda_k}{I_T}$$

Dans la pratique les données sont standardisées avant l'implémentation de l'ACP. La standardisation d'une variable consiste à lui retirer sa moyenne puis la diviser par son écart-type. Cette transformation permet de neutraliser l'effet des différences d'échelle entre les variables, effet qui pourrait biaiser les résultats de l'analyse.

Application : Une des propriétés importantes dans une étude géographique est la proximité des observations pour des communes très proches. Il serait anormal, par exemple, d'observer des valeurs météorologiques très différentes sur deux communes adjacentes. Les résultats de l'ACP permettront de vérifier que les données de la base *Smart Road Data* respectent cette propriété.

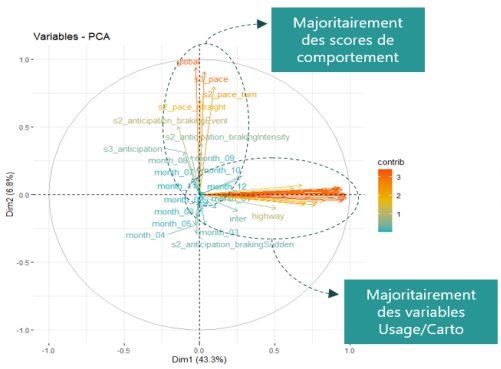

FIGURE 2.17 – *Représentation des variables dans le premier plan factoriel*

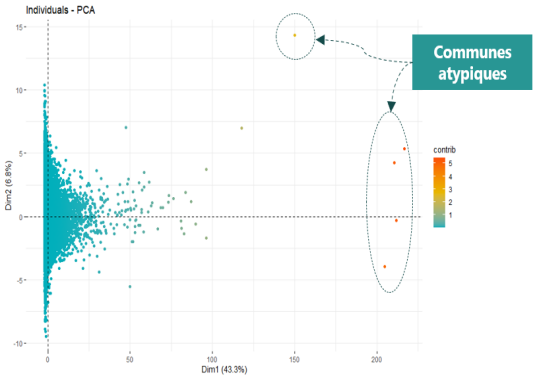

FIGURE 2.18 – *Représentation du nuage de points des communes dans le premier plan factoriel*

Une information communément présentée par les deux graphiques ci-dessus est la part d'inertie expliquée par le premier plan factoriel qui est de 50,1%. En effet, le premier plan factoriel se compose des deux premières composantes principales (dim 1 et dim 2) expliquant respectivement 43,3% et 6,8% de l'inertie totale. Cela signifie que la moitié de l'information contenue dans la base initiale se résume à partir de ces deux

premiers axes. Il est donc pertinent d'observer les communes de cette base dans ce plan.

La figure 2.17 présente les variables ayant contribué à la construction des deux axes. La première composante principale (dim 1, Axe horizontal) est corrélée positivement aux variables concernant l'usage du véhicule et aux variables cartographiques. Cela signifie que la construction de cet axe s'est principalement basée sur les liens entre ces différentes variables (liens déjà présentés dans la sous-section sur les corrélations). De même, la deuxième composante principale (dim 2, Axe vertical) est corrélée positivement aux scores de comportement. Les variables de saisonnalité contribuent très faiblement à la construction de ce premier plan.

La figure 2.18 présente le nuage de points des communes dans le premier plan factoriel. Les observations des communes sont majoritairement regroupées dans un même secteur du plan à l'exception de quelques unes d'entre elles. Cela s'interprète par le fait que dans l'ensemble, les communes très proches ont des valeurs de variables assez similaires. Les données de la base *Smart Road Data* respectent donc la propriété géographique énoncée plus haut. Aussi, convient-il d'expliquer les positions atypiques de quelques communes dans le plan.

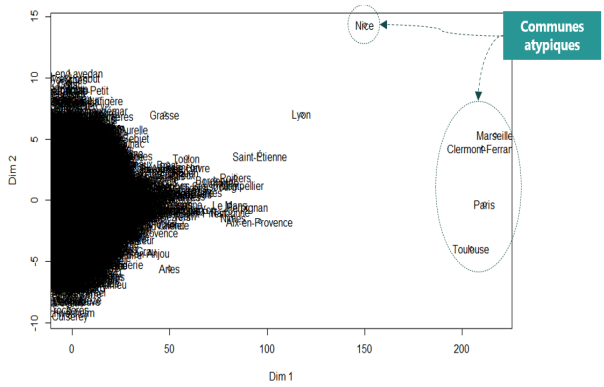

FIGURE 2.19 – *Représentation du nuage de points des communes dans le premier plan factoriel (avec leur nom)*

Les communes ayant des comportements atypiques sont : Marseille, Clermont-Ferrand, Paris, Toulouse et Nice. Les quatre premières ont les valeurs les plus élevées sur la première composante principale et Nice a la valeur la plus élevée sur la deuxième composante principale. La première composante principale étant corrélée positivement aux kilomètres parcourus, aux durées des trajets et aux distances des routes, les valeurs de ces variables sont évidemment très élevées sur ces quatre communes. Le graphique suivant (Figure 2.20) présente en guise d'exemple la valeur du total de kilomètres parcourus sur Paris (point rouge).

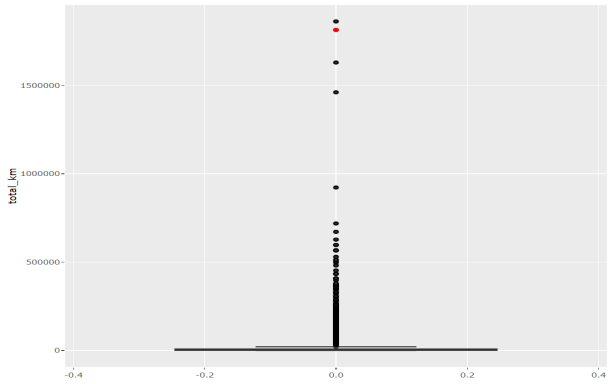

FIGURE 2.20 – *Boîte à moustache du total de kilomètres parcourus avec en rouge la valeur sur Paris*

La deuxième composante principale est, quant à elle, corrélée positivement avec la majeur partie des scores de comportement. La valeur élevée de Nice sur cet axe dénote donc un bon comportement de conduite sur cette commune.

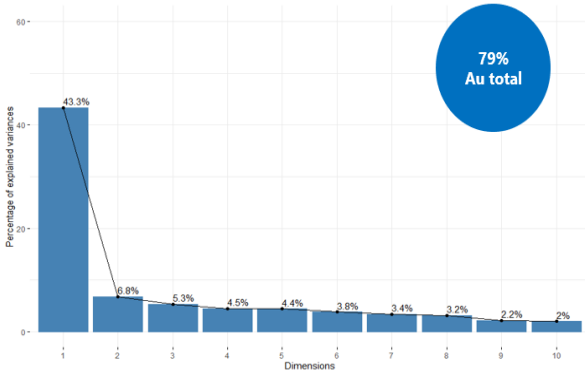

FIGURE 2.21 – *Part de variance expliquée par chacune des dix premières composantes principales*

Enfin, force est de constater que les dix premières composantes principales résument à elles seules environ 80% des informations de la base initiales. Il serait tentant de continuer l'étude avec ces dix nouvelles variables decorrélées et contenant près de la quasi totalité des informations. Cependant l'interprétabilité des données serait compromise. Dans la section suivante, l'objectif sera donc de créer de nouvelles combinaisons de variables qui comporteront une majeur partie des informations initiales, qui seront pour la plupart decorrélées entre elles et qui garderont cet aspect interprétable.

2.4 *Feature Engineering* et sélection non supervisée

2.4.1 *Feature Engineering*

Le processus de *Feature Engineering* consiste à créer, à partir des données initialement présentes dans une base, de nouvelles variables moins corrélées, plus pertinentes

et interprétables. Ce processus est effectué sur les différentes catégories de variables, à l'exception des scores de comportement. En effet, ces variables de comportement ont déjà subi des transformations puisque certaines d'entre elles sont le résultat de combinaisons.

Les variables de saisonnalité :

L'étude des corrélations sur cette catégorie a révélé que l'information majeure apportée par ces douze variables se résumait au niveau du trafic routier sur les différentes zones territoriales suivant les différentes saisons de l'année. Il convient donc de créer quatre nouvelles variables hiver, printemps, été et automne en remplacement des douze variables initiales et qui seront les moyennes des colonnes représentant les mois respectifs de ces saisons.

Les variables concernant l'usage du véhicule :

Ce bloc se décompose suivant différents contextes. Ici seront uniquement présentées les transformations effectuées sur les variables suivant le contexte météorologique. Les transformations des variables suivant les autres contextes respectent la même logique.

Le contexte météorologique dans la base *Smart Road Data* se décline sous la forme d'un système complet d'évènements qui sont : l'évènement "temps sec" et l'évènement "temps humide". En théorie des probabilités, on dit qu'un système d'évènements est complet, si les évènements qui le composent sont deux à deux incompatibles et si leur réunion donne l'univers des possibles.

L'intérêt d'avoir un système complet d'évènement est qu'avec une partie de l'information, il est possible de retrouver l'autre partie. Par exemple, le total de kilomètres parcourus en temps sec se retrouve en soustrayant le total de kilomètres parcourus en temps humide du total de kilomètres parcourus en général. Cela sous-entend qu'il est possible de diminuer la quantité de données sans pour autant perdre de l'information. Ainsi, conserver les deux évènements pour la suite paraît superflu.

Considérons donc maintenant un des deux évènements, par exemple l'évènement "temps humide". Pour cet évènement, la base propose deux variables à savoir : le total de kilomètres parcourus en temps humide et la durée totale des trajets en temps humide. L'étude des corrélations sur ce type de variables a montré qu'elles apportaient des informations assez similaires. Il faut donc réussir à trouver une ou deux transformations qui permettront de palier à ce problème.

Une première transformation se dessine automatiquement lorsqu'on parle de kilomètres et de durées, celle des vitesses. En effet, il est possible d'obtenir une estimation de la vitesse en temps humide en faisant le ratio du nombre total de kilomètres parcourus en temps humide par la durée totale des trajets en temps humide :

$$wet\_weather\_mean\_speed = \frac{wet\_weather\_total\_km}{wet\_weather\_total\_duration}$$

Une seconde transformation qui paraît aussi intuitive que la première est celle des valeurs relatives. En effet, il est quelque fois préférable d'utiliser des pourcentages en lieu et place des valeurs brutes de comptage. Dans notre cas, une estimation du taux de parcours en temps humide s'obtient par le ratio du total de kilomètres parcourus en temps humide par le total de kilomètres parcours en général :

$$wet\_weather\_part\_of\_km = \frac{wet\_weather\_total\_km}{total\_km}$$

Finalement l'ensemble de ces transformations peut se résumer à l'aide du tableau suivant :

	Général	Humide	Sec
Km	✓	✗	✗
Durée	✗	✗	✗
Vitesse	✓	✓	✗
Taux de km		✓	✗
Taux de durée		✗	✗

FIGURE 2.22 – *Tableaux récapitulatif des transformations effectuées (contexte météo)*

A l'intérieur du tableau, les v verts représentent les variables à conserver et les croix oranges représentent les variables supprimables. Il est aisé de vérifier par de petits calculs que les variables supprimables sont retrouvables à partir de combinaisons des variables conservées.

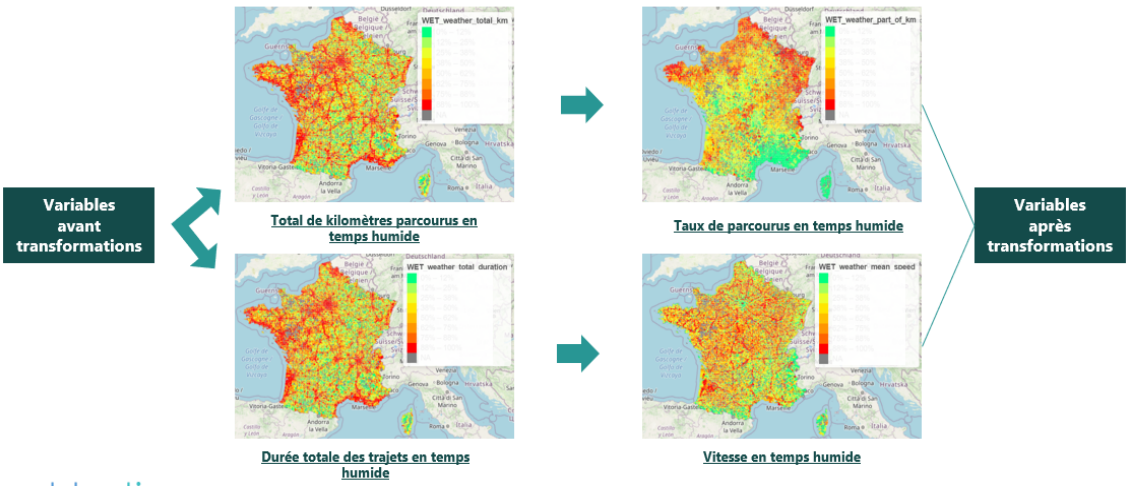

FIGURE 2.23 – *Cartographie des variables avant/après transformations (contexte météo)*

Sur les Cartographies présentées ci-dessus (Figure 2.23), les gradients de couleurs passent du vert signifiant des valeurs très faibles au rouge représentant des valeurs élevées.
Il est possible d'observer que les Cartographies des variables initiales sont assez similaires du fait de la forte corrélation qui les lie. Cependant, l'application des transformations retenues permet d'aboutir à de nouvelles variables ayant des représentation très différentes et donc apportant des informations probablement décorrélées. De plus, il a déjà été démontré que les transformations effectuées n'entraînaient aucune perte d'informations.

Le passage du total de kilomètres parcourus en temps humide au taux de kilomètres parcourus en temps humide est un excellent indice qui confirme la pertinence des transformations. En effet, l'information apportée par la variable renseignant sur le compteur kilométrique en temps humide est pareille que celle qu'apporte la variable renseignant sur le total de kilomètre en général. L'aspect météorologique qu'est censé contenir cette variable n'est pas mis en relief. Le fait de passer au taux de parcours en temps humide permet de révéler ce contexte météorologique de la donnée. La nouvelle information transmise est la probabilité de faire un parcours en temps humide suivant les différentes zones. Cette information est très visible sur la Cartographie du taux de parcours en temps humide. Il est par exemple possible de remarquer que la probabilité de faire un parcours en temps humide dans le sud de la France est très faible, tandis qu'elle est très élevée en Bretagne. Cette observation paraît assez objective lorsqu'on se réfère à la pluviométrie dans ces deux zones. En outre, il convient de noter le volet interprétable de ces nouvelles données post-transformations.

Dans le cadre de cet exemple, l'évènement conservé etait "temps humide". Dans la suite de cette section, des arguments seront avancés afin de justifier les choix des différents évènements à conserver ou à rejeter.

Les variables cartographiques :

L'idée développée ici est assez similaire à celle présentée ci-dessus. Considérons par exemple la courbure des routes à l'intérieur des communes : ligne droite ou virage. C'est encore une fois une structure de système complet. Il serait donc possible de conserver l'information que sur un seul de ces types. La transformation en relatif est encore appliquée. Elle permet par exemple d'aboutir aux taux de virage à l'intérieur des communes, c'est à dire le ratio des distances des virages par la somme de celles-ci et des distances des lignes droites :

$$part\_of\_curve = \frac{Curve}{Curve + Straight}$$

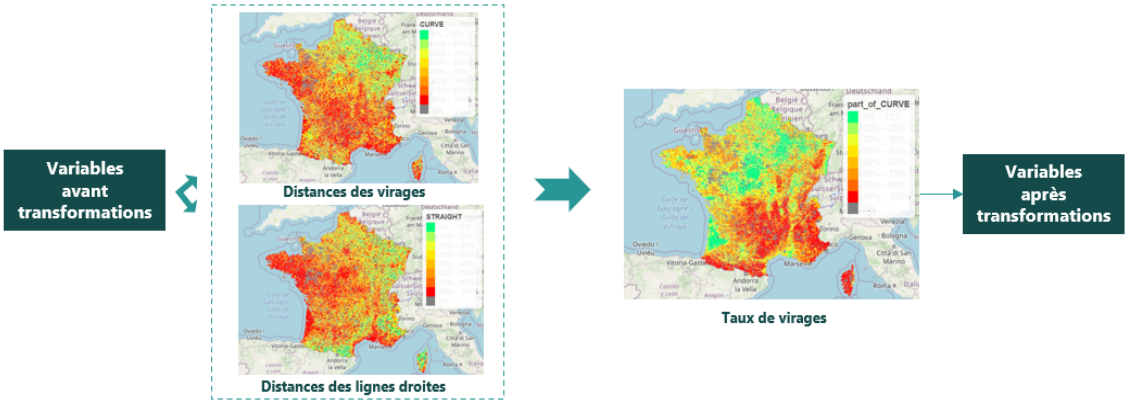

FIGURE 2.24 – *Cartes des variables avant/après transformations (courbure de routes)*

L'information apportée par les deux variables de départ se résume à la quantité de routes en général à l'intérieur des communes. La courbure de la route est moins mis en avant. Pourtant, en passant au taux de virage, cette information devient la probabilité d'emprunter un virage dans une commune donnée. Il est par exemple notable qu'il est plus probable d'emprunter un virage dans le massif central que dans les flandres.

2.4.2 Sélection non supervisée des variables de la base *Smart Road Data*

L'étape de *Feature Engineering* a permis d'aboutir à de nouvelles variables pertinentes, interprétables et qui semblent decorrélées. Concernant les corrélations entre ces nouvelles données, elles sont étudiées à la fin de cette sous section.
Cette étape de transformations de données a aussi soulevé un point très important qui est le choix des informations à conserver à l'intérieur des différents systèmes complets d'évènements. Ce choix a été effectué par avis métier et par réflexion sur la sinistralité de la garantie RCM. C'est donc une sélection par a priori et de manière non supervisée. En guise d'illustration, considérons encore le cas des variables concernant l'usage du véhicule suivant le contexte météorologique. Le choix est à faire entre le contexte "temps humide" et "temps sec". Pour la suite des travaux, il a été décidé de conserver l'évènement "temps humide" puisqu'il favorise naturellement l'accentuation de la sinistralité en automobile et donc représente un facteur de risque plus important pour un assureur.

Cependant, il a déjà été démontré que le choix de cet évènement n'occasionnait aucune perte d'informations. Par exemple, pour retrouver le taux de parcours en temps sec, il suffit de soustraire de 1 le taux de parcours en temps humide :

$$dry\_weather\_part\_of\_km = 1 - wet\_weather\_part\_of\_km$$

De manière plus formelle, cela signifie qu'il suffit de considérer l'effet inverse des taux de parcours en temps humide sur la sinistralité pour connaître celui des taux de parcours en temps secs.

Finalement, après la sélection non supervisée il ne reste plus que 37 variables (contre 56 au départ) à conserver pour la suite. La liste de ces variables est consultable en Annexe A. Le graphique suivant montre les niveaux de corrélation avant et après transformation/sélection.

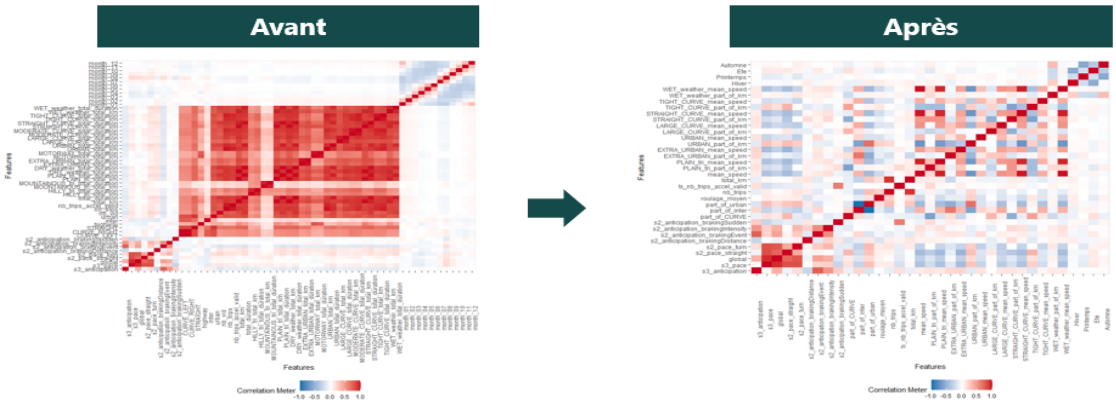

FIGURE 2.25 – *Corrélogrammes avant et après transformation - sélection*

La comparaison avant et après transformation/sélection montre une baisse des niveaux de corrélation entre les variables présentes en bases. Certes, les variables conservées ne sont pas totalement decorrélées comme dans l'ACP mais elles sont toutes pertinentes et interprétables. Dans la suite des travaux, une approche supervisée permettra d'affiner davantage cette sélection.

2.5 Fusion des deux sources de données

Tout au long de ce chapitre ont été séparément présentées les deux sources de données de l'étude. Cette section explicite la méthodologie appliquée afin de les relier.

La base de l'assureur contient comme clé géographique le code IRIS. Tandis que les donnée externes de la base *Smart Road Data* sont à la maille commune (code INSEE). Les IRIS étant des parties de communes, il convient dans un premier temps de faire correspondre les codes IRIS de l'assureur à leur code INSEE respectif. Une fois cette étape de correspondance IRIS-INSEE terminée, la fusion des deux sources de données est effectuée avec pour clé de jointure les codes INSEE.

A l'issu de cette fusion, il ressort qu'environ 92% des communes présentes dans la base de l'assureur se retrouvent dans la base externe. Ces 92% de communes bien reliées représentent environ 95% de l'exposition sur le portefeuille, soit environ 5% d'exposition perdue.

Que retenir des chapitres 1 et 2 ?

Ces deux premiers chapitres ont permis d'introduire de manière contextuelle le sujet de ce mémoire, de présenter ses enjeux et de décrire les différentes sources de données utilisées pour la réalisation de ces travaux.

Dès lors, l'évaluation de l'apport des données télématiques dans la modélisation du risque géographique peut être entamée. Celle-ci passe d'abord, par la conception de premiers modèles, privés volontairement d'informations géographiques et qui serviront de point de départ pour cette modélisation du risque géographique. Le chapitre suivant détaillera la procédure pour obtenir ces modèles initiaux.

Chapitre 3

Modélisation classique de la prime pure

« Plus elle se prête au changement, plus l'œuvre classique est vitale. »

Azorin, Lectures espagnoles

Afin de modéliser le risque géographique par le biais de données télématiques, il convient dans un premier temps de construire des modèles tarifaires de référence démunis de toute information géographique. Dans le cadre des travaux de ce mémoire, ces modèles tarifaires sont des modèles linéaires généralisés (GLM). Jusqu'à présent, les GLM restent les modèles classiques les plus utilisés pour la tarification des produits d'assurance IARD du fait de leur simplicité et de leur interprétabilité.

Dans ce chapitre, la théorie sur ce type de modèles est présentée ainsi que son application sur le portefeuille de l'assureur.

3.1 Modèle collectif

Soient les variables aléatoires suivantes :

- N : Les nombres de sinistres observés au cours d'un exercice (variable discrète à valeur dans $\mathbb{N}$),
- $(Y_i)_{i\geq 0}$: la suite de coûts engendrés par chacun de ces sinistres (variables continues à valeur dans $\mathbb{R}$).

La charge totale de sinistres S s'écrit :

$$S = \sum_{i=1}^{N} Y_i$$

La prime pure se définit comme étant le montant que l'assureur devra payer en moyenne afin de tenir ses engagements, c'est à dire couvrir les charges de sinistres. C'est la prime technique estimée avant toutes retouches commerciales. Partant de cette définition, le calcul de la prime pure se résume au calcul de l'espérance mathématique de la variable aléatoire S présentée ci-dessus.

Sous les hypothèses d'indépendance et d'identité des distributions des coûts $(Y_i)_{i\geq 0}$, et de plus en les supposant indépendants aux nombres N de sinistres, l'espérance mathématique de la variable aléatoire S se décompose de la façon suivante :

$$\mathbb{E}(S) = \mathbb{E}(N).\mathbb{E}(Y_1) = \mathbb{E}(N.Y_1) \tag{3.1}$$

Dans la pratique l'assureur dispose d'informations (âge, catégorie socio-professionnelle,...) lui permettant de mieux estimer les valeurs de ces différentes espérances. C'est le principe de la tarification dite "a priori". En effet, l'assureur affine son calcul du risque grâce à des informations qu'il récupère en amont de la réalisation de celui-ci lors de la souscription du contrat.

Soit Ω l'ensemble de ces informations, les égalités (3.1) deviennent :

$$\mathbb{E}(S|\Omega) = \mathbb{E}(N|\Omega).\mathbb{E}(Y_1|\Omega) = \mathbb{E}(N.Y_1|\Omega)$$

Finalement, le calcul de la prime pure peut donc s'obtenir de deux manières : soit en faisant le produit des estimations des espérances des nombres de sinistres et des coûts

(modélisation fréquence-sévérité) soit en estimant l'espérance du produit des nombres et coûts de sinistres (modélisation du *burning cost*). Dans l'optique d'obtenir ces différentes estimations, l'outil le plus utilisé jusqu'à ce jour en assurance IARD restent les modèles linéaires généralisés (GLM). Le succès de ces modèles est dû à leur lisibilité, c'est à dire leur facilité d'interprétation, mais aussi à la simplicité de leur implémentation.

3.2 Théorie des Modèles Linéaires Généralisés (GLM)

Les modèles linéaires généralisés ou *Generalized Linear Models (GLM)*, ont pour la première fois été proposés dans la littérature scientifique par Nelder et Wedderburn en 1972.

L'idée de ces types de modèles est d'estimer l'espérance mathématique d'une variable à modéliser, conditionnellement à un ensemble de caractéristiques appelées variables explicatives et en supposant connue la loi de cette variable à modéliser sachant ces variables explicatives.

Soit Y la variable à modéliser et $\Omega = \{X_1,X_2,...,X_p\}$ l'ensemble des variables explicatives. Le modèle GLM s'écrit :

$$g(\mathbb{E}(Y|\Omega)) = X\beta \iff g(\mathbb{E}(Y|X_1,X_2,...,X_p)) = \beta_0 + \sum_{k=1}^{p} \beta_k X_k$$

avec g une fonction strictement monotone et dérivable appelée fonction de lien, et $(\beta_i)_{0\leq i\leq p}$ les paramètres à estimer du modèle.

La loi de $Y|\Omega$ doit appartenir à une famille spécifique de loi qui est la famille exponentielle. La loi d'une variable aléatoire Y appartient à la famille exponentielle si sa fonction de densité peut s'écrire sous la forme :

$$f_{\theta,\psi}(y) = exp\left(\frac{y\times\theta - b(\theta)}{a(\psi)} + c(y,\psi)\right)$$

avec :

- θ le paramètre naturel de la famille de loi exponentielle,
- ψ le paramètre de dispersion de la distribution,
- b une fonction définie sur $\mathbb{R}$ trois fois dérivable et de dérivée première injective,
- a et c deux fonctions définies et respectivement dérivable et différentiable sur $\mathbb{R}$ et $\mathbb{R}^2$.

Ainsi grâce à l'expression de la fonction de densité de $Y|\Omega$, il se déduit une autre expression du calcul de son espérance :

$$\mathbb{E}(Y|\Omega) = b'(\theta) \iff \theta = H(\beta_0 + \sum_{k=1}^{p} \beta_k X_k)$$

où H est la composée de la fonction inverse de la dérivée de b et de la fonction inverse de la fonction de lien g.

