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

Par : Monsieur Aurélien ROUSSEAU

Titre du mémoire :

Construction d'un zonier en assurance grêle sur cultures

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

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

Membres présents du jury de la filière : Signature :	Entreprise : Nom : Pacifica Signature : 
	Directeur de mémoire en entreprise Nom : Rogier Claire Signature : 
Membres présents du jury de l'Institut des Actuaires : Signature :	Invité : Nom : Signature :
	Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité) Signature du responsable entreprise :  Signature du candidat : 

Résumé

Une tempête de grêle est un phénomène climatique pouvant causer d'importants dégâts pour de nombreux acteurs. C'est le cas en particulier des agriculteurs qui ont leurs cultures très exposées à ce risque. Le 20 juin 2022, la Dordogne a par exemple été touchée par une tempête de grêle d'une grande violence. Celle-ci a occasionné des pertes pouvant se compter en centaines de milliers d'euros pour certaines exploitations. Néanmoins, une partie des agriculteurs n'avaient pas assuré leurs récoltes. Malgré le besoin de plus en plus croissant de se couvrir contre le risque climatique, le taux de pénétration de l'assurance climatique reste faible. C'est pourquoi commercialiser un produit Grêle adapté au risque de chaque agriculteur est un enjeu majeur.

La grêle a un caractère très hétérogène, que ce soit de par son intensité ou sa localisation. Dans cet objectif de mettre en place un produit qui s'adapte au mieux à l'exposition de chaque agriculteur au risque grêle, la problématique de ce mémoire est la construction d'un zonier en assurance grêle sur cultures.

La première étape de notre étude consiste à mieux comprendre le risque grêle sur le territoire, à déterminer des variables explicatives du phénomène ainsi que de quantifier leur relation au taux de prime pure observé sur le produit Grêle. Ensuite, grâce à ce premier travail, nous avons cherché à segmenter finement le risque grêle sur le territoire. Enfin, l'utilisation d'outils de classification nous permettra de mutualiser ce risque caractérisé par une part aléatoire, tout en conservant une partie de la segmentation obtenue.

Mots clés : assurance climatique, grêle, zonier, taux de prime pure, classification, modèle linéaire généralisé, résidus, krigeage, stabilité, segmentation.

Abstract

Hail can cause serious damage to various parties, especially farmers, whose crops are very sensitive to this hazard. A recent example occurred on June 20, 2022, when severe hail struck the Dordogne, causing significant economic losses to a number of farms. Unfortunately, some farmers have not insured their crops. Indeed, despite the growing need to protect themselves against climate risks, the share of insured farmers remains low. As such, creating a Hail product that takes into account the risks of each farmer becomes a challenging task.

Hail is a complex weather phenomenon that varies in intensity and location. The purpose of this thesis is to build a hail insurance zone system for crops that best adapts to each farmer's hail risk level.

To achieve this, the initial phase of the research is to understand the threat of hail on French territory, find out the factors affecting the phenomenon and calculate the net premium rate of the Hail product. Next, using this data, the study aims to accurately segment the risk of hail in the region. Finally, the use of classification tools will allow us to pool this risk characterized by a random component while preserving the achieved segmentation.

Keywords : climate insurance, hail, zoning, net premium rate, classification, general linear model, residuals, kriging, stability, segmentation.

Note de synthèse

Contexte

La grêle constitue un aléa climatique pouvant occasionner d'importants dégâts auprès de plusieurs types d'acteurs de la société. Un acteur spécifique va être sensible aux tempêtes de grêle : l'agriculteur. Il est exposé via ses locaux et ses véhicules, mais le risque réside surtout dans la fragilité de ses cultures face à la grêle. Ce mémoire s'inscrit dans le cadre de l'assurance climatique au sein du secteur agricole.

La grêle est un phénomène dont la fréquence et l'intensité varient selon les endroits. Cette différence d'exposition est visible à l'échelle d'un département, où les communes au sein de celui-ci peuvent être sinistrées de manière très différente sur l'historique. Dans ce mémoire, nous allons essayer d'établir un zonier qui permet de capter, au mieux, ces différences d'exposition au risque grêle selon la localisation de la récolte.

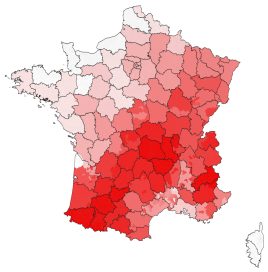


FIGURE 1 – Zonier actuel utilisé par Pacifica

Le zonier actuel, représenté en Figure 1, est majoritairement à la maille département. Quelques zones d'exception existent, mais elles sont rares et représentent une faible partie des capitaux assurés. Un objectif de ce mémoire sera d'effectuer un zonier à une maille plus fine. Cette optimisation permettra une meilleure segmentation du risque et une tarification plus adaptée à l'ensemble des communes du territoire.

Le zonier n'a pas pour seul objectif de segmenter au maximum le risque entre les assurés. Il est le fruit d'un arbitrage entre finesse du maillage géographique et mutualisation, correspondant à un dilemme courant du statisticien : l'arbitrage biais et variance. Un des enjeux principaux dans la construction du zonier sera d'éviter le surapprentissage, c'est-à-dire éviter de construire un zonier incapable de s'adapter à de nouveaux scénarios de sinistralité.

Données

Pour construire le zonier, nous disposons de différentes bases de données pouvant améliorer notre connaissance du risque grêle par commune. Nous constituons donc un ensemble de données le plus exhaustif possible (première étape de la Figure 2). Le Tableau 1 récapitule toutes ces sources de données utilisées :

Type de base de données	**Description**	**Exemples de variables** (par commune)
Données internes	L'historique des différents produits Pacifica donne des informations sur le risque grêle. Cette information peut être incomplète.	— Fréquence des sinistres Multirisques Climatiques — Fréquence des sinistres Habitation — ...
Données externes disponibles	Des bases externes à Pacifica, disponibles gratuitement, peuvent nous apporter des informations sur chaque commune. Les variables présentes dans ces bases sont potentiellement corrélées au risque grêle.	— Altitude — Pluviométrie — Températures Moy/Min/Max — Typologie des sols — ...
Données externes privées	Cette base de données recense les tempêtes de grêle. Elle peut être pertinente pour compléter l'information des données de sinistralité grêle.	— Nombre de tempêtes — Diamètre grêlons Moy/Max — Type de grêlons

TABLE 1 – Récapitulatif des données à disposition

Nous analyserons dans ce mémoire si l'achat d'une base de données recensant les tempêtes de grêle est pertinent. Les données météorologiques, figurant dans la base de données externes disponible gratuitement, ne sont pas connues pour l'ensemble des communes, seulement pour celles où figurent une station. Nous utilisons alors le krigeage afin de propager à l'ensemble du territoire les informations de pluviométrie ou de températures contenues dans chaque station.

Modélisation

Pour construire notre zonier, nous effectuons plusieurs étapes afin de traiter, le mieux possible, l'information contenue dans les bases de données. Ce processus est représenté en Figure 2 :

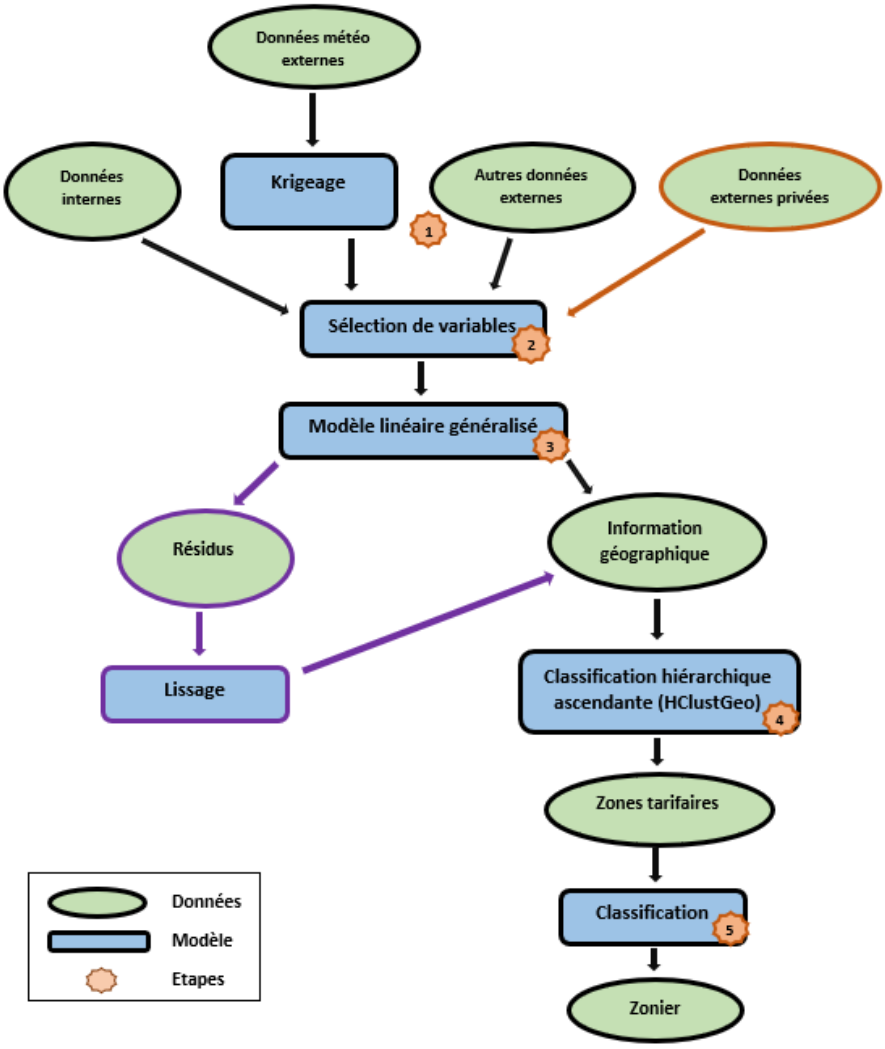
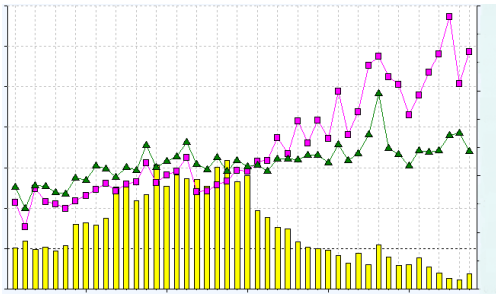


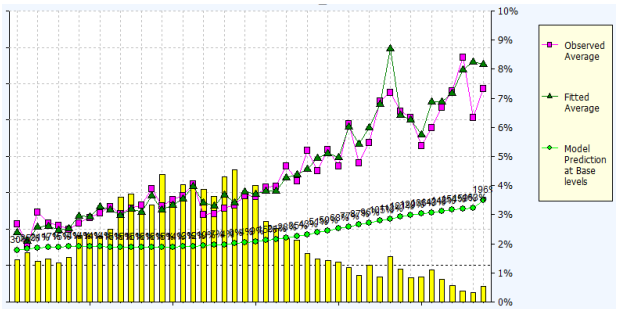
FIGURE 2 – Étapes de construction des zoniers

La première étape a été présentée en même temps que les données. La deuxième étape du processus de construction correspond à la sélection des variables explicatives pertinentes qui figureront dans le modèle d'ajustement. En Figure 3 est représentée une variable retenue, l'*Altitude*. Des variations du taux de prime pure sont captées grâce à cette dernière. Nous retenons, parmi les variables contenant une information géographique : l'*Altitude*, l'*Amplitude de températures moyennes mensuelles d'avril à septembre* , la *Fréquence des sinistres Habitation* et la *Fréquence des sinistres Multirisques Climatiques*.

La troisième étape utilise les modèles linéaires généralisés dans l'objectif d'ajuster le taux de prime pure. À partir de cette étape, nous sommes en capacité de calculer un zonier brut sur l'ensemble du territoire. Le coefficient tarifaire d'une commune correspond au produit des coefficients calculés pour les quatre variables géographiques.



(a) Variable non utilisée dans le modèle



(b) Variable utilisée dans le modèle

FIGURE 3 – Ajustement de la variable *Altitude*

La quatrième étape correspond à l'implémentation d'un modèle de classification hiérarchique ascendante. Nous utilisons le package HClustGeo du logiciel R. Cette fonction regroupe des communes selon leur distance géographique et leur distance en termes de coefficients. Deux paramètres sont utilisés pour générer des zones tarifaires :

— k, le nombre de zones à créer

— α, le poids accordé à la distance géographique par rapport à celle des coefficients.

La cinquième étape correspond au regroupement des zones tarifaires en classes de risque. Nous regroupons les zones en classes de risque afin de palier au problème de zones faiblement représentées en capitaux assurés et ainsi améliorer la stabilité du zonier.

Deux étapes intermédiaires dans le processus de construction du zonier sont également testées. Premièrement, l'ajout de l'ensemble des variables explicatives issues de la base de données grêle privées. Le second ajout est l'utilisation de l'information géographique contenue dans les résidus. L'ensemble des variables explicatives n'est pas exhaustif, tout comme notre connaissance de la grêle sur le territoire, il est donc possible que les résidus contiennent une information pertinente supplémentaire. Pour retirer le bruit présent dans ceux-ci, nous allons les lisser géographiquement.

Au final, nous avons construit cinq zoniers différents :

— Un premier à partir de l'équation tarifaire en vigueur à Pacifica

— Un deuxième reclassant les zones du premier à partir de l'historique des sinistres

— Un troisième à partir des données complémentaires gratuites disponibles

— Un quatrième reprenant le troisième zonier et utilisant l'information des résidus

— Un cinquième ayant pour socle le troisième et qui utilise des données grêles privées.

Comparaison des modèles

Dans l'objectif de comparer ces zoniers, nous avons utilisé différents indicateurs. Afin de juger de la capacité de segmentation des zoniers, nous avons notamment regardé le rapport des coefficients extrêmes ainsi que la dispersion inter-zones. Afin de juger de l'homogénéité des classes de chacun des zoniers, nous utilisons l'écart-type intra-classe. Enfin, afin de juger de la stabilité, nous utilisons la validation croisée à travers un ensemble de test et un ensemble d'entraînement. Pour cette dernière comparaison, nous calculons la différence entre les coefficients calculés sur un modèle d'entraînement et ceux d'un modèle de test. Ces indicateurs et leurs résultats sont présentés, pour chaque zonier, dans le Tableau 2 :

Type de zonier	**Rapport des coefficients extrêmes**	**Distance Inter-Classes**	**Écart-type Intra-Classe**	**Différence Test/Train**
Actuel utilisé par Pacifica	8,4	48,6%	39,8%	21,34%
Actuel optimisé	19,6	52,4%	37,1%	16,10%
Utilisant les données complémentaires disponibles	28,1	52,1%	31,4%	14,24%
Utilisant les résidus	29,1	56,4%	31,5%	14,13%
Utilisant les données grêles privées	27,2	61,1%	31,1%	13,26%

TABLE 2 – Comparaison statistiques des zoniers construits

Les zoniers construits à partir des données complémentaires offrent des meilleurs résultats pour l'ensemble des indicateurs. Ces trois zoniers segmentent mieux le risque et ont une meilleure stabilité. Le zonier utilisant les données complémentaires disponibles ainsi que la base de données grêle privée donne la meilleure dispersion inter-classes tout en ayant une stabilité contrôlée par rapport aux autres zoniers.

L'utilisation de la base de données grêle payante offre donc bien les meilleurs résultats. Cependant, son achat ne permet pas de créer un zonier aux performances significativement améliorées par rapport à celles du zonier utilisant seulement les données complémentaires. Son utilisation ne semble pas indispensable pour construire un zonier qui satisfasse nos contraintes et nos objectifs.

Executive summary

Context

Hail is a weather phenomenon that poses a significant threat to various members of society. Farmers in particular are at risk of hail damage to their crops, property and vehicles. This work is part of the climate insurance framework for the agricultural sector.

Hail frequency and intensity vary by location. This difference in exposure is visible at the scale of a department, where the municipalities within it can be damaged differently over time. The purpose of this work is to create a zoning system that takes into account these different exposures to hail risks based on growing sites.

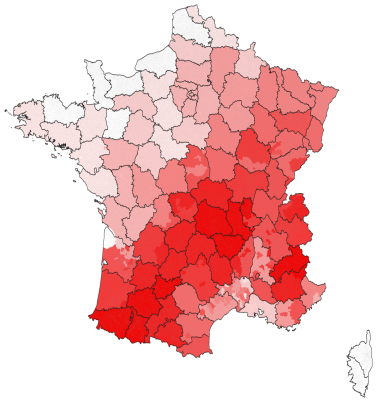


FIGURE 4 – Zoning currently used by Pacifica

Currently, the zonal system shown in 4 is primarily at the departmental level, with some rare exceptions representing a small portion of the insured capital. This paper attempts to optimize the system by creating a finer mesh to improve risk segmentation and pricing for all municipalities.

The zoning system aims to balance geographic sophistication and commonality with a focus on avoiding overfitting. This creates the most comprehensive dataset possible (first step in Figure 5). Table 3 summarizes all data sources used.

Data

To build the zoning, we have different databases that can improve our knowledge of hail risk by municipality. We therefore constitute a set of data as exhaustive as possible (first step in Figure 5). The Table 3 summarizes all these data sources used :

Type of database	**Description**	**Examples of variables** (per municipality)
Internal data	The history of various Pacifica products provides information on hail risk. This information may be incomplete.	— Frequency of Climate Multi-risk Claims — Frequency of Home Insurance Claims — ...
External data available	External databases available for free can provide us with information on each municipality. The variables present in these databases are potentially correlated with the hail risk.	— Altitude — Pluviometry — Mean/Min/Max temperatures — Soil typology — ...
Private external data	This database lists hailstorms and can be relevant to supplement information on hail claims data.	— Number of hailstorms — Mean/Max Hailstone Diameter — Type of hailstones

TABLE 3 – Summary of available data

We also analyze whether obtaining a hail database is necessary, as weather data is only available in communities with weather stations. This data is used in conjunction with kriging to extend precipitation and temperature information to surrounding areas. The feasibility and effectiveness of this approach will be assessed in this work.

Modeling

To build our zoning model, we undertake several steps to process the information contained in the databases as efficiently as possible. This process is illustrated in Figure 5.

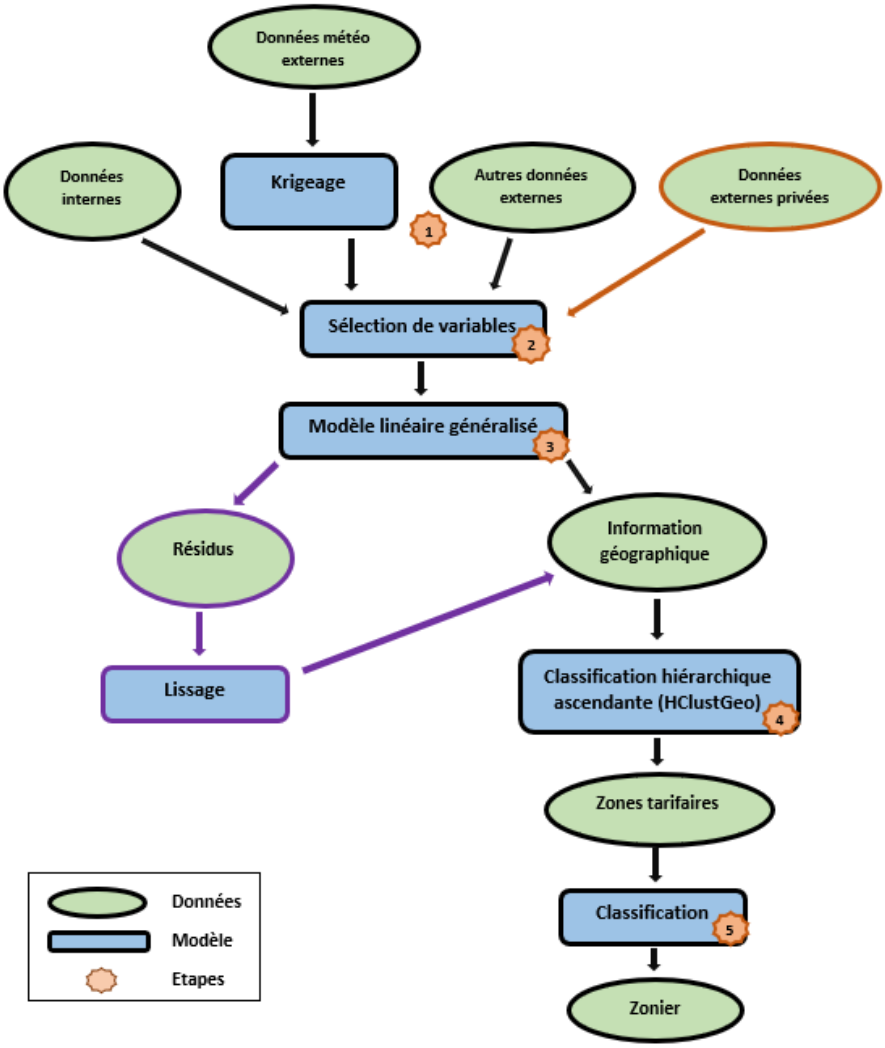
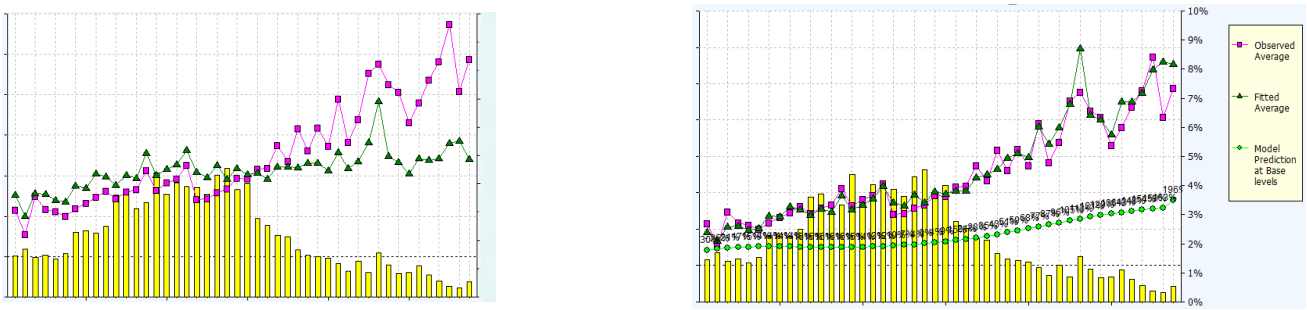


FIGURE 5 – zoning construction stage

The first step was presented simultaneously with the data. The second step in the construction process involves selecting relevant explanatory variables that will be included in the fitting model. Figure 6 represents a selected variable, the *Altitude*. This variable captures variations of the net premium rate. Ultimately, we choose, among those containing geographic information : the *Altitude*, the *Monthly Mean Temperature Amplitude from April to September*, the *Home Insurance Claims Frequency*, and the *Catastrophic Multi-Risk Claims Frequency*.

The third step uses generalized linear models to adjust the pure premium rate. From this step, we are able to calculate a zoning variable over the entire territory. The tariff coefficient for a municipality corresponds to the product of the coefficients calculated for the four geographic variables.



(a) Variable not used in the model

(b) Variable used in the model

FIGURE 6 – Adjusting of the *Altitude* variable

The fourth step involves implementing an ascending hierarchical clustering model. We use the HClustGeo package in the R software. This function groups municipalities according to their geographic distance and their distance in terms of coefficients. Two parameters are used to generate tariff zones :

— k, the number of zones to create

— α, the weight given to geographic distance compared to that of coefficients.

The fifth step involves grouping tariff zones into risk classes. To improve the stability of the zoning model, we group these zones into risk classes to overcome the problem of zones with low insured values.

Two intermediate steps in the zoning model construction process are also tested. First, adding all the explanatory variables from the private hail database. The second addition is using the geographic information contained in the residuals. The set of explanatory variables is not exhaustive, just like our knowledge of hail in the territory, so it is possible that the residuals contain additional relevant information. To remove the noise present in these residuals, we will smooth them geographically.

In the end, we have constructed five different zoning models :

— A first one using Pacifica's current tariff equation

— A second one reclassifying the zones from the first one based on claims history

— A third one using available complementary free data

— A fourth one that uses the information contained in the residuals from the third model

— A fifth one that incorporates the third model and uses private hail data.

Comparison of models

Various metrics were used to compare these zoning models. The segmentation ability of the zoning model was evaluated based on extreme coefficient ratios and variability between zones. The within-class standard deviation is used to determine class homogeneity for each zoning model. Finally, we assessed stability by cross-validation using both training and test sets. This last comparison calculated the difference between the training model coefficients and the test model coefficients. The results for each zone model and their respective metrics are shown in Table 4 :

Zoning Type	**Extreme Coefficient Ratio**	**Inter-Class Distance**	**Intra-Class Standard Deviation**	**Test/Train Difference**
Currently used by Pacifica	8.4	48.6%	39.8%	21.34%
Currently optimized	19.6	52.4%	37.1%	16.10%
Using available complementary data	28.1	52.1%	31.4%	14.24%
Using residuals	29.1	56.4%	31.5%	14.13%
Using private hail data	27.2	61.1%	31.1%	13.26%

TABLE 4 – Statistical comparison of constructed zoning

Using additional data in the construction zone model yields excellent performance on all metrics. These three zone models demonstrate improved risk segmentation and increased stability. Among them, the zoning model that combines available ancillary data and a private hail database has the best within-class variance and well-tuned stability compared to the other zoning models.

The use of the paying hail database therefore makes it possible to improve the results. However, its purchase does not allow to create a zoning with significantly improved performance compared to those of the zoning using only the complementary data.

Abréviations

S/C : Ratio de la charge des sinistres par le montant des cotisations

A0 : Grêlons de taille comprise entre un demi centimètre et un centimètre
A1 : Grêlons de taille comprise entre un centimètre et deux centimètres
A2 : Grêlons de taille comprise entre deux centimètres et trois centimètres
A3 : Grêlons de taille comprise entre trois centimètres et quatre centimètres
A4 : Grêlons de taille comprise entre quatre centimètres et cinq centimètres
A5 : Grêlons de taille supérieure à cinq centimètres

GLM : Modèles linéaires généralisés (Generalized linear model)

CAH : Classification ascendante hiérarchique

Remerciements

Je souhaite adresser mes remerciements à l'ensemble des personnes ayant contribué, de près ou de loin, à la réalisation de ce mémoire.

Tout d'abord, je remercie mes responsables Marie Tien, Yann Mercuzot et Lionel Ferraud, pour m'avoir fait confiance dans la réalisation de mon alternance au sein de la Direction des particuliers de Pacifica. Je tiens également à les remercier pour leur expertise ainsi que l'accompagnement qu'ils m'ont offert durant l'élaboration de ce mémoire.

Tout particulièrement, je tiens à remercier Claire Rogier, ma tutrice, pour sa rigueur et ses connaissances techniques. Elle m'a guidé durant ce mémoire en me transmettant son expérience sur les produits d'assurance Climatiques. J'ai ainsi pu bénéficier d'une expertise métier conséquente, très appréciable pour construire un zonier.

De manière générale je souhaite remercier Clément Courtillat pour son concours dans l'implémentation des modèles de classification. Audrey Mahuzier, Julien Chemin, Xavier Miglietti ainsi que tous les membres de la direction des particuliers, des agricoles et professionnels de Pacifica pour leur disponibilité et les connaissances qu'ils m'ont transmises.

Je souhaite aussi remercier mon tuteur académique, Maud Thomas, pour ses différents conseils dans la structuration de cette étude.

Un grand merci également à François Thooris, Lucas Jullia, Chaïmaa Benabbou, Kilian Soares, et Arthur Corda, mes relecteurs, pour leurs retours pertinents, aussi bien sur la forme que sur le fond du mémoire.

Ce mémoire marquant la fin de mes études, je tiens à remercier l'ensemble des enseignants et professeurs qui m'ont transmis leurs connaissances de manière inspirante.

Je souhaite enfin remercier mes parents et mes proches pour leur soutien sans faille.

Table des matières

Introduction

Le changement climatique ainsi que ses conséquences sur les activités humaines représentent une grande problématique pour les compagnies d'assurance. Chaque année, différents phénomènes tels que par la sécheresse, le gel ou les tempêtes engendrent d'importants dégâts. Il est devenu nécessaire pour les différents acteurs de la société de trouver des solutions pour se couvrir contre ces risques. Les compagnies d'assurance constatent des résultats dégradés pour les produits couvrant les risques climatiques. Il apparaît donc essentiel de bien appréhender les caractéristiques de ces risques pour commercialiser un produit répondant aux besoins des assurés et qui ne pose pas des problèmes de solvabilité et de rentabilité.

En juin 2022, une sinistralité record a été observée par Keraunos [1]. A l'image de la Dordogne précédemment évoquée, le Département du Doubs a connu quant à lui, le 20 juillet 2022, un épisode de grêle d'une très grande violence qui impacta tous les acteurs de la zone. Cette tempête est symbolisée à travers les 800 appels reçus par les pompiers à cette occasion. Les articles de journaux rapportent notamment les perspectives compliquées pour une partie des professionnels, particuliers et surtout des agriculteurs sinistrés n'ayant pas souscrit à une assurance contre la grêle.

La garantie contre la grêle peut intervenir auprès des particuliers pour couvrir leurs véhicules ou leurs habitations, de la même façon pour les professionnels. Elle peut également intervenir auprès des agriculteurs en couvrant ces derniers contre les conséquences d'une tempête de grêle sur leurs cultures. Dans ce mémoire, nous allons nous intéresser au produit Grêle dans le cas de l'assurance sur les récoltes.

La grêle n'est pas considérée comme une catastrophe naturelle, mais comme ces dernières elle va apparaître de manière hétérogène sur le territoire. Il est important dans le cadre de la tarification d'affecter le bon risque à chacune des parties du territoire pour éviter tout problème d'antisélection. La construction d'un zonier sera l'objectif de ce mémoire.

Notre étude a été réalisée au sein de la Direction des particuliers, à l'intérieur du service actuariat, produits et réassurance. Cette direction a pour vocation de réaliser la tarification de l'ensemble des produits d'assurance non-vie et de garantir l'équilibre des résultats techniques. Ce mémoire vient s'inscrire dans cet objectif : adapter la tarification du produit Grêle. Nous allons venir tester les performances du zonier actuel et le mettre en concurrence avec d'autres modèles.

1. Observatoire français des tornades et orages violents qui répertorie les tempêtes de grêle depuis 2006.

La première partie de ce mémoire permettra de définir le cadre d'étude. Nous évoquerons le cadre assurantiel dans lequel se situe le produit Grêle sur le marché de l'assurance climatique. Nous expliquerons ensuite le phénomène physique qu'est la grêle afin de mieux l'appréhender dans notre étude. Nous évoquerons enfin les objectifs et les contraintes pour construire un zonier.

Dans un deuxième temps, l'ensemble des données utilisées pour mener à bien cette étude seront décrites. Nous définirons notamment l'historique de données sur le produit Grêle de l'assurance grêle de Pacifica. Nous présenterons également les données complémentaires qui permettront d'expliquer la sinistralité : celles internes, celles externes gratuites et celles externes payantes.

La troisième partie viendra présenter les trois principaux outils mathématiques utilisés dans ce mémoire : le krigeage, les modèles linéaires généralisés et la classification ascendante hiérarchique.

Au sein de la quatrième partie, nous décrirons les différents zoniers que nous aurons construit. Nous obtiendrons au final cinq zoniers. Ceux-ci seront construits à partir de techniques ou de variables explicatives différentes.

Ces cinq zoniers seront enfin comparés dans la cinquième partie. Nous analyserons le découpage, la capacité de segmentation, l'homogénéité des classes et la stabilité de chacun des zoniers. Nous recenserons enfin dans cette partie les axes d'améliorations possibles à notre étude.

Chapitre I Cadre d'étude

I.1 Cadre assurantiel

L'inversion du cycle de production dans le secteur assurantiel et la segmentation du risque entre assurés constituent deux problématiques pour les compagnies d'assurance non-vie. Des défis auxquels l'actuaire en tarification doit être en mesure de répondre.

Dans ce mémoire, nous étudierons le produit Grêle en assurance agricole et nous essaierons de répondre à l'objectif suivant : l'optimisation du zonier utilisé par ce produit. Pour ce faire, nous allons d'abord décrire l'environnement assurantiel dans lequel s'inscrit le produit Grêle, ensuite décrire le phénomène qu'est la grêle, et enfin présenter quelques notions de tarification primordiales dans notre étude.

Le risque dans le secteur de l'assurance non-vie est défini comme l'éventualité qu'un sinistre ou un préjudice se déclare. Différents produits d'assurance vont être commercialisés par une compagnie afin de couvrir les assurés face à la conséquence d'un risque qu'ils ne peuvent assumer individuellement. Nous allons regarder ici comment est traité le risque grêle sur les cultures par les assureurs.

I.1.1 Biens sinistrés par la grêle

La grêle constitue un aléa climatique pouvant occasionner d'importants dégâts pour plusieurs types d'acteurs de la société. Elle va affecter les particuliers à travers les dégâts qu'elle peut créer, par exemple sur les voitures ou les habitations. De même, elle va impacter les professionnels qui peuvent être grandement exposés via leur flotte ou leurs locaux. Il y a donc un intérêt important pour l'assureur de proposer une garantie grêle.

