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 22 Septembre 2021

Par : Cédric Denniel
Titre : Lissage des résidus par Krigeage dans la création d'un zonier : Application sur un portefeuille MRH

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Sophie BOURDET
Romain LAILY
Bertrand DESCHAMPS
Signature :

Entreprise :
Addactis France
Signature :

Membres présents du jury de l'EURIA :
Daniel BOIVIN

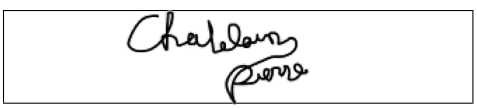
Directeur de mémoire en entreprise :
Pierre Chatelain
Signature :

Invité :

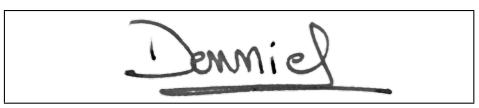
Signature :

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

Résumé

L'objectif du mémoire est de revisiter la méthodologie classique de construction d'un zonier, notamment en y intégrant une nouvelle approche pour le lissage des résidus : le Krigeage. On propose de passer d'un lissage classique à un lissage « prédictif » où il sera possible à la fois de choisir une structure de lissage adéquate aux données, mais également d'ajouter des informations complémentaires à la géolocalisation du portefeuille lors du lissage.

La démarche de ce mémoire est de comparer le lissage par Krigeage avec le lissage par crédibilité et la prédiction des résidus par XGBoost (et des données externes), deux méthodes classiques pour le zonier. Cette comparaison se basera sur un portefeuille Multi-Risques Habitation et permettra de mettre en lumière les atouts du Krigeage, mais aussi les liens avec les autres méthodes de lissage et ses limites.

Ce mémoire propose une méthodologie d'optimisation des paramètres pour avoir un meilleur lissage des résidus et un moyen de réduire les temps de calcul du modèle.

Mots clefs: Zonier, Krigeage, Multi-Risque habitation, Garantie Dégâts des Eaux, Résidus, GLM, XGBoost, Lissage par crédibilité, Lissage prédictif, Données Externes

Abstract

The aim of the thesis is to refine the classic way to make a zoning, including a new appoach for the residuals smoothing : Kriging models. It suggests to move from a classical smoothing to a "predictive" smoothing, where it will be possible to choose both an appropriate smoothing structure for the data and additional information about the geopositioning of the portfolio during the smoothing.

The methodology of this paper is to compare the smoothing by Kriging with the smoothing using a credibility approach and the residuals prediction by a XGBoost algorithm (and external data), two classical ways to make a zoning. This comparison is based on a home insurance portfolio and highlights the Kriging assets, the links with the other smoothing approaches and its limits.

This thesis puts forward an optimization methodology of parameters to get a better residuals smoothing and a way to reduce the computing time of the model.

Keywords: Zoning, Kriging, House insurance, Water damage, Residuals, GLM, XGBoost, Smoothing by credibility, Predictive smoothing, External data

Remerciements

Je tiens, en premier lieu, à remercier mon tuteur Pierre CHATELAIN, pour sa patience, sa pédagogie et pour tous ses conseils sur le mémoire.

Je tiens également à remercier Nabil RACHDI pour avoir suivi et aidé ce mémoire, notamment sur le plan technique du Krigeage.

Mes remerciements vont également à toute la Practice "Pricing & Data" du pôle "P&C", pour ces déjeuners, ces bonnes idées et ces conseils partagés.

Je remercie mon tuteur académique, Franck VERMET, pour son intérêt pour le sujet de ce mémoire.

Je remercie aussi l'EURIA de m'avoir instruit et m'avoir donné goût à l'actuariat pendant 3 ans.

Enfin, je remercie Camille DELLOYE pour son soutien sans faille et ses conseils avisés lors des relectures.

Note de Synthèse

Dans un milieu aussi concurrentiel que celui de l'assurance, proposer le meilleur tarif est primordial. Cela ne signifie pas de proposer le tarif le plus bas, mais plutôt de réussir à modéliser au mieux le risque sous-jacent aux assurés. L'assurance Multi-Risque Habitation (ou MRH) n'y échappe pas. Ce mémoire propose donc de revisiter une variable apportant beaucoup d'informations sur le risque des assurés : le zonier.

Un zonier est une répartition de toutes les communes françaises en un certain nombre de groupes représentant un risque géographique similaire. Cette variable regroupe l'ensemble des informations sur la météorologie, la démographie ou encore la géographie des communes. Ceci permet une meilleure segmentation du portefeuille, pour une meilleure tarification du produit MRH.

Ce produit MRH est représenté sur une base d'ADDACTIS France, où seule la garantie Dégâts des Eaux est étudiée. Cette garantie possède une forte fréquence de sinistres sur un large échantillon d'assurés. Une base de données externes sera également nécessaire et provient d'anciens travaux d'ADDACTIS France sur le sujet. Elle contient notamment l'altitude maximale, la population, la pluviométrie ou encore le pourcentage de maisons de 36000 communes environ. Elle est formée de données prises sur des sites gouvernementaux comme l'INSEE ou sur des sites en Open Source comme OpenStreetMap.

À partir de ces bases, les zoniers peuvent être réalisés en suivant deux méthodologies classiques décrites par la figure **1**.

Le point de départ est la base de données internes. Elle représente les caractéristiques des assurés, de leur bien et des potentiels sinistres passés. Avec un Modèle Linéaire Généralisé (ou "GLM"), la fréquence des sinistres est prédite à partir des caractéristiques de l'assuré et de son bien. Comme le zonier doit capter toute l'information géographique, les variables géographiques (comme la commune où se situe le bien ou le niveau de ruralité de la commune du bien) sont retirées des données d'entrée du GLM.

Le zonier se base sur les erreurs de prédiction du GLM, ou autrement appelées résidus. Ces résidus sont souvent des résidus additifs, c'est-à-dire la simple différence entre la fréquence observée et la fréquence prédite par le modèle. Ici, les résidus d'Anscombe sont préférés. Ces résidus sont calculés différemment pour approcher au maximum une loi normale centrée réduite. Une cartographie des résidus est représentée figure **3.15**.

Ces résidus ont besoin d'être lissés. En effet, comme il s'agit d'erreur sur un historique

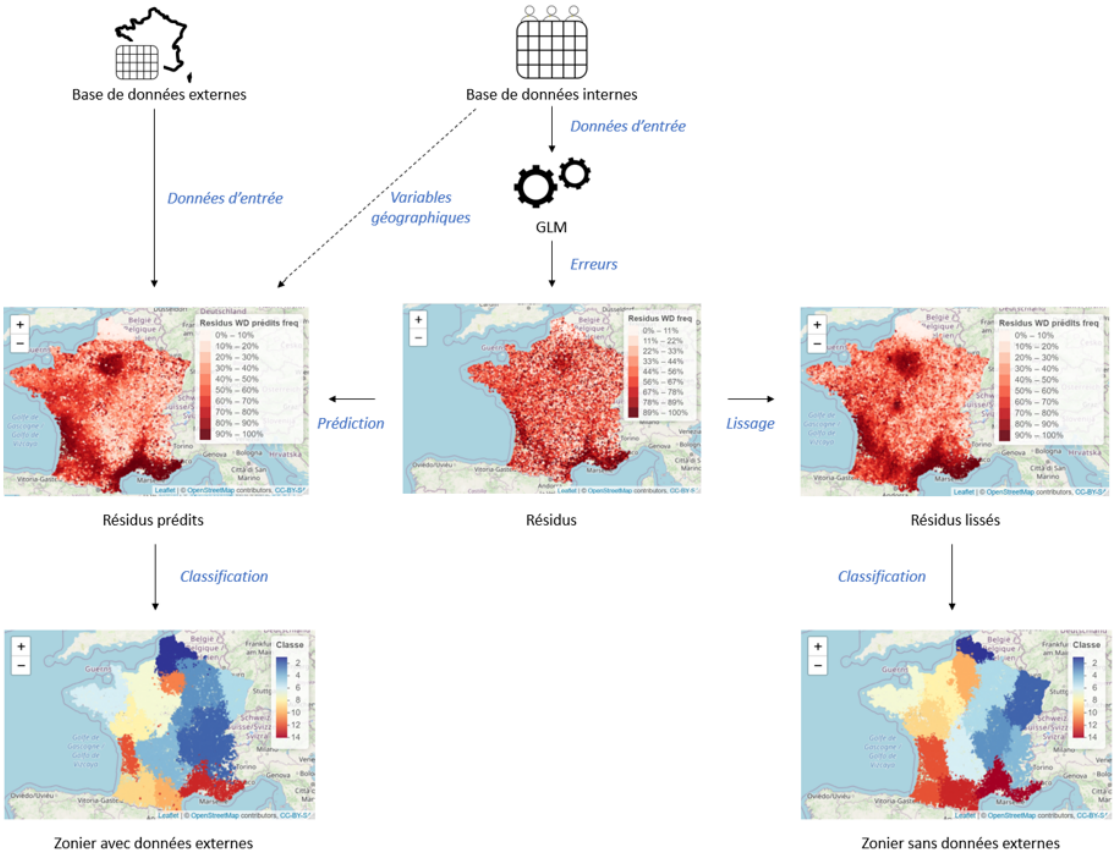


FIGURE 1 – Schéma de deux méthodes classiques de création d'un zonier pour la fréquence Dégâts des Eaux

de sinistres, les résidus ont une grande part d'aléatoire. Le but est alors de réduire cet aléatoire en rapprochant les résidus des communes proches. Pour ce faire, il existe deux méthodes classiques, réalisées pour ce mémoire : le lissage par crédibilité ou la prédiction des résidus par Machine Learning.

Le lissage par crédibilité part d'un principe : une commune peu exposée aura un résidu plus éloigné que celui d'une commune fortement exposée. Pour combler cela, la notion de crédibilité intervient. Si une commune est fortement exposée, sa crédibilité sera plus importante et son résidu lissé se rapprochera de son résidu initial. Si une commune est peu exposée, son résidu lissé se rapprochera de la moyenne des résidus des communes voisines. Cette moyenne est pondérée par l'exposition des communes avoisinantes et par la distance entre celles-ci et la commune pour laquelle le résidu lissé est calculé. Une cartographie des résidus lissés par crédibilité est représentée figure **3**.

Cette méthode comporte des limites. Premièrement, son efficacité est relative. Ensuite, la géographie des lieux n'est pas prise en compte. A cause de ces limites, la méthode

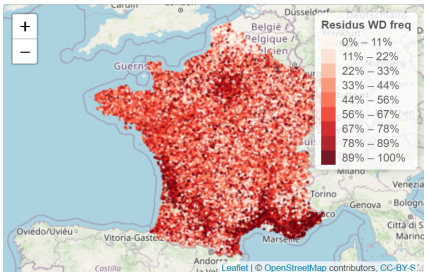


FIGURE 2 – Carte des résidus d'Anscombe en fréquence sur la garantie Dégâts des Eaux

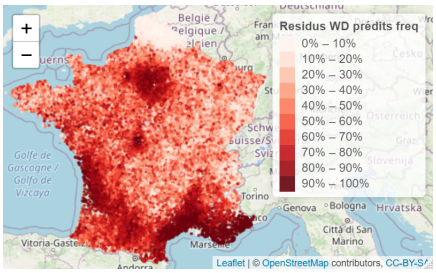


FIGURE 3 – Carte des résidus lissés par crédibilité en fréquence sur la garantie Dégâts des Eaux

prédictive est préférée par beaucoup d'assureurs aujourd'hui.

Cette méthode prédictive ressemble à ce qui a été fait avec le GLM : prédire une variable aléatoire avec des données déterministes. Pour prédire les résidus, l'algorithme d'XGBoost est plébiscité. C'est un algorithme rapide et très performant, souvent utilisé pour les concours de Data Science. La prédiction se fait à partir des données externes récupérées par ADDACTIS France sur la météorologie, la démographie des lieux, etc. Les résultats sont meilleurs que pour le lissage par crédibilité et les données externes peuvent expliquer ce risque géographique porté par les résidus. Une cartographie des résidus prédits par XGBoost est représentée figure **4**.

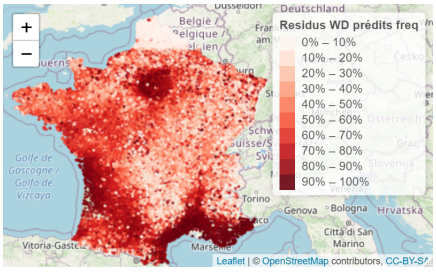


FIGURE 4 – Carte des résidus prédits par XGBoost en fréquence sur la garantie Dégâts des Eaux

Avec l'importance des données externes, on peut se demander : Le lissage par crédibilité peut-il être amélioré en ajoutant une prédiction de données externes ? La réponse se trouve dans le modèle du Krigeage.

Le Krigeage propose de transformer la distance spatiale utilisée dans le lissage par crédibilité par une distance entre les variables des données externes. Le modèle ne se base plus seulement sur la longitude et la latitude pour calculer la distance entre deux communes. Par exemple, les différences entre les altitudes, entre les pluviométries ou encore entre les nombres d'habitants constituent une nouvelle distance exploitable. Le but est de trouver la dépendance spatiale entre les communes pour ensuite prédire au mieux les résidus des prochaines communes.

Un résidu prédit par Krigeage est la somme d'une fonction déterministe, représentant sa tendance et d'un processus gaussien centré avec une structure de dépendance décrite plus tôt. Cette structure de dépendance prend la forme d'un des cinq noyaux décrits dans le mémoire. Le noyau choisi est trouvé empiriquement en les testant tous sur la base de données et en sélectionnant finalement celui qui convient le mieux.

Pour prédire une nouvelle valeur, les propriétés sur les processus gaussiens sont utilisées. L'espérance est alors estimée et sert de valeur prédite pour le résidu. La variance est, elle aussi, estimée et sert d'information sur la volatilité du résultat. Les paramètres restants sont estimés par maximum de vraisemblance.

Cette méthode est appliquée à la base MRH d'ADDACTIS France, avec d'excellents résultats sur la base d'apprentissage de 4000 points. Le modèle du Krigeage prédit mieux les données que le modèle de l'XGBoost. Une cartographie des résidus prédits par Krigeage sur une base d'entraînement de 4000 points est réalisée sur la figure **5**. Cependant, lorsque le modèle apprend l'entièreté de la base de données, le Krigeage devient plus difficile à utiliser. Les temps de calcul importants rendent difficiles l'entraînement sur la base entière.

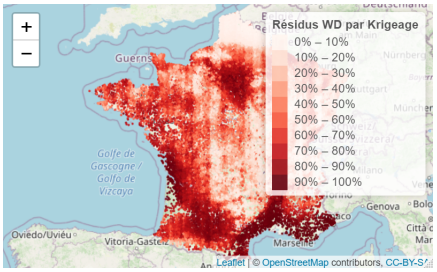


FIGURE 5 – Carte des résidus prédits par Krigeage en fréquence sur la garantie Dégâts des Eaux, avec une base d'entraînement de 4000 points

En effet, le temps de calcul du modèle évolue exponentiellement avec le nombre de points dans la base d'entraînement. Pour remédier à cette limite, il existe différentes solutions. Celle développée dans ce mémoire divise la base en plusieurs "sous-bases d'ap-

prentissage" et calcule un modèle de Krigeage sur chacune pour ensuite faire la moyenne des prédictions. L'efficacité sur la base de 4000 points est un peu en dessous de la méthode de base, mais le temps de calcul est divisé par quatre.

Cependant, si cette méthode de division est intéressante pour les petites bases, le problème subsiste pour les bases importantes. En effet, le résidu semble trop lissé lorsque les métriques d'erreur sont comparées. Pour expliquer cela, le modèle ne sur-apprend quasiment pas, puisque chaque commune est la moyenne de prédiction de plusieurs modèles différents. Une cartographie des résidus lissés par Krigeage avec la méthode de division est présentée figure **6**.

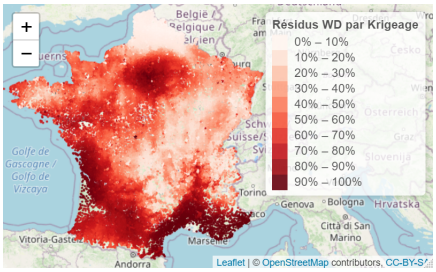


FIGURE 6 – Carte des résidus lissés par Krigeage en fréquence sur la garantie Dégâts des Eaux, avec la méthode de division

Cependant, lorsque les trois zoniers obtenus par crédibilité, XGboost et Krigeage sont ajoutés au GLM, le GLM avec le zonier par Krigeage obtient de meilleures métriques (Déviance, AIC, coefficient de Gini), comme montré dans le tableau 1. Pour ce portefeuille, le zonier par Krigeage est donc meilleur.

GLM	Déviance	AIC	Gini
GLM sans zonier	648714	793702	0,1576
GLM avec zonier crédibilité	638357	783361	0,2729
GLM avec zonier XGBoost	638371	783385	0,2595
GLM avec zonier Krigeage	637757	782768	0,2764

TABLE 1 – Tableau des métriques des GLM (avec le zonier par Krigeage)

Ici, l'étude se concentre sur un portefeuille, pour une garantie. Appliquer le Krigeage pour un autre portefeuille, sur une autre année ou sur une autre garantie serait un bon moyen de vérifier en pratique la robustesse du modèle.

Summary

In a environment as competitive as the insurance one, offering the best rate is critical. This does not mean offering the lowest rate, but rather successfully modeling the underlying risk to policyholders. Home insurance is no exception. Therefore, this paper proposes to revisit a variable that provides a lot of information on the risk of insured people : the zoning.

The zoning variable is the distribution of all French towns into a number of groups representing a similar geographical risk. This variable groups together all the information on weather, demographics and geography of the municipalities. This allows a better segmentation of the portfolio, for a better pricing of the Home insurance product.

This House insurance product is represented for an ADDACTIS France base, where only the water damage coverage is studied. This coverage has a high frequency of claims on a large panel of insured persons. An external database will also be necessary and comes from former works of ADDACTIS France on the subject. It contains, for instance, the maximum altitude, the population, the rainfall or the percentage of houses over 36000 municipalities. It is made of data taken from governmental sites like INSEE or from Open Source sites like OpenStreetMap.

With these bases, zonings can be realized by using two classical methodologies described in figure **7**.

The starting point is the internal database. It represents the characteristics of the insured, their property and potential past claims. With a Generalized Linear Model (or "GLM"), the frequency of claims is predicted from the characteristics of the insured and his property. Since the zonings must capture all geographic information, geographic variables (such as the municipality in which the property is located or the rurality level of the property's municipality) are removed from the GLM input data.

The zonings are based on the GLM's prediction errors, or otherwise known as residuals. These residuals are often additive residuals, wich are the simple difference between the observed frequency and the frequency predicted by the model. Here, Anscombe residuals are preferred. These residuals are computed differently to approximate as closely as possible a normal distribution with reduced center. A mapping of the residuals is shown in figure **8**.

This method has limitations. The first is its relative efficiency. The second is the fact

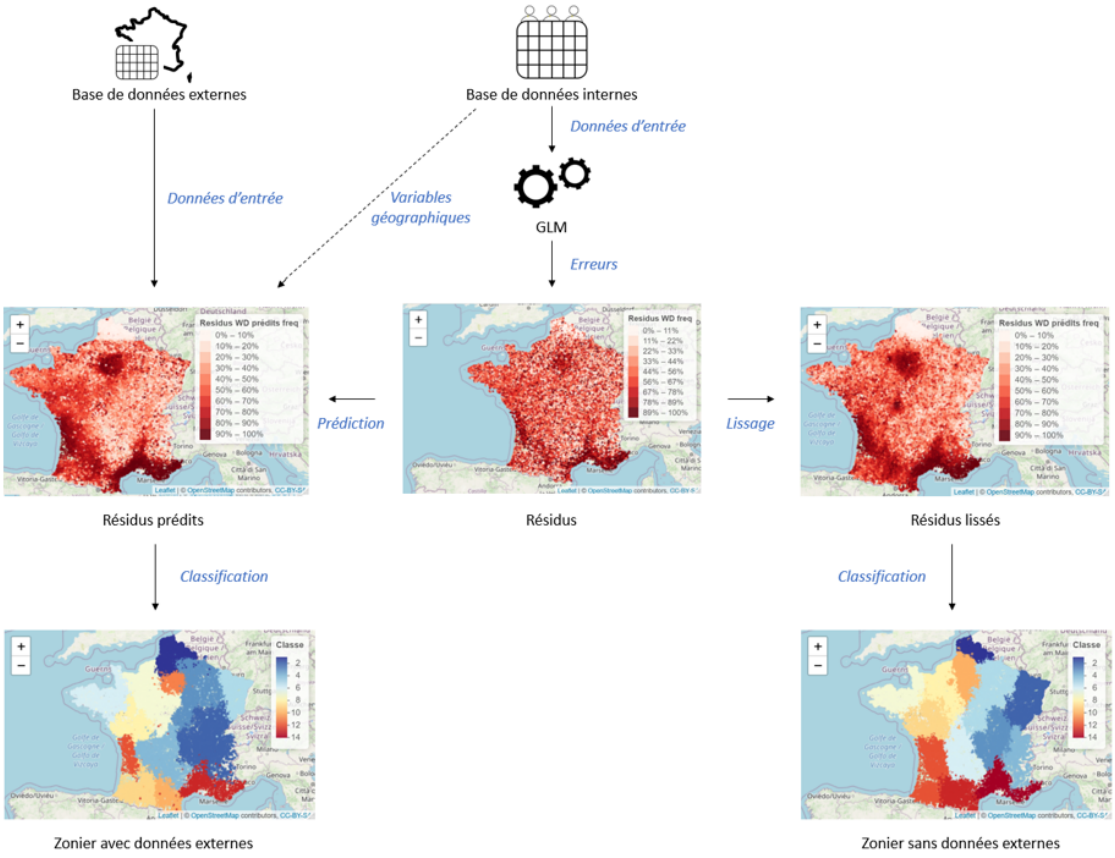


FIGURE 7 – Diagram of two classical methods to create a zonings for the Water Damage frequency

that the geography of the location is not taken into account. Because of these limitations, the predictive method is preferred by many insurers nowadays.

This predictive method is similar to what was done with GLM : predicting a random variable with deterministic data. To predict the residuals, the XGBoost algorithm is popular. It is a fast and very powerful algorithm, often used for Data Science competitions. The prediction is made from external data retrieved by ADDACTIS France on meteorology, demography of the places, etc. The results are better than for credibility smoothing and the external data can explain the geographical risk borne by residuals. A mapping of the residuals predicted by XGBoost is shown in figure **9**.

With the importance of external data, one may ask : Can credibility smoothing be improved by adding external data prediction ? The answer lies in the Kriging model.

Kriging proposes to transform the spatial distance used in credibility smoothing by a distance between the variables of the external data. The model no longer relies solely on longitude and latitude to calculate the distance between two municipalities. For instance,

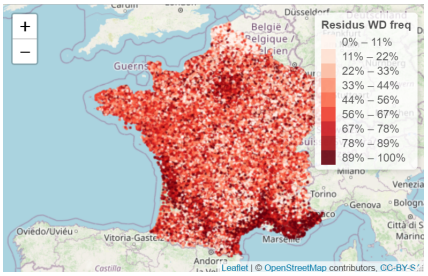


FIGURE 8 – Anscombe's residual frequency map on water damage coverage

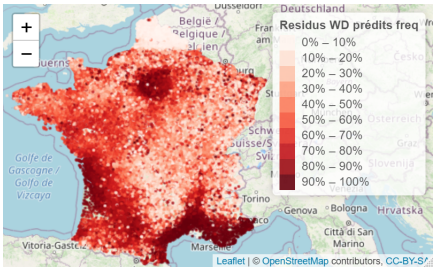


FIGURE 9 – Map of residuals predicted by XGBoost in frequency on water damage coverage

differences between the altitudes, the rainfall or the number of inhabitants constitute a new exploitable distance. The goal is to find the spatial dependency between the towns in order to predict the residuals of the next towns.

A residual predicted by Kriging is the sum of a deterministic function, representing its trend and a centered Gaussian process, and a dependence structure described earlier. This dependence structure takes the form of one of the five kernels described in the thesis. The chosen kernel is found empirically by testing all of them on the database, to find which one fits best.

To predict a new value, properties about Gaussian processes are used. The expectation is then estimated and used as a predicted value for the residual. The variance is also estimated and serves as information on the volatility of the result. It remains to estimate the remaining parameters by maximum likelihood.

This method is applied to the house insurance database of ADDACTIS France, with excellent results on the training base of 4000 points. The Kriging model predicts the data better than the XGBoost model. A mapping of the residuals predicted by Kriging on a training base of 4000 points is done on figure **10**. However, when the model learns the entire database, Kriging becomes more difficult to use. The long computation times make it difficult to compute on the whole database.