Le paramètre θ de la loi de $Y|\Omega$ est donc lié aux paramètres $(\beta_i)_{1\leq i\leq n}$ du modèle. Il est donc possible d'estimer ces $(\beta_i)_{1\leq i\leq n}$ par une méthode faisant intervenir la densité de $Y|\Omega$. l'une des méthodes couramment utilisée dans ce cas est la méthode du maximum de vraisemblance. Dans la pratique, la maximisation s'effectue plutôt sur la log-vraisemblance.

La log-vraisemblance de $Y|\Omega$ s'écrit :

$$\mathcal{L}(y,\beta) = \sum_{i=1}^{n} ln(f_{\theta_i,\psi}(y_i))$$

L'estimateur $\hat{\beta} \in \mathbb{R}^{p+1}$ du vecteur des β est tel que $\hat{\beta} = \underset{\beta\in\mathbb{R}^{p+1}}{argmax}\ \mathcal{L}(\beta)$.

3.3 Sélection supervisée de variables et qualité de modèle

Pour des raisons de précision, de stabilité et de consistance du modèle, il convient de conserver comme variables explicatives, uniquement les informations pertinentes influençant significativement la variable à expliquer. Ce processus de tri d'informations s'appelle la sélection de variables. Ici, cette dernière est dite supervisée car elle est orientée par la variable à prédire. Ce type de sélection s'appuie sur des critères mathématiques permettant d'évaluer la qualité du modèle.

3.3.1 Méthodes de Sélection supervisée de variables

Il en existe plusieurs dans la littérature scientifique. Les plus courantes sont les méthodes "pas à pas" : *backward*, *forward* et *stepwise*, et l'algorithme de régression pénalisée LASSO.

Méthode *forward*

La méthode *forward* ou sélection en avant se base sur l'ajout successif de variables dans le modèle. Au départ, celui-ci ne contient aucune variable explicative. En partant de ce modèle "vide", les variables sont rajoutées une par une. Si l'ajout d'une variable dégrade la qualité du modèle, celle-ci est écartée.

Le problème dans cette méthode de sélection est qu'une variable peut être significative à une étape et ne plus l'être aux étapes suivantes. Pourtant, aucune élimination n'a lieu a posteriori.

Méthode *backward*

La méthode *backward* ou sélection en arrière se base sur le retrait successif de variables dans le modèle. Le modèle de départ est "complet", c'est à dire qu'il contient toutes les variables explicatives à disposition. En partant de ce modèle "complet", les variables sont éliminées une par une jusqu'à aboutir à un modèle ne contenant que des variables significatives.

Pour rappel le modèle GLM s'écrit :

$$g(\mathbb{E}(Y|X_1,X_2,...,X_p)) = \beta_0 + \sum_{k=1}^{p} \beta_k X_k$$

Le test statistique de significativité opéré à chaque itération est le test de Wald. Son principe est le suivant :

Soit une variable explicative X_j du modèle, les hypothèses du test pour celle-ci sont :

$$H_0 : \beta_j = 0 \quad contre \quad H_1 : \beta_j \neq 0$$

L'hypothèse (H_0) signifie qu'il est possible d'annuler le coefficient de la variable X_j et donc de la supprimer du modèle. Si cette hypothèse est rejetée avec une probabilité $1-\alpha$ (de l'ordre de 95%) alors la variable est à conserver et est donc significative.

Le problème de la méthode *backward* est qu'une variable peut ne pas être significative dans un modèle du fait de la présence d'une autre variable à l'intérieur de celui-ci. Pourtant une fois supprimée, aucune réintégration de variables n'a lieu a posteriori.

Méthode *stepwise*

La méthode *stepwise* ou sélection bidirectionnelle est pensée de sorte à résoudre les problèmes des deux précédentes méthodes. Son principe est d'intercaler des sélections *backward* à chaque itération d'une sélection *forward*. Ainsi, une variable significative à une étape pourra être supprimée dans les étapes suivantes si elle ne l'est plus et vice versa.

C'est la meilleure des méthodes de sélection dites "pas à pas". Cependant, elle peut rapidement devenir très chronophage en présence d'un jeu de données de grande dimension (beaucoup de lignes et/ou de colonnes).

Régression LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) est un modèle de régression tout comme les GLM, mais qui effectue une pénalisation des coefficients β à l'intérieur de celui-ci. C'est cet aspect de pénalisation qui fait que cet algorithme est très souvent utilisé pour la sélection de variables. En effet, en réduisant les valeurs de certains β associés à des variables, il suggère leur suppression du modèle puis-qu'influençant très faiblement les prédictions de celui-ci. Cette pénalisation passe par la minimisation de la fonction :

$$Lasso(\beta,\lambda) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

avec λ un paramètre fixé arbitrairement.

Le choix du λ est déterminant pour la qualité de la sélection finale de variables en sortie de ce modèle. Assurément, plus la valeur du λ est élevé et plus les valeurs $\hat{\beta}_j$ minimisant la fonction $Lasso(\beta,\lambda)$ seront faibles voir nulles, impliquant donc un faible nombre de variables conservées. Ainsi, choisir un λ trop grand induirait l'exclusion de toutes les variables du modèle. Inversement, du choix d'un λ très faible découlerait une sélection conservant l'entièreté des variables. Il faut donc trouver la valeur de λ qui favoriserait un compromis entre complexité et performance du modèle.

3.3.2 Indicateurs de qualité d'un modèle

En vue d'évaluer la qualité d'un modèle construit à partir des processus précédemment présentés, plusieurs indicateurs et méthodes peuvent être utilisés. Dans cette sous-section sont décrits les indicateurs et les méthodes utilisés dans le cadre des travaux de ce mémoire.

Deviance du modèle

La deviance est un indicateur mesurant l'écart entre le modèle GLM saturé et un modèle GLM construit. Le modèle GLM saturé est le modèle ayant autant de paramètres que d'observations. C'est une utopie puisque dans la pratique cela n'est quasiment jamais réalisable. La deviance permet donc de quantifier la distance entre ce modèle "parfait" et le modèle GLM construit. Statistiquement, cela se traduit par le calcul de la différence des log-vraisemblances de ces deux modèles.

Soient $\mathcal{L}(y,\beta)$ et $\mathcal{L}_s(y,\psi)$ les log-vraisemblances respectives des modèles GLM construit et saturé. La deviance D s'obtient par :

$$D = 2[\mathcal{L}_s(y,\psi) - \mathcal{L}(y,\beta)]$$

Plus la deviance est faible, meilleure est l'adéquation du modèle construit aux données observées.

Le problème est que plus on ajoute de variables explicatives au GLM, plus le modèle tend à être saturé et plus la déviance diminue. Ainsi, la déviance peut dans certains cas ne pas être un bon critère de comparaison de la qualité de deux ou plusieurs modèles.

Akaike Information Criterion (AIC)

L'AIC est aussi un critère de mesure de la qualité d'un modèle GLM. Cependant à la différence de la deviance, le calcul de l'AIC prend en compte la complexité du modèle, c'est à dire le nombre de paramètres estimés. De ce fait, ce critère peut légitimement être utilisé afin de comparer la qualité de deux modèles GLM sans se préoccuper de leur nombre respectif de variables explicatives. La valeur de l'AIC s'obtient par la formule suivante :

$$AIC = 2[p - \mathcal{L}(y,\beta)]$$

où p est le nombre de paramètres estimés et $\mathcal{L}(y,\beta)$ est la log-vraisemblance du modèle GLM construit. Plus l'AIC du modèle est faible, meilleur est la qualité de celui-ci en terme d'apprentissage d'informations.

Bayesian Information Criterion (BIC)

Le BIC est un critère inspiré de l'AIC. Il se calcule de la manière suivante :

$$BIC = \log(n)p - 2\mathcal{L}(y,\beta)$$

avec n la taille de l'échantillon d'observations. En faisant intervenir la taille de l'échantillon d'observations dans son calcul, le BIC accentue la pénalisation des modèles GLM à très grande complexité. Plus le BIC est faible, meilleur est la qualité du modèle évalué en terme d'apprentissage d'informations.

Indice de Gini

L'indice de Gini est un indicateur de la dispersion d'une distribution dans une population. Il est souvent utilisé en économie pour mesurer l'inégale répartition des richesses au sein de la population d'un pays. En statistique, il sert à mesurer le pouvoir discriminant d'un modèle. En d'autres mots, il informe sur la capacité du modèle à segmenter les risques des moins importants aux plus importants.

Son calcul se base sur la courbe de Lorenz. Cette dernière est la représentation graphique de la fonction qui, à la part x des détenteurs d'une grandeur, associe la part y de la grandeur détenue. La figure 3.1 permet de visualiser un exemple de courbes de Lorenz/Gini.

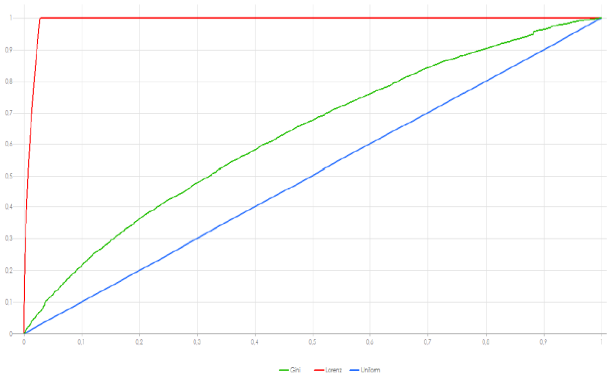

FIGURE 3.1 – *Exemple de courbes de Lorenz/Gini*

Sur ce graphique, la première bissectrice (ligne bleue) représente une situation d'égalité parfaite au sein de l'échantillon. Aucune discrimination n'est donc opérée.
La courbe rouge représente la courbe de Lorenz associée aux observations du modèle. L'aire entre cette courbe et la première bissectrice traduit l'inégalité de distribution du risque au sein de l'échantillon observé.
Enfin, la courbe verte est la courbe du Gini du modèle. Elle représente la quasi-courbe de Lorenz associée aux observations, c'est à dire la courbe de Lorenz obtenue à partir de ces observations ordonnées selon les valeurs de leurs prédictions correspondantes. Plus la courbe du Gini se rapproche de la courbe de Lorenz, plus le modèle est discriminant.

De manière plus formelle, l'indice de Gini du modèle s'obtient soit en faisant :

$$Gini = \frac{A}{0.5}$$

avec A l'aire entre la courbe du Gini et la première bissectrice (formule standard du Gini), soit en faisant :

$$Gini_{norm} = \frac{A}{B}$$

où B est l'aire entre la courbe de Lorenz et la première bissectrice (formule du Gini normalisé).

De façon équivalente à ce qui a été dit plus haut, plus la valeur de ces indices est élevée, plus le modèle est discriminant.

Validation croisée

L'une des plus importantes qualités que doit avoir tout modèle de prédiction est la capacité à généraliser le phénomène modélisé sur de nouvelles données. En effet, un

modèle qui ne fonctionne que sur les données présentes n'est pas très efficient puisqu'il ne pourra servir en cas de disponibilité de nouvelles données. Par exemple, un modèle d'assureur qui ne donne de "bonnes prédictions" de fréquences de sinistres que pour les assurés déjà enregistrés ne pourra être utilisé lors de la souscription de nouveaux clients. L'expression "bonnes prédictions" utilisée dans la phrase précédente, fait allusion au fait que les prédictions du modèle soient très proches des valeurs observées de la variable à prédire.

En vue de construire un modèle capable de généraliser le phénomène modélisé sur de nouvelles données, une stratégie classiquement utilisée est la validation croisée *Hold-out*. Cette stratégie consiste à scinder la base de données à disposition en deux sous-bases. Sur l'une de ces sous-bases est construit le modèle, elle est appelée base d'apprentissage. Sur l'autre sous-base est testée la capacité de généralisation de celui-ci, elle est appelée base de validation. La validation croisée *Hold-out* est la plupart du temps utilisée pour l'évaluation de modèles finaux. Cependant, elle peut quelques fois fournir des résultats assez instables quand elle est appliquée lors d'une sélection de variables ou lors du choix des paramètres d'un modèle. Pour ces dernières situations, il est préférable d'utiliser une validation croisée *k-fold*. La validation croisée *k-fold* est une généralisation de la validation croisée *Hold-out*. Cette fois, la base de départ est divisée en k sous-ensembles ($k \geq 2$ et choisi arbitrairement). A chaque itération du processus *k-fold*, une des k sous-bases est choisie comme la base de validation et les $k-1$ sous-bases restantes forment la base d'apprentissage. La figure 3.2 est une illustration du fonctionnement de la validation croisée *k-fold*.

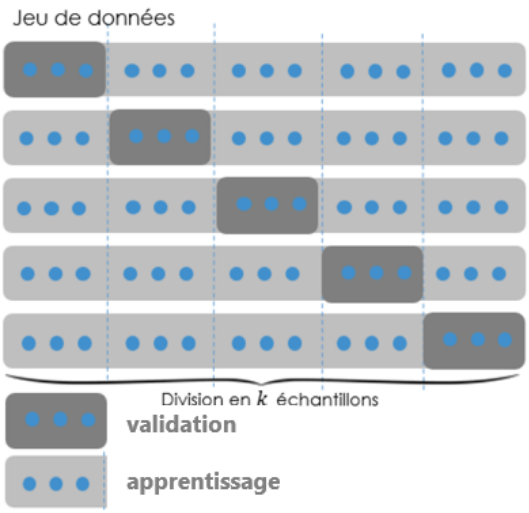

FIGURE 3.2 – *Fonctionnement de la validation croisée k-fold*

Une métrique permettant d'évaluer la proximité entre les valeurs observées et les valeurs prédites de la variable à expliquer est la racine de l'erreur quadratique moyenne

ou *Root Mean Square Error* (RMSE) qui s'écrit :

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

avec y_i la i ième valeur observée de la variable à expliquer et $\hat{y}_i$ sa prédiction par le modèle.

Un modèle qui a une RMSE très faible sur une base d'apprentissage comparée à celle obtenue sur une base de validation, est un modèle qui généralise difficilement le phénomène à expliquer. Cette situation s'appelle le sur-apprentissage. A contrario, un modèle qui a une RMSE trop élevée sur une base d'apprentissage comparée à celle obtenue sur une base de validation, est un modèle sous-appris. Il faut donc rechercher un modèle qui fait un compromis entre qualité d'apprentissage et qualité de généralisation. En statistique, cela s'appelle le compromis biais-variance. Une illustration de ce compromis est observable sur la Figure 3.3.

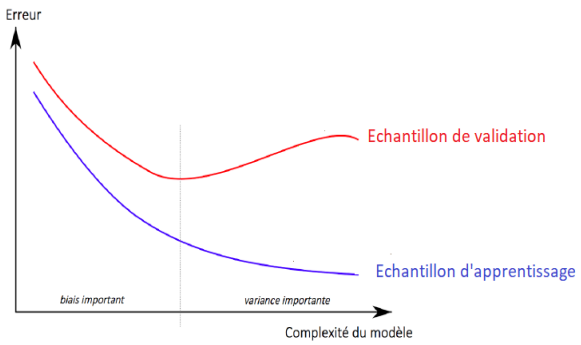

FIGURE 3.3 – *Illustration du compromis biais-variance*

3.4 Application sur le portefeuille d'assurance de l'étude

Les sections ci-dessus ont montré que le calcul de la prime pure passait par des estimations d'espérances mathématiques par le biais de modèles de régression tels que les GLM. Dans cette section, l'objectif est donc de construire des modèles GLM de référence, démunis de toutes informations géographiques et qui permettent d'obtenir des premières valeurs de primes pures. Dans les prochains chapitres, ces modèles sont modifiés suivant différentes approches de sorte à évaluer l'apport des données télématiques externes dans la modélisation du risque géographique.

La méthodologie adoptée pour la construction des modèles GLM est la suivante :

1. les choix de la loi et de la fonction de lien ;
2. la sélection supervisée des variables ;

3. le traitement des variables sélectionnées ;
4. l'implémentation et l'évaluation du modèle GLM construit.

3.4.1 Modélisation de la survenance de sinistres

La modélisation de la survenance de sinistres consiste à estimer l'espérance des nombres de sinistres sachant des variables explicatives. Dans la suite des travaux, les deux approches de modélisation du risque géographique par des données télématiques contextualisées sont basées sur ce modèle. Ce choix est dû au fait qu'il est généralement plus aisé d'observer l'impact géographique sur les nombres de sinistres que sur les coûts moyens.

Choix de la loi et de la fonction de lien

La construction d'un modèle GLM passe en premier lieu par le choix de la loi et de la fonction liant l'espérance de la variable à prédire aux variables explicatives. En assurance IARD, il est classique de choisir pour la modélisation de la survenance de sinistres une distribution de **poisson** comme loi et de choisir le **logarithme népérien** comme fonction de lien. La loi de poisson a l'avantage d'être très adaptée pour modéliser les processus de comptage. Quant au choix du logarithme népérien comme fonction de lien, il implique que :

$$\mathbb{E}(N|X_1,X_2,...,X_p) = \exp(\beta_0).\prod_{k=1}^{p}\exp(\beta_k X_k) \tag{3.2}$$

Le modèle adopte donc une structure multiplicative. Cela simplifie son interprétation et sa mise en production.
Savoir interpréter ce modèle (3.2) est primordiale pour la bonne compréhension de la suite des travaux. Il convient donc de l'expliquer :

- La constante $\exp(\beta_0)$ est la valeur qui aurait été retournée si le modèle ne contenait aucune variable explicative. Elle représente donc le nombre moyen de sinistres (attritionnels dans notre cas) observés sur le portefeuille de l'assureur. En statistique, le β_0 est appelé intercepte ou effet commun.

- Les valeurs des variables $\exp(\beta_k X_k)$ avec $(k = 1,..,p)$ constituent des facteurs multiplicatifs de majoration ou de minoration du nombre moyen de sinistres $\exp(\beta_0)$. En effet, si $\exp(\beta_k X_k) < 1$, alors multiplier le nombre moyen de sinistres par cette quantité revient à le faire baisser, sinon il en résulte l'effet inverse. Ainsi, Pour connaître l'effet d'une variable X_k dans un modèle GLM, il suffit de connaître la valeur du facteur multiplicatif qui lui est associé.

- Parmi les variables explicatives du modèle $\{X_1,X_2,...,X_p\}$ se trouve le logarithme népérien de l'exposition (on supposera : $X_1 = \log(Exposition)$ afin d'alléger les

écritures). Spécialement, pour cette variable, le coefficient β_1 qui lui est associé est contraint à valoir 1. Le fait de contraindre la valeur du coefficient β_k associé à une variable explicative X_k ($k = 1,...,p$) à valoir 1 s'appelle la "mise en *offset*" de cette variable. Le modèle (3.2) peut donc se réécrire :

$$\mathbb{E}(\tilde{N}|X_1,X_2,...,X_p) = \frac{\mathbb{E}(N|X_1,X_2,...,X_p)}{Exposition} = \exp(\beta_0).\prod_{k=2}^{p}\exp(\beta_k X_k) \qquad (3.3)$$

L'exposition est donc utilisée comme une pondération des nombres de sinistres à l'intérieur du modèle. Cela permet de crédibiliser les valeurs prédites par celui-ci. Finalement, la variable à modéliser devient la fréquence de sinistres, qui pour rappel est le rapport du nombre de sinistres par l'exposition.

Sélection supervisée des variables

Une fois les choix de la loi et de la fonction de lien effectués, il faut réussir à déterminer parmi toutes les variables à disposition, celles qui influencent le plus significativement la variable à prédire. Cette dernière étant ici la fréquence de sinistres, il s'agit donc de trouver des facteurs permettant d'expliquer au mieux les variations de celle-ci.

Dans la pratique, le nombre de variables internes peut être très élevé. De plus, ces variables sont très souvent en majorité qualitatives. C'est le cas de la base interne de l'assureur de l'étude.

Dans cette situation, le processus de tri des variables démarre d'abord par une présélection par avis métier. Ces avis métier permettent d'écarter un premier lot de variables qui a priori n'influencent pas la fréquence de sinistres.

Ensuite, une étude des corrélations est menée sur les variables conservées. Elle met en lumière les potentielles redondances d'informations apportées par certaines variables. Dans la mesure où deux variables corrélées sont retenues par la méthode de sélection finale, des tests statistiques sont utilisés afin de n'en garder qu'une des deux.

Enfin, la sélection finale des variables est implémentée. La méthode choisie est la méthode *stepwise* s'appuyant sur une validation croisée. le *stepwise* est la plus robuste des méthodes de sélection dites "pas à pas".

A l'issue de ce processus de sélection, les variables de la base interne de l'assureur conservées sont :

les variables concernant le profil de l'assuré :

- âge du conducteur ;
- catégorie socio-professionnelle ;
- coefficient novice [1]

les variables concernant le véhicule de l'assuré :

- âge du véhicule ;
- puissance du véhicule ;
- classe SRA. [2]

Les variables géographiques internes sont volontairement exclues du processus de sélection en vue de construire un premier modèle démuni de toutes informations géographiques.

Les variables retenues sont supposées être explicatives de la fréquence de sinistres. Une analyse des variations de cette fréquence de sinistres en fonction de chacune d'entre elles permet donc de comprendre les facteurs d'accentuation du risque. Ici, pour des raisons de confidentialité, cette analyse n'est présentée uniquement que pour deux variables sélectionnées : l'âge du conducteur pour les variables concernant le profil de l'assuré et l'âge du véhicule pour les variables concernant le véhicule de l'assuré.

Sur les figures 3.4 et 3.5 suivantes, la courbe de couleur rouge représente la fréquence moyenne de sinistres pour chacune des modalités de la variable explicative analysée et les barres de couleur verte, les niveaux d'exposition sur ces différentes modalités.

1. C'est une classification des conducteurs régie par le code des assurances. Ce code définit un conducteur comme novice s'il est concerné par l'une des situations suivantes :
   - jeune conducteur ayant obtenu son permis de conduire depuis moins de 3 ans ;
   - conducteur occasionnel dont le nom n'apparaît sur aucun contrat d'assurance auto comme conducteur principal ou secondaire depuis plus de 3 ans ;
   - conducteur ayant repassé son permis de conduire pour cause d'annulation de celui-ci ;
   - conducteur étant assuré dans un autre pays.

2. C'est une catégorisation des véhicules en fonction de leur valeur neuf TTC (hors option et remise). Les modalités possibles de cette variable varient entre A et V avec A = faible valeur, V = forte valeur, HC = hors classe. Cette classification des véhicules provient de l'organisme professionnel de Sécurité et Réparation Automobile (SRA).

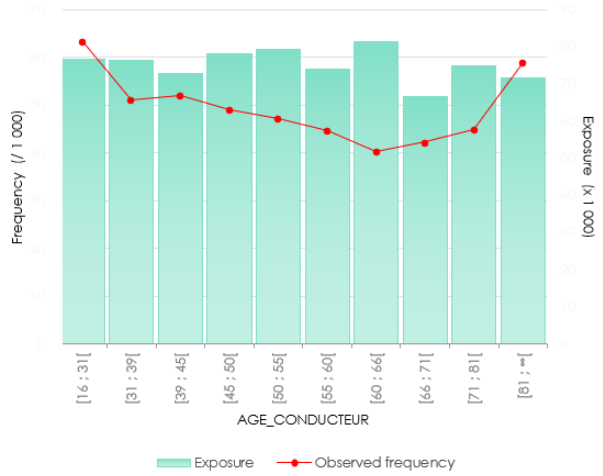

FIGURE 3.4 – *Évolution de la fréquence de sinistres en fonction de l'âge des conducteurs*

D'après la figure 3.4, la fréquence de sinistres a tendance à baisser jusqu'à un certain âge de conducteur "seuil", puis à remonter au delà de ce seuil.

La première partie de ces évolutions (tendance à la baisse) peut s'expliquer par le fait qu'au fil des années, les conducteurs gagnent en expérience de conduite (phénomène de rodage), diminuant ainsi leur probabilité de faire des accidents de la route.

La seconde partie de ces évolutions (tendance à la hausse) est due entre autres aux effets de la vieillesse. En effet, à partir d'un âge très avancé, certains conducteurs rencontrent quelques soucis de santé tels que la baisse de la vue, qui entraînent une augmentation de leur probabilité d'avoir un sinistre.

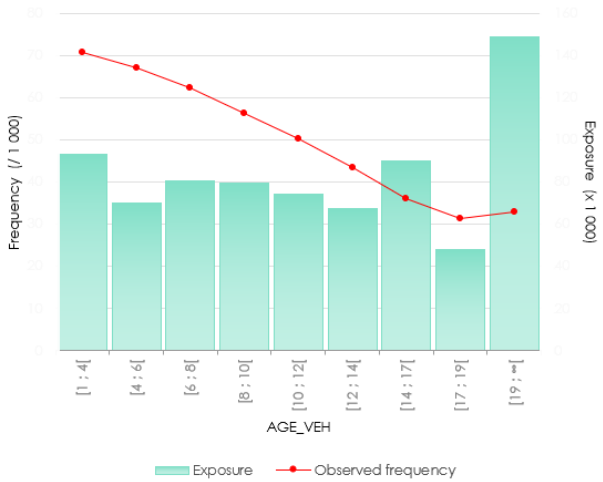

FIGURE 3.5 – *Évolution de la fréquence de sinistres en fonction de l'âge de véhicules des assurés*

D'après la figure 3.5, la fréquence de sinistres baisse quasi-linéairement avec l'augmentation de l'âge du véhicule. Une explication de cette observation est que l'utilisation du véhicule décroît avec le temps. En d'autres termes, plus le véhicule vieilli, moins il est utilisé par son propriétaire.

Traitement des variables sélectionnées

En aval du processus de sélection, il est souvent important de traiter les variables retenues en vue d'accroître leur significativité dans le modèle final.

Au niveau des variables qualitatives :

Ces variables renferment en leur sein plusieurs autres variables. En effet, chacune des modalités de ce type de colonnes constituent elles mêmes des variables explicatives. Ainsi, conserver dans un modèle un facteur qualitatif ayant un grand nombre de modalités peut engendrer une hyper segmentation du risque occasionnant des biais dans l'estimation des indicateurs de sinistralité. Il faut donc, si la situation se présente, procéder à des regroupements de certaines modalités.
Par exemple, la variable catégorie socio-professionnelle de la base interne de l'étude contient une cinquantaine de modalités. Ces dernières sont finalement regroupées en six modalités suivant les niveaux d'utilisation du véhicule et suivant les niveaux d'exposition dans chacune des catégories.

Au niveau des variables quantitatives :

Le fait de conserver une variable quantitative dans un modèle GLM revient à considérer une même monotonie du risque pour chacune des valeurs possible de cette variable. Pourtant, la figure 3.4 présentant l'évolution de la fréquence de sinistres en fonction de l'âge des conducteurs a par exemple montré que le risque pouvait évoluer suivant différents sens de variation sur différents groupes de valeurs d'âges. Partant de cette observation, garder cette variable sous forme quantitative dans le modèle paraît inefficient.

En vue de pallier à ce problème, les variables quantitatives retenues sont toutes discrétisées. La discrétisation se fait soit en quantiles, c'est à dire en plusieurs classes de même effectif (cas des variables : âge du véhicule et puissance du véhicule), soit par *splines* dans le cas d'une variable non strictement monotone telle que l'âge du conducteur. La théorie sur les splines est disponible en annexe B.

Implémentation et évaluation du modèle GLM construit

Post validation des trois premières étapes, le modèle GLM de fréquence de sinistres est implémenté. Le tableau suivant (figure 3.6) présente les indicateurs de la qualité de celui-ci en fonction de ceux obtenu par le modèle trivial (modèle ne contenant aucune variable explicative).