La garantie grêle va prendre en charge les sinistres causés par l'action mécanique du choc des grêlons sur les éléments assurés. Elle intervient, par exemple, dans les produits d'assurance Habitation et Automobile, afin de prendre en charge la restauration de la toiture ou de la carrosserie, ainsi que la mise à disposition d'un hébergement ou d'un véhicule provisoire...

Un acteur spécifique va également être sensible aux tempêtes de grêle : l'agriculteur. Lui aussi est exposé via ses locaux et ses véhicules, mais le risque réside surtout dans la fragilité de ses cultures face à la grêle. Ces dernières représentent, pour une grande partie d'entre eux, leur unique source de revenus. Les récoltes de l'agriculteur pourront subir une perte simultanée en quantité et

en qualité lors d'une tempête de grêle. Il peut exister certains produits d'assurance agricole qui couvrent seulement une perte de quantité. Il peut être opportun, afin de s'adapter aux spécificités fortes de chaque agriculteur, de commercialiser plusieurs produits différents proposant la garantie grêle.

I.1.2 Protection des cultures contre le risque grêle

Nous recensons trois grandes solutions possibles pour les agriculteurs afin de se prémunir des conséquences d'une importante tempête de grêle sur leurs récoltes.

I.1.2.1 Auto-assurance

Peu d'agriculteurs sont assurés face au risque grêle, seulement 18% d'après les chiffres du ministère de l'agriculture. Il existe une grande disparité entre les types de cultures de la surface assurée, par exemple seulement 3% des arboricoles sont assurés. Pour une majorité d'agriculteurs, le frein à souscrire à une assurance contre les risques climatiques est culturel. "Pour une majorité d'exploitants, la gestion du risque se fait avant tout sur le terrain et non sur des contrats financiers", rapporte Joran Chambolle dans un article d'Alix Coutures[5]. Joran Chambolle appartient à l'équipe agricole de l'entreprise Béssé conseillant les entreprises sur leurs risques et assurances. Cette article faisait suite à d'importantes tempêtes de grêle constatées sur le territoire français à l'été 2022.

Les agriculteurs utilisent ainsi l'auto-assurance comme premier moyen pour se protéger contre une potentielle baisse de production. Cela signifie que si le rendement de leurs récoltes est déficitaire, à cause d'une tempête de grêle ou pour toute autre raison, l'agriculteur a constitué une réserve qui lui permettra de faire face à cet aléa. Cette réserve est souvent financière. Elle peut également provenir de stocks constitués par le passé si le type de récolte le permet ; c'est le cas notamment des cultures de blé et des viticultures. Cette stratégie est limitée dans la mesure où le déficit persisterait sur plusieurs années consécutives.

Le deuxième moyen pour l'agriculteur est de couvrir ses cultures via des filets grêlifuges. Ces filets peuvent combiner plusieurs utilités comme protéger des attaques d'oiseaux en plus des intempéries. Ils sont idéaux pour protéger les cultures viticoles, celles arboricoles, mais ne s'adaptent pas vraiment aux grandes étendues que sont les cultures de céréales par exemple. A noter que, même si leur coût est abordable, couvrir ses cultures représente un investissement important pour

l'agriculteur. L'installation peut également être délicate et entraîner une baisse de l'ensoleillement sur les végétaux protégés.

L'agriculteur peut aussi privilégier certaines cultures plus résistantes pour se prémunir des conséquences des tempêtes de grêle. Le risque grêle n'est malheureusement pas le seul auquel doit faire face l'agriculteur, les cultures doivent en effet être résistantes, en plus de la grêle, au gel, à la sécheresse, aux maladies...

L'agrivoltaïsme, consistant à combiner la production agricole et la production d'énergie solaire sur un même terrain, peut être une solution face au risque grêle. Toutefois, il ne peut garantir une protection absolue contre la grêle car ces panneaux ne couvrent pas entièrement la culture. De plus, les panneaux sont sensibles aux tempêtes composées de gros grêlons. Cette solution permet, en revanche ,de lutter contre la sécheresse, tout en produisant de l'énergie.

Enfin, les agriculteurs ont la possibilité de recourir aux canons à iodure d'argent. Ceux-ci qui vont venir modifier les conditions météorologiques de la zone. Toutefois, les canons à iodure d'argent ne semblent pas pertinents car ils ont une efficacité limitée, non prouvée, et ils représentent de potentiels dangers pour la santé.

Cette auto-assurance est importante et va de pair avec la souscription à une assurance visant à se prémunir contre les risques climatiques. En effet, le système d'assurance n'est pas soutenable sans prévention ni actions des agriculteurs ; à l'inverse, l'auto-assurance n'est pas suffisante face au niveau de risque auquel font face les agriculteurs actuellement.

I.1.2.2 Aides publiques

L'État joue un rôle important dans la protection de l'agriculture contre les risques climatiques. L'indemnisation des pertes de récolte reposait, jusqu'à 2022, sur le fonctionnement parallèle de deux régimes :

— Le premier, celui des calamités agricoles, existait depuis les années 1960. Cofinancé par les agriculteurs et l'État, il excluait certains pans de l'agriculture (viticulture et grandes cultures) et ses délais étaient jugés trop longs.

— Le second était le système assurantiel, privé, mais subventionné à 65% par l'État. Le système était déficitaire et encore peu souscrit par les agriculteurs comme évoqué.

En février 2022, un texte a été adopté visant à créer "un régime universel d'indemnisation" à trois étages[13] :

1. Un premier niveau qui relève de l'agriculteur. Ce dernier devra assumer les pertes les plus modestes, jusqu'à un seuil de franchise.
2. Un deuxième niveau qui relève de l'assureur.
3. Un troisième niveau qui incombe à l'État, qui, au-dessus d'un seuil de pertes, mobilisera les fonds publics pour répondre à des situations de catastrophe.

Ce texte s'inscrit dans un souci de simplification pour inciter les agriculteurs à souscrire à une assurance. Pour répondre à cela, un guichet unique va être créé pour simplifier les démarches. Le texte prévoit également la création d'un pool d'assureurs pour mutualiser notamment le risque et les données, comme il est indiqué dans l'article de l'AFP[10].

De plus, l'épisode de gel tardif du printemps 2021 a montré les limites du système passé, obligeant notamment l'État à intervenir massivement. Ce texte vise à créer un système adapté aux problématiques actuelles. L'objectif est de créer une relation de confiance entre agriculteurs et assureurs, et de les inciter à s'assurer contre les risques climatiques, afin de tendre vers une importante part de surface assurée sur tout le territoire français d'ici 2030.

"Les agriculteurs fonctionnent avec la logique de l'État providence. Auparavant, le fonds calamité suffisait à couvrir les intempéries qui surviennent tous les 10 ans", abonde Dominique Chargé. Ce fonds n'est plus suffisant ces dernières années avec l'accélération d'un changement climatique planétaire ; il est ainsi nécessaire de réformer le système et cela va de pair avec le fait de devoir modifier la perception et les habitudes des agriculteurs en termes d'assurance. .

I.1.2.3 Produits d'assurance privée

Nous venons de voir que le rôle de deux acteurs, l'État et les agriculteurs, rôle primordial pour se prémunir contre le risque climatique encouru par les récoltes. Ces deux acteurs doivent agir conjointement avec l'assureur. Ce dernier doit avoir pour fonction de commercialiser des produits qui vont permettre la mutualisation du risque et s'adapter aux besoins des agriculteurs. Enfin, les agriculteurs doivent avoir un intérêt à souscrire à ce contrat.

L'assureur peut établir différents types de contrats d'assurance visant à garantir les récoltes. Il est possible d'assurer à la culture ou à l'exploitation. Dans le premier cas, la sinistralité est jugée

pour chaque récolte de l'exploitant. Dans le second cas, la sinistralité est jugée sur l'ensemble de l'exploitation.

L'assureur va venir assurer un capital, équivalent au chiffre d'affaire annuel. La durée du contrat d'assurance sur les cultures est quant à elle différente de l'assurance habituelle. Normalement, la couverture correspond à une année complète de risque à la partir de la date d'effet du contrat. Dans le cas de la couverture des cultures agricoles, la date d'effet du contrat peut commencer parfois au 1^{er} mars et cesser à partir de la fin de la récolte : nous parlons d'année de récolte. Des exceptions existent, notamment pour les arbres fruitiers, où les contrats commencent à des dates variables, dépendant du moment de floraison.

L'agriculteur aura le choix entre deux types de produit selon ses besoins : Multirisques Climatiques et Aléas Climatiques. Dans le cadre du premier, il sera assurer contre plusieurs climatiques (gel, sécheresse, grêle...), alors que dans le second cas il pourra s'assurer contre un risque précis.

I.1.3 Assurabilité des cultures face au risque grêle

Ce questionnement est abordé dans le mémoire de L.Batisse et Y.Mercuzot[2] sur l'étude du risque climatique et a tout son intérêt dans le cadre du produit Grêle.

I.1.3.1 Assurabilité juridique

Au moment de la souscription du contrat, la grêle constitue un événement futur non réalisé, aléatoire dans l'espace et le temps, indépendant de la volonté de l'assuré. A première vue, il est impossible d'anticiper plusieurs semaines à l'avance et avec exactitude une tempête. La grêle peut donc être assurée légalement car les sinistres qu'elle cause ont des conséquences quantifiables pouvant donner lieu à des indemnités.

Nous pouvons par ailleurs nous demander si la grêle n'est pas un phénomène annuel systématique. Est-ce que, pour une partie du territoire, la grêle ne tombe pas chaque année et vient endommager les cultures ? C'est le cas actuellement de la sécheresse, qui, dans certaines zones, est devenue récurrente. Le caractère aléatoire disparaît alors et l'assurabilité des cultures est remise en question. La grêle demeure un phénomène gardant un caractère très aléatoire. La fréquence des fortes tempêtes est très faible sur la plupart des communes. La grêle est bien un phénomène assurable juridiquement.

I.1.3.2 Assurabilité économique

L'aléa moral représente la situation où un assuré, protégé d'un risque par la compagnie, va se comporter autrement que s'il était lui-même totalement exposé au risque. Cette situation n'existe pas dans le cadre de la garantie grêle car :

— L'agriculteur ne peut anticiper l'occurrence d'une tempête de grêle durant l'année.

— Il est difficile pour l'agriculteur de se prémunir contre la grêle, voire impossible pour beaucoup de cultures.

Le comportement de l'agriculteur face au risque de grêle, qu'il soit assuré ou non, ne change pas car il ne peut ni s'en prémunir, ni l'anticiper.

L'antisélection apparaît lorsqu'existe une asymétrie d'information entre l'assuré et l'assureur au sujet d'un certain risque. Dans ce cas, l'assuré a une meilleure connaissance du risque. Ce problème est un enjeu du mémoire : nous chercherons à mieux connaître la grêle, notamment les couloirs de tempêtes, pour pallier à cette asymétrie potentielle.

I.1.3.3 Assurabilité actuarielle

L'assurabilité sur le plan actuariel repose sur trois notions clés :

— Les évènements assurés et la charge moyenne à supporter doivent être identifiables et quantifiables.

— Les risques doivent être mutualisables.

— Les pertes maximales ne doivent pas mettre en péril la solvabilité de l'assureur.

Pour le premier point, la compagnie est en mesure d'estimer le capital assuré et d'inscrire ceux-ci dans le contrat. Ainsi, la charge est quantifiable et identifiable. L'évènement assuré, la tempête de grêle, est lui aussi identifiable. Le sujet de ce mémoire portera sur le fait de quantifier la probabilité d'apparition de cet évènement.

Concernant le deuxième point, la mutualisation revient à une application de la loi des grands nombres. Cette dernière est vérifiée puisque nous avons : un grand nombre d'observations (de l'ordre du millier), une indépendance à l'échelle nationale (la grêle est un phénomène très localisé) et une distribution identique de la probabilité du risque au sein des groupes (la segmentation du risque vient corriger les différences de risque entre observations).

Concernant le troisième point, la compagnie est en mesure de construire un produit qui, selon les différents scénarios, puisse répondre aux contraintes de solvabilité nécessaires. Nous verrons, à

travers les différents ratios de sinistralité qui seront présentés, que le produit Grêle est rentable sur le marché français.

Au final, nous allons donc étudier dans ce mémoire un produit assurable sur les plans juridique, économique et actuariel.

I.1.4 Garantie grêle sur le marché français

En octobre 2021, la Fédération Française de l'Assurance a publié un rapport sur l'assurance agricole : "L'assurance agricole en 2020"[11], au sein de laquelle apparaît une étude sur le produit Grêle.

Lors de l'exercice 2020, le marché de l'assurance grêle comprenait 74 200 contrats, pour 177,5 millions de cotisations acquises. A noter que le volume de cotisations et plus fortement le nombre de contrats décroissent. Ceci est lié au transfert des contrats vers l'offre multirisques climatiques.

Les cotisations acquises ne sont pas uniformément réparties sur le territoire, comme nous pouvons le voir à travers la carte en Figure 7 :

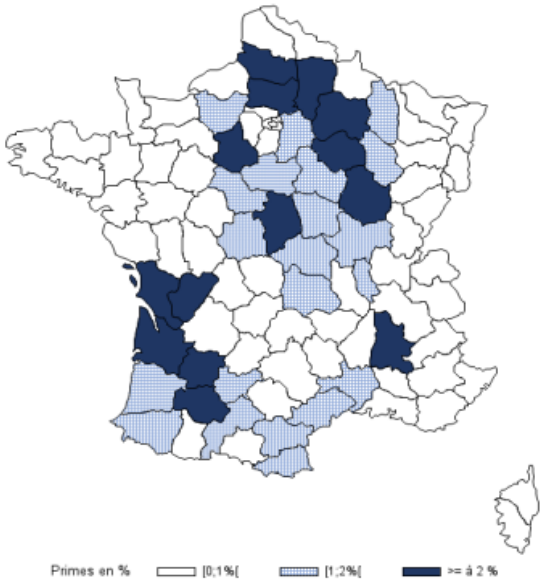


FIGURE 7 – Répartition des cotisations Grêle par département en 2020 (en %) (rapport FFA)

Des zones comme le Nord-Ouest, l'Alsace et la Provence-Alpes Côte d'Azur, semblent très faiblement représentées en terme de cotisations à l'assurance Grêle. Ceci implique un volume de données moins conséquent dans ces régions, et donc une moins bonne connaissance du risque.

Parmi les départements fortement représentés, figurent : les départements viticoles comme le Bordelais, la Côte d'Or et la Champagne, également la terre arboricole qu'est la Drôme. Ces zones, à cause du type de culture assuré, ont une forte valeur et sont très sensibles à la grêle. En Picardie, à l'inverse, la majorité des terres sont agricoles et on y trouve notamment des cultures de blés tendres. Globalement, la garantie grêle est inégalement répartie sur le territoire : que ce soit en montant assuré, et aussi en cultures assurées.

A travers les deux graphiques en Figure 8, nous observons la répartition des capitaux assurés et des cotisations des groupes de récoltes qui ont souscrit à la garantie grêle :

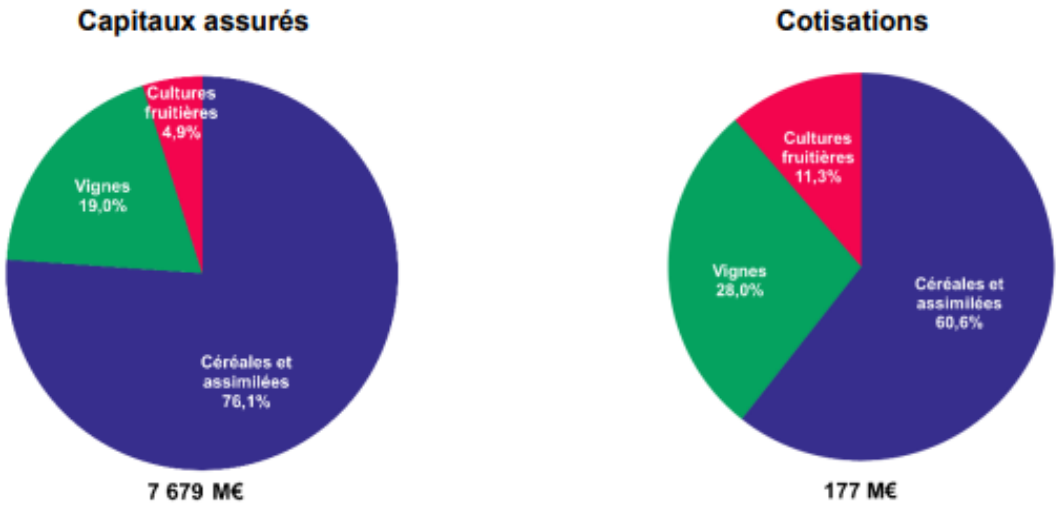


FIGURE 8 – Répartition des capitaux assurés et des cotisations en 2020 (rapport FFA)

Une forte représentation des vignes est visible, à hauteur de 28% des cotisations. Toutefois, la majorité des capitaux assurés, 75%, sont des céréales. Nous pouvons remarquer également que les cultures fruitières représentent 11,3% des cotisations pour seulement 4,9% du risque, symbolisant un risque plus important sur les cultures arboricoles. Un phénomène similaire est visible sur la vigne, mais moins significatif.

La Figure 9 montre une importante volatilité du ratio de S/P selon les exercices. Par exemple, en 2015, nous observons un S/P de 19%, alors qu'il était de 87% en 2014, soit 4,5 fois plus important l'année précédente. Toutefois, notons que le S/P cumulé historique est stable autour de 64% depuis 2016.

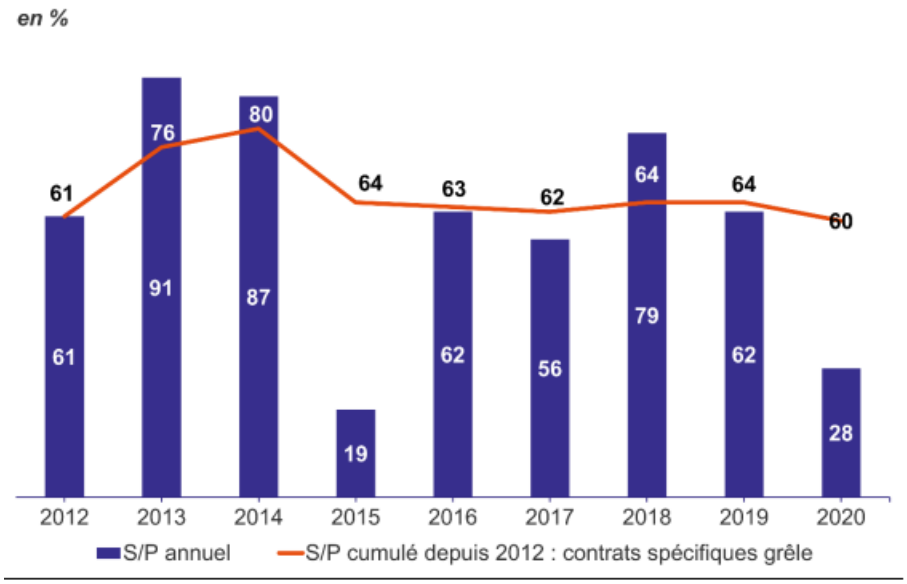


FIGURE 9 – Évolution du ratio sinistres à primes de l'assurance grêle (rapport FFA)

Certaines cultures dépassent le seuil de rentabilité, comme le montre le graphique extrait ci-dessous :

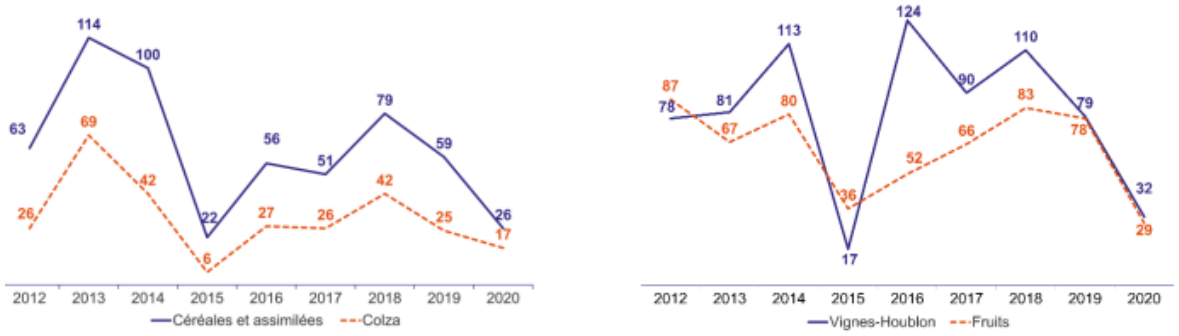


FIGURE 10 – Évolution du ratio S/P depuis 2012 selon les types de cultures (rapport FFA)

Les cultures de colza et les exploitations arboricoles ont de bons S/C, inférieurs à 100% chaque année. Les cultures de céréales et assimilés présentent également de bons ratio de sinistralités annuels, constamment inférieurs à 100% depuis 2015. Ce n'est pas le cas en revanche des vignes, dont le S/C a dépassé le seuil des 100% trois fois au cours des neuf exercices étudiés. Nous constatons à travers ces précédents graphiques que la vigne nécessitera un traitement différent des autres cultures pour être mieux appréhendée. En effet, la vigne présente un caractère atypique du fait de sa localisation très précise sur le territoire, sa fragilité à la grêle et une représentation non négligeable en portefeuille.

Enfin, certains départements présentent des résultats très dégradés. C'est le cas par exemple de la Creuse, la Lozère et les Alpes de Hautes-Provence sur la figure ci-dessous :

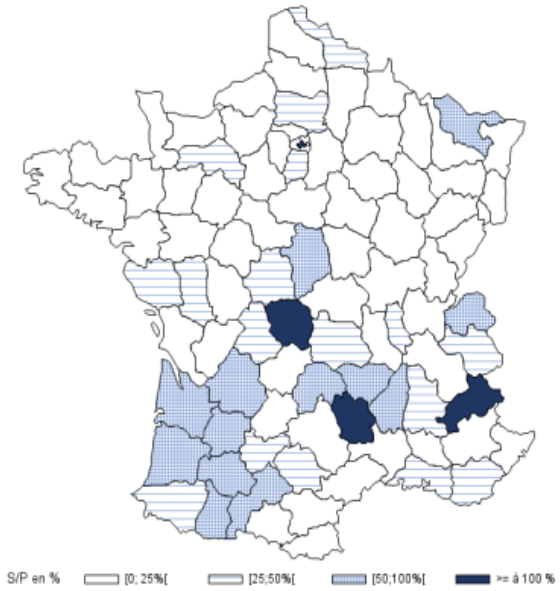


FIGURE 11 – Évolution du ratio S/P depuis 2012 (rapport FFA)

Notons que pour les trois départements cités précédemment, le volume de capitaux assurés est très faible. Nous remarquons une zone avec un résultat dégradé dans le Sud-Ouest, zone réputée grêligène.

Nous venons d'analyser la représentation et les résultats de la garantie grêle sur le marché agricole français. Intéressons-nous maintenant à notre socle d'étude, la garantie grêle chez Pacifica.

I.1.5 Produit grêle chez Pacifica

I.1.5.1 Offre grêle

Le produit grêle a pour garantie principale la grêle, il est complété de la garantie tempête. D'autres garanties optionnelles existent comme gel de printemps ou vent de sable.

Le produit grêle couvre une perte de quantité et de qualité, donc de rendement, sur la culture, consécutivement à une tempête de grêle. Contrairement à l'assurance Multirisques Climatiques, le produit grêle assure au cours de toute une année de récolte. La garantie va intervenir à la parcelle : chaque étendue de terre d'un seul tenant de l'exploitant sera assurée individuellement. Aucune compensation à l'échelle de l'exploitation ou de parcelles voisines ne sera alors possible. Toutes

les cultures sont assurables par le produit grêle, à l'exception du bois des arbres, à différencier du bourgeon et du fruit, et des cultures sous serres.

L'indemnisation de l'assuré pour ce produit correspond à un pourcentage de perte appliqué au capital assuré total. Ce montant assuré correspond au produit du capital assuré par hectare et de la surface assurée. Pacifia fait le choix dans son produit grêle d'assurer une perte de quantité, mais également une perte de qualité. Cette perte de qualité est traduite en terme de rendement grâce à un coefficient de dépréciation qualitative. Le calcul de la perte de rendement se fait à dire d'expert selon une procédure préétablie.

Enfin, l'assuré a la possibilité de modifier sa franchise de base. Celle-ci est une information importante à prendre en compte pour juger de la fréquence : la grêle peut parfois ne ravager qu'une petite partie de la culture, pour un seuil ne dépassant pas la franchise. Par exemple, dans le cas de tempêtes lors desquelles des grêlons de moins de deux centimètres sont tombés : la culture n'est probablement pas sinistrée aux yeux de l'assureur, mais une partie a tout de même pu subir des pertes.

I.1.5.2 Équation tarifaire

Pour ce produit, la cotisation d'un contrat est une fonction linéaire du volume de capitaux assurés. Ces derniers se calculent comme suit :

$$\text{Capitaux assurés} = \text{Surface (ha)} \cdot \text{Valeur de la récolte (€/ha)} \tag{1}$$

où la valeur de la récolte correspond au produit du rendement par hectare et du prix d'une unité de récolte en euros.

Nous retrouvons dans cette équation, des variables non géographiques comme la franchise, décrite brièvement précédemment. De plus, la famille de natures de récolte permet de différencier la sensibilité de la culture à la grêle. Figurent également dans cette équation des variables géographiques, qui vont intervenir dans le cadre du zonier.

Dans ce mémoire, nous étudierons plus en détail ces variables présentes dans l'équation. Nous viendrons également compléter l'équation tarifaire avec de nouvelles variables que nous jugerons pertinentes.

I.2 Risque grêle

I.2.1 Orage de grêle

La grêle est un phénomène météorologique frappant de manière brève une zone géographique restreinte durant certains mois de l'année. Les dégâts engendrés peuvent grandement varier d'une tempête à l'autre. La grêle possède un caractère grandement aléatoire. Cependant, ses composantes physiques, que nous allons détailler ici, vont nous aider à mieux appréhender le phénomène et à mettre en évidence les zones les plus risquées et celles moins risquées.

I.2.1.1 Définitions

Le phénomène physique de formation de la grêle utilise des termes techniques qui lui sont propres et qui doivent être définis en préambule d'explications plus détaillées.

a) Termes relatifs à la formation de la grêle

* Un noyau glaçogène (ou glacigène) correspond à tout grain de matière qui permettra de donner naissance à de la glace, que ce soit par condensation solide de la vapeur d'eau (noyau de sublimation) ou que ce soit par congélation de l'eau (noyau de congélation).

* Le cumulonimbus est lui un type de cumulus fortement chargé en énergie et composé de cristaux de glace sur sa partie supérieure. Nous retrouvons à l'intérieur de celui-ci les noyaux glaçogènes et une forte quantité d'eau en condensation.

* Le courant d'air ascendant, autre notion très importante pour comprendre le phénomène, va directement former le cumulonimbus. Ce courant se forme suite à la rencontre de deux poches d'air aux températures très différentes, provoquant une montée des bulles d'air chaud qui formeront un cumulus.

Les grêlons sont issus de noyaux glaçogènes présents dans les cumulonimbus, et vont se transformer en glace. Cette transformation se produit dans la troposphère, couche inférieure de l'atmosphère, au sein de laquelle les températures avoisinent les -10°C propices à la congélation. La solidification de ces noyaux se fait par deux mécanismes :

— Condensation solide, c'est-à-dire passage de l'état gazeux à l'état solide de la vapeur d'eau présente dans le nuage.

— Congélation de l'eau surfondue présente dans le nuage.

Le grêlon ainsi formé est défini comme un grain de glace de diamètre variable, formé dans les nuages, dont la chute en abondance constitue la grêle. Une fois la phase de formation du grêlon effectuée, il y a ensuite l'accrétion : l'accroissement d'un corps par apport avec ou sans agglomération de matière. Le grêlon va subir, selon les cas, une accrétion plus ou moins importante afin de former, dans des cas extrêmes, des grêlons de plusieurs centimètres.

b) Termes relatifs à la tempête de grêle et ses dégâts

Les tempêtes qui produisent des grêlons atteignant le sol sont appelées tempêtes de grêle. Elles ne durent généralement pas plus de quinze minutes mais peuvent blesser des personnes et endommager des bâtiments, des véhicules et des cultures. Ces tempêtes vont frapper une zone que nous nommerons "couloir de grêle". En effet, la grêle va souvent frapper une zone précise du fait d'un chemin préférentiel des cumulonimbus. Ces couloirs ne sont pas totalement aléatoires.

La tempête de grêle varie de par sa durée, sa localisation mais également de par son intensité. L'intensité d'une tempête de grêle est matérialisée par le diamètre du grêlon, variant de cinq à cinquante millimètres et pouvant atteindre les dix centimètres.

I.2.1.2 Explications physiques

Nous avons précédemment défini les termes propres à la grêle. Nous allons à présent voir comment se forme cette précipitation atmosphérique, de type solide, apparaissant dans les cumulonimbus. Ceci nous permettra de mieux comprendre pourquoi certaines zones du territoire sont significativement plus sinistrées par le phénomène que d'autres.

La première phase de formation correspond à la solidification du noyau glaçogène déjà évoquée. Une fois celle-ci effectuée, le grêlon va pouvoir atteindre différentes tailles variant de plusieurs centimètres selon les cas. Ce grossissement du noyau en grêlon s'effectuer en plusieurs étapes. Celles-ci sont décrites en Figure 12 et à travers l'énumération suivante[15] :

1. Au départ, le noyau glaçogène se situe dans des zones aux températures très négatives, du fait de l'altitude très élevée. Dès lors, toute vapeur d'eau venant en contact avec le grêlon va se solidifier et contribuer à faire grossir progressivement ce dernier : c'est ce qu'on appelle la croissance « sèche ».

2. Le grossissement du grêlon va contribuer à l'alourdir. La particule ainsi formée dans le nuage va descendre un peu plus que les gouttelettes d'eau surfondue. Elle va donc les collecter et les faire geler instantanément. Le grêlon est alors de quelques millimètres.

3. Le grêlon continue sa descente, descente de plus en plus rapide à force de grossir. A partir d'une certaine altitude, les particules se réchauffent et peinent à geler au contact du grêlon. La particule prend alors une apparence transparente et sa densité augmente. Nous parlons alors de croissance « humide ».

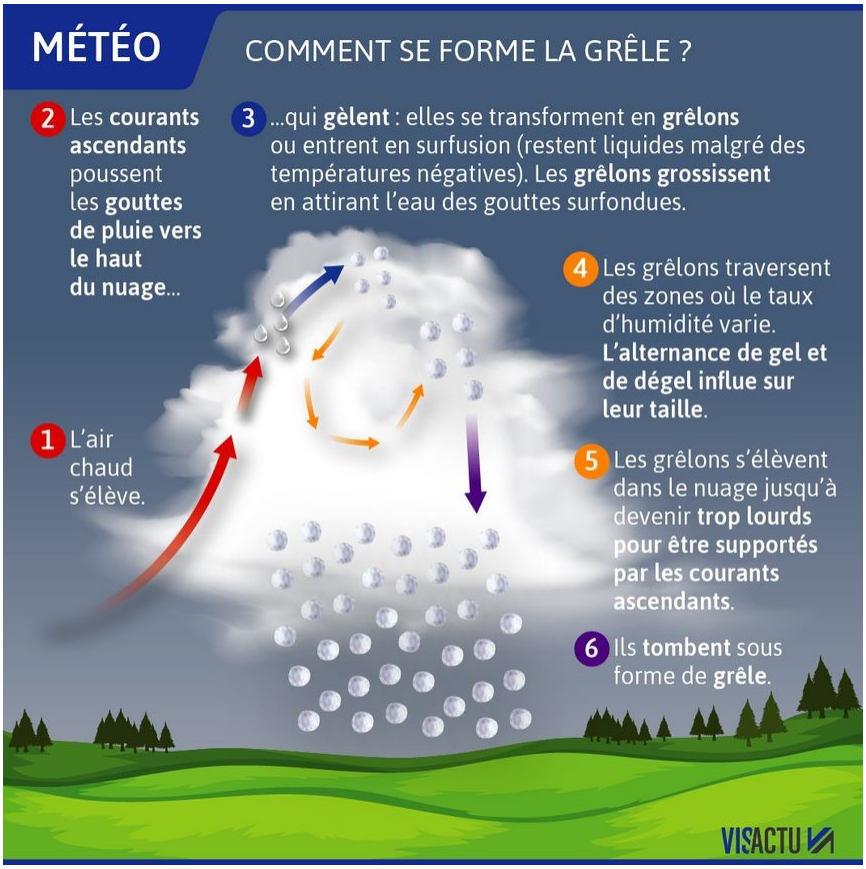


FIGURE 12 – Schéma descriptif de la formation d'une tempête de grêle (source : VISACTU)

Les accrétions «sèches» et celles «humides» ne sont pas uniques au cours de la formation du grêlon, elles vont s'alterner. Ceci explique que le grêlon est formé de plusieurs couches, couches différentes en apparence, comme le montre la Figure 13. Le nombre de répétitions de ces cycles de croissance dépend de la vitesse de chute et des caractéristiques de l'environnement du grêlon, notamment de la température et de la concentration en eau surfondue. Nous observons déjà une caractéristique importante favorisant la grêle : une atmosphère avec une température assez froide pour permettre la solidification.