Indeed, the computation time of the model evolves exponentially with the number of points in the training base. To remedy this limitation, there are different solutions. The

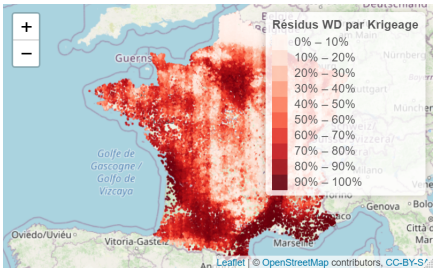


FIGURE 10 – Map of residuals predicted by frequency Kriging on Water Damage coverage, with a training base of 4000 points

one developed in this thesis divides the database into several "training sub-bases" and computes a Kriging model on each of them to then average the predictions. The efficiency on the 4000 points base is a bit lower than the basic method, but the computation time is divided by four.

However, if this division method is interesting for small bases, the problem remains for large bases. Indeed, the residual seems too smooth when the error metrics are compared. To explain this, the model hardly over-learns, since each commune is the average prediction of several different models. A mapping of the residuals smoothed by Kriging with the split method is shown in Figure **11**.

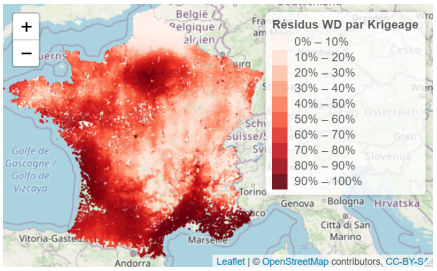


FIGURE 11 – Map of residuals smoothed by frequency Kriging on the Water Damage coverage, with the division method

However, when the three zonings obtained by credibility, XGboost and Kriging are added to the GLM, the GLM with the zoning by Kriging obtains better metrics (Deviance, AIC, Gini coefficient), as shown in the table 2 For this portfolio, the Kriging zoning is therefore better.

Here, the study focuses on a portfolio, for a guarantee. Applying Kriging to another portfolio, another year or another guarantee would be a good way to check the robustness of the model in practice.

GLM	Deviance	AIC	Gini
GLM without zoning	648714	793702	0.1576
GLM with credibility zoning	638357	783361	0.2729
GLM with XGBoost zoning	638371	783385	0.2595
GLM with Krigeage zoning	637757	782768	0.2764

Table 2 – Table of GLM metrics (with Kriging zonier)

Table des matières

Chapitre 1

Introduction

L'assuré paie sa première prime d'assurance avant même que lui ou l'assurance ne connaisse la valeur de ce qui va lui être remboursé. C'est la difficulté du travail de l'actuaire lorsqu'il propose un tarif aux assurés. Pour ce faire, il estime la sinistralité en fonction des caractéristiques des assurés et de leur bien assuré pour une assurance Multi-Risque Habitation. Puis, il crée des groupes de risque équivalent pour segmenter son tarif et ainsi proposer les meilleurs primes aux futurs assurés.

Avec la concurrence accrue, due à la réglementation française, aux moyens de souscription, aux nouveaux acteurs du marché, aux marges négatives imposées par le marché, la segmentation se veut plus fine, sans pousser à l'extrême.

Le zonier, un regroupement de communes de risque similaire, permet cette segmentation plus fine. En effet, les variables géographiques, regroupées dans le zonier, permettent d'avoir une meilleure connaissance du risque. De plus, le regroupement en une seule variable des données géographiques permet d'éviter le sur-apprentissage.

Le but de ce mémoire est de présenter deux méthodes classiques de construction de zonier, puis de les comparer avec une autre méthode, le lissage des résidus par Krigeage. Ce papier tend à montrer les atouts et les limites du Krigeage dans ce domaine, ainsi que des manières d'optimiser son paramétrage et ses temps de calcul. Cette comparaison entre les méthodes se fait avec une base de données d'ADDACTIS France sur la garantie Dégâts des Eaux de l'assurance Multi-Risque Habitation.

La démarche de ce mémoire est la suivante :

— Le premier chapitre est une introduction au sujet.

— Le second chapitre revient sur les bases de l'assurance, la concurrence sur le marché et la nécessité d'avoir un bon zonier. Ensuite, la méthodologie du zonier y est présentée et les différentes bases de données utilisées y sont décrites.

— Le troisième chapitre explicite la création du zonier en revenant sur chaque étape. L'aspect théorique y est détaillé et les premiers résultats des deux méthodes classiques y sont présentés.

— Le quatrième chapitre analyse le modèle du Krigeage. Il y est formalisé mathématiquement et son optimisation est discuté. Ses atouts et ses limites sont exposés et des solutions sont apportées à ses limites.

— Le cinquième chapitre compare les modèles et leurs apports. La question de la métrique d'erreur pour les zoniers est abordée. Enfin, des ouvertures sur ce qui peut être encore fait sur le sujet sont évoquées.

Chapitre 2

Le zonier dans l'assurance Multi-Risque Habitation

Les assurances ont un mode de fonctionnement particulier : il est nécessaire de mettre un prix sur un service dont on ne connaît pas la valeur à l'avance. La mission des actuaires est donc de trouver un moyen d'estimer au mieux la prime des futurs assurés, afin d'engranger le plus de bénéfice et de résister à la concurrence toujours plus forte. L'assurance Multi-Risque Habitation n'y échappe pas, comme décrit dans la section **2.1**.

Pour ce faire, les données géographiques des communes des assurés sont regroupées en une variable : le zonier. Ce zonier apporte beaucoup d'information sur le risque sous-jacent aux assurés, tout en étant compact pour faciliter les calculs sur la base de données. Sa méthodologie est développée dans la section **2.2**.

La base d'assurés utilisée ici est une base propre à ADDACTIS France sur la sinistralité sur les maisons en Multi-Risque Habitation. Cette base importante, représentative du marché, est présentée dans la section **2.3**.

Enfin, ce chapitre se clôture sur la présentation de la base de données externes, dans la section **2.4**. Cette base est issue de données publiques et possède des données sur la météorologie, sur la démographie ou encore sur l'altitude de près de 36000 communes en France.

2.1 L'assurance Multi-Risque Habitation

L'assurance Multi-Risque Habitation (abrégé "MRH") est un type d'assurance concernant les biens immobiliers des assurés (**2.1.1**). Comme pour tout type d'assurance, l'assuré paie une cotisation, en amont de la prestation de l'assureur. En anticipant cela, l'assureur fixe ce tarif en fonction des caractéristiques de l'assuré (**2.1.2**). Cependant, l'assureur doit également prendre en compte les autres assurés et le marché concurrentiel qui l'entoure. Une forte compétition amène les assureurs à faire des choix dans leur politique et à optimiser au maximum leur tarification (**2.1.3**).

2.1.1 L'assurance Multi-Risque Habitation dans le milieu assurantiel

Dans une entreprise classique, pour donner un prix à un produit, le coût de celui-ci et la demande des acheteurs rentrent en compte. D'autres paramètres comme la réglementation ou la concurrence influent également sur le processus. Ce sont toutes des informations connues avant la commercialisation du produit. Cependant, en assurance, la manière de faire est différente. Le produit d'assurance est la promesse de verser une indemnisation si un événement aléatoire survient, afin d'en protéger celui qui est assuré. Le coût final de ce produit n'est donc pas connu lors de sa commercialisation. Il s'agit du **cycle inversé de production**.

Cette particularité d'effectuer un paiement pour se protéger d'un sinistre futur est retrouvée dans la civilisation babylonienne avec le code d'Hammourabi, servant de code juridique. Lorsqu'un marin faisait un prêt pour financer son voyage et qu'il avait versé une certaine somme en amont au préteur, le remboursement de ce prêt était annulé si le navire est pillé ou coulé. L'assurance des navires marchands se popularisera au V^{e} siècle avant J.C chez les Grecs et en Italie vers le XII^{e} siècle.

D'autres types d'assurance verront le jour plus tard et seront classés en deux principaux types : l'assurance **de personnes** et l'assurance **dommages**. L'assurance de personnes regroupe les assurances santé, les assurances vie, les prévoyances, les assurances dépendances, c'est-à-dire lorsque ce qui est assuré est une personne. Contrairement à l'assurance dommages ou IARD (Incendie, Accident, Risques Divers) qui assure un bien comme une automobile ou une habitation.

L'**assurance Multi-Risque Habitation** fait en grande partie de l'assurance dommage. Elle garantit une somme à l'assuré, lorsque son habitation est touchée par un sinistre. Cependant, une assurance responsabilité civile doit être signée en plus de l'assurance d'un bien immobilier, ce qui fait que l'assurance MRH comporte aussi de l'assurance de personnes.

En 2020, les produits d'assurance MRH ont généré 11.6 Mds d'€ de cotisations, ce qui représente environ 19,5% des cotisations des assurances dommages [1]. Ces produits trouvent leur origine au XVII^{e} siècle suite à l'incendie de près de 13000 bâtiments à Londres. La garantie incendie représente encore aujourd'hui une proportion importante de la sinistralité de l'assurance MRH, aux côtés d'autres garanties comme :

— Dégâts des eaux
— Vol
— Bris de glace
— Responsabilité civile
— Catastrophes naturelles
— Tempête/Grêle/Neige (TGN)
— Dommages électriques

1. [Lustman, 2021] Conférence de presse de la FFA du 24 mars 2021

De façon générale, chaque garantie implique une tarification du produit, car la nature des sinistres, les critères de risques et les taxes sont différents. Par exemple, pour la garantie Incendie, les dégâts sont souvent élevés, comparé à la garantie Bris de glace. Le travail de l'actuaire est donc de trouver le tarif adéquat à chaque garantie, en fonction des critères de risque et des taxes.

2.1.2 La tarification dans l'assurance

La **prime pure** d'un produit d'assurance reflète ce que l'assuré devra, en moyenne, coûter à l'assureur. L'assureur doit alors estimer cette prime, notamment en s'appuyant sur la base des assurés et sur l'historique des sinistres associés. La base des assurés est en partie composée des informations du questionnaire fourni pendant la souscription. Avec ces informations, on peut discerner les profils d'assurés plus risqués que d'autres. Par exemple, une cheminée dans un bien augmente le risque d'avoir un incendie. L'assuré qui n'a pas de cheminée paiera alors moins cher sa garantie Incendie, toute chose égale par ailleurs, que celui qui en a une. En attirant plus de profils peu risqués, l'assureur évite l'**antisélection** qui pourrait mettre à mal ses profits (voir figure **2.1**).

Il est donc nécessaire de détecter les caractéristiques des assurés qui les rendent plus ou moins risqués en créant une **segmentation**. Pour ce faire, l'actuaire dispose de modèles prédictifs, comme le **Modèle Linéaire Généralisé**. Ce modèle, très souvent utilisé dans la tarification de produit d'assurance en France, rend compte de l'influence de chaque caractéristique sur la sinistralité. La section **3.1** détaille le fonctionnement de ce modèle. La tarification permet alors d'attirer les profils souhaités avec des primes qui leur correspondent. La figure **2.1** illustre cette notion de segmentation.

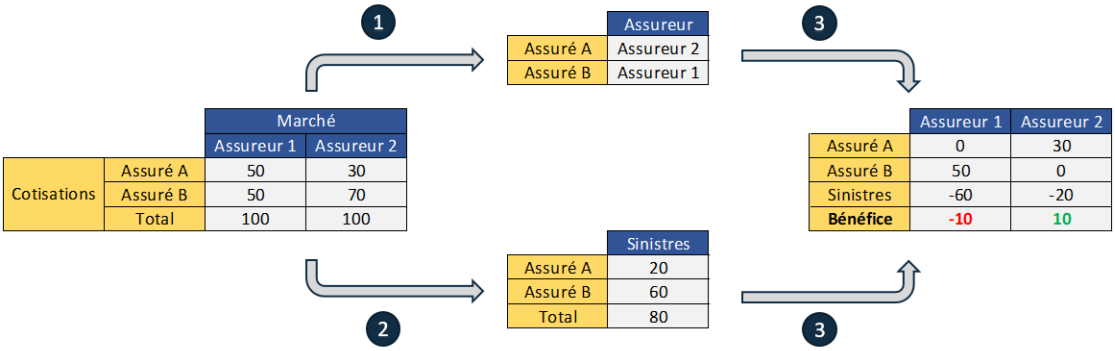


FIGURE 2.1 – Schéma de l'impact de la segmentation

Sur la figure **2.1**, le cas simple d'un marché assurantiel avec deux acteurs est supposé. L'assureur 1 propose le même tarif pour les deux assurés A et B. L'assureur 2 propose une segmentation entre A et B en proposant un tarif moins cher pour A et plus cher pour B. Lorsque A et B sont assurés chez le même assureur, celui-ci reçoit 100 de cotisations.

Cependant, les assurés prennent généralement l'assurance la moins chère pour eux. En

effet, comme montré en (1), l'assuré A préfère s'assurer chez l'assureur 2 qui lui propose la cotisation la plus faible (30 contre 50 pour l'assureur 1) et l'assuré B préfère s'assurer chez l'assureur 1 (50 contre 70 pour l'assureur 2).

Au niveau des sinistres en (2), l'assuré A a seulement 20 de sinistres comparés aux 60 de sinistres de l'assuré B. La segmentation de l'assureur 2 était donc supposée juste : l'assuré A avait un profil moins risqué que celui de l'assuré B et cela s'est avéré sur les sinistres de l'année. Lorsque les assurés sont assurés chez le même assureur, celui-ci doit prendre en charge 80 de sinistres.

Seulement, les assurés ont choisi un assureur différent, celui qui leur propose la prime la plus faible. L'assureur 1 encaisse 50 de cotisations et l'assureur 2 encaisse 30 de cotisations en tout. Sans compter la sinistralité, l'assureur 1 possède un meilleur chiffre d'affaires. Néanmoins, en ayant mieux segmenté, l'assureur 2 possède un meilleur bilan avec 10 de bénéfices, tandis que l'assureur 1 possède un bilan de -10. Il est à noter que sans notion de marché, chaque assureur aurait fait un bénéfice de 20.

Cet exercice illustre simplement ce qu'il se passe lorsqu'un assureur approche mieux le risque que d'autres. Dans le cas réel, les assureurs segmentent sur plusieurs variables, complexifiant l'analyse.

La tarification dépend donc de l'assuré, mais aussi des garanties assurées. Les sinistres de chaque garantie peuvent être séparés en sinistres attritionnels, graves ou climatiques. Les sinistres attritionnels correspondent à un coût faible et une fréquence élevée. Les sinistres graves correspondent à un coût élevé et une fréquence faible. Les sinistres climatiques relèvent de régimes spéciaux selon la classification de l'origine du sinistre en catastrophe naturelle ou non. Chacune de ces catégories possède un risque spécifique, un traitement des sinistres particulier et donc d'une modélisation différente.

Enfin, cette prime pure ne reflète que le risque de l'assuré. Pour calculer la **prime commerciale**, la prime proposée à l'assuré, l'assureur rajoute classiquement :

— Les chargements de sécurité, pour surestimer les sinistres s'il y a une mauvaise année.

— Les chargements de frais de gestion, pour rémunérer les courtiers ou agents généraux et pour couvrir les autres frais de gestion des sinistres.

— Les différentes taxes que l'assurance doit payer.

2.1.3 Un milieu concurrentiel

En plus de ces chargements, des types de sinistres, du profil de l'assuré et de son bien, il faut prendre en compte le milieu dans lequel opère l'assureur. Un client potentiel peut passer par différents **moyens de souscription** pour une assurance :

— Un agent général de l'assurance ou un conseiller bancaire pour les bancassureurs
— Un courtier
— Un site internet

L'agent général et le courtier, pour fidéliser leur clientèle, peuvent effectuer des remises pour leurs clients (ce qui n'est normalement pas le cas sur internet). Pour ce faire, il est nécessaire qu'ils connaissent la manière dont le produit est tarifé et que le tarif soit explicable. Cette notion d'interprétabilité du tarif est très importante, car l'assuré doit pouvoir comprendre ce qu'il paie, s'il le souhaite.

En France, les assurances automobiles et MRH sont obligatoires. La demande est donc forte. De plus, avec l'arrivée des bancassureurs, les prospects ont le choix entre de nombreuses offres différentes : 11 compagnies se partageaient 90% du marché MRH en 2020 [2].

La concurrence se retrouve également amplifiée avec la **législation française**. Par exemple, la loi Hamon, votée en 2014, permet aux assurés de changer d'assurance plus facilement, au bout d'une année. Le prochain assureur se charge de la partie administrative auprès du précédant assureur. Cela amène plus de concurrence entre les assureurs qui peuvent attirer plus facilement les assurés d'autres compagnies. Avoir un premier tarif encore plus attractif est devenu la priorité chez les assureurs.

Pour ce faire, une segmentation encore meilleure est fondamentale. Les actuaires vont alors chercher une segmentation toujours plus fine, retraçant des groupes de risques toujours plus précis. Avec les outils d'apprentissage statistiques et le Big Data (c'est-à-dire les grandes bases d'assurés et les grandes bases de données disponibles sur internet, à partir des objets connectés,...), les actuaires peuvent aller jusqu'à l'**hypersegmentation**. Il s'agit d'un terme désignant une segmentation extrêmement fine, voire personnalisée (un tarif unique pour chaque assuré).

Cette hypersegmentation présentée comme le futur de l'assurance, possède en réalité des problématiques sous-jacentes. La fin de la mutualisation, et donc de la notion d'assurance, présente d'abord un problème éthique. En effet, si on adapte tous les tarifs aux assurés, les profils plus risqués devront payer beaucoup plus, voire, beaucoup plus que leurs moyens. Cela pourrait conduire à l'exclusion d'une partie de la population, ne pouvant pas se payer une assurance. Ceci va à l'encontre du partage des sinistres de l'assurance.

Un souci d'ordre statistique peut également être formulé. L'assurance travaille sur un grand groupe d'assurés pour réduire la variabilité des sinistres provoqués par ceux-ci. Plus il y a de personnes assurées, plus la moyenne de sinistres se rapproche de l'espérance des sinistres calculée. Si un nombre plus important de sinistres arrive, il y a un grand nombre de cotisations pour y faire face. Ce n'est pas le cas chez les assureurs qui pratiquent l'hypersegmentation, comme ils travaillent avec un petit groupe d'assurés.

Également, on stipule que les assurés chercheront toujours le prix le plus faible pour leur assurance. Or, cela n'est pas toujours vrai. Par exemple, dans les communes moins peuplées, il n'y a pas d'agence pour tous les assureurs du marché. Un futur assuré peut donc aller voir la seule agence près de chez lui, même si elle n'est pas la moins chère. Aussi, d'autres critères comme la qualité du service client peuvent influencer le choix de

2. L'Argus de l'Assurance, Classement Auto-MRH 2020 : coup d'arrêt sur les marchés !

l'assuré. Encore, un assuré peut prendre une assurance automobile chez un assureur pour son prix attractif et rester chez cet assureur pour prendre une assurance habitation, bien que son prix soit plus élevé que celui de la concurrence. L'assurance automobile fait alors office de **produit d'appel**.

Utiliser un produit comme produit d'appel est une chose courante dans le monde assurantiel. Les assureurs ayant une marge négative sur ce produit cherchent à avoir le tarif le plus compétitif. Si ce n'est pas comme un produit d'appel pour un autre produit, il peut s'agir d'un produit avec une marge négative la première année, qui repasse positive les années suivantes. Si la définition classique de la prime commerciale, vu précédemment, indique qu'il est impossible d'avoir une prime pure supérieure à une prime commerciale, la forte concurrence du marché impose d'autres règles. Pour les assureurs, la prime commerciale est composée de la prime pure et d'une marge, négociée selon les tarifs proposés par la concurrence et en fonction des clients. Une marge peut donc être négative si le marché est très concurrentiel sur un segment.

L'impact de la concurrence est très important sur la manière dont est tarifé un produit d'assurance MRH. Pour rester compétitif, les assureurs sont donc obligés d'avoir une excellente connaissance des risques sous-jacents de leur portefeuille d'assurés. Cela se traduit souvent par une segmentation plus fine, mais pas à l'extrême avec une hyper-segmentation.

2.2 Définition et création du zonier

Pour réaliser cette segmentation plus fine, la variable géographique apporte une connaissance supplémentaire du risque de l'assuré. Facilement récupérable avec l'adresse de l'assuré, son traitement pour le lier au risque assurantiel est plus compliqué. Dans ce cadre, un zonier est préconisé (**2.2.1**). La méthodologie pour obtenir les résidus qui servent pour le zonier est toujours la même, mais leur utilisation est différente selon le zonier à réaliser (**2.2.2**).

2.2.1 Contexte et définition

L'information géographique s'obtient sans difficulté avec l'adresse de l'assuré dans le questionnaire de l'assurance. Avec cette donnée, l'assureur peut savoir, par exemple, si le bien se trouve dans une ville peuplée, dans une région inondable ou encore dans une région à fortes pluies. Avec l'étude adéquate sur une base de données géographiques, l'assureur dispose alors d'une segmentation fine par commune, améliorant l'approximation du risque.

Néanmoins, cette base géographique nécessite un travail supplémentaire, décrit dans la section **2.4**. De même, une étude de risque sur une commune peu exposée biaiserait le modèle, alors que sur une commune fortement exposée, l'appréciation du risque serait plus juste. Enfin, deux communes proches auront un risque différent selon l'historique des sinistres propre à chaque commune, bien qu'elles partagent quasiment la même position

géographique.

Pour contrecarrer ces problèmes, l'assureur peut créer un **zonier** par garantie. Il s'agit d'une variable qui divise le pays en plusieurs zones de risque proche. Ainsi, on regroupe en une variable tout le risque géographique sur la sinistralité. Cela allège considérablement les modèles, en ne considérant pas toutes les données externes. Aussi, dans la création des zones, l'exposition des communes peut-être prise en compte et donc donner plus d'importance aux communes fortement exposées.

2.2.2 Méthodologie du zonier

La méthodologie classique de création d'un zonier peut se résumer avec le schéma de la figure **2.2**.

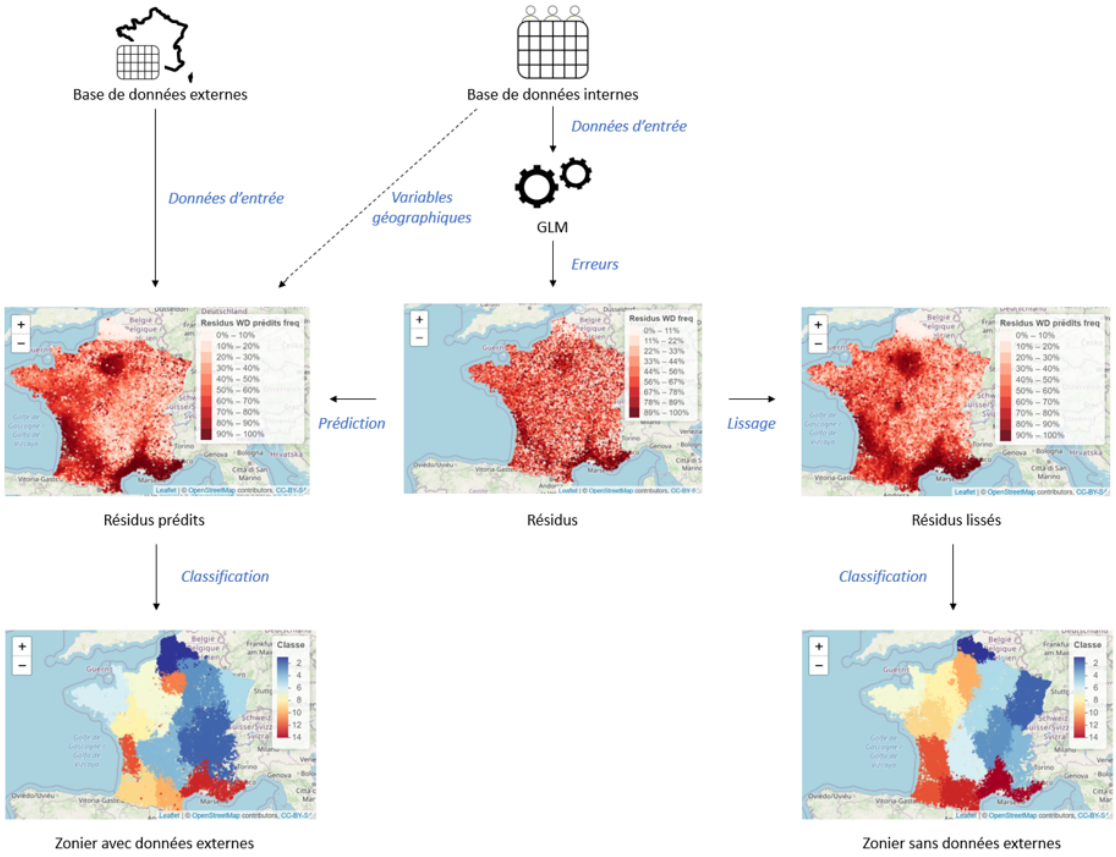


FIGURE 2.2 – Schéma de deux méthodes de création d'un zonier pour la fréquence Dégâts des Eaux

L'assureur part de la base des assurés avec les sinistres associés. Cette base contient les caractéristiques des assurés, de leur bien et les sinistres passés. Il y retire les variables géographiques, comme l'adresse, le code postal, la zone urbaine...