	AIC	BIC	Normalized Gini coefficient	Normalized Gini coefficient (Validation Set)	Root Mean Square Error (Validation Set)
Trivial	100	100	100	100	100
GLM sans var Géo	**98,42**	**98,61**	**1626,00**	**1386,44**	**99,92**

FIGURE 3.6 – *Comparaison des indicateurs de qualité du modèle GLM initial de fréquence de sinistres et du modèle trivial (base 100)*

Le modèle GLM initial de fréquence de sinistres construit est de meilleur qualité que le modèle GLM ne contenant aucune variable explicative. Dans la suite, les variations des indicateurs de la qualité du modèle initial par rapport à ceux des modèles contenant des données télématiques externes intégrées suivant les deux approches de modélisation du risque géographique proposées, constitueront un moyen d'évaluation technique de l'apport de ces données télématiques.

3.4.2 Modélisation des coûts moyens

La modélisation de la fréquence de sinistres conduit à mi-chemin du parcours pour obtenir les valeurs des primes pures. L'autre moitié du chemin s'effectue par la modélisation des coûts moyens de sinistres. Pour rappel, le calcul de la prime pure en fréquence-sévérité est le suivant :

$$Prime\ pure = \mathbb{E}(N|\Omega).\mathbb{E}(Y|\Omega)$$

La modélisation des coûts moyens de sinistres consiste à estimer l'espérance mathématique des coûts de sinistres conditionnellement à des variables explicatives par le biais d'un modèle GLM. On parle de "coûts moyens" car dans ce modèle, le nombre de sinistres est une variable mise en *offset* de manière à modéliser le ratio des coûts par les nombres de sinistres.

Choix de la loi et de la fonction de lien

En assurance IARD, Il est classique d'utiliser pour la modélisation des coûts attritionnels une loi *Gamma*. La fonction de lien utilisée reste le logarithme népérien pour les mêmes raisons citées plus haut.

Sélection supervisée de variables

Les variables sélectionnées doivent être explicatives des coûts de sinistres. La méthode de sélection utilisée est la méthode *stepwise*. La liste des variables retenues est la suivante :

les variables concernant le profil de l'assuré :

- âge du conducteur

les variables concernant le véhicule de l'assuré :

- âge du véhicule ;
- puissance du véhicule ;
- classe de réparation du véhicule

une variable supplémentaire :
l'année d'exercice qui permet de capter les effets de l'inflation sur les coûts de sinistres.

Qualité du modèle

	AIC	BIC	Normalized Gini coefficient	Normalized Gini coefficient (Validation Set)	Root Mean Square Error (Validation Set)
Trivial	100	100	100	100	100
Modele CM	**99,95**	**100,01**	**592,99**	**472,77**	**99,78**

FIGURE 3.7 – *Comparaison des indicateurs de qualité du modèle GLM de coûts moyens et du modèle trivial (base 100)*

Le modèle GLM de coûts moyens construit est globalement de meilleur qualité que le modèle GLM ne contenant aucune variable explicative. Ce modèle GLM de coûts moyens sera utilisé dans le dernier chapitre du mémoire, dans lequel sont présentées des analyses sur les primes pures.

3.4.3 Modélisation du burning cost

La modélisation du *burning cost*[3] est une estimation plus directe de la prime pure :

$$Prime\ pure = \mathbb{E}(NY|\Omega)$$

Elle est brièvement présentée ici parce qu'elle est uniquement utilisée dans une section de l'avant dernier chapitre de ce mémoire. Son principe est d'obtenir les valeurs de primes pures en estimant à partir d'un modèle GLM l'espérance du produit des nombres et des coûts de sinistres sachant des variables explicatives. Cela revient dans la pratique à modéliser le ratio des coûts sur les expositions, résultant du produit de la fréquence de sinistres et des coûts moyens.

3. *burning cost* : ce sont les primes pures historiques qui auraient été réellement payées si le cycle de production assurantiel n'était pas inversé

Choix de la loi et de la fonction de lien

La loi utilisée pour la construction de ce modèle est une distribution *twedie*. Cette distribution a l'avantage d'être très adaptée pour modéliser des variables continues contenant une quantité importante de zéro. Ici encore, la fonction de lien utilisée est le logarithme népérien de manière à rendre le modèle multiplicatif, donc plus simple à comprendre et à mettre en production.

Sélection supervisée de variables

Les variables sélectionnées doivent être à la fois explicatives des nombres de sinistres et des coûts de sinistres. La méthode de sélection utilisée est la méthode *stepwise*. La liste des variables retenues est la suivante :

les variables concernant le profil de l'assuré :

- âge du conducteur ;
- catégorie socio-professionnelle ;
- coefficient novice

les variables concernant le véhicule de l'assuré :

- âge du véhicule ;
- puissance du véhicule ;
- classe de réparation du véhicule ;
- valeur du véhicule ;

une variable supplémentaire :
l'année d'exercice qui permet de capter les effets de l'inflation sur les coûts de sinistres.

Qualité du modèle

	AIC	BIC	Gini coefficient	Normalized Gini coefficient	Normalized Gini coefficient (Validation Set)	Root Mean Square Error (Validation Set)
Trivial	100	100	100	100	100	100
Tweedie sans var géo	**99,37**	**99,45**	**1022,30**	**1022,30**	**1271,71**	**99,94**

FIGURE 3.8 – *Comparaison des indicateurs de qualité du modèle GLM initial de burning cost et du modèle trivial (base 100)*

Le modèle GLM initial du *burning cost* construit est de meilleur qualité que le modèle GLM ne contenant aucune variable explicative. Cette modélisation du *burning cost* est utilisée à la fin du chapitre 4 en vue d'évaluer l'impact d'une combinaison de données *Open data* et des données télématiques dans la modélisation du risque géographique.

Chapitre 4

Modélisation du risque géographique 1 : approche naïve

« Il faut beaucoup de naïveté pour faire de grandes choses. »
René Crevel, L'esprit contre la raison

L'idée générale de cette approche est d'intégrer directement dans la structure tarifaire de l'assureur, certaines données télématiques géospatialisées, minutieusement sélectionnées. Cette intégration s'effectue précisément à l'intérieur du modèle GLM de fréquence de sinistres construit au chapitre précédent.

L'objectif est de comparer les performances statistiques de ces différents modèles (ancien modèle sans information géographique et nouveaux modèles contenant des informations télématiques géographiques) en vue d'en déduire un premier avis sur l'apport des données télématiques dans la modélisation du risque géographique.

4.1 Présentation théorique de l'approche naïve

L'approche naïve consiste à matérialiser le risque géographique à l'intérieur du modèle GLM de fréquence de sinistres en y ajoutant directement quelques unes des variables télématiques géographiques de la base externe *Smart Road Data*.

Le modèle GLM de survenance de sinistres devient donc :

$$\mathbb{E}(\tilde{N}|X_1,X_2,...,X_p,X_{p+1},...,X_{p+q}) = \exp(\beta_0).\prod_{k=2}^{p}\exp(\beta_k X_k)\prod_{j=p+1}^{p+q}\exp(\beta_j X_j)$$

avec les $(X_j)_{p+1\leq j\leq p+q}$ représentant une sélection de q ($q \geq 1$) variables télématiques géographiques de la base externe.

Cela revient à considérer que l'impact des comportements dans les différentes zones géographiques sur la fréquence de sinistres (et par conséquent sur la prime pure) RCM peut être mathématiquement formalisé par une combinaison de variables géographiques reflétant les habitudes de conduite dans ces différentes zones.

Un des avantages de cette approche est qu'elle permet d'observer très facilement les effets marginaux de chacune des q variables télématiques sur la fréquence de sinistres. De plus, elle favorise l'étude des interactions entre ces données télématiques géographiques et les variables internes de l'assureur. Ces effets d'interactions peuvent dans bien de cas représenter de nouvelles informations dont la prise en compte pourrait engendrer une amélioration du modèle.

4.2 Sélection supervisée et analyses des variables

4.2.1 Sélection supervisée des variables

L'approche naïve implique une sélection de variables beaucoup plus rigide, au risque de trop complexifier le modèle de départ.

Cette sélection de variables s'effectue cette fois sur la base obtenue après la fusion des bases interne et externe. Dans cette base commune, il y a 43 variables pouvant potentiellement intégrer le nouveau modèle à construire, à savoir : les 6 variables de la base interne déjà présentes dans le modèle initial et les 37 variables de la base *Smart Road Data* présélectionnées de manière non supervisée.

Première sélection par l'algorithme LASSO

Malgré les transformations et la sélection non supervisée opérées sur les données de la base *Smart Road Data*, les 37 variables restantes ne sont pas totalement décorrelées.

De ce fait et dans l'optique d'en sélectionner quelques unes, il faudrait utiliser une méthode de sélection qui pénaliserait les facteurs corrélés. L'algorithme LASSO paraît très adapté pour la réalisation de cette tâche. En effet, en pénalisant les coefficients du modèle, le LASSO diminue la probabilité de conserver deux ou plusieurs variables apportant pratiquement les mêmes informations.

la sélection des variables débute donc par l'implémentation de l'algorithme LASSO s'appuyant sur une validation croisée.

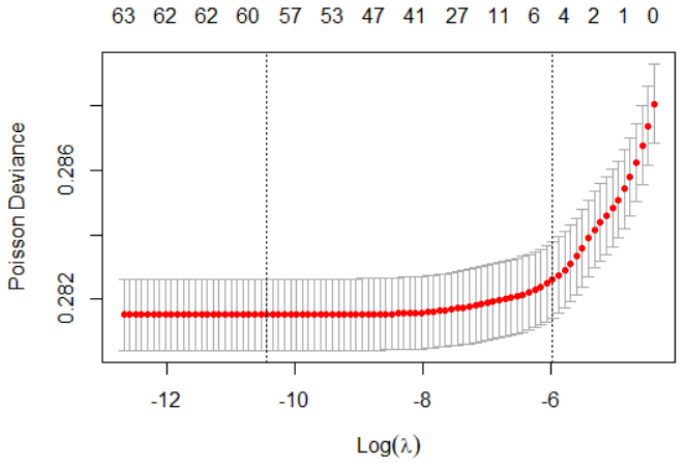

FIGURE 4.1 – *Détermination du paramètre λ par validation croisée*

La figure 4.1 présente l'évolution de la déviance en fonction des valeurs du $\log(\lambda)$. La déviance est minimale pour une valeur de $\log(\lambda)$ avoisinant -10 (première barre verticale en pointillée). Cependant, pour cette valeur, le nombre de variables à retenir est très élevé. Toutefois, il est possible d'observer que les valeurs de la déviance restent quasiment constantes pour les valeurs de $\log(\lambda)$ comprises entre -10 et -8. Au delà de -8, les valeurs de la déviance deviennent fortement croissantes.

La valeur choisie pour le paramètre est donc : $\lambda = \exp(-8)$. Pour cette valeur de λ, 20 variables sur les 43 sont sélectionnées. Ce lot de 20 variables se compose des 6 variables internes initialement présentes dans le modèle et de 14 variables provenant de la base *Smart Road Data*. Cette sélection de 20 variables étant encore conséquente, il serait bénéfique de l'affiner davantage.

Deuxième sélection par la méthode stepwise

La méthode *stepwise* est implémentée sur ces 20 variables restantes en s'appuyant sur une validation croisée et en considérant comme critère de qualité l'AIC. Finalement, cette méthode conduit à ne conserver que 13 variables sur les 20 précédemment sélectionnées. Parmi ces 13 variables se retrouvent les 6 variables internes présentes dans le modèle GLM de départ de la fréquence de sinistres et 7 variables issues de la base externe *Smart Road Data*. La liste des 13 variables retenues est dressée dans le tableau de la figure 4.2.

Nom de la variable	Description
AGE_CONDUCTEUR	Age du conducteur
AGE_VEH	Age du véhicule
CLASSE_SRA	Classe SRA
COEFF_NOVICE	Coefficient novice
CSP	Categorie socio-professionnelle
PUISSANCE	Puissance du véhicule
s3_anticipation	Score global de freinage
s3_pace	Score global d'allure de la conduite
total_km	Total des kilomètres parcourus sur la commune
roulage_moyen	Fréquence de conduite
WET_weather_part_of_km	Taux de kilomètres parcourus en temps humide
part_of_urban	Taux de route urbaine dans la commune
part_of_CURVE	Taux de virage dans la commune

FIGURE 4.2 – *Liste des 13 variables sélectionnées pour l'approche naïve*

Les sept variables issues de la base *Smart Road Data* se composent de deux scores de comportement, de deux variables concernant l'usage général du véhicule, d'une variable concernant l'usage du véhicule suivant le contexte météorologique et de deux variables en lien avec les types de routes à l'intérieur des communes.

Avant d'intégrer ces variables dans un nouveau modèle GLM de fréquence de sinistres, deux types d'analyse sont menés.

Le premier type consiste à expliquer l'évolution de la fréquence de sinistres en fonction de ces nouvelles variables. il permet de connaître a priori les effets marginaux de ces dernières dans le nouveau modèle.
Le second consiste à vérifier que les informations apportées par ces nouvelles variables n'ont pas déjà été apprises par le modèle initial. Ce type permet de limiter la redondance d'informations à l'intérieur du nouveau modèle.

Ces deux types d'analyse peuvent être conjointement réalisés à partir d'un même graphique. La sous-section suivante présente ces différents types d'analyse pour chacune des sept variables de la base externe sélectionnées.

4.2.2 Analyses des variables retenues par sélection supervisée

Toutes les analyses de variables présentées dans cette sous-section se basent sur une vision géographique de la fréquence de sinistres. Ainsi, dans un premier temps, l'analyse de la fréquence moyenne de sinistres suivant les différentes communes est effectuée.

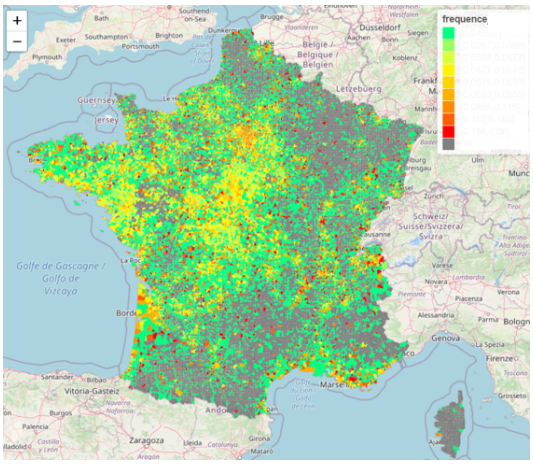

FIGURE 4.3 – *Cartographie de la fréquence moyenne de sinistres observée par commune*

Sur la cartographie ci-dessus (Figure 4.3), la palette de couleur passe du vert représentant les communes ayant une fréquence moyenne de sinistres très faible au rouge symbolisant les communes ayant une fréquence moyenne de sinistres très élevée. Les zones en gris sont les communes non couvertes par l'assureur.

Cette cartographie met en lumière la non-homogénéité de la fréquence de sinistres RCM sur le territoire. En effet, cette dernière est relativement élevée dans certaines zones telles que l'Île-de-France, la Gironde, les Bouches du rhône et l'Auvergne-Rhone-Alpes, et est plus faible dans d'autres zones telles que la Bretagne, la Normandie et le Nord-pas-de-calais.

Sur les graphiques qui suivront,

- la courbe de couleur rouge représente la fréquence moyenne de sinistres **observée** sur chacune des modalités de la variable analysée ;
- la courbe de couleur bleue représente la fréquence moyenne de sinistres **estimée** par le modèle GLM **initial** (sans variables externes) sur chacune des modalités de la variable analysée ;
- les barres de couleur verte représentent les niveaux d'exposition sur chacune des modalités de la variable analysée.

Les valeurs des variables analysées sont normalisées sur l'intervalle $[0; 1]$.

Score global de freinage

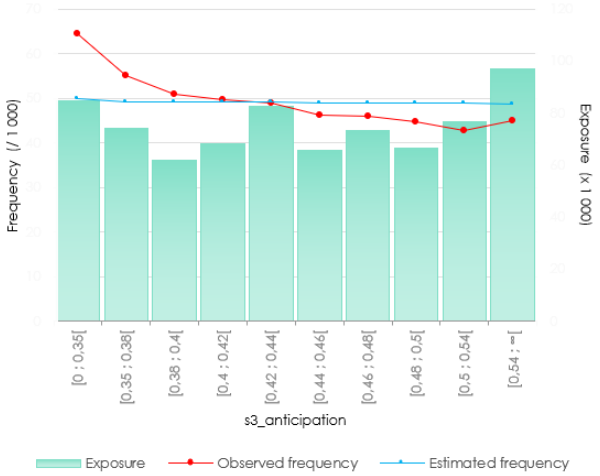

FIGURE 4.4 – *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Score global de freinage*

D'après la figure 4.4, la fréquence de sinistres observée décroît lorsque la note de freinage s'améliore. Cela peut s'expliquer par le fait qu'une bonne manière de freiner limite le risque de faire des accidents de la circulation.

Score global d'allure de conduite

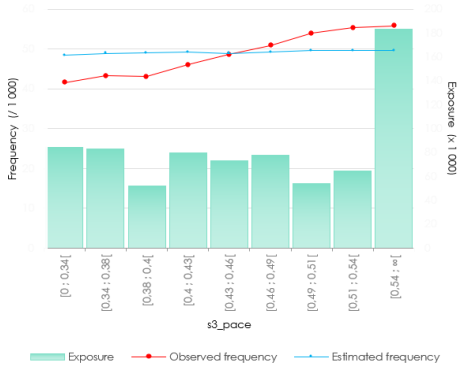

FIGURE 4.5 – *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Score global d'allure de conduite*

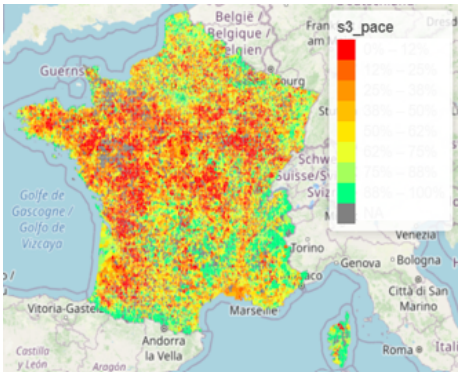

FIGURE 4.6 – *Cartographie du score global d'allure de conduite*

La figure 4.5, montre que la fréquence de sinistres observée croît lorsque le score global d'allure de conduite augmente. La compréhension de cette observation nécessite la comparaison de la cartographie de la fréquence de sinistres et de celle du score étudié. En effet, en comparant ces deux cartes, il est possible de remarquer que sur les zones

où la fréquence de sinistres est élevée telles que Paris, Bordeaux et Marseille, le score global d'allure de conduite est aussi élevé. Parallèlement, sur les zones où la fréquence de sinistres est faible telles que dans le Finistère, le score global d'allure de conduite est aussi faible.

Total de kilomètres parcourus

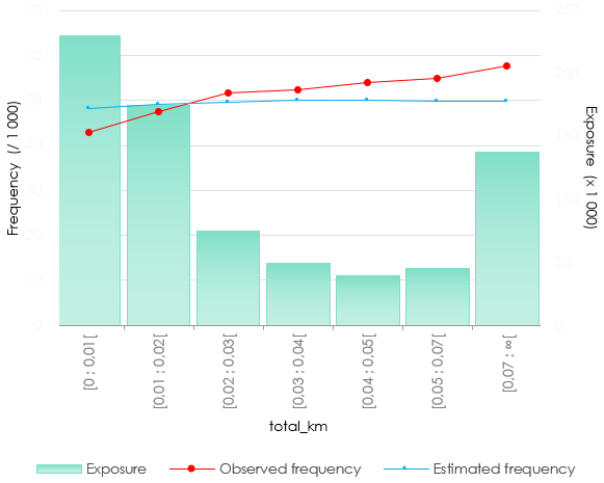

FIGURE 4.7 – *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Total de kilomètres parcourus*

D'après la figure 4.7, la fréquence de sinistres observée croit lorsque le total de kilomètres parcourus s'accentue. Cette observation paraît objective puisque plus on parcourt de kilomètres avec son véhicule, plus la probabilité de faire un accident de la route augmente.

Fréquence de conduite

La fréquence de conduite est le résultat du ratio des nombres de trajets sur une commune et des nombres de véhicules qui sont passés par cette commune :

$$roulage\_moyen = \frac{nb\_trips}{nb\_vid}$$

Par exemple, la fréquence de conduite pour une commune sur laquelle il y a eu dix trajets effectués par cinq véhicules vaut deux. Cela signifie qu'en moyenne un véhicule effectue deux trajets sur la commune en question.

La figure 4.8 montre que la fréquence de sinistres observée et la fréquence de conduite évoluent dans le même sens. Cela prouve que la probabilité de faire des accidents est fonction du niveau d'utilisation des véhicules.

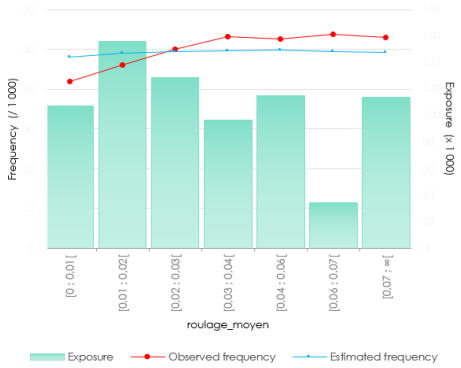

FIGURE 4.8 – *Analyse des valeurs observées et estimées de la fréquence de sinistres sur la Fréquence de conduite*

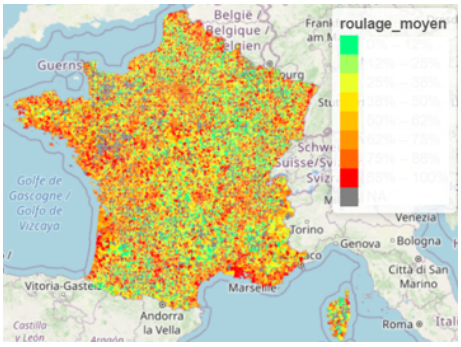

FIGURE 4.9 – *Cartographie de la fréquence de conduite*

Taux de parcours en temps humide

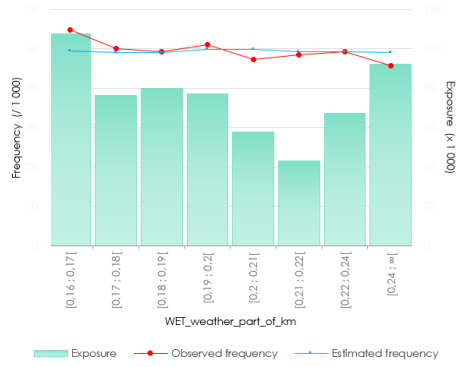

FIGURE 4.10 – *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Taux de parcours en temps humide*

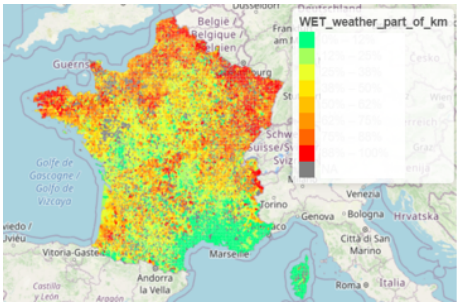

FIGURE 4.11 – *Cartographie du taux de parcours en temps humide*

D'après la figure 4.10, plus les taux de parcours en temps humide augmente, plus la fréquence de sinistres a tendance à baisser. Cette observation peut paraître d'emblée contre-intuitives. Cependant, elle peut s'expliquer par le fait qu'en temps humide les conducteurs ralentissent leur vitesse et sont beaucoup plus vigilants.

Taux de routes urbaines

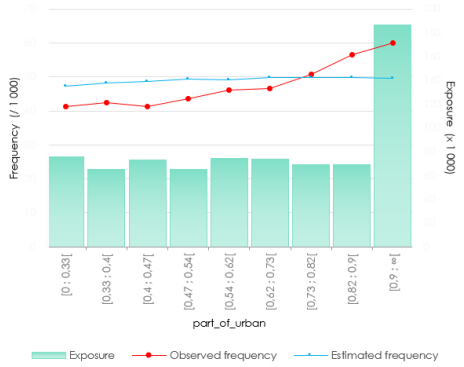

FIGURE 4.12 – *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Taux de routes urbaines*

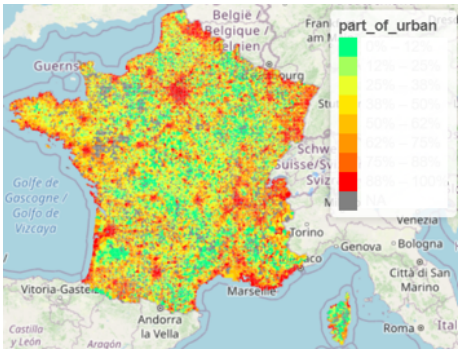

FIGURE 4.13 – *Cartographie du taux de routes urbaines*

La figure 4.12 montre que la fréquence de sinistres et les taux de routes urbaines augmentent simultanément. Cette observation peut s'expliquer en analysant la cartographie 4.13. Sur cette dernière, les zones où les taux de routes urbaines sont élevés sont des zones où se trouvent des périphériques. De plus, sur les périphériques il est très courant d'observer un fort taux d'accidentalité pouvant expliquer ces hausses de fréquences de sinistres.

Taux de virages

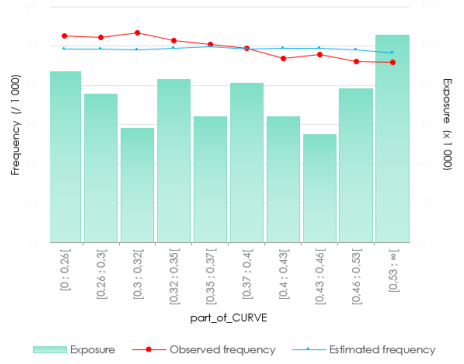

FIGURE 4.14 – *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Taux de virages*

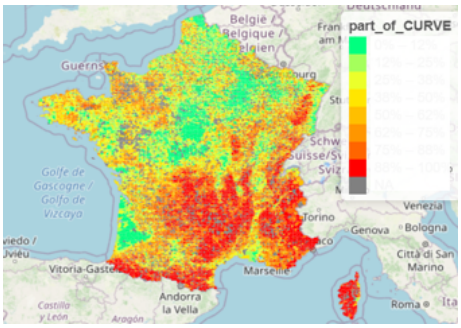

FIGURE 4.15 – *Cartographie du taux de virages*

D'après la figure 4.14, la fréquence de sinistres décroît lorsque les taux de virages à l'intérieur des communes augmentent. Cela peut s'expliquer par les règles strictes de conduite dans les virages, entre autres le ralentissement de la vitesse.

En complément de tout ce qui a été présenté, il est possible d'observer sur les graphiques précédents que pour chacune des variables étudiées, les estimations de fréquences de sinistres du modèle initial (courbe de couleur bleue) ne suivent pas les tendances des valeurs observées. Cela démontre que ces nouvelles variables sélectionnées apportent des informations encore inconnues de ce modèle.

Après avoir effectué toutes ces analyses validant la pertinence de la prise en compte de ces variables, un nouveau modèle GLM de fréquence de sinistres est construit avec cette fois les 13 variables sélectionnées. Ce nouveau modèle est appelé symboliquement **"modèle naïf"** en référence à l'approche ayant permis sa construction. Toutefois, ce modèle naïf conserve une structure linéaire classique. La section suivante consiste à étudier la possibilité d'introduire des interactions à l'intérieur de celui-ci. La structure de ce modèle deviendra donc polynomiale. Cette complexification peut, suivant certains critères améliorer la qualité du modèle.