FIGURE 13 – Couches d'un grêlon, symbole des étapes de formation de celui-ci

Enfin, la grêle quitte le cumulonimbus pour tomber au sol lorsque le grêlon est suffisamment lourd par rapport à ce que le courant d'air ascendant du nuage peut supporter pour le soulever une nouvelle fois. Les gros grêlons sont favorisés par des courants d'air ascendants assez puissants pour continuer à soulever ces derniers. En effet, ceci permettra de renouveler plusieurs fois les phases d'accrétion et ainsi d'additionner les couches autour du grêlon.

Intéressons-nous maintenant aux facteurs favorisant la formation d'un orage de grêle. Le phénomène apparaît dans une masse d'air très instable, chaude et humide, bien au-dessus du point de congélation. Les tempêtes de grêle pourront donc se produire plus facilement dans les zones où de telles masses d'air apparaissent, phénomène survenant généralement l'été.

Nous pouvons au final retenir que certaines régions, du fait de leur atmosphère, sont plus favorables à la grêle. C'est le cas par exemple des régions en altitude, propices au faibles températures et pouvant être des lieux où l'air est particulièrement instable. Les régions connaissant des fortes variations de températures au cours de la journée ou du mois peuvent être également favorables à la grêle car nous y constatons des courants d'air ascendants importants.

I.2.2 Orages de grêle en chiffres et en cartes

Nous venons d'analyser le phénomène physique de la grêle. Attardons-nous maintenant sur sa répartition géographique et sa répartition au cours de l'année à travers des études empiriques. Le phénomène de grêle étant de faible fréquence et volatile sur le territoire, son étude nécessite un historique conséquent et une collecte des données rigoureuse. Il existe certaines études réalisées sur le territoire français. Toutefois, afin de mieux comprendre les zones propices aux tempêtes de grêle, nous avons également analysé des études réalisées sur le territoire américain, plus abondamment documentées.

I.2.2.1 Analyse de la grêle dans certaines zones du monde

La National Oceanic and Atmospheric Administration (NOAA) est l'agence américaine responsable de l'étude de l'océan et de l'atmosphère. Elle précise que ce phénomène atmosphérique est plus probable à certains moments et endroits que d'autres. Il n'en demeure pas moins qu'il peut frapper chaque zone à tout moment de l'année.

En parcourant d'autres études, nous recensons que les tempêtes de grêle sont plus fréquentes :

— Au-delà des tropiques, malgré une fréquence d'orages plus forte, car l'atmosphère a une tendance à être plus chaude dans cette zone.[6]

— A l'intérieur des terres, car la formation de grêle est considérablement plus probable lorsque le niveau de congélation est inférieur à 4000 mètres d'altitude.[3]

— Le long des chaînes de montagne à cause du soulèvement orographique.[14]

— Pour des altitudes intermédiaires, comprises entre 300m et 1000m : les études semblent s'accorder que la grêle apparaît moins souvent aux altitudes basses et très hautes. Ces zones ont une constitution atmosphérique peu propice à la formation de la grêle. Á basse altitude, la grêle a notamment le temps de fondre.

Un cas d'étude a également retenu notre attention, celui de la Hail Alley (Allée de grêle). Cette zone se situe aux États-Unis et comprend les États du Colorado, du Nebraska et du Wyoming. Nous constatons, dans cette large zone, certaines communes extrêmement sujettes à la grêle.

Nous remarquons, par exemple, le cas de la ville de Cheyenne, dans le Wyoming. Cette dernière est connue comme la ville la plus touchée par la grêle aux États-Unis. La grêle a encore fortement sinistré la commune en 2021. Au total, une dizaine de tempêtes de grêle ont été constatées cette année-là lors desquelles des grêlons de plusieurs centimètres sont tombées.

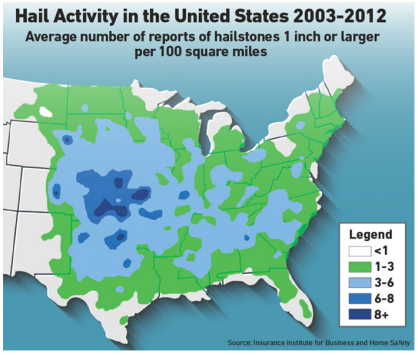
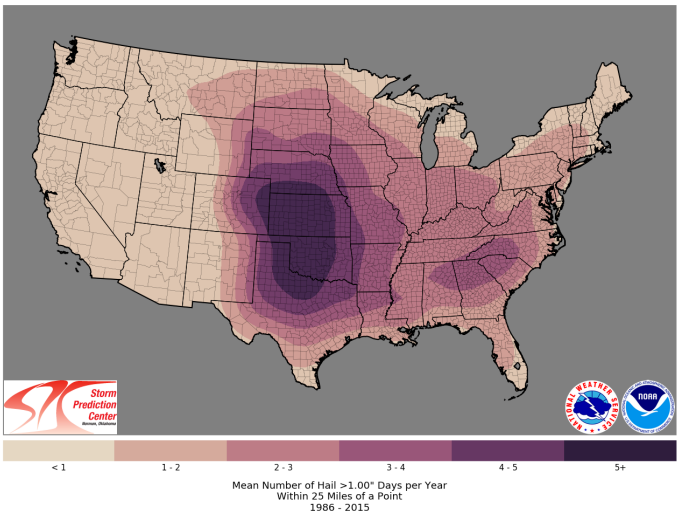


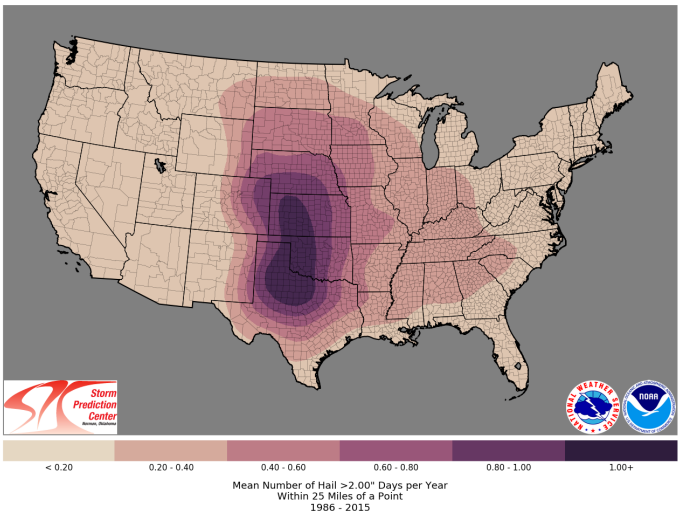
FIGURE 14 – Répartition du nombre de tempêtes de grêle sur le territoire américain de 2003 à 2012 par 100 Miles^2 (Agrégation des données assurantielles américaines)

Sur la Figure 14 est représentée l'occurrence de la grêle pour une partie des communes du territoire américain : l'Ouest, faiblement touché, n'est pas représenté ici. Nous constatons une très forte sinistralité dans les terres, au pied des rocheuses, à une altitude intermédiaire. Cette carte confirme donc les facteurs favorisant la grêle cités précédemment. Toutefois, cette carte comporte des petites zones, d'une ou quelques communes isolées, fortement sinistrées de manière inexpliquée, à l'image du cas de Cheyenne.

Les cartes, ci-dessous, sont formées par un relevé quotidien, de 1986 à 2015. Elles représentent des orages de grêle dont la taille minimale des plus gros grêlons dépasse un pouce pour la première, deux pouces pour la seconde. A partir de ces relevés, la NOAA a utilisé des outils mathématiques de lissage dans le temps et dans l'espace pour gommer les effets aléatoires de la grêle. Sur la Figure 15a apparaissent toutes les tempêtes de grêles pouvant causer des dégâts sur les cultures, c'est-à-dire celles où des grêlons de taille supérieure à un pouce (2,5 cm) sont recensées :



(a) Grêlons de plus de 1 pouce de diamètre (source : NOAA)



(b) Grêlons de plus de 2 pouces de diamètre (source : NOAA)

FIGURE 15 – Répartition du nombre moyen de jours par an avec de la grêle sur le territoire américain (source : NOAA)

La Figure 15a, établie grâce à un large historique, vient confirmer les hypothèses précédemment avancées, notamment une concentration de la sinistralité dans les terres en moyenne montagne.

Sur la Figure 15b apparaissent les zones touchées par de très fortes tempêtes de grêle, grêlons de deux pouces et plus. A partir de ce diamètre, tout type de culture est considéré comme ravagé suite à la tempête. Une corrélation entre fréquence et intensité de la grêle est visible. En effet, la représentation géographique des fréquences est similaire sur les deux cartes précédentes.

Notre objectif sera de constituer un zonier similaire à la représentation visible sur la Figure 14.

I.2.2.2 Analyse de la grêle sur le territoire français

Nous allons vérifier si les études américaines peuvent être extrapolées au cas français, si les phénomènes avancés précédemment favorisent la grêle.

Sur la carte en Figure 16, les tempêtes de grêle sont sur-représentées dans les terres, aux altitudes intermédiaires. Enfin, ces zones semblent bien se situer au pied des chaînes de montagnes françaises (Alpes, Pyrénées, Massif central), comme c'était le cas sur le territoire américain avec la Cordillère des Andes.

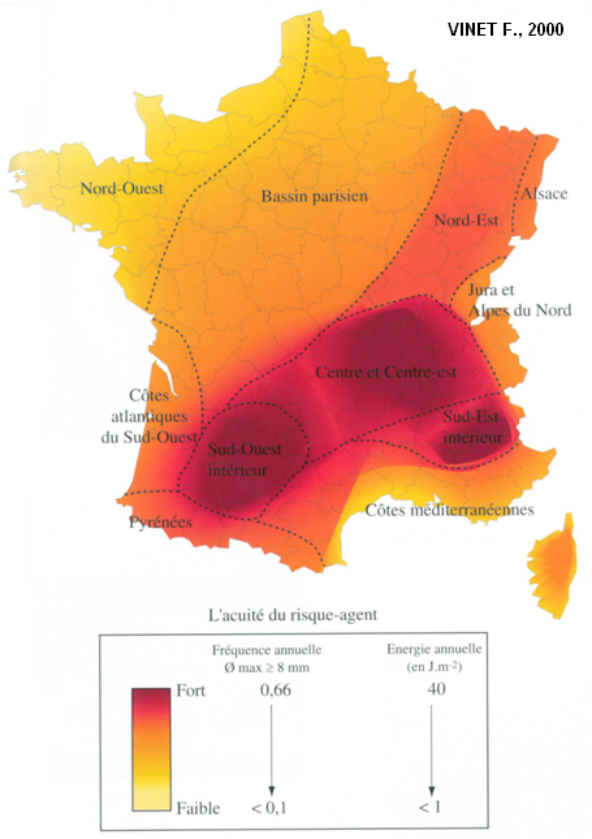


FIGURE 16 – Répartition de la fréquence des tempêtes de grêle sur le territoire français (F.VINET 2000)

Nous notons donc quelques points de divergence entre les études américaines et l'étude de F.VINET et P.PIGEON[16] en 2000, étude française la plus importante sur le sujet. Les deux points pour l'instant inexpliqués concernent la forte sinistralité dans le Sud-Ouest intérieur, la région toulousaine approximativement, et dans les Hautes-Alpes. Elles peuvent, par exemple, résulter de fortes variations de températures, propices à la formation de cumulonimbus.

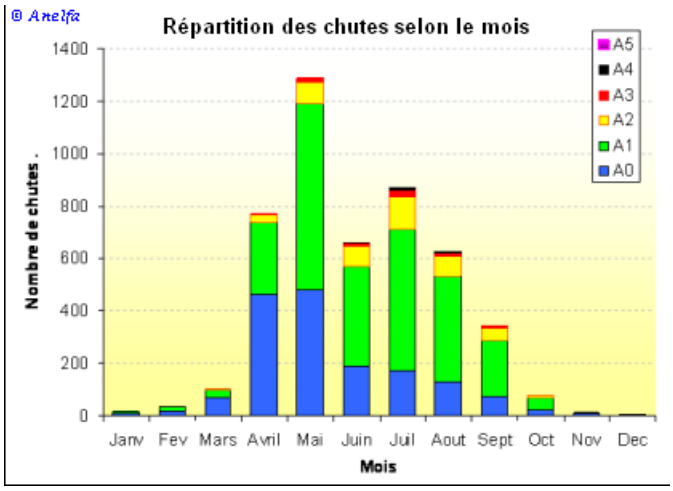


FIGURE 17 – Répartition des chutes de grêle selon le mois de l'année en France et selon le type (source Anelfa)

2

Enfin, la Figure 17 illustre que la grêle tombe nettement plus souvent et plus intensément au printemps et durant l'été. Il sera judicieux de retenir les données météo observées seulement au printemps et en été, d'autant plus que c'est à cette période que les différentes récoltes sont les plus sensibles à la grêle. L'hiver et l'automne peuvent donc être retirés du cadre d'étude.

I.3 Quelques concepts de la tarification non-vie

Nous allons décrire ici les notions importantes de la tarification, socles de la suite du mémoire qui vont permettre de lutter contre le phénomène d'anti-sélection.

I.3.1 Segmentation et mutualisation du risque

La mutualisation permet aux assurés de transférer l'aléa, qu'ils ne peuvent supporter individuellement, vers l'assureur. L'assureur pourra supporter l'aléa grâce à l'agrégation de tous les contrats du portefeuille.

Pour se prémunir du risque, les différents acteurs économiques ont recours à un contrat d'assurance. Il est construit de telle sorte que l'assuré paie une cotisation auprès de l'assureur et ce dernier verse à l'assuré une prestation lorsqu'un sinistre survient. Toutefois, chaque assuré possède une probabilité d'occurrence de sinistre qui lui est propre. Les prestations potentielles à régler en cas de sinistre ne sont également pas les mêmes pour chaque assuré. Les assurés sont tous définis par des caractéristiques qui leur sont propres et qui vont participer à définir leur risque auquel ils

font face. Au final, le tarif va être construit pour mutualiser un portefeuille d'assurés, mais ce portefeuille contient des assurés avec différents niveaux de risque.

Dans un contexte d'information parfaite et d'existence de modèles parfaitement discriminants, il serait possible d'attribuer à chaque assuré son risque et ainsi avoir une segmentation exactement adaptée à chacun. Dans le cas, du produit grêle, ceci correspond à être en capacité d'attribuer à chaque parcelle de terrain le niveau de risque qui lui est propre. Or, l'information à disposition n'est pas parfaite et n'est pas suffisamment complète sur le territoire. Il est alors compliqué de tarifer exactement chaque assuré.

L'actuaire va utiliser toutes les informations disponibles des caractéristiques de l'assuré afin de créer des groupes d'assurés en fonction des similitudes de risque. Il sera ainsi possible de créer des segments d'assurés plus homogènes. Cette homogénéisation s'inscrit dans le besoin d'éviter l'antisélection.

L'anti-sélection en assurance apparaît lorsqu'un assureur, mal informé sur le risque de ses assurés, tarifie un assuré «A» faiblement risqué au même prix qu'un assuré «B» fortement risqué. «A», connaissant mieux que l'assureur la valeur de son risque, jugera que son tarif est trop élevé et résiliera son contrat. Le portefeuille va alors être déséquilibré et sera porté par des assurés trop risqués. Le tarif du portefeuille sera également sous-estimé. Ce phénomène, dans le cas d'un zonier grêle, apparaît lorsqu'un groupe d'agriculteurs possède l'information que la grêle ne tombera jamais, ou presque, sur sa parcelle. Ce groupe n'aura alors aucun intérêt à s'assurer au même tarif que les autres. Au contraire, leurs voisins, eux, sinistrés très souvent, auront intérêt à cotiser auprès de l'assureur. Ces derniers déséquilibreront le portefeuille et le rendront trop risqué.

Il est donc important pour l'assureur de construire un tarif qui différencie au mieux les individus. Cette différenciation se fait sous la contrainte de l'information disponible et des outils à disposition. Des disparités au sein des segments créés subsisteront, mais l'anti-sélection sera supprimée.

I.3.2 Variables cibles

a) Notion de prime pure

La prime pure d'un produit est le montant permettant à l'assureur de régler les sinistres frappant l'ensemble des assurés. La prime pure est appelée également « prime de risque » ou encore « prime d'équilibre ». C'est elle qui va évaluer le risque auquel fait face la compagnie. Les frais et divers chargements ne sont pas inclus dans la prime pure et ne seront pas traités dans ce mémoire.

Pour une compagnie d'assurance ayant N contrats exposés sur tout l'historique, la formule de la prime pure moyenne correspond à :

$$Prime\ pure = \frac{\sum_{i=1}^{N} Charge\ grêle_i}{N} \tag{2}$$

La charge déclarée sur la garantie grêle au cours de l'année est bien la variable que nous souhaitons être en mesure de prédire pour chaque assuré. Plus particulièrement, dans le cas d'un zonier, nous souhaitons être en mesure de connaître la charge de sinistres due à la grêle pour une zone donnée, tout autre facteur tarifaire égal par ailleurs.

Dans notre cas, nous retiendrons la notion de taux de prime pure. Le recourt à un poids, ici l'inverse du volume de capitaux assurés de l'observation, va permettre de retirer l'effet lié à la valeur de la parcelle dans la charge. Les observations deviennent alors comparables entre elles.

Pour une compagnie d'assurance ayant N contrats exposés sur tout l'historique, la formule de la prime pure moyenne correspond à :

$$Taux\ prime\ pure = \frac{\sum_{i=1}^{N} charge\ grêle_i}{\sum_{i=1}^{N} capital\ assuré_i} \tag{3}$$

b) Notion de fréquence/coût

Ici, nous décomposons la prime pure via deux variables à expliquer : la fréquence et le coût moyen. La fréquence correspond au ratio du nombre de sinistres sur les années assurances. Le coût moyen correspond quant à lui à la charge totale de sinistralité divisée par le nombre de sinistres. Nous obtenons donc la formule ci-dessous pour le même cadre que précédemment :

$$Prime\ pure = \sum_{i=1}^{N} Coût\ Moyen_i * Fréquence_i \tag{4}$$

Cette notion de fréquence-coût semble plus pertinente pour segmenter car les caractéristiques de chaque assuré n'expliquent pas de la même façon le coût moyen et la fréquence. Toutefois, dans ce mémoire, nous utiliserons le taux de prime pure comme variable à expliquer dans nos modèles par simplification. De plus, un modèle de coût moyen nous poserait souci car il concerne seulement les observations déjà sinistrées. Or, pour le cas du portefeuille grêle, une majorité des contrats n'ont pas subi de sinistre sur l'historique.

I.3.3 Zonier

I.3.3.1 Construction d'un zonier

Le zonier en assurance correspond à un découpage géographique d'un ensemble où chaque zone se voit attribuer un coefficient tarifaire. Le découpage est effectué selon l'exposition au risque de chacune des zones, tout facteur non géographique égal par ailleurs.

Il existe plusieurs mailles possibles pour un zonier, comme les mailles issues du découpage administratif du territoire : parcelle, commune, département ou région. Ce type de découpages a l'avantage d'offrir une simplicité dans la commercialisation du tarif et dans la communication autour de celui-ci. Toutefois, de telles mailles n'ont pas vraiment de justification statistique : par exemple, une région comme le Rhône-Alpes est très hétérogène face au risque grêle. De plus avec un tel maillage, nous pouvons avoir deux communes pour des départements limitrophes qui auront un tarif nettement différent. Pour pallier cela, la mise en place de zones exceptionnelles peut être une solution. Le lissage spatial peut également en être une autre. Ces zones serviront à corriger les incohérences du maillage administratif actuel.

Une autre possibilité est d'établir nous-même le dessin du découpage du territoire. Dans ce cas, les frontières du zonier sont basées par une étude empirique. Ce découpage pourra induire des zones larges, si nous jugeons que l'exposition au risque est similaire au sein de celles-ci. A l'inverse, il pourra induire des zones plus petites, plus précises, si nous jugeons que ces zones limitrophes ont une exposition au risque sensiblement différente.

Nous pouvons mieux justifier le contour de chaque zone avec un découpage réalisé par nos études empiriques. Toutefois, nous ne résolvons que partiellement le problème de l'écart de tarif entre deux communes limitrophes appartenant à la même zone. C'est pourquoi un lissage spatial peut être opéré sur le territoire.

I.3.3.2 Objectif et limites du zonier

Le zonier a pour objectif d'attribuer à chaque zone géographique un tarif qui correspond à son niveau de risque. La difficulté réside dans le fait que le risque est différent du taux de prime pure constaté sur la zone.

Le taux de prime pure d'une observation peut-être décomposé en trois parties :

— Une partie expliquée par les caractéristiques de l'observation, caractéristiques dont nous savons quantifier la corrélation avec le taux de prime pure.

— Une partie expliquée par les caractéristiques de l'observation, caractéristiques dont nous ne connaissons pas la relation avec le taux de prime pure. Cette partie est confondue avec la suivante dans un modèle : elle correspond au résidu.

— Une partie inexpliquée : l'aléa.

Notre objectif est de réduire la deuxième partie, celle dont nous ignorons la relation des caractéristiques avec le taux de prime pure. Plus l'historique est faible et incomplet, plus il sera difficile de réduire cette partie, donc les résidus. Un ensemble de variables explicatives faible empêchera également de réduire les résidus. A l'inverse, une mauvaise utilisation de l'information peut conduire à utiliser la troisième partie, l'aléa, comme information explicative, à tort.

Plusieurs problématiques interviennent alors : comment extraire au mieux la part inexpliquée lors de nos estimations de la part expliquée ? Devons-nous la minimiser ? Comment traiter l'inexpliqué ? En effet, est-ce réellement le hasard qui a touché cette zone ou bien est-ce un facteur inconnu réel qui influence la sinistralité de cette zone ? Cette partie inexpliquée peut être assimilée à de la variance [3].

Le zonier est le fruit d'un arbitrage entre finesse du maillage géographique et consistance de l'estimation construite pour chaque zone. Ceci correspondant à un dilemme courant du statisticien : arbitrage biais [4] et variance. La finesse du maillage géographique va permettre de réduire le sous apprentissage et donc de réduire le biais. Toutefois, cette finesse va induire un sur apprentissage et donc une importante variance. Idéalement, nous allons chercher à tarifier très finement le risque grêle car celui-ci est très localisé. Notre objectif est bien d'avoir un maillage géographique le plus fin possible, mais sous la contrainte d'une variance faible. Cette contrainte doit être maîtrisée car le possibilité d'apprendre d'une information erronée est très importante avec le risque grêle.

3. La variance est l'erreur due à la sensibilité aux petites fluctuations de l'échantillon d'apprentissage. C'est-à-dire modéliser le bruit aléatoire des données d'apprentissage plutôt que les sorties prévues.

4. Le biais statistique est la différence qui se produit entre un estimateur mathématique et sa valeur numérique, une fois qu'une analyse a été effectuée.

I.4 Sujet : Construction d'un zonier en assurance grêle sur cultures

Pour clore cette introduction, l'enjeu majeur de ce mémoire est d'estimer au mieux le risque sous-jacent propre à chaque zone géographique dans le cadre du risque grêle en agriculture. Ceci nous permettra d'associer le bon coefficient tarifaire géographique à chaque contrat.

Nous avons vu, au cours de cette introduction, plusieurs problèmes auxquels nous devons faire face :

— L'arbitrage biais/variance dans la finesse de la maille choisie : quelle maille nous permettra d'avoir une variance faible et un biais satisfaisant sur les parcelles ? Nous chercherons à apprendre un maximum de l'information à disposition tout en évitant le surapprentissage.

— L'utilisation de l'erreur de prédiction : est-ce que l'erreur d'ajustement est porteuse d'information que nos facteurs explicatifs ne captent pas ? Comment l'utiliser pour mieux estimer ?

— Comment découper un territoire pour générer un zonier optimal ?

Un enjeu de ce mémoire, dépassant le simple cadre de la grêle, est de réaliser une estimation des coefficients tarifaires dans le cadre d'un historique de données bruité.

Le bruit dans un modèle d'apprentissage est défini comme une variabilité non désirée ou inattendue dans les données d'entraînement pouvant altérer la qualité des prédictions du modèle. Le bruit peut se manifester de différentes manières, telles que des erreurs de mesure, un historique trop faible, des données manquantes, des valeurs aberrantes... Il est important de réduire le bruit dans les données d'entraînement pour améliorer la qualité de l'ajustement réalisé par le modèle.

Le bruit peut provoquer du surapprentissage (overfitting) ou un sous-apprentissage (underfitting). Le surapprentissage se produit lorsque le modèle s'adapte trop étroitement aux données d'entraînement et capture le bruit en plus de la relation sous-jacente entre les variables. Ceci peut conduire à une mauvaise généralisation des résultats. Le sous-apprentissage se produit quant à lui lorsque le modèle ne parvient pas à capturer la relation sous-jacente entre les variables.

Pour répondre à toutes ces questions, nous disposons de plusieurs jeux de données et d'outils mathématiques que nous allons décrire respectivement dans les deux parties suivantes. Une fois ces ressources mieux appréhendées, nous allons construire plusieurs zoniers grêle dans le cadre de l'assurance climatique en France. Dans la dernière partie, nous comparerons les différents zoniers construits et évoquerons les axes d'améliorations de ceux-ci.

Chapitre II Présentation des données utilisées

II.1 Données du produit Grêle de Pacifica

L'historique de données du produit Grêle, constitué grâce aux contrats souscrits par les agriculteurs auprès de Pacifica, représente pour nous une source importante et disponible d'informations. Cette ressource nous permet d'étudier le phénomène climatique sur le territoire français : son évolution dans le temps et sa représentation géographique. Nous allons décrire cette base de données, présenter ses bénéfices et limites dans le cadre de la construction d'un zonier.

II.1.1 Description de la base

Les données de sinistralité et de capitaux assurés sont issues du produit grêle. L'historique est disponible depuis 2005. Le volume annuel de contrats a crû fortement jusqu'à 2020, dernière année de notre étude. Le nombre d'exercices nous permet de bénéficier de plusieurs scénarios de sinistralité différents. Toutefois, la rareté des tempêtes de grêle nécessite un nombre d'années d'étude conséquent pour construire un zonier suffisamment robuste. Par conséquent, 16 années seront-t-elles suffisantes ?

Dans cette base, chaque observation correspond à une parcelle durant un exercice. Une parcelle va apparaître autant de fois dans la base qu'elle aura d'années de souscription au produit Grêle. Nous avons, au cours de notre étude, agréger les parcelles ayant le même numéro de contrat, la même commune et la même récolte. Ce regroupement est dû au fait que Pacifica ne dispose pas, à date, de moyen technique pour localiser précisément les parcelles lors de la tarification. De plus, notre historique est trop faible et par conséquent trop volatile pour effectuer une tarification à la parcelle.

Dans notre base d'étude, une observation correspondra donc à un regroupement de parcelles qui ont en commun le type de récolte, la commune, le contrat et l'exercice couvert.

Pour chaque observation, nous avons une charge et un montant de capitaux assurés associé. La charge correspond au montant que l'assureur a versé à l'assuré au titre d'un dédommagement suite à une tempête de grêle. Les capitaux assurés correspondent au montant maximum, hors effet franchise, pouvant être octroyés au titre d'un dédommagement d'un sinistre grêle.

Nous chercherons, comme évoqué en première partie, à ajuster un taux de prime pure pour chacune de ces observations. Nous retirons ainsi l'effet taille de l'observation, biaisant la prime pure.

Sur les deux cartes en Figure 18, nous observons que notre portefeuille de clients agricoles sur la garantie grêle est globalement bien représentatif du territoire agricole français. La Corse, ainsi que les départements d'Outre-Mer, ne sont pas ouverts à la souscription chez Pacifica : ils ne sont par conséquent pas représentés dans notre étude.

(a) Proportion de territoires agricoles par commune (Source : base Corine Land Cover)

(b) Volume de capitaux assurés par la garantie grêle cumulés durant l'historique, selon la commune (données Pacifica)

FIGURE 18 – Répartition des terres agricoles

En revanche, le portefeuille n'est pas réparti uniformément au sein du territoire. Nous notons la présence de zones faiblement représentées en capitaux assurés et d'autres très fortement représentées. Cette dissemblance pourra engendrer du bruit lors des estimations pour les zones avec peu de capitaux assurés. C'est le cas par exemple de la Bretagne, peu représentée, alors que le territoire est agricole.

Un exemple différent est celui des Hautes-Alpes, zone propice à la grêle et faiblement représentée en portefeuille. Cette zone, normalement très peu agricole, n'est pas pour autant à négliger. En effet, diverses raisons pourraient rendre cette zone propice aux arbres fruitiers ou à la viticulture par exemple.

II.1.2 Répartition des contrats au cours du temps

La courbe rose, sur la Figure 19, représente l'évolution du taux de prime du produit Grêle pure par an. Logiquement, en tant que produit climatique, nous observons une importante volatilité de la sinistralité au fil des exercices.

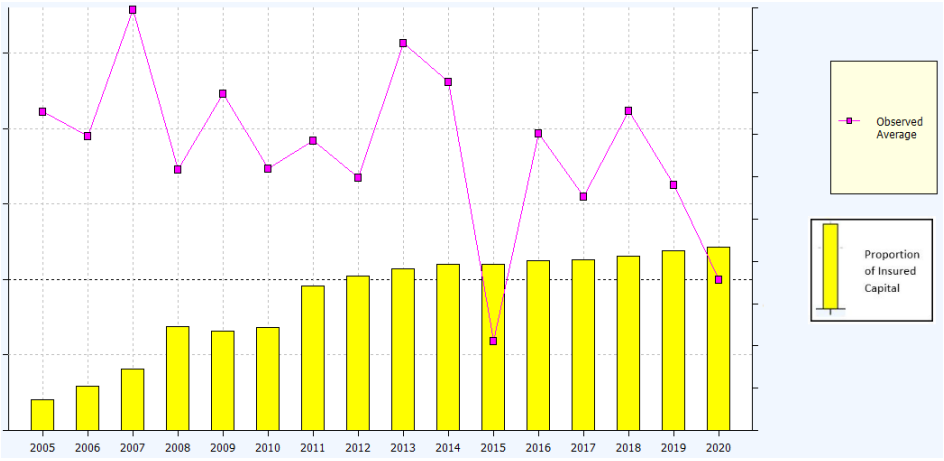
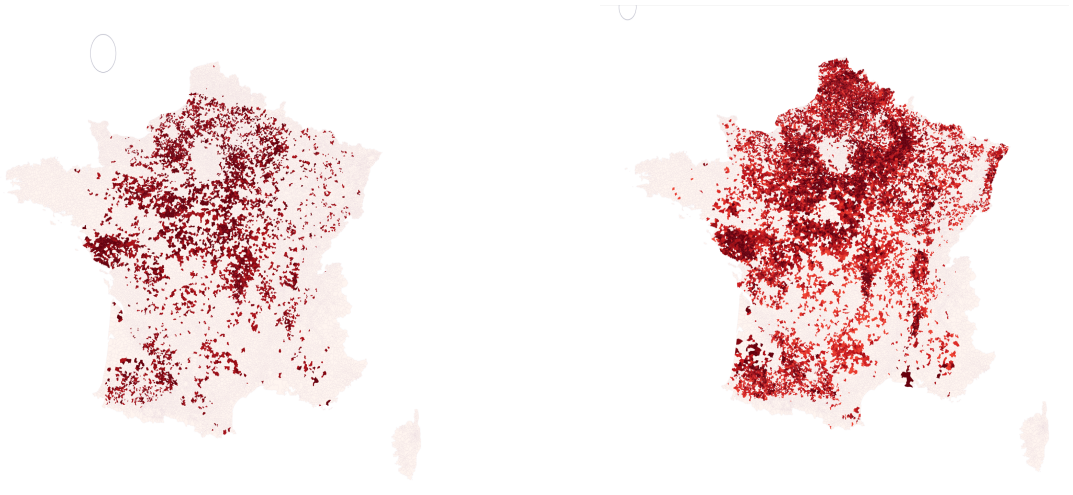


FIGURE 19 – Répartition des capitaux assurés (en jaune) et évolution du taux de prime pure (en rose) par exercice

Chez Pacifica, le produit Grêle est apparu en 2005 auprès des agriculteurs. Le portefeuille a fortement crû les cinq premières années, pour connaître une évolution moins importante la décennie suivante, visible en Figure 19 à travers les bâtons jaunes. Cette forte croissance constatée implique également une représentation géographique du portefeuille différente au cours du temps. Les deux cartes de la Figure 20 viennent montrer ceci :



(a) En 2005 (b) En 2020

FIGURE 20 – Évolution de la répartition géographique du volume de capitaux assurés sur le produit Grêle

Notre portefeuille est non uniforme au cours du temps. Les contrats sont beaucoup plus nombreux en 2020 et mieux représentatifs du territoire agricole français. Le Sud-Ouest est plus densément représenté en 2020. C'est le cas également de la vallée du Rhône, du Nord et du Grand-Est.

Des communes du portefeuille ne sont présentes qu'une ou quelques années. Certaines communes peuvent avoir été présentes en portefeuille seulement les années de fortes grêles, ou à l'inverse, n'étaient pas en portefeuille les années de forte sinistralité. Ceci engendre un important bruit car une commune peut envoyer des informations opposées selon la souscription des assurés.