Tout d'abord, on entraîne un **Modèle Linéaire Généralisé** (ou abrégé GLM), détaillé dans la section **3.1**, à prédire le nombre et le coût moyen des sinistres des assurés, à partir des caractéristiques de ceux-ci.

Les résidus sont ensuite récupérés, c'est-à-dire la différence entre la fréquence (ou le coût moyen) des sinistres observés et la fréquence (ou le coût moyen) prédite par le GLM. Les résidus additifs sont obtenus par une simple différence, mais en appliquant différentes fonctions sur la fréquence observée et prédite, d'autres résidus peuvent être obtenus, comme les **résidus d'Anscombe** (voir section **3.2**). L'assureur peut alors lisser les résidus ou les prédire avec des données externes, les deux méthodes classiques pour réaliser un zonier.

Le lissage des résidus (à droite sur le schéma **2.2**) se fait en moyennant les résidus des communes avec les communes proches. Ce lissage est appelé **lissage par crédibilité**, car un poids de "crédibilité" est donné au résidu initial de la commune en fonction de son exposition. Plus la commune est exposée, plus son résidu sera pris en compte dans le lissage. Également, la distance des communes par rapport aux autres est prise en compte dans la pondération de la moyenne, pour privilégier les communes proches (voir section **3.3**).

Une autre méthode pour créer un zonier consiste à prédire les résidus avec une base de données externes et les variables géographiques de la base d'assurés. Ces données externes proviennent de sites gouvernementaux ou d'Open Sources (données pouvant être utilisées comme l'utilisateur le souhaite) et portent sur la météorologie, la démographie ou encore la géographie des lieux. Pour prédire les résidus, un algorithme de **Machine Learning** est utilisé, ici il s'agit d'un **XGBoost** (voir section **3.4**).

Maintenant que les résidus des communes sont lissés, c'est-à-dire que la variabilité des résidus est réduite, il est plus simple de regrouper les communes. Pour ce faire, un algorithme de **classification ascendante hiérarchique** est utilisé (voir section **3.5**). Le principe est de comparer les distances entre toutes les communes, au niveau de leur résidu et de leur position, pour déterminer des premiers petits groupes, puis de regrouper encore et encore ces groupes pour arriver au nombre de zones voulues (ici au nombre de 10). Le zonier est alors utilisable en tant que variable pour le GLM d'une tarification d'un produit MRH par exemple.

2.3 Présentation de la base ADDACTIS MRH

Afin de réaliser un zonier, il est donc nécessaire de disposer d'une base de données internes de caractéristiques d'assurés ainsi que de l'historique de leurs sinistres, comme point de départ. ADDACTIS France a développé en interne une base de données représentative du marché MRH français dont la partie Dégâts des Eaux sera utilisée pour ce mémoire. Quelques variables seront présentées dans la section **2.3.1**. L'exposition, variable très importante dans la tarification, sera étudiée plus en profondeur dans la section **2.3.2**. Enfin, la fréquence des sinistres en Dégâts des Eaux, étant la variable utilisée pour

le zonier de ce mémoire, sera également étudiée dans une autre section : la section **2.3.3**.

2.3.1 Présentation générale de la base

Ici, la base de données internes utilisée est une base MRH d'ADDACTIS France, sur la garantie **Dégâts des Eaux** pour seulement des maisons. Elle dispose d'environ 1400000 contrats et de 30 variables. Parmi ces variables, les caractéristiques d'assurés se retrouvent, ainsi que les caractéristiques de leur sinistralité (Nombre de sinistres, Montant total des sinistres, Type de sinistres,...). Les caractéristiques des assurés servent à expliquer au maximum la fréquence et le coût des sinistres, avec un GLM. Il faut donc avoir des variables intéressantes et récupérables simplement pour allier précision de la prédiction et efficacité du questionnaire de souscription.

Quelques exemples de variables sont présentés ci-dessous :

— **L'âge de l'assuré** permet de regrouper les assurés selon des tranches d'âge qui correspondent à des modes de vie similaires et donc à des sinistralités similaires. Également, c'est une variable dont l'interaction avec une autre peut permettre d'avoir une meilleure modélisation du risque. Pour ce portefeuille, l'âge minimum est de 18 ans, l'âge moyen est de 42 ans et l'âge maximum est de 100 ans.

— **Le nombre de pièces du bien** est une variable donnant une bonne information sur le bien de l'assuré. La surface habitable est également utilisée dans le GLM, ce qui enrichit l'information sur la superficie et la disposition du bien.

— **La qualité de l'assuré** représente le fait que l'assuré soit un propriétaire occupant, un propriétaire non occupant ou un locataire du bien assuré. Ces trois types de qualité impliquent des fonctionnements différents de l'utilisation du bien et donc une sinistralité différente.

— **Le nombre de piscines du bien** est une variable qui concerne un petit groupe de biens dans le portefeuille (5% des biens ont une piscine ou plus). Elle possède un impact direct sur le nombre de sinistres en Dégâts des Eaux (augmentation de 20% de la fréquence en moyenne lorsque le bien possède une piscine), mais elle donne aussi une information sur le prix du bien et la qualité de vie des assurés puisque ce sont des assurés plus aisés qui possèdent une piscine.

— **Le code INSEE et la zone urbaine du bien** sont deux variables géographiques utilisées dans le portefeuille. Elles permettent de connaître la ville dans laquelle est situé le bien et de donner une information sur la qualité de vie dans celle-ci. La zone urbaine sépare les INSEE en trois catégories : Urbain, Banlieue, Rural. L'INSEE est également utile pour réaliser la jointure avec la base de données externes, présentée dans la section **2.4**.

2.3.2 Focus sur l'exposition

L'exposition représente la période de temps, sur une année, durant laquelle le bien est assuré. Ici, elle est comprise entre 0 et 1, avec par exemple, une valeur de 1 représentant

un contrat annuel et 0.5 un contrat sur 6 mois. C'est une donnée importante pour le contrat, puisque plus l'exposition est grande, plus le bien est exposé au risque.

Sur la cartographie figure **2.3**, la somme de l'exposition des contrats pour chaque ville est décrite :

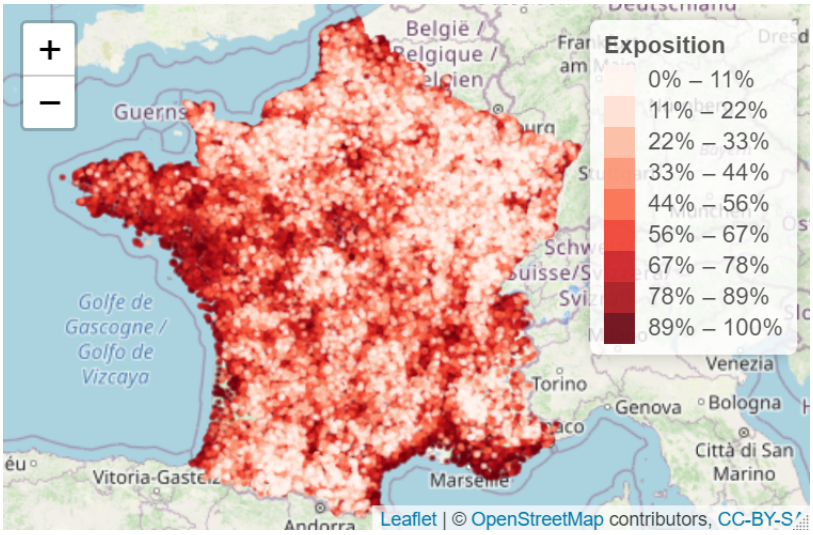


FIGURE 2.3 – Carte de la somme de l'exposition des contrats par commune, par quantile

Le choix de représenter l'exposition **par quantile** provient du fait que cette variable possède une densité particulière : Comme représentée dans le graphique **2.4**, cette densité est concentrée autour des petites expositions. De fait, le quantile à 90% de l'exposition est égal à 190, alors que le maximum est égal à 3700. Or, la palette de couleurs classique répartit équitablement l'ensemble des valeurs de l'exposition des communes en une dizaine de couleurs. Comme sur la carte figure **2.5**, la légende montre que les communes de 0 à 190 d'exposition ont une couleur similaire et ainsi la carte est illisible à cause de cette masse unicolore. Pour remédier à cela, on représente l'exposition en quantile.

Le portefeuille ne possède pas d'exposition dans la région Corse. La particularité du marché de l'assurance habitation dans cette région (beaucoup de logements secondaires, prix beaucoup plus élevés qu'en métropole, géographie particulière,...) font qu'ADDACTIS ne s'est pas risqué à simuler des contrats dans cette région. Hormis cette région, le portefeuille possède une exposition dans quasiment la totalité des INSEE de France (36000 INSEE dans le portefeuille).

Les villes les plus habitées sont généralement les villes les plus exposées dans le portefeuille. Pourtant, les trois plus grandes villes de France (Paris, Marseille et Lyon) ne sont pas représentées. En effet, le portefeuille ADDACTIS ne représente que le marché de l'assurance habitation maison. Le choix a été fait de séparer les marchés des maisons et des appartements, car représentant des risques différents.

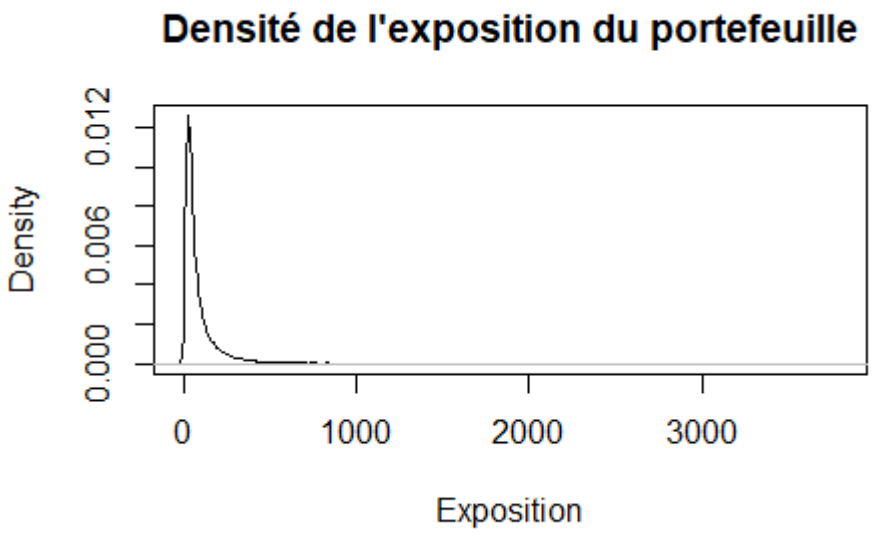


FIGURE 2.4 – Graphique de la densité de l'exposition

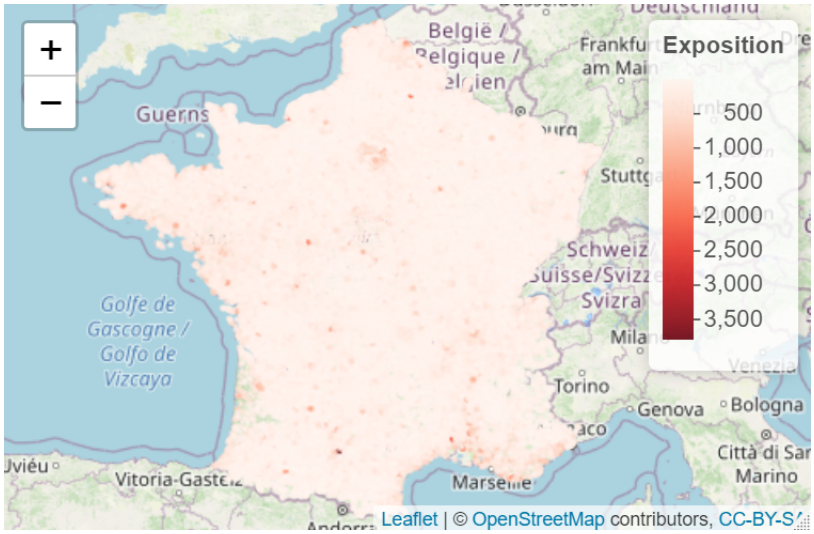


FIGURE 2.5 – Carte de la somme de l'exposition des contrats par commune, sans utiliser les quantiles

2.3.3 Focus sur la fréquence des sinistres

La fréquence des sinistres est le résultat de la division du nombre de sinistres de l'année par l'exposition sur l'année. Par exemple, un assuré ayant eu 1 sinistre sur 6 mois de contrat (donc une exposition de 0,5) aura une fréquence de : $\frac{1}{0,5} = 2$. Lié l'exposition

au nombre de sinistres permet de mieux appréhender le risque réel de l'assuré.

Cette base de données internes se veut proche des sinistres attritionnels marchés, dont voici une cartographie de la fréquence pour la garantie Dégâts des Eaux, figure **2.6** :

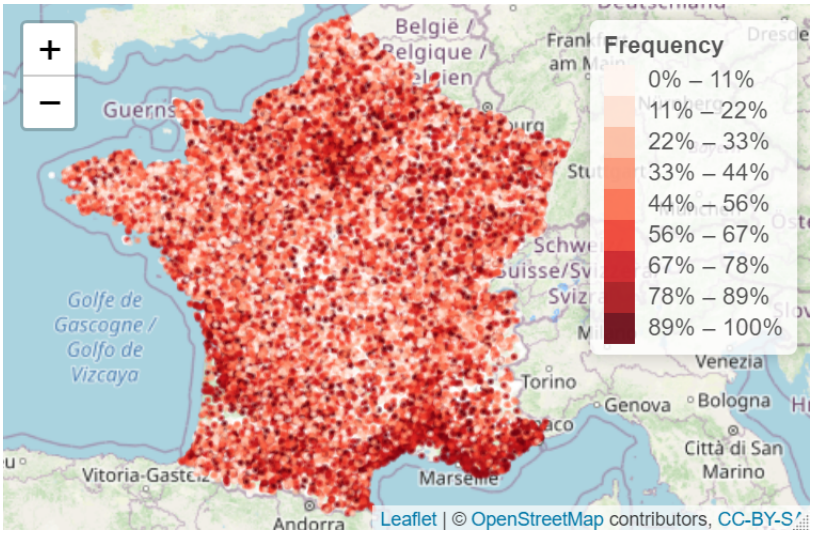


FIGURE 2.6 – Carte de la fréquence moyenne de la garantie Dégâts des Eaux par commune

La cartographie figure **2.6** montre que la fréquence est assez volatile sur le territoire français. En effet, la sinistralité est plutôt aléatoire, ce qui fait que deux communes peu exposées et voisines vont avoir des fréquences moyennes très différentes. La situation est inversée pour deux villes très exposées, pour qui la loi des grands nombres fait que leur fréquence moyenne s'approche quasiment de l'espérance de la fréquence.

De plus, cette cartographie montre que la sinistralité n'est pas répartie complètement aléatoirement sur tout le territoire français. En effet, des zones plus sombres peuvent se faire remarquer dans le bassin méditerranéen, dans l'estuaire de la Gironde ou encore dans l'Île-de-France. Ces zones à fortes fréquences se retrouveront dans le zonier si le GLM ne parvient pas à expliquer ces zones par des corrélations entre les variables internes et leur répartition géographique. Par exemple, le nombre de piscines influe sur la sinistralité en Dégâts des Eaux, mais est également sur-représenté dans le sud de la France. Peut-on alors dire que le bassin méditerranéen est risqué à cause de son nombre de piscines plus important ou est-il risqué en soi ? D'où l'importance de travailler d'abord avec un GLM sur les données internes pour réaliser un zonier.

La fréquence s'explique donc en partie par les caractéristiques des assurés et de leur bien, mais aussi par leur localisation en France. Pour cela, on utilise notamment une base de données externes dont celle utilisée est présentée dans la section **2.4** qui suit.

2.4 Présentation des données externes

Pour réaliser un zonier par XGBoost ou par Krigeage, des données externes sont à utiliser. Ces données externes sont des variables concernant une unité géographique comme la commune ou l'adresse. Il s'agit d'informations supplémentaires, ne dépendant pas des assurés.

La section **2.4.1** présente globalement la base de données externes. Certaines variables y seront explicitées avec leur définition, leur utilité et quelques cartographies.

La section **2.4.2** présente la méthode et les sources pour obtenir cette base. Ces travaux étant réalisés en amont de ce mémoire par d'autres personnes d'ADDACTIS France, seul un aperçu de la méthode sera donné.

2.4.1 Présentation générale de la base

Les données externes sont donc des informations extérieures au portefeuille des assurés. C'est-à-dire que les données ne sont pas issues du questionnaire de souscription que remplissent les assurés. L'intérêt de disposer de ces données est donc de pouvoir apporter une information supplémentaire à la sinistralité d'une garantie.

La base de données externes d'ADDACTIS France contient 130 variables pour environ 36000 codes INSEE. La maille INSEE est utilisée ici, c'est-à-dire que les variables ont une valeur pour chaque code INSEE, qui représente une commune. On distingue alors plusieurs catégories de variables :

— **Les variables de localisation** comme la longitude, la latitude ou l'altitude, permettent de situer la commune dans l'espace. Elles permettent également de regrouper les communes de montagne, de littoral ou encore du sud de la France qui partagent une sinistralité similaire. L'altitude est représentée sur la carte figure **2.7**.

— **Les données météorologiques et climatiques** comme la température, l'ensoleillement, les précipitations ou la vitesse du vent sont exprimées en maximum, en minimum et en moyenne pour chaque commune. Ce sont des données importantes, mais disponibles seulement pour les communes disposant de stations météorologiques. Il est donc nécessaire de faire un traitement adéquat pour lisser les valeurs pour toutes les communes françaises, comme expliqué dans la section **2.4.2**.

— **Les données routières** comme le nombre de radars, d'intersections ou la longueur moyenne des rues sont plus utiles pour l'assurance automobile que pour l'assurance habitation. Néanmoins, elles apportent une information sur l'urbanisation de la commune, ce qui peut avoir son importance pour un zonier habitation.

— **Les variables géographiques** comme la superficie, l'aire urbaine associée ou encore le nombre de communes voisines apportent des informations sur l'entourage de la commune. On peut alors savoir si la commune est une grande ville urbaine, ou si elle est plutôt située en banlieue ou alors si elle est loin de toute autre commune.

— **Les données économiques et sociales** comme le niveau de vie médian, le nombre d'habitants ou encore le nombre de commerces de proximité sont des données qui permettent d'avoir des informations sur la vie des habitants au sein de la commune. La population (c'est-à-dire le nombre d'habitants) est représentée sur la carte figure **2.8**.

— **Les données sur l'emploi** comme le transport des actifs, le taux de chômage ou encore le nombre de cadres donnent des informations sur le rythme professionnel des habitants. Certaines données se trouvent dans la base de données internes, mais individuellement. Ici, les catégories socio-professionnelles sont moyennées par âge et sur toute la commune, ce qui permet de comprendre le type de population habitant dans la commune.

FIGURE 2.7 – Carte de l'altitude moyenne des communes

2.4.2 Méthodologie de récupération et de traitement des données

Plusieurs sources ont été utilisées pour récupérer ces données :

— **L'INSEE** ou l'Institut National de la Statistique et des Études Économiques est un institut officiel fournissant un nombre conséquent de statistiques sur les communes françaises notamment. Elle apporte des données fiables sur la localisation, la géographie et certaines données économiques et sociales.

— **data.gouv.fr** est une plateforme officielle de diffusion de données publiques françaises. Elle apporte des données fiables sur l'emploi et certaines données économiques et sociales.

— **La NOAA** ou la National Oceanic and Atmospheric Administration est l'agence américaine officielle responsable de l'étude de l'océan et de l'atmosphère. Elle fournit des données fiables sur la météorologie et le climat en France.

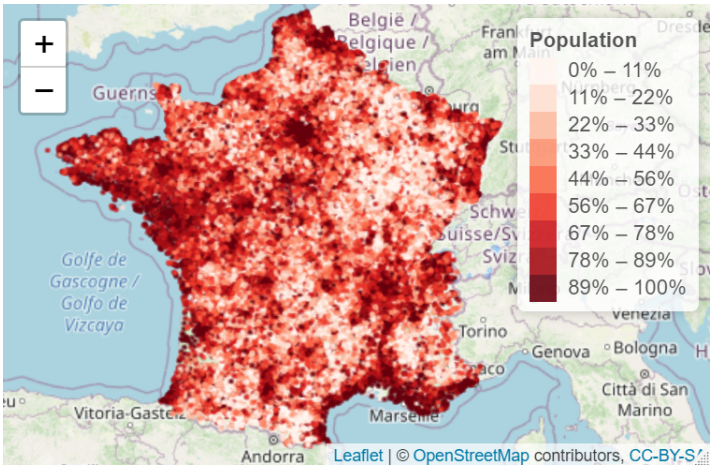


FIGURE 2.8 – Carte de la population (nombre d'habitants) des communes, par quantile

— **OpenStreetMap** est une plateforme en Open Source (NDLR : dont l'accès et l'utilisation sont libres) dont les utilisateurs de toutes parts enrichissent les bases de données. Chacun peut donc apporter ses connaissances sur la commune, ses outils de reconnaissance d'images satellites ou ses bases de données de sources diverses pour alimenter la plateforme. Les données routières proviennent d'OpenStreetMap.

Ces données récupérées forment alors des données brutes. Elles ne sont pas utilisables de suite.

On procède d'abord à un **nettoyage** après analyse de la pertinence, de la fiabilité, du rafraichissement des différentes sources et traitement des données manquantes. En effet, certaines sources peuvent être datées et d'autres moins fiables, car non officielles.

Ensuite, il peut y avoir une **création** de nouvelles variables pertinentes, stables et adaptées à chaque situation à partir des données brutes collectées (exploitables sur de nombreuses données brutes) comme la densité d'intersection, la densité de rues vis-à-vis de la population ou de la superficie de la commune, le calcul de distance à une zone d'emploi, etc.

Après, il y a des **analyses** de données extrêmes, atypiques et manquantes. Cela permet de retraiter les variables afin de les rendre utilisables. Les données manquantes sont plutôt nombreuses, puisque chaque commune n'a pas forcément de données collectées, particulièrement les communes avec peu d'habitants. Les données sont alors lissées et complétées à partir des données des communes voisines. Plusieurs **lissages/complétions** existent, allant du plus simple (remplacer les données manquantes par la moyenne française des autres données), au plus complexe (lissage par Krigeage par exemple).

La figure **2.9** donne un exemple de la gestion de la donnée sur la température. Cette donnée est issue de stations météorologiques (carte de gauche). Or, chaque commune ne dispose pas de station météorologique (162 en France). Par conséquent, il est nécessaire de

lisser la température sur l'ensemble des communes françaises. Pour ce faire, le **Krigeage** est utilisé avec d'autres données externes comme l'altitude, pour modéliser la température dans chaque commune, à partir de la localisation des stations météorologiques (carte du milieu). Le principe du Krigeage est expliqué dans la section **4**. D'autres lissages comme les *K plus proches voisins* ou le *Support Vector Machine* sont également montrés, mais sont moins précis (cartes de droite).

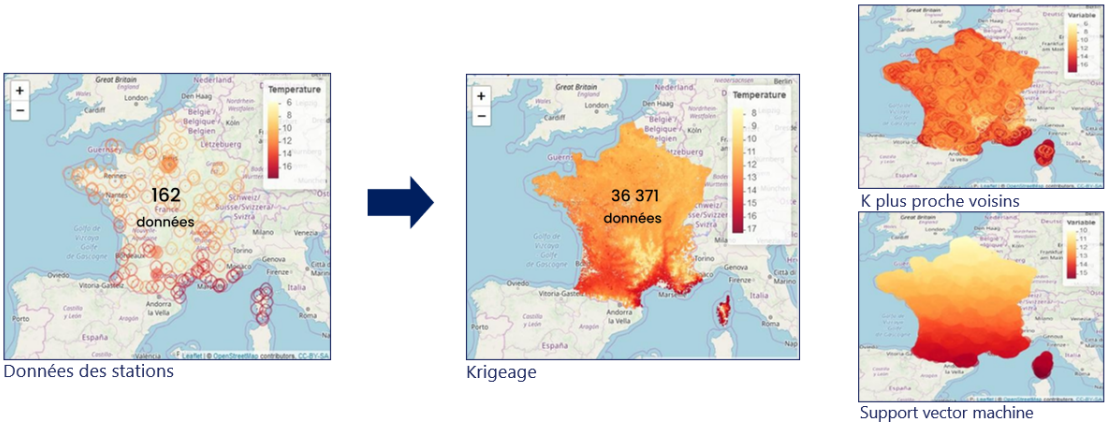


FIGURE 2.9 – Schéma du lissage de la température sur l'ensemble des communes

Chapitre 3

Méthodes classiques de construction d'un zonier

Ce mémoire a pour but d'observer l'apport du Krigeage dans la création d'un zonier MRH. Il convient donc d'étudier, au préalable, la création "classique" d'un zonier, afin de pouvoir la comparer au zonier par Krigeage. Ce chapitre se propose donc de revenir sur la méthodologie du zonier afin d'en exposer les aspects théoriques.