4.3 Étude des interactions entre les variables sélectionnées

Dans la section précédente, les effets marginaux des variables télématiques retenues sur la fréquence de sinistres ont été observés. Cependant, l'effet marginal d'une variable explicative sur une variable à expliquer peut, quelque fois, varier en fonction des valeurs prises par une ou plusieurs autres variables explicatives. Dans ce genre de situations, les variables explicatives impliquées sont dites en interaction. La prise en compte de cette interaction à l'intérieur d'un modèle GLM peut participer à l'amélioration de sa qualité.

Cette section est consacrée à l'étude des possibles interactions entre les variables sélectionnées. Elle s'intéresse aux interactions entre les données télématiques elles-mêmes et aux interactions entre les données télématiques et les données internes. Le but est d'exploiter davantage les informations apportées par les données télématiques sélectionnées. Cette étude des interactions s'effectue suivant différentes étapes :

1. La détection des interactions à partir d'un arbre de décision ;
2. Le test de la significativité des interactions détectées ;
3. Le choix de(s) l'interaction(s) significative(s) à intégrer dans le modèle.

4.3.1 Théorie des arbres de décision CART

Un arbre de décision est une suite de partitions de plus en plus fines de l'ensemble de tous les individus observés vis à vis d'une variable à expliquer.

Introduits dans la littérature scientifique par Léo BREIMAN en 1984, les arbres de décision de type CART sont un moyen simple mais efficace de visualisation de la décomposition d'une variable à expliquer à partir de plusieurs variables explicatives.

Le fonctionnement de l'algorithme CART est le suivant : partant de la base d'apprentissage complète appelée **racine**, l'arbre se construit en séparant la population initiale en plusieurs sous-groupes. A chaque itération, une séparation donne naissance à un **noeuds-fils** jusqu'à ce qu'un critère d'arrêt soit vérifié. Une fois ce critère vérifié, une partition finale est alors obtenue et les éléments de cette partition finale constituent une **feuille** de l'arbre.

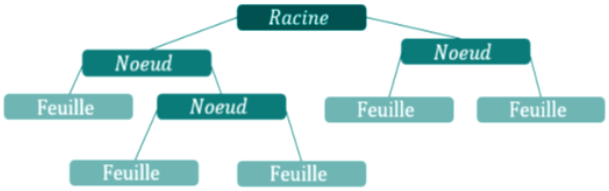

FIGURE 4.16 – *Illustration schématique d'un arbre CART*

Critère de division

Les différentes subdivisions effectuées pendant la construction d'un arbre CART vise à créer des groupes homogènes au sein des différentes feuilles. Pour atteindre cet objectif, l'algorithme cherche à chacune de ses itérations à minimiser l'erreur d'approximation. Il cherche donc le couple (j^*,s^*) tel que :

$$(j^*,s^*) = \underset{(j,s)}{argmin}\{\sum_{x_i \in \{X|X_j<s\}} (y_i - \hat{y}\{X|X_j < s\})^2 + \sum_{x_i \in \{X|X_j \geq s\}} (y_i - \hat{y}\{X|X_j \geq s\})^2\}$$

Critère d'arrêt

Au cours de sa construction, un arbre peut s'étendre en segmentant de plus en plus les données en partitions jusqu'à ce que chacune des observations constitue une feuille. Un arbre qui atteint cette profondeur est sujet inéluctablement à des effets de sur-apprentissage (sur-ajustement du modèle aux données d'apprentissage impliquant une moins bonne capacité de généralisation). Afin d'éviter cette situation, il convient de définir des conditions d'arrêt de l'algorithme. ces conditions peuvent consister à définir entre autres :

- Le nombre minimal d'individus par feuille ;
- La profondeur maximale de l'arbre.

Avantages et inconvénients

L'avantage majeur des arbres de décision CART est leur facilité d'interprétation. En effet, il est possible de représenter l'arbre construit, ce qui permet de comprendre le mécanisme interne d'explication de la variable cible par les variables explicatives. C'est donc un outil d'une grande utilité dans le cadre d'une analyse supervisée multidimensionnelle. Toutefois, cet algorithme est en général relativement instable pour la régression.

4.3.2 Application du CART pour la détection d'interactions

La figure 4.17 présente l'arbre de décision de la fréquence de sinistres en fonctions des variables explicatives sélectionnées pour l'approche naïve. Sur cet arbre, chaque subdivision constitue une interaction introduite par le modèle. Par exemple, à la profondeur 1, il y a interaction entre la variable âge du véhicule (variable interne) et le score général de freinage (variable externe). Pour rappel, cette étude ne considère pas les interactions entre deux variables internes. La liste des interactions détectées est disponible dans le tableau sur la figure 4.18.

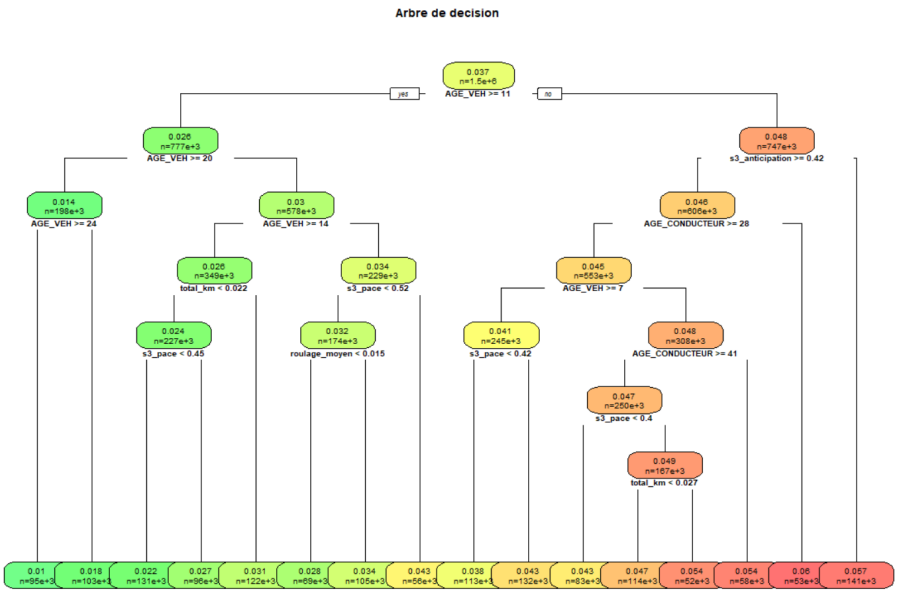

FIGURE 4.17 – *Arbre de décision de la fréquence de sinistres en fonction des variables sélectionnées (approche naïve)*

Intercations Détectées
AGE_VEH x s3_anticipation
AGE_CONDUCTEUR x s3_anticipation
AGE_VEH x s3_pace
AGE_CONDUCTEUR x s3_pace
AGE_VEH x total_km
s3_pace x total_km
s3_pace x roulage_moyen

FIGURE 4.18 – *Liste des interactions détectées à partir de l'arbre CART*

4.3.3 Test de significativité et sélection des interactions

Une fois les différentes interactions détectées, une étude est menée afin d'évaluer leur significativité. La raison de cette étude est que certaines interactions peuvent souvent ne pas être significatives à l'intérieur d'un modèle. Soit, cela est dû au fait que l'information apportée par ces croisements de variables est déjà connue du modèle, soit parce que ces interactions sont "artificielles". Une interaction "artificielle" est une fausse interaction créée par la présence de collinéarités subsidiaires entre deux ou plusieurs variables d'un modèle.

Il existe différentes méthodes pour tester la significativité d'une interaction de variables à l'intérieur d'un modèle. Elles peuvent être sophistiquées et automatisées (*S. BUCCI* cf.[4]), comme elles peuvent aussi être plus simplistes et manuelles. Ici, c'est une méthode manuelle qui est utilisée afin d'effectuer ces tests de significativité du fait du nombre raisonnable d'interactions détectées. Cette méthode s'appuie sur le test de Wald (déjà présenté au chapitre 3). Le principe reste le même sauf que l'élément testé est ici le croisement de variables.

Soit $X_1 * X_2$ l'interaction détectée et rajoutée à l'intérieur du modèle. Ce dernier se réécrit :

$$g(\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_1 * X_2$$

Les hypothèses du test sont $H_0 : \beta_k = 0$ contre $H_1 : \beta_k \neq 0$. Si la probabilité $1 - \alpha$ de rejeter H_0 est élevée (de l'ordre de 95%) alors l'interaction $X_1 * X_2$ est significative à l'intérieur du modèle.
La méthode manuelle consiste donc à répéter ce test pour chacune des interactions détectées et à conserver celles qui sont significatives.

A l'issue de ce processus, les interactions révélées significatives pour le nouveau modèle de fréquence de sinistres construit (modèle naïf) sont les suivantes :

- **(I_1) : Âge du véhicule x Score global d'allure de conduite** ;
- **(I_2) : Âge du véhicule x Total de kilomètres parcourus**.

Cependant, intégrer deux interactions dans un même modèle GLM peut le rendre très complexe (trop de paramètres à estimer, difficulté dans la mise en production du modèle). Il est donc nécessaire de faire un choix entre ces deux interactions significatives. Ce choix peut être guidé par l'analyse de la surface de réponse des modèles contenant chacune de ces deux interactions. En effet, l'ajout d'interactions dans un modèle GLM contribue à la déformation de la linéarité de sa surface de réponse. Ainsi, plus la déformation de cette surface de réponse est prononcée, plus l'interaction apporte une quantité importante de nouvelles informations au modèle.

Analyse de l'interaction (I_1)

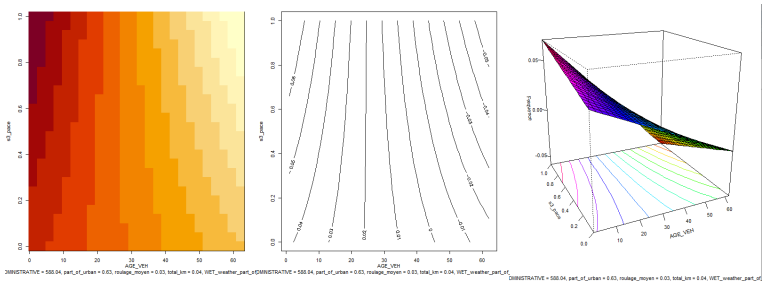

FIGURE 4.19 – *Représentation de la surface de réponse du GLM naïf de fréquence de sinistres contenant l'interaction (I_1)*

Les graphiques de la figure 4.19 montrent que l'ajout de l'interaction I_1 à l'intérieur du modèle GLM naïf déforme de manière relativement forte la linéarité de départ de la surface de réponse de ce modèle. A partir de cette figure, il est possible de lire les informations apportées par cette interaction au modèle. Par exemple, lorsque la note générale d'allure de conduite diminue et que simultanément l'âge du véhicule augmente, cela induit une baisse de la fréquence de sinistres (effet diagonale allant de l'extrémité supérieure gauche à l'extrémité inférieure droite sur les graphiques de contour).

Analyse de l'interaction (I_2)

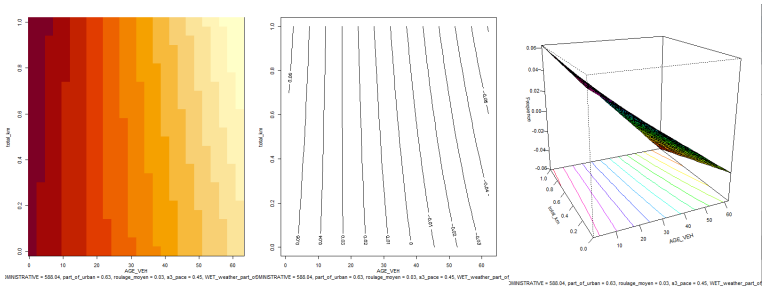

FIGURE 4.20 – *Représentation de la surface de réponse du GLM naïf de fréquence de sinistres contenant l'interaction (I_2)*

Comparativement aux graphiques de la figure 4.19, ceux de la figure 4.20 montrent que l'ajout de l'interaction I_2 à l'intérieur du modèle GLM naïf déforme de manière moins prononcée la linéarité de départ de la surface de réponse de ce modèle par rapport à l'ajout de l'interaction I_1.

Il ressort de cette analyse comparative, que l'interaction à intégrer dans le modèle GLM naïf est l'interaction (I_1) **: Âge du véhicule x Score global d'allure de conduite**. Un troisième modèle GLM de fréquence de sinistres est donc construit en intégrant l'interaction I_1 à l'intérieur du modèle naïf.

4.4 Première évaluation de l'apport des données télématiques contextualisées

L'approche naïve a consisté à matérialiser le risque géographique à l'intérieur du modèle GLM de départ de fréquence de sinistres en lui ajoutant des informations télématiques géospatialisées et contextualisées pour certaine d'entre elles. A la fin de la mise en application de cette approche, trois modélisations de la fréquence de sinistres RCM sont donc à disposition : le modèle GLM initial, le modèle GLM naïf et le modèle GLM naïf avec interaction. La comparaison de la qualité statistique de ces trois modèles constitue un premier moyen d'évaluation de l'apport des données télématiques dans la modélisation du risque géographique. Ainsi, dans cette dernière section, les variations des performances du modèle initial, par rapport à celles des nouveaux modèles contenant les données télématiques sont analysées. Une comparaison des cartographies des fréquences de sinistres prédites par ces différents modèles est présentée en complément des analyses sus-mentionnées.

4.4.1 Comparaisons des performances statistiques des modèles : initial et naïfs

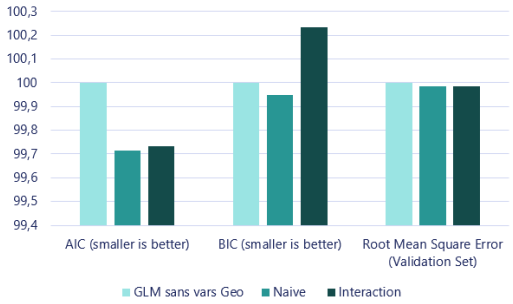

FIGURE 4.21 – *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle initial sans variables géographiques)*

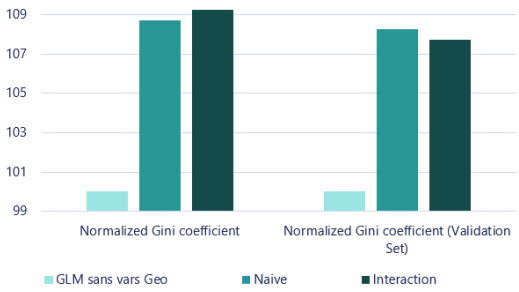

FIGURE 4.22 – *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle initial sans variables géographiques)*

Les graphiques des figures 4.21 et 4.22 mettent en lumière une amélioration générale des indicateurs de la qualité statistique du modèle initial (sans variables géographiques) lorsqu'on lui fournit des informations télématiques.

Au niveau des critères AIC et BIC :

Le modèle GLM naïf a une meilleure qualité d'apprentissage en terme d'AIC et de BIC que le modèle initial. Le modèle naïf conserve cette très bonne qualité d'apprentissage en terme d'AIC lorsqu'on y introduit les effets d'interactions entre l'âge du véhicule et le score général d'allure de conduite. Toutefois, la complexité induite par ces effets

d'interaction contribue significativement à la dégradation de la valeur du BIC de ce modèle.

Au niveau des indices de Gini :

Il a été décidé de ne présenter que les indices de Gini normalisés du fait de la proximité de leurs valeurs et de celles des indices de Gini standards. La figure 4.22 montre qu'en présence d'informations télématiques géospatialisées le modèle discrimine mieux la fréquence de sinistres. Cela s'observe via l'amélioration de plus de 7% de la valeur du Gini normalisé lors du passage du modèle initial aux nouveaux modèles naïf et naïf contenant une interaction.

En terme d'erreurs de prédiction RMSE, l'amélioration observée est moins accentuée qu'au niveau des précédents indicateurs. Cependant, ne regarder que l'erreur de prédiction globale des modèles afin d'évaluer la qualité de leur précision limite la prise en compte de l'aspect géographique dans cette évaluation. En effet, du point de vue géographique, un modèle dont les valeurs prédites conservent les mêmes tendances que les valeurs observées sur les différentes zones est préférable à un modèle qui a des tendances de prédictions complètement différentes de celles des valeurs observées, quand bien même ces deux modèles décrits auraient des valeurs très proches de RMSE. Cela introduit une notion d'erreurs géographiques. L'évaluation de cette dernière peut s'effectuer en comparant les cartographies des prédictions des différents modèles à celle des valeurs observées.

4.4.2 Comparaison visuelle des prédictions des modèles : initial et naïf

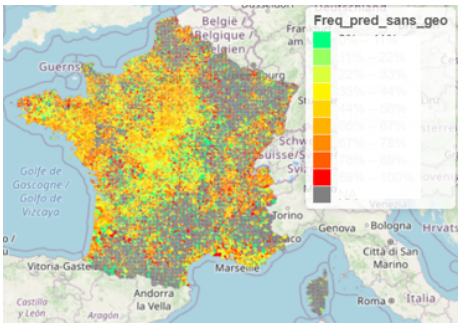

FIGURE 4.23 – *Cartographie des prédictions de fréquences de sinistres du modèle initial sans variables géographiques*

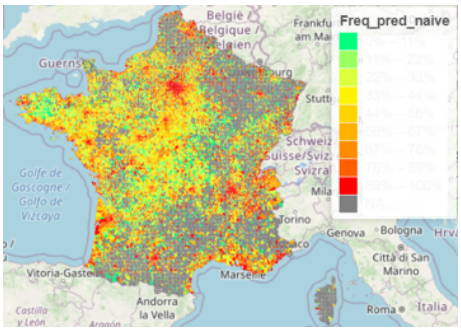

FIGURE 4.24 – *Cartographie des prédictions de fréquences de sinistres du modèle naïf*

En comparaison à la cartographie des fréquences de sinistres observées (figure 4.3), les cartographies 4.23 et 4.24 montrent une surestimation des fréquences de sinistres prédites par les modèles : initial et naïf.
Toutefois, les prédictions du modèle initial semblent assez homogènes sur les zones cou-

vertes par l'assureur contrairement aux valeurs observées des fréquences de sinistres. Pourtant, au niveau du modèle naïf, les prédictions respectent les tendances des valeurs observées des fréquences de sinistres. En effet, les prédictions du modèle naïf sont plus hétérogènes avec des valeurs accentuées sur des zones telles que l'Île-de-France, la Gironde, les Bouches du rhône et l'Auvergne-Rhone-Alpes où la fréquence de sinistres observée est élevée.

Que retenir des chapitres 3 et 4 ?

Les chapitres 3 et 4 ont présenté une première évaluation de l'apport des données télématiques dans la modélisation du risque géographique. Cette évaluation s'est appuyée sur une approche nommée naïve, consistant à intégrer directement des facteurs télématiques externes dans les équations tarifaires de l'assureur.

Il ressort de cette première évaluation que la modélisation naïve du risque géographique par une combinaison de variables télématiques permet d'améliorer la qualité statistique et la précision des modèles de tarification. Les habitudes de conduite dans les différentes zones seraient donc des facteurs permettant de comprendre le risque géographique en assurance automobile.

Cependant, dans la pratique, la mise en production de cette approche peut être assez difficile pour un assureur. En effet, rajouter plusieurs variables à la structure tarifaire de base de ce dernier constitue une trop grande modification à gérer.

Afin de palier cette contrainte opérationnelle, l'idéal serait de synthétiser toutes les informations géographiques apportées par ces données télématiques externes en une unique variable. C'est le principe même de l'approche qui est développée au chapitre suivant.

Chapitre 5

Modélisation du risque géographique 2 : Zonier

« L'innovation systématique requiert la volonté de considérer le changement comme une opportunité. »

Peter Drucker, Innovation et entreprenariat

La seconde approche de modélisation du risque géographique proposée dans ces travaux est la construction d'un zonier. Un zonier est une classification des différentes zones géographiques suivant les niveaux de risques qui leur sont rattachés. Le zonier est l'outil par excellence utilisé par les assureurs pour capter l'impact géographique de leur portefeuille sur la sinistralité.

La construction d'un zonier peut s'effectuer de manière traditionnelle (sans apport de données externes) ou de manière moderne (avec apport de données externes). Dans ce chapitre, ces deux méthodes sont exploitées. Dans un premier temps, un zonier est construit suivant la méthode traditionnelle. Dans un second temps, un autre zonier est construit suivant la méthode moderne par apprentissage statistique des données télématiques externes. La comparaison des performances des modèles contenant ces deux zoniers, constitue un nouveau moyen d'évaluer l'apport des données télématiques dans la modélisation du risque géographique.

Toutefois, au cours de ces dernières années, plusieurs travaux ont révélé l'utilité des données *Open data* dans la construction d'un zonier suivant la méthode moderne. Ainsi, à la fin de ce chapitre, une comparaison de performances est effectuée entre un modèle contenant un zonier construit uniquement par apprentissage statistique de données *Open data* et un autre modèle contenant un zonier construit par apprentissage statistique d'une combinaison de données *Open data* et de données télématiques. Cette comparaison consolidera les précédentes évaluations de la pertinence des données télématiques dans la modélisation du risque géographique.

5.1 Présentation des étapes de la construction d'un zonier

La méthodologie de construction d'un zonier n'est pas universelle. Elle peut varier suivant les attentes du modélisateur. Dans le cadre de ces travaux, deux méthodologies sont utilisées : une méthodologie traditionnelle et une méthodologie moderne. Ici, ces deux méthodologies se basent sur une hypothèse de départ commune qui est que les erreurs des modèles initiaux (sans données géographiques) sont dues en partie au manque d'informations géographiques dans leur apprentissage. Cette hypothèse implique que les erreurs encore appelées résidus, renferment un signal géographique inexpliqué.

La première étape de la construction du zonier est donc le calcul des résidus et leur agrégation à la maille géographique (INSEE pour cette étude). La seconde étape consiste à traiter ces résidus agrégés. Ce traitement du signal géographique est traditionnellement effectué par un lissage géospatial des résidus agrégés. Cependant, une seconde méthode plus moderne consiste à expliquer d'abord ce signal géographique contenu dans les résidus à partir de données externes et par le biais d'un modèle d'apprentissage statistique puis à appliquer le lissage géospatial sur les prédictions de ce modèle. Une fois ces différents traitements effectués, la dernière étape consiste à regrouper les valeurs des résidus traités en classes de risque afin d'obtenir le zonier.

Ce zonier est finalement intégré au modèle initial (duquel ont été extraits les résidus) en tant que variable explicative de sorte à combler la carence en informations géographiques de ce dernier.

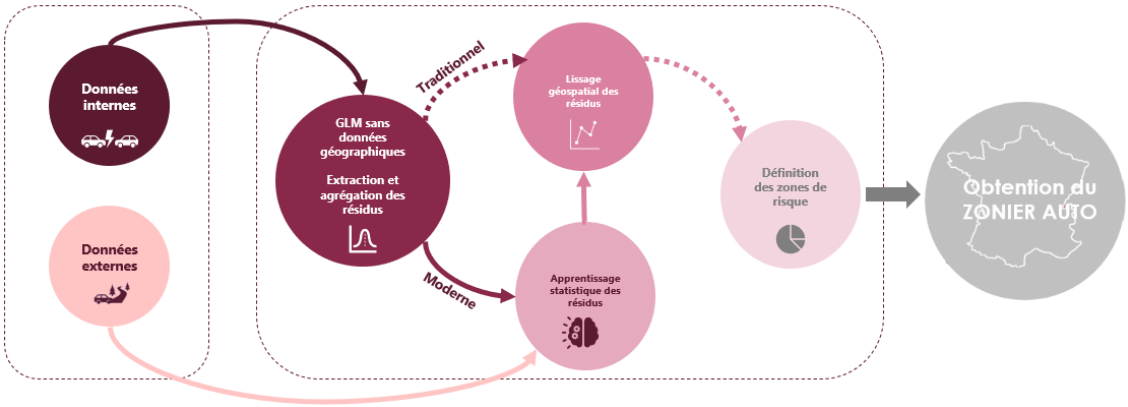

FIGURE 5.1 – *Méthodologies de construction d'un zonier*

5.2 Calcul et agrégation géographique des résidus

Un résidu est une quantification de l'erreur de prédiction d'un modèle (*McCULLAGH et NELDER* cf.[13]). Dans le cas du modèle GLM initial (sans donnée géographique) de fréquence de sinistres, c'est une mesure de l'écart entre la fréquence de sinistres observée

et la fréquence de sinistres prédite.

Soit y_i la valeur de la fréquence de sinistres observée sur la i ième police d'assurance et $\hat{y}_i$ la prédiction de celle-ci par le modèle initial de fréquence de sinistres.

La manière la plus simple et la plus intuitive d'obtenir l'écart entre y_i et $\hat{y}_i$ est de calculer la différence de ces deux quantités. Le résidus r_i résultant de ce calcul est dit **additif** et est tel que :

$$r_i = y_i - \hat{y}_i \quad \forall i \in \{1,...,n\}$$

avec n le nombre total de polices d'assurance à l'intérieur du portefeuille de l'assureur.

Cependant, fondamentalement, les résidus additifs issus d'une régression linéaire doivent vérifier les hypothèses suivantes :

- (H_1) : les résidus sont distribués suivant une loi normale ;
- (H_2) : les résidus sont de moyenne nulle et sont homoscédastiques (c'est à dire de variance constante)

Pourtant, dans le cadre d'un GLM de type log-poisson ces hypothèses ne sont pas vérifiées. En effet, il est vrai que la moyenne théorique des r_i est nulle puisque $\mathbb{E}(Y) = \hat{Y}$.

Cependant, le fait que Y suive une loi de poisson implique que :

- les r_i ne sont pas distribués suivant une loi normale ;
- $\mathbb{V}ar(Y) = \mathbb{E}(Y) = \hat{Y}$ et donc les r_i ne sont pas homoscédastiques.

Dans l'optique d'obtenir des résidus qui vérifient les hypothèses (H_1) et (H_2), une première transformation couramment utilisée est le calcul des **résidus de Pearson**. Suivant ce calcul, les r_i deviennent :

$$r_i = \frac{y_i - \hat{y}_i}{\sigma_{y_i}} = \frac{y_i - \hat{y}_i}{\hat{y}_i^{\frac{1}{2}}} \quad \forall i \in \{1,...,n\}$$

avec $\sigma_{Y_i} = \sqrt{\mathbb{V}ar(Y_i)} = \hat{Y}_i^{\frac{1}{2}}$.

Ce calcul revient en quelque sorte à centrer la variable Y à partir d'une estimation de sa moyenne et à la réduire par une estimation de son écart-type. Les résidus de Pearson ont donc l'avantage de vérifier par construction l'hypothèse (H_2). Toutefois, l'hypothèse (H_1) de normalité des résidus n'est toujours pas vérifiée.