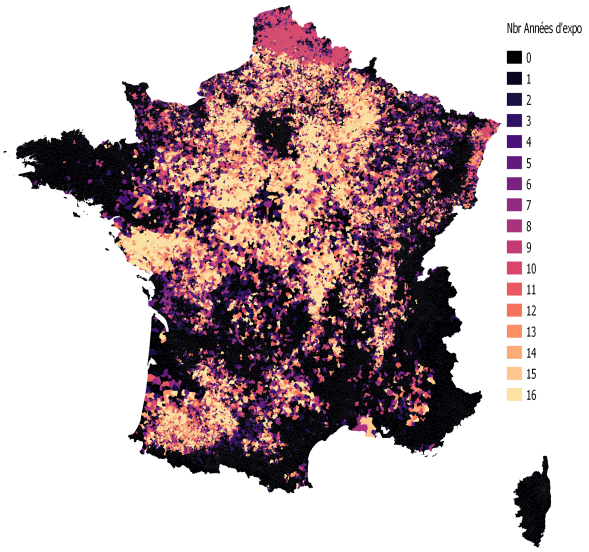


FIGURE 21 – Nombre d'années d'exposition en portefeuille par commune

La Figure 21 montre que le bassin parisien, hors Île-de-France, est très représenté durant tout l'historique, ainsi que le Sud-Ouest de la France et tout particulièrement le Gers. Ce département est jugé risqué dans les études : nous avons donc une zone sinistrée et bien représentée e à étudier. En revanche, l'historique dans le Nord-Est est faible, alors qu'il y existe des zones avec un risque grêle non négligeable. C'est le cas également du Limousin qui est faiblement exposé en terme de capitaux assurés, alors que c'est une région exposée aux tempêtes de grêle. Enfin, le Nord n'était pas représenté, jusqu'à ce que nous constations des entrées en portefeuille massives à partir de la septième année.

II.1.3 Répartition des récoltes assurées sur le territoire

Notons tout d'abord que les types de récoltes peuvent être définis à plusieurs mailles. Une première maille : *Nature de récolte*, très fine, qui est composée de 196 modalités. Une deuxième variable, *Groupe de récoltes*, plus agrégée avec onze modalités, qui est représentée en Figure 22 :

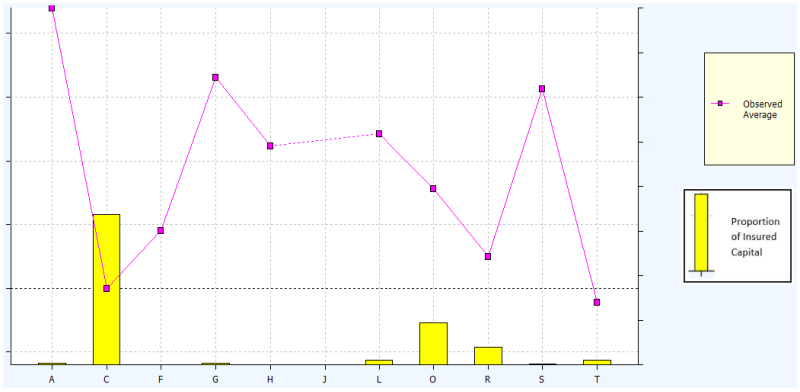


FIGURE 22 – Répartition des capitaux assurés et évolution du taux de prime pure par *groupe de récoltes*

Nous observons que certaines cultures sont fortement représentées en portefeuille, alors que d'autres, comme la classe A [5], représentent très peu de capitaux assurés.

Cette maille semble trop agrégée avec des modalités très importantes que nous pouvons diviser en plusieurs modalités comme la classe Céréales (C sur le graphique). A l'inverse, la variable *Nature de récolte* comprend un nombre très important de modalités faiblement représentées, pouvant engendrer une explosion de la variance. C'est pourquoi, nous avons créé la variable *Classe de récoltes*. Une variable au maillage intermédiaire représentée en Figure 23 :

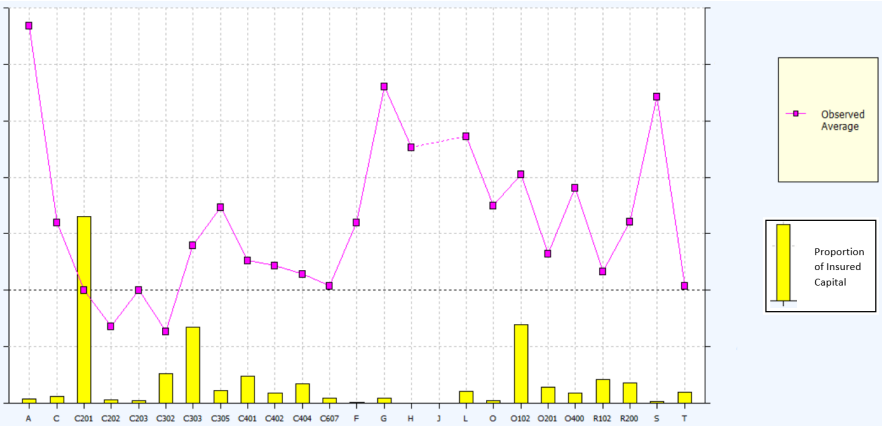


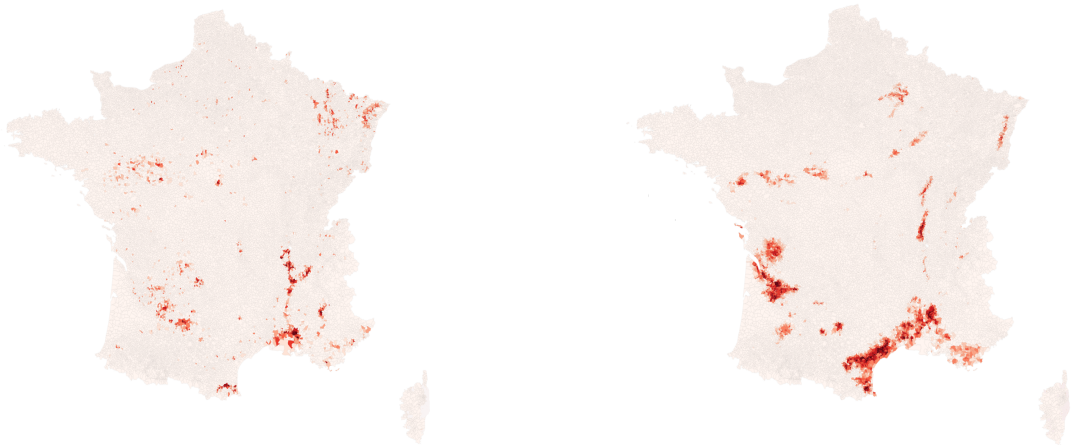
FIGURE 23 – Répartition des capitaux assurés et évolution du taux de prime pure par *classe de récoltes*

5. classe A : Arboriculture

Les capitaux assurés sont répartis sur un plus grand nombre de modalités. Pour chacune des récoltes, les différences de sensibilités à la grêle sont donc mieux captées. Le problème de volatilité, lié à un trop grand nombre de modalités faiblement représentées, est également limité.

Nous remarquons sur les deux précédentes figures, Figure 22 et Figure 23, que la courbe rose est en dent de scie, signe d'une grande variabilité des taux de prime pure entre les types de récoltes. Un pic est notamment visible pour les arboricultures, cultures qui semblent très sensibles à la grêle. Cette partie contribue à nous montrer l'importance de capter la sensibilité d'un type de récolte face aux tempêtes de grêle.

Certaines récoltes sont représentées seulement sur quelques zones restreintes du territoire, induisant des difficultés à appréhender l'information renvoyée par celles-ci. Par exemple, les cultures d'arbres fruitiers sont très localisées : il peut être compliqué de distinguer le risque géographique du risque de la sensibilité de la culture à la grêle.



(a) Cultures arboricoles (b) Terres viticoles

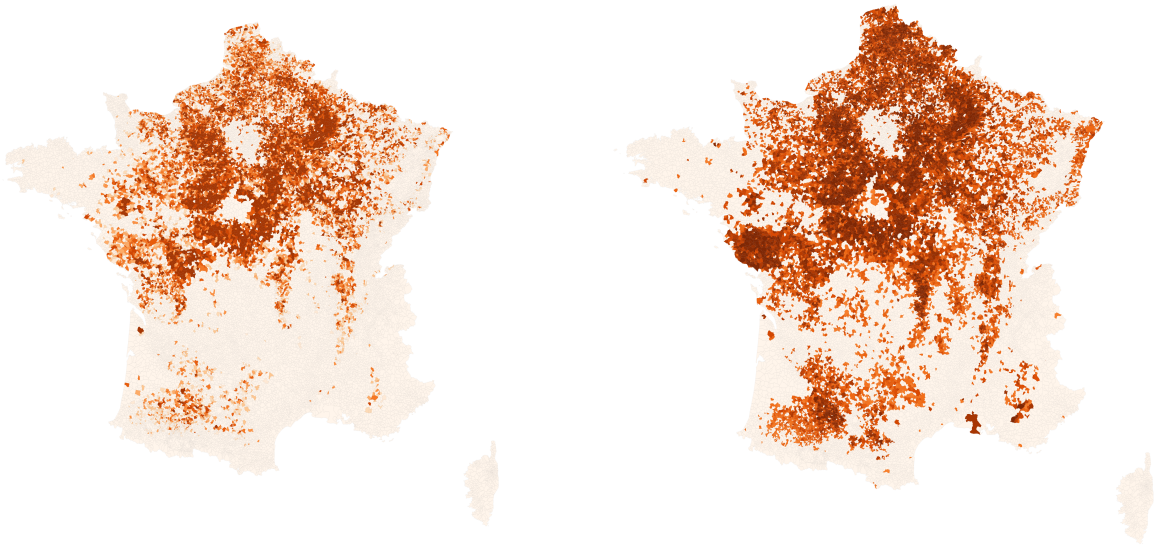
FIGURE 24 – Proportion de surface assurée pour différents groupes de récoltes

Le biais géographique de la localisation des récoltes est bien visible sur la Figure 24. Sur les deux cartes, ces cultures apparaissent en grande proportion sur certaines communes, en termes de capitaux assurés. C'est la cas par exemple des cultures viticoles : elles sont très concentrées dans la vallée du Rhône. Dans cette zone, il sera difficile de différencier si la sinistralité est liée à la zone ou à la culture. En effet, ces cultures sont peu représentées en nombre de communes et les communes représentées ont très peu d'autres cultures assurées pouvant servir de comparaison. Ce problème est visible également à travers les cultures viticoles de Gironde : plusieurs communes de cette zone sont représentées à plus de 80% par des vignes.

Les modèles d'ajustement calculent des coefficients de risque pour chaque groupe de récoltes. Toutefois, pour lutter contre la volatilité de certains coefficients, l'avis d'un expert agronome pourra venir confirmer nos hypothèses et gommer le biais potentiel évoqué.

Pour la vigne, qui est une culture avec un risque qui lui est propre et une localisation restreinte des capitaux assurés sur le territoire, nous allons construire un modèle à part. Cela afin d'affiner au mieux les estimations de la sensibilité de la vigne à la grêle. Ce modèle ne sera pas évoqué dans notre analyse car similaire, dans sa construction, au zonier global. Notre estimation portera donc sur toutes les cultures à l'exception de celles viticoles.

Par ailleurs, deux natures de récoltes constituent une grande partie du portefeuille : le blé d'hiver et le colza d'hiver. L'exposition de ces deux natures de récoltes est représentée géographiquement dans la Figure 25.



(a) Colza d'hiver (b) Blé d'hiver

FIGURE 25 – Représentation géographique des volumes de capitaux assurés par commune

Le colza d'hiver est principalement présent dans des zones où la sinistralité grêle est faible. Cette récolte est présente dans la moitié Nord du territoire français, comme le montre la localisation des communes orangées sur la Figure 25a. Cette partie du territoire est moins fréquemment touchée. En revanche, le blé d'hiver est représenté sur une plus grande partie du territoire, au sein de zones aussi bien fortement que faiblement sinistrées par la grêle. Le blé d'hiver sera un socle de nos estimations car cette récolte aide à comparer toute chose égale par ailleurs le taux de prime pure d'une commune. Ce n'est pas le cas du colza car le modèle peut avoir des difficultés à savoir si la faible sinistralité de la récolte est due à sa résistance à la grêle ou à sa localisation géographique.

II.1.4 Sinistralité des assurés

II.1.4.1 Taux de prime pure par commune

Précédemment, nous avons remarqué que beaucoup de communes, voire de zones, du territoire français ne possédaient que quelques, voire aucune, années d'historique au sein du portefeuille. Ceci implique une volatilité importante du taux de prime pure, caractérisée en Figure 26 :

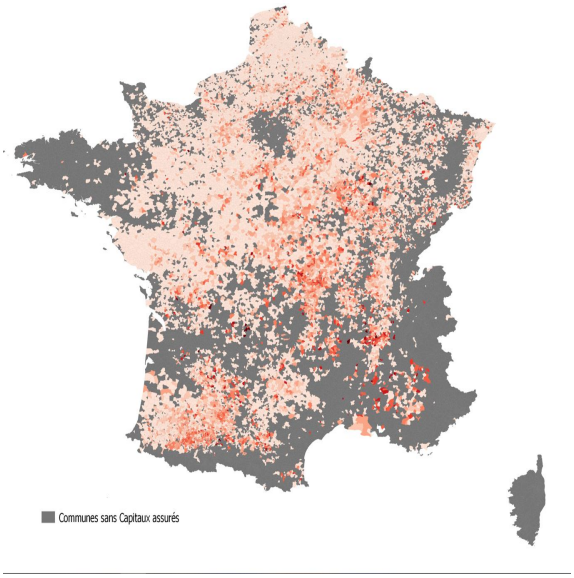


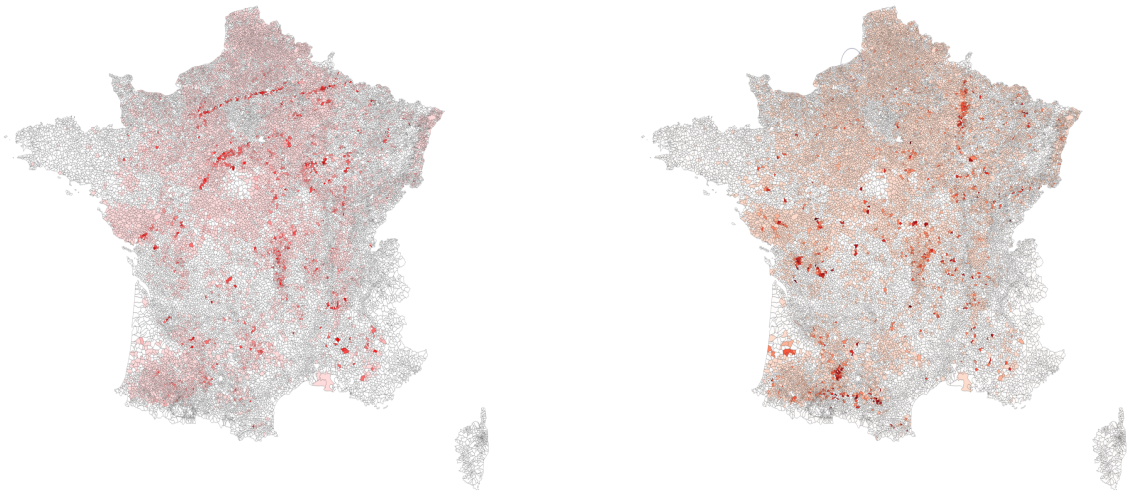
FIGURE 26 – Taux de prime pure par commune constaté entre 2005 et 2020

La zone à l'intérieur des terres, partant de la base des Pyrénées jusqu'à la Bourgogne et comprenant le contour alpestre, est plus sinistrée. A l'inverse, la partie Nord française, et surtout le quart Nord-Ouest, n'est quasiment pas sinistrée. Nous retrouvons pour les régions, voire ensembles de régions, les mêmes conclusions qu'en première partie quant à leur exposition à la grêle.

Nous observons de la volatilité parmi les territoires sinistrés. Le Sud-Ouest, suffisamment représenté sur l'historique, semble avoir une volatilité plus limitée du taux de prime pure entre communes, et ce malgré sa forte sinistralité. Ce n'est en revanche pas le cas du contour alpestre, très faiblement représenté, avec des communes très sinistrées et d'autres, voisines, non sinistrées. Les départements de l'Allier, de la Loire et de la Haute-Loire ont également ce problème de volatilité. Nous aurons donc une attention toute particulière sur ces zones dans notre étude. Nous essaierons de limiter la variance au mieux, pour éviter tout apprentissage à partir d'un quelconque bruit.

II.1.4.2 Taux de prime pure au cours du temps

Nous avons observé sur la Figure 19 [6], que les variations du taux de prime pure du portefeuille sont importantes d'un exercice à l'autre. Notamment, nous constatons un creux important de la courbe rose lors de l'exercice 2015, correspondant à un très faible taux de prime pure.
La volatilité du taux de prime par commune est très évidente lorsque nous analysons ce dernier par exercice. L'impact d'une tempête y est fortement visible. Voici, en Figure 27, deux exercices 2013 et 2018 permettant d'illustrer les différents taux de prime pure sur le territoire français :



(a) 2013 (b) 2018

FIGURE 27 – Taux de prime pure sur le territoire français selon certains exercices

Des zones sont très sévèrement touchées sur un des deux exercices et totalement épargnées lors de l'autre. C'est le cas du Nord du Centre Val de Loire par exemple, épargné en 2018. Nous remarquons cependant que la Marne, terre de Champagne, est touchée lors des deux exercices, symbolisant que les tempêtes ne sont pas seulement des phénomènes rares.

II.1.4.3 Exposition et taux de prime pure selon d'autres variables

Une autre variable importante à l'élaboration du tarif est la franchise. Le taux de prime pure et l'exposition en capitaux assurés selon la franchise du contrat apparaissent sur la Figure 28 :

6. page 50

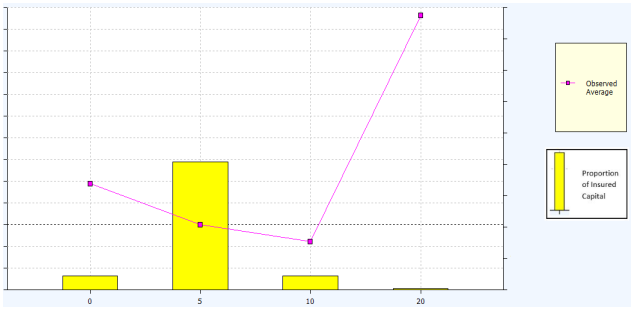


FIGURE 28 – Répartition des capitaux assurés et évolution du taux de prime pure par *franchise*

La franchise est représentée très majoritairement sur une seule modalité : 5%. Nous remarquons que le taux de prime est bien décroissant de la franchise de 0% à 10%. A noter que la franchise 20% est appliquée exclusivement aux cultures arboricoles, groupe de récoltes très sinistrés. Ceci explique l'important taux de prime pure associé à ce niveau de franchise.

Nous avons enfin créé une variable regroupant des contrats selon leur volume de capitaux assurés. Ceci afin de voir si la sinistralité pouvait différer selon le volume de la récolte.

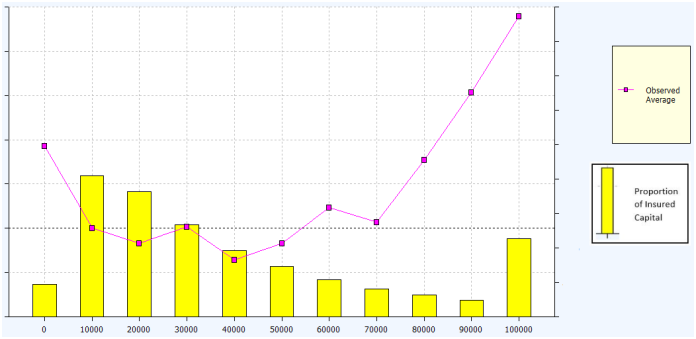


FIGURE 29 – Répartition des capitaux assurés et évolution du taux de prime pure selon le volume assuré

La Figure 29 fait apparaître une courbe rose croissante, indiquant une plus forte sinistralité pour les contrats avec de forts capitaux assurés. Il sera intéressant de vérifier cette corrélation lors de la construction des modèles d'ajustement. Cette variable, non géographique, n'interviendrait pas dans l'élaboration du zonier, mais pourrait expliquer une partie du taux de prime pure.

L'historique de données que constitue le portefeuille de clients agricoles Grêle chez Pacifica nous permet de mieux comprendre l'impact de la grêle en France car nous retrouvons bien les zones grêligènes et celles très faiblement sinistrées. Toutefois, nous avons remarqué que l'historique d'étude est incomplet géographiquement et dans le temps. Ce constat nous oblige à recourir à des outils mathématiques et des données complémentaires afin de mieux appréhender la grêle sur ces zones volatiles. Nous allons à présent décrire ces données complémentaires.

II.2 Données complémentaires

II.2.1 Bases de données internes

D'autres produits que celui étudié peuvent nous fournir une information sur la sinistralité consécutive à la grêle. C'est le cas des produits Automobile et Habitation, et également du produit Multirisques Climatiques en assurance agricole. Chacun de ces produits couvre les assurés contre le risque grêle et ont un volume de données très important.

Nous ne retiendrons toutefois pas le produit Automobile car la localisation des sinistres grêle subis par les assurés demeure inconnue dans nos bases. En effet, le lieu du sinistre n'est pas forcément renseigné, contrairement au lieu de souscription. L'information du produit Automobile serait alors trop biaisée pour être retenue. Il reste, dès lors, deux produits candidats : Habitation et Multirisques Climatiques.

II.2.1.1 Produit Habitation en assurance des particuliers

a) Description du produit

Le produit Habitation propose différentes garanties :

— Les dommages aux biens : incendie, dégâts des eaux, dommages électriques, gel, événements climatiques, attentats, catastrophes naturelles, vandalisme, vol, bris de glace.

— La protection des personnes : protection corporelle, garantie scolaire.

— La responsabilité civile : vie privée, risque locatif, propriétaire d'immeuble.

Le risque grêle est donc bien couvert dans ce produit, parmi la catégorie des dommages aux biens.

Le produit Habitation en assurance des particuliers concerne un vaste marché : plus de 35 millions de logements en France métropolitaine. Concernant Pacifica, le produit représente environ quatre millions de clients pour un milliard d'euros de cotisations acquises. Une très forte concurrence existe sur ce marché : que ce soit du fait de la présence de nombreux acteurs, ou du fait de la "loi Hamon" [7]. Une tarification précise constitue donc un enjeu important, et pour cela des données nombreuses et de qualité sont nécessaires.

L'important volume d'information du produit Habitation et la qualité de ce volume vont apporter une meilleure connaissance de la sinistralité grêle sur le territoire français. Ce produit est représenté sur l'ensemble sur le territoire, permettant de bénéficier d'une information complète.

7. Loi Hamon (2015) : un client a la possibilité de résilier son contrat Habitation à tout moment, une fois la première année d'engagement écoulée

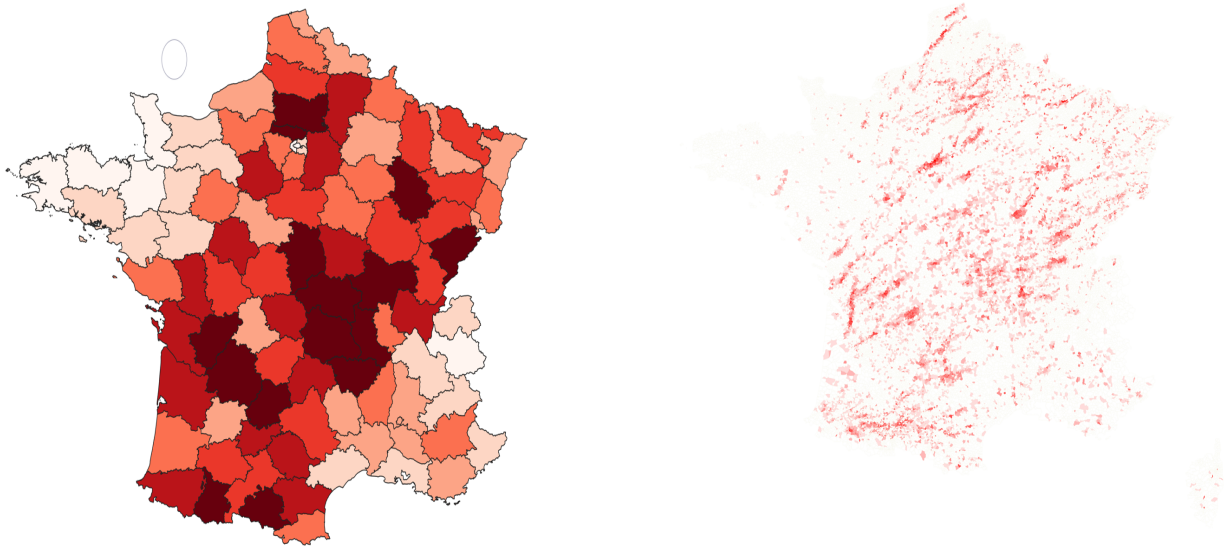
b) La fréquence des sinistres

Nous calculons la fréquence des sinistres par commune pour ce produit, selon le nombre d'années de souscription constaté. Nous prendrons nos données entre 2005 et 2020. La fréquence sera notre indicateur car il est simple d'interprétation et d'usage. L'utilisation de la prime pure nécessiterait d'autres retraitements pour rendre les communes comparables entre elles.

Ce produit met en évidence les grosses tempêtes de grêle car l'habitat est plus résistant que les cultures face au risque grêle. Les vitres et les toitures sont sensibles aux grêlons de plus de 25 millimètres environ, alors que les cultures le sont à partir de 15 millimètres. Le portefeuille Habitation est mieux représenté à la fois sur l'historique et mieux représenté sur le territoire. Il nous permet ainsi de compléter nos connaissances sur la sinistralité des communes faiblement exposées sur le produit Grêle.

En revanche, à travers cette variable, les tempêtes de moindre gravité, celles détruisant en partie une parcelle, ne sont pas mises en évidence.

La fréquence peut être calculée à différentes mailles, départementale ou communale par exemple. Nous avons retenu un maillage par commune du fait de la bonne représentation du produit à la fois sur l'historique et sur le territoire.



(a) Par département (b) Par commune

FIGURE 30 – Fréquence de sinistralité sur le produit Habitation des particuliers de 2005 à 2020

Sur la carte départementale, Figure 30a, nous retrouvons bien les zones sensibles à la grêle vues précédemment. Toutefois, nous voyons en comparant avec la Figures 30b, que la maille départementale agrège trop la fréquence, générant une perte d'information importante.

Les fortes tempêtes, constatées au cours de ces 16 dernières années, sont reconnaissables sur la carte communale, ainsi que leur localisation précise. Nous observons, par exemple en Haute-Marne ou dans l'Oise, des zones peu grêligènes ainsi que zones sinistrées s'apparentant à l'empreinte laissée par une tempête. L'inconvénient de cette variable est qu'elle va parfois surpondérer les zones, peu risquées théoriquement, mais ayant subi une grosse tempête durant les 16 ans.

II.2.1.2 Produit Multirisques Climatiques en assurance agricole

Nous avons également à disposition les données du produit Multirisques Climatiques venant compléter l'information offerte par le produit Habitation.

a) Description du produit

Le contrat Assurance Récolte permet de garantir un rendement minimum pour les cultures végétales non fourragères sur pied, pendantes ou en andain, en cas de survenance d'un dommage garanti occasionné par un ou plusieurs évènements climatiques.

Le champ des événements garantis est fixé par les pouvoirs publics, afin que le contrat puisse répondre à l'éligibilité aux subventions. Il est commun à l'ensemble des assureurs. Pour le cas de Pacifica, ce produit couvre un plus grand nombre d'évènements que le produit Grêle. En effet, il couvre environ une dizaine d'évènements dont principalement les sécheresses, les orages, les grêles, les gels, les inondations, les excès d'eau et les tempêtes.

Le produit Multirisques Climatiques couvre différents groupes de cultures dont principalement :

— Céréales, oléagineux.

— Cultures légumières de plein champ.

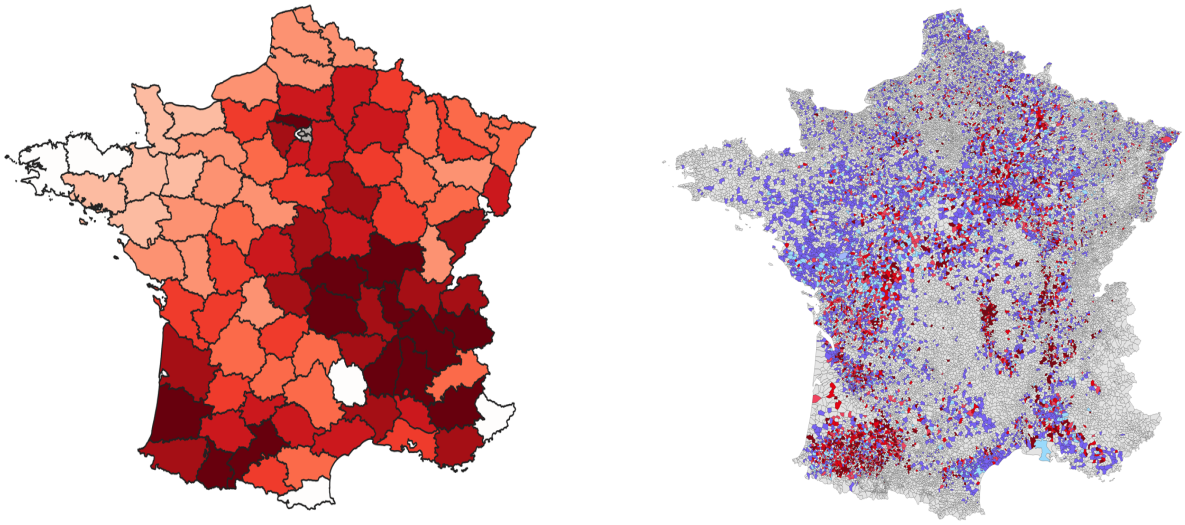
— Cultures à racines et tubercules.

— Vignes et cultures arboricoles.

Étant donné que le produit Multirisques Climatiques couvre des récoltes communes au produit Grêle, la sensibilité aux risques climatiques est donc comparable entre les deux produits, ce qui n'était pas le cas pour le produit Habitation.

b) La fréquence des sinistres

Contrairement au produit Habitation, ce produit n'est pas bien représenté sur le territoire, comme le montre la Figure 31b avec les zones grises, notées -1 en légende, correspondant aux communes sans contrat durant ces 16 dernières années.



(a) Par département

(b) Par commune

FIGURE 31 – Fréquence de sinistralité sur le produit Multirisques Climatiques en assurance agricole de 2005 à 2020

Compte tenu d'un nombre très important de communes absentes en portefeuille, nous pourrions retenir, par simplicité, la variable agrégée à la maille départementale. Celle-ci est complète comme il est visible en Figure 31a. Au final, nous utilisons une fréquence communale que nous complétons en remplaçant les données manquantes par la fréquence départementale. La méthode reste biaisée, mais permet d'ajouter de l'information sur les communes aux fréquences inconnues, sans modifier l'information sur les communes présentes en portefeuille.

Cette variable permettra de diffuser notre connaissance de la grêle à des codes INSEE agricoles peu ou non représentés dans le portefeuille.

II.2.2 Bases de données externes en accès libre

Nous allons ici construire un jeu de variables explicatives par le biais de sources externes à Pacifica. Nous avons vu en première partie que certaines caractéristiques d'une zone, comme l'altitude ou les variations de températures constatées sur celle-ci, étaient corrélées au risque grêle. Ces constats sur le phénomène vont orienter nos recherches. Certaines variables s'avéreront pertinentes et seront conservées dans les modèles. En revanche, d'autres n'apporteront pas d'informations pertinentes et seront logiquement retirées.

Nous utiliserons, pour construire nos estimations, des modèles linéaires généralisés via le logiciel Emblem. Celui-ci nécessite d'avoir un nombre maximal de 250 modalités pour chaque variable. Les variables continues ne sont donc pas utilisables : il faut créer, pour chacune d'elles, des classes de modalités. De plus, par soucis de volatilité, nous créerons nettement moins de 250 modalités afin qu'elles soient suffisamment représentées en terme de capitaux assurés. Pour effectuer ces regroupements, plusieurs méthodes sont possibles. Nous avons notamment utilisé le package R "quand.cut" qui va découper une variable quantitative en une variable qualitative et créer des modalités aux effectifs similaires. Pour découper, nous regarderons également le niveau de capitaux assurés par modalités, plutôt que les effectifs, car ceux-ci étant plus représentatifs du portefeuille.

II.2.2.1 Données altitude

Nous avons représenté la variable altitude, à partir des données de l'Insee, sur la Figure 32 :

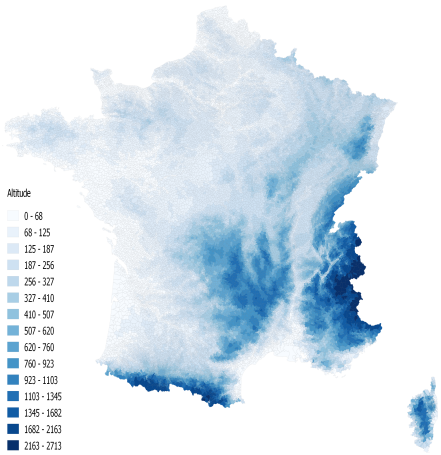


FIGURE 32 – Altitude par code Insee

Les zones grêligènes se trouvent bien en moyenne montagne comme dans l'Ouest du Massif central ou autour des Alpes et des Pyrénées. Toutefois, des zones comme la région toulousaine, sont à de basses altitudes, bien que grêligènes. Par conséquent, d'autres variables sont à analyser.