Le **GLM**, l'outil principal de la tarification MRH, est un modèle de régression statistique entre la variable aléatoire à prédire et les variables déterministes qui servent à prédire. Son principe est évoqué dans la section **3.1** Dans le cadre du zonier, le GLM est utilisé pour prédire la fréquence et le coût des sinistres avec seulement les données des assurés, en enlevant les données géographiques. L'objectif est de garder les résidus, c'est-à-dire la différence entre la variable aléatoire prédite et la variable observée. Ces résidus, agrégés par commune, donnent une information sur l'impact géographique de la sinistralité.

Pour expliquer au mieux cet impact géographique, il convient d'étudier plutôt les **résidus d'Anscombe**, comme dans la section **3.2**. En effet, les résidus d'Anscombe sont une transformation des résidus qui s'approche de la loi normale centrée réduite. Comparés aux autres résidus, ils sont facilement interprétables et calculables, ce qui fait qu'ils sont souvent privilégiés dans la littérature sur le zonier. Par la suite, deux méthodes de création de zonier vont d'abord être réalisées sur ces résidus : le lissage par crédibilité et la prédiction par XGboost.

Pour le **lissage par crédibilité**, détaillé dans la section **3.3**, le résidu lissé d'une commune est calculé en fonction du résidu initial de la commune et des résidus des autres communes, pondéré par l'exposition de celles-ci et leur distance à la commune. Ainsi, les communes proches ont un résidu quasiment égal, représentant un risque géographique similaire. Également, les communes moins exposées possèdent un résidu proche de celui de la commune très exposée la plus proche. Cela permet de réduire la variabilité des résidus. Cependant, cette méthode comporte des limites assez importantes, comme sa faible performance ou sa non-prise en compte de la géographie globale. Cette méthode

n'est plus utilisée maintenant que par quelques assureurs. Elle s'est fait remplacer par les méthodes prédictives.

Le but des méthodes prédictives est de prédire les résidus des communes, à l'aide de données externes sur la géographie, la démographie ou encore la météorologie des lieux. Les algorithmes d'eXtreme Gradient Boosting (ou "**XGBoost**") ou de Random Forest sont les plus utilisés. Ici, l'efficacité et la rapidité de l'algorithme XGBoost ont été des arguments privilégiés pour choisir l'algorithme à utiliser. Le fonctionnement de l'algorithme est résumé dans la section **3.4**. Le résidu est alors représenté comme une résultante de l'impact géographique sur les sinistres. La prédiction par des données externes permet également d'expliquer pourquoi certaines régions sont plus risquées que d'autres, en raison de leur altitude, population, température moyenne... De plus, cette méthode est bien plus performante et fine que le lissage par crédibilité, ce qui en fait le principal outil pour réaliser un zonier aujourd'hui. Il existe quand-même des limites à cette méthode, comme le sur-apprentissage et les prédictions dans les zones à faible exposition.

Après avoir discuté des deux méthodes pour regrouper les résidus de communes proches, il convient de réaliser le zonier à proprement parler. C'est-à-dire que les communes vont être regroupées dans un certain nombre de zones, correspondant à un risque similaire. Ce regroupement des communes se fait par **classification ascendante hiérarchique**, décrit dans la section **3.5**. Le principe est de comparer les distances entre les communes, puis de regrouper celles qui sont les plus proches, ensuite de regrouper les groupes les plus proches et ainsi de suite. Le fait de regrouper les communes permet d'avoir une information géographique simple du risque, avec une dizaine de modalités seulement. Cela évite d'utiliser toutes les variables externes dans la base de données internes et évite le sur-apprentissage. Ce zonier, une fois créé peut être incorporé comme une variable explicative dans le GLM qui servira pour la tarification d'un produit MRH.

3.1 Modèle Linéaire Généralisé

Les modèles linéaires généralisés sont les modèles les plus utilisés dans la tarification MRH en France, pour leur facilité d'utilisation et d'interprétabilité. Avec le GLM, il est possible d'établir un premier lien entre une variable aléatoire (comme la fréquence d'un sinistre, le coût d'un sinistre et la probabilité de souscription) et des variables déterministes (comme l'âge de l'assuré, la nature du bien et la localisation du bien). Il s'agit ici de faire un rapide rappel sur le GLM et les hypothèses sous-jacentes dans la section **3.1.1**. Ensuite, les métriques classiques de validation d'un GLM seront rappelées dans la section **3.1.2**. Ensuite, il sera question d'appliquer le GLM sur la base de données MRH dans la section **3.1.2**.

3.1.1 Aspects théoriques

Le but de la tarification d'un produit d'assurance est de trouver comment les caractéristiques des assurés influent sur la sinistralité. C'est-à-dire établir un lien entre

une variable **aléatoire** Y (la fréquence ou le coût d'un sinistre) et des variables **déterministes** $X_1, X_2, ..., X_n$(les caractéristiques de l'assuré). Pour ce faire, le modèle GLM propose d'étudier Y comme :

$$g(\mathbb{E}[Y]) = X\beta$$

.

avec

— Y la variable aléatoire à modéliser. L'hypothèse que les Y_i sont indépendants est faite puisque la sinistralité est indépendante entre les assurés. Cette variable aléatoire doit appartenir à la famille exponentielle, c'est-à-dire que sa densité f est de la forme :

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

où :

- y appartient au domaine de Y
- $\theta \in \mathbb{R}$ le paramètre canonique de la loi de distribution
- $\psi \in \mathbb{R}$ le paramètre de dispersion de la loi de distribution
- b une fonction de classe $\mathcal{C}^3$ avec sa première dérivée inversible
- a et c deux fonctions dérivables

Des lois discrètes comme la loi de Bernoulli ou de Poisson et des lois continues comme la loi normale ou Gamma font partie de la famille des lois exponentielles.

— $X_1, X_2, ..., X_n$ les variables déterministes qui servent à expliquer Y. Ces variables sont à choisir pour éviter le sur-apprentissage (lorsque beaucoup de variables sont prises), tout en évitant les modèles peu robustes (lorsque peu de variables sont prises). La sélection des variables peut se faire par *forward* (ou *backward*) *stepwise* ou encore par de la pénalisation LASSO. Le choix ne se fait pas seulement informatiquement, car une interprétation de ce qui fait que telle variable influe sur tel risque est nécessaire.

— g, la **fonction de lien**, doit être déterministe, bijective, strictement monotone et définie sur $\mathbb{R}$. Elle fait le lien entre la variable Y et les variables X. Pour la loi de Poisson, la fonction de lien canonique est la fonction ln.

— β le vecteur des paramètres à estimer. Ces coefficients sont estimés par **maximum de vraisemblance**. Les coefficients $\beta_1, \beta_2, ..., \beta_n$, correspondant aux variables explicatives $X_1, X_2, ..., X_n$, maximisent la log-vraisemblance du modèle :

$$\mathcal{L}(\beta) = ln \prod_{i=1}^{p} f(yi, \theta_i, \psi_i) = \sum_{i=1}^{p} ln(f(yi, \theta_i, \psi_i))$$

Si l'implication de β dans la formule précédente ne semble pas évident, il est lié au paramètre θ par les égalités :

$$\mathbb{E}[y_i|X] = b'(\theta_i) = g^{-1}(X_i^t\beta)$$

3.1.2 Métriques de validation

Après avoir choisi les variables explicatives et calculer les β, il convient de vérifier la qualité du modèle avec quelques métriques.

— **Le Root Mean Squared Error (ou RMSE)**

Le RMSE est l'erreur quadratique moyenne du modèle, mise à la racine carrée :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{p}(Y_i - \hat{Y}_i)^2}{p}}$$

avec Y les valeurs observées et $\hat{Y}$ les valeurs prédites par le modèle.

C'est une métrique classiquement utilisée pour la comparaison entre des modèles. Plus le RMSE d'un modèle est proche de 0, plus l'écart entre les valeurs prédites et observées est faible et donc plus la qualité du modèle est élevé.

Il est important de noter que cette métrique pénalise fortement les grands écarts entre valeurs observées et prédites, avec les écarts mis à la fonction carré.

— **La Déviance**

Lorsqu'un modèle possède autant de paramètres que d'observations, il est appelé un **modèle saturé**. Ce modèle est le modèle "parfait", en prédisant exactement les valeurs observées. La Déviance $\mathcal{D}$ compare ce modèle avec le modèle utilisé, en comparant leur log-vraisemblance :

$$\mathcal{D} = -2(\mathcal{L} - \widetilde{\mathcal{L}})$$

avec $\widetilde{\mathcal{L}}$ la log-vraisemblance du modèle saturé.

Plus la Déviance est proche de 0 plus l'écart entre le modèle utilisé et le modèle "parfait" est faible et donc plus la qualité du modèle utilisé est élevée.

Cependant, la Déviance diminue d'elle-même lorsque le nombre de variables explicatives augmente, même si le modèle n'est pas amélioré. Pour éviter cela, l'AIC et le BIC sont préférés.

— **L'AIC et le BIC**

L'AIC pénalise la Déviance avec le nombre de paramètres libres dans le modèle :

$$AIC = -2\mathcal{L} + 2k$$

avec k le nombre de paramètres libres du modèle.

Le BIC pénalise la Déviance avec le nombre de paramètres libres et la taille de l'échantillon utilisés dans le modèle :

$$BIC = -2\mathcal{L} + ln(p)k$$

avec k le nombre de paramètres libres du modèle et p la taille de l'échantillon.

Si deux modèles sont comparés, celui qui a l'AIC et le BIC le plus faible sera de meilleure qualité.

3.1.3 Application du GLM à la base de données

Le but est de prédire la fréquence et le coût moyen des sinistres des assurés, en dégâts des eaux, à l'aide de leurs caractéristiques et d'un modèle GLM. Pour cela, on utilise l'outil de tarification de ADDACTIS France, ADDACTIS Pricing©. Ce logiciel permet de gérer les variables d'une base de données, réaliser des GLM sur celle-ci et de générer les graphiques adéquats pour vérifier le modèle.

Par exemple, la figure **3.1** représente pour l'âge, en pointillés, la fréquence prédite et en ligne continue, la fréquence observée pour la garantie Dégâts des Eaux. Le groupement de l'âge s'est fait par rapport à l'exposition, représenté en vert. Il est à noter qu'un **spline** quadratique a été réalisé pour correspondre au mieux aux valeurs observées (les nœuds du spline sont situés aux âges 40 et 68).

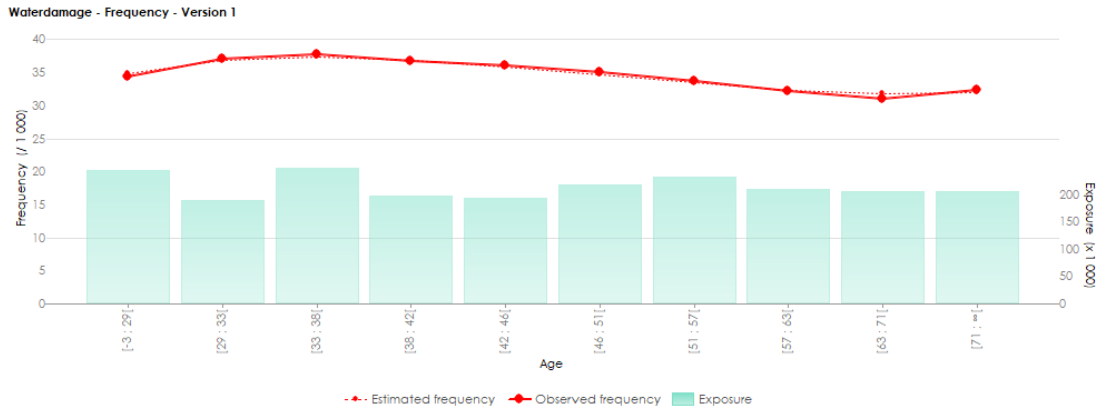


FIGURE 3.1 – Graphique de la comparaison entre le modèle GLM et les valeurs observées sur la fréquence de la garantie Dégâts des Eaux en fonction de l'âge des assurés

Les variables explicatives du GLM, pour la garantie dégâts des eaux, sont : l'année en cours, l'âge de l'assuré, l'ancienneté de l'assuré, le nombre d'enfants à charge de l'assuré, le nombre de pièces du bien assuré, la surface habitable du bien assuré, le statut socio-professionnel de l'assuré, la qualité de l'assuré, la formule du contrat d'assurance, l'extra activité d'hébergement dans le bien assuré et la présence d'une piscine dans le bien assuré.

La base de données est divisée en deux parties : une base d'entrainement (ou "**base de train**") et une base de validation (ou "**base de test**"). Le modèle GLM vient s'entrainer sur cette base de train, puis on vérifie sur la base de test le bon fonctionnement du modèle. Cela permet notamment de vérifier qu'il n'y a pas de **sur-apprentissage**, c'est-à-dire que le modèle prédit mal la base de test, en correspondant trop à la base de train. 70% des individus sont mis dans la base de train.

Pour la fréquence des sinistres, un **GLM Poisson** avec une fonction de lien $ln()$ sera privilégié et pour le coût moyen des sinistres, un **GLM Gamma** avec une fonction de lien $ln()$ sera utilisé.

3.2 Résidus d'Anscombe

Avec le GLM réalisé, les erreurs peuvent être récupérées. Ces erreurs, appelées résidus, représentent le risque qui ne dépend plus des variables explicatives du GLM. En théorie, ces résidus représentent donc le risque géographique. Ces résidus, agrégés par communes, vont être étudiés sous plusieurs formes, détaillées dans la section **3.2.1**. Enfin, les résidus d'Anscombe seront calculés et justifiés pour leur emploi dans le calcul du zonier, dans la section **3.1.2**.

3.2.1 Différents types de résidus

Les **résidus additifs** sont la simple différence entre les valeurs prédites et les valeurs observées tels que :

$$R_i = Y_i - \hat{Y}_i$$

avec R_i les résidus additifs, Y_i les valeurs observées et $\hat{Y}_i$ les valeurs prédites, pour l'individu i.

Ces résidus rendent compte de l'erreur de prédiction du modèle par individu et sont calculables assez facilement. Cependant, ces résidus restent assez simples et peu applicable au GLM Poisson.

Les **résidus multiplicatifs** ont ceci d'intéressant qu'ils gardent une forme multiplicative, en association avec la fonction de lien $ln()$:

$$R_i = \frac{Y_i}{\hat{Y}_i}$$

Cependant, dans une régression linéaire, les résidus devraient suivre une loi normale. Ici, dans le GLM Poisson, les résidus additifs suivent une loi de Poisson et les résidus multiplicatifs ne suivent pas une loi normale. Pour ce faire, une normalisation des résidus s'impose.

Les **résidus de Pearson** sont la division des résidus additifs par l'écart-type :

$$R_i = \frac{Y_i - \hat{Y}_i}{\sigma_{Y_i}}$$

Sa normalité en fait un outil classique de validité d'un modèle GLM. Cependant, un autre type de résidus est préféré dans la littérature sur le zonier.

3.2.2 Les résidus d'Anscombe

Les **résidus d'Anscombe** normalisent aussi les résidus, mais d'une autre manière : Comme Y suit une loi de Poisson, la transformée $Y^{1/3}$ va permettre de s'approcher d'une loi normale centrée réduite. En effet, les résidus d'Anscombe sont calculés comme suit :

$$R_i = \frac{3}{2}\frac{Y_i^{2/3} - \hat{Y}_i^{2/3}}{\hat{Y}_i^{1/6}}$$

La formule s'explique mieux en reprenant : $Z_i = Y^{1/3}$ (et $\hat{Z}_i = \hat{Y}_i^{1/3}$). Ainsi, la formule devient :

$$R_i = \frac{3}{2}\frac{Y_i^{2/3} - \hat{Y}_i^{2/3}}{\hat{Y}_i^{1/6}} = \frac{3}{2}\frac{Z_i^2 - \hat{Z}_i^{\,2}}{\sqrt{\hat{Z}_i}} = \frac{3}{2}\frac{Z_i - \hat{Z}_i}{\sqrt{\hat{Z}_i}}(Z_i + \hat{Z}_i)$$

On remarque alors que la formule d'Anscombe ressemble à un résidu de Pearson multiplié par $(Z_i + \hat{Z}_i)$. Avec ceci, on peut approximer une loi normale centrée réduite pour les résidus. L'un des atouts de cette méthode est de rassembler les résidus, de réduire l'écart-type pour avoir des résidus qui peuvent être comparés. Pour preuve, voici un aperçu de l'écart-type des différents résidus dans le tableau **3.1** :

Type de résidus	Écart-type
Résidus additifs	1,942
Résidus de Pearson	198,278
Résidus d'Anscombe	0,426

TABLE 3.1 – Tableau de l'écart-type des résidus

De plus, comparé aux résidus de Pearson, l'écart-type n'a pas besoin d'être estimé et les résidus d'Anscombe approchent une loi normale centrée réduite. Son calcul et son interprétabilité sont simples, tout en rentrant dans le cadre du GLM. C'est pour cela que ces résidus seront calculés et utilisés pour la suite de ce mémoire.

Ces résidus sont représentés graphiquement sur la figure **3.2**. On y distingue clairement certaines zones de résidus élevés (en rouge foncé) et certaines zones de résidus faibles (en blanc). Par exemple, le bassin méditerranéen, l'estuaire de la Gironde ou encore l'Île-de-France représentent ici des risques plus élevés que la moyenne, tandis que le Nord-Pas-de-Calais représente un risque plus faible que la moyenne. Ces zones sont construites selon le portefeuille d'origine, donc elles ne sont pas représentatives de la réalité du marché français. De plus, le résidu dépendant de la différence entre la fréquence prédite par le GLM et la fréquence observée, un résidu élevé stipule simplement que la commune a un risque sous-estimé par le GLM. L'important ici est que ces zones se

retrouvent dans le zonier, afin que le futur GLM puisse avoir cette information géographique. Également, il est important que les zones les plus risquées et les moins risquées se retrouvent aux bons extrêmes du zonier.

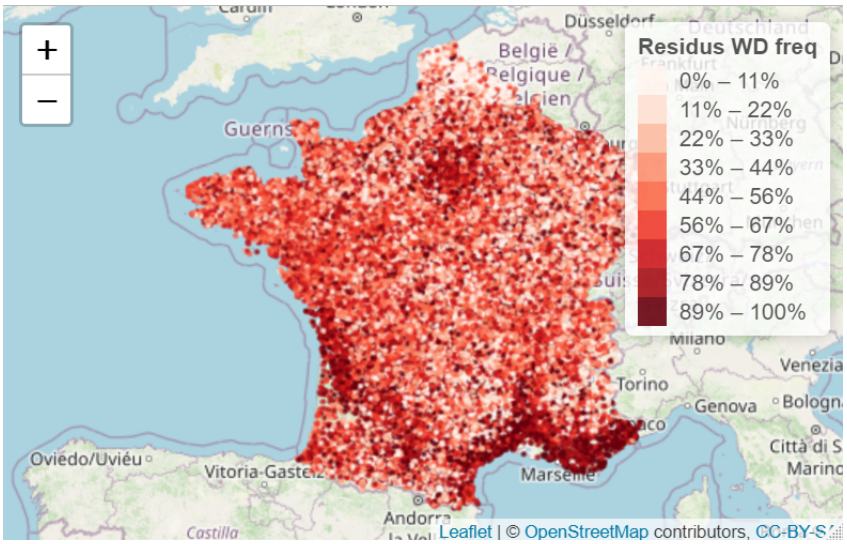


FIGURE 3.2 – Carte des résidus d'Anscombe de la fréquence de la garantie Dégâts des Eaux par quantile

3.3 Lissage des résidus par Crédibilité

Un zonier est la répartition de toutes les communes en un nombre défini de groupes de risques homogènes. Pour regrouper ces communes, les résidus des individus de celles-ci sont regroupés et étudiés. Cependant, la faible exposition de certaines communes fausse la vision du risque de celles-ci. Comme les résidus ne sont pas représentatifs pour ces communes, il convient de s'appuyer sur les communes les plus exposées pour qualifier le risque des communes peu exposées. De plus, les communes proches partagent la même géographie et lorsqu'on s'éloigne, de moins en moins de caractéristiques similaires rapprochent les communes. Les communes les plus proches devraient avoir des résidus similaires.

C'est dans ce contexte qu'un lissage par crédibilité peut-être effectué. En effet, le résidu lissé de chaque commune est en partie composé du résidu de la commune et de la moyenne des résidus des autres communes, pondérée par l'exposition et la distance à la commune de celles-ci. La modélisation mathématique détaillée se trouve à la section **3.3.1**.

Les paramètres, liés à la distance et à l'exposition, sont à optimiser pour obtenir le degré de lissage souhaité. Les métriques à optimiser et les résultats seront exposés à la section **3.3.2**.

Enfin, avec les bons paramètres, les résidus en fréquence de la garantie Dégâts des

Eaux seront lissés. Des cartographies compareront les résidus avant et après ce lissage, dans la section **3.3.3**.

3.3.1 Modélisation mathématique

La **crédibilité** est l'importance que l'on donne à quelque chose, comparativement aux autres. Ici, la crédibilité représente la proportion du résidu initial dans le calcul du résidu lissé. Cette crédibilité dépend de l'exposition, puisque les communes les plus exposées ont un résidu correspondant plus au risque géographique sous-jacent. Pour les communes à faible exposition, leur résidu lissé correspond, en grande partie, aux résidus des communes plus exposées proches. Plus la distance entre deux communes est grandes, moins leurs résidus s'influenceront.

Dans ce contexte, le résidu lissé $\bar{R}_i$ pour la commune i est :

$$\bar{R}_i = c(E_i)R_i + (1 - c(E_i))\frac{\sum_{j=1}^{p-1} R_j E_j f(d_{ij})}{\sum_{j=1}^{p-1} E_j f(d_{ij})} \tag{3.1}$$

avec :

- E_i, la somme des expositions des individus de la commune i
- $c(E_i) = \frac{E_i}{E_i+a}$, la **fonction de crédibilité**. Elle est comprise entre 0 et 1 et décroit lorsque l'exposition de la commune i décroit. Elle dépend d'un paramètre a à optimiser.
- R_i, la moyenne des résidus d'Anscombe des individus de la commune i.
- d_{ij}, la distance entre les communes i et j
- $f(d_{ij}) = \frac{1}{d_{ij}^n}$, la **fonction inverse à la distance**. Cette fonction décroit lorsque la distance augmente. Elle dépend d'un paramètre n à optimiser.

Les fonctions c et f permettent de lisser les résidus selon l'exposition et la distance. Leurs formes se veulent simples pour aider à la compréhension du modèle, bien qu'amenant moins de performance de prédiction. Ici, nous avons considéré qu'une seule fonction ; d'autres fonctions pourraient aussi être utilisées pour remplir le même rôle comme $f(d_{ij}) = e^{-d_{ij}^n}$, pour exprimer le fait que plus la distance est grande, moins la commune sera comptée dans le calcul du résidu. De plus, elles dépendent d'un paramètre chacun, qu'il faut optimiser.

3.3.2 Optimisation des paramètres

Ces paramètres a et n jouent un rôle primordial dans les fonctions c et f. En effet, lorsque a augmente, la fonction de crédibilité c diminue et le résidu initial sera moins pris en compte. Lorsque n augmente, la fonction inverse à la distance f diminue et les résidus des communes éloignées seront moins pris en compte.

Pour trouver la valeur de ces paramètres, on souhaite minimiser une métrique d'erreur. Pour ce faire, la base des résidus des communes est divisée en une base de train et une base de test.

La base de train contient 4000 communes. 80% de ces communes ont été choisies aléatoirement avec la fonction R *sample()* et un poids correspondant à leur exposition. Cela permet de favoriser l'apprentissage du modèle sur les communes possédant un résidu proche du risque géographique réel. 20% de la base de train a été choisi au hasard sur le reste des communes. Pour la base de train, la formule de la crédibilité est appliquée. Pour la base de test, on pose $c(E_i) = 0$ et seules les communes de la base de train sont dans la moyenne pondérée. Pour valider cela, le $\mathbf{R^2}$ et le $\mathbf{Q^2}$ sont utilisés.