En vue de rapprocher la distribution des résidus à celle d'une loi normale tout en conservant les avantages de la formule de Perason, le mathématicien Anscombe proposa

en 1953 une formule de calcul des résidus faisant intervenir une fonction $A(y)$ en lieu et place de y. La fonction A est telle que :

$$A(x) = \int_{-\infty}^{x} V^{\frac{-1}{3}}(t)dt$$

où $V(t)$ est la fonction de variance de la loi utilisée. La formule de calcul des **résidus d'Anscombe** est la suivante :

$$r_i = \frac{A(y_i) - A(\hat{y}_i)}{A'(\hat{y}_i)\sqrt{V(\hat{y}_i)}} \quad \forall i \in \{1,...,n\}$$

avec $A'(x)$ la fonction dérivée de $A(x)$. Dans le cas d'une loi de poisson, elle se réécrit :

$$r_i = \frac{3}{2}\frac{y_i^{\frac{2}{3}} - \hat{y}_i^{\frac{2}{3}}}{\hat{y}_i^{\frac{1}{6}}} \quad \forall i \in \{1,...,n\}$$

C'est finalement cette formule qui est utilisée pour la suite. L'interprétation des valeurs des résidus est la même pour les trois formules présentées. Si $r_i < 0$, alors la fréquence de sinistres prédite est supérieure à la fréquence de sinistres observée, il s'agit d'une surestimation de la fréquence de sinistres. Dans le cas où $r_i > 0$, la fréquence de sinistres prédite est inférieure à la fréquence de sinistres observée, c'est une sous-estimation de la fréquence de sinistres. Sous-estimer la fréquence de sinistres fait encourir un risque à l'assureur. Ainsi, plus la valeur du résidus augmente, plus le risque pour l'assureur est élevé.

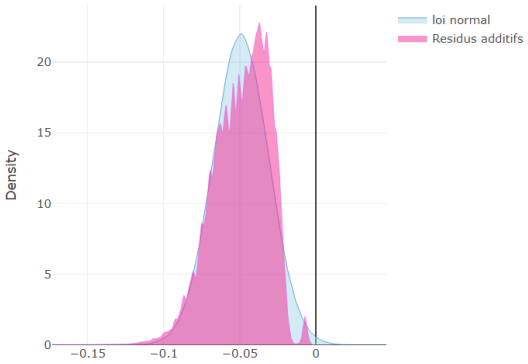

FIGURE 5.2 – *Superposition de la courbe de la densité empirique des résidus additifs et d'une loi normale*

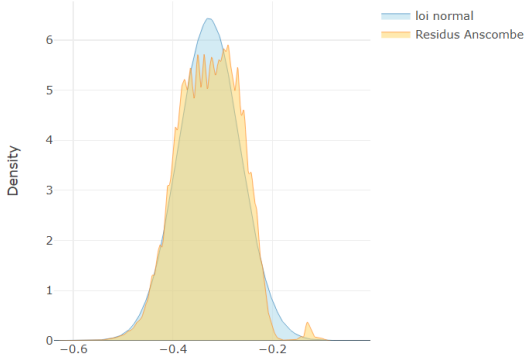

FIGURE 5.3 – *Superposition de la courbe de la densité empirique des résidus d'Anscombe et d'une loi normale*

Les figures 5.2 et 5.3 montrent que de manière empirique, la répartition des résidus d'Anscombe se rapproche plus d'une loi normale que celle des résidus additifs. La variance empirique des résidus d'Anscombe est plus élevée que celle des résidus additifs, ce qui est un avantage pour le traitement du signal géographique qu'ils contiennent. Enfin, la moyenne empirique de ces deux types de résidus est strictement inférieure à

zéro. Cela est dû à la sur-représentation des fréquences de sinistres nulles à l'intérieur du portefeuille.

Une fois les résidus individuels calculés à l'aide de la formule d'Anscombe, ceux-ci sont agrégés à une maille géographique dans le but de connaître le risque sur les différentes zones. Pour rappel, la maille géographique de cette étude est la commune. La formule de calcul du résidu sur une commune j donnée est la suivante :

$$R_j = \frac{\sum_{k=1}^{n_j} e_{k,j} r_{k,j}}{\sum_{k=1}^{n_j} e_{k,j}}$$

avec n_j le nombre total de polices rattachées à la commune j, $e_{k,j}$ et $r_{k,j}$ respectivement l'exposition et la valeur du résidu sur la police k rattachée à la commune j. Les résidus R_j sur les communes s'interprètent pareillement que les résidus individuels.

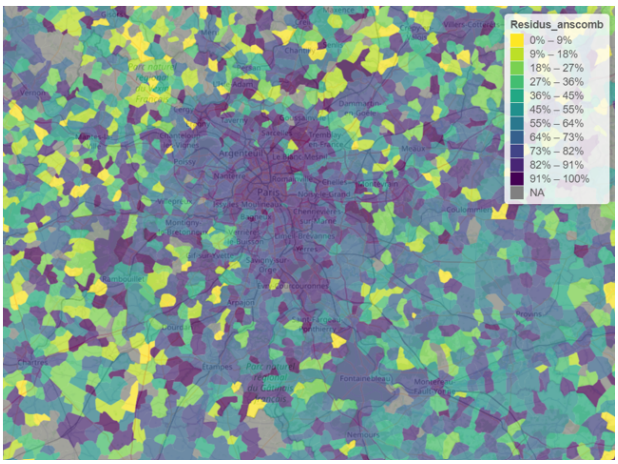

FIGURE 5.4 – *Cartographie des résidus agrégés : Zoom sur les communes d'Île de France*

Sur cet extrait de cartographie (Figure 5.4), le gradient de couleur passe du jaune représentant les communes sur lesquelles les résidus sont très faibles, au bleu nuit symbolisant les communes sur lesquelles les résidus sont très élevés. Les zones en gris sont des communes non couvertes par l'assureur.

Ici aussi, il est possible d'observer que le risque encouru par l'assureur n'est pas homogène sur le territoire. En effet, il se dégage des communes à risques plus élevés que d'autres.

Cependant, en leur état actuel, les résidus géographiques ont deux insuffisances majeures. premièrement, leur caractère bruité (basé sur les hypothèses H_1 et H_2) rend difficilement distinguable les différentes zones de risque. Deuxièmement, ces résidus ne sont pas exhaustifs sur le territoire. Pourtant, l'assureur doit être en mesure de proposer un tarif à une nouvelle police rattachée à une commune non-couverte par celui-ci. Pour

cette raison, il a besoin de connaître les niveaux de risque sur l'entièreté des communes du territoire.

En vue de combler ces insuffisances, les résidus géographiques subissent des traitements suivant deux méthodes différentes : une traditionnelle et l'autre moderne. Ces méthodes sont théoriquement présentées et mises en pratique dans les sections suivantes.

5.3 Méthodes de traitements du signal géographique

5.3.1 Théorie du lissage géospatial par crédibilité

Le lissage géospatial est une technique mathématique utilisée afin d'interpoler la valeur d'une variable sur une zone d'un espace à partir des valeurs de cette variable sur des zones voisines. En assurance, cette interpolation est objectivée par la prise en compte d'un facteur de crédibilité qui est l'exposition. En effet, la confiance de l'assureur quant à la valeur du risque (résidus) sur une commune dépend du niveau total d'exposition des polices rattachées à cette commune. De ce fait, la formule du lissage géospatial est structurée de sorte à prendre en compte l'exposition sur les communes. On parle dans ce cas de lissage par crédibilité.

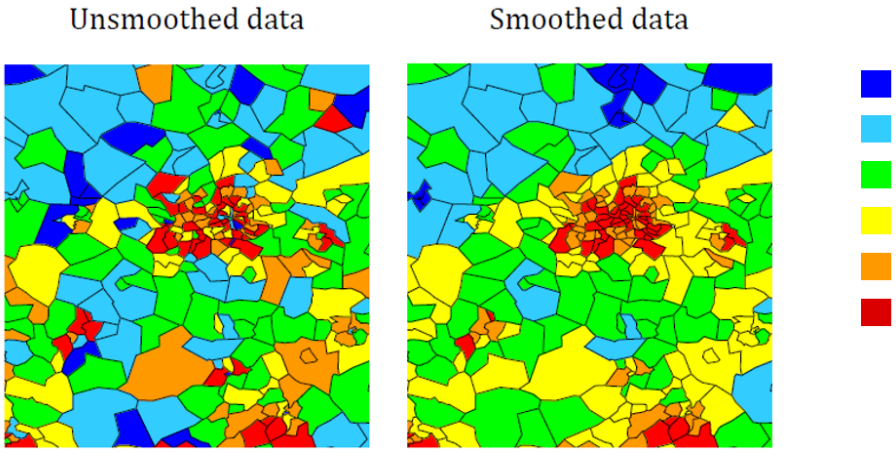

FIGURE 5.5 – *Illustration de l'impact d'un lissage géospatial (cas fictif)*
source : Territory Analysis Updates to the Traditional Methods, Gary Wang

La figure 5.5 montre l'impact d'un lissage géospatial d'une variable sur des zones d'un espace.
Avant l'application du lissage, il est possible d'observer des changements de valeur très drastiques entre des zones très proches. En effet, sur certaines zones voisines la couleur passe du bleu nuit au rouge. Cette situation est la plus part du temps le signe de la présence d'un bruit dans la donnée géographique. Pour un assureur, elle n'est pas très optimale dans la mesure où il serait par exemple difficile de justifier une grande différence de prime entre deux assurés du même profil habitant dans des communes adjacentes du

fait d'une grande différence du risque géographique sur ces communes.
Après lissage, les changements de valeur entre les zones sont moins drastiques et plus transitoires. Il y a donc eu réduction du bruit contenu initialement dans la donnée. Cela permet de mieux discerner des groupes de zones en fonction de la variable analysée.

Avantages

- Le lissage géospatial permet de diminuer le bruit contenu dans les résidus géographiques.
- La prise en compte de l'exposition dans la formule du lissage permet de crédibiliser les valeurs du risque sur les communes faiblement exposée.
- Le lissage géospatial permet d'obtenir par extrapolation les valeurs des résidus géographiques des communes non-couvertes par l'assureur.

Inconvénients

- Le lissage géospatial restreint l'explication des différences de risque sur les différentes zones aux distances entre celles-ci. Pourtant, plusieurs autres facteurs peuvent permettre d'expliquer la répartition du risque sur le territoire.
- Dans la pratique, la qualité d'un lissage par crédibilité est très subjective.

Formule théorique

La formule du lissage par crédibilité s'écrit :

$$\hat{R}_i = c(E_i)R_i + (1 - c(E_i))\frac{\sum_{j\neq i} R_j E_j f(d_{ij})}{\sum_{j\neq i} E_j f(d_{ij})}$$

avec :

- $\hat{R}_i$ la valeur estimée du résidu par le lissage sur la commune i
- E_i l'exposition totale des polices rattachées à la commune i
- R_i la valeur initiale du résidu sur la commune i
- d_{ij} la distance entre les communes i et j, ici il s'agit de la distance euclidienne entre les centroïdes des communes i et j.
- c est la **fonction de crédibilité** telle que : $c(E_i) = \frac{E_i}{E_i+a}$ où a est un paramètre à optimiser. Cette fonction permet de calibrer l'influence de l'exposition dans l'application du lissage ;
- f est une **fonction décroissante de la distance**. Ici, elle est telle que : $f(d_{ij}) = (\frac{1}{d_{ij}})^b$ où b est un paramètre à optimiser. Cette fonction permet de régler l'influence de la proximité entre la commune i et les communes j l'avoisinant dans l'application du lissage.

Optimisation des paramètres a et b

Mathématiquement parlant, il est difficile d'établir un réel critère afin de juger la qualité d'un lissage géospatial par crédibilité. En effet, quand il est très accentué le lissage peut complètement dénaturer le sens initial de la donnée. Dans le cas contraire, lorsqu'il est très faible le lissage redonne quasiment les valeurs initiales. L'objectif n'est donc pas de trouver une borne optimale, mais plutôt d'atteindre un équilibre convenable.

Du point de vue assurantiel, un bon lissage par crédibilité se définit de façon formelle comme un lissage dont les valeurs lissées sont très proches des valeurs initiales uniquement sur les zones fortement exposées.

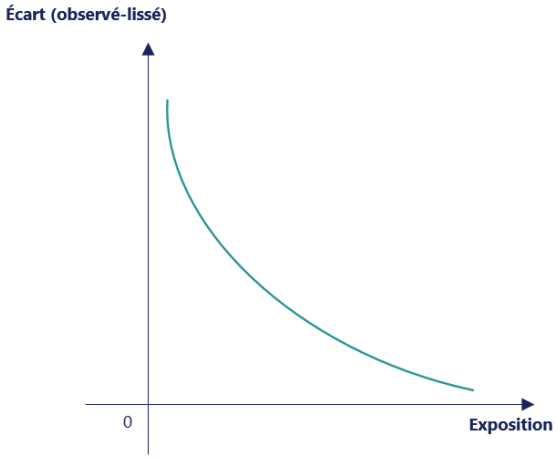

FIGURE 5.6 – *Illustration d'un bon lissage du point de vue assurantiel*

Partant de cette définition assurantielle, une métrique permettant d'évaluer la qualité d'un lissage est proposée. Soit Q^2 cette métrique de qualité :

$$Q^2 = 1 - \frac{\sum_i E_i(R_i - \hat{R}_i(a,b))^2}{\sum_i E_i(R_i - \overline{R})^2}$$

avec $\overline{R}$ le résidu géographique moyen. Plus le Q^2 est élevé, meilleure est la qualité du lissage. Il faut donc trouver le couple (a^*,b^*) pour lequel la valeur du Q^2 est maximale.

Cette optimisation rencontre un problème. Le Q^2 tel que définit privilégie les valeurs très faibles de a puisque :

$$\forall b \in \mathbb{R}^+, \quad \lim_{a\to 0} \hat{R}_i(a,b) = R_i \implies \lim_{a\to 0} Q^2 = 1$$

Il convient de corriger cette formule du Q^2 afin de pénaliser cet avantage des faibles valeurs de a sans pour autant favoriser la prise de grandes valeurs. La nouvelle expression du Q^2 proposée est :

$$Q^2_{corrige} = \log(1+a)Q^2$$

Dans cette nouvelle formule, la décroissance du Q^2 en fonction des valeurs de a est corrigée par la croissance de $\log(1+a)$. Il est possible d'étudier les comportements limites du rapport des fonctions $c(E_i,a)$ (qui fait décroître le Q^2 en fonction de a) et $\log(1+a)$. Soit la fonction G telle que pour une valeur de E_i fixée :

$$E_i \in]0, \max(E_i)] \quad et \quad \forall a \in \mathbb{R}^+, \quad G(a) = \frac{c(E_i,a)}{\log(1+a)} = \frac{\frac{E_i}{E_i+a}}{\log(1+a)}$$

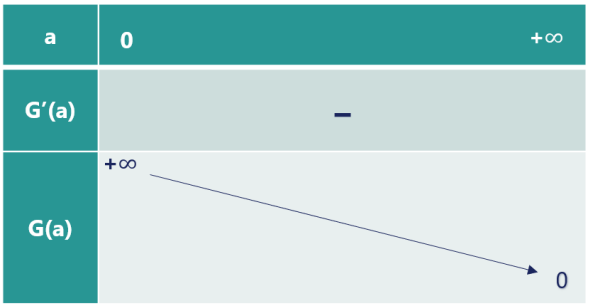

FIGURE 5.7 – *Tableau des variations de la fonction G*

Le tableau des variations de la fonction G montre effectivement que l'influence de la fonction $c(E_i,a)$ décroît progressivement lorsque les valeurs de a augmentent du fait du correcteur $\log(1+a)$ appliqué au Q^2.

Une méthodologie d'optimisation des paramètres a et b est proposée. La base des communes couvertes par l'assureur est scindée en deux sous-bases : une base des connus et une base des inconnus. Le pourcentage de communes dans la base des connus est égal au taux de communes couvertes par l'assureur. Sur cette sous base, l'exposition totale et les résidus géographiques sur les communes sont considérés connus. La base des inconnus contient les communes restantes. Sur cette dernière, l'exposition totale et les résidus géographiques sur les communes sont considérés inconnus.

Étape 1 : Optimisation du paramètre b
Le paramètre a n'ayant aucune influence sur la base des inconnus (les expositions étant considérées inconnues), l'optimisation du paramètre b s'effectue sur cette sous-base. Une valeur arbitraire de a est choisie et plusieurs lissages sont testés sur la base complète contenant les deux sous-bases en faisant varier les valeurs du b. A la fin de chaque lissage, la valeur du Q^2 est calculée sur la base des inconnus grâce aux valeurs extrapolées et aux valeurs initialement considérées inconnues. La valeur optimale b^* du paramètre b est celle pour laquelle la valeur du Q^2 est la plus élevée.

Étape 2 : Optimisation du paramètre a
Avec la valeur b^* trouvée, plusieurs lissages sont à nouveau testés sur la base complète contenant les deux sous-bases en faisant varier cette fois les valeurs du a. A la fin de

chaque lissage, la valeur du $Q^2_{corrige}$ est calculée sur la base des connus grâce aux valeurs lissées et aux valeurs initiales. La valeur optimale a^* du paramètre a est celle pour laquelle la valeur du $Q^2_{corrige}$ est la plus élevée.

5.3.2 Théorie des forêts aléatoires

La sous-section précédente a mis en lumière les avantages d'une première méthode de traitement du signal géographique contenu dans les résidus à savoir le lissage géospatial par crédibilité. Cependant, elle a souligné un de ses inconvénients majeurs qui est la restriction de l'explication du risque géographique sur les communes aux distances qui les séparent. Pourtant, au delà des distances, plusieurs autres facteurs peuvent expliquer la répartition de ce risque sur le territoire. Ainsi, avant l'application du lissage, il serait bénéfique d'expliquer, de prime abord, les résidus géographiques à partir de facteurs externes.

Au chapitre 4, il a été expliqué que les arbres de décision étaient des outils très efficaces pour la réalisation de cette tâche. Toutefois, il a aussi été dit qu'en terme de régression, les prédictions de ces arbres étaient assez instables. C'est en vue d'exploiter leur capacité d'explication tout en s'assurant de la robustesse et de la stabilité des prédictions, qu'est né l'algorithme de **forêt aléatoire**.
Une très belle métaphore serait d'assimiler le principe de cet algorithme à un proverbe qui dit : « si une personne te dit que tu es un cheval, n'y prête pas attention. Si deux personnes te disent que tu es un cheval, commence à dresser l'oreille. Si trois personnes te disent que tu es un cheval, cours t'acheter une selle ». En effet, la réponse finale d'une forêt aléatoire se base sur la réponse majoritaire parmi les réponses de plusieurs arbres de décision. Ce style de fonctionnement est appelé le *Bagging*.

Bagging

Le terme *Bagging* est un anglicisme né de la fusion des mots *Bootstrap* et *aggregating*. Le *Bagging* est une méthode ensembliste qui consiste dans un premier temps à construire plusieurs arbres sur différents échantillons d'individus tirés aléatoirement dans la base initiale. Cet échantillonage aléatoire effectué est le volet *Bootstrap* du *Bagging*. Dans un second temps, une réponse finale est calculée soit en moyennisant les réponses des différents arbres (cas d'une régression), soit en sélectionnant la réponse prépondérante parmi les réponses des différents arbres (cas d'une classification). Cette dernière étape est le volet *aggregating* du *Bagging*.

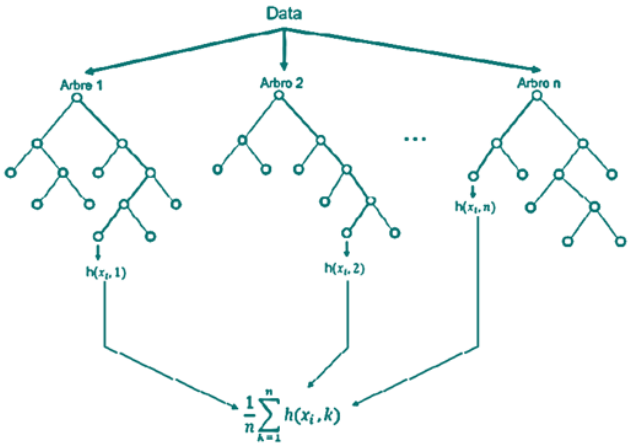

FIGURE 5.8 – *Illustration d'une forêt aléatoire : cas d'une régression*

Erreur *Out Of Bag (OOB)*

soit i un individu de la base initiale et y_i sa valeur observée de la variable à prédire. Soit $\mathbb{O}$ la famille des échantillons *Bootstrap* ne contenant pas l'individu i. La prédiction OOB pour l'individu i est la réponse moyenne ou majoritaire donnée par les arbres construits sur les échantillons contenus dans $\mathbb{O}$. Dans le cas d'une régression, elle s'écrit :

$$\hat{z}_i = \frac{1}{Card(\mathbb{O})} \sum_{k \in \mathbb{O}} h_i(x_{k_p},...,x_{k_q},k)$$

avec les $x_{k_.}$ les valeurs des variables explicatives utilisées pour la construction de l'arbre k et $h_i(x_{k_p},...,x_{k_q},k)$ la réponse de l'arbre k pour l'individu i.

L'erreur OOB d'une forêt aléatoire est une quantification de l'écart moyen entre les valeurs observées Y et les prédictions OOB $\hat{Z}$:

$$E_{OOB} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{z}_i)^2$$

La minimisation de L'erreur OOB permet de réduire les erreurs de prédictions finales du modèle de forêt aléatoire.

Importance des variables

Il est possible d'attribuer un score d'importance aux variables explicatives d'une forêt aléatoire. Ces scores d'importance permettent de détecter et de retirer du modèle les variables n'ayant aucun impact dans l'explication de la variable à prédire. Cet aspect de l'algorithme fait qu'il est souvent utilisé pour la sélection supervisée de variables.

Pour connaître l'importance de la variable j, il faut premièrement calculer E^j_{OOB} l'erreur OOB moyenne rattachée à la variable j, c'est à dire la moyenne des erreurs OOB calculées à partir des prédictions des arbres contenant la variable j comme variable explicative. Deuxièmement, ce même calcul est effectué mais cette fois-ci pour la variable j stressée notée j_s. La variable j_s est obtenue en permutant de manière aléatoire les observations de la variable j. Finalement l'importance de la variable j est l'écart absolu entre ces deux erreurs calculées :

$$Imp_j = |E^j_{OOB} - E^{j_s}_{OOB}|$$

Plus Imp_j est élevé, plus la variable j est importante puisqu'une perturbation de ses observations entraîne une forte modification des prédictions du modèle.

Avantages

- l'algorithme de forêt aléatoire allie la simplicité du fonctionnement des arbres de décisions et la robustesse des prédictions.
- la modélisation par cet algorithme ne fait aucune hypothèse sur la loi ou sur la linéarité de la relation entre la variable à prédire et les variables explicatives.
- En construisant plusieurs arbres à partir de différents lots de variables explicatives, la forêt aléatoire prend naturellement en compte les effets d'interaction entre les variables explicatives.

Inconvénients

- Pour des jeux de données de grande taille, le temps d'implémentation de l'algorithme peut être conséquent.
- Le nombre élevé de paramètres dont dispose celui-ci peut rendre son calibrage quelque peu complexe.
- Il est difficile de tracer l'algorithme. Dans le jargon statistique, on parle de "modèle boîte noire".

Optimisation des paramètres

Afin d'éviter que le modèle de forêt aléatoire sur-apprenne les données d'apprentissage, il convient de l'hyper-paramétrer. Cela revient à déterminer le jeux de paramètres pour lequel le modèle réalise le meilleur compromis biais-variance. Dans ces travaux, cette optimisation porte sur quatre paramètres influençant fortement le modèle. Ce sont :

- *ntree* : le nombre d'arbres construits au cours de l'implémentation de l'algorithme ;
- *mtry* : le nombre de variables sélectionnées pour la construction de ces arbres ;
- *nodesize* : le nombre minimum d'individus dans les feuilles des arbres construits ;
- *maxnodes* : le nombre maximum de feuilles pour chacun des arbres construits.

5.3.3 Méthodes de clustering

Les deux premières méthodes présentées ci-dessus permettent de traiter le signal géographique contenu dans les résidus. Une fois cette étape achevée, les résidus traités sont regroupés en classe de manière à définir des zones (*clusters*) de risque. C'est l'ensemble de ces zones de risque qui forme la variable appelée **zonier**.

Dans ces travaux, deux méthodes de *clustering* sont testées : le découpage par quantiles et la Classification Ascendante Hiérarchique (CAH).

découpage par quantiles

Découper une variable par quantiles revient à effectuer une équipartition de celle-ci. Cette technique s'appuie donc sur une égale répartition des nombres de communes à l'intérieur des classes constituées sans tenir compte de la proximité des valeurs de leur résidu.

Classification Ascendante Hiérarchique (CAH)

La CAH, quant à elle, est une méthode de partitionnement qui cherche à constituer des groupes tels que :

- au sein d'un même groupe, les valeurs des résidus sur les communes sont très proches (homogénéité intra-groupe) ;
- deux communes de deux groupes différents ont des valeurs de résidu très éloignées (hétérogénéité inter-groupe).

Cette méthode de classification est dite hiérarchique puisqu'algorithmiquement elle consiste à construire une suite de partitions imbriquées. A chaque itération l'algorithme crée une sous partition de l'ensemble précédent afin de maximiser l'homogénéité intra-groupe et l'hétérogénéité inter-groupe en s'appuyant sur la proximité des valeurs des résidus géographiques. La distance couramment utilisée pour quantifier cette proximité est la distance de Ward. Soient A et B deux ensembles :

$$d_{Ward}(A,B) = \frac{n_A n_B}{n_A + n_B} d(G_A G_B)^2$$

avec n_A et n_B les nombres respectifs d'éléments à l'intérieur des ensembles A et B et $d(G_A,G_B)$ la distance euclidienne entre les centres de gravité respectifs des ensembles A et B.

Choix du nombre optimal de groupes

Le choix du nombre de groupes à constituer dépend des attentes du modélisateur. Un nombre de groupes très élevé induira des zones de risque trop fines, quand un nombre de groupes très faible conduira à des zones de risque trop triviales.

Le plus important étant d'avoir des zones très hétérogènes, il est possible de choisir le nombre de groupes à constituer en fonction de la diminution de l'hétérogénéité inter-classe que celui-ci implique. En mathématique cette hétérogénéité se quantifie par le calcul de l'inertie. Le choix du nombre k de groupes à constituer peut donc s'opérer en représentant la perte de l'inertie inter-classe en fonction de k.

5.4 Application de ces méthodes pour la construction de zoniers

Dans cette section, deux zoniers sont construits à partir des résidus géographiques issus du modèle initial de fréquence de sinistres.

Le premier zonier construit est le **zonier traditionnel**. Il est obtenu par un lissage géospatial crédibilisé de ces résidus. Ce zonier représente le zonier standard d'un assureur et son intégration dans le modèle initial de fréquence de sinistres permet d'obtenir un **modèle de référence** d'assureur.

Le second zonier construit est le **zonier innovant**. Son obtention passe d'une part par une étape de modélisation statistique des résidus avec pour variables explicatives les données télématiques externes et de l'autre par un lissage géospatial crédibilisé des prédictions de ce modèle. Ce zonier est dit innovant car sa construction combine à la fois une méthodologie moderne et l'intégration d'informations rarement utilisées pour l'établissement d'un zonier sur le marché de l'assurance automobile. L'intégration de ce zonier dans le modèle initial de fréquence de sinistres permet d'obtenir un **modèle innovant** d'assurance automobile.

5.4.1 Construction d'un zonier traditionnel

La construction du zonier traditionnel débute par le choix des paramètres a et b du lissage géospatial par crédibilité. La méthode d'optimisation de ces paramètres, déjà présentée théoriquement, est appliquée sur la base des communes couvertes par l'assureur.

La valeur maximale du Q^2 sur la sous-base des inconnus est atteinte lorsque le paramètre b vaut 1,8 (Figure 5.9).