Il s'avère, que pour le cas de l'altitude, l'utilisation de la méthode des quantiles pour créer des classes n'est pas judicieuse. En effet, les basses valeurs, en-dessous de 125 mètres d'altitude, sont très représentées sur le territoire et correspondent à des zones peu grêligènes. Il est plus intéressant de différencier les communes de plus de 400 et 500 mètres d'altitude, plutôt que de différencier les communes de 15 et 18 mètres d'altitude. Un regroupement en quantile différencierait seulement le second cas et regrouperait le premier. Nous avons donc décider de créer 50 modalités s'adaptant au mieux à nos besoins.

II.2.2.2 Données météo

Beaucoup de données météorologiques sont disponibles sur Internet : certaines sont gratuites, d'autres payantes, l'historique peut être plus ou moins long. Le nombre de stations météo y figurant fluctue d'une source à l'autre. Enfin, la qualité des relevés et la structure de la base peuvent être plus ou moins fiables.

Dans cette étude, nous avons fait le choix d'utiliser des données de l'association Infoclimat qui répertorie une grande quantité de données météo sur le site internet **infoclimat**. Cette association a plus de vingt ans d'existence. Elle collecte et valorise plus de six milliards de données. Ce site permet de combiner les données météo France historiques disponibles et des données issues de leurs propres stations météorologiques. Les relevés ont une ancienneté variant de plusieurs dizaines d'années à quelques années, selon les stations. Ces relevés vont nous permettre d'affiner notre connaissance météo sur le territoire français.

Nous retrouvons sur ce site des variables potentiellement explicatives de la grêle comme la vitesse des vents, la pluviométrie, les températures minimales, maximales ou moyennes mesurées sur chaque station. Une observation de cette base correspond à un relevé mensuel sur une station. Pour une observation, plusieurs informations sont complétées. Certaines variables sont manquantes selon la période et la station. Enfin, d'autres indicateurs peuvent être construits à partir de ceux disponibles et il est possible d'effectuer diverses statistiques à partir des relevés.

La Figure 33 représente la carte des stations météorologiques qui génèrent les relevés de la base construite par l'association :

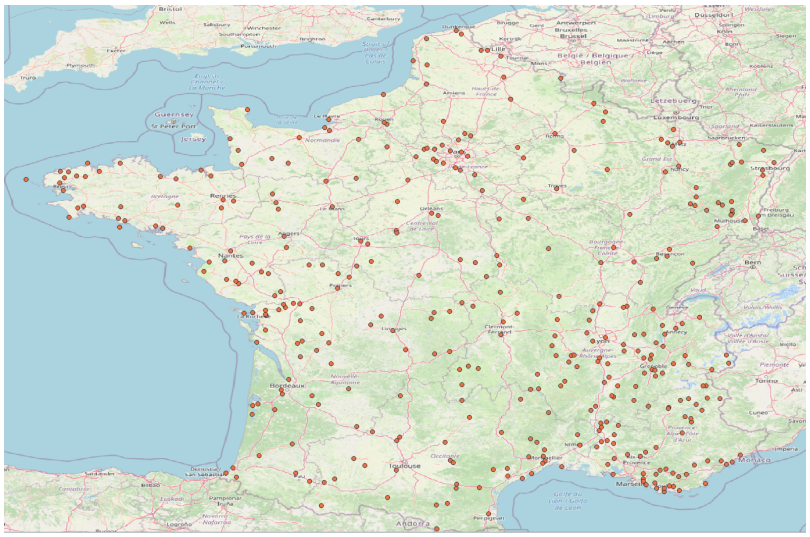


FIGURE 33 – Localisation des stations Infoclimat sur le territoire français

Seulement une petite partie des communes possède une station. Nous observons cependant que la répartition sur le territoire de ces stations est assez uniforme, à l'exception des stations côtières très représentées. Cette base de données, combinée à l'utilisation d'outils mathématiques de propagation, est en mesure de nous offrir des informations sur l'ensemble du territoire. Un biais d'estimation subsistera pour les communes montagneuses, là où les températures varient plus facilement d'une commune à l'autre.

Comme évoqué, l'ancienneté des relevés peut se compter en dizaines d'années. Compte tenu de la durée de notre historique, nous retiendrons seulement les 16 dernières années. Nous pourrons créer des variables selon divers niveaux d'ancienneté : entre 3 et 15 ans. Le premier cas s'adapte mieux au changement climatique, cependant il demeure plus volatile car possède moins d'historique.

Les variables disponibles dans la base de données Infoclimat sont décrites précisément en annexe A. Certaines sont incomplètes ou peu renseignées. Beaucoup de variables sont très corrélées, elles offrent donc une information redondante. Nous allons décrire plus en détail, dans les prochaines parties, le choix des variables, leur corrélation avec la grêle et notre façon de rendre disponible l'information de quelques points à tout l'ensemble du territoire.

II.2.2.3 Données d'occupation des sols

La base CORINE Land Cover est une base de données s'inscrivant dans le cadre du programme européen CORINE, de COoRdination de l'InformatioN sur l'Environnement concernant 38 états. En France, le Service de l'Observation et des Statistiques (SOeS), au sein du Meeddat [8], est chargé d'en assurer la production, la maintenance et la diffusion. Nous utilisons dans notre cas le millésime 2018. D'autres comme 1990, 2000, 2006 et 2012 sont disponibles, nous retenons le plus récent.

La base que nous utilisons recense, pour chaque commune, la surface occupée pour chaque type de sols. Il existe plusieurs granularités différentes de types de sols. Il est notamment répertorié, dans de cette base, cinq grands types d'occupation du territoire : les zones artificialisées, les zones agricoles, les forêts, les zones humides et les surfaces en eau. La granularité peut-être beaucoup plus fine [9] et il est possible de distinguer de quel type de forêt est composée la commune.

8. MEEDDAT = Ministère de l'Ecologie, de l'Energie, du Développement Durable et de l'Aménagement du Territoire

9. La nomenclature de la base est disponible sur le site www.statistiques.developpement-durable.gouv.fr

II.3 Bases de données privées recensant les tempêtes de grêle

Le recensement des tempêtes de grêle peut se faire par l'intermédiaire de plusieurs méthodes : relevé manuel, relevé radar ou encore relevé d'articles de presse. Les différentes techniques comportent des inconvénients. Chacune d'elles a, par exemple, un manque d'exhaustivité. A noter que, la technique par radar est améliorée continuellement. Le retard technologique des détections et relevés est principalement dû au faible enjeu économique et au faible enjeu préventif, en comparaison à d'autres phénomènes climatiques. Il est en effet difficile de se prémunir contre le risque grêle.

Nous allons faire ici un inventaire succinct des bases existantes en y dressant les avantages et inconvénients. A noter que l'association Infoclimat, présentée pour les données météorologiques précédentes, ne fournit pas de données grêle. Météo France n'offre également pas de solutions satisfaisantes et abordables car celles-ci sont très coûteuses, de l'ordre de plusieurs dizaines de milliers d'euros.

II.3.1 Les différentes sources disponibles

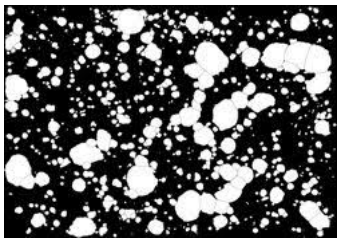
a) Anelfa

L'Anelfa [10] est une association fondée en 1951 à l'initiative d'un groupe d'agriculteurs, d'agronomes, de physiciens et d'élus. Ils profitent donc d'études et de connaissances acquises sur 70 ans. L'association est composée de plusieurs salariés. Leurs objectifs sont de développer les recherches scientifiques dans le domaine de la physique des nuages et de perfectionner une méthode de traitement des orages afin d'améliorer la prévention. Par leurs travaux, ils ont pu constituer plusieurs bases de données que nous allons décrire ici.

Les relevés de ces bases sont effectués par le biais d'un grêlimètre. C'est un appareil simple mis au point par des chercheurs canadiens. Il permet d'enregistrer la trace des impacts de grêlons arrivant au sol. Il est constitué d'un piquet métallique de 1,50 m de haut supportant une tôle horizontale sur laquelle est disposée une plaque de polystyrène extrudée de 30 cm x 40 cm. Un grêlimètre est illustré en Figure 34a :

10. http://www.anelfa.asso.fr

(a) Grêlimètre



(b) Scan d'un grêlimètre suite à une tempête de grêle

Après la chute de grêle, les impacts de grêlons sont rendus visibles par un encrage noir de la plaque, visible sur la Figure 34b .

Les mesures ne sont effectuées que dans les départements partenaires : une large partie du sud-ouest, le sud et le centre de la France. De plus, les grêlimètres ne pouvant pas être installés partout sur le territoire, les relevés ne correspondent qu'à un nombre limité de points sur l'espace étudié. Ces zones d'études sont visibles en orange sur la Figure 35 :

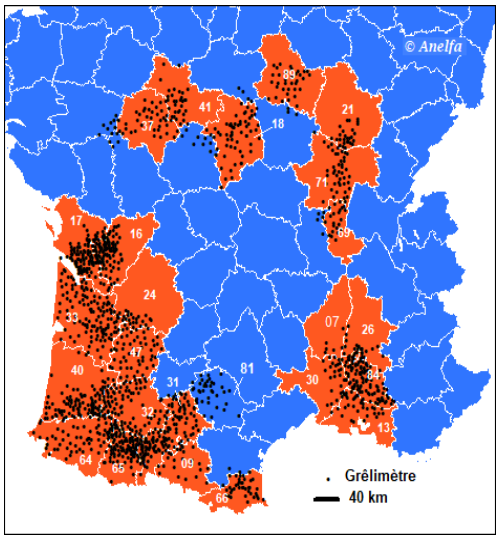


FIGURE 35 – Zones grêligènes selon l'Anelfa

Les points noirs correspondent aux grêlimètres installés. La densité est suffisante dans le Sud-Ouest, alors que cette dernière est nulle sur une grande partie des départements français.

Cette base de données contient un premier axe intéressant : elle semble offrir un jeu de données très détaillé pour les départements les plus touchés par la grêle. De plus, elle permet d'offrir un relevé fidèle à la réalité. Toutefois, elle a deux problèmes qui nous freineront grandement : ces données sont privées et hétérogènes, empêchant une modélisation de qualité.

b) Keraunos

Keraunos [11] est l'observatoire français des tornades et orages violents. Fondée en 2006, cette organisation regroupe des compétences spécialisées dans le domaine des risques météorologiques. Les activités de Keraunos sont essentiellement concentrées dans les domaines de la prévision, de l'alerte, de la mesure du risque et de l'expertise. Parmi les phénomènes météorologiques qui constituent leur périmètre de spécialité figurent notamment la grêle, les tornades, la foudre, les rafales de vent et les pluies diluviennes. C'est à partir de ces bases que Keraunos fournit un ensemble de services dédiés aux entreprises de tous les secteurs d'activité (agriculture, assurance, industrie,...), ainsi qu'aux collectivités.

Concernant la base de données grêle : plus de 70 000 chutes de grêle y sont recensées. Celles-ci sont expertisées par une procédure multi-sources exclusive, dont l'essentiel couvre la période de 2006 à nos jours. Cette base de données est actualisée quotidiennement et plus de 5 000 entrées supplémentaires y sont ajoutées chaque année. Elle compte également des cas des XIX^{e} et XX^{e} siècles, afin d'assurer une climatologie sur une longue période. Les informations contenues dans cette base de données incluent la localisation des chutes de grêle à la commune près, ainsi que les diamètres des grêlons observés.

Les bases de données de Keraunos sont alimentées par deux moyens :

— D'une part, une veille permanente est assurée afin de recenser et de référencer tous les phénomènes météorologiques sévères en temps réel. Ces phénomènes sont vérifiés et validés par recoupement de sources diverses, pouvant provenir de la presse ou d'expertises directes sur site.

— D'autre part, une activité spécifique de recherches historiques est assurée afin de compléter les phénomènes actuels par des archives de cas passés. Ceci permet de reconstituer une climatologie des phénomènes météorologiques violents en France sur plusieurs siècles.

L'ensemble constitue une base de données particulièrement large, entièrement expertisée, qui fait référence en France dans le domaine des phénomènes météorologiques sévères en général, et orageux en particulier.

c) CatNat.net

CATastrophesNATurelles.net [12] est un observatoire permanent des catastrophes naturelles. Il est le premier site francophone de veille permanente sur l'actualité internationale des catastrophes na-

11. https://www.keraunos.org
12. https://www.catnat.net

turelles, ainsi que de la gestion des risques naturels et des changements climatiques. CatNat.net est un site géré par Ubyrisk Consultants : structure spécialisée dans l'expertise et le conseil en risques naturels.

Le catalogue d'événements est entretenu quotidiennement depuis janvier 2001. Afin d'exploiter statistiquement ce catalogue, l'ensemble des informations sur les événements a été retranscrit sous la forme de bases de données comme : la BD CatNat, BD Grêle France, BD Grêle Europe, Base arrêtés Cat' Nat'...

La base de données BD Grêle France qui nous intéresse est décrite sur leur site internet. Elle combine des informations sur près de 200 000 tempêtes de grêle depuis 2000. Les tempêtes correspondent aux chutes de grêle supérieures ou égales à un centimètre. Cette base de données, la plus exhaustive en la matière, a été constituée à partir des détections réalisées par leur outil Grêle Warning France. Celui-ci se base sur l'analyse d'image radar Doppler à haute résolution, croisée avec l'utilisation d'un modèle atmosphérique simulant les conditions des basses couches (températures, humidité...). Ce croisement permet de caractériser le diamètre des grêlons au sol. Le traitement des données, via un algorithme spécialement développé en interne, a fait l'objet de vérification quant à sa robustesse lors de son développement, notamment grâce à des grêlimètres.

Les relevés de chutes de grêle ont été effectués de manière uniforme sur l'ensemble du territoire depuis fin 2005. Pour la période précédente, les événements ont été reconstitués à partir de relevés météorologiques de stations disposant directement de grêlimètre, de données provenant de grêlimètres agricoles et aussi à partir de revues de presse.

Il est possible d'acheter l'historique à date ou de souscrire à un abonnement permettant de recevoir régulièrement la base actualisée.

II.3.2 La base de données retenue

Nous avons fait le choix de retenir la base de données issue du site CatNat.net. La base CatNat dispose d'un nombre plus important de tempêtes de grêle et utilise une technologie qui nous semblait plus avancée pour recenser ces tempêtes.

a) Descriptif de la base de données utilisée

Dans cette base de données recensant un historique de grêle sur le territoire français, une observation correspond à une commune sinistrée par une tempête de grêle à une date définie.

La base de données comprend quelques spécificités concernant la méthode de recueil. Leur technologie n'a été mise en évidence seulement en 2005 et l'évolution forte de celle-ci ces dernières années rend les données hétérogènes au cours du temps. Depuis 2020 et le dernier bond technologique, davantage de tempêtes sont détectées grâce à une résolution plus fine de l'imagerie radar. Cette évolution permet de détecter un nombre plus important de faibles tempêtes, soit celles générant des grêlons de petite taille. Sur l'historique, les tempêtes de forte intensité semblent assez bien détectées. Toutefois, comme le symbolisent les bonds technologiques successifs, cet outil est en amélioration et remet ainsi en cause l'exhaustivité des premières années. Depuis 2020 et le dernier bond, le nombre de tempêtes détectées a fortement augmenté et CatNat.net semble tendre vers l'exhaustivité.

La base de données est composée de toutes les variables utiles pour connaître la localisation des tempêtes comme : le code Insee, le nom de la commune, la date. Elle contient également des variables pertinentes afin de mieux connaître l'historique de sinistralité grêle, comme des variables relatives à la taille des grêlons recensés lors de la tempête (classe de diamètre, diamètre minimal et maximal du grêlon lors de la tempête...). Il est possible de retrouver en Annexe B un extrait de cette base.

b) Création d'indicateurs

Des retraitements sont nécessaires pour rendre cette base exploitable. En effet, cette hétérogénité au fil du temps demeure un problème. Un plus grand nombre de tempêtes est détecté en 2020 suite au bond technologique. En effet, beaucoup de tempêtes survenues en 2020, figurant dans la base grâce au bond technologique, n'auraient pas été détectées auparavant. Les communes touchées par ces tempêtes sont alors surpondérées dans la base.

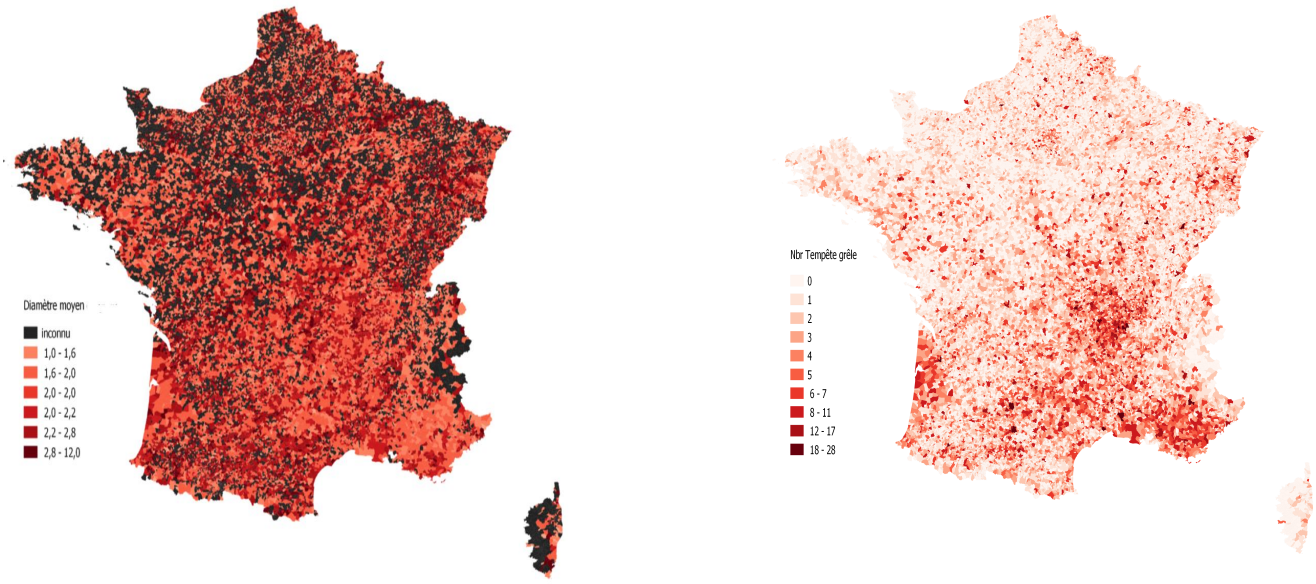
Différentes variables ont été créées sur l'ensemble de l'historique : diamètre minimal moyen, diamètre maximal moyen, nombre total de tempêtes constatées, nombre total de tempêtes par classe de taille de grêlons. Du fait du bond technologique survenu la dernière année, nous créons également ces mêmes variables en retirant l'année 2020 pouvant biaiser l'étude. Nous analyserons par la suite quelles variables sont pertinentes dans le cadre de la construction du zonier.

Deux variables, créées à partir de cette base, caractérisent des informations majeures sur la sinistralité causée par la grêle :

— Le diamètre maximal moyen, retranscrivant la puissance des tempêtes subies.

— Le nombre de fois où la grêle touche la commune, retranscrivant la fréquence.

La représentation géographique de ces deux variables est visibles en Figure 36 :



(a) Diamètre maximum moyen du grêlon (b) Nombre de tempêtes de grêle

FIGURE 36 – Statistiques par commune à partir de la base CatNat

La variable du nombre total de tempêtes présente une carte similaire aux études analysées en introduction et proche de la fréquence vue précédemment sur notre portefeuille. A noter que le Bordelais et la Côte d'Azur semblent très touchés, ce qui n'avait pas été identifié auparavant.

Le diamètre maximal moyen ne semble pas être une variable très discriminante sur le territoire. En effet, nous ne distinguons pas clairement de régions avec des grêlons de petites ou grandes tailles ressortir. L'Allier, le Gers et le Tarn semblent être sinistrés par des grêlons de plus grande taille.

Finalement, nous avons à disposition une base de données permettant d'améliorer notre connaissance des tempêtes de grêle sur le territoire français. Elle vient en complément des bases de données internes et externes en accès libre que nous avons présentées. Nous testerons la pertinence de toutes ces variables par la suite et nous jugerons de l'apport d'information de la base de données grêle achetée.

c) La capacité de détection des tempêtes

Nous avons enfin analysé la capacité de la base de données achetée à détecter les tempêtes de grêle constatées dans le portefeuille Pacifica.

La grêle est un phénomène très bref. Par conséquent, la date de survenance d'une tempête est précise et ne prête pas à confusion. Nous sommes donc en mesure de croiser, de manière fiable, les

date des sinistres de notre portefeuille avec les dates des tempêtes de la base de données. La localisation des tempêtes, en revanche, est moins précise. Il peut être compliqué de délimiter précisément une tempête, d'autant que selon les endroits les diamètres des grêlons vont varier. Pour croiser les deux bases de données, nous utiliserons donc, en plus de la date de survenance, le département qui est à une maille assez large.

Sur l'historique, nous nous observons que seulement 15% des sinistres ont une tempête associée dans la base de données achetée. De plus, 26% de la charge des sinistres est affectée à une tempête. Ce constat montre tout d'abord que la technologie de CatNat détecte mieux les fortes tempêtes car une plus grande part de montants est détectée.

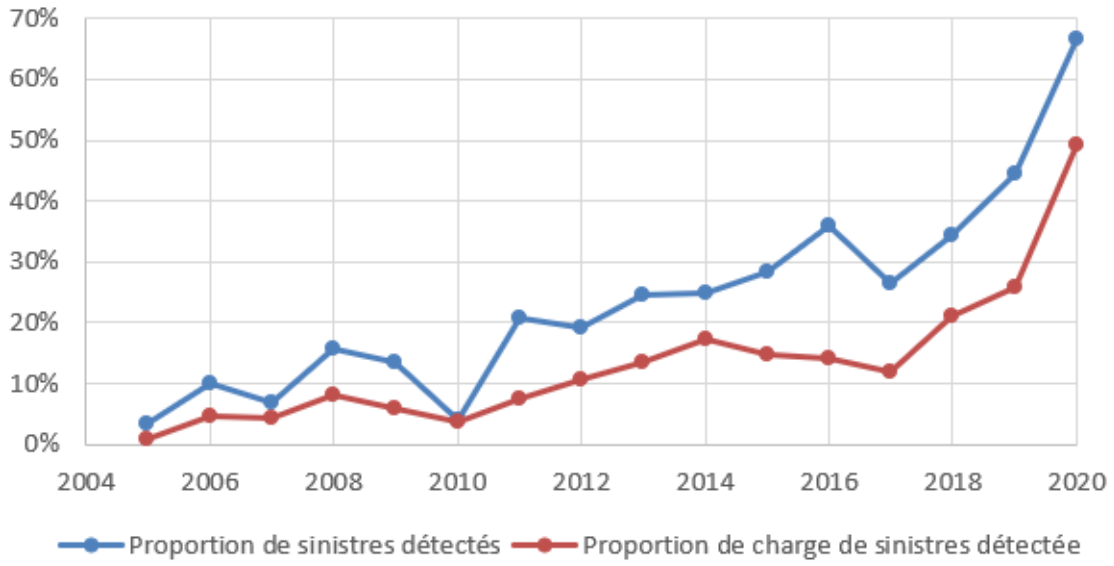


FIGURE 37 – Évolution de la capacité de détection des sinistres grêle par la technologie CatNat

La Figure 37 fait apparaître la très faible capacité de détection de la technologie jusqu'à 2012. Ensuite, jusqu'à 2018, cette détection reste limitée. Enfin, depuis 2018, nous observons une évolution très importante, symbolisant le potentiel de cette base de données. Nous pouvons considérer que jusqu'à 2012 l'apport d'information de cette base est non significatif, il devient ensuite de plus en plus important au fil des années.

Le croisement du portefeuille Pacifica et de la base de données donne également des informations sur la sensibilité des récoltes aux différents niveaux de tempêtes de grêle. Le tableau 5 montre la répartition des sinistres et de la charge des sinistres pour l'année 2020 par classe de grêlons :

Type de grêlon observés	Proportion de sinistres	Proportion de la charge des sinistres
A1 [1cm ;2cm[	75,5%	67,8%
A2 [2cm ;3cm[	22,1%	28,8%
A3 [3cm ;4cm[	2,3%	3,3%
A4 [4cm ;5cm[	0%	0%
A5 [5cm ;6cm[	0,1%	0,1%

TABLE 5 – Répartition des sinistres et de la charge des sinistres pour l'année 2020 par classe de grêlons, parmi ceux détectés par CatNat

Nous remarquons bien que les tempêtes de grêlons de toutes les tailles impactent les cultures. A titre de comparaison, les grêlons de type A1 ne causent pas de dégâts sur les toitures. Cependant, les tempêtes A1 représentent au moins 67,8% de la charge des sinistres grêle. Les récoltes sont des biens assurables très sensibles au risque grêle.

A noter que lorsque la capacité de détection des tempêtes se rapprochera de l'exhaustivité, il sera possible d'améliorer notre connaissance de la sensibilité des différentes cultures au risque grêle. Nous pourrons regarder par exemple quelles cultures sont sensibles aux tempêtes de type A1, quelles cultures résistent la majeure partie du temps au tempêtes A2. Nous pourrons voir également dans quelle proportion chaque type de tempêtes de grêle sinistrent les cultures. Cette base de données nous permettrait alors de raisonner toute chose égale par ailleurs lors de l'analyse de la sensibilité des cultures.

Chapitre III Outils mathématiques

Nous disposons d'un nombre importants de variables, il faut retenir seulement celles pertinentes et optimiser l'utilisation de l'information contenue par ces dernières. La base de données est volatile, il est nécessaire de dissocier le bruit de l'information pertinente. La base de données est également partielle, il est judicieux de compléter l'information à l'ensemble du territoire.

Pour pallier à ces contraintes, il est possible de recourir à différents outils mathématiques. Ceux-ci permettront une meilleure utilisation de l'information à disposition. Nous pourrons ainsi construire un zonier grêle adapté à chaque commune limitant la volatilité.

III.1 Krigeage

III.1.1 Utilisation de l'information spatiale

Lors de la présentation de nos données, nous avons pu voir que certaines variables sont incomplètes géographiquement : l'information est disponible seulement pour une partie des communes du territoire. C'est le cas des données météorologiques ou bien encore de la fréquence des sinistres sur le produit multirisque climatique.

D'autre part, nous pourrons être confrontés à des variables complètes sur le territoire, mais exposées à une forte volatilité. C'est le cas des résidus de nos modèles d'ajustement. La forte volatilité de la sinistralité grêle, pour certaines communes, peut aboutir à des résidus très importants lors de nos ajustements. Il sera intéressant de travailler ceux-ci afin d'en extraire une information géographique, information qui nous est inconnue avec nos variables à disposition.

III.1.1.1 Autocorrélation spatiale

Divers outils mathématiques peuvent nous permettre de compléter ou extraire au mieux l'information géographique. L'utilisation de ces méthodes pose une nouvelle problématique : une autocorrélation spatiale est-elle suffisamment importante pour la variable ?

L'autocorrélation spatiale est définie comme la corrélation, positive ou négative, d'une variable avec elle-même du fait de la localisation spatiale des observations. Si l'autocorrélation spatiale d'une variable est trop faible, alors le recours à un outil de propagation n'est pas pertinent. En effet, cela reviendrait à compléter l'information manquante de manière aléatoire. Pour mesurer l'auto-

corrélation spatiale, différents indicateurs existent comme la diagramme de Moran, mais souvent, une simple visualisation cartographique suffit pour s'apercevoir du phénomène.

Nous utiliserons la visualisation cartographique. Nous comparerons une carte illustrant une distribution géographiquement aléatoire sur le territoire à une carte représentant la distribution constatée. Un exemple d'un tel procédé apparaît en Figure 38, il retranscrit la répartition des salaires par quartiers sur Paris :

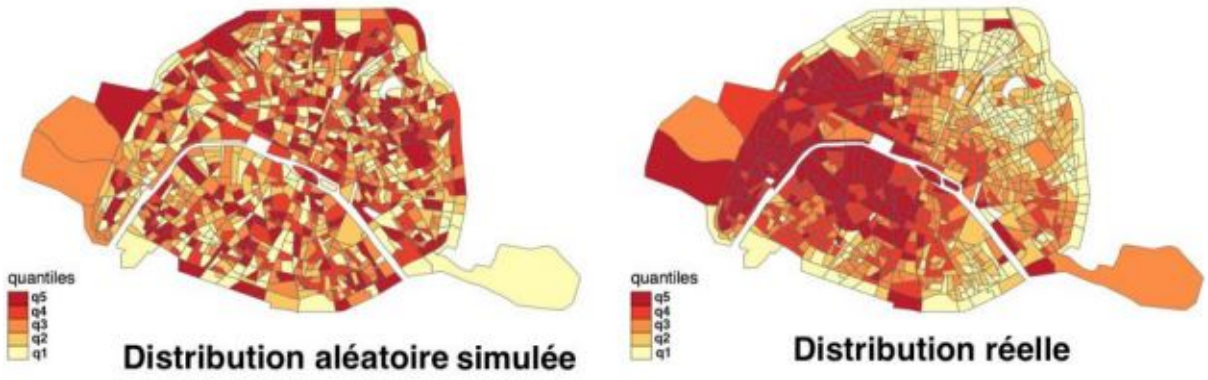


FIGURE 38 – Illustration, sur les Iris parisiens, de l'écart entre une distribution aléatoire et une distribution autocorrélée spatialement (Source : Insee, Revenus Fiscaux Localisés 2010).

La représentation géographique de la distribution réelle montre l'existence d'une autocorrélation spatiale des salaires à Paris. Si l'information de salaire d'un quartier est manquante, nous pouvons utiliser celle des quartiers voisins pour l'estimer. Bien que l'information des voisins soit imparfaite, elle reste très pertinente. Son utilisation ne le serait pas si nous nous trouvions dans le cas d'une distribution aléatoire, comme sur la première carte.

Certaines autocorrélations spatiales sont évidentes visuellement, comme nous avons pu le voir précédemment. Lorsque ce n'est pas le cas, elle peut également être confirmée à partir d'indicateurs précis. Il est possible de définir un seuil à partir duquel l'autocorrélation sera jugée suffisante. Enfin, celle-ci, une fois constatée, nécessite un traitement adéquat.

III.1.1.2 Les outils de mathématiques spatiales de propagation

Supposons qu'une autocorrélation existe. Nous allons nous intéresser aux outils mathématiques de propagation, d'estimation ou encore de lissage spatial disponibles qui nous permettront d'utiliser les informations offertes par les voisins.

Tout d'abord, nous allons présenter la régression spatiale, également appelée régression géographiquement pondérée. Il existe au moins autant de modèles de régression spatiale que de types d'interaction spatiale. Elle peut être utilisée par exemple dans le cas d'une corrélation spatiale liées à des caractéristiques inobservées, inconnues ou même négligées, cas typique des résidus dans le cadre de la construction d'un zonier. La régression spatiale s'utilise donc très bien dans un but exploratoire afin de découvrir ou de vérifier des phénomènes spatiaux. Ces types de modèles peuvent être compliqués à mettre en place puisqu'ils nécessitent l'implémentation de modèles locaux afin d'en obtenir un global. C'est pourquoi nous n'avons pas retenu cette méthode ici, mais elle représente un outil majeur dans l'analyse spatiale avec des résultats concluants.

Deuxièmement, il existe l'outil d'interpolation spatiale. Ce dernier consiste à estimer des valeurs en différents points de l'espace à partir de valeurs connues en un nombre limité de points. Cette technique est essentiellement utilisée pour modéliser des phénomènes physiques continus comme la climatologie. Il peut donc apporter parfaitement son utilité dans la résolution de notre problème de données météorologiques incomplètes sur le territoire. Néanmoins, nous ne retiendrons pas non plus cette méthode car l'interpolation va certes prendre en considération les distances entre le point à prédire et les points connus et peut prendre aussi en considération une pondération souhaitée, mais ne prendra pas en compte la distance entre les points connus eux-mêmes.

Supposons un cas extrême et absurde : nous souhaitons connaître la température moyenne à Lyon et nous disposons seulement des données de Marseille, Toulon, Montpellier et Paris. De plus, nous sommes dans l'incapacité de mettre une pondération individuelle adaptée. L'algorithme ne différenciera alors pas ces villes équidistantes de Lyon. Un biais existera car il ne notera pas que les trois premières citées apportent une information redondante du fait de leur proximité.

Troisièmement, nous avons relevé l'existence du lissage spatial comme outil de statistiques spatiales. Celui-ci consiste à filtrer l'information pour révéler des structures spatiales sous-jacentes et délocaliser l'information. La combinaison de l'interpolation spatiale et du lissage spatial permet souvent de s'affranchir des problèmes de maillage en ramenant les phénomènes à des grilles. Cette utilisation peut s'avérer pertinente dans la construction d'un zonier.