Le R^2 et le Q^2 sont des métriques validant la prédiction sur, respectivement, la base de train et la base de test :

$$R^2 = 1 - \frac{\sum_{i=1}^{p}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{p}(Y_i - \bar{Y})^2}$$

avec

— Y_i les valeurs observées de la base de train

— $\hat{Y}_i$ les valeurs prédites de la base de train

— $\bar{Y}$ la moyenne des valeurs observées de la base de train

Le Q^2 possède le même calcul, mais sur la base de test. On pose R^2_{tot} la même métrique lorsque le modèle est entraîné sur toute la base. Il est à remarquer que le MSE (MSE = $RMSE^2$) se trouve au numérateur et est comparé à l'écart entre les valeurs observées et la moyenne de celles-ci. Le R^2 et le Q^2 sont compris généralement entre 0 et 1. Lorsque le modèle prédit parfaitement les valeurs observées, les deux métriques valent 1. Plus ils se rapprochent de 1, plus le modèle est d'une bonne qualité. Si l'une de ses métriques est négative, cela veut dire que le modèle est moins efficace que si les valeurs prédites étaient simplement remplacées par la moyenne. Enfin, le modèle possèdera un meilleur R^2 que Q^2, dû au fait que le modèle est entraîné sur les résidus de la base de train. Cependant, on souhaite avoir le Q^2 maximum, sinon le modèle sur-apprend et donnera de mauvaises prédictions pour de nouvelles valeurs.

On souhaite donc que les paramètres a et n maximisent le Q^2 plutôt que le R^2. En effet, lorsque $a = 0$, $C(E_i) = 1$ et $\bar{R}_i = R_i$, ce qui fait que $R^2 = 1$. Prendre les valeurs observées comme valeurs prédites n'a aucun sens. Il convient donc de trouver un moyen d'optimiser ces paramètres sans ce biais.

Comme le but est de comparer le zonier par Krigeage et le zonier par crédibilité, il est utile de comparer les Q^2 pour des R^2 similaires, afin de comparer l'apprentissage. Également, il est utile de comparer les R^2 pour la base entière, afin de comparer la qualité globale des modèles.

La démarche est donc la suivante :

— Lorsque le zonier par Krigeage a été optimisé sur la base de train de 4000 points, un R^2 a été obtenu.

— À l'aide d'une grille, c'est-à-dire un ensemble de couples (a, n), on trouve les paramètres a et n tels que le R^2 de la crédibilité se rapproche de celui du Krigeage, tout en maximisant le Q^2.

— Avec ces paramètres obtenus, on compare les R^2 sur la base entière.

3.3.3 Application du lissage aux résidus en fréquence de la garantie Dégâts des Eaux

Cette démarche est appliquée aux résidus en fréquence de la garantie Dégâts des Eaux. Après l'optimisation du Krigeage pour le zonier (développé dans la section **4**), le R^2 à approcher est de 59,25%. Pour ce R^2, le Q^2 maximum de 5,87% est obtenu avec $a = 200$ et $n = 2$. Pour la base totale, le R^2_{tot} est de 44,47%.

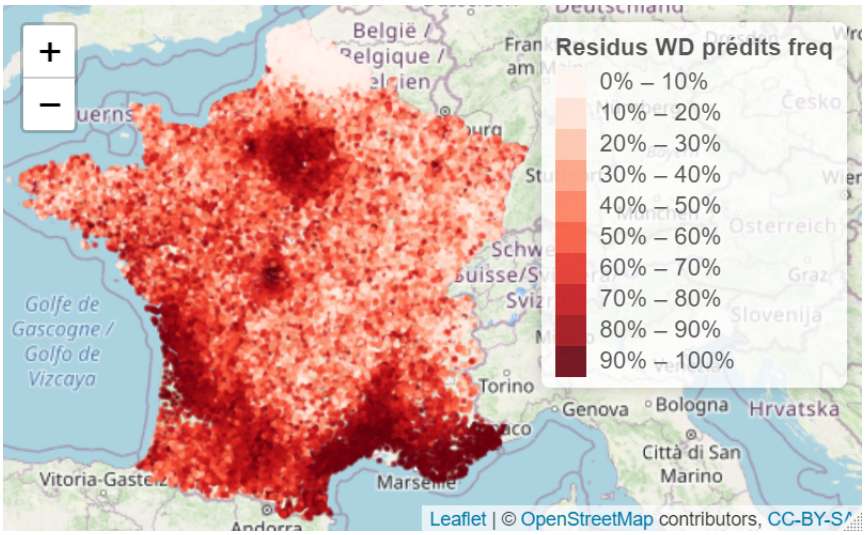


FIGURE 3.3 – Carte des résidus lissés de la fréquence de la garantie Dégâts des Eaux par quantile

La carte **3.3** rend compte du lissage des résidus. Les communes proches partagent des résidus similaires, pour faciliter le regroupement des communes en zone. Il est à noter que les résidus plus foncés sur la carte **3.2** des résidus initiaux, comme le sud-est de la France, se retrouvent plus groupés, formant quasiment des zones. Cette méthode est facilement réalisable sur R, en suivant la formule **3.1**.

Bien que le calcul soit plutôt rapide (environ 1 min pour la base entière), les performances prédictives peuvent-être améliorées. De plus, les zones formées ne donnent aucune information supplémentaire que leur localisation. Pouvoir expliquer pourquoi ces zones sont créées serait un atout supplémentaire.

3.4 Prédiction des résidus par eXtreme Gradient Boosting

Pour palier à ces limites ou ces manquements, les actuaires préfèrent utiliser des outils de Machine Learning et des données externes (avec aussi les données géographiques de la base de données internes). Comme pour le GLM, les algorithmes de Machine Learning

cherchent à prédire les résidus à partir des données externes. Ici, l'eXtreme Gradient Boosting (ou "**XGBoost**") est utilisé, pour sa rapidité et ses performances. Son principe est rappelé rapidement dans la section **3.4.1**. L'outil possède de nombreux paramètres à optimiser ainsi que les données externes à choisir en entrée. Cette optimisation est détaillée dans la section **3.4.2**. Enfin, l'XGBoost prédira les résidus en fréquence de la garantie Dégâts des Eaux, afin d'en voir les avantages et les inconvénients, expliqués dans la section **3.4.3**.

3.4.1 Principe de l'algorithme d'XGBoost

Connu pour ces performances lors des concours de data-science, notamment les compétitions Kaggle, cet algorithme est l'outil idéal pour prédire efficacement une variable avec un ensemble de données de départ. Les travaux de [Chen et Guestrin, 2016] expliquent parfaitement le fonctionnement de cet algorithme. Le principe est ici résumé et le lecteur est invité à parcourir cette source pour plus d'informations.

Le principe de l'**eXtreme Gradient Boosting** provient des trois mots composant son nom : "extreme", "gradient" et "boosting".

Le **boosting** propose de combiner plusieurs prédicteurs de faible qualité (ou "*weak learners*") afin d'avoir un prédicteur de bonne qualité. Ces modèles individuels sont calculés itérativement, c'est-à-dire qu'à chaque fois qu'ils sont entraînés sur la base de données, les modèles les moins performants reçoivent un poids plus lourd et les modèles sont de nouveau ré-entraînés sur la base où les poids sont corrigés et ainsi de suite. La priorité de ces modèles est de prédire les valeurs difficiles à prédire, ce qui peut entrainer un sur-apprentissage. Enfin, ces prédicteurs sont regroupés pour former un modèle fiable. Pour l'XGBoost, chaque itération utilise un arbre de régression de type **CART** dans la régression.

Le **gradient** est utilisé par l'algorithme pour calculer ces poids à chaque itération. En effet, à chaque itération, l'algorithme observe la **fonction de perte**, c'est-à-dire l'erreur entre la valeur prédite et la valeur observée. Avec le gradient de cette fonction, qui est sa dérivée partielle, la pente de la fonction est obtenue et donne des informations sur comment changer les paramètres pour converger vers le minimum. Cette méthode s'appelle la **descente de gradient**.

Le terme "**Extreme**" du nom de l'XGBoost provient du fait que la descente de gradient est "poussée à l'extrême". Là où le gradient de la fonction de perte apporte une information sur le minimum, l'algorithme d'XGBoost dérive partiellement le gradient, pour obtenir une seconde dérivée partielle de la fonction de perte. Ceci amène encore plus d'informations sur comment obtenir le minimum d'erreurs. De plus, des calculs parallèles sont effectués pour accélérer grandement le processus.

3.4.2 Optimisation des paramètres

Un aspect intéressant du XGBoost est la multitude de paramètres qui agissent sur la qualité de l'apprentissage, le sur-apprentissage, la complexité ou encore la vitesse de

convergence. Ces paramètres sont facilement changeables dans la fonction R *xgboost()* du package *xgboost*. Par exemple, le paramètre *eta*, compris entre 0 et 1, règle la vitesse de convergence de l'algorithme. Plus ce paramètre est élevé, plus l'algorithme sera robuste au sur-apprentissage, mais plus il sera lent à converger. Ainsi, tout est réglable selon ce qui est souhaité par l'utilisateur. Les paramètres utilisés sont listés ci-dessous :

- *nrounds*, le nombre d'itérations de l'algorithme.
- *max_ depth*, la longueur maximale d'un arbre.
- *colsample_ bytree*, le pourcentage de colonnes prises lors de la construction d'un arbre.
- *eta*, la vitesse de convergence.
- *gamma*, le minimum de la fonction de perte requis pour créer une nouvelle partition sur le nœud d'un arbre, pour réduire le sur-apprentissage.
- *min_ child_ weight*, la somme minimale de poids nécessaire pour créer une nouvelle partition.
- *subsample*, le pourcentage de la base concerné par la base de train.

Cependant, comparé à d'autres modèles, de nombreux paramètres sont à optimiser, ce qui peut augmenter le temps de calcul pour obtenir un modèle correct. Ici, une grille de paramètres est utilisée, grâce au package R *caret* et la fonction *expand.grid()*. Issues du même package, les fonctions *train()* et *trainControl()* permettent d'obtenir les meilleurs paramètres en un temps plus court. La méthode utilisée dans la fonction *trainControl()* est la **Cross Validation K-fold**.

La Cross Validation K-fold (ou "Validation Croisée sur K sous-ensembles") utilise le concept d'échantillonnage. La figure **3.4** montre le principe avec 5 sous-ensembles. En divisant la base de données en K sous-ensembles, l'algorithme considère d'abord le premier sous-ensemble comme une base de test et les K-1 autres sous-ensembles comme une base de train. Ensuite, le second sous-ensemble est considéré comme une base de test et ainsi de suite. A chaque itération, l'algorithme calcule le résidu additif R_i, puis, lorsque le dernier sous-ensemble est traité, il calcule R, la moyenne des résidus R_i. En réalisant cette démarche pour plusieurs valeurs d'un paramètre, les R sont comparés et la valeur du paramètre possédant un R minimum est gardé.

Enfin, le choix des variables externes est aussi important. Certaines variables inutiles peuvent perturber le bon fonctionnement du modèle. Garder les variables essentielles est important pour réduire le temps de calcul et éviter le sur-apprentissage. Pour vérifier l'importance des variables, il est utile d'utiliser un **Variable Importance Plot** (ou "VIP"). Le package *vip* fournit ces graphiques avec la fonction *vip()*. Le VIP des variables externes prises en compte est représenté sur la figure **3.5**.

3.4.3 Prédiction des résidus en fréquence de la garantie Dégâts des Eaux

Cette démarche est appliquée aux résidus en fréquence de la garantie Dégâts des Eaux. Pour la même base de train de 4000 points et un R^2 de 59,68%, un Q^2 de 7,63% est

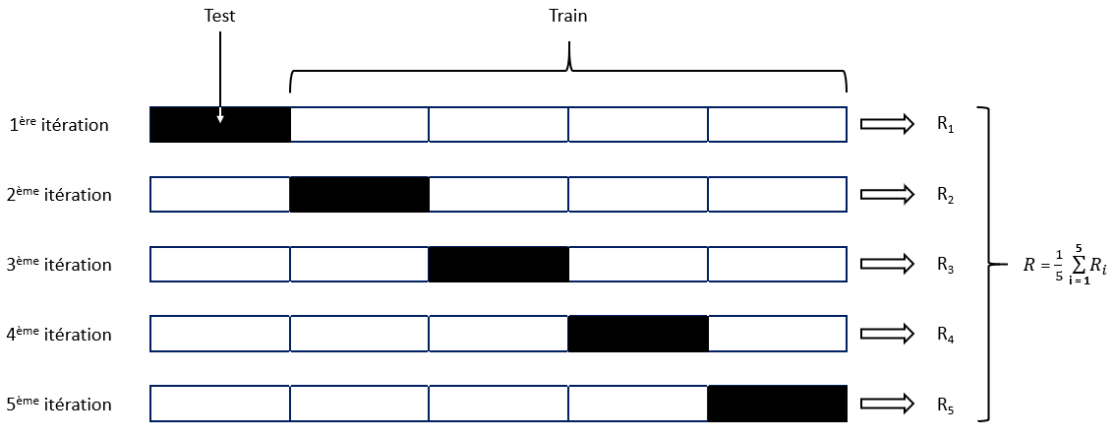


FIGURE 3.4 – Schéma du principe de la Validation Croisée K-fold avec K = 5

obtenu. Le Q^2 est largement supérieur à celui du lissage par crédibilité. L'apprentissage est meilleur pour l'algorithme d'XGBoost. Par ailleurs, le R^2_{tot} est de 49,62% ce qui est supérieur à celui de la crédibilité. L'XGBoost est donc plus performant.

Les paramètres finaux utilisés sont dans le tableau **3.2**.

nrounds	125
max_depth	5
colsample_bytree	0,6
eta	0,1
gamma	0
min_child_weight	25
subsample	0,9

TABLE 3.2 – Tableau des paramètres de l'XGBoost

La liste des variables externes utilisées se trouve dans le tableau **3.3** Les variables externes sont assez variées, certaines portent sur la météorologie des lieux, d'autres la démographie et d'autres sur la position géographique. Cependant, toutes n'ont pas la même importance, comme le démontre le VIP figure **3.5**.

Le temps de calcul est relativement court (30 sec), mais le fait de devoir optimiser les 7 paramètres augmente considérablement le temps de calcul (une dizaine de minutes pour obtenir un bon modèle sans compter la sélection des variables explicatives qui rallonge d'autant plus le temps).

La carte des résidus prédits par XGBoost de la fréquence de la garantie Dégâts des Eaux est représentée par la figure **3.6**. La carte ressemble beaucoup à celle de la crédibilité avec deux importantes différences cependant :

— Les communes proches ont parfois des résidus très différents. Par exemple, dans

0	Latitude
1	Longitude
2	Pourcentage de maisons
3	Zone urbaine
4	Taux de chomage
5	Altitude minimale
6	Altitude maximale
7	Nombre d'heures d'ensoleillement par an
8	Nombre de jours de pluie par an
9	Mediane du niveau de vie
10	Nombre d'habitants
11	Densité de population
12	Exposition
13	Pourcentage de la population entre 0 et 14 ans
14	Pourcentage de la population entre 75 et 89 ans
15	Pourcentage de la population entre 60 et 74 ans

TABLE 3.3 – Tableau des variables explicatives de l'XGBoost et de leur numéro dans le VIP

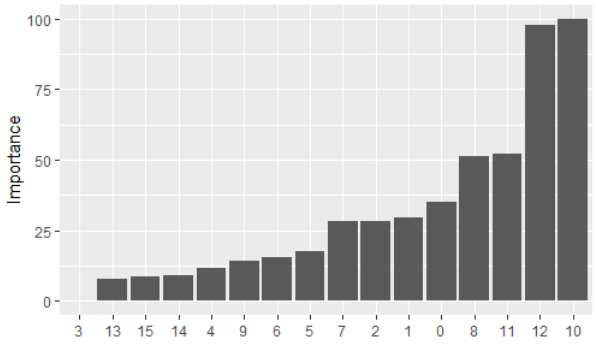


FIGURE 3.5 – Exemple de Variance Importance Plot

le Nord-Pas-De-Calais, des communes possèdent des résidus largement supérieurs (en rouge) à leurs voisins (en blanc). Cela peut correspondre à des villes qui ont un risque différent des communes rurales. Mais dans le cadre d'un zonier où on souhaite regrouper les communes proches, cela peut poser problème.

— L'impact des variables externes peuvent être aperçu. L'urbanisation est évoquée plus tôt, mais l'altitude est également bien visible avec par exemple le Massif Central, au centre de la France, dont les communes ont un risque moindre, comparé aux communes voisines, hors de la chaîne de montagnes.

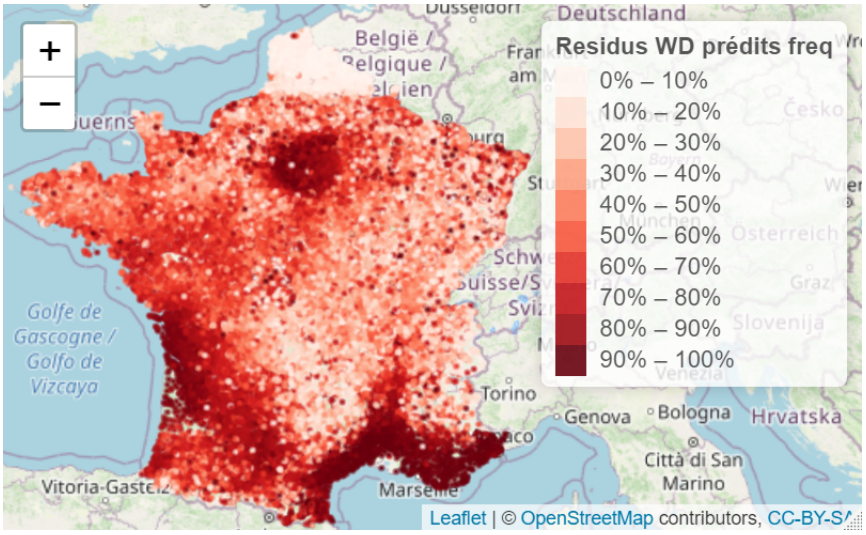


FIGURE 3.6 – Carte des résidus prédits par XGBoost de la fréquence de la garantie Dégâts des Eaux par quantile

3.5 Classification ascendante hiérarchique

Il reste une dernière étape au zonier : le regroupement des communes en zones. Pour ce faire, un algorithme de classification ascendante hiérarchique (ou "CAH") est utilisé. Son principe est décrit dans la section **3.5.1**. Dans la section **3.5.2**, cet algorithme est utilisé sur les résidus lissés obtenus auparavant avec les deux méthodes classiques.

3.5.1 Principe de la classification

Certains assureurs créent leurs zones "à la main", en observant la carte des résidus agrégés et en remarquant les communes voisines dont les résidus sont similaires. Par exemple, sur la figure **4** où les résidus sont prédits par XGBoost, certaines zones se voient facilement : l'Île-de-France en couleur très foncée, le Nord en blanc, le bord de la Méditerranée en couleur très foncée, le Massif Central en blanc, les bords de la Gironde en couleur foncée. Pour le reste de la France, il est plus compliqué de se prononcer. C'est pour cela que l'utilisation d'un algorithme est nécessaire.

Cet algorithme fait de la **classification ascendante hiérarchique**. Son principe est simple. Au début, toutes les communes forment chacune un groupe. Pour la base utilisée ici, il s'agit de 31940 groupes au début. Ensuite, quand deux communes ont un résidu proche et une distance entre elles courte, elles forment un nouveau groupe et ainsi de suite jusqu'à ce qu'il y ait plus qu'un seul groupe commun.

Ici, la critère à maximiser pour regrouper deux communes est défini selon la **méthode de Ward**. Deux groupes sont fusionnés si ce nouveau groupe maximise l'inertie inter-

groupes I telle que :

$$I = \frac{1}{n} \sum_{i=1}^{k} n_i d(g, g_i)^2$$

avec :

— n, le nombre de communes au total

— k, le nombre de groupes

— n_i, le nombre de communes dans le groupe $i, i \in [1, k]$

— g, le centre de gravité de l'ensemble des communes

— g_i, le centre de gravité du groupe $i, i \in [1, k]$

La notion de distance est importante ici. Plus tôt dans cette section, elle comprenait la distance spatiale et la différence de résidus. Pour calculer cette distance de classification, on récupère la distance spatiale entre toutes les communes et la distance entre tous les résidus des communes. Ensuite, on applique un **paramètre de distance** qui vient prendre une proportion de la distance spatiale et le reste de la distance entre les résidus. Ce paramètre ressemble au paramètre de crédibilité vu précédemment, sauf que, comme les distances ont des ordres de grandeur différentes, il sert aussi de coefficient de proportionnalité.

Ce paramètre est estimé empiriquement : c'est en testant un paramètre et en observant le zonier obtenu sur une carte qu'il est décidé s'il doit être modifié ou non.

3.5.2 Application aux précédents lissages

La CAH est appliquée aux résidus lissés obtenus par crédibilité et par XGBoost. Le paramètre de distance est testé pour différentes valeurs. Par exemple, sur la figure **3.7**, $t = 1$ est utilisé pour avoir un effet trop prononcé de la distance entre les résidus et $t = 0,5$ est utilisé pour avoir un effet trop prononcé de la distance spatiale. Ces cartes sont réalisées sur les résidus prédits par XGBoost. Pour $t = 0,5$, les zones "évidentes" vues précédemment (comme l'Île-de-France, le bord de la Méditerranée, etc.) ne figurent plus. Pour $t = 1$, il n'y a même plus de zones spatiales, comme les communes sont regroupées seulement par leur résidu. Pour $t = 0,97$, les zones correctes sont dessinées. C'est donc cette valeur qui est prise pour le zonier XGBoost.

Le nombre k de groupes est également important, comme montré sur la figure **3.8**. Avec un k trop faible, comme $k = 5$, les groupes ne sont pas assez précis et n'auront pas assez d'impact sur le zonier. Avec un k trop important, comme $k = 25$, le risque de sur-apprentissage est plus important. Le nombre final de groupes est de 14, puisque le zonier correspond aux attentes.

Les zoniers issus de la crédibilité et de l'XGBoost se trouve respectivement figure **3.9** et figure **3.10**. Les groupes sont classés dans l'ordre croissant de risque. Plus la commune se trouve dans un groupe élevé, plus elle possède un risque de sinistralité lié à la géographie des lieux élevé. Le zonier par XGBoost semble plus proche du zonier imaginé après la vue de la carte des résidus d'Anscombe en France. Les zones à haut

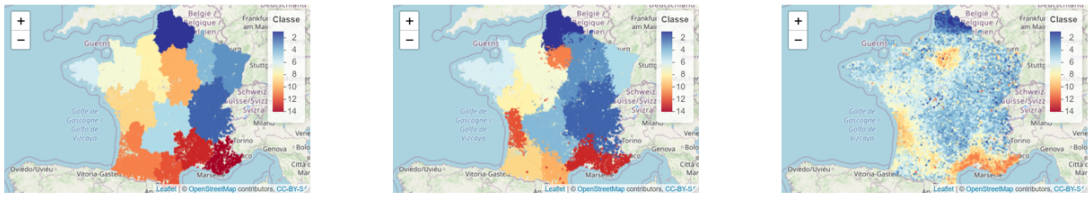


FIGURE 3.7 – Effet du paramètre de distance sur la CAH

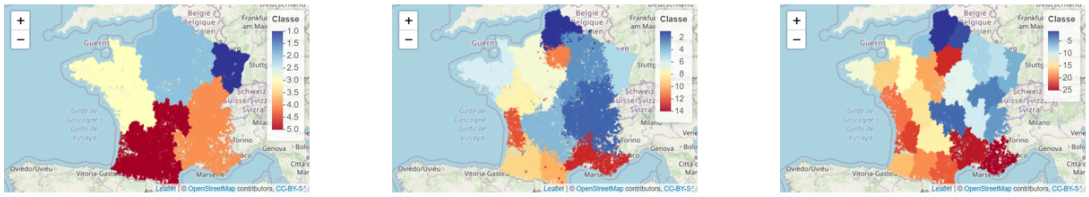


FIGURE 3.8 – Effet du nombre de groupes sur la CAH

risque comme l'Île-de-France, le bord de la Méditerranée ou encore le bord de la Gironde sont bien mieux représentées que pour le zonier par crédibilité.

3.6 Application des zoniers au GLM

Le but premier du zonier est d'être une variable du GLM, c'est-à-dire d'améliorer la prédiction des sinistres avec l'information géographique. Pour ce faire, il faut donc associer les différentes zones aux contrats et réaliser un nouveau GLM avec ces zones, comme dans la section **3.6.1**. Ensuite, les GLM avec le zonier par crédibilité et le zonier par XGBoost seront comparés avec différentes métriques et cartographies, comme dans la section **3.6.2**.

3.6.1 Application des zoniers au GLM

La première étape est de "nettoyer" les zoniers bruts pour qu'ils soient utilisables dans le GLM. Pour ce faire, chaque commune n'étant pas dans le zonier se verra attribuer la zone de la commune la plus proche. En effet, pour calculer la matrice de distance nécessaire à la CAH, il a fallu considérer 85% de la base ou la mémoire de l'ordinateur n'était pas suffisante. Il suffit après de rassembler les communes qui n'étaient pas comptées dans le zonier. Cependant, si deux communes de deux zones différentes sont équidistantes, une des deux zones sera prise au hasard. Cela permet d'inclure simplement tous les contrats dans le zonier.