L'optimisation du paramètre a confirme les démonstrations présentées dans la partie théorique. En effet, sans l'application du correcteur au Q^2, la valeur optimale du paramètre a est la plus petite des valeurs testées (décroissance du Q^2 en fonction de a, Figure 5.10). Cependant, en corrigeant les valeurs du Q^2 par $\log(1+a)$, la valeur du $Q^2_{corrige}$ atteint un pique sur la sous-base des connus lorsque a vaut 40 (Figure 5.11).

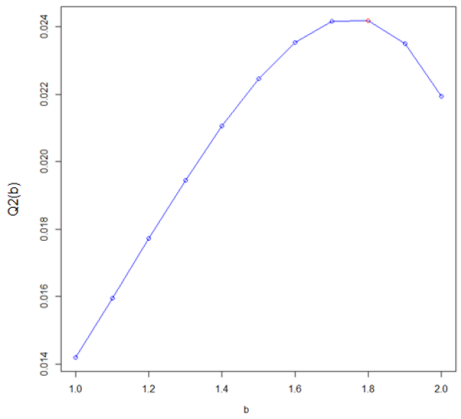

FIGURE 5.9 – *Optimisation du paramètre b sur la base des inconnus en utilisant le* Q^2

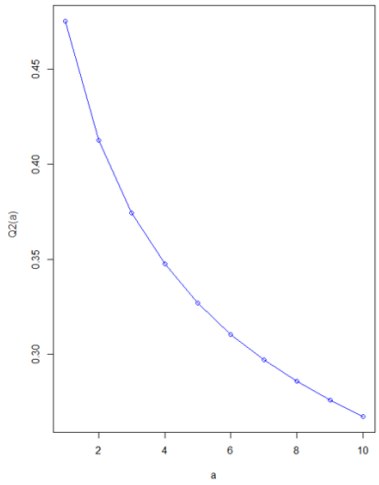

FIGURE 5.10 – *Optimisation du paramètre a sur la base des connus en utilisant le* Q^2

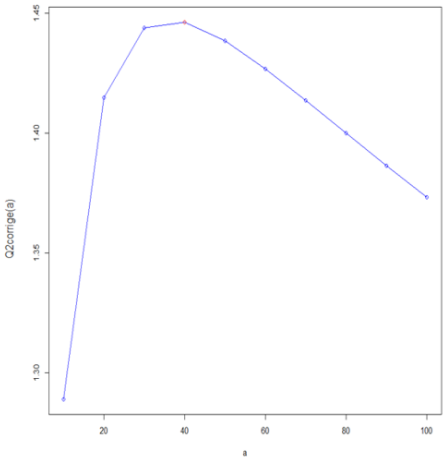

FIGURE 5.11 – *Optimisation du paramètre a sur la base des connus en utilisant le* $Q^2_{corrige}$

Cette méthode d'optimisation des paramètres a et b proposée, a certes des avantages, mais elle a aussi des inconvénients. Par exemple les valeurs maximales du Q^2 et du $Q^2_{corrige}$ obtenues peuvent être des maximum locaux. Elle doit donc être utilisée avec un minimum de réserve.

Avec le couple de paramètres $(a^*,b^*) = (1.8,40)$ obtenu, le lissage par crédibilité "optimal" est appliqué aux résidus géographiques. Après cette étape de lissage, arrive le choix du nombre de zones à constituer à partir de ces résidus lissés.

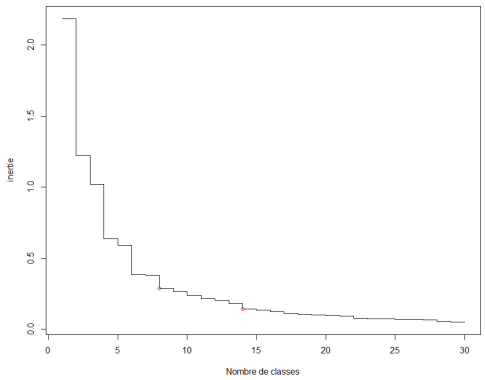

FIGURE 5.12 – *Perte d'inertie inter-groupe en fonction du nombre de zones*

Sur la figure 5.12, il est possible de remarquer que la perte d'inertie inter-groupe est faible et tend à se stabiliser pour les nombres k de zones compris entre 8 et 14 (points rouges sur la figure). Ainsi, il a été décidé de construire 11 zones de risque sur les résidus lissés. Pour la création de ces 11 zones de risque, le découpage par quantiles et la CAH sont testés.

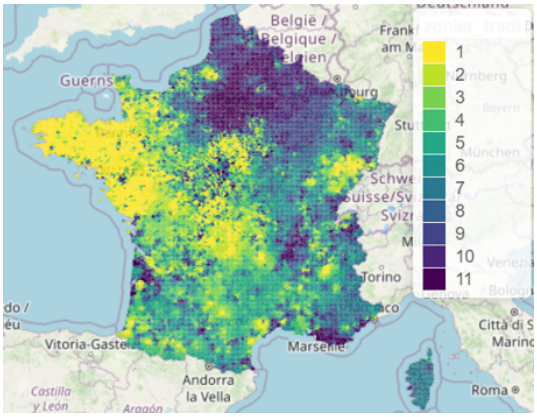

FIGURE 5.13 – *Zonier traditionnel obtenu en découpant les résidus lissés par quantiles*

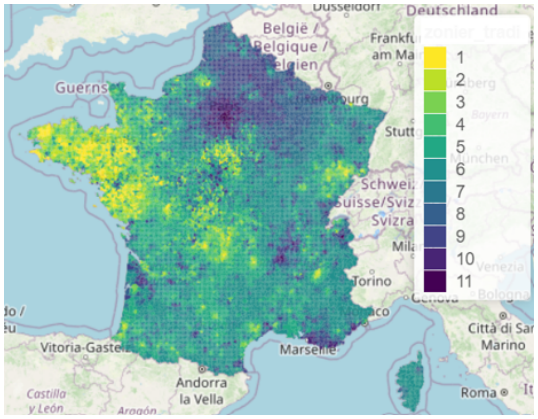

FIGURE 5.14 – *Zonier traditionnel obtenu en appliquant la CAH sur les résidus lissés*

C'est finalement le zonier traditionnel obtenu en subdivisant les résidus lissés par quantiles (Figure 5.13) qui est retenu. Ce choix se justifie par le reflet de zones plus discernables sur ce dernier.

Sur le zonier traditionnel, les zones sont disposées par ordre croissant de niveau de risque. En d'autres termes par ordre croissant des résidus géographiques lissés moyens sur celles-ci. La zone 1 représente donc les communes les moins risquées pour l'assureur et la zone 11 renferme les communes les plus à risque pour ce dernier. Par exemple, la cartographie du zonier traditionnel montre que la Bretagne est une zone à risque relativement faible pour l'assureur, tandis que l'île de France est une zone beaucoup plus

risquée pour celui-ci.

5.4.2 Construction d'un zonier innovant

La construction du zonier innovant démarre avec la modélisation des résidus géographiques prenant pour variables explicatives des données télématiques externes s'effectuant ici par le biais d'un modèle de forêt aléatoire. Une sélection supervisée des variables télématiques externes s'impose en amont de cette modélisation.

Cette sélection de variables s'opère en deux phases. La première phase consiste à faire un premier tri à l'aide de la régression LASSO. Le but de cette première sélection est de minimiser les corrélations entre les variables qui seront finalement conservées. La seconde phase consiste à faire un premier modèle de forêt aléatoire pour les résidus géographiques avec les variables sélectionnées par LASSO et de garder celles ayant une importance significative à l'intérieur de ce modèle.
A l'issue de cette étape, 13 variables externes sont sélectionnées. Parmi ces 13 variables, 10 sont des variables télématiques et 3 sont des données sur les types de routes à l'intérieur des communes. La liste des 13 variables est consignée dans la figure 5.15.

Nom de la variable	Description
s2_anticipation_brakingIntensity	Score evaluant l'intensité de freinage
s2_anticipation_brakingSudden	Score evaluant les freinages brusques
s2_pace_straight	Score evaluant l'allure de la conduite en ligne droite
total_km	Total des kilomètres parcourus sur la commune
WET_weather_part_of_km	Taux de kilomètres parcourus en temps humide
STRAIGHT_CURVE_part_of_km	Taux de kilomètres parcourus en ligne droite
PLAIN_tri_part_of_km	Taux de kilomètres parcourus dans les plaines
TIGHT_CURVE_part_of_km	Taux de kilomètres parcourus dans les virages sérrés
TIGHT_CURVE_mean_speed	Vitesse moyenne dans les virages sérrés
EXTRA_URBAN_mean_speed	Vitesse moyenne sur les routes extra-urbaines
part_of_CURVE	Taux de virage dans la commune
part_of_inter	Taux de route inter-urbaine dans la communne
part_of_urban	Taux de route urbaine dans la commune

FIGURE 5.15 – *Liste des 13 variables sélectionnées pour la construction du zonier innovant*

Après la sélection des variables explicatives du modèle, vient l'étape de son hyperparamétrage. L'objectif de cette autre étape est de déterminer le jeu de paramètres pour lequel celui-ci a un taux d'erreurs de prédiction faible sans pour autant sur-apprendre les données d'apprentissage. Il s'agit donc d'analyser les variations des erreurs de prédiction en fonction des différentes valeurs des paramètres à optimiser. Le cas de l'optimisation du paramètre *ntree* est présenté en guise d'exemple.

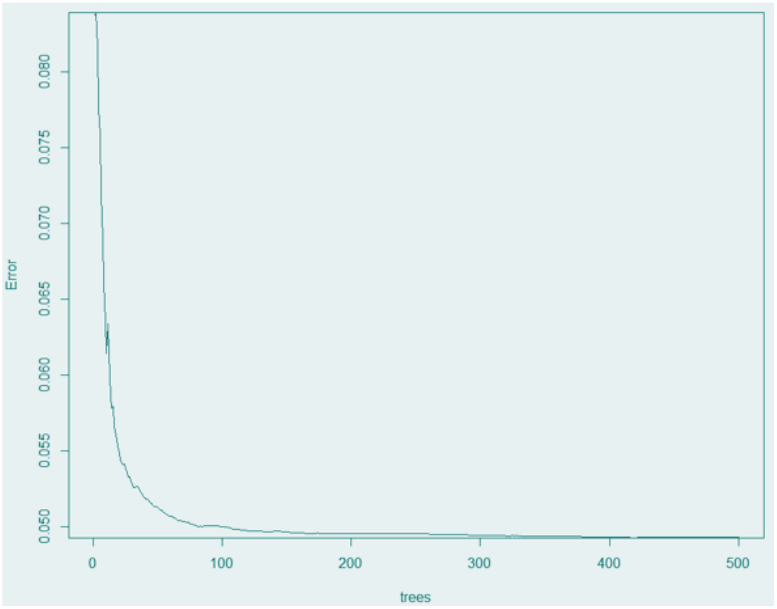

FIGURE 5.16 – *Évolution de l'erreur de prédiction en fonction du nombre d'arbres ntree*

La valeur par défaut du paramètre *ntree* est 500 arbres. Pourtant, la figure 5.16 montre que l'erreur de prédiction du modèle de forêt aléatoire ne baisse quasiment plus à partir de 200 arbres. La valeur optimale du paramètre *ntree* est de ce fait fixée à 200 arbres.

Le tableau suivant résume les résultats de l'étape de l'hyper-paramétrage du modèle.

	paramètres par défaut	paramètres optimisés
ntree	500	200
mtry	4	3
nodesize	5	100
maxnodes	12000	5000
RMSE base de train	**0,103**	**0,188**
RMSE base de test	**0,212**	**0,21**

FIGURE 5.17 – *Tableau récapitulatif de l'hyper-paramétrage du modèle de forêt aléatoire*

Maintenant que les variables explicatives et les paramètres sont connus, le modèle est entraîné et ses prédictions de résidus géographiques sont extraites. L'importance de chacune des 13 variables explicatives dans la construction de ces prédictions est présentée à la figure 5.18.

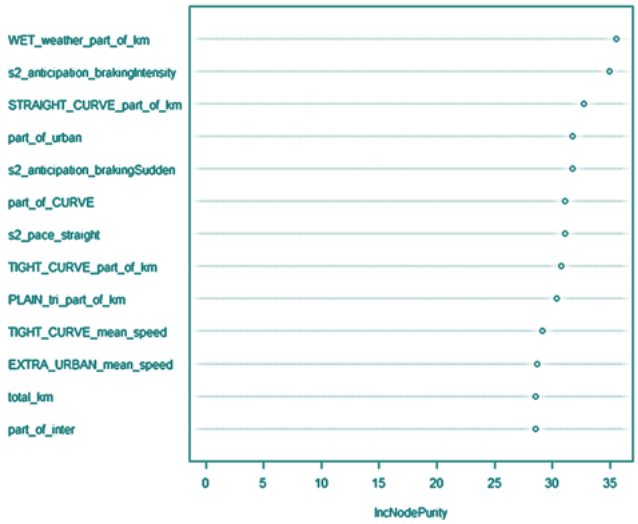

FIGURE 5.18 – *Importance des 13 variables dans le modèle de forêt aléatoire*

Les prédictions des résidus géographiques extraites sont lissées par crédibilité afin de raffiner l'apport des informations télématiques externes par les distances entre les communes. Enfin, le zonier innovant est obtenu, avec 11 zones constituées par quantiles parallèlement au découpage du zonier traditionnel.

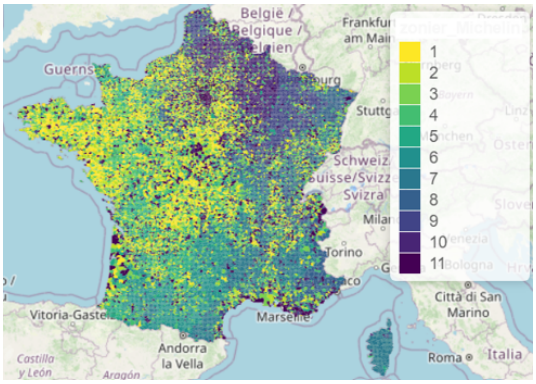

FIGURE 5.19 – *Zonier innovant obtenu en découpant les résidus prédits puis lissés par quantiles*

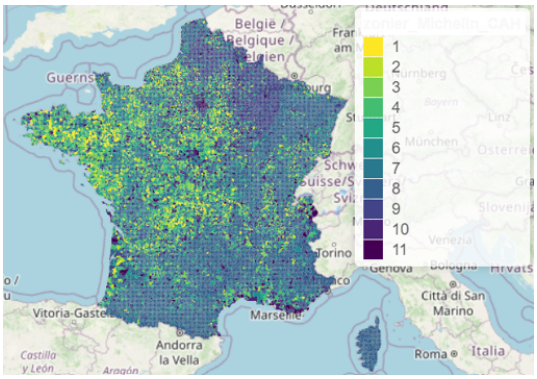

FIGURE 5.20 – *Zonier innovant obtenu en appliquant la CAH sur les résidus prédits puis lissés*

Sur la cartographie du zonier innovant (Figure 5.19), il se dégagent certaines zones presque similaires à celles du zonier traditionnel. Par exemple, sur ce zonier innovant, la Bretagne est aussi une zone où le risque pour l'assureur est relativement faible. Néanmoins, il est possible de remarquer des différences sur certaines zones. En île de France notamment, le zonier innovant considère de manière beaucoup plus restreinte que la zone à risque élevé est seulement le noyau central de ce département.

5.5 Deuxième évaluation de l'apport des données télématiques contextualisées

La précédente section a été consacrée à la construction de deux zoniers. Le premier zonier s'est construit de manière traditionnelle en vue d'obtenir une segmentation géographique standard. L'intégration de ce zonier traditionnel dans le modèle initial de fréquence de sinistres permet d'obtenir un modèle typique d'un assureur automobile. Ici, ce modèle typique est appelé modèle référent. Le second zonier s'est construit de manière plus innovante, synthétisant les informations apportées par certaines données télématiques géospatialisées. L'intégration de cette segmentation géographique innovante dans le modèle initial de fréquence de sinistres permet d'obtenir un modèle assez atypique pour les assureurs automobile. Ce nouveau modèle est nommé modèle innovant.

La comparaison des performances statistiques de ces deux modèles : référent et innovant, constitue un deuxième moyen d'évaluer l'apport des données télématiques contextualisées dans la modélisation du risque géographique. Cette évaluation est complétée par une analyse du pouvoir de discrimination du risque intrinsèquement aux deux zoniers construits.

5.5.1 Comparaisons statistiques des performances des modèles : référent et innovant

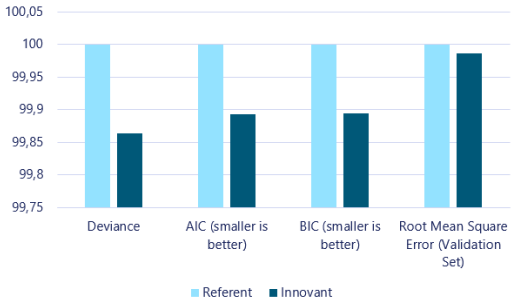

FIGURE 5.21 – *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle référent)*

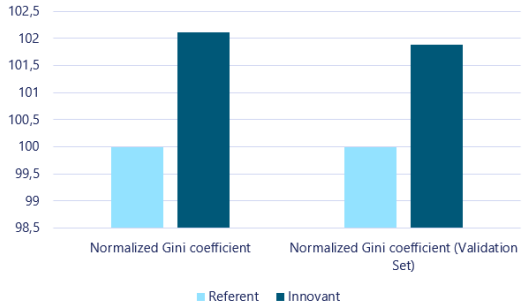

FIGURE 5.22 – *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle référent)*

Les figures 5.21 et 5.22 semblent dire que le modèle innovant est statistiquement de meilleure qualité que le modèle référent. Cependant, les écarts de performances entre ces deux modèles sont relativement moins conséquents que ceux obtenus lors des comparaisons dans l'approche naïve. Par exemple, ici, les indices de Gini évoluent au maximum de 2% en passant du modèle référent au modèle innovant. Ce constat est totalement objectif puisque dans cette nouvelle approche, les deux modèles comparés contiennent une couche de segmentation géographique. Un trop grand écart de performances entre ces derniers aurait pu signifier la mauvaise construction d'un des zoniers au profit de

l'autre.

5.5.2 Comparaison du spread des zoniers : traditionnel et innovant

Au delà de la comparaison des indicateurs de qualité des modèles, il est aussi possible de comparer la manière dont les zoniers contenus dans ces modèles discriminent le risque géographique. Cela revient à comparer le poids du risque associé aux classes extrêmes de chacun de ces zoniers. En effet, plus les poids du risque associés au zones extrêmes d'un zonier sont éloignés, plus les zones de ce zonier sont discriminantes du risque. Mathématiquement, le poids du risque associé à une zone est le facteur multiplicatif que le modèle GLM lui attribue.

Spread d'un zonier en théorie

Le *spread* d'un zonier est une métrique qui permet de quantifier sa puissance de discrimination du risque. Le calcul du *spread* se base sur les facteurs multiplicatifs des zones extrêmes. soient A et Z respectivement la zone la moins risquée et la zone la plus risquée d'un zonier, de manière standard :

$$spread = 1 - \frac{\exp(\beta_A)}{\exp(\beta_Z)}$$

avec $\exp(\beta_A)$ et $\exp(\beta_Z)$ les facteurs multiplicatifs respectifs des zones A et Z.

Plus la valeur du *spread* se rapproche de 1 (ou de 100 en pourcentage), plus la variation du poids du risque entre les zones extrêmes du zonier est élevée. Cela implique qu'entre deux zoniers, celui qui discrimine le mieux le risque est celui ayant le plus grand *spread*.

Pour éviter que les différences des expositions sur les zones extrêmes n'influencent fortement le calcul, une formule normalisée du *spread* est proposée :

$$spread_{norm} = spread * \frac{\min(Expo_A; Expo_Z)}{\max(Expo_A; Expo_Z)}$$

avec $Expo_A$ et $Expo_Z$ respectivement l'exposition totale sur les zones A et Z.

Cette nouvelle formule pénalise le *spread* par l'équilibre des expositions sur les zones extrêmes. L'interprétation de ce nouvel indicateur est pareille à celle du *spread* standard.

Application : calcul et comparaison du spread des zoniers traditionnel et innovant

Sur les graphiques suivants, [R] représente la zone de référence et les chiffres au dessus des courbes représentent les variations entre le facteur multiplicatif de la zone de référence et les facteurs multiplicatifs des autres zones.

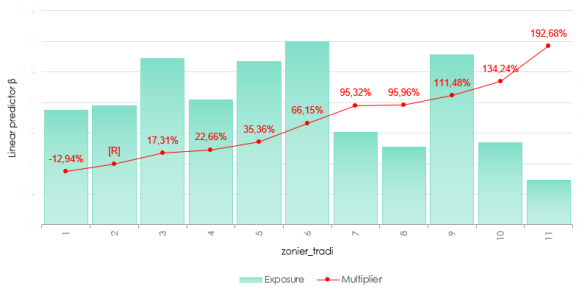

FIGURE 5.23 – *Courbe des facteurs multiplicatifs des modalités du zonier traditionnel*

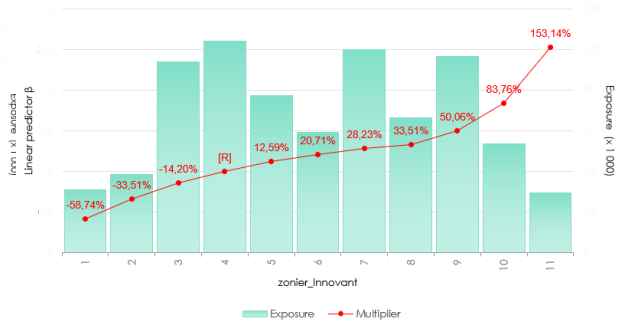

FIGURE 5.24 – *Courbe des facteurs multiplicatifs des modalités du zonier innovant*

Les Figures 5.23 et 5.24 montrent que les zones de risque des deux zoniers analysés sont cohérentes avec la fréquence de sinistres. En effet, le passage d'une zone moins risquée à une zone plus risquée entraîne une variation positive entre les facteurs multiplicatifs de ces deux zones. Cela explique la pente positive de ces deux zoniers. Ainsi, passer d'une zone moins risquée d'un de ces zoniers à une zone plus risquée du même zonier augmente la probabilité d'avoir un sinistre.

Le tableau suivant présente les différentes valeurs des *spread* des zoniers traditionnel et innovant.

	Zonier traditionnel	zonier innovant
Multiplier 1	0,8706	0,4126
Multiplier 11	2,9268	2,5314
Exposition 1	64425	30870
Exposition 11	25207	29420
spread standard	**70,25%**	**83,70%**
spread normalisé	**27,49%**	**79,77%**

FIGURE 5.25 – *Tableau récapitulatif des comparaisons de spread des zoniers traditionnel et innovant*

D'après le tableau récapitulatif (Figure 5.31), quelque soit la formule de calcul utilisée, le *spread* du zonier innovant domine celui du zonier traditionnel. Cela signifie qu'en terme de puissance de discrimination du risque, le zonier innovant est nettement meilleur que le zonier traditionnel.

Cette analyse des *spread* couplée aux comparaisons des performances statistiques précédemment présentées prouve une nouvelle fois la significativité de l'apport des données télématiques dans la modélisation du risque géographique en assurance automobile.

5.6 Open data et Smart road data, une union envisageable ?

Jusqu'ici, les résultats obtenus confirment le fait que le risque géographique en assurance automobile peut être expliquer à partir d'informations sur les habitudes de conduite dans les différentes zones du territoire. Toutefois, depuis quelques années, plusieurs travaux en actuariat ont montré qu'il était possible de modéliser ce risque à l'aide d'un autre type de données appelées *Open data*.

Les Open data sont des données numériques, provenant généralement de structures publiques, dont l'accès, l'usage, la modification et la rediffusion sont librement ouverts à tous les usagers. Elles peuvent être de divers ordres. En assurance automobile, les plus utilisées concernent :

- l'environnement socio-économique (données sur les infrastructures, sur le niveau de vie, sur la population,...)
- l'emploi (données sur les taux de chômage, sur les taux d'actifs, sur les lieux de travail,...)
- l'environnement routier (données sur les nombres de radars, sur les type de routes, sur le réseau routier,...)
- la météo (données sur les températures, sur les précipitations, sur l'ensoleillement, ...)

La question à laquelle répond cette dernière section est la suivante : la fusion *Open data* - données télématiques peut-elle permettre d'obtenir une meilleure segmentation géographique que l'utilisation exclusive des données *Open data* ?

Pour répondre à cette interrogation, une dernière comparaison de deux modèles est proposée. La racine de ces modèles est la modélisation du *burning cost* (modèle *tweedi*) déjà présentée au chapitre 3. La différence entre ces modèles se situe au niveau des zoniers.

En effet, dans l'un de ces derniers se trouve un zonier construit (en dehors de ces travaux) de manière moderne avec pour variables explicatives trois données *Open data* : La densité de la population par commune, le niveau de vie des populations par commune et l'ensoleillement. Cependant, la technique de lissage utilisée pour obtenir ce zonier diffère de celle présentée dans ce chapitre. Cette technique, appelée *Krigeage* et récemment utilisée en assurance ne sera pas abordée ici. Le modèle contenant ce zonier décrit est nommé **modèle *Open data***.

Dans l'autre modèle se trouve un zonier construit (dans le cadre de ces travaux) aussi de manière moderne mais cette fois en complétant les trois variables *Open data* citées par une sélection de données télématiques extraite de la base externe *Smart road data*. L'intégration de ce zonier hybride dans le modèle initial de *burning cost* permet

d'obtenir **le modèle hybride**.

L'importance des variables dans le modèle de forêt aléatoire ayant permis la construction du zonier hybride est présentée sur la figure 5.26.

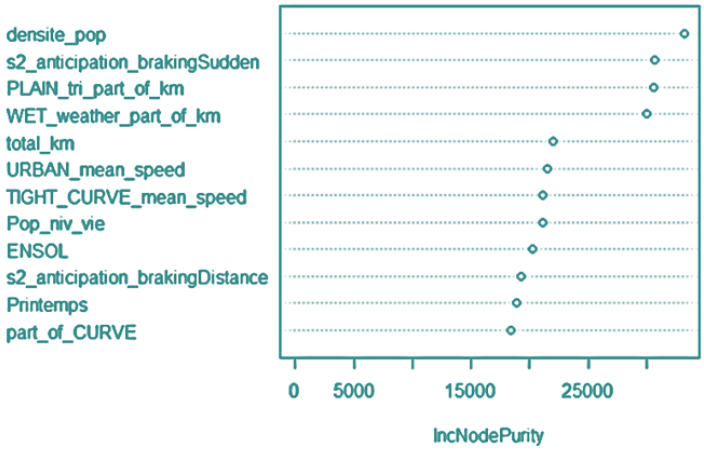

FIGURE 5.26 – *Importance des variables dans le modèle de forêt aléatoire pour la construction du zonier hybride*

En termes d'importance, quatre variables se démarquent le plus : une variable *Open data* concernant la densité des populations par commune, trois variables télématiques renseignant le score de freinage brusque, les taux de parcours dans les plaines et les taux de parcours en temps humides. La diversité de ces variables importantes pour l'explication des résidus géographiques laisse pressentir les bénéfices de l'union entre les *Open data* et les données télématiques dans la modélisation du risque géographique.