Le principe du lissage spatial est de représenter, non pas la valeur observée en un point, mais une moyenne pondérée des valeurs observées au voisinage de ce point dans un rayon prédéfini. Dans cette moyenne pondérée, les lieux les plus proches sont davantage pris en compte. Plus la taille du rayon est augmentée, plus le lissage est important, et ainsi plus le biais augmente. A l'inverse, en réduisant le rayon, la variance diminue. L'objectif du statisticien, comme souvent, est alors d'effectuer un arbitrage entre les deux. Son objectif peut-être par exemple de minimiser une mesure d'erreur, telle que l'erreur quadratique.

Nous retenons au final comme solution le lissage spatial qui pourra nous être utile à la fois pour compléter l'information de nos variables explicatives, ainsi que pour extraire une potentielle information géographique de nos résidus. Parmi les méthodes de lissage spatial existantes, nous avons retenu le krigeage que nous allons décrire ci-après.

III.1.2 Présentation théorique du krigeage

Le krigeage est une méthode géostatistique de modélisation spatiale qui donne la possibilité d'obtenir une représentation homogène des informations étudiées. Le krigeage nous permettra donc par exemple, à l'aide de relevés obtenus sur des stations météorologiques, d'estimer ces mêmes variables sur le reste du territoire.

Marie Hennequi relate que : « Les méthodes statistiques classiques telles que la régression linéaire se basent sur une hypothèse fondamentale qui est l'indépendance des variables. Or, lorsqu'une variable est spatialement autocorrélée, cette hypothèse n'est plus vérifiée. Ainsi, le krigeage va se baser sur cette nouvelle hypothèse : l'autocorrélation spatiale des données. »[9]

Le modèle de base du krigeage reprend celui d'une régression linéaire, mais cette fois les résidus sont supposés dépendants spatialement :

$$Z(s) = \mu(s) + \delta(s) \tag{5}$$

$\forall s \in D$, où $\mu(.)$ constitue la structure déterministe pour l'espérance de Z. $\delta(.)$ est une fonction aléatoire stationnaire.

L'idée de base du krigeage est d'estimer la valeur d'une variable à partir de son voisinage par une combinaison linéaire des données ponctuelles adjacentes. Le krigeage veut retranscrire une information non biaisée et limiter la variance de l'estimation. Ces objectifs se résument en quatre contraintes qui servent de socle à la méthode de krigeage :

La contrainte de linéarité :
L'estimateur est une combinaison linéaire pondérée des données. Ainsi, dans le cas où nous souhaitons estimer la valeur de la variable régionalisée, Z, au point s_0, nous aurons :

$$\hat{Z}(s_0) = a + \sum_{i \in V(s_0)} \Lambda_i Z_i(s_i) \tag{6}$$

où les s_i $(i = 1...n)$ sont les points utilisés pour l'estimation, et où la constante a et les poids Λ_i sont les inconnues du problème.

La contrainte d'autorisation :
L'erreur d'estimation doit être une combinaison linéaire autorisée, c'est-à-dire que son espérance et sa variance doivent exister.

La contrainte de non-biais :
L'espérance de l'erreur d'estimation doit être nulle. Si l'estimation de krigeage est réalisée un grand nombre de fois, avec la même paramétrisation, les erreurs commises se compensent. Ceci rendra l'analyse des résidus des modèles construits impossible.

La contrainte d'optimalité :
Enfin, il faudra rechercher des pondérateurs qui minimisent la variance de l'erreur d'estimation, sous les contraintes précédentes. Si l'estimation est reproduite un grand nombre de fois, avec la même paramétrisation, alors la variance des erreurs est la plus faible possible.

Au final, quel que soit le type de krigeage utilisé, le système de krigeage est toujours obtenu au terme de ces quatre étapes.

III.1.3 Application du krigeage

L'implémentation du krigeage peut se faire sur le logiciel R, à partir de la fonction "km" du package "DiceKriging". La paramétrisation de ce modèle de lissage est complexe, elle suit plusieurs étapes :

— Le variogramme : le variogramme est une fonction qui décrit la dépendance spatiale de la variable géostatistique. La fonction variogramme doit être estimée à partir des données disponibles en utilisant une méthode appropriée comme celle des moindres carrés ordinaires. La fonction variogramme peut ensuite être utilisée pour estimer les poids optimaux des observations dans le processus de krigeage.

— Le modèle de krigeage : le modèle de krigeage est une fonction mathématique qui décrit comment les observations sont combinées pour prédire les valeurs à des emplacements non observés. Les modèles de krigeage couramment utilisés comprennent notamment le krigeage ordinaire, le krigeage universel, le krigeage avec dérive.

— Le choix de l'interpolateur : R propose plusieurs fonctions d'interpolation pour effectuer le krigeage, telles que la fonction krige, la fonction predict.kriging, la fonction gstat, ... Le choix de l'interpolateur dépend des besoins spécifiques de l'utilisateur.

— La sélection du nombre de voisins : le nombre de voisins est le nombre d'observations utilisées pour estimer la valeur de la variable à un emplacement donné. Le nombre de voisins intervient directement sur l'intensité du lissage.

— Il est enfin important d'évaluer la qualité de l'ajustement en utilisant des indicateurs tels que le coefficient de corrélation ou l'erreur moyenne d'ajustement sur un ensemble de test.

Nous devrons être précautionneux dans l'application du krigeage comme le décrit Frédéric Sémércurbe :

≪ Derrière la qualité esthétique des cartes lissées se cache néanmoins un piège majeur. Mes méthodes de lissage atténuent les ruptures et les frontières et induisent des représentations continues des phénomènes géographiques. Les cartes lissées font donc apparaître localement de l'autocorrélation spatiale. Deux points proches par rapport au rayon de lissage ont mécaniquement des caractéristiques comparables dans ce type d'analyse. De ce fait, commenter à partir d'une carte lissée des phénomènes géographiques dont l'ampleur spatiale est de l'ordre du rayon de lissage n'a guère de sens.≫ [7]

Ainsi, si nous souhaitons comparer des communes voisines après avoir effectué un lissage à partir d'un rayon de plusieurs kilomètres, cela n'a que peu d'intérêt. En effet, le lissage rendra la discrimination entre voisins impossible car son objectif est contraire. Il nous faudra donc vérifier si le krigeage s'adapte à nos besoins en terme de segmentation de l'information.

III.2 Modèles Linéaires Généralisés (GLM)

Les modèles linéaires généralisés permettent d'établir un lien entre des variables explicatives choisies et la variable à expliquer. Nous pourrons, à partir de ces modèles, isoler et quantifier les différents effets qui conduisent à une telle sinistralité estimée pour chaque assuré. Les modèles linéaires généralisés pourront donc nous permettre par exemple, pour chaque commune, d'établir un coefficient géographique de risque à partir des différentes caractéristiques de celle-ci.

Les modèles linéaires généralisés [13] sont fréquemment utilisés dans le cadre de la tarification en assurance car ils sont contraints par un faible nombre d'hypothèses difficilement vérifiables, contrairement à d'autres modèles. De plus, ces modèles offrent une facilité d'interprétation. Ils présentent également l'avantage, contrairement à leurs homologues, les modèles linéaires simples, que la relation entre la variable estimée et les variables explicatives ne soit pas obligatoirement linéaire.

13. ou GLM pour Generalized Linear Model

III.2.1 Présentation théorique

III.2.1.1 Définition

Les modèles linéaires généralisés sont définis par l'équation suivante :

$$g(E[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{7}$$

Avec : $Y = Y_i, i = 1, \ldots, n$ un vecteur de R_n contenant les variables à estimer. Sa densité appartient à une famille exponentielle.
Pour i allant de 1 à n : $(X_{1i}, \ldots, X_{pi})$ un p-uplet contenant l'ensemble des p variables explicatives pour l'observation i.
$\beta_0, \beta_1, ..., \beta_p$ les paramètres de la régression estimés grâce à la méthode du maximum de vraisemblance. La combinaison linéaire des paramètres et des variables explicatives représente la composante déterministe du modèle.

La loi de Y doit appartenir à la famille exponentielle, c'est-à-dire qu'il faut trouver :

(i) $\theta \in \mathbb{R}$ (paramètre canonique)

(ii) $\phi \in \mathbb{R}$ (paramètre de dispersion)

(iii) a une fonction définie sur $\mathbb{R}^*$

(iv) b une fonction définie sur $\mathbb{R}$ et deux fois dérivables

(v) c une fonction définie sur $\mathbb{R}^2$ tels que la densité de Y peut s'écrire sous la forme :

$$f(\theta, \phi) = exp(exp(\frac{y\theta - b(\theta)}{a(\theta)} + c(y, \phi))) \tag{8}$$

Dans ce cas, les moments de Y peuvent s'exprimer en fonction des différents paramètres :

$$E[Y] = \mu = b'(\theta) \qquad \text{et} \qquad Var[Y] = \sigma^2 = b''(\theta) * a(\phi) \tag{9}$$

Parmi les lois appartenant à la famille de lois exponentielles nous retrouvons par exemple :

— Les lois de Poisson (utilisées pour estimer les fréquences de sinistres)

— Les lois Gamma (utilisées elles dans le cadre de l'estimation du coût des sinistres)

— Les lois Binomiales ou de Weibull

Il nous faut alors, après avoir choisi la distribution adaptée, définir une fonction de lien. Cette fonction doit être dérivable et strictement croissante. Nous avons le choix parmi les fonctions suivantes :

— La fonction identité, pour le modèle linéaire classique :

$$g(E[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{10}$$

— La fonction logit telle que : $g(x) = ln(\frac{x}{1-x})$, donnant le modèle logarithmitique suivant :

$$E[Y] = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)} \tag{11}$$

— La fonction inverse telle que $g(x) = \frac{1}{x}$, donnant :

$$E[Y] = \frac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p} \tag{12}$$

— La fonction logarithme népérien que nous privilégierons dans le cadre de notre tarification et nous permettant ainsi d'avoir un modèle multiplicatif. A partir de $g(x) = ln(x)$, nous retrouvons donc :

$$E[Y] = \exp(\beta_0) * exp(\beta_1 X_1) * exp(\beta_2 X_2) * ...exp(\beta_p X_p) \tag{13}$$

III.2.1.2 Estimation des paramètres de la régression

Les paramètres à estimer sont β et Φ. Φ étant un paramètre de nuisance qui n'influence pas la maximisation de la vraisemblance.

Étant donné une certaine loi de probabilité Y et des observations $(y1, y2, \ldots, yn)$, la vraisemblance quantifie la probabilité que les observations proviennent d'un échantillon de la loi de Y. Ainsi, plus la vraisemblance est proche de 0, moins l'adéquation est bonne. C'est pourquoi nous cherchons à maximiser cette vraisemblance, afin d'en déduire les paramètres optimaux. La vraisemblance correspond au produit des probabilités de chacune des réalisations, d'où nous avons la

vraisemblance notée L et définie par :

$$L(\beta) = \prod_{i=1}^{n} f(y|\theta_i(\mu_i(\beta)), \phi) \tag{14}$$

En pratique, il est souvent plus simple de passer par le logarithme népérien afin de transformer le produit en somme :

$$l(\beta) = ln(\prod_{i=1}^{n} f(y|\theta_i(\mu_i(\beta)), \phi)) = \sum_{i=1}^{n} ln(f(y|\theta_i, \phi)) \tag{15}$$

Les paramètres optimaux, que nous appelons estimateur du maximum de vraisemblance et que nous notons $\widehat{\beta}$, sont les paramètres qui permettent de maximiser la fonction log de vraisemblance, d'où :

$$\widehat{\beta} = argmax_l l(y, \beta) \tag{16}$$

En pratique, nous résolvons ce problème avec des méthodes itératives telles que l'algorithme de Newton-Raphson. Une fois les paramètres estimés, nous pouvons finalement passer à l'étape de prédiction de l'espérance pour l'ensemble de nos observations i :

$$\widehat{\mu_i} = \mu_i(\widehat{\beta}) = g^{-1}(\sum_{k=1}^{p}(\widehat{\beta_k} x_{ki}) \tag{17}$$

De manière plus pratique, les coefficients vont servir de base à notre modèle et représentent la contribution des variables explicatives dans l'ajustement. Nous sommes dans le cas d'un modèle multiplicatif, et non additif, impliquant une interprétation des coefficients adaptée. La valeur du coefficient interprétée seule et de manière brute n'a aucun intérêt, ce sont seulement les écarts relatifs que nous interpréterons.

III.2.2 Indicateurs de qualité et méthode de construction

III.2.2.1 Indicateurs de qualité d'un modèle

La paramétrisation d'un modèle et la sélection des variables sont des étapes obligatoires dans la quête d'une estimation optimisée. Pour réaliser ces choix, il nous faut être en mesure de comparer les modèles. Ceci est permis grâce à des indicateurs comme le R² ajusté, le BIC ou l'AIC par exemple. Ces indicateurs permettent de rendre compte de l'information apportée, tout en pénalisant l'ajout trop important de variables. L'ajout de variables permet naturellement d'augmenter la quantité d'informations apportée à l'estimation, mais ces ajouts viennent alourdir le modèle .

Il nous faudra prendre soin de vérifier la significativité des coefficients pour les variables explicatives retenues. Divers indicateurs comme le V de Cramer ou les tests de significativité des coefficients permettent de réaliser cette vérification. Il faudra également vérifier les potentielles corrélations existantes entre les variables explicatives elles-mêmes, afin d'éviter tout apport redondant d'information dans le modèle.

Enfin, pour juger le surapprentissage, il sera intéressant d'analyser la capacité de notre modèle à s'ajuster fidèlement à un jeu de données différent de celui d'entraînement. Pour réaliser cela nous découperons notre base en deux ensembles : un ensemble d'entraînement et un ensemble de test. Un modèle qui ne surapprend pas est un modèle dont l'erreur d'estimation sur l'ensemble de test est proche de l'erreur d'ajustement sur l'ensemble d'entraînement.

III.2.2.2 Méthode de construction de la régression

Pour éviter de réaliser une estimation sur des informations potentiellement erronées, redondantes ou inutiles, il nous faudra établir une stratégie dans la construction du modèle. Deux méthodes sont possibles :

— La méthode ascendante (forward selection) : nous allons dans un premier temps implémenter plusieurs modèles avec une seule variable en sélectionnant celui offrant les meilleurs résultats. Les meilleurs modèles sont ceux qui optimisent le R^2 et l'AIC. Ensuite, nous construisons des modèles à deux variables composés de celle sélectionnée précédemment et d'une autre. Nous conservons le modèle offrant les meilleurs résultats. Nous construisons alors des modèles à trois variables à partir des deux précédentes de la même manière. Ainsi le processus se poursuit jusqu'à ce que l'ajout de variable ne soit plus bénéfique au modèle.

— La méthode descendante (backward selection) : nous démarrons l'implémentation de l'algorithme avec l'ensemble des variables. Á chaque étape sera retirée la variable apportant le moins d'information. Le processus prend fin lorsque le retrait d'aucune variable ne permet pas une diminution de l'AIC ou du BIC.

III.2.3 Compréhension des visuels du logiciel Emblem

Emblem est un logiciel développé par l'entreprise de conseil Towers Watson, spécialisée dans les assurances. Le logiciel permet d'ajuster facilement des modèles composés d'un nombre impor-

tant de variables. Il permet aussi de vérifier les colinéarités et de réaliser différents tests sur les variables explicatives conservées. Emblem offre surtout divers affichages interactifs pour analyser la pertinence du modèle ainsi que des pistes d'améliorations. Une fois l'analyse réalisée, il est possible d'adapter les paramètres de la modélisation rapidement, comme par exemple regrouper des modalités, changer le degré d'ordre du coefficient d'une variable afin de lisser la pente des coefficients d'une variable.

Lorsque nous présenterons nos modèles dans la suite de ce mémoire, nous utiliserons les visuels du logiciel Emblem pour appuyer nos propos. Les visuels mettent en valeur plusieurs indicateurs. Pour une meilleure compréhension future, nous allons les présenter brièvement par le biais des deux figures suivantes :

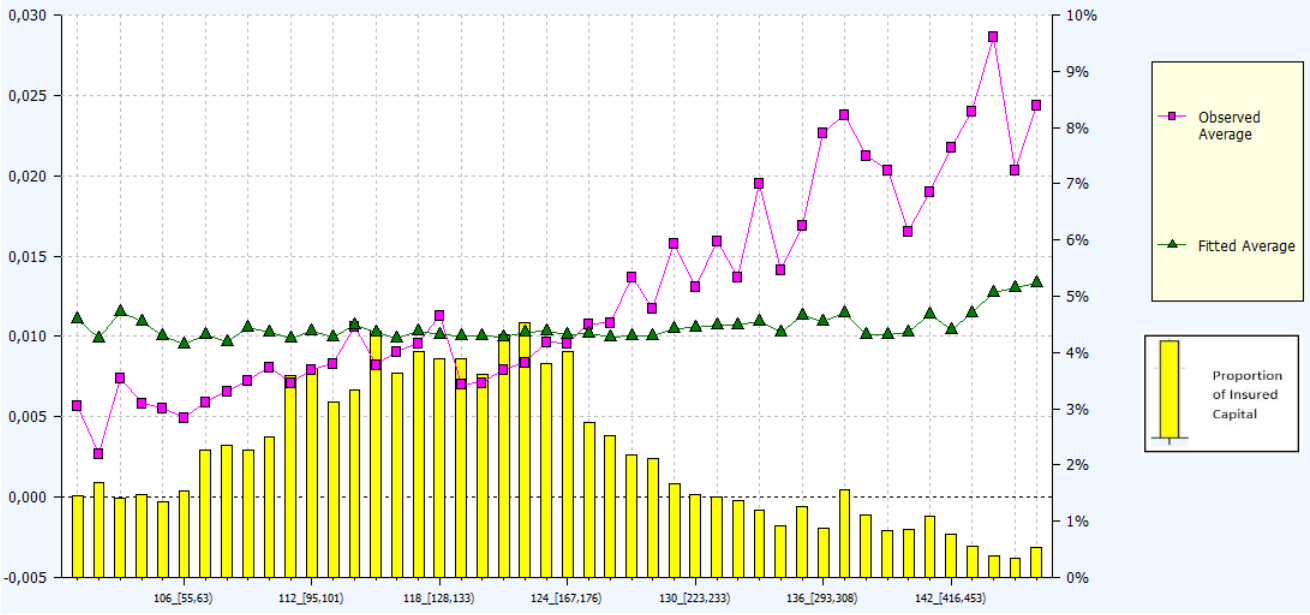


FIGURE 39 – Exemple de visuel Emblem représentant les taux de prime pure observés et ajustés

Sur la Figure 39 sont représentés les résultats pour une variable qui n'est pas utilisée dans le modèle. Sur cette figure apparaissent deux courbes et un diagramme bâton. En abscisse, sont représentées les différentes modalités, classées par ordre alphabétique.

Tout d'abord, le diagramme bâton en jaune indique le volume de capitaux assurés pour chaque modalité. Nous pouvons interpréter cela comme le poids de chaque modalité. La courbe rose correspond au taux de prime pur observé sur l'historique pour chaque modalité de la variable explicative concernée. La courbe verte foncée fait apparaître quant à elle le taux de prime pure ajusté par le modèle pour chaque modalité. Enfin, nous observons deux échelles pour les ordonnées : celle de gauche affiche le taux de prime et va donc concerner les courbes vertes foncées et roses ; celle de droite représente le poids en terme d'exposition (capitaux assurés) pour chaque modalité,

légendant ainsi le diagramme bâton jaune. L'échelle des taux de prime pure sera masqué sur tous les graphiques pour des raisons de confidentialité.

Notre objectif est que la courbe verte foncée et la courbe rose, respectivement l'ajusté et l'observé, se rapprochent. Nous pouvons voir sur la Figure 39 que les deux courbes ont des tendances assez éloignées. Il pourrait être intéressant d'ajuster le taux de prime pure par la variable présentée, le modèle capterait alors les variations liées à cette variable explicative.

Un ajustement parfait entre les deux courbes n'est pas non plus un objectif. Pour un ajustement idéal, la différence entre les deux courbes devrait correspondre uniquement au bruit. Ce bruit est l'information de l'observé que nous ne voulons pas capter dans notre modèle. Un ajustement trop important capterait ce bruit.

Pour ajuster au mieux sur l'observé, nous pourront être amenés à utiliser différents degrés de polynômes pour effectuer un lissage. Le passage polynomiale permet d'éviter les sauts de coefficients d'une modalité à l'autre et permet de limiter la volatilité pour les modalités faiblement exposées. L'augmentation des degrés est fonction positive de la variance mais négative du biais. De plus, les polynômes permettent d'obtenir des coefficients plus cohérents et interprétables, tendant vers une certaine monotonie, afin d'éviter par exemple une évolution en dent de scie.

Sur la Figure 40, nous pouvons constater l'ajustement de la variable *Altitude* sur la prime pure observée :

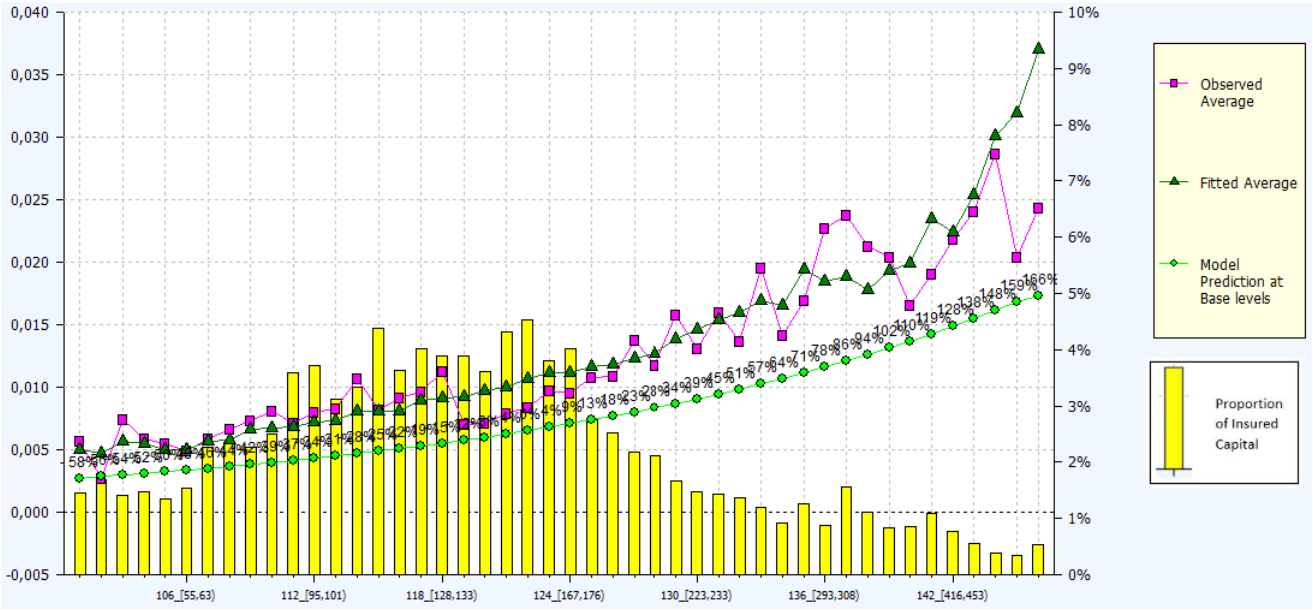


FIGURE 40 – Exemple du taux de prime pure ajusté sur la variable *Altitude*

La courbe verte claire représente la valeur du coefficient estimé pour chaque modalité d'une variable. Nous appellerons tout au long du mémoire cette courbe la pente des coefficients.

L'objectif est de différencier les modalités entre elles, cette différenciation doit pouvoir être expliquée et comprise. Sur la Figure 40, plus la valeur prise par la variable augmente, plus le taux de prime pure élevé, ce qui est en accord avec ce que nous avons constaté en première partie.

La modalité mode des capitaux assurés de chaque variable explicative sert de coefficient de référence. Cette modalité référence aura un coefficient de 100%. De plus, sur la Figure 40, nous remarquons que la classe la plus élevée a un coefficient de 196%. Le taux de prime pure est ainsi 2.96 fois plus élevé qu'au niveau de la modalité de référence, toute chose égale par ailleurs (196% + 100%).

III.2.4 Avantages et limites dans le cadre de la tarification

Les coefficients nous permettent de connaître les effets pour chaque variable d'une modalité par rapport à une autre, toute chose égale par ailleurs. Ils nous permettent de différencier de manière claire chaque modalité d'une variable. D'un point de vue interprétation, ce type de modèle est donc parfaitement adapté à la tarification. Il répond à notre objectif principal qui est de chercher à connaître les effets d'une modalité sur la prime pure afin de réaliser l'équation tarifaire.

De plus, les modèles linéaires généralisés sont aisés à mettre en place et à modifier. Il est possible d'optimiser plusieurs paramètres facilement et rapidement d'autant plus grâce à Emblem. Ces modèles ont également l'avantage d'être robustes.

Les modèles linéaires généralisés sont probabilistes. Ils nous offrent la possibilité d'effectuer des tests sur les coefficients en vérifiant que ceux-ci sont significativement différents de 1. Ce test valide ainsi la présence de la variable dans l'ajustement. De la même manière, il est possible de construire des intervalles des coefficients estimés.

Plusieurs inconvénients sont toutefois à noter. Le premier est que nous devons regrouper les valeurs des variables quantitatives continues en classes car le modèle ne peut traiter qu'un nombre fini de modalités pour une variable. Il y a donc une perte d'information avec ces regroupements. Ces types de modèles nécessitent également des hypothèses fortes sur la loi de distribution de données, ce qui peut être compliqué à vérifier.

III.3 Algorithmes de classification hiérarchique

Il serait possible d'attribuer un coefficient géographique propre à chaque commune en fonction du risque grêle que nous lui associons. Toutefois, compte tenu des données à disposition, de l'importante volatilité du phénomène de grêle dans l'espace et dans le temps, il est nécessaire pour nous de tarifer à une maille plus large. A partir de ce constat, quelle maille choisir ?

Plusieurs méthodes statistiques permettent de partitionner une population en différentes classes, ou dans notre cas en zones. La classification ascendante hiérarchique est l'une d'entre elles. Nous allons la décrire dans cette partie. L'objectif sera d'obtenir des zones homogènes en leur sein, et les plus hétérogènes possibles entre elles. Ceci nous permettra de mener à bien l'objectif de différencier au mieux les individus en s'adaptant aux différentes contraintes.

III.3.1 Présentation théorique de la classification hiérarchique ascendante

III.3.1.1 Classification ascendante hiérarchique (CAH)

Le principe de la classification ascendante hiérarchique est de rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distance, distance entre chaque individu pris deux à deux comme le précise Joseph Larmarange dans un article sur la CAH[12]. Ce dernier décrit également que "Deux observations identiques auront une distance nulle. Plus les deux observations seront dissemblables, plus la distance sera importante."

La classification ascendante hiérarchique va rassembler les individus de manière itérative afin de produire un dendrogramme ou arbre de classification. La classification est ascendante car elle part des observations individuelles. Elle est hiérarchique car elle produit des classes ou groupes de plus en plus vastes. Pour obtenir une partition s'adaptant à nos contraintes, il faut choisir à quel endroit nous souhaitons couper l'arbre construit."

La première étape pour implémenter cet algorithme est de calculer une matrice de distance où chaque individu est représenté par n coordonnées. Ce sont les n variables explicatives, centrées réduites, retenues, qui permettront de différencier les individus.

Il faut donc choisir quelles variables retenir. Ensuite, il faudra choisir l'indice de dissimilarité le plus adapté à notre échantillon. Cette dissimilarité est obtenue par le biais d'une distance. Il en existe plusieurs différentes pour les variables quantitatives. Nous présentons ci-dessous quelques

distances parmi les plus courantes :

— La Distance Euclidienne entre les individus I_i et I_j :

$$d(I_i, I_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{in} - x_{jn})^2} \tag{18}$$

— L'Indice de Gower :
Si nous notons S_g l'indice de similarité de Gower comme le fait Antonio Falco[1], la distance de Gower D_g s'obtient simplement de la manière suivante : $D_g = 1 - S_g$. Ainsi, la distance sera nulle entre deux individus identiques et elle sera égale à 1 entre deux individus totalement différents. S_g est tel que :

$$S_g(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} (s_{12j}) \tag{19}$$

s_{12j} représente la similarité partielle entre les individus 1 et 2 concernant le descripteur j. Cette similarité partielle se calcule différemment s'il s'agit d'une variable qualitative ou quantitative. Pour des variables quantitatives, cas qui nous intéresse pour le zonier, la différence absolue entre les valeurs des deux variables est d'abord calculée : soit $|y_{1j}$-$y_{2j}|$. Puis l'écart maximum observé sur l'ensemble du fichier est déterminé et noté R_j. Dès lors, la similarité partielle vaut :

$$S_{12j} = \frac{1 - |y_{1j} - y_{2j}|}{R_j} \tag{20}$$

— La distance du Φ^2 :
Il s'agit de la distance utilisée dans les Analyses de Correspondances Multiples (ACM). C'est une variante de la distance du χ^2. Nous considérons ici que nous avons Q variables. Nous associons à chaque individu une combinaison de valeur pour toutes les Q variables. Des individus avec des combinaisons similaires auront une distance nulle. La distance du Φ^2 peut s'exprimer comme suit :

$$d^2_{\Phi^2}(L_i, L_j) = \frac{1}{Q} \sum_{k=1}^{Q} \frac{(\delta_{ik} - \delta_{jk})^2}{f_k} \tag{21}$$

Ensuite, la deuxième étape consiste à choisir la bonne méthode d'agrégation de ces distances calculées afin de construire le dendrogramme visible en Annexe C. De nombreuses solutions existent comme la moyenne pondérée, la distance maximum, le saut minimum ou encore la méthode de

Ward.

La méthode de Ward, ou méthode du moment d'ordre deux, se distingue de toutes les autres car elle utilise une analyse de la variance approchée afin d'évaluer les distances entre classes. Cette méthode tente de minimiser la somme des carrés de tous les couples potentiels de classes pouvant être formés à chaque étape. Les indices d'agrégation sont à nouveau calculés à chaque étape à l'aide de la règle suivante : supposons qu'une classe M soit obtenue en regroupant les classes K et L. Nous obtiendrons alors, suite à ce regroupement, la distance de la classe M à la classe J par la formule suivante :

$$D(M,J) = \frac{(N_J + N_K) * D(K,J) + (N_J + N_L) * D(L,J) - N_{JD}(K,L)}{N_J + N_k + N_L} \tag{22}$$

où N correspond aux effectifs d'une classe.

En résumé, la méthode de Ward consiste à réunir les deux clusters dont le regroupement fera le moins baisser l'inertie interclasse. De plus, la distance euclidienne sera utilisée pour ce type d'agrégation. Nous calculerons donc la distance au carré entre le barycentre de deux classes pondérée par les effectifs des clusters.

La troisième et dernière étape consiste à choisir le nombre de classes que nous souhaitons conserver, ou plus théoriquement, à quel endroit du dendrogramme souhaitons nous effectuer notre partition. Pour effectuer ce choix, nous pouvons tout d'abord regarder le dendrogramme. Nous analyserons surtout les sauts d'inertie du dendrogramme selon le nombre de classes. Ces sauts symbolisent la quantité d'information perdue pour chaque regroupement effectué au sein du dendrogramme. L'inertie d'un jeu de données est la somme des carrés des distances des points au centre de gravité. Plus l'inertie intra classe est faible, plus les observations sont proches du centre de gravité. Dans le cas où tous les individus sont séparés, cette inertie intra devient donc nulle.
Nous chercherons donc à couper le dendrogramme avant un niveau de regroupement induisant une importante perte d'inertie intra classe.

III.3.1.2 Description de la fonction HClustGeo

La fonction HClustGeo s'utilise sur R. Cette fonction est présentée en détail sur la plateforme R-CRAN[4]. Elle permet d'implémenter un algorithme de regroupement hiérarchique de type Ward avec une contrainte de contiguïté souple. Les arguments principaux de la fonction sont :

— Une matrice $D0$ des dissemblances dans l'espace des "caractéristiques"

— Une matrice $D1$ des dissemblances dans l'espace « contrainte », dans notre cas une matrice des dissemblances géographiques

— Un paramètre de mixage α prenant ses valeurs entre 0 et 1. Le paramètre de mixage définit l'importance de la contrainte dans la procédure de clustering

— Un paramètre de mise à l'échelle, avec une valeur logique. Si $TRUE$, les matrices de dissemblance $D0$ et $D1$ sont mises à l'échelle entre 0 et 1

Nous remarquons déjà que cette fonction, contrairement à la classification ascendante hiérarchique classique, ne calcule pas qu'une seule matrice de distance, mais bien deux. Nous aurons ici une matrice classique, D0, avec les variables retenues pour calculer la dissimilarités entre les observations. Nous aurons également une matrice de dissimilarités géographiques, D1, qui n'est pas mélangée aux autres distances. Par le biais du paramètre α, nous pourrons choisir quelle importance donner à cette matrice de distance géographique dans l'implémentation de l'algorithme sur notre jeu de données.

Enfin, le dernier paramètre permet d'avoir des distances comparables. En effet, il se peut que l'ordre de grandeur d'une des deux distances soit plus important que l'autre avec des dispersions différentes. Ce paramètre va donc pallier ce problème en mettant les deux matrices à la même échelle.