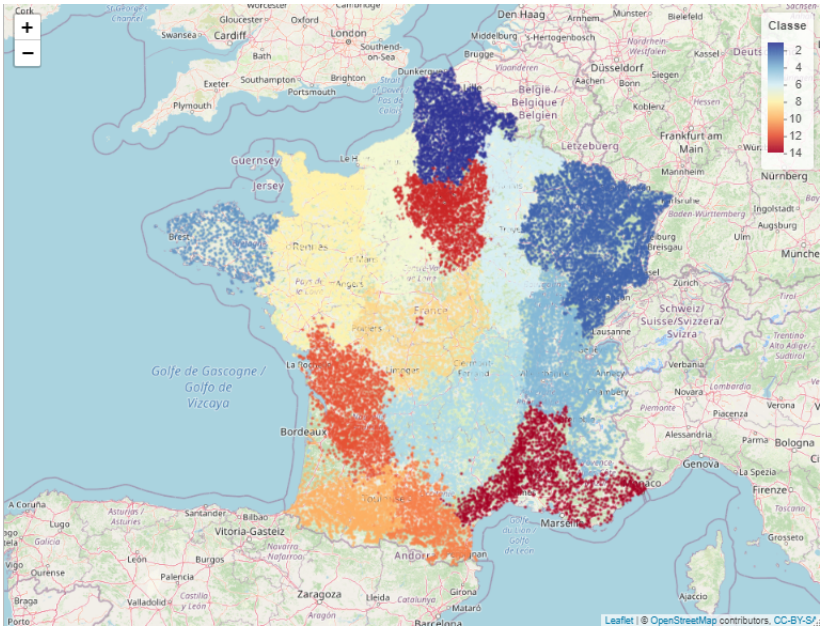


FIGURE 3.9 – Zonier pour le lissage des résidus par crédibilité, en fréquence sur la garantie Dégâts des Eaux

La seconde étape est de regrouper les zones d'exposition très faible. L'algorithme de CAH a pu considérer quelques communes dans une zone extrême. Si cela est sûrement vrai pour le portefeuille à l'instant T, il faut vérifier si cela n'est pas un sur-apprentissage. Un travail à la main est alors utile pour déceler les zones trop extrêmes ou trop petites. Ces zones sont associées à une zone voisine de taille plus importante.

Les zoniers finalement utilisés sont sur la figure **3.11** et la figure **3.12**. Les différences avec les zoniers bruts se voient assez peu sur la cartographie, mais sont nécessaires pour la bonne utilisation du zonier dans le GLM.

La troisième étape est de réaliser une jointure entre le zonier et la base de données internes. Le code INSEE est encore utilisé comme clé de jointure. La base de données internes est maintenant prête pour une nouvelle modélisation de la fréquence de sinistres en Dégâts des Eaux.

3.6.2 Comparaison des zoniers avec le GLM

Trois GLM sont alors réalisés : Un premier sans zonier (comme au départ), un second avec la variable zonier par crédibilité et un troisième avec la variable zonier par XGBoost. Les mêmes autres données internes sont utilisées pour comparer au mieux les zoniers.

Les premières métriques à regarder sont la Déviance et l'AIC. Ces métriques classiques du GLM permettent de comparer la qualité de deux GLM, car plus la Déviance et l'AIC sont faibles, plus le modèle est qualitatif.

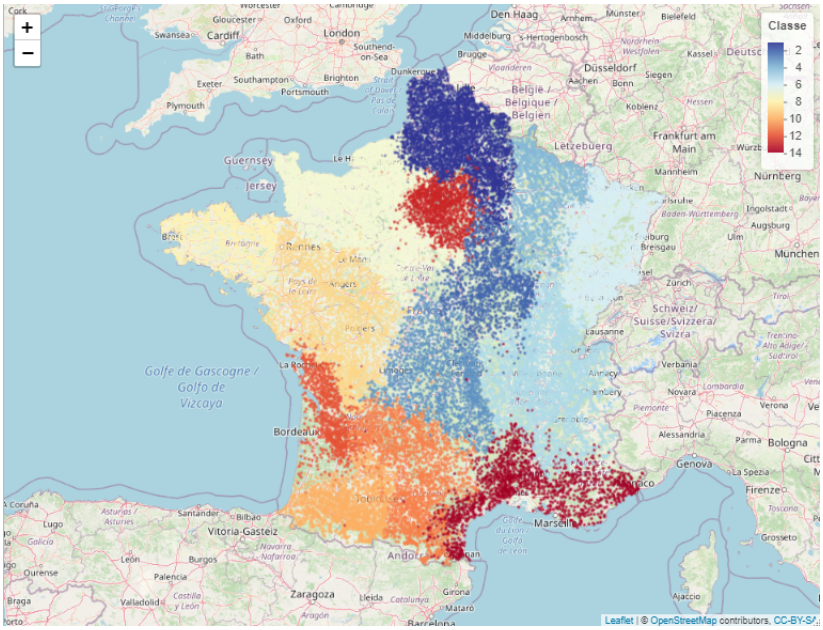


FIGURE 3.10 – Zonier pour le lissage des résidus par XGBoost, en fréquence sur la garantie Dégâts des Eaux

Le tableau **3.4** résume le résultat des métriques pour les 3 GLM. En plus de la Déviance et de l'AIC, le coefficient de Gini est calculé. C'est une autre métrique intéressante pour comparer les modèles, notamment sur leur pouvoir discriminatoire. Un coefficient proche de 100% signifie que le modèle est meilleur d'un point de vue discriminant et donc d'un point de vue prédiction.

GLM	Déviance	AIC	Gini
GLM sans zonier	648714	793702	0,1576
GLM avec zonier crédibilité	638357	783361	0,2729
GLM avec zonier XGBoost	638371	783385	0,2595

TABLE 3.4 – Tableau des métriques des GLM

Les conclusions de ce tableau sont que le zonier apporte toujours au modèle de base : Déviances plus faibles, AIC plus faibles et Gini plus élevés. Les écarts entre les GLM avec zonier sont assez faibles, hormis pour le Gini où le zonier par crédibilité semble plus impactant.

La seconde métrique à observer est la courbe des Bêtas. Les Bêtas sont les coefficients multiplicateurs associés à chaque zone par le GLM. L'intérêt est de regarder si la courbe est croissante, c'est-à-dire que l'association des communes pour chaque zone est pertinente par rapport aux autres. Plus la courbe est croissante, plus les zones sont pertinentes. Une autre chose à observer est l'écart entre le coefficient multiplicateur de la

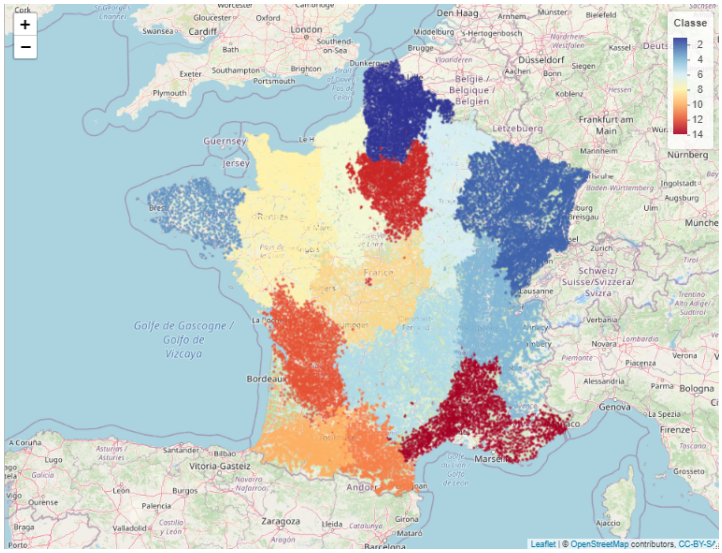


FIGURE 3.11 – Zonier final pour le lissage des résidus par crédibilité, en fréquence sur la garantie Dégâts des Eaux

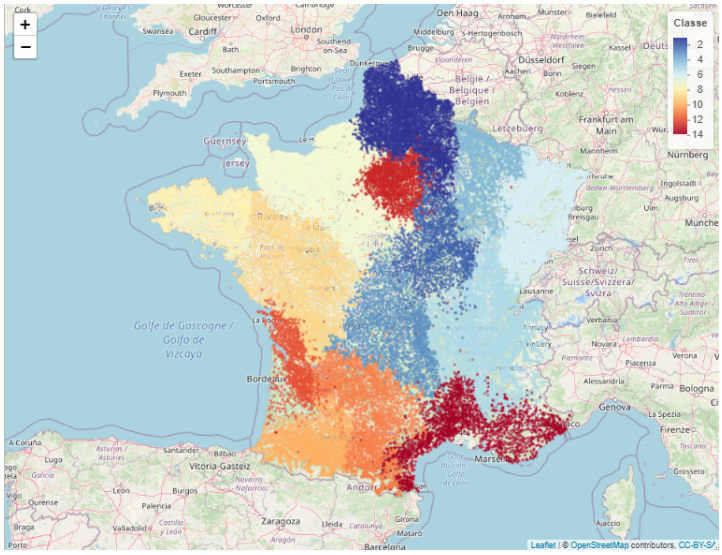


FIGURE 3.12 – Zonier final pour le lissage des résidus par XGBoost, en fréquence sur la garantie Dégâts des Eaux

première zone et celui de la dernière zone. Plus l'écart est grand, plus la discrimination l'est également.

Les graphiques figure **3.13** et figure **3.14** montrent les courbes de Bêtas pour le zonier par crédibilité et le zonier par XGBoost.

Les deux courbes de Bêtas sont croissantes, avec quelques zones mal-placées : Pour

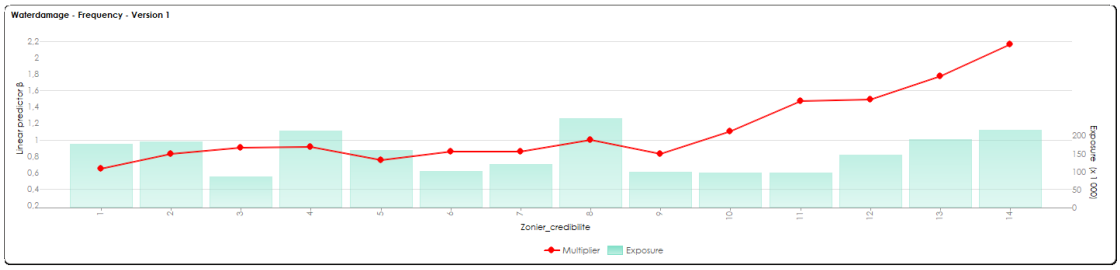


FIGURE 3.13 – Courbe de Bêtas pour le zonier par crédibilité

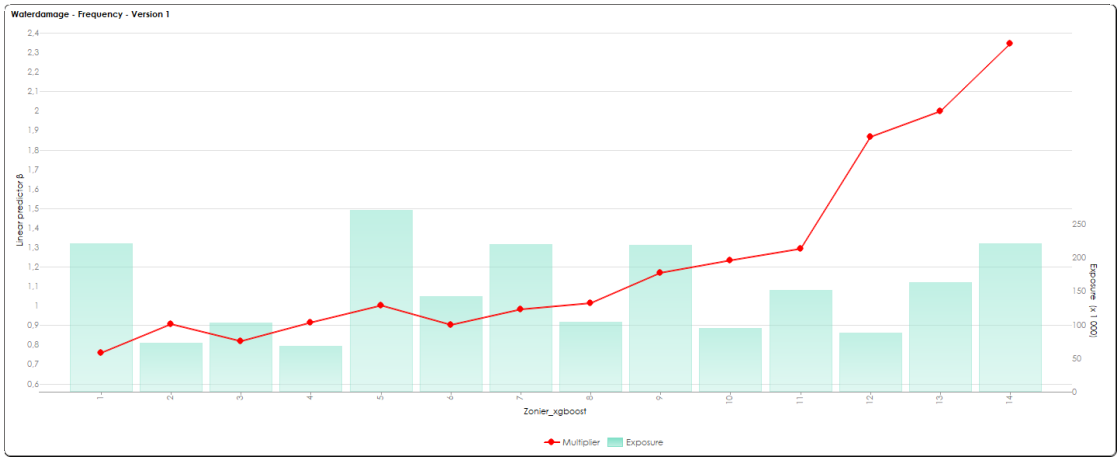


FIGURE 3.14 – Courbe de Bêtas pour le zonier par XGBoost

le zonier par crédibilité, la zone 5 possède un Bêta plus petit que celui de la zone 4 et pour le zonier par XGBoost, la zone 3 possède un Bêta plus petit que celui de la zone 2. Ce sont des exemples de zones mal placées, mais il est possible que l'année suivante, ces zones soient correctes. Pour comparer encore les modèles, l'écart des extrêmes est similaire et les courbes ont une croissance similaire.

La troisième métrique à étudier est la répartition des résidus des GLM sur une carte. En effet, ces résidus sont censés être indépendants des variables explicatives du GLM, donc du zonier. Si la cartographie des résidus montre qu'ils sont répartis totalement aléatoirement sur la France, alors le zonier a capté l'ensemble de l'information géographique.

Ces cartographies sont visibles figures **3.15**, **3.16** et **3.17** et représentent respectivement les cartographies des résidus du GLM sans zonier, du GLM avec le zonier crédibilité et du GLM avec le zonier XGBoost. On remarque qu'il reste encore des zones où les résidus sont globalement élevés et l'impact des zones est observable, mais les résidus semblent

plus aléatoires sur al France, notamment dans les zones foncées de la cartographie des résidus du premier GLM. En effet, comme les résidus prennent la fréquence prédite et que celle-ci est similaire sur toute une région, les résidus se ressembleront. Cela montre même l'importance du zonier dans le GLM. Les zones restantes résultent d'informations non captées par des données externes comme des corrélations entre des données internes et la localisation. Les cartographies sont similaires pour les deux modèles.

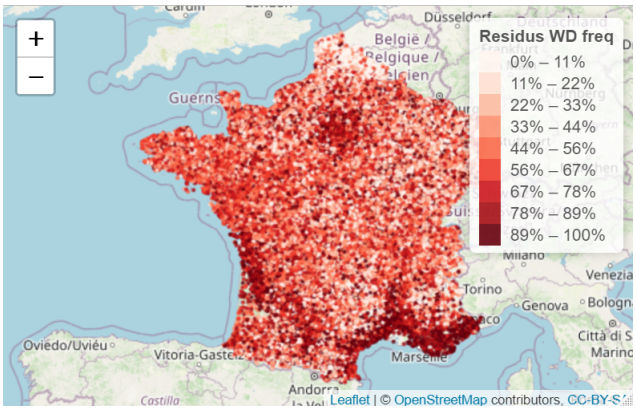


FIGURE 3.15 – Cartographie des résidus d'Anscombe en fréquence pour la garantie Dégâts des Eaux

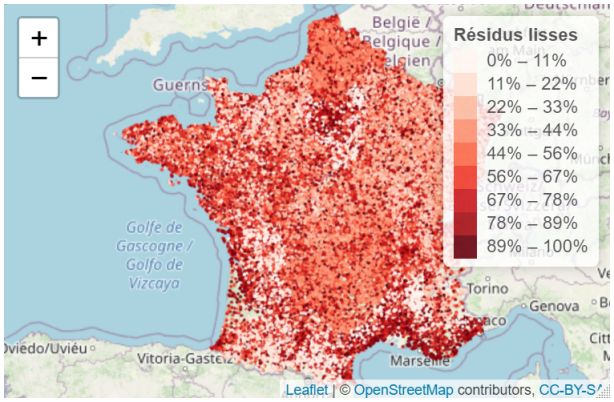


FIGURE 3.16 – Cartographie des résidus en fréquence pour la garantie Dégâts des Eaux après le GLM avec le zonier par crédibilité

En conclusion, pour ce portefeuille, pour ces modèles choisis, pour ces zoniers pris, les zoniers par crédibilité et par XGBoost se valent. Le zonier par crédibilité pourrait être meilleur avec un coefficient de Gini supérieur ($0{,}2729 > 0{,}2595$). Il faut prendre en compte que ce résultat serait peut-être différent pour un autre portefeuille, une autre garantie, une autre répartition sur la France, etc.

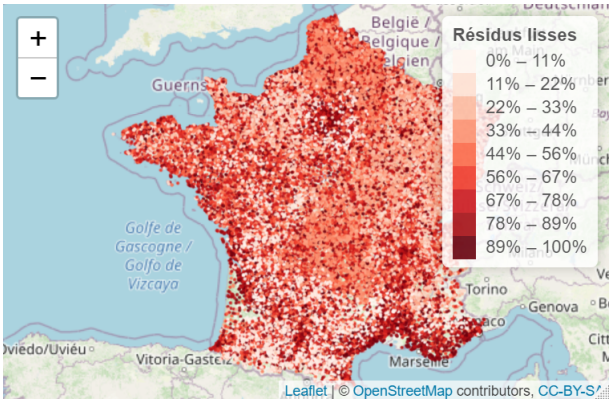


FIGURE 3.17 – Cartographie des résidus en fréquence pour la garantie Dégâts des Eaux après le GLM avec le zonier par XGBoost

Chapitre 4

Le lissage « prédictif » du Krigeage

Le lissage par crédibilité possède un concept intéressant : Calculer le résidu d'une commune en moyennant les résidus des communes proches, tout en favorisant celles qui sont le plus exposées. Cependant, les algorithmes de Machine Learning ont recalé ce type de lissage, en étant plus performant en termes de prédiction et en utilisant des données externes pouvant expliquer le risque géographique. L'idée est donc de chercher un moyen d'améliorer le lissage en le rendant plus performant et en utilisant des données externes. C'est alors qu'intervient le Krigeage, dont le lien avec le lissage par crédibilité est décrit dans la section **4.1**.

Le Krigeage est une méthode d'interpolation spatiale. L'idée est que pour connaître la valeur en un point, le Krigeage effectue une combinaison linéaire des valeurs connues des points adjacents. Les poids de la combinaison linéaire dépendent donc de la localisation des observations et de leur structure de dépendance spatiale. Cela veut dire que l'on ne considère pas seulement la distance entre les points, comme pour le lissage par crédibilité, mais également les données externes concernant les points.

Le modèle de base du Krigeage possède une partie déterministe, correspondant à son espérance, et une partie aléatoire, dont la covariance dépend des liens spatiaux entre les points. Cette covariance se modélise à partir de l'un des cinq noyaux usuels. La matrice de covariance ainsi que les noyaux seront détaillés dans la section **4.2**.

S'appuyant sur les dépendances spatiales des points, le Krigeage semble avoir une appétence pour une utilisation dans la création d'un zonier. Dans la section **4.3**, l'utilisation pratique du Krigeage sera exposée. Les paramètres, les valeurs externes et les résultats y seront détaillés.

Cependant, le Krigeage possède une limite majeure : des temps de calcul très coûteux. Cela pose beaucoup de problèmes dans la modélisation, puisque l'utilisateur doit se limiter pour pouvoir calculer et utiliser ces modèles. C'est pour cela que des solutions seront proposées dans la section **4.4**.

4.1 Du lissage par crédibilité au Krigeage

Pour classifier les résidus afin de réaliser un zonier, il est nécessaire de les lisser. La première idée, celle du lissage par crédibilité, est de faire une moyenne entre les résidus d'une commune et de ses communes voisines, pondérée par l'exposition de celles-ci et la distance avec la commune. Mais les limites de ce lissage, exposées dans la section **4.1.1**, poussent ce modèle au second plan.

En effet, les outils de Machine Learning, comme l'XGBoost, sont les plus utilisés pour agréger les résidus, grâce à leur performance prédictive supérieure et à l'apport des données externes. Ce sont ces dernières qui vont apporter au lissage par crédibilité une nouvelle dimension et surtout une performance accrue, comme décrit dans la section **4.1.2**.

Vouloir lisser avec des données externes, c'est en quelque sorte le principe du Krigeage. En effet, ce modèle propose que le résidu inconnu d'une commune soit écrit comme la combinaison linéaire des résidus des communes adjacentes. Ici, les coefficients linéaires sont calculés en fonction de la distance spatiale entre les communes, mais aussi en fonction des valeurs des données externes entre les communes. Ce point fait du Krigeage un modèle avantageux, comme expliqué dans la section **4.1.3**.

4.1.1 Les limites du lissage par crédibilité

Le lissage des résidus est important dans la création d'un zonier. Comme les résidus possèdent une part aléatoire, car ils sont dépendants des sinistres qui ont eu lieu, les résidus des communes peu exposées sont peu fiables pour étudier le risque géographique sous-jacent. Pour y remédier, le lissage par crédibilité propose de prendre une proportion du résidu de la commune et le reste comme une moyenne des résidus des communes autour.

Le plus important ici est que plus une commune est exposée, plus elle garde son résidu initial. A l'inverse, moins une commune est exposée, plus son résidu lissé est composé de la moyenne des résidus des communes autour. Suivant le même principe, les résidus des communes les plus exposées "influenceront" plus les résidus lissés des communes autour. Enfin, plus la distance entre deux communes est grande, moins leurs résidus seront pris en compte dans la moyenne. Pondérer par l'exposition et la distance entre les communes sont les grands principes du lissage par crédibilité.

Cependant, cette méthode possède de vrais défauts, qui sont la faible performance prédictive et la non-prise en compte de la géographie globale. Les petites communes en banlieue possèdent un résidu similaire à celui de la grande ville voisine, puisque celle-ci prend une place importante dans le calcul du résidu lissé, alors que le risque en ville et en banlieue est différent. Aussi, deux petites communes proches, l'une en amont d'une montagne et l'autre en aval, ont un résidu lissé similaire, puisque influencées par les mêmes villes alentours, alors que les risques sont différents. C'est pour cela que les actuaires utilisent principalement des algorithmes de Machine Learning avec des données externes.

4.1.2 L'apport des données externes

L'XGBoost, un outil puissant de Machine Learning, permet de prédire avec une précision correcte les sinistres à partir de données externes. Ces données externes, provenant d'organismes comme l'INSEE ou de sites en Open Source, concernent la sociodémographie, la météorologie ou encore la géographie des lieux. Elles sont un véritable atout puisqu'elles apportent des informations sur le risque géographique, pour comprendre ce qui influe sur la sinistralité. L'idéal serait de pouvoir les ajouter au lissage par crédibilité.

Si on revient à la formule **3.1** du lissage par crédibilité pour le calcul du résidu lissé $\bar{R}_i$:

$$\bar{R}_i = c(E_i)R_i + (1 - c(E_i))\frac{\sum_{j=1}^{p-1} R_j E_j f(d_{ij})}{\sum_{j=1}^{p-1} E_j f(d_{ij})}$$

Il existe un endroit où les données externes pourraient être considérées : la fonction inverse à la distance $f(d_{ij})$. En effet, ici la distance est seulement fonction de la longitude et de la latitude des communes. En prenant en compte l'altitude, la population, le nombre de maisons ou encore la pluviométrie, les résidus lissés seraient plus fiables. Alors le calcul d'un résidu lissé $\bar{R}_i$ pourrait être généralisé en :

$$\bar{R}_i = a_i + \sum_{j=1}^{p} R_j w(x_i, x_j)$$

avec $w(x_i, x_j) = w((lon_i, lat_i, alt_i, expo_i, ...), (lon_j, lat_j, alt_j, expo_j, ...))$ et a_i une constante finie.

Ici la longitude, la latitude, l'altitude et l'exposition sont pris en compte, mais plus de données externes sur les communes peuvent être ajoutées.

Cette modélisation est celle du **Krigeage**.

4.1.3 Le Krigeage, un modèle avantageux

Le **Krigeage** prend ses racines en 1951 avec les travaux de D.G. Krige, un ingénieur des mines sud-africain. Ses travaux furent repris par le français G. Matheron qui lui donnera le nom "Krigeage" (ou "Kriging" en anglais) en l'honneur de son prédécesseur. Si ce modèle stochastique est toujours utilisé dans la géostatistique, les autres branches des statistiques appliquées comme la météorologie s'y intéressent. Dans le cadre de l'assurance, le sujet se développe avec des mémoires sur son utilisation dans la réassurance [Picabea, 2019], dans l'assurance vie [Zurfluh, 2019] ou encore dans l'assurance automobile [Chamoulaud, 2020].