5.6.1 Comparaisons statistiques des performances des modèles : Open data et hybride

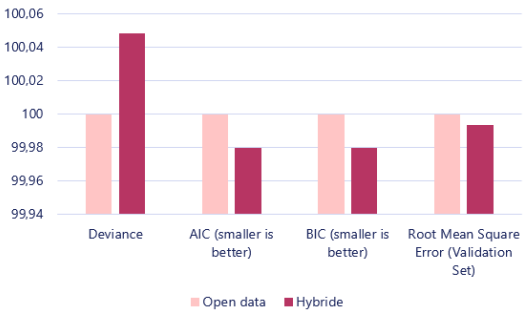

FIGURE 5.27 – *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle Open data)*

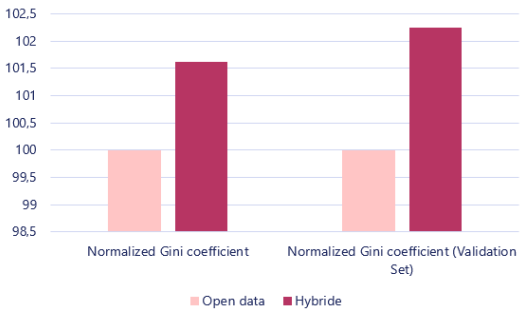

FIGURE 5.28 – *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle Open data)*

Les figures 5.27 et 5.28 montrent que le modèle hybride est globalement de meilleure qualité que le modèle *Open data*. Toutefois, il convient de noter que les différences de performances entre ces deux modèles sont véritablement très faibles et qu'en terme de déviance les tendances des performances s'inversent.

5.6.2 Comparaison du spread des zoniers : Open data et hybride

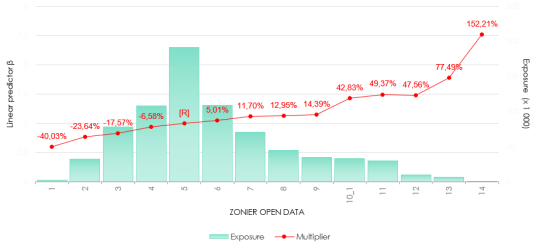

FIGURE 5.29 – *Courbe des facteurs multiplicatifs des modalités du zonier Open data*

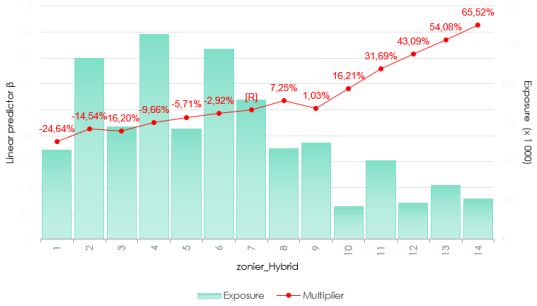

FIGURE 5.30 – *Courbe des facteurs multiplicatifs des modalités du zonier hybride*

Les tendances croissantes des facteurs multiplicatifs associés respectivement aux zones de chacun des zoniers prouvent la cohérence de ceux-ci vis à vis du *burning cost*. Ils représentent donc effectivement des variables significatives dans l'explication de ce dernier. Le tableau suivant (Figure 5.31) permettra néanmoins de déterminer lequel des deux zoniers discrimine le mieux le risque impactant ce *burning cost*.

	Zonier Open Data	zonier Hybride
Multiplier 1	0,5997	0,7536
Multiplier 14	2,5221	1,6552
Exposition 1	1917	45803
Exposition 14	540	21004
spread standard	**76,22%**	**54,47%**
spread normalisé	**21,47%**	**24,98%**

FIGURE 5.31 – *Tableau récapitulatif des comparaisons de spread des zoniers Open data et hybride*

Au niveau des écarts de *spread*, la différence entre les deux zoniers reste encore très infime. Se fier aux *spreads* standards reviendrait à considérer que le zonier *Open data* discrimine mieux le risque que le zonier hybride. Pourtant en prenant en compte l'équilibre des expositions sur les zones extrêmes, la formule normalisée du *spread* donne l'avantage au zonier hybride face au zonier *Open data*.

L'analyse menée dans cette dernière section présente plusieurs limites, entre autres :

- Les comparaisons effectuées se basent sur une seule approche de modélisation du risque géographique qui est le zonier.
- Le nombre de données *Open data* utilisées est assez réduit par contrainte de réutilisation du zonier *Open data* déjà à disposition.

- Le modèle racine utilisé est une modélisation du *burning cost* par contrainte de réutilisation du zonier *Open data* déjà à disposition. Cette modélisation met moins en valeur l'impact géographique qu'une modélisation de fréquence de sinistres.

Cependant, malgré ces limites, cette analyse a permis de montrer que la fusion *Open data* - données télématiques est une alliance très prometteuse. Elle mérite donc d'être approfondie afin d'affiner la connaissance du risque géographique en assurance automobile et de profiter des avantages de ces deux types de sources de données très complémentaires.

Chapitre 6

Évaluation actuarielle : Quels impacts sur la prime de l'assureur ?

« Eclairer les risques, tracer l'avenir »

Devise de l'Institut des Actuaires de France

Dans les chapitres 4 et 5, l'évaluation de l'apport des données télématiques dans la modélisation du risque géographique s'est consacrée à l'analyse des indicateurs statistiques de la qualité des modèles.

Cependant, au delà de ces indicateurs statistiques, ce qui importe pour l'assureur c'est d'évaluer l'impact opérationnel d'une modélisation intégrant ces nouvelles données télématiques, sur sa segmentation du territoire et sur son tarif.

L'objectif de ce chapitre est donc d'étudier les déformations de la segmentation et du tarif engendrées par la prise en compte des données télématiques dans la modélisation du risque géographique de l'assureur.

Cette étude se basera sur les modèles référent et innovant développés au chapitre 5. Les primes pures analysées ici, sont les résultantes du produit des prédictions de ces différents modèles et des prédictions du modèle de coûts moyens présenté au chapitre 3.

6.1 Analyse des migrations entre les zoniers

Supposons que l'assureur de l'étude utilise pour sa modélisation du risque géographique un zonier standard tel que le zonier traditionnel construit au chapitre précédent. Pour rappel, le zonier traditionnel a été construit en traitant l'information géographique contenue dans les résidus du modèle de fréquence de sinistres par un lissage géospatial par crédibilité.

Cet assureur, désireux d'améliorer sa segmentation géographique du risque, décide de construire un nouveau zonier beaucoup plus innovant dont la construction se base cette fois-ci sur l'apprentissage statistique de données télématiques géospatialisées externes.

Une fois la construction de ce nouveau zonier achevée et après avoir apprécié statistiquement l'apport de celui-ci dans son modèle de fréquence de sinistres, l'attention de l'assureur se porte sur la déformation de ces zones de risque lors du passage de son zonier traditionnel actuel à son nouveau zonier innovant.

D'un côté, une grande quantité de changements importants de zones (exemple : passage de la zone la moins risquée du zonier traditionnel à la zone la plus risquée du zonier innovant) serait très difficile à expliquer et à accepter par l'assureur dans le sens où le nouveau zonier changerait complètement la structure de segmentation géographique de ce dernier. D'un autre côté, une quantité conséquente de changements très faibles de zones (exemple : décalage d'une zone) pourrait inciter l'assureur à conserver son zonier traditionnel. La correction apportée par le zonier innovant ne doit donc être ni trop prononcée, ni trop faible.

6.1.1 Switch entre deux zoniers

Le *Switch* est un indicateur proposé dans ces travaux, dont le but est de quantifier les migrations des communes entre les différentes zones de deux zoniers. Pour faciliter son calcul, les deux zoniers comparés doivent respecter certains critères :

- ils doivent avoir le même nombre de zones de risque ;
- leurs zones de risque doivent être ordonnées dans le même sens suivant les valeurs du risque sur chacune d'entre elles (ordre croissant de préférence) ;
- les noms des zones de risque doivent être des numéros.

Soient Z et Z' deux zoniers à comparer. Le *Switch* de la commune i entre ces deux zoniers s'obtient en faisant la différence des numéros des zones dans lesquelles cette commune se trouve dans ceux-ci :

$$Switch_i = z_i - z'_i$$

avec z_i et z'_i les numéros respectifs des zones auxquelles appartient la commune i dans les zoniers Z et Z'.

Premier cas : $switch_i = 0$

Dans ce cas, les deux zoniers comparés s'accordent sur la classification du risque sur la commune i.

Second cas : $switch_i = k$ **et** $k < 0$ **(respectivement** $k > 0$**)**

Dans ce cas, il y a divergence entre les deux zoniers comparés sur la classification du risque sur la commune i. Le zonier Z' considère que le risque sur la commune i mérite d'être catégorisé dans une classe de risque plus élevé (respectivement moins élevé) que celle proposée par le zonier Z. Il y a donc sur-classement (respectivement déclassement) du risque de k classes.

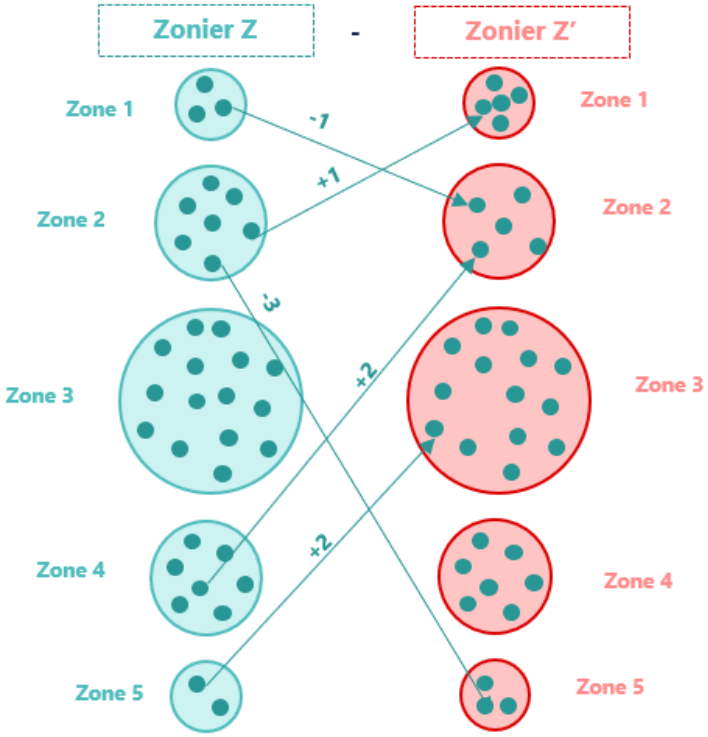

FIGURE 6.1 – *Illustration du switch entre deux zoniers Z et Z'*

6.1.2 Application : calcul et évaluation du switch entre les zoniers traditionnel et innovant

Les zoniers traditionnel et innovant respectant les critères précédemment énumérés, l'assureur procède aux calcul du *Switch* entre ces derniers.

$$Switch = zonier_{traditionnel} - zonier_{innovant}$$

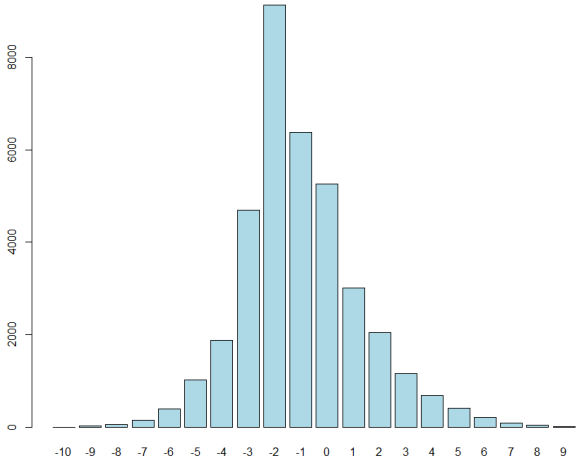

FIGURE 6.2 – *Distribution du switch entre les zoniers traditionnel et innovant*

La figure 6.2 montre que la répartition du *Switch* entre les zoniers traditionnel et innovant est concentrée entre -3 et +2 avec un pique en -2. La déformation apportée par le zonier innovant n'est donc ni trop forte ni trop faible. Le mode de cette série statistique discrète valant -2, cela signifie que sur la plupart des communes, le zonier innovant suggère un sur-classement du risque de deux classes par rapport à la classe proposée par le zonier traditionnel. Il est possible d'analyser le *Switch* à l'aide d'une cartographie.

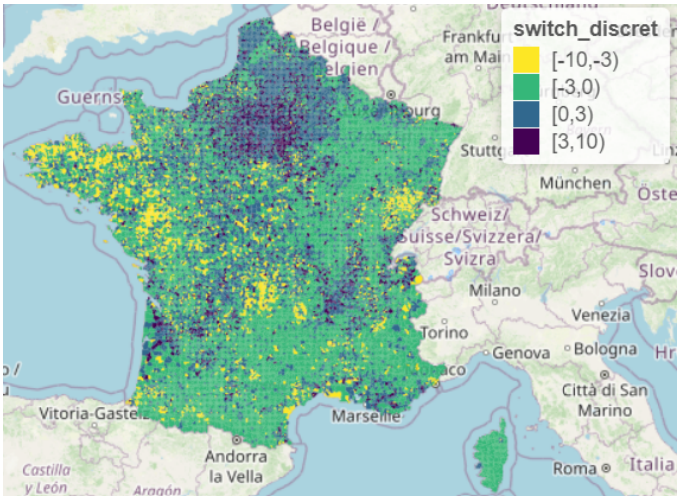

FIGURE 6.3 – *Cartographie du switch entre les zoniers traditionnel et innovant*

D'après la cartographie (Figure 6.3), les classes de *Switch* extrêmes se localisent principalement sur la Bretagne (sur-classement extrême) et sur l'île de France (déclassement extrême). Dans la majeure partie du reste de la France, le *Switch* est assez modéré, variant entre -3 et +2.

Face à cette cartographie, l'assureur reste un peu perplexe. En effet, il est vrai que le *Switch* semble assez modéré sur une grande partie du territoire. Néanmoins, les quelques parties touchées par les *Switch* extrêmes bien que très faibles en nombre, semblent représenter une part non négligeable de l'exposition de celui-ci, eu égard à leur localisation.

Cette dernière analyse motive l'assureur à évaluer l'impact de ces différents *Switch* sur sa prime pure afin de se convaincre de leur réelle importance.

6.2 Impact des changements de zones sur la prime pure de l'assureur

Pour mesurer l'impact des *Switch* entre les zoniers traditionnel et innovant sur sa prime pure, l'assureur commence par calculer cette dernière suivant ses différents modèles.

Il effectue les calculs suivants :

$$Prime_{obs} = Coût_{obs}/Exposition$$

$$Prime_{ref} = Freq_{ref} * CM$$

$$Prime_{inno} = Freq_{inno} * CM$$

où :

- $Prime_{obs}$ représente les *burning cost*, c'est à dire les valeurs de primes pures qui auraient été réellement payées si le cycle de production assurantiel n'était pas inversé ;
- $Coût_{obs}$ représente les valeurs de coûts réellement endossés par l'assureur ;
- CM représente les coût moyens estimés (voir chapitre 3) ;
- $Prime_{ref}$ représente les primes pures estimées par la structure tarifaire contenant le zonier traditionnel ;
- $Freq_{ref}$ représente les prédictions du modèle référent de fréquence de sinistres ;
- $Prime_{inno}$ représente les primes pures estimées par la structure tarifaire contenant le zonier innovant ;
- $Freq_{ref}$ représente les prédictions du modèle innovant de fréquence de sinistres.

Puis il moyennise ces différentes primes calculées sur les communes qu'il couvre.

6.2.1 Delta entre deux primes

Le delta entre deux primes est la quantification de la variation entre celles-ci. Dans ces travaux, le delta calculé est multiplicatif de façon à l'interpréter plus facilement. Soient P et P' deux structures de primes, le delta entre celles-ci sur une commune i s'obtient en faisant :

$$Delta_i = \frac{p_i}{p'_i}$$

où p_i et p'_i sont les primes moyennes sur la commune i estimées respectivement par les structures P et P'.

Premier cas : $Delta_i = 1$

Dans ce cas, il y a accord entre les deux structures de primes quant à la prime moyenne estimée sur la commune i.

Second cas : $Delta_i = \alpha$ **avec** $\alpha < 1$ **(respectivement** $\alpha > 1$**)**

Dans ce cas, il y a divergence entre les deux structures de primes quant à la prime moyenne estimée sur la commune i. La structure P considère que la prime moyenne p'_i estimée par la structure P' sur la commune i devrait être diminuer (respectivement augmenter) d'un facteur α.

6.2.2 Application : calcul et évaluation du delta entre les primes innovantes et les primes de référence

L'assureur procède au calcul du delta entre les primes innovantes et les primes de référence :

$$Delta = \frac{Prime_{inno}}{Prime_{ref}}$$

La figure 6.4 présente la distribution du delta entre les primes innovantes et les primes de référence. Le delta varie entre 0.25 et 3.75 (bornes extrêmes très faiblement représentées). Il se concentre dans l'intervalle [0.5 ;1.5] et sa valeur moyenne est de 1. cette moyenne de 1 signifie que, les structures tarifaires innovante et de référence sont en accord sur les primes proposées sur la plupart des communes couvertes par l'assureur. Néanmoins, la concentration du delta dans l'intervalle [0.5 ;1.5] montre qu'en moyenne, la minoration (respectivement majoration) maximale suggérée par la structure innovante revient à diviser(respectivement multiplier) le tarif de référence par 2, ce qui paraît acceptable.

Cependant l'analyse des *switch* avait montré que les cas extrêmes n'étaient pas à négliger. L'assureur décide donc de représenter sur une carte le delta, afin de vérifier si les delta extrêmes correspondent aux zones de *switch* extrême.

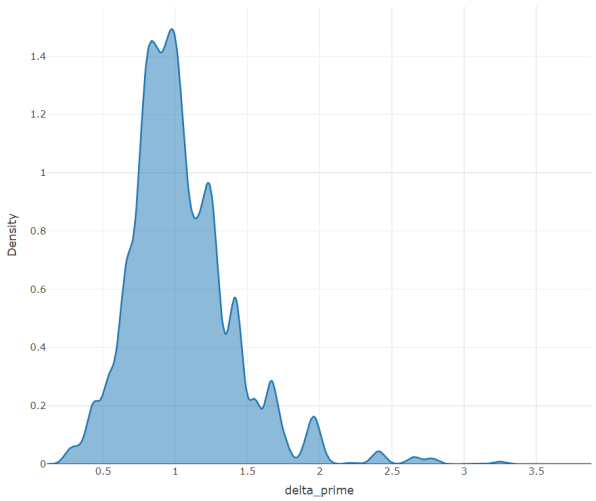

FIGURE 6.4 – *Distribution du delta entre les primes innovantes et les primes de référence*

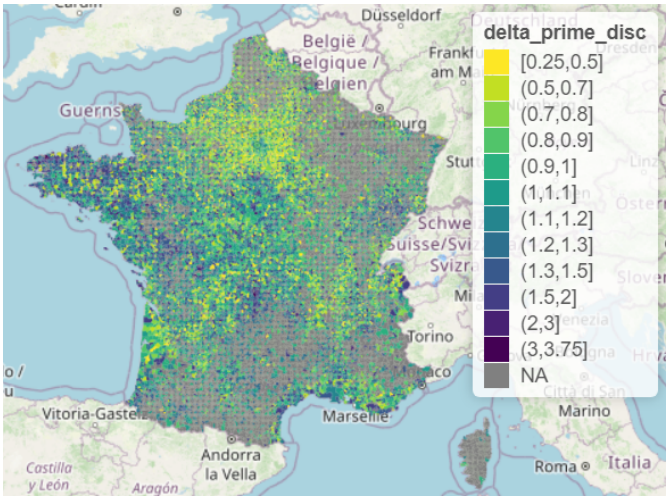

FIGURE 6.5 – *Cartographie du delta entre les primes innovantes et les primes de référence*

La cartographie du delta (Figure 6.5) confirme ce qui a été dit plus haut. En effet, sur la majeure partie des communes couvertes par l'assureur, le delta varie effectivement entre 0.5 et 1.5. Toutefois, les zones sur lesquelles le delta est extrême semblent correspondre exactement aux zones où les *switch* sont aussi extrêmes.

Face à ce dilemme, l'assureur décide de mener une dernière analyse qui lui permettra de tirer des conclusions définitives. Dans cette dernière analyse, il introduit deux indicateurs très importants. Ces sont :

- les expositions totales qui lui servent d'indicateurs d'importance ;
- les *burnig cost* qui représentent les "vraies" valeurs de primes auxquelles doivent se

rapprocher les primes estimées par les différentes structures tarifaires.

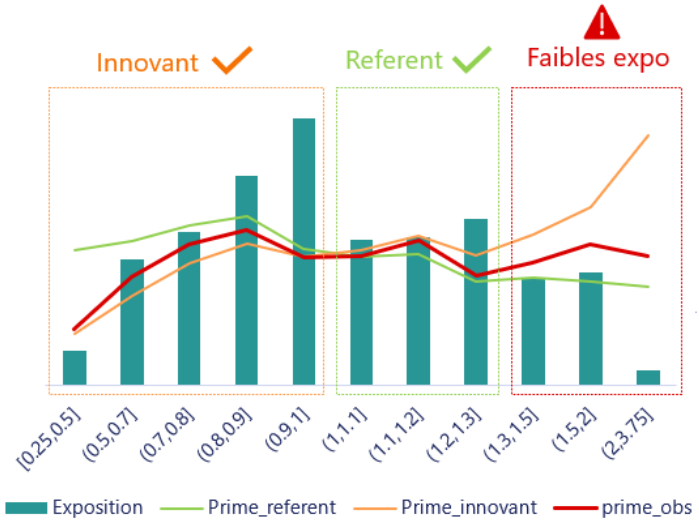

FIGURE 6.6 – *Comparaison des primes innovantes et de référence relativement aux burning cost*

Sur la figure 6.6 :

- les barres de couleur vert-foncée représentent les expositions totales sur les différentes classes de delta ;
- la courbe de couleur rouge passe par les *burning cost* moyens sur les différentes classes de delta ;
- la courbe de couleur orange passe par les primes innovantes moyennes sur les différentes classes de delta ;
- la courbe de couleur vert-claire passe par les primes de référence moyennes sur les différentes classes de delta.

Sur cette figure, trois grands intervalles de delta se dégagent.

les valeurs de delta inférieures à 1

C'est l'intervalle de delta le plus exposé. Sur cet intervalle, la structure tarifaire innovante suggère de baisser les primes de référence. Cette suggestion paraît objective puisque la courbe des primes innovantes se rapproche plus de celle des *burning cost* que celle des primes de référence. Sa prise en compte est d'autant plus importante du fait du niveau d'exposition très élevé sur cet intervalle. Cela montre que les *switch* extrêmes (déclassements extrêmes du risque) et les delta très faibles observés sur certaines communes de l'Île-de-France paraît justifier. La structure tarifaire de référence surestime beaucoup trop les niveaux de primes sur ces zones.

les valeurs de delta comprises entre 1 et 1.3

C'est un intervalle de delta moyennement exposé. Sur cet intervalle, la structure tarifaire innovante suggère de majorer les primes de référence. Cette suggestion est moins pertinente que la précédente. En effet, sur cet intervalle, c'est la courbe des primes de référence qui se rapproche le plus de celle des *burning cost*. Les primes proposées par la structure tarifaire de référence semblent donc plus appropriées sur cet intervalle que celles proposées par la structure tarifaire innovante.

les valeurs de delta supérieures à 1.3

L'exposition totale sur cet intervalle de delta est très faible en comparaison aux deux premiers intervalles. Se décider sur ce genre de zones est très délicat pour l'assureur vue que sa marge de confiance est très faible. D'une part, Privilégier la structure tarifaire de référence qui proposerait des valeurs de primes très faibles comparées aux valeurs de *burning cost* pourrait être préjudiciable pour le ratio S/P [1] de l'assureur sur cet intervalle. D'autre part, se fier à la structure tarifaire innovante qui proposerait des valeurs de primes très élevées comparées aux valeurs de *burning cost* pourrait faire encourir un risque à l'assureur, dans la mesure où cette sur-tarification n'est pas fiable compte tenu du manque d'exposition.

Décision finale de l'assureur

l'analyse du graphique 6.6 a permis à l'assureur de tirer une conclusion très importante. Le but de la modélisation innovante n'est pas de remplacer drastiquement sa modélisation de référence, mais plutôt de l'améliorer là où elle souffre d'un manque de précision. A partir de cette analyse, l'assureur est en mesure de proposer une structure de primes pures corrigées et plus adaptées. Une vision économique de cette correction de la prime de référence par la prime innovante pourrait par exemple être une moyenne de ces deux types de primes. Les primes pures corrigées résultantes sont observables sur la figure 6.7.

1. ratio S/P est le rapport des charges totales de sinistres décaissées par les primes encaissées

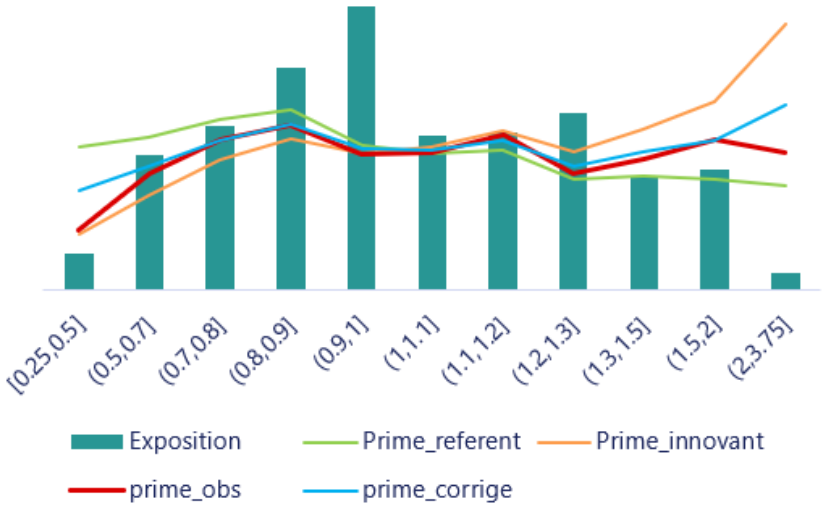

FIGURE 6.7 – *Représentation des primes pures corrigées*

Conclusion

Bilan des travaux

Ce mémoire s'inscrit dans une dynamique de recherche d'outils innovants permettant d'améliorer la compréhension du risque en assurance automobile. Le sujet traité a posé comme problématique l'évaluation de pertinence et de la significativité d'une modélisation du risque géographique basée sur des données télématiques, reflétant les habitudes de conduite dans les différentes zones du territoire.

Cette évaluation de l'apport des données télématiques dans la modélisation du risque géographique s'est articulée autour de deux principaux aspects : l'aspect technique et l'aspect opérationnel.

L'évaluation technique a consisté à analyser les performances statistiques de plusieurs modèles suivant différentes approches de modélisation du risque géographique intégrant les données télématiques.

Dans un premier temps, cette évaluation technique s'est appuyée sur une modélisation naïve du risque géographique. Dans cette approche, ce risque a été formalisé par une combinaison de variables télématiques intégrées directement dans la modélisation initiale (hors données géographiques) de la fréquence de sinistres. A l'issu de cette première évaluation technique, il est ressorti que l'intégration de variables télématiques géospatialisées à l'intérieur du modèle initial a favorisé l'amélioration de sa qualité statistique et de sa précision géographique.

Dans un second temps, l'évaluation technique de l'apport des données télématiques a porté sur une modélisation du risque géographique par zonier. Deux types de zonier ont été construits : un zonier traditionnel, résultat d'un lissage géospatial des résidus géographiques et un zonier innovant, conçu par apprentissage statistique d'une sélection de données télématiques. La comparaison des modèles contenant ces deux zoniers a montré que le zonier innovant permettait de mieux discriminer le risque géographique que le zonier traditionnel.