III.3.2 Intérêts et Application

III.3.2.1 Avantages et limites d'une telle méthode

a) Avantages

Ce découpage de l'espace va permettre de prendre en compte la valeur du risque grêle que nous attribuons à une commune et la distance géographique des communes entre elles, afin d'éviter d'obtenir un zonier donnant une tarification trop différente entre communes voisines. Il pourra être judicieux de donner plus d'importance à la géographie, notamment si nous souhaitons mutualiser fortement le risque. A l'inverse, il faudra donner moins d'importance aux distances de cette matrice géographique si nous voulons fortement segmenter la tarification.

Une tempête de grêle est un phénomène très localisé et peut frapper très rarement, voire jamais, une commune durant l'historique à disposition. Par conséquent, certaines communes peuvent avoir été épargnées durant l'historique, mais subiront l'année suivante une importante tempête. D'autres communes pourront avoir été fortement sinistrées durant l'historique, mais seront épargnées les

prochaines années. Á l'heure actuelle, nous sommes incapables d'estimer avec précision les couloirs de grêle. L'information contenue par les voisins peut nous aider à diminuer le bruit de certaines observations. Un zonier ne prenant pas en compte la distance géographique entre communes ne mutualiserait pas le risque. Il pourrait également réaliser un surapprentissage.

b) Limites

Les limites de la classification ascendante hiérarchique peuvent être déduites à partir du paragraphe précédent : la paramétrisation est complexe. Il est en effet souvent difficile de trouver la coupure significative sur le dendrogramme. De plus, du fait de la structure hiérarchique, le partitionnement n'est pas optimal puisqu'il dépend des précédents regroupements.

Enfin, cet algorithme a un fort coût de calcul lorsque le nombre d'observations devient grand car le nombre de distances à prendre en compte dans le partitionnement explose rapidement.

III.3.2.2 Choix des paramètres

La fonction $choicealpha$ implémente une procédure pour aider l'utilisateur dans le choix d'une valeur appropriée du paramètre de mélange alpha. Les deux fonctions HClustGeo et $choicealpha$ peuvent être combinées pour trouver une partition optimale avec contrainte de contiguïté géographique. Les deux étapes de la procédure sont :

— Trouver la partition de k clusters des N communes utilisant la matrice de dissemblance D0. Les clusters de cette partition sont homogènes sur les variables explicatives retenues et aucune contrainte de contiguïté n'est utilisée.

— Choisir un paramètre de mixage α, afin d'augmenter la cohésion géographique des clusters, en recourant à la matrice $D1$, sans trop détériorer l'homogénéité obtenue en première étape.

Des itérations peuvent être réalisées dans le cas où plusieurs valeurs k sont à tester. En effet, chaque k n'a pas forcément le même α optimal qui lui est associé.

Dans un premier temps, pour choisir le nombre de partitions, k, nous allons regarder la perte d'inertie suite à un regroupement, perte visible sur le dendrogramme notamment. Cette perte d'inertie est également visible graphiquement, à travers une courbe que l'on peut dessiner à partir de la formule suivante :

$$Perte_k = \frac{I_k - I_{k-1}}{I_k} \tag{23}$$

où I_k est l'inertie constatée pour k partitions effectuées.

Différentes valeurs de k pourront être retenues. Ce sont les différents regroupements à la suite duquel nous observons une forte perte d'inertie. Il faudra retenir celui le plus en cohérence avec nos besoins et contraintes.

Dans un second temps, nous nous intéressons au choix de α, le paramètre qui indique au package l'importance à donner à la distance géographique. Plus α sera proche de 1, plus on accordera d'importance à la distance géographique. Inversement, plus α proche de 0, plus nous donnerons d'importance à la distance de coefficients entre deux communes. La fonction $choicealpha$ sur R nous aidera dans ce choix. Toutefois, notre expertise sera essentielle compte tenu des diverses contraintes. En effet, la fonction a des difficultés à gérer une trop grande volatilité entre voisins, ce qui implique qu'elle suggère dans ces cas-là de ne pas suffisamment tenir compte de la distance géographique entre les individus.

Chapitre IV Construction de zoniers grêle

Dans les trois premières parties de ce mémoire, nous avons décrit les ressources à notre disposition pour construire un zonier sur le risque grêle en assurance agricole. À présent, nous avons une meilleure connaissance du phénomène physique qu'est la grêle, ainsi que de ses conséquences. Nous avons également évoqué les outils mathématiques qui nous permettront d'estimer le risque grêle propre à chaque commune. Nous allons maintenant appliquer ces ressources afin de construire un zonier optimal compte tenu des diverses contraintes évoquées.

IV.1 Analyse et optimisation du zonier actuellement utilisé par Pacifica

IV.1.1 Distribution du taux de prime pure

En préambule des constructions de zonier, nous allons définir le paramétrage du modèle linéaire généralisé. La sélection et la spécification d'une distribution appropriée est un élément important de l'implémentation. Le taux de prime pure, variable à modéliser, suit une distribution précise que nous allons chercher à connaître. En Figure 41 est représenté l'histogramme de l'historique des sinistres du produit Grêle :

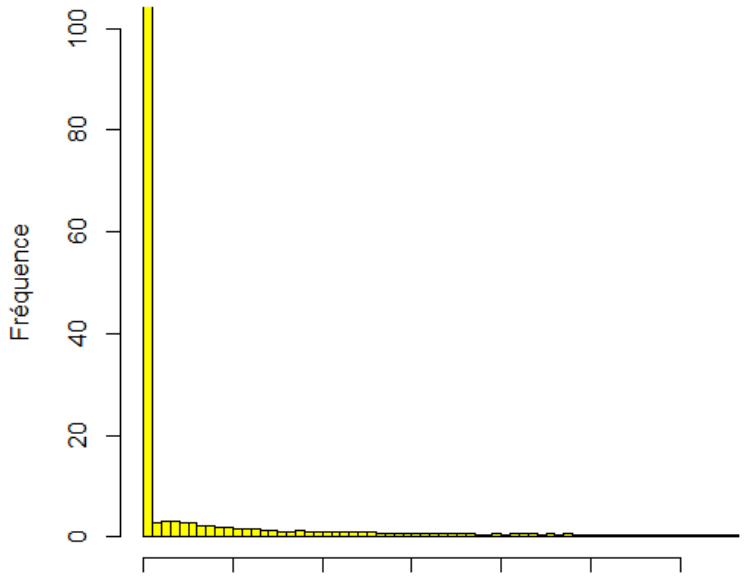


FIGURE 41 – Histogramme des taux de prime pure par classe, divisé en 2000 blocs

La loi de Poisson n'est pas un candidat approprié car nous ne traitons pas uniquement les

données de comptage. D'autre part, la distribution Gamma ne prend pas de valeurs nulles, alors que, comme nous le voyons, un nombre très important d'observations possèdent un taux de prime pure nul. La loi normale n'est également pas appropriée. Nous pouvons évoquer la distribution de Tweedie comme potentiel candidat pour ajuster le modèle.

La distribution de Tweedie est souvent utilisée dans les modèles d'analyse de données pour lesquels la variance des données dépend de la moyenne, tels que les modèles d'assurance, les modèles financiers et les modèles de survie. Elle est idéale pour des variables représentées par beaucoup de zéro. Elle a la particularité de changer d'aspect selon son paramétrage. Lorsque le paramètre de puissance prend une valeur proche de 1, la distribution de Tweedie ressemble à une distribution de Poisson. Lorsque le paramètre de puissance prend une valeur proche de 2, la distribution de Tweedie ressemble à une distribution Gamma. Les valeurs intermédiaires du paramètre de puissance donnent des distributions qui sont des combinaisons de ces deux extrêmes. La sélection de la puissance optimale est réalisée par la comparaison de différentes distribution en Figure 42 :

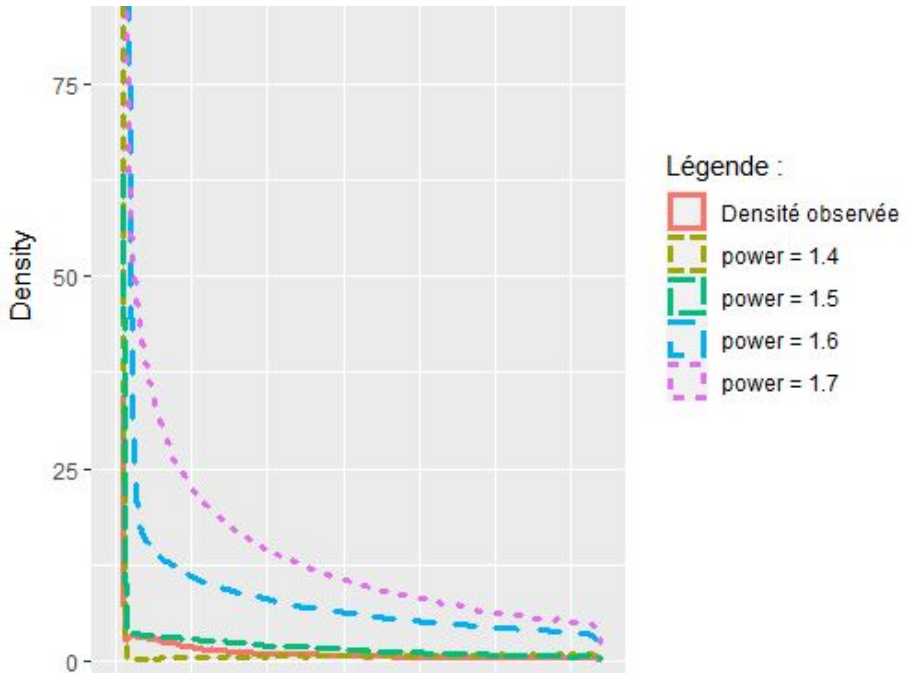


FIGURE 42 – Ajustement de la distribution du taux de prime pure observé par différentes distribution de Tweedie

Chacune des distributions de Tweedie représentées ont une valeur de 1 pour le paramètre ϕ de dispersion. Nous remarquons que le meilleur ajustement est réalisé par une Tweedie ayant une puissance de 1,5. Nous retiendrons donc cette loi comme distribution utilisée pour l'implementation du modèle linéaire généralisé visant à ajuster le taux de prime pure.

IV.1.2 Description du zonier actuel

Avant d'entamer la construction d'un zonier, nous allons analyser celui actuellement utilisé par Pacifica afin d'en observer les limites et les axes d'amélioration. Dans l'équation tarifaire précédemment évoquée, apparaît une variable qui segmente les observations selon le risque grêle lié à la localisation de ces dernières. Nous nommerons cette variable le *zonier actuel*.

IV.1.2.1 Analyse cartographique

Le zonier utilisé par Pacifica est représenté géographiquement sur la Figure 43. Sur cette figure et les suivantes, l'échelle de valeur correspond aux coefficients tarifaires appliqués à chaque commune. La valeur du coefficient tarifaire est traduit graphiquement par l'intensité du rouge.

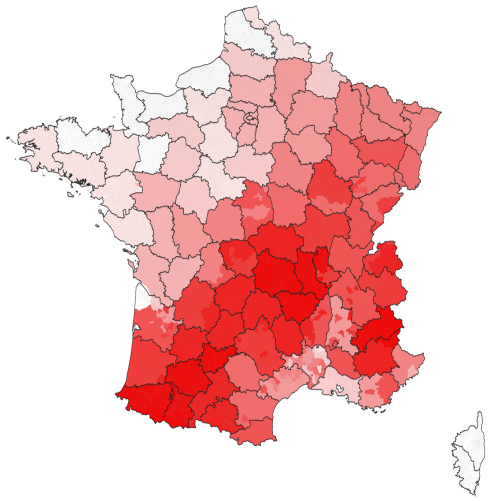


FIGURE 43 – Zonier actuel utilisé par Pacifica

Le zonier actuel est majoritairement à la maille département, c'est-à-dire que les communes d'un département ont un coefficient tarifaire commun qui leur est affecté. Toutefois, des zones d'exception existent pour quelques départements. Ces zones d'exception sont des groupements de communes qui vont avoir un coefficient inférieur ou supérieur à celui du département. Ces zones se situent en Gironde, dans la Drôme ou en Ardèche, comme le montre la Figure 43. Elles ont pour objectif de retranscrire une hétérogénéité du risque grêle au sein du département.

Ce zonier représente le risque grêle sur le territoire français d'une manière globalement semblable à ce qui a été vu en introduction. La diagonale de départements risqués, des Pyrénées au Massif Central, est bien retrouvée. Toutefois, une segmentation plus fine sera à explorer. Il faudra également vérifier la validité des zones d'exception.

IV.1.2.2 Analyse des coefficients

Nous souhaitons maintenant savoir si les coefficients associés pour chaque zone sont cohérents avec ce que nous constatons durant l'historique, que ce soit en terme de valeur et en terme de classement.

Pour ce faire, nous allons ajuster sur Emblem un modèle de taux de prime pure utilisant les variables explicatives telles que : le *code récolte*, la *franchise*, le *zonier actuel* et la variable *sur-zones/sous-zones*. La variable *zonier actuel* est à la maille département et va affecter à chacune des communes de chaque département le même coefficient, qu'elles appartiennent à une zone d'exception ou non. Ensuite, la variable *sur-zones/sous-zones* s'applique aux zones d'exception, elle représente de combien est décalée la zone d'exception par rapport au département. Grâce à ce modèle d'ajustement, nous pourrons comparer les valeurs des coefficients d'ajustement de chaque zone. Nous pourrons également comparer la pente tarifaire commercialisée et celle obtenue avec l'ajustement d'un modèle linaire généralisé.

Le zonier actuel associe un coefficient géographique propre à chaque département. Pour que chaque modalité soit suffisamment représentée en terme de capitaux assurés et ainsi réaliser un ajustement suffisamment stable, il faut regrouper les départements en classes. Nous faisons le choix de regrouper les départements selon la valeur de leur coefficient de risque. Nous obtenons une variable avec 11 modalités tarifaires, classées par ordre croissant des coefficients de risque moyen de chaque classe. A noter que nous avons veillé à créer des groupes homogènes en terme de coefficients tarifaires et assez hétérogènes entre eux, au détriment de créer des modalités plus faiblement représentées parfois. De plus, chaque modalité est composée de plus de 100 millions de capitaux assurés constatés sur l'historique.

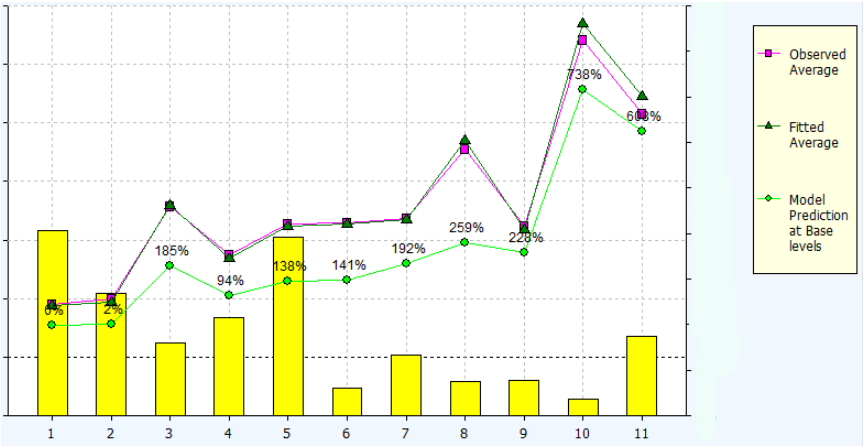


FIGURE 44 – Courbe des coefficients du zonier commercialisé, générée par le GLM

L'ajustement, par un modèle linéaire généralisé, des deux variables géographiques clés est représenté en Figure 44 et en Figure 45. La pente de la courbe de coefficients du zonier est globalement croissante, visible en Figure 44. Ceci montre que le zonier utilisé actuellement offre une segmentation pertinente du risque, avec notamment des zones risquées bien ciblées avec les classes 8, 9, 10 et 11. Par ailleurs, nous observons que la pente des coefficients n'est pas strictement croissante, signe d'une mauvaise affectation du risque pour certains départements. C'est le cas des classes 10 et 3 qui semblent avoir un risque sous-estimé dans le zonier actuel. A l'inverse, la classe 9 semble avoir un risque sur-estimé. À noter que les zones 9 et 10 sont sujettes à plus de volatilité car elles représentent moins de capitaux en portefeuille, respectivement entre 4% et 2% du portefeuille.

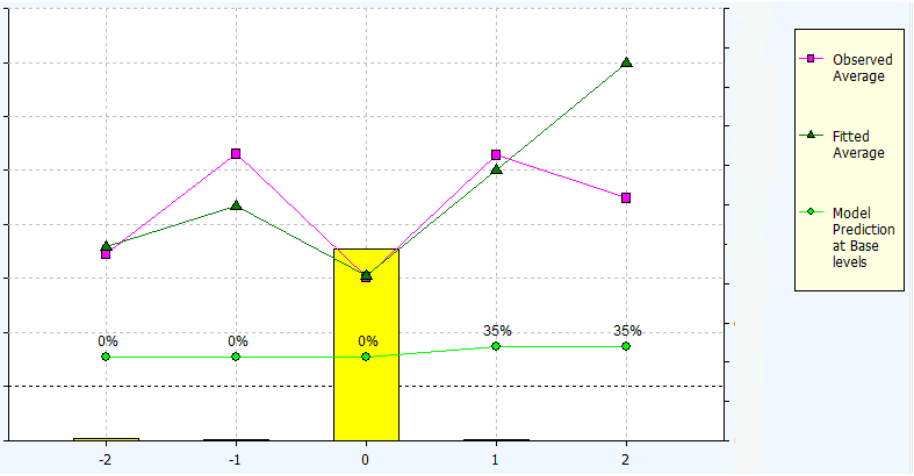


FIGURE 45 – Courbe des coefficients des sur-zones/sous-zones, générée par le GLM

Les zones d'exceptions sont regroupées en cinq modalités. La modalité 0 correspond aux communes ne figurant pas parmi la liste des communes d'exception. Les classes négatives représentent les zones d'exception minorées, -2 étant la classe dont le coefficient tarifaire est le plus minoré par rapport au niveau du département. Les classes positives représentent les zones d'exception majorées, 2 étant la classe dont le coefficient tarifaire est le plus majoré par rapport au niveau du département. La Figure 45 montre que le volume de capitaux assurés sur chaque zone d'exception est très faible. La volatilité des coefficients calculés pour les classes -2, -1, 1, 2 est importante.

Dans le but d'éviter une trop grande volatilité, nous avons effectué un regroupement afin de former trois classes : sous-zones, zone normale et sur-zones. Les sous-zones, celles où le coefficient de la commune est inférieur à celui du département n'ont pas une prime pure significativement différente de leur département. Finalement, la variable zones d'exception n'apporte pas une information significative et discrimine très peu d'assurés.

Il est intéressant de comparer la pente tarifaire du zonier commercialisé avec celle ajustée dans le modèle linéaire généralisé que nous venons de construire.

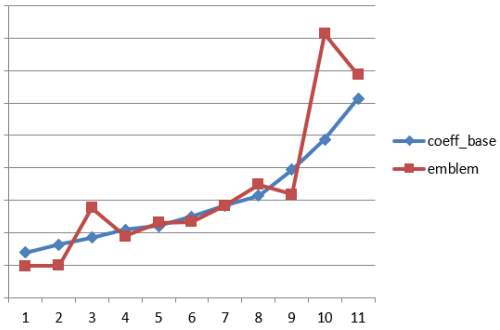


FIGURE 46 – Comparaison de la courbe tarifaire commercialisée et des coefficients calculés par le GLM

La Figure 46 montre que la tarification appliquée offre une segmentation limitée du risque entre les communes. Elle va plutôt mutualiser le risque entre les zones. L'ajustement effectué sur le zonier possède une plus grande amplitude entre la première et la dernière zone que la pente tarifaire. Pour les modalités intermédiaires, le coefficient tarifaire est similaire à celui du modèle.

La volatilité du produit Grêle est importante. Les résultats peuvent être portés par certains exercices aux résultats très dégradés. Il est important donc de juger de la stabilité dans le temps des coefficients calculés. La Figure 47 nous permet de vérifier la stabilité du modèle d'ajustement sur le zonier actuel. Sur le graphique présenté, chaque point d'un axe vertical correspond à un même exercice, chaque courbe de couleur correspond à une classe tarifaire du zonier actuel. Dans le cas d'une stabilité parfaite au cours du temps : les courbes ne se croisent jamais, la classe la plus risquée est la plus haute. Pour un risque aussi volatile que celui des produits climatiques, cette stabilité parfaite est impossible, une forte tempête pourra faire exploser le taux de prime pure d'une classe. Nous chercherons tout de même à réduire au maximum cette instabilité entre exercices.

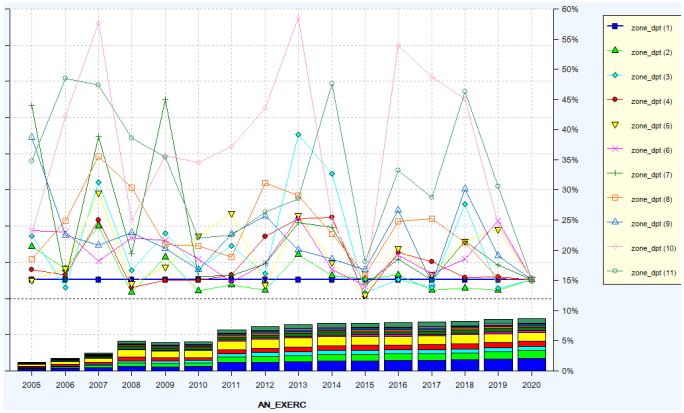


FIGURE 47 – Stabilité au cours du temps du zonier utilisé par Pacifica

Sur la Figure 47, nous remarquons premièrement une stabilité en terme de capitaux assurés par zone au cours du temps, visible grâce aux diagrammes bâtons. Même si notre portefeuille a fortement grandi au fil du temps, la cohérence demeure quant à la proportion de capitaux assurés par zone. Nous remarquons notamment pour 2013, année de forte sinistralité, que la hiérarchie entre les classes de risque n'est pas respectée, témoin d'une instabilité potentielle du zonier.

Une certaine volatilité du taux de prime pure selon les zones existe au cours du temps. Nous constatons des évolutions très différentes entre les zones 10 et 11 par exemple, également entre la zone 6 et la zone 5. Le taux de prime pure de la zone 2 est majoritairement inférieur à celui de la zone 1, mais cette zone 2 subit quelques exercices exceptionnels perturbant la stabilité du zonier.

Nous venons de constater que le zonier utilisé actuellement était cohérent dans sa construction globale. Toutefois, nous avons observé plusieurs axes d'amélioration, axes sur lesquels nous allons essayer de travailler ci-après.

IV.1.3 Reclassement des zones

En reclassant le niveau de risque affecté à chaque département, nous allons chercher à construire un zonier qui offre une pente des coefficients strictement croissante. Nous chercherons également à mieux contrôler la stabilité de l'ajustement au cours du temps. Pour ce faire, chaque zone devra totaliser au moins 5% du portefeuille en capitaux assurés. L'objectif est d'obtenir une segmentation plus importante et plus juste du risque grêle par commune.

IV.1.3.1 Étapes d'optimisation

Pour tenter d'obtenir un meilleur zonier nous procédons par étape :

1. Ajustement du taux de prime pure des départements durant les seize années d'historique.
2. Comparaison de la prime pure ajustée à celle observée sur sa zone d'affectation.
3. Changement de zone :
   - Si un département a un taux de prime pure prédit qui diffère relativement de plus de +20% (respectivement -20%) du taux de prime pure observé, nous diminuons (respectivement augmentons) de deux zones le département.
   - Si un département a un taux de prime pure prédit qui diffère relativement entre +10% et +20% (respectivement -10% et -20%) du taux de prime pure observé, nous diminuons (respectivement augmentons) d'une zone le département. Ces seuils sont arbitraires, ils s'adaptent à notre besoin de reclasser les départements mal affectés sans modifier l'entièreté du zonier actuel.

— La valeur des zones étant bornée par 1 et 11, nous ne reclasserons alors pas des départements au-delà de ces classes. Nous ne déplacerons pas également les départements avec de trop faibles capitaux assurés. Nous faisons confiance au zonier faute d'information suffisante pour le contredire.

4. Ajustement de la prime pure sur les variables habituelles et ce nouveau zonier et analyse de la pente des coefficients.
5. Répétition du processus jusqu'à ce que les reclassements ne soient plus bénéfiques, c'est-à-dire qu'il est impossible d'obtenir une pente strictement croissante et dont la différence de coefficient entre les zones ne peut être significativement augmentée.

L'ajustement du modèle, après un seul processus de reclassement, est représenté en Figure 48 :

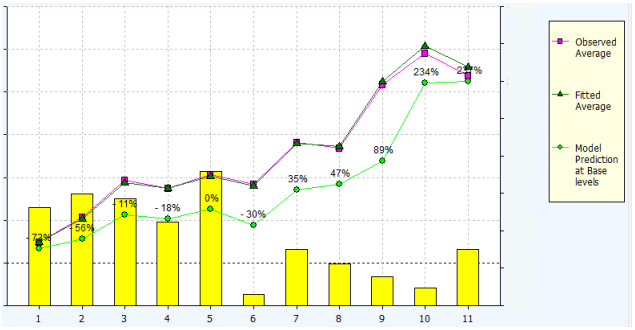


FIGURE 48 – Courbe des coefficients du zonier utilisé par Pacifica, après une étape d'optimisation

Nous remarquons un premier problème : la répartition des capitaux assurés doit être améliorée. La classe 6 est trop faiblement représentée, et de plus son coefficient vient casser la monotonie de la pente des coefficients en vert clair sur la figure ci-dessus. De nouveaux reclassements sont nécessaires pour obtenir une pente strictement croissante.

Pour améliorer notre processus, nous ajoutons un test lors de la comparaison de la prime pure du département avec celui de la zone. Nous allons regarder la moyenne olympique [14], afin d'effectuer un reclassement qui ne sera pas porté uniquement par une valeur extrême. Si la moyenne olympique est similaire à la moyenne de la zone, aucun reclassement sera nécessaire.

De plus, lors ne nos analyses, nous nous apercevons que la classe 10 a des résultats dégradés liés principalement à un exercice. Est-ce que cette sinistralité est due à un très fort aléa ou est-ce bien une zone plus risquée ?

14. moyenne olympique : moyenne sans le minimum et le maximum

IV.1.3.2 Zonier actuel optimisé

Nous avons réitéré plusieurs fois le processus décrit précédemment, tout en ayant un regard sur la moyenne olympique. La pente des coefficients obtenues pour les deux variables géographiques est représentée en Figures 49 et 50 :

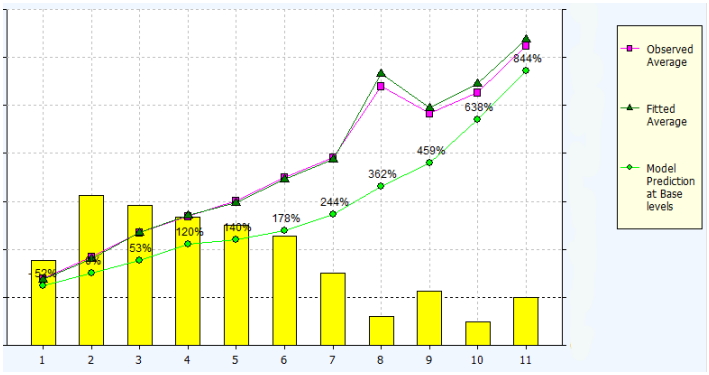


FIGURE 49 – Courbe des coefficients du zonier optimisé

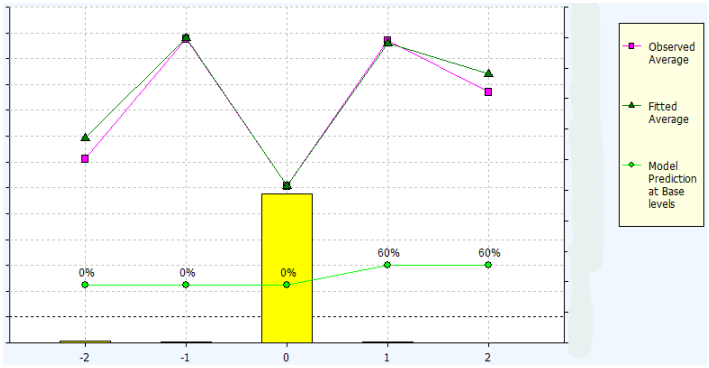


FIGURE 50 – Courbe des coefficients de la variable sur-zone/sous-zone du zonier optimisé

La pente des coefficients du zonier optimisé est strictement croissante. De plus, chaque classe semble bien différente de ses classes voisines, contrairement au zonier actuel où certaines classes voisines ont des coefficients similaires. Nous notons une plus grande amplitude pour ce zonier. Le coefficient tarifaire le plus grand est 20 fois plus élevé que le plus petit, à titre de comparaison il n'est que 7 fois plus élevé dans le zonier actuellement utilisé.

Nous avons également construit des zones mieux réparties en terme de capitaux assurés, même si deux zones, la 8 et la 10, sont légèrement en dessous de la ligne des 5% de capitaux assurés, seuil escompté.

Sur la Figure 50, simplement en reclassant les départements, nous avons réussi à donner un peu plus d'importance aux sur-zones. Les sous-zones n'apportent pas d'information et sont soumises à

une forte volatilité, nous ne retenons donc pas ces modalités au final. Toutefois, ces zones sont si faiblement représentées qu'il est difficile de conclure à une significative amélioration.

Nous avons enfin représenté, en Figure 51, la carte du zonier optimisé à partir du zonier actuellement utilisé par Pacifica. De la même manière que la précédente carte : plus le département est rouge, plus son coefficient tarifaire est élevé.

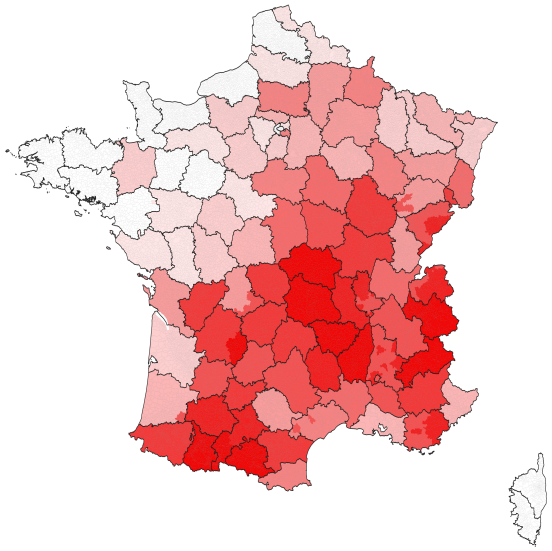


FIGURE 51 – Zonier actuel optimisé

Nous observons que la diagonale de risque n'est plus continue entre les Pyrénées et le Massif Central. Un département comme l'Ardèche a nettement été réévalué et se trouve plus proche de ses voisins en terme de coefficients.

Les reclassements n'ont pas généré de grandes disparités entre départements voisins, dans la majeure partie des cas. Toutefois, certains départements après les reclassements possèdent plusieurs classes de décalages avec leurs voisins. C'est le cas du Jura et du Tarn, jugés nettement moins risqués, de l'Ille-et-Vilaine et du Val-de-Marne, jugés eux plus risqués. Ces départements évoqués ont une faible représentation en portefeuille, moins de 0,2% des capitaux assurés pour chacun d'eux. Face au manque d'information et pour éviter le surapprentissage, nous ne prenons pas en compte le reclassement de ces départements et nous laissons ceux-ci dans leur classe de risques d'origine.

IV.1.4 Possibilités d'amélioration et limites

a) Problème de stabilité

L'instabilité est moins importante que dans le zonier de départ. Comme nous pouvons le voir en Figure 52, la démarcation entre les courbes est plus significative. Cependant, la hiérarchie des taux de prime pure par zone n'est pas respectée pour tous les exercices.