Cet intérêt pour le Krigeage provient du fait qu'il fait le lien entre interpolation, probabilités et données externes. L'**interpolation** diffère des modèles de régression utilisés habituellement, comme les régressions linéaires ou les réseaux de neurones. La figure **4.1** montre la différence entre les deux principes :

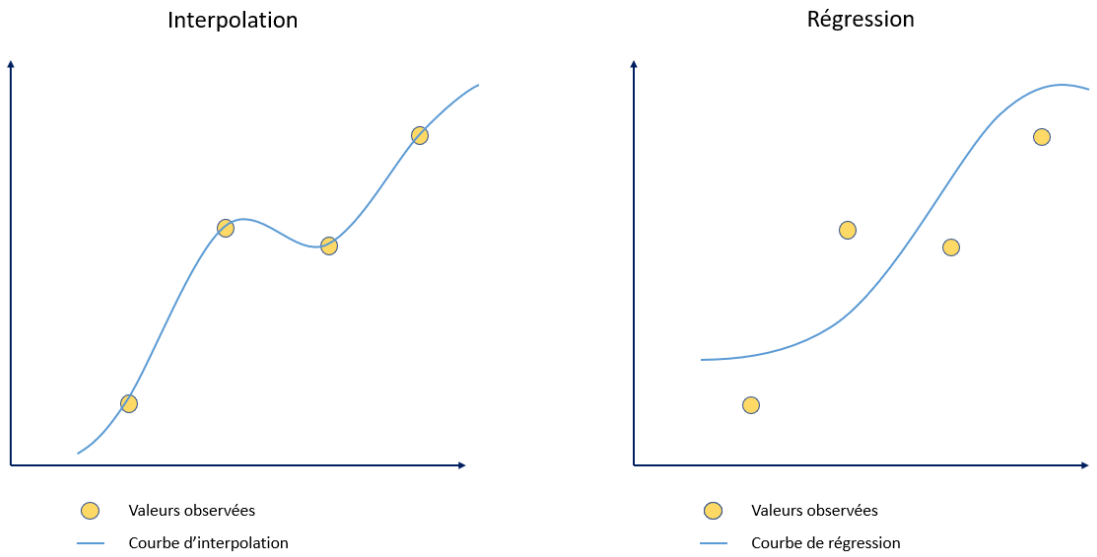


FIGURE 4.1 – Schéma de la différence entre l'interpolation et la régression

L'interpolation propose un modèle qui coïncide exactement avec les valeurs observées. Lorsque les valeurs observées sont issues d'une variable aléatoire, comme des résidus, ce modèle pousse au sur-apprentissage. En effet, si la valeur change ou qu'une commune voisine non apprise est testée sur le modèle, le modèle serait alors trop éloigné de la réalité. C'est pour cela que les outils de Machine Learning préfèrent la régression. Le modèle est ajusté selon les valeurs observées de la variable aléatoire. Bien que moins précis sur les valeurs observées, il permet de mieux prédire, en moyenne, les valeurs non-observées.

Le Krigeage se base pourtant sur un modèle d'interpolation. Mais ce modèle est particulier : Le Krigeage est plus précisément un modèle d'interpolation spatiale stochastique, c'est-à-dire que les valeurs observées sont considérées comme des réalisations d'une variable aléatoire, plus précisément d'un processus gaussien. Le modèle va simuler un grand nombre de processus gaussiens conditionnés à avoir les valeurs observées en les points observés. Ensuite, la moyenne de cet échantillon de processus sert d'estimateur pour les points non observés et l'ensemble des valeurs des processus pour un point non observé sert d'intervalle de confiance.

Ce principe est résumé dans le schéma **4.2** :

4.2 Formalisation mathématique

Le modèle du Krigeage, en tant qu'interpolation spatiale stochastique, approxime la variable aléatoire avec une multitude de processus gaussiens. Il modélise la variable à expliquer avec une partie déterministe et une partie aléatoire. La partie déterministe

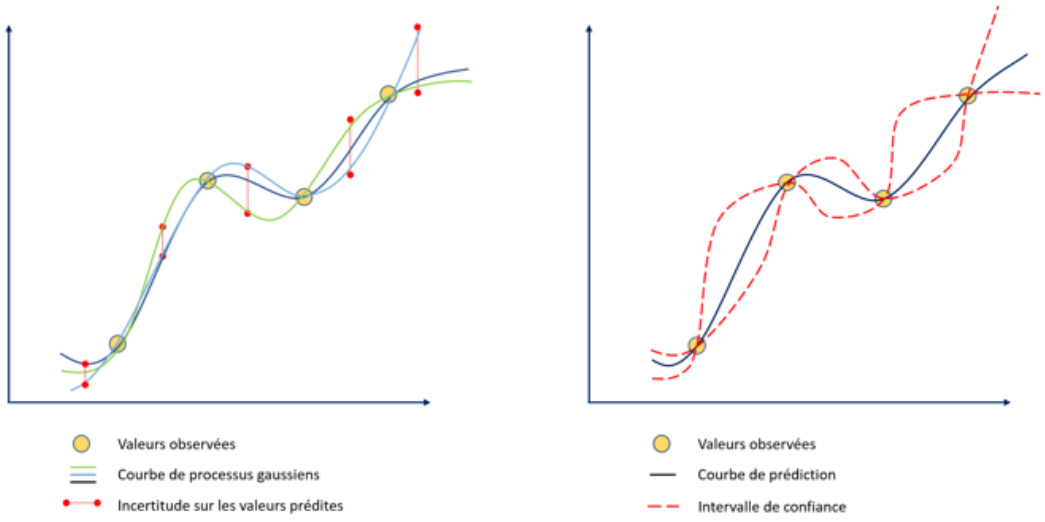


FIGURE 4.2 – Schéma du principe du Krigeage

correspond à son espérance et peut s'écrire et s'étudier de plusieurs manières, qui sont les différentes formes de Krigeage, décrites dans la section **4.2.1**.

La partie aléatoire est un processus aléatoire dont la structure de dépendance doit être calculée avec les valeurs observées. Sa covariance est modélisée à partir d'un des cinq noyaux usuels, décrits dans la section **4.2.2**.

La modélisation de la valeur d'un nouveau point est expliqué dans la section **4.2.3**. Grâce aux propriétés sur les processus gaussiens, on peut exprimer l'espérance et la variance de cette nouvelle valeur (conditionnellement aux valeurs observées). L'espérance de la nouvelle valeur correspond à la somme entre une constante et une combinaison linéaire des résidus, comme évoqué plus tôt. Ici, les poids sont explicités.

Enfin, l'effet pépite est discuté dans la section **4.2.4**. Comme les résidus sont aléatoires, une interpolation exactement sur les valeurs observées est inutile. Pour se préserver de l'aléatoire, le modèle du Krigeage apporte un effet pépite qui modélise la variance non nulle pour une distance minimale entre deux points. Cependant, la notion de bruit est préférée ici.

4.2.1 Les différentes formes du Krigeage

Le Krigeage n'est pas un modèle de régression, comme les modèles de Machine Learning par exemple. Il s'agit d'un modèle d'**interpolation spatiale stochastique**. Cela veut dire que le modèle considère que toutes les valeurs observées sont des réalisations d'une variable aléatoire. Le Krigeage va même plus loin en posant que cette variable aléatoire est un **processus gaussien** de la forme :

$$Z(x) = \mu(x) + \delta(x)$$

avec :

— x, les données explicatives d'une commune

— Z, la variable à expliquer qui est un processus gaussien

— μ une fonction déterministe pour l'espérance de Z

— δ un processus aléatoire stationnaire, d'espérance nulle, de variance σ^2 et de fonction de corrélation r, tel que :

- $\mathbb{E}[\delta(x)] = 0$
- $Cov(\delta(x), \delta(\widetilde{x})) = k(x, \widetilde{x}) = \sigma^2 r(x, \widetilde{x})$

Les trois types de Krigeage sont définis par la tendance $\mu(x)$ choisie pour le modèle.

— Le Krigeage **simple** où $\mu(x) = m$ est une constante connue

— Le Krigeage **ordinaire** où $\mu(x) = \mu$ est une constante inconnue

— Le Krigeage **universel** où $\mu(x) = \sum_j f_j(x)\beta_j$ est une combinaison linéaire de fonctions de x

Le Krigeage universel est le plus "complet". Cependant, le temps de calcul est déjà assez long pour optimiser le Krigeage pour considérer le Krigeage universel. Le Krigeage ordinaire est donc considéré et la constante inconnue est estimée par maximum de vraisemblance.

En résumé, la variable à expliquer est la somme entre sa tendance et la corrélation spatiale entre les données, comme le montre le schéma **4.3** :

4.2.2 Covariance et noyaux

La forme de la covariance est, elle aussi, spécifiée. En effet, pour un nombre d de variables explicatives, la covariance est modélisée comme :

$$k(x, \widetilde{x}) = \sigma^2 r(x, \widetilde{x}) = \prod_{j=1}^{d} k_j(x^j, \widetilde{x}^j)$$

avec les fonctions k_j de l'une des cinq formes suivantes, appelées **noyaux** et $\alpha \in \mathbb{R}^+$:

— Noyau Exponentiel : $k_j(y, y') = \sigma^2 e^{\frac{-|y-y'|}{\alpha}}$

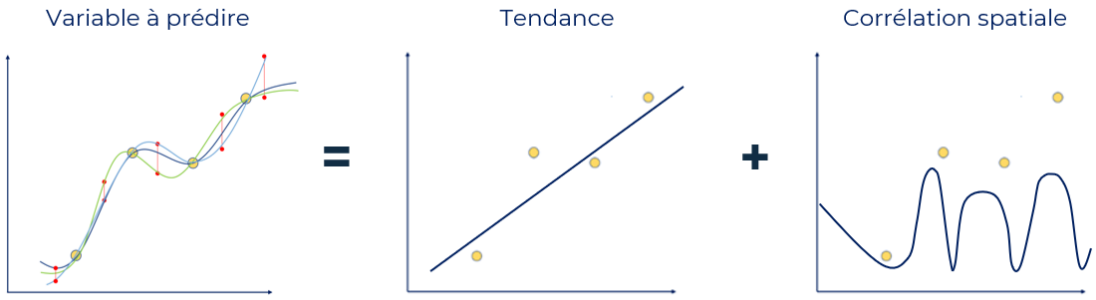


FIGURE 4.3 – Schéma du modèle mathématique du Krigeage

— Noyau Puissance Exponentielle : $k_j(y, y') = \sigma^2 e^{\frac{-|y-y'|^m}{\alpha^m}}$ et $m \in]0, 2[$

— Noyau Matern 3/2 : $k_j(y, y') = \sigma^2(1 + \frac{\sqrt{3}|y-y'|}{\alpha})e^{\frac{-\sqrt{3}|y-y'|}{\alpha}}$

— Noyau Matern 5/2 : $k_j(y, y') = \sigma^2(1 + \frac{\sqrt{5}|y-y'|}{\alpha} + \frac{5(y-y')^2}{3\alpha^2})e^{\frac{-\sqrt{5}|y-y'|}{\alpha}}$

— Noyau Gaussien : $k_j(y, y') = \sigma^2 e^{\frac{-(y-y')^2}{2\alpha^2}}$

Le choix du noyau est empirique : il faut essayer les différents noyaux et prendre celui qui modélise le mieux les données sur lesquels on apprend le modèle. Pour aider dans le choix du noyau, il faut savoir que les classes des noyaux sont différentes et donc que leurs régularités sont aussi différentes. Ainsi, les noyaux Exponentiel et Puissance Exponentielle sont de classe $\mathcal{C}^0$; le noyau Matern 3/2 est de classe $\mathcal{C}^1$; le noyau Matern 3/2 est de classe $\mathcal{C}^2$ et le noyau Gaussien est de classe $\mathcal{C}^\infty$. Plus le degré de la classe est élevé, plus le processus sera régulier. Le choix empirique du noyau sur la base de données s'appelle l'**analyse variographique**.

Les paramètres μ, a et σ (ainsi que m si le noyau Puissance Exponentielle est utilisé) sont à estimer par maximum de vraisemblance sur les données d'apprentissage, pour pouvoir prédire une nouvelle valeur.

4.2.3 Prédiction d'une nouvelle valeur

Avec la tendance et la covariance du processus, il est possible de prédire les résidus des communes dans la base de test. Pour ce faire, il est essentiel de rappeler qu'on fait l'hypothèse que la variable à prédire Z est un processus gaussien de formule : $Z_n \sim \mathcal{N}(\mu_n, K_n)$ avec : $Z_n = (Z(x_1), ..., Z(x_n))^t$, $\mu_n = (\mu(x_1), ..., \mu(x_n))^t$ et $K_n = (k(x_i, x_j))_{i,j=1,...,n}$ où $k(x_i, x_j)$ prend la forme d'un produit de noyaux vu précédemment.

Pour prédire le résidu en une nouvelle commune $x*$, le modèle utilisé s'écrit :

$$\begin{pmatrix} Z_n \\ Z(x*) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_n \\ \mu(x*) \end{pmatrix}, \begin{pmatrix} K_n & k(x_n, x*) \\ k(x_n, x*)^t & \sigma^2 \end{pmatrix} \right)$$

avec $k(x_n, x*) = (k(x_1, x*), ..., k(x_n, x*)^t)$.

En utilisant la **stabilité par conditionnement des vecteurs gaussiens**, $Z(x*)|Z_n$ est un vecteur gaussien de formule :

$$Z(x*)|Z_n \sim \mathcal{N}(m(x*), \theta^2(x*))$$

avec :

— $m(x*) = \mu(x*) + k(x_n, x*)^t K_n^{-1}(Z_n - \mu_n)$

— $\theta^2(x*) = \sigma^2 - k(x_n, x*)^t K_n^{-1} k(x_n, x*)$

L'estimateur du résidu pour la commune $x*$, qui est le résidu prédit utilisé en pratique, est $m(x*)$. Le lien avec la crédibilité se fait lorsque $m(x*)$ est réécrit comme une combinaison linéaire des résidus des communes adjacentes :

$$m(x*) = a(x*) + \sum_{i=1}^{n} \lambda_i(x*) Z(x_i)$$

avec : $a(x*) = \mu(x*) - k(x_n, x*)^t K_n^{-1} \mu_n$ et $\lambda(x*) = k(x_n, x*)^t K_n^{-1} = (\lambda_1(x*), ..., \lambda_n(x*))$

De plus, l'incertitude autour de la prédiction est quantifié par $\theta^2(x*)$. On dispose alors d'une information sur la fiabilité de la prédiction. Cette incertitude est même représentable sur une carte pour voir quelles zones sont fiables et quelles zones sont à considérer autrement.

4.2.4 L'effet pépite

Le modèle de Krigeage propose d'ajouter un **effet pépite**. Il s'agit du minimum de variance entre deux points. Cela sert notamment à ne pas interpoler sur les points, mais dans un voisinage très proche des points. Le package *DiceKriging* propose d'estimer lui-même cet effet pépite, en demandant à l'utilisateur seulement un effet pépite initial.

Cependant, de meilleurs résultats sont obtenus autrement. En effet, les résidus observés sont aléatoires et donc il faudrait quantifier leur volatilité qui dépend de chaque commune. On suppose que le résidu d'Anscombe de chaque individu r_l suit une loi normale centrée de variance σ^2. On souhaite calculer la variance du résidu d'une commune, c'est-à-dire la variance de la moyenne des résidus des assurés de la commune. En posant n_j l'exposition de la commune et en supposant que les individus d'une même commune

sont indépendants :

$$\begin{aligned} Var(r_{commune_j}) &= Var(\frac{1}{n_j}\sum_{l=1}^{n_j} r_l) \\ &= \frac{1}{n_j^2} Var(\sum_{l=1}^{n_j} r_l) \\ &= \frac{1}{n_j^2}\sum_{l=1}^{n_j} Var(r_l) \\ &= \frac{n_j}{n_j^2}\sigma^2 = \frac{\sigma^2}{n_j} \end{aligned}$$

Cette variance pour la commune est plutôt logique : si la variance des résidus individuels augmente, la variance du résidu de la commune augmente et si l'exposition augmente, la variance du résidu de la commune diminue. L'hypothèse d'indépendance des résidus dans une commune, n'est pas toujours vraie : dans le cas d'une catastrophe naturelle, les sinistres des individus de la commune seront liés. Mais dans la base de sinistres utilisée, seuls les sinistres attritionnels en Dégâts des Eaux sont comptés, donc l'hypothèse fonctionne.

En sachant cela, on ajoute du bruit autour des résidus, équivalent à la variance du résidu de la commune. Des essais ont été réalisés avec $\frac{1}{\sqrt{exposition}}$ ou $\frac{1}{exposition^2}$, mais le résultat était moins convaincant. Comme la variance par résidu individuel est estimé par la variance sur R, il est possible de poser λ un facteur correcteur, mais les résultats étant assez pertinent sans ce paramètre, il ne sera pas appliqué. Le bruit autour des variables observées s'ajoute avec le paramètre *noise.var* de la fonction *km*.

4.3 Optimisation sur les résidus en fréquence de la garantie Dégâts des Eaux

Ce modèle est appliqué aux résidus en fréquence de la garantie Dégâts des Eaux. Pour la même base de train de 4000 points, un R^2 de 59,25% et un Q^2 de 12,71% sont obtenus. Le Q^2 est largement supérieur à celui du lissage par crédibilité et de l'XGBoost. L'apprentissage est meilleur pour le modèle du Krigeage.

Les performances des noyaux sur 3000 points sont représentées dans le tableau **4.1** :

Le noyau exponentiel propose un bon Q^2 pour un temps de calcul moins long que celui des autres noyaux. C'est donc le noyau exponentiel qui sera choisi.

Pour choisir les variables externes à utiliser, les variables sont mises une part une dans le modèle. La variable dont le modèle possède le meilleur Q^2 est gardée. Puis, les variables sont à nouveau rajoutées une par une, etc. La liste des variables externes utilisées est : "Longitude", "Latitude", "Densité de population", "Nombre d'heures d'ensoleillement

Noyaux	Q^2	Temps de calcul (en min)
Exponentiel	10,4	7,9
Puissance Exponentielle	10,8	18,7
Matern 3/2	10,1	9,8
Matern 5/2	10,1	15
Gaussien	0,9	11,8

TABLE 4.1 – Tableau des paramètres de l'XGBoost

par an", "Nombre de jours de pluie par an", "Altitude minimale", "Pourcentage de maisons", "Nombre d'habitants". Il y a moins de variables externes que pour l'XGBoost pour réduire le temps de calcul. En effet, plus il y a de variables, plus la distance est longue à calculer, plus la matrice de covariance est longue à exprimer et l'algorithme fonctionne donc moins vite. La section **4.4** discute du temps de calcul et de solutions pour le réduire.

La carte des résidus prédits par Krigeage de la fréquence de la garantie Dégâts des Eaux est représentée par la figure **4.4**.

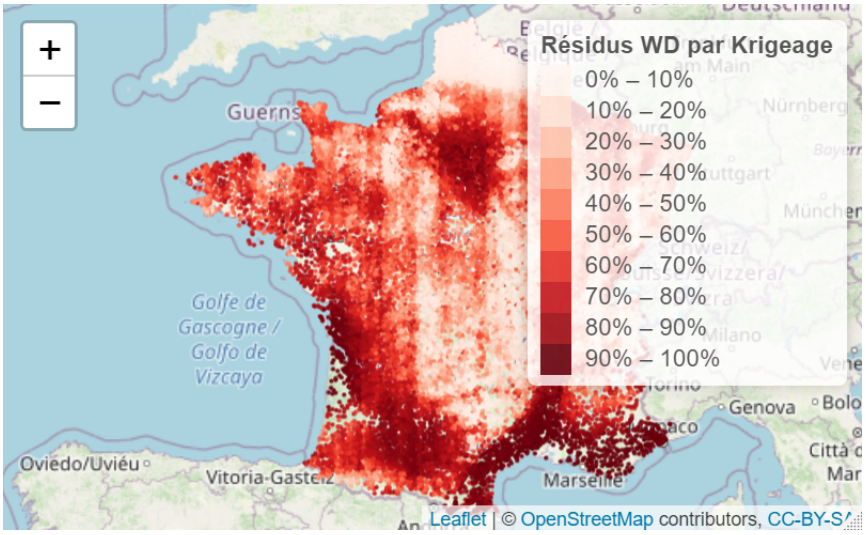


FIGURE 4.4 – Carte des résidus en fréquence de la garantie dégâts des Eaux prédits par le modèle du Krigeage

Les résidus sont plus lissés que pour le lissage par crédibilité et l'XGBoost. C'est dû au fait que les résidus sont prédits en n'étant pas dans la base de train.

4.4 Réduction du temps de calcul

Si le modèle est d'une bonne qualité, il comporte une importante limite : de longs temps de calcul. Ces temps de calcul augmentent de manière exponentielle par rapport au nombre de points de la base de train, comme expliqué dans la section **4.4.1**. Dans cette section, il est question de voir des solutions pour y remédier (section **4.4.2** et la solution utilisé pour ce mémoire (section **4.4.3**).

4.4.1 Un temps de calcul exponentiel

La valeur du résidu en une nouvelle commune est estimé par :

$$m(x*) = \mu(x*) + k(x_n, x*)^t K_n^{-1}(Z_n - \mu_n)$$

Il est donc nécessaire d'inverser la matrice de covariance K_n de taille n*n. Plus la valeur de n est grande, plus la taille de la matrice est importante et plus son inversion est longue. De plus, la matrice augmente en ligne et en colonne, ce qui fait que de passer de n = 1000 à n = 1001 est plus facile que de passer de n = 4000 à n = 4001. En effet, une inversion de matrice se fait en complexité $\mathcal{O}(n^3)$ pour l'algorithme de classique de Gauss-Jordan. L'algorithme de Strassen permet d'inverser une matrice avec une complexité $\mathcal{O}(n^{log_2(7)})$.

Cela a donc une répercussion sur le calcul du modèle de Krigeage. Les temps de calcul, en minutes et en fonction du nombre de points dans la base de train, ont été compilé dans le tableau **4.2** :

Nombre de points	Temps de calcul (en min)
1000	0,5
2000	3,5
3000	17
4000	65

TABLE 4.2 – Tableau des temps de calcul par rapport au nombre de points

Une régression a aussi été réalisé sur Excel pour rendre compte de l'augmentation **exponentielle** du temps de calcul, sur la figure **4.5**.

Au-delà de 4000 points, le temps pris par la modélisation n'est plus acceptable. Surtout qu'il s'agit d'un modèle à optimiser et dont les variables externes sont à choisir en faisant apprendre un nouveau modèle à chaque fois. C'est pour cela que la base de train est composée de 4000 points.

Il est à noter également que les modèles prennent beaucoup de place (2.8 Go pour 4000 points) et que cela impacte énormément la mémoire vive et les prochains calculs.

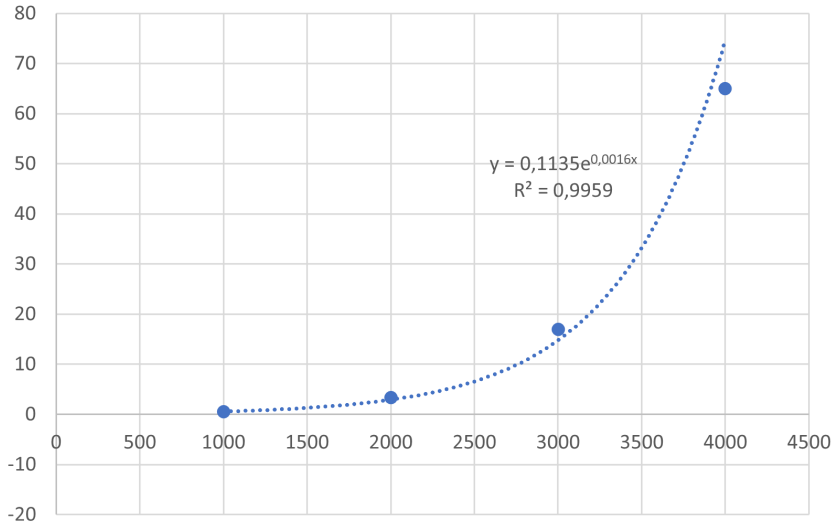


FIGURE 4.5 – Régression exponentielle du temps de calcul par rapport au nombre de points

4.4.2 Quelques solutions

Les temps de calcul trop importants et la place prise sur la mémoire vive font du Krigeage un outil particulier à utiliser. Cependant, plusieurs travaux de recherche ont essayé d'améliorer ces limites, tout en gardant les bonnes performances du modèle. Quelques solutions sont présentées ici avec des sources pour le lecteur voulant les approfondir.

La solution la plus simple serait de faire calculer le modèle sur un ordinateur plus puissant, voire une machine virtuelle. Avec ceci, les bonnes performances de prédiction du Krigeage seraient gardées. L'ordinateur sur lequel ont été réalisés les modèles dans ce mémoire n'était clairement pas assez puissant pour modéliser sur plus de 4000 points. Cependant, en plus de nécessiter un outil informatique plus puissant, le problème n'est pas réglé en soi.

Une deuxième solution serait d'utiliser des **matrices creuses** comme dans [Furrer, 2009]. Il s'agit d'une matrice contenant une part importante de zéros, réduisant ainsi la taille de la matrice informatiquement. Les algorithmes doivent être réadaptés, mais le gain de temps est présent. Le package *spam* sur R permet de travailler avec les matrices creuses et le package *fields* permet de réaliser des modèles de Krigeage sur des base de données plus importantes.

Une troisième solution est d'utiliser des champs aléatoires gaussiens de Markov (**GMRF**). Au lieu d'estimer une valeur avec un modèle de Krigeage sur une base de données, on estime la base de données avec un GMRF, pour ensuite récupérer la prédiction exacte par Krigeage. Cette méthode est développée dans [Hartman et Hössjer, 2008].