Ces résultats techniques corroborent la pertinence de l'utilisation des données télé-

matiques pour la modélisation du risque géographique en assurance automobile.

Au niveau du volet opérationnel, il était question d'évaluer l'impact qu'une modélisation du risque géographique se basant sur les données télématiques pouvait avoir sur le tarif proposé par l'assureur. Pour y parvenir, une analyse des variations des primes induites par les migrations de zones entre le zonier traditionnel et le zonier innovant a été menée. Cette analyse a montré qu'il était possible d'ajuster les primes pures proposées par l'assureur en se basant sur les informations apportées par les données télématiques synthétisées dans le zonier innovant.

Cette évaluation opérationnelle confirme donc la significativité d'une modélisation du risque géographique basée sur les habitudes de conduite dans les différentes zones.

En somme, toute ces analyses semblent converger vers une unique et même conclusion : la télématique est bel et bien un outil innovant auquel les assureurs automobile peuvent s'adosser afin de mieux appréhender le risque géographique.

Limites et perspectives d'amélioration

Il convient cependant, de nuancer cette conclusion générale. D'une part, les travaux ayant été réalisés uniquement sur la garantie RCM d'un seul portefeuille d'assureur, il ne serait pas objectif de les considérer comme une tautologie. D'autre part, il est important de souligner certaines limites intrinsèques aux données utilisées dans le cadre de ces travaux. D'abord, au niveau de la base de données externes *Smart Road Data*, la non connaissance des formules ayant permis la construction et l'agrégation des différents scores de comportement constitue un réel frein dans la transparence de leur utilisation. Ensuite, la version de la base à disposition pour ces travaux correspondant à la période 2020-2021 révèle deux principales insuffisances :

- cette période est fortement impactée par la crise sanitaire due à la pandémie de la COVID-19.
- cette période ne correspond pas à celle des données internes utilisées.

Enfin, au niveau des données internes de l'assureur, il aurait été plus pertinent de posséder et d'utiliser les codes IRIS rattachés aux lieux des différents sinistres enregistrés par police plutôt que des codes IRIS en lien avec leur domicile.

Toutefois, il n'empêche que ces résultats restent très encourageants et que le sujet mérite d'être investigué davantage. En guise de pistes d'amélioration, il serait par exemple possible :

- de mener la même étude sur d'autres garanties d'assurance automobile afin de juger de la pertinence des résultats suivant différentes visions ;
- d'affiner la maille géographique des données télématiques en la faisant passer de la maille INSEE à une maille au trajet ;

- de tester une approche naïve de modélisation du risque géographique basée sur les données télématiques mais qui cette fois-ci utilisera comme modèles de prédiction des algorithmes d'apprentissage statistique ;
- d'améliorer la technique de lissage utilisée au niveau de la construction des zoniers, en privilégiant des algorithmes de lissage prédictif tel que le krigeage.

Ouverture

Au delà de la modélisation du risque géographique, les données télématiques ne pourraient-elles pas être considérées comme une solution de simplification des parcours de souscription en assurance automobile ?

Annexe A

Tables descriptives des variables

Nom de la variable	Description
s2_anticipation_brakingDistance	Score evaluant la distance de freinage
s2_anticipation_brakingEvent	Score evaluant la fréquence de freinage
s2_anticipation_brakingIntensity	Score evaluant l'intensité de freinage
s2_anticipation_brakingSudden	Score evaluant les freinages brusques
s2_pace_straight	Score evaluant l'allure de la conduite en ligne droite
s2_pace_turn	Score evaluant l'allure de la conduite dans les virages
s3_anticipation	Score global de freinage
s3_pace	Score global d'allure de la conduite
global	score global de conduite

FIGURE A.1 – *Tableau des différents scores de comportement*

Contexte	Nom de la variable	Description
GENERAL	nb_vid	Nombre de véhécules passés sur la commune
	nb_trips	Nombre de trajets effectués sur la commune
	nb_trips_accel_valid	Nombre de trajets effectués sur la commune avec un accélérateur valide
	duration_total	Total des temps de trajets sur la commune
	km_total	Total des Kilomètres parcourus sur la commune
METEO	WET_weather_total_duration	Total des temps de trajets en temps humide
	WET_weather_total_km	Total des Kilomètres parcourus en temps humide
	DRY_weather_total_duration	Total des temps de trajets en temps sec
	DRY_weather_total_km	Total des Kilomètres parcourus en temps sec
RELIEF	PLAIN_tri_total_duration	Total des temps de trajets dans les plaines
	PLAIN_tri_total_km	Total des Kilomètres parcourus dans les plaines
	HILLY_tri_total_duration	Total des temps de trajets dans les vallées
	HILLY_tri_total_km	Total des Kilomètres parcourus dans les vallées
	MOUNTAINOUS_tri_total_duration	Total des temps de trajets dans les montagnes
	MOUNTAINOUS_tri_total_km	Total des Kilomètres parcourus dans les montagnes
URBANITE	URBAN_total_duration	Total des temps de trajets sur les routes urbaines
	URBAN_total_km	Total des Kilomètres parcourus sur les routes urbaines
	EXTRA_URBAN_total_duration	Total des temps de trajets sur les routes extra urbaines
	EXTRA_URBAN_total_km	Total des Kilomètres parcourus sur les routes extra urbaines
	MOTORWAY_total_duration	Total des temps de trajets sur les autoroutes
	MOTORWAY_total_km	Total des Kilomètres parcourus sur les autoroutes
COURBURE DE ROUTE	LARGE_CURVE_total_duration	Total des temps de trajets dans les grands virages
	LARGE_CURVE_total_km	Total des Kilomètres parcourus dans les grands virages
	MODERATE_CURVE_total_duration	Total des temps de trajets dans les virages modérés
	MODERATE_CURVE_total_km	Total des Kilomètres parcourus dans les virages modérés
	TIGHT_CURVE_total_duration	Total des temps de trajets dans les virages sérrés
	TIGHT_CURVE_total_km	Total des Kilomètres parcourus dans les virages sérrés
	STRAIGHT_CURVE_total_duration	Total des temps de trajets en ligne droite
	STRAIGHT_CURVE_total_km	Total des Kilomètres parcourus en ligne droite

FIGURE A.2 – *Tableau des différentes variables concernant l'usage du véhicule*

Aspect	Nom de la variable	Description
URBANITE	URABN	Total des distances des routes urbaines dans la commune
	INTER	Total des distances des routes inter-urbaines dans la commune
	HIGHWAY	Total des distances des autoroutes dans la commune
COURBURE DE ROUTE	CURVE_RIGHT	Total des distances des virages à droite dans la commune
	CURVE_LEFT	Total des distances des virages à gauche dans la commune
	STRAIGHT	Total des distances des lignes droites dans la commune

FIGURE A.3 – *Tableau des variables cartographiques*

Catégorie	Nom de la variable	Description
SCORES DE COMPORTEMENT	s2_anticipation_brakingDistance	Score evaluant la distance de freinage
	s2_anticipation_brakingEvent	Score evaluant la fréquence de freinage
	s2_anticipation_brakingIntensity	Score evaluant l'intensité de freinage
	s2_anticipation_brakingSudden	Score evaluant les freinages brusques
	s2_pace_straight	Score evaluant l'allure de la conduite en ligne droite
	s2_pace_turn	Score evaluant l'allure de la conduite dans les virages
	s3_anticipation	Score global de freinage
	s3_pace	Score global d'allure de la conduite
	global	score global de conduite
USAGE	roulage_moyen	Ratio entre le nombre de trajets effectués et le nombe de véhicules de passages
	nb_trips	nombre de trajets effectués sur la commune
	tx_nb_trips_accel_valid	Taux de trajets effectués avec un accéléromètre valide
	total_km	Total des kilomètres parcourus sur la commune
	mean_speed	Vitesse moyenne sur la commune
	PLAIN_tri_part_of_km	Taux de kilomètres parcourus dans les plaines
	PLAIN_tri_mean_speed	Vitesse moyenne dans les plaines
	EXTRA_URBAN_part_of_km	Taux de kilomètres parcourus dans sur le routes extra-urbaines
	EXTRA_URBAN_mean_speed	Vitesse moyenne sur les routes extra-urbaines
	URBAN_part_of_km	Taux de kilomètres parcourus sur les routes urbaines
	URBAN_mean_speed	Vitesse moyenne parcourus sur les routes urbaines
	LARGE_CURVE_part_of_km	Taux de kilomètres parcourus dans les grands virages
	LARGE_CURVE_mean_speed	Vitesse moyenne parcourus dans les grands virages
	STRAIGHT_CURVE_part_of_km	Taux de kilomètres parcourus en ligne droite
	STRAIGHT_CURVE_mean_speed	Vitesse moyenne en ligne droite
	TIGHT_CURVE_part_of_km	Taux de kilomètres parcourus dans les virages sérrés
	TIGHT_CURVE_mean_speed	Vitesse moyenne dans les virages sérrés
	WET_weather_part_of_km	Taux de kilomètres parcourus en temps humide
	WET_weather_mean_speed	Vitesse moyenne en temps humide
SAISONNALITE	hiver	Densité du trafic en hiver
	printemps	Densité du trafic au printemps
	été	Densité du trafic en été
	automne	Densité du trafic en automne
CARTOGRAPHIQUE	part_of_CURVE	Taux de virage dans la commune
	part_of_inter	Taux de route inter-urbaine dans la communne
	part_of_urban	Taux de route urbaine dans la commune
	highway_presence	Presence d'autoroute
	mountain_presence	Presence de montagne

FIGURE A.4 – *Tableau des variables conservées après la sélection non-supervisée*

Annexe B

Splines

Définition mathématique

Soit $[a,b]$ un intervalle de $\mathbb{R}$ et soit une partition de cet intervalle en $k+1$ sous-intervalles :

$$a = t_0 < t_1 < ... < t_k < t_{k+1} = b$$

Une *spline* S de degré n avec k noeuds, est une fonction de $[a,b] \to \mathbb{R}$ qui vérifie les deux propositions suivantes :

1. La restriction de S à un sous-intervalle i est un polynôme de dégré n

   $$P_i : [t_{i-1},t_i] \to \mathbb{R}$$

   tel que :

   $$\begin{aligned} S(t) &= P_1(t) && \forall t \in [t_0,t1] \\ S(t) &= P_2(t) && \forall t \in [t_1,t2] \\ &\cdots \\ S(t) &= P_{K+1}(t) && \forall t \in [t_K,tk+1] \end{aligned}$$

2. S est C_{n-1} sur $]a,b[$, c'est à dire continûment dérivable jusqu'à l'ordre $n-1$ sur tout intervalle ouvert et en particulier aux noeuds intérieurs $\{t_1,t_2,\ldots,t_k\}$:

   $$S^j(t_i^-) = S^j(t_i^+) \quad \forall j \in [|0,n-1|] \quad \forall i \in [|1,k|].$$

Utilisation en régression

L'utilisation d'une transformation par *spline* $x \mapsto S(x)$ dans un modèle de régression, passe par l'estimation d'un coefficient par dimension, ou degré de liberté de la *spline*. Pour cela, il est nécessaire de la décomposer en B-*spline* élémentaires, à l'aide d'un algorithme tel que celui de *Boor* :

$$S = \sum_{i+1}^{k+n+1} \lambda_i B_i$$

En outre, il convient ici de noter que l'une des composantes de la *spline* correspond à la constante et ne dépend donc pas de x. Or cette constante est déjà matérialisée dans le modèle par l'intercepte. Le nombre total de composantes à considérer est donc diminué de 1.

L'intégration d'une *spline* de degré n à k noeuds au modèle se traduit donc par l'ajout de $k + n$ variables numériques : $(B_1(x), B_2(x), \ldots, B_{k+n}(x))$.

Bibliographie

Livres et articles

[1] Institut des Actuaires. *La Gender Directive un an après*. Actuariel. Mars 2014.

[3] France Assureurs. *Le marché de l'assurance automobile des particuliers en 2020 / Assurances de biens et de responsabilité*. Juin 2021.

[5] Arthur CHARPENTIER et Christophe DUTANG. *L'Actuariat avec R*. Décembre 2012.

[6] Christophe CHESNEAU. *Introduction aux arbres de décision (de type CART)*. Master. FRANCE, 2020. isbn : cel-02281064v3.

[8] P.A. CORNILLON et al. *R pour la statistique et la science des données*. Pratique de la statistique. Presses Universitaires de Rennes. isbn : 978-2-7535-7573-8.

[9] Pierre-Adre CORNILLON et Eric MATZNER-LOBER. *Régression avec R*. FRANCE : Springer-Verlag, 2011. isbn : 978-2-8178-0183-4.

[13] P. McCULLAGH et J.A. NELDER. *Generalized Linear Models*. 2e éd. Monographs on Statistics and Applied Probability 37. Chapman et Hall.

[14] Eliade MICU. *Territorial ratemaking*. EagleEye Analytics. Mars 2012.

[16] Frédéric PLANCHET et Maxime BEN-BRIK. *Quelques réflexions sur la segmentation en assurance*. Juin 2019.

[20] ITN Consultants SA. *Livre blanc « PAY AS YOU DRIVE »*. Décembre 2008.

[25] Gary WANG. *Territory analysis updates to the traditionnal methods*. EagleEye Analytics. Mars 2012.

Cours et mémoires

[2] Pierre AILLIOT. *Cours Analyse de Données*. UBO. Février 2020.

[4] Silvia BUCCI. *Étude et implémentation de techniques d'analyse de sensibilité dans les modèles de tarification Non-Vie. Application à la tarification à l'adresse*. Mémoire d'actuariat. ENSAE. Mars 2021.

[7] Christian CHOW. *Utilisation de données télématiques pour l'analyse de la sinistralité automobile*. Mémoire d'actuariat. ISUP. 2019.

[10] J-P CROISILLE. *Interpolation Spline*. Université Paul Verlaine-Metz. 2008.

[11] Sophie KRANZLIN. *Modélisation du risque géographique en assurance automobile*. Mémoire d'actuariat. ENSAE. Mars 2017.

[12] Kevin LECOMTE. *Automatiser la comparaison de modèles : Application sur l'amélioration d'un modèle de fréquence par des techniques de machine learning*. Mémoire d'actuariat. ENSAE. février 2018.

[15] Claire NICOLLE. *Tarification au trajet à l'aide de l'Open Data*. Mémoire d'actuariat. EURIA. Sept. 2017.

[17] Frédéric PLANCHET et Antoine MISERAY. *Tarification IARD Introduction aux techniques avancées*. ISFA. Mars 2017.

[18] Henri QIU. *Étude sur l'apport des données et du score de conduite en assurance automobile télématique*. Mémoire d'actuariat. ENSAE. Nov. 2016.

[21] Kevin SADOUN. *Apport des télématiques dans la segmentation tarifaire en assurance automobile*. Mémoire d'actuariat. Université Paris-Dauphine. Jan. 2016.

[23] Franck VERMET. *Arbres de décision et méthodes ensemblistes*. EURIA. 2020.

[24] Edouard VICAIRE. *L'open data et les réseaux neuronaux : vers une amélioration de la prédictibilité des sinistres ?* Mémoire d'actuariat. ENSAE. Nov. 2017.

[26] Fatima-Zohra ZOUGGAGH. *Tarification automobile à l'aide de modèles de machine learning et apport des données télématiques*. Mémoire d'actuariat. EURIA. Sept. 2018.

Liens web pour en savoir plus sur la télématique

[19] *Qu'est-ce que la télématique ?* GEOTAB. URL : `https://www.geotab.com/fr/blog/qu-est-ce-que-la-t%5C%C3%5C%A9l%5C%C3%5C%A9matique/` (visité le 19/05/2021).

[22] *télématique*. L'Argus de l'assurance. URL : `https://www.argusdelassurance.com/telematique/`.

Table des figures

1 *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle initial sans variables géographiques)* . . . vii
2 *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle initial sans variables géographiques)* vii
3 *Zonier traditionnel* . . . viii
4 *Zonier innovant* . . . viii
5 *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle référent)* . . . viii
6 *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle référent)* . . . viii
7 *Courbe des facteurs multiplicatifs des modalités du zonier traditionnel* . . . ix
8 *Courbe des facteurs multiplicatifs des modalités du zonier innovant* . . . ix
9 *Tableau récapitulatif des comparaisons de spread des zoniers traditionnel et innovant* . . . ix
10 *Distribution du switch entre les zoniers traditionnel et innovant* . . . x
11 *Cartographie du switch entre les zoniers traditionnel et innovant* . . . x
12 *Distribution du delta entre les primes innovantes et les primes de référence* xi
13 *Comparaison des primes innovantes et de référence relativement aux burning cost* . . . xii
14 *Représentation de la prime pure corrigée* . . . xii
15 *Comparison of the quality indicators of the models (base 100 with the values of the indicators of the initial model without geographical data as reference)* . . . xv
16 *Comparison of standardized Gini indices (base 100 with the values of the indicators of the initial model without geographical data as reference)* . . . xv
17 *Traditional risk zoning* . . . xvi
18 *Innovative risk zoning* . . . xvi
19 *Comparison of the quality indicators of the models (base 100 with the values of the indicators of the benchmark model as reference)* . . . xvi
20 *Comparison of normalized Gini indexes (base 100 with the values of the indicators of the benchmark model as reference)* . . . xvi
21 *Curve of multiplicative factors of the modalities of the traditional risk zoning* xvii

22 *Curve of multiplicative factors of the modalities of the innovative risk zoning* xvii
23 *Summary table of spread comparisons of traditional and innovative risk zonings* . . . xvii
24 *Switch distribution between traditional and innovative risk zonings* xviii
25 *Cartography of switch between traditional and innovative risk zonings* . . . xviii
26 *Delta distribution between innovative and benchmark technical premiums* . xix
27 *Comparison of innovative and benchmark premiums relative to burning costs* . . . xx
28 *Representation of technical premium corrected* . . . xx

1.1 *Les familles d'assurance* . . . 6
1.2 *Évolution du trafic routier en France entre Janvier 2020 et Septembre 2021 Source de la donnée : Cerema - Indicateur du trafic routier* . . . 8
1.3 *Niveaux de fréquence de sinistres par garanties et suivant l'année (Base 100 2019)* . . . 9
1.4 *Niveaux de coûts moyens par garanties et suivant l'année (Base 100 2019)* 9
1.5 *Montants de primes proposés par chacun des assureurs* . . . 10
1.6 *Sinistralité observée au cours de l'exercice* . . . 10
1.7 *Exemples de mailles géographiques* . . . 12
1.8 *Illustration shématique du fonctionnement de la télématique automobile* . 13

2.1 *Histogramme des âges des conducteurs sur le portefeuille* . . . 20
2.2 *Histogramme des âges des véhicules sur le portefeuille* . . . 21
2.3 *Évolution de l'exposition et des coûts moyens de la garantie RCM sur le portefeuille suivant l'année (2017 à 2019)* . . . 22
2.4 *Évolution des coûts moyens de la garantie RCM sur le marché de l'assurance automobile en base 100 de 2017 source : Données clés de l'assurance française en 2019 FFA* . . . 22
2.5 *Évolution de l'exposition et des fréquences de sinistres de la garantie RCM sur le portefeuille suivant l'année (2017 à 2019)* . . . 23
2.6 *Évolution des fréquences de sinistres de la garantie RCM sur le marché de l'assurance automobile en base 100 de 2017 (source : Données clés de l'assurance française en 2019 FFA)* . . . 23
2.7 *Les différents niveaux de scores de comportement* . . . 24
2.8 *Cartographie du score S3_anticipation* . . . 25
2.9 *Cartographie du total de kilomètres parcourus à l'intérieur des communes* 25
2.10 *Cartographie du niveau de trafic en Juin* . . . 26
2.11 *Cartographie du niveau de trafic en Août* . . . 26
2.12 *Cartographie des communes traversées par des autoroutes* . . . 27
2.13 *Corrélogramme des données de la base Smart Road Data* . . . 28
2.14 *Corrélogramme des 12 variables de saisonnalité* . . . 29
2.15 *Corrélogramme des variables d'usage et de cartographie* . . . 29
2.16 *Corrélogramme des scores de comportement* . . . 30
2.17 *Représentation des variables dans le premier plan factoriel* . . . 31

2.18 *Représentation du nuage de points des communes dans le premier plan factoriel* . . . 31
2.19 *Représentation du nuage de points des communes dans le premier plan factoriel (avec leur nom)* . . . 32
2.20 *Boîte à moustache du total de kilomètres parcourus avec en rouge la valeur sur Paris* . . . 33
2.21 *Part de variance expliquée par chacune des dix premières composantes principales* . . . 33
2.22 *Tableaux récapitulatif des transformations effectuées (contexte météo)* . . . 35
2.23 *Cartographie des variables avant/après transformations (contexte météo)* . 35
2.24 *Cartes des variables avant/après transformations (courbure de routes)* . . 37
2.25 *Corrélogrammes avant et après transformation - sélection* . . . 38

3.1 *Exemple de courbes de Lorenz/Gini* . . . 48
3.2 *Fonctionnement de la validation croisée k-fold* . . . 49
3.3 *Illustration du compromis biais-variance* . . . 50
3.4 *Évolution de la fréquence de sinistres en fonction de l'âge des conducteurs* 54
3.5 *Évolution de la fréquence de sinistres en fonction de l'âge de véhicules des assurés* . . . 54
3.6 *Comparaison des indicateurs de qualité du modèle GLM initial de fréquence de sinistres et du modèle trivial (base 100)* . . . 56
3.7 *Comparaison des indicateurs de qualité du modèle GLM de coûts moyens et du modèle trivial (base 100)* . . . 57
3.8 *Comparaison des indicateurs de qualité du modèle GLM initial de burning cost et du modèle trivial (base 100)* . . . 58

4.1 *Détermination du paramètre λ par validation croisée* . . . 61
4.2 *Liste des 13 variables sélectionnées pour l'approche naïve* . . . 62
4.3 *Cartographie de la fréquence moyenne de sinistres observée par commune* 63
4.4 *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Score global de freinage* . . . 64
4.5 *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Score global d'allure de conduite* . . . 64
4.6 *Cartographie du score global d'allure de conduite* . . . 64
4.7 *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Total de kilomètres parcourus* . . . 65
4.8 *Analyse des valeurs observées et estimées de la fréquence de sinistres sur la Fréquence de conduite* . . . 66
4.9 *Cartographie de la fréquence de conduite* . . . 66
4.10 *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Taux de parcours en temps humide* . . . 66
4.11 *Cartographie du taux de parcours en temps humide* . . . 66
4.12 *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Taux de routes urbaines* . . . 67

4.13 *Cartographie du taux de routes urbaines* . . . 67
4.14 *Analyse des valeurs observées et estimées de la fréquence de sinistres sur le Taux de virages* . . . 67
4.15 *Cartographie du taux de virages* . . . 67
4.16 *Illustration schématique d'un arbre CART* . . . 69
4.17 *Arbre de décision de la fréquence de sinistres en fonction des variables sélectionnées (approche naïve)* . . . 70
4.18 *Liste des interactions détectées à partir de l'arbre CART* . . . 70
4.19 *Représentation de la surface de réponse du GLM naïf de fréquence de sinistres contenant l'interaction (I_1)* . . . 72
4.20 *Représentation de la surface de réponse du GLM naïf de fréquence de sinistres contenant l'interaction (I_2)* . . . 72
4.21 *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle initial sans variables géographiques)* . . . 73
4.22 *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle initial sans variables géographiques)* 73
4.23 *Cartographie des prédictions de fréquences de sinistres du modèle initial sans variables géographiques* . . . 74
4.24 *Cartographie des prédictions de fréquences de sinistres du modèle naïf* . . 74

5.1 *Méthodologies de construction d'un zonier* . . . 78
5.2 *Superposition de la courbe de la densité empirique des résidus additifs et d'une loi normale* . . . 80
5.3 *Superposition de la courbe de la densité empirique des résidus d'Anscombe et d'une loi normale* . . . 80
5.4 *Cartographie des résidus agrégés : Zoom sur les communes d'Île de France* 81
5.5 *Illustration de l'impact d'un lissage géospatial (cas fictif) source : Territory Analysis Updates to the Traditional Methods, Gary Wang* . . . 82
5.6 *Illustration d'un bon lissage du point de vue assurantiel* . . . 84
5.7 *Tableau des variations de la fonction G* . . . 85
5.8 *Illustration d'une forêt aléatoire : cas d'une régression* . . . 87
5.9 *Optimisation du paramètre b sur la base des inconnus en utilisant le Q^2* . 91
5.10 *Optimisation du paramètre a sur la base des connus en utilisant le Q^2* . . 91
5.11 *Optimisation du paramètre a sur la base des connus en utilisant le $Q^2_{corrige}$* 91
5.12 *Perte d'inertie inter-groupe en fonction du nombre de zones* . . . 92
5.13 *Zonier traditionnel obtenu en découpant les résidus lissés par quantiles* . . 92
5.14 *Zonier traditionnel obtenu en appliquant la CAH sur les résidus lissés* . . 92
5.15 *Liste des 13 variables sélectionnées pour la construction du zonier innovant* 93
5.16 *Évolution de l'erreur de prédiction en fonction du nombre d'arbres ntree* . 94
5.17 *Tableau récapitulatif de l'hyper-paramétrage du modèle de forêt aléatoire* . 94
5.18 *Importance des 13 variables dans le modèle de forêt aléatoire* . . . 95

5.19 *Zonier innovant obtenu en découpant les résidus prédits puis lissés par quantiles* . . . 95
5.20 *Zonier innovant obtenu en appliquant la CAH sur les résidus prédits puis lissés* . . . 95
5.21 *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle référent)* . . . 96
5.22 *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle référent)* . . . 96
5.23 *Courbe des facteurs multiplicatifs des modalités du zonier traditionnel* . . . 98
5.24 *Courbe des facteurs multiplicatifs des modalités du zonier innovant* . . . 98
5.25 *Tableau récapitulatif des comparaisons de spread des zoniers traditionnel et innovant* . . . 98
5.26 *Importance des variables dans le modèle de forêt aléatoire pour la construction du zonier hybride* . . . 100
5.27 *Comparaison des indicateurs de qualité des modèles (base 100 avec pour référence les valeurs des indicateurs du modèle Open data)* . . . 100
5.28 *Comparaison des indices de Gini normalisés (base 100 avec pour référence les valeurs des indicateurs du modèle Open data)* . . . 100
5.29 *Courbe des facteurs multiplicatifs des modalités du zonier Open data* . . . 101
5.30 *Courbe des facteurs multiplicatifs des modalités du zonier hybride* . . . 101
5.31 *Tableau récapitulatif des comparaisons de spread des zoniers Open data et hybride* . . . 101

6.1 *Illustration du switch entre deux zoniers Z et Z'* . . . 105
6.2 *Distribution du switch entre les zoniers traditionnel et innovant* . . . 106
6.3 *Cartographie du switch entre les zoniers traditionnel et innovant* . . . 106
6.4 *Distribution du delta entre les primes innovantes et les primes de référence* 109
6.5 *Cartographie du delta entre les primes innovantes et les primes de référence* 109
6.6 *Comparaison des primes innovantes et de référence relativement aux burning cost* . . . 110
6.7 *Représentation des primes pures corrigées* . . . 112

A.1 *Tableau des différents scores de comportement* . . . 117
A.2 *Tableau des différentes variables concernant l'usage du véhicule* . . . 117
A.3 *Tableau des variables cartographiques* . . . 118
A.4 *Tableau des variables conservées après la sélection non-supervisée* . . . 118