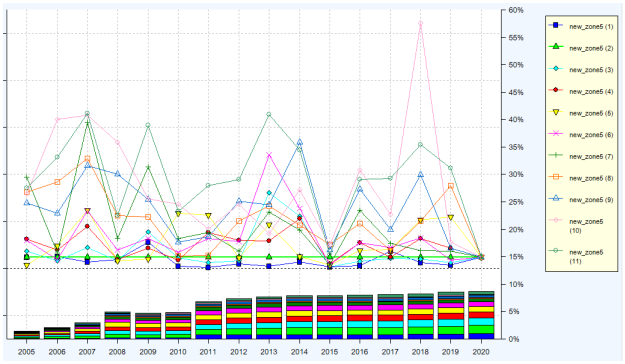


FIGURE 52 – Stabilité au cours des exercices du modèle optimisé

Cette fois la stabilité entre les classes de risque sur l'année 2013, fortement sinistrée, est mieux respectée. Nous constatons toujours cependant que certaines classes de risque sont portées par certaines années fortement dégradées. Ceci symbolise un potentiel surapprentissage. En effet, ce reclassement se base uniquement sur notre historique, il en capte ainsi fortement les bruits.

b) Axes d'amélioration de ce zonier

Parmi les données externes que nous avons recueillies, certaines pourraient permettre de mieux expliquer la sinistralité grêle sur le territoire français. Nous analysons ici pour certaines variables si les variations de prime pure entre les modalités sont captées par le zonier optimisé construit juste avant. Sur la Figure 53 est représenté l'ajustement de la variable "Amplitude maximale de température au court du mois" par le modèle lorsque celle-ci n'est pas utilisée comme variable explicative :

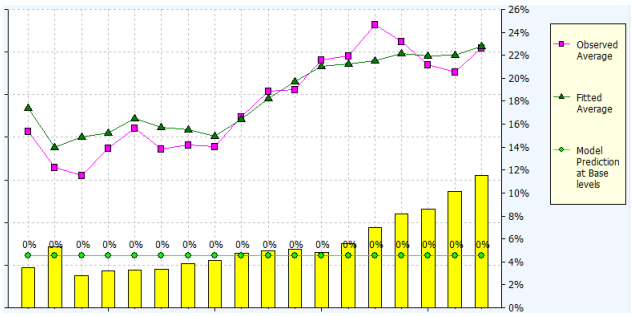


FIGURE 53 – Ajustement de la variable "Amplitude de températures sur la commune", par le zonier optimisé

L'ajustement de la variable "Diamètre maximal moyen des grêlons sur la commune", est visible en Figure 54 :

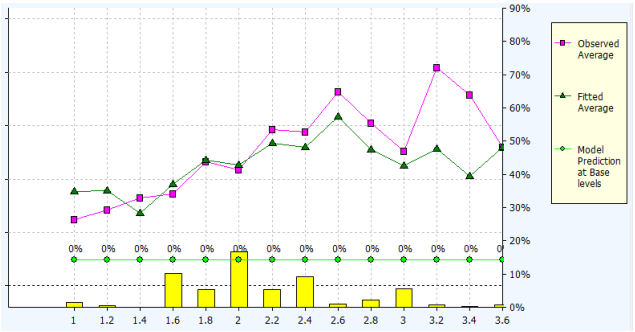


FIGURE 54 – Ajustement de la variable "Diamètre maximal moyen des grêlons sur la commune", par le zonier optimisé

Pour les modalités 1.6cm, 1.8cm et 2cm, le niveau du taux de prime pure est bien ajusté par le modèle par rapport au taux observé. En revanche, ce n'est pas le cas pour les tempêtes avec des grêlons de diamètre maximal supérieur à 2cm, le modèle sous-estime le taux de prime pure pour ceux-ci. Les variations du taux de prime pure pourraient également être mieux captées grâce à la variable "Fréquence de sinistres en Habitation sur la commune", comme l'illustre la Figure 55 :

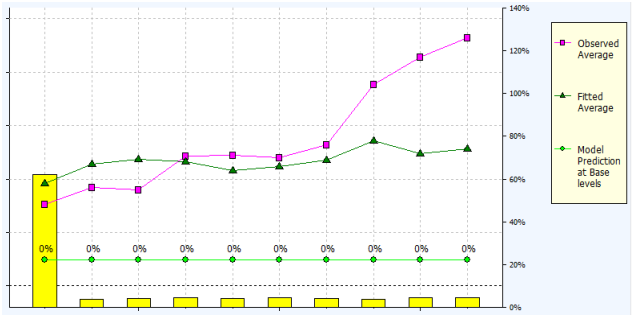


FIGURE 55 – Ajustement de la variable "Fréquence de sinistres en Habitation sur la commune", par le zonier optimisé

Des améliorations significatives du zonier sont donc possibles, notamment en ajustant avec la variable de fréquence des sinistres sur le produit Habitation.

Enfin, hormis les rares zones d'exceptions, nous utilisons dans ce zonier des frontières départementales. Or, rien n'indique que ces frontières soient les plus adaptées pour construire un zonier sur le risque grêle. Un autre axe d'amélioration sera donc de définir des frontières de zones plus adaptées aux coefficients de risque attribués aux communes.

IV.2 Construction d'un zonier sans l'utilisation de données externes privées

Nous allons ici construire un zonier qui n'inclut pas de données externes, privées, recensant les tempêtes de grêles. Ces dernières étant payantes, nous souhaitons savoir si leur achat, régulier ou ponctuel, est pertinent pour améliorer la qualité du zonier. Dans ce modèle, les données internes complémentaires et celles externes disponibles gratuitement seront utilisées pour estimer au mieux le risque grêle sur le territoire français.

IV.2.1 Propagation de l'information des variables explicatives

IV.2.1.1 Propagation des données météorologiques

Plus de 90% des codes Insee n'apparaissent pas dans la base de données "Infloclimat". Le Krigeage est une solution pour diffuser l'information à l'ensemble du territoire.

La Figure 56 représente la France lorsque la variable n'est pas autocorrélée spatialement.

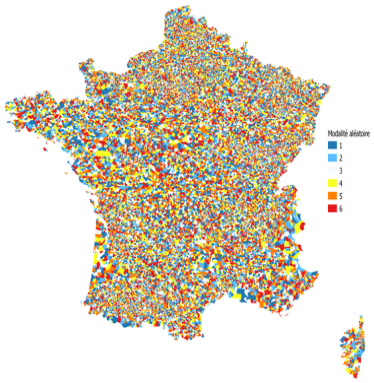
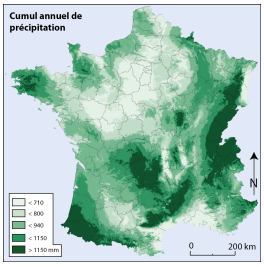
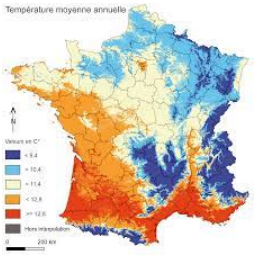


FIGURE 56 – Représentation d'une variable avec une autocorrélation spatiale nulle sur le territoire français



(a) *Volume annuel des précipitations*

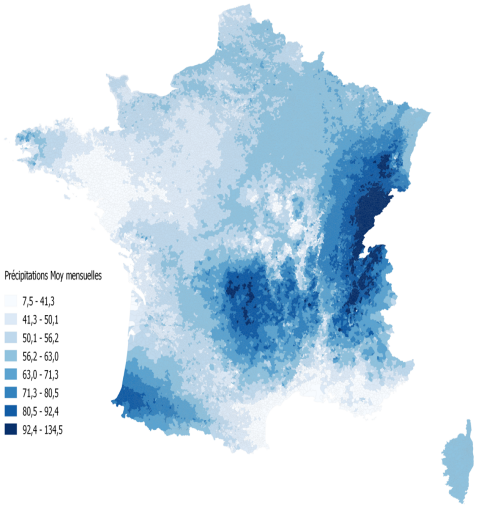


(b) *Températures moyennes annuelles*

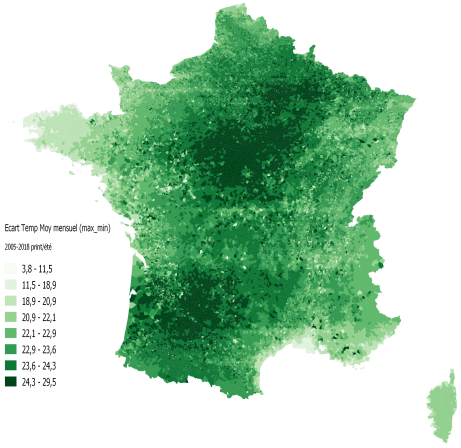
FIGURE 57 – Représentation par commune de variables météorologiques

Sur les cartes en Figure 57 sont représentées les *Températures moyennes annuelles* et le *Volume annuel des précipitations*. Une autocorrélation spatiale existe pour ces variables : les cartes en Figure 57, contrairement à la Figure 56, ne représentent pas une variable aléatoire sur le territoire français. Nous allons donc utiliser le krigeage pour propager les valeurs des stations sur les codes Insee qui nous sont inconnues car une autocorrélation spatiale est bien constatée.

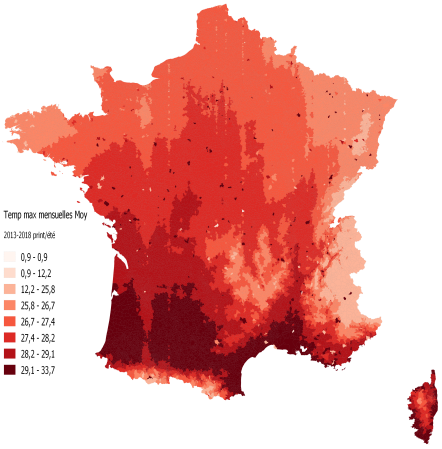
La figure 58 fait apparaître les résultats, après application du Krigeage, sur les données de pluviométrie, d'amplitudes de températures et de températures maximales :



(a) *Volume de précipitions mensuel moyen de 2005 à 2018*



(b) *Amplitude de température maximale mensuelle moyenne sur les mois d'avril à septembre de 2005 à 2018*



(c) *Température maximale mensuelle moyenne sur les mois d'avril à septembre de 2013 à 2018*

FIGURE 58 – Propagation des données météorologiques sur l'ensemble du territoire

Ces lissages restent des approximations de la réalité, mais permettent d'étoffer notre jeu de variables explicatives. Chaque commune possède maintenant une information météorologique. Nous effectuons ce même lissage pour l'ensemble des variables météorologiques disponibles.

IV.2.1.2 Propagation des données du produit Multirisques Climatiques

Nous disposons également de la fréquence des sinistres grêle du produit Multirisques Climatiques. Celui-ci peut nous apporter des informations importantes sur la sinistralité grêle pour chaque commune, notamment pour les communes où nous n'avons pas de contrats Grêle sur l'historique. Cependant, le portefeuille de ce produit, contrairement à celui du produit Habitation, est fortement incomplet géographiquement.

Il est intéressant de regarder si cette variable peut être complétée géographiquement de la même façon que nous l'avons fait pour les températures. La Figure 59 représente la fréquence des sinistres du produit Multirisques Climatiques :

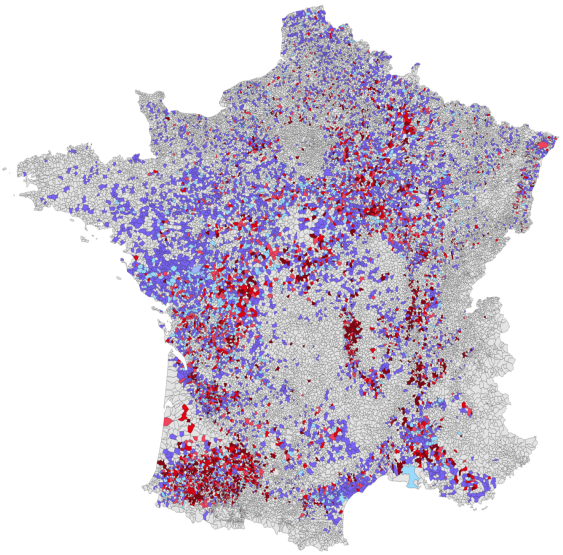


FIGURE 59 – *Fréquence des sinistres u produit Multirisques climatiques en agricole de 2005 à 2020 par commune*

Cette figure met en évidence des zones en moyenne plus grêligènes que d'autres. Nous constatons bien une certaine auto corrélation spatiale car la représentation est nettement différente du cas aléatoire, visible en Figure 56. Toutefois, au sein de ces zones, il existe d'importantes disparités dues à la volatilité du phénomène.

Le principal problème dans le lissage est que, contrairement aux stations météorologiques qui occupent le territoire de manière uniforme, le portefeuille Multirisques Climatiques n'est pas présent uniformément sur le territoire. En effet, nous observons des zones très denses et d'autres totalement vides. La propagation par le krigeage devient compliquée car le paramètre du nombre de voisin devrait alors s'adapter à différents cas.

Finalement, le krigeage n'apparaît pas comme une solution adaptée dans ce cas-ci. Nous utiliserons donc dans notre modèle d'estimation la variable fréquence du produit Multirisques Climatiques à la maille Insee, sans lissage, afin de garder l'information intacte et suffisamment fine.

IV.2.2 Ajustement de la prime pure

La prime pure sera ajustée par modèle linéaire généralisé. Nous allons présenter ici les variables explicatives retenues dans le modèle final, ainsi que les raisons motivant ces choix.

IV.2.2.1 Variables apportant une information non géographique

Ces variables, donnant une information non géographique sur le risque de l'assuré, sont présentes dans l'équation tarifaire. Il s'agit de la franchise, du nombre de rotations et surtout du type de récolte. Ces dernières constituent une partie importante du tarif car selon la récolte, la sensibilité à la grêle diffère grandement.

La franchise et le nombre de rotations n'ont que très peu de modalités et la plupart des observations sont concentrées sur quelques modalités seulement. De plus, elles sont facilement interprétables. Nous insérons ces variables dans notre modèle sans retraitement.

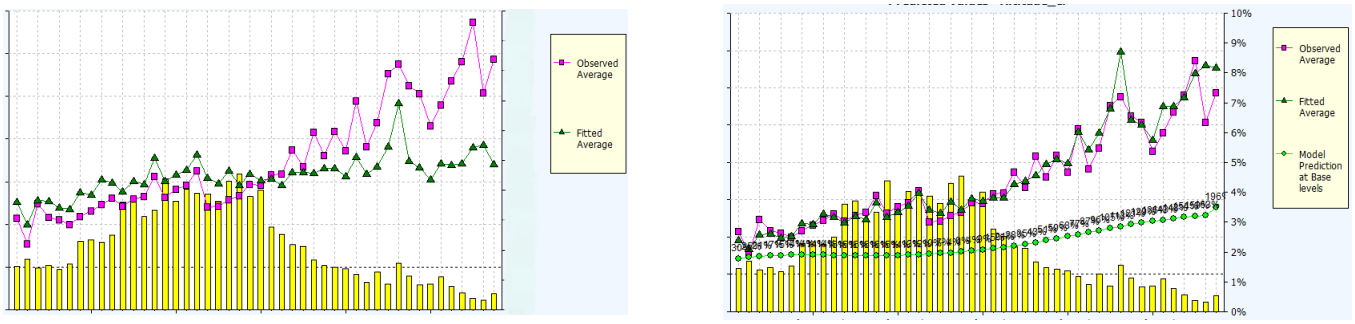
A l'inverse, la variable relative au type de récolte nécessite une compréhension particulière : quelle maille choisir ? comment gérer la volatilité ? faire confiance à l'expert ou à l'historique ?

Nous avions le choix de prendre une variable à la maille groupe de récoltes, avec 11 modalités, ou à la maille nature de récolte avec 197 modalités. La première offre une trop faible segmentation, alors que la seconde fait apparaître une volatilité trop importante avec beaucoup de natures de récolte très faiblement représentées. Nous avons alors fait le choix au final d'ajuster le taux de prime pure par le biais d'une variable intermédiaire aux deux présentées, construite à partir du code récolte. Pour les modalités trop faiblement représentées, nous les regroupons selon leur culture, sinon nous gardons la nature de récolte.

Nous obtenons au final 46 modalités pour la variable relative au type de récolte. Nous nommerons cette dernière *Classe de récoltes*. Certaines cultures demeurent très faiblement représentées et sont sujettes à la volatilité. Nous ne pouvons pas regrouper les cultures à d'autres : leur sensibilité diffère trop. Concernant sensibilité d'une culture, pour éviter d'apprendre à partir d'une information polluée par l'aléa, nous avons échangé avec l'expert agronome de Pacifica en charge du produit. Ce dernier a pu confirmer la cohérence des coefficients obtenus lors de l'ajustement de nos modèles.

IV.2.2.2 Variables apportant une information géographique

Ces variables explicatives vont indiquer différentes caractéristiques des communes ou départements de chaque parcelle en portefeuille. Nous connaîtrons ainsi pour chaque observation : *l'Altitude, l'Amplitude de températures moyennes, la Pluviométrie mensuelle,...* Ces variables sont potentiellement corrélées à la grêle, d'après notre analyse physique et empirique du phénomène. Nous allons étudier les liens de ces variables à la sinistralité grêle du portefeuille et regarder s'il serait pertinent de les intégrer dans notre modèle.



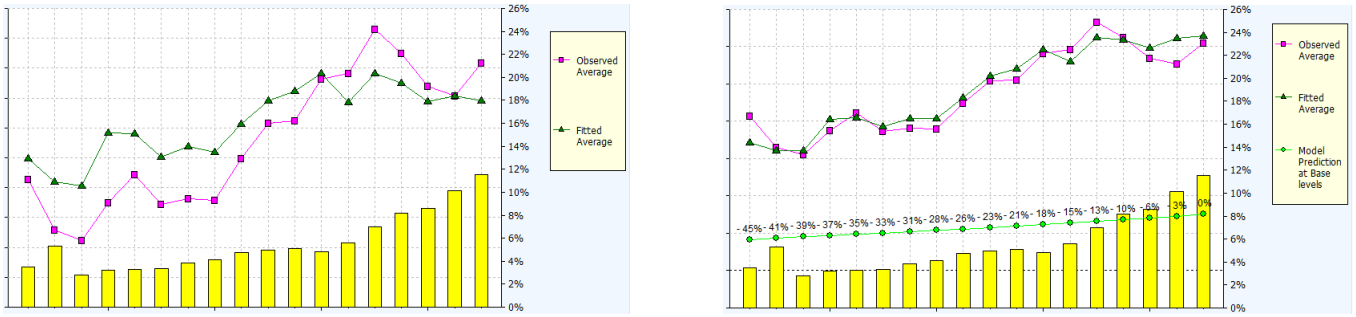
(a) Variable non utilisée dans le modèle

(b) Variable utilisée dans le modèle

FIGURE 60 – Ajustement de la variable *Altitude*

Sur la Figure 60 est comparé l'ajustement du taux de prime pure sur la variable *Altitude* : lorsque celle-ci n'est pas utilisée dans le modèle (Figure 60a) et lorsque celle-ci est utilisée dans le modèle (Figure 60b). L'*Altitude* a un réel intérêt d'apparaître dans le modèle car elle apporte un meilleur ajustement. De plus, la pente des coefficients de cette variable est strictement croissante.

L'ajustement du taux de prime pure comparé en Figure 61 selon si la variable *Amplitude de températures moyennes mensuelles d'avril à septembre* figure dans le modèle. L'historique de la variable correspond à 2005 à 2018. Un lissage polynomiale d'ordre un est effectué sur la pente des coefficients. Nous obtenons une courbe des coefficients strictement croissante.

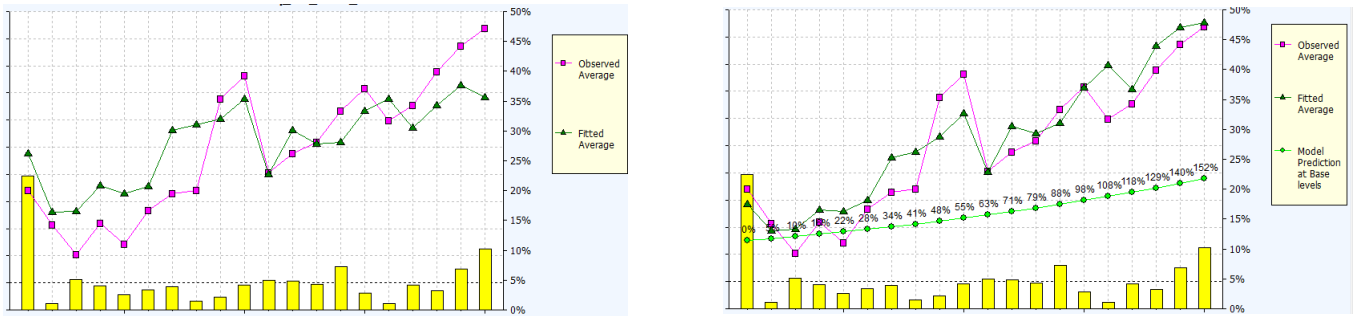


(a) Variable non utilisée dans le modèle

(b) Variable utilisée dans le modèle

FIGURE 61 – Ajustement de la variable *Amplitude de températures moyennes mensuelles d'avril à septembre*

L'amplitude de température permet de réaliser un meilleur ajustement. Nous remarquons bien sur la Figure 61a : une sous-estimation du risque pour les communes où est constatée une faible amplitude, et inversement, une surestimation pour les communes avec une grande amplitude de température.



(a) Variable non utilisée dans le modèle

(b) Variable utilisée dans le modèle

FIGURE 62 – Ajustement de la variable *Fréquence des sinistres grêle Multirisques Climatiques par commune*

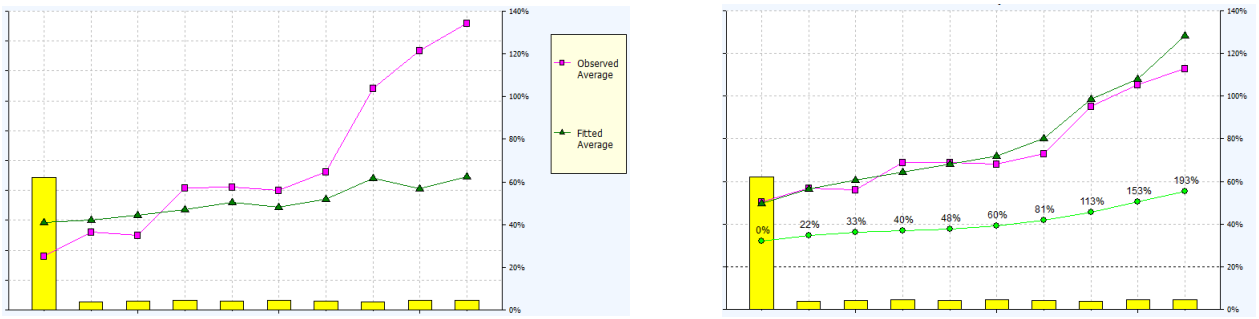
La comparaison des ajustements de la *Fréquence des sinistres grêle Multirisques Climatiques*, représentée en Figure 62, illustre également une information non captée par le modèle. Ces graphiques montrent que l'ajustement est plus fidèle à l'observé grâce à l'utilisation de la variable. De la même manière que précédemment, nous effectuons un lissage d'ordre 1, venant gommer la volatilité. Enfin, l'ajustement pour cette variable n'est pas parfait car ne prend pas en compte toutes les variations, notamment aux extrémités, pour deux raisons :

— Cette variable est fortement incomplète. La première modalité, représentant les communes non présentes dans le portefeuille Multirisque Climatiques, correspondent à 25% de l'exposition.

— Les communes à fortes fréquences ont été, en grande partie, sujettes à un important aléa durant les 15 ans d'historique. Cet aléa, nous ne souhaitons pas le capter.

La *Fréquence des sinistres grêle Multirisques Climatiques* permet de propager, à des communes non représentées dans le portefeuille du produit Grêle, l'information de l'exposition au risque grêle de récoltes.

La variable *Fréquence des sinistres grêle Habitation* apporte une information sur la fréquence des sinistres graves. En Figure 63 apparaît un nombre important de communes ayant une fréquence nulle : 60% des communes du territoire. Nous constatons également que cette variable permet de réaliser un ajustement nettement amélioré sur 15% des communes, celles où les habitations sont souvent sinistrées.

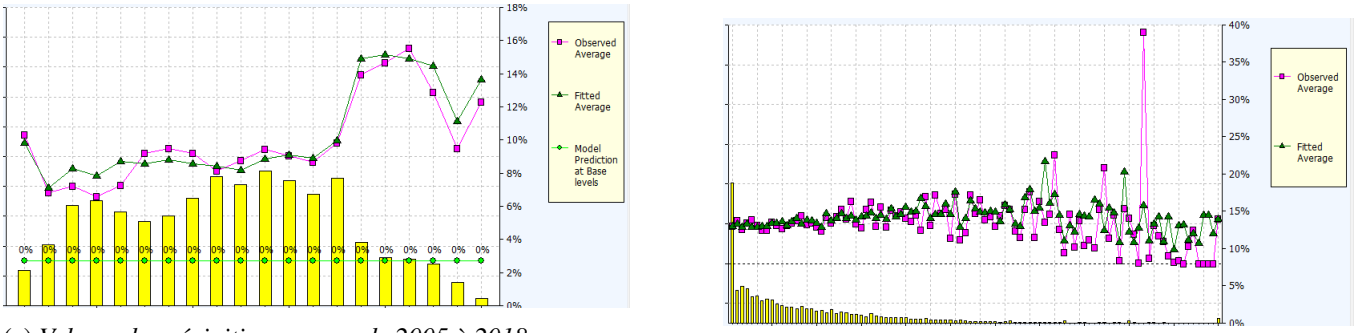


(a) Non utilisée dans le modèle

(b) Utilisée dans le modèle

FIGURE 63 – Ajustement de la variable *Fréquence des sinistres Habitation par commune*

À l'inverse des variables précédemment présentées, beaucoup de variables n'apportent pas d'information pertinentes à l'ajustement du taux de prime pure observé. C'est le cas de la *Pluviométrie* et de la *Proportion de terres forestières*, représentées sur la Figure 64. Les variations du taux de prime pure, sur chacun des graphiques, sont bien captées pour la plupart des modalités. Ces deux variables, ainsi que toutes celles n'apportant pas d'informations pertinentes, ne seront pas retenues dans le modèle d'ajustement.



(a) *Volume de précipitions moyen de 2005 à 2018* par commune

(b) *Proportion de terres forestières* sur la commune

FIGURE 64 – Variables dont le modèle capte déjà les variations

De manière générale, les variables d'occupation des sols de la base Corine Land Cover ne seront pas retenues car elles n'apportent pas d'information. Celles-ci figuraient dans l'étude à titre exploratoire, sans réelle fondement théorique.

Enfin, nous regardons si les variables explicatives, que nous jugeons pertinentes pour notre étude, ne sont pas trop corrélées entre elles. Nous souhaitons, en effet éviter toute redondance dans le modèle et tout biais potentiel à cause de ces relations entre variables explicatives. En cas de très forte corrélation, le modèle est incapable d'associer les variations aux bonnes variables.

La Figure 65 fait apparaître le V de Cramer des variables explicatives retenues dans le modèle d'ajustement du taux de prime pure :

Factor (#Levels)	)	ranch_cl (4)	recolt_cl (25)	ltitude_cl (46)	ns_print_ete_cl (18)	MC_insee_cl (20)	_nh_insee_cl (10)
RANDOM (3)	)	0,000	0,000	0,000	0,000	0,000	0,000
AN_EXERC (16)	)	0,000	0,000	0,000	0,000	0,000	0,000
DEPT (92)	5	0,000	0,000	0,000	0,000	0,000	0,000
franch_cl (4)	2	0,000	0,000	0,000	0,000	0,000	0,000
recolt_cl (25)	7	0,625	0,000	0,000	0,000	0,000	0,000
Altitude_cl (46)	5	0,154	0,110	0,000	0,000	0,000	0,000
TXX_TNN_13ans_print_e	3	0,121	0,078	0,174	0,000	0,000	0,000
freq_MC_dpt_cl (6)	3	0,189	0,272	0,330	0,254	0,000	0,000
freq_MC_insee_cl (20)	2	0,149	0,112	0,151	0,125	0,000	0,000
freq_nh_dpt_cl (5)	5	0,120	0,161	0,259	0,187	0,456	0,000
freq_nh_insee_cl (10)	3	0,065	0,068	0,130	0,083	0,097	0,000

FIGURE 65 – V de Cramer du modèle final

Au-dessus du 0.3 dans le V de Cramer, la littérature s'accorde à dire qu'il y a une relation forte entre les variables explicatives. C'est le cas par exemple de la variable *Classe de Récoltes* (recolt_cl) et de la *Franchise* (franch_cl). La franchise à 20% correspond à une unique culture, les arbres fruitiers, et la modalité 20% de la variable *Franchise* ne contient que des arboricoles. Nous maîtrisons cette corrélation et l'acceptons.

De plus, nous remarquons sur la Figure 65 pour la variable *Fréquence des sinistres Multirisques Climatiques par département* (freq_MC_dpt_cl) que des corrélations fortes existent avec d'autres variables comme : l'*Altitude* (altitude_cl) ou l'*Amplitude de températures moyennes mensuelles d'avril à septembre* (TXX_TNN_13ans_print_ete). Ceci nous conforte dans notre choix de retenir les variables de fréquences des sinistres, sur les autres produits, à la maille communale.

Il est indiqué également qu'une relation moyenne existe entre deux variables lorsque la valeur dans le V de Cramer dépasse 0.2. Nous ne dépassons pas non plus ce seuil pour les variables explicatives retenues dans le modèle d'ajustement :

— *Franchise* (corrélation comprise sur la modalité 20%)

— *Classe de récoltes*

— *Altitude*

— *Amplitude moyenne des températures mensuelles de 2005 à 2018 durant l'été et le printemps*

— *Fréquence des sinistres grêle Multirisques Climatiques*

— *Fréquence des sinistres grêle Habitation*

L'équation utilisée pour réaliser l'ajustement du modèle linéaire généralisé, sans recourt aux données privées sur les tempêtes de grêle, est la suivante :

$$g(E[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{24}$$

Avec : $Y = Y_i, i = 1, \ldots, n$ un vecteur de R_n contenant le taux de prime pure estimé pour les n observations, de l'ordre du million dans notre cas.
Pour i allant de 1 à n, nous avons $X_{1i}, \ldots, X_{6i}$, un 6-uplet contenant l'ensemble des six variables explicatives pour l'observation i.

Les coefficients de risque grêle associés à chaque commune, définissant notre zonier, sont issus du produit des coefficients calculés par le modèles pour les variables explicatives retenues.

IV.2.3 Utilisation du package HClustGeo pour générer des zones

Nous venons de choisir précédemment les quatre variables explicatives retenues pour construire le zonier. Le produit des coefficients ajustés permet de construire le zonier en Figure 66 :

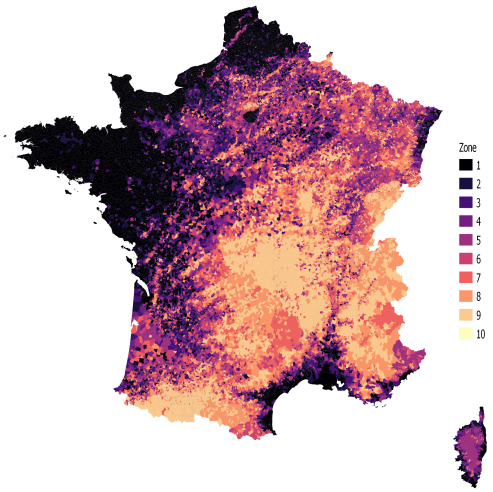


FIGURE 66 – Zonier brut, regroupé en dix classes

Nous avons regroupé les coefficients en dix classes. Ce choix est réalisé dans le but d'obtenir des classes suffisamment représentées en terme de capitaux assurés. Les zones grêligènes déjà évoquées sont retrouvées sur ce zonier. Toutefois, sur la figure ci-dessus, nous remarquons que tarifer à la commune, sans regroupement géographique, est sujet à une variance trop importante. Il semblerait que nous sur-apprenions de notre historique, ceci matérialisé par le dessin des différentes grosses tempêtes de grêle depuis 2005, visibles notamment sur le quart Nord-Ouest avec des trajectoires diagonales. Ce zonier-ci ne mutualise donc pas suffisamment le risque.

La volatilité du risque grêle et la faiblesse de notre historique perturbent notre ajustement à la maille commune. La finesse d'un tel maillage engendre un surapprentissage pour un nombre important de communes. Pour celles-ci, nous n'avons pas réussi, pour l'instant, à différencier l'aléa de l'information pertinente. Ce surapprentissage implique également que la variance de notre ajustement n'est pas contrôlée. Pour pallier à ce problème, une solution est de regrouper des communes voisines et similaires en terme de risque grêle.

IV.2.3.1 Optimisation des paramètres du package HClustGeo

Nous allons chercher à regrouper des communes voisines, qui ont un risque similaire face à la grêle. Ce regroupement fait apparaître plusieurs questions :

— Quelle finesse de maillage conserver ?

— Quel le poids accordé à la *Distance spatiale*, par rapport à la *Distance des coefficients* ?

Ces deux questions correspondent aux deux paramètres du package HClustGeo, déjà présenté. Nous allons optimiser ceux-ci, afin d'obtenir un zonier le plus adapté possible à nos objectifs.

a) Optimisation de l'importance accordée à la distance géographique

Le paramètre α indique au package l'importance à accorder à la distance géographique. Plus α sera proche de 1, plus nous accorderons de l'importance à la distance géographique. Inversement, plus α sera proche de 0, plus nous donnerons d'importance à la distance de coefficients.

Sur la Figure 67 est représenté les niveaux d'explicabilité de la proportion d'inertie de la distance géographique (D1) et celle des coefficients (D0), selon différentes valeurs de α :

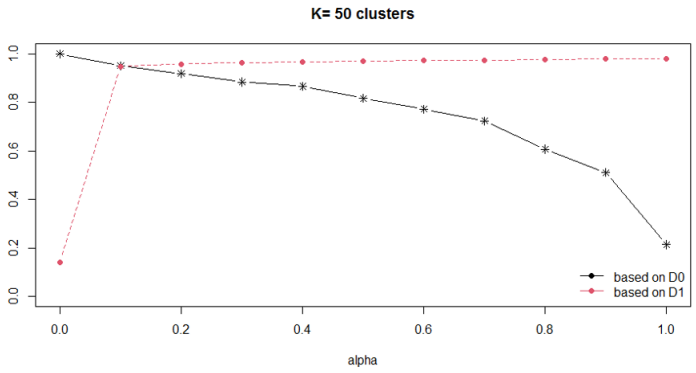