4.4.3 Diviser pour mieux régner

Une autre solution a été ici développée. La méthode est de diviser la base de train et d'entraîner un modèle de Krigeage sur chacune des sous-divisions. Le temps de calcul évoluant exponentiellement, il sera donc plus rapide de traiter 4 sous-parties d'une base plutôt que la base entière.

A ce principe, il est important d'ajouter que les communes les plus exposées sont les plus intéressantes pour entraîner le modèle. De ce fait, la méthode de division prend les 1000 communes les plus exposées et associe, à chaque itération, 1000 autres communes prisent aléatoirement. Enfin, la valeur du résidu prédit pour une commune est la moyenne des prédictions faites par le modèle à chaque itération. Le temps de calcul est ainsi énormément réduit, comme le montre le schéma sur la figure **4.6**.

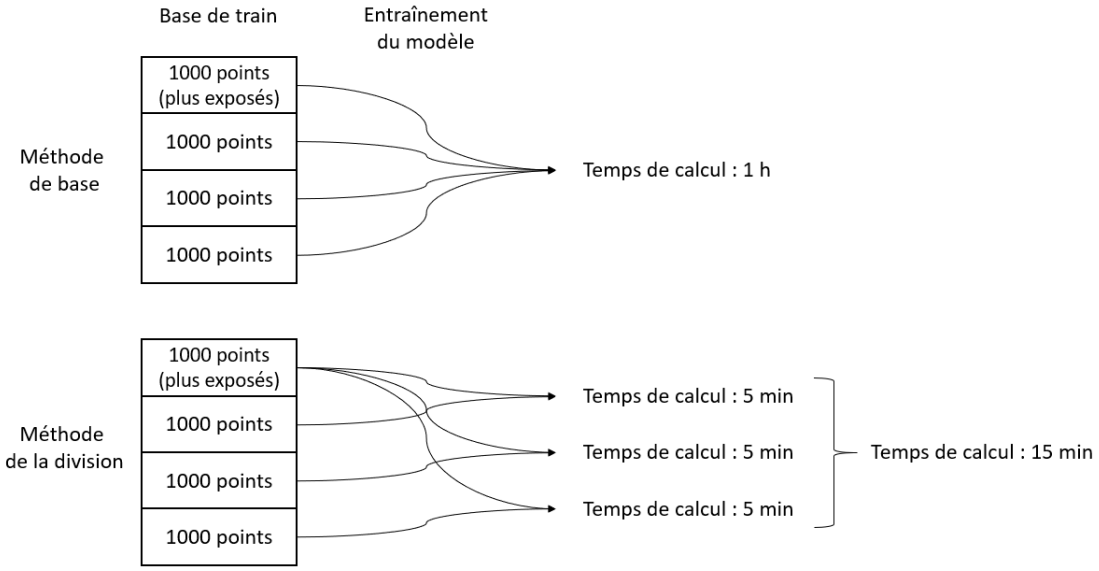


FIGURE 4.6 – Schéma de la différence entre la méthode de base et la méthode de division

Le modèle global perd peu en qualité suite à cette méthode. En effet, pour une base de train de 4000 points, les Q^2 sont presque équivalents : 12,71% pour la méthode de base et 12,08% pour la méthode de division. De plus, un R^2_{tot} a pu être calculé avec cette nouvelle méthode.

Pour le calcul de celui-ci, les 1000 communes les plus exposées ont été encore sorties de la base pour être utilisée avec le reste des sous-divisions. Ces sous-divisions sont constituées de 2000 communes chacune, prises aléatoirement. Cela permet d'avoir une bonne répartition des communes sur la France, pour capter le plus d'informations possible.

Cependant, la limite de cette méthode est le fait que toute la base n'est pas prise en compte. Ainsi, le résultat est moins précis, au vu des métriques utilisées, avec cette

méthode qu'avec la méthode du Krigeage classique. S'il était possible d'utiliser toute la base pour entraîner le modèle, le modèle du Krigeage serait le plus performant au vu des métriques utilisées. C'est pour cela qu'avec cette méthode on obtient seulement un R^2_{tot} de 15,85%. Pour améliorer la méthode, un coefficient de crédibilité pourrait être utilisé pour mettre davantage de poids sur le modèle qui utilise la commune dans la base de train.

Ce n'est pas pour autant un mauvais lissage. À regarder la carte figure **4.7**, le lissage semble même plutôt juste. Lorsque le zonier finalement réalisé, visible figure **4.8**, est incorporé dans le GLM, les résultats sont même plutôt bons. La métrique d'erreur est alors le problème.

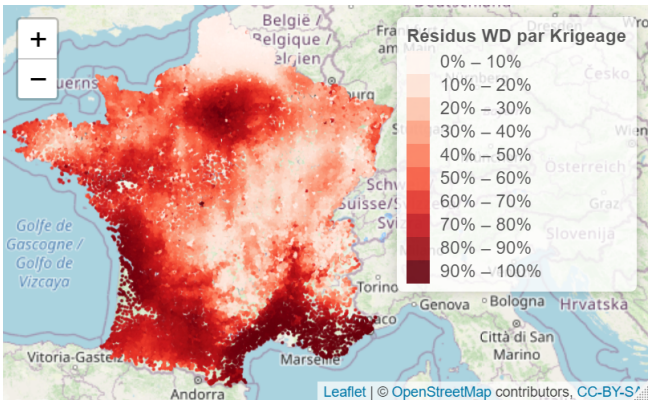


FIGURE 4.7 – Carte des résidus en fréquence de la garantie dégâts des Eaux prédits par le modèle du Krigeage, avec la méthode de division

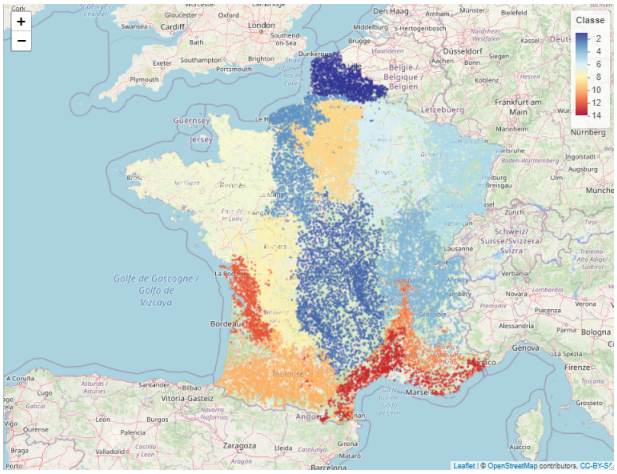


FIGURE 4.8 – Zonier de la garantie dégâts des Eaux prédits par le modèle du Krigeage, avec la méthode de division

En effet, les erreurs de prédiction d'un lissage sont "normales", puisque son but est de lisser les résidus et donc de ne pas prédire exactement les résidus. Le tout est de ne pas avoir une erreur vraiment importante, où le lissage deviendrait inutile. L'apprentissage sur une petite base de train est important, puisqu'il montre la qualité de prédiction du modèle. Sur ce point, le Krigeage est le meilleur modèle testé ici.

4.5 Application du zonier au GLM

Comme pour les zoniers précédemment faits, le zonier par Krigeage est inutile tant qu'il n'est pas appliqué à un GLM. C'est ce qui est réalisé dans la section **4.5.1**. La section **4.5.2** propose de comparer le GLM obtenu avec l'ajout du zonier par Krigeage avec les autres GLM. Cela permettra de conclure sur la pertinence du Krigeage dans la réalisation d'un zonier.

4.5.1 Application du zonier par Krigeage au GLM

Les mêmes étapes préliminaires que pour l'intégration des zoniers par crédibilité et par XGBoost sont utilisées pour le zonier par Krigeage :

— Une complétion par plus proche voisin pour les communes ne disposant pas de zone.

— Un regroupement des zones de faible exposition

— Une jointure entre la base de données interne et le zonier finalisé.

La figure 4.9 montre le zonier final, utilisable pour le GLM. Le nombre de zones n'a pas été touché puisqu'il y a assez d'exposition dans chacune d'elle.

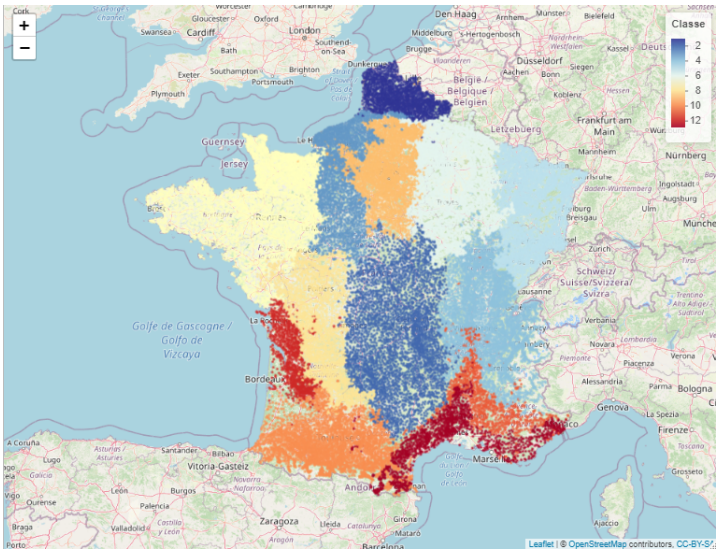


FIGURE 4.9 – Zonier final de la garantie dégâts des Eaux prédits par le modèle du Krigeage, avec la méthode de division

4.5.2 Comparaison des modèles

Les métriques importantes sont de nouveau calculées. La Déviance, l'AIC, le coefficient de Gini se trouvent dans le tableau **4.3**. On remarque que le GLM avec le zonier par Krigeage possède une meilleure Déviance, un meilleur AIC et un meilleur coefficient de Gini que les autres GLM.

GLM	Déviance	AIC	Gini
GLM sans zonier	648714	793702	0,1576
GLM avec zonier crédibilité	638357	783361	0,2729
GLM avec zonier XGBoost	638371	783385	0,2595
GLM avec zonier Krigeage	637757	782768	0,2764

TABLE 4.3 – Tableau des métriques des GLM (avec le zonier par Krigeage)

La courbe de Bêtas du zonier par Krigeage, figure **4.10** est croissante dans l'ensemble, ce qui est correct pour un zonier. Les courbes de Bêtas confirment toutes que les zones ont été choisies correctement.

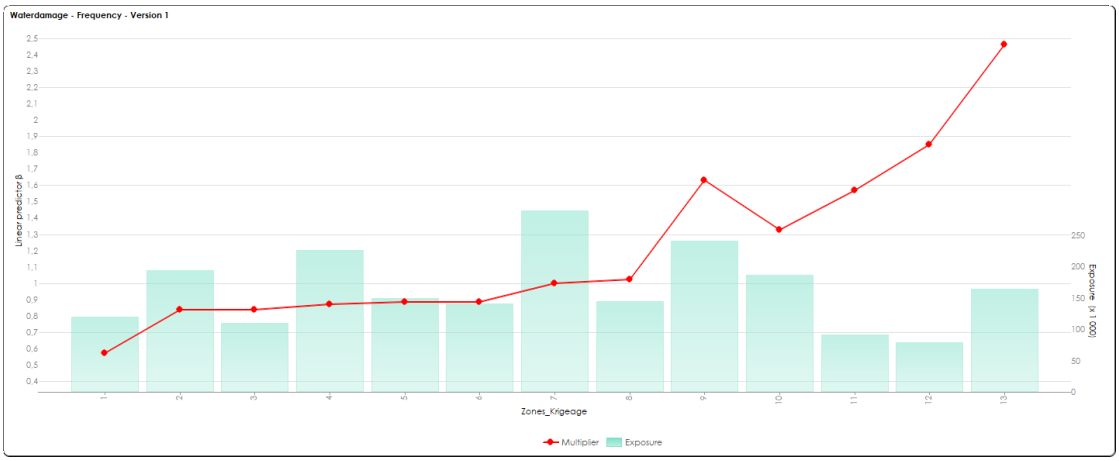


FIGURE 4.10 – Courbe de Bêtas pour le zonier par crédibilité

Enfin, la figure **4.11** montre la cartographie des résidus issus du GLM avec le zonier par Krigeage. La cartographie ressemble à celles pour les résidus issus du GLM avec le zonier par crédibilité et par XGBoost. Dans l'ensemble, on peut conclure de ces cartographies que les zoniers ont servi pour la modélisation.

La conclusion de cette comparaison est que pour ce portefeuille, le zonier le plus performant est celui par Krigeage. En effet, il possède de meilleures métriques numériques (Déviance, AIC, coefficient de Gini). Cependant, l'écart est faible et les autres zoniers sont aussi utilisables. De plus, pour un autre portefeuille, une autre garantie, une autre répartition géographique du portefeuille, la conclusion aurait peut-être été différente.

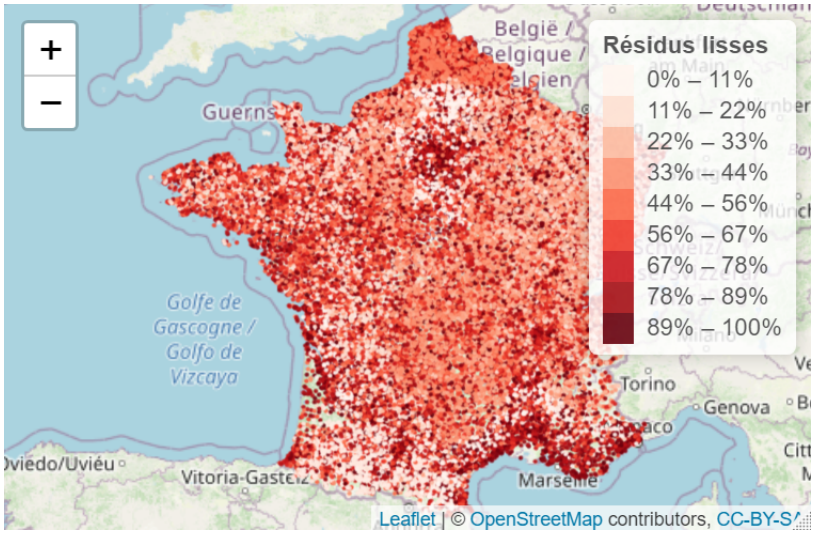


FIGURE 4.11 – Cartographie des résidus en fréquence pour la garantie Dégâts des Eaux issus du GLM avec le zonier par Krigeage

4.6 Conclusion de l'application du Krigeage au zonier

Le but de ce mémoire est d'apporter une nouvelle méthode de lissage de résidus dans la création d'un zonier, puis de comparer cette méthode avec deux autres méthodes classiques. Le Krigeage, modèle alliant lissage et données externes, semble être une bonne idée pour concurrencer les autres modèles. Le modèle de Krigeage dispose d'une prédiction très supérieure aux autres modèles, mais d'un temps de calcul très long. Ses autres avantages et limites techniques sont précisés dans la section **4.6.1**. D'un point de vue assurantiel, il y a ici une opportunité à saisir, mais qui n'est pas forcément adaptée pour toutes les situations. Ses autres avantages et limites assurantiels sont précisés dans la section **4.6.2**.

4.6.1 Avantages et limites techniques

Le principal atout du Krigeage est sa très bonne prédiction des modèles. Les métriques d'erreur vont dans ce sens avec un Q^2 de 12,71% sur 4000 points dans la base de train, comparé aux 5,87% et 7,63% obtenus par crédibilité et XGBoost. Le Krigeage permet donc de mieux modéliser le résidu issu du GLM afin d'en sortir une meilleure information géographique.

Cela est dû à son deuxième atout : le mélange des méthodes. En effet, avec la théorie de la crédibilité, on a un lissage, mais celui-ci manque de précision par manque de données externes. Avec le Machine Learning, les données externes sont utilisées, mais le résultat n'est pas lissé. Avec le Krigeage, le lissage et les données externes sont utilisés afin d'obtenir un résultat optimal.

Cependant, ces atouts ont un coût technique : le temps de calcul. Évoluant exponentiellement avec le nombre de communes dans la base d'entraînement, il devient difficile d'entraîner le Krigeage sur de grandes bases. Il est nécessaire de travailler sur quelques points et de prédire sur les autres.

Pour résoudre ce problème, la méthode de la division a été crée. Elle se base sur le principe que si le temps de calcul évolue exponentiellement, il est intéressant de diviser la base et de réaliser un calcul sur toutes les sous-divisions. Après plusieurs optimisations, notamment sur la prise en compte de l'exposition, le résultat est convaincant. En effet, celui-ci est plus rapide et subit moins de sur-apprentissage, puisqu'il résulte d'une moyenne de plusieurs prédictions.

D'autres atouts peuvent être cités, comme le nombre d'opportunités d'optimisation, comme le travail réalisé sur le bruit autour des variables ou le travail réalisé sur la méthode de la division. D'autres limites peuvent être citées comme la grande place que prend le modèle du Krigeage sur la mémoire, à cause des importantes matrices de corrélation utilisée.

4.6.2 Avantages et limites assurantiels

Ces avantages et ces limites techniques ont un effet sur les avantages et les limites d'un point de vue assurantiel. L'application du zonier au GLM permet d'appréhender les atouts et les contraintes plus pratiques.

Pour le portefeuille utilisé pour ce mémoire, le zonier par Krigeage semble meilleur : Meilleure Déviance, meilleur AIC et meilleur coefficient de Gini. D'un point de vue assurantiel, un zonier plus performant signifie une meilleure segmentation et donc une meilleure rentabilité. De plus, avec les données externes, le Krigeage apporte une information supplémentaire pour expliquer la sinistralité (les zoniers par Machine Learning le font également).

Bien que le zonier soit meilleur, le temps mis pour le réaliser est beaucoup plus élevé par rapport aux autres méthodes. La contrainte n'est pas que technique, puisque le temps passé à réaliser un zonier est un temps passé à ne pas passer sur un autre sujet. Pour un assureur, cette problématique doit se poser lorsque le Krigeage est utilisé. Surtout pour un assureur possédant des contrats dans beaucoup de communes différentes, son temps de calcul ne sera que beaucoup plus élevé s'il souhaite réaliser un lissage par Krigeage sur sa base entière.

Pour diminuer ce problème, la méthode de la division a été crée pour réduire ce temps de calcul. De plus, cette méthode lisse beaucoup plus que le Krigeage sans cette méthode. Un gain de temps précieux est alors obtenable pour l'assureur. Mais ce gain de temps reste minime par rapport aux temps de calcul des autres modèles.

Chapitre 5

Conclusion

La nécessité de segmenter toujours plus finement rend le zonier incontournable, comme expliqué dans la section **5.1**. En effet, celui-ci permet d'obtenir des informations sur le risque géographique qui influe beaucoup sur la sinistralité. Deux méthodologies de création de zonier sont présentées dans ce mémoire : le lissage des résidus par crédibilité et la prédiction des résidus par XGBoost. Le Krigeage, méthode d'interpolation statistique, représente une sorte de mélange des deux méthodologies.

Le but de ce mémoire est d'examiner l'apport du Krigeage dans la méthodologie du zonier. Pour ce faire, il a été comparé à la théorie de la crédibilité et à la prédiction par XGBoost, deux méthodes classiques de lissage des résidus dans le zonier. Ce mémoire a mis en lumière deux choses : le Krigeage est un excellent modèle prédictif, mais les temps de calcul sont vraiment longs. La section **5.2** rappelle ces faits avec d'autres atouts et limites.

La section **5.3** dresse les limites de l'étude. En effet, celle-ci n'est pas à généraliser, elle montre simplement une nouvelle méthode de lissage de résidus pour un zonier, avec de bons résultats à la fin.

Pour ces limites, des ouvertures sont proposées dans la section **5.4**. Elles sont pistes pour continuer le sujet et pour inviter le lecteur à réaliser d'autres zoniers par Krigeage dans d'autres situations.

5.1 Le zonier : une variable incontournable de la tarification

La concurrence toujours plus forte pousse les assureurs à trouver de nouveaux moyens de prédire au mieux les futurs sinistres de leurs assurés. Le zonier, c'est-à-dire le regroupement de toute l'information géographique en une variable, permet une segmentation plus fine, grâce à l'apport de données externes. Avoir un meilleur zonier est donc synonyme de meilleure prédiction des sinistres et donc un meilleur chiffre d'affaires.

Pour faire un zonier, ce mémoire revoit deux méthodes classiques de lissage de résidus. La première lisse les résidus en utilisant la théorie de la crédibilité. La seconde, plus populaire car plus performante, prédit les résidus par un algorithme de Machine Learning

(ici un XGBoost). Ces deux méthodes classiques sont comparées à une troisième : le lissage des résidus par Krigeage. Cette comparaison a pour but de mettre en lumière les atouts et les limites du Krigeage dans son application pour un zonier.

5.2 Le Krigeage : un excellent modèle prédictif, mais difficile à mettre en pratique

Lorsque le modèle du Krigeage apprend sur la base de train de 4000 communes, ses performances sont supérieures aux autres modèles présentés. Le Q^2, la principale métrique d'erreur utilisée, est de 12,7 %, comparé à 7,6 % pour l'XGBoost et 5,9 % pour la crédibilité. Cette différence est issue de la structure même du Krigeage d'une part et d'autre part l'optimisation de celui-ci, décrite lors de ce mémoire.

Ces atouts se sont révélés précieux pour réaliser un zonier en MRH. Une fois importé dans le GLM, celui-ci obtient de meilleures métriques (Déviance, AIC, coefficient de Gini) que les autres GLM avec les autres zoniers classiques.

Cependant, le temps de calcul pour obtenir ce zonier est très long comparé aux autres méthodes de création de zoniers. Celui-ci évolue exponentiellement avec le nombre de communes à apprendre. Pour 4000 communes, un modèle calcule pendant une heure. La totalité de la base est alors impossible à utiliser comme base d'entraînement.

Pour outrepasser ce problème, la méthode de la division a été créée. Celle-ci permet de réduire le temps de calcul en calculant sur des sous-divisions de la base et d'éviter un maximum le sur-apprentissage en réalisant une moyenne sur l'ensemble des prédictions des sous-divisions. Le temps de calcul est diminué, mais reste encore long.

5.3 Les axes d'amélioration du mémoire

D'un point de vue assurantiel, le ratio entre temps de calcul supplémentaire et gain obtenu avec le zonier s'étudie. Celui-ci est important en pratique, mais n'est pas quantifiable par une étude comme celle-ci.

La robustesse du modèle est également à prouver. En effet, en théorie, le sur-apprentissage est évité. Mais en pratique, il faudrait vérifier que pour l'année suivante, le zonier reste cohérent.

De même, le portefeuille utilisé possède des spécificités comme des zones assez accentuées dans le Nord, l'Île-de-France, etc. Le but est de retrouver ces zones, mais également de regarder leur impact sur les communes alentours.

5.4 Ouvertures sur les axes d'amélioration

Un portefeuille plus réaliste permettrait d'obtenir des résultats plus concrets. Aussi, une autre garantie, une autre répartition en France, un nombre plus faible ou plus élevé

d'individus seraient intéressants pour voir quelle méthode est la meilleure. Une vision sur plusieurs années pourrait être utile également.

Pour aller plus loin, certains détails pourraient être développés. Par exemple, le Krigeage universel, développé dans la section **4.2.1**, est plus général et apporterait peut-être plus d'informations au modèle que le Krigeage ordinaire utilisé. Pour la méthode de division, un facteur de crédibilité pourrait être ajouté pour permettre au modèle apprenant sur la commune d'avoir plus de poids dans la moyenne.

Bibliographie

[Anderson, 2007] Anderson, Feldblum, M. (2007). *A Practitioner's Guide to Generalized Linear Models*. Towers Watson.

[Candillier, 2016] Candillier, L. (2016). *Contextualisation, visualisation et évaluation en apprentissage supervisé*. Thèse à l'université de Lille 3.

[Chamoulaud, 2020] Chamoulaud, F.-X. (2020). Réalisation d'un véhiculier à l'aide d'outils de machine learning. *Mémoire d'actuariat*.

[Chen et Guestrin, 2016] Chen, T. et Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[Deville, 2012] Deville, O. R. . D. G. . Y. (2012). Dicekriging, diceoptim : Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*.

[Furrer, 2009] Furrer, R. (2009). Sparse matrices methods and kriging : Applications to large spatial data sets. *University of Zurich*.

[Hartman et Hössjer, 2008] Hartman, L. et Hössjer, O. (2008). Fast kriging of large data sets with gaussian markov random fields. *Elsevier*.

[Lustman, 2021] Lustman, F. (2021). Les assureurs, acteurs de la relance durable. Conférence de presse de la FFA.

[Pariente, 2017] Pariente, J. (2017). Modélisation du risque géographique en assurance habitation. *Mémoire d'actuariat*.

[Picabea, 2019] Picabea, R. (2019). Construction d'un métamodèle d'efficience de réassurance par différentes méthodes d'interpolation spatiale. *Mémoire d'actuariat*.

[Zurfluh, 2019] Zurfluh, E. (2019). Utilisation du machine learning dans l'estimation du ratio de solvabilité d'un assureur vie et application aux reverses stress tests. *Mémoire d'actuariat*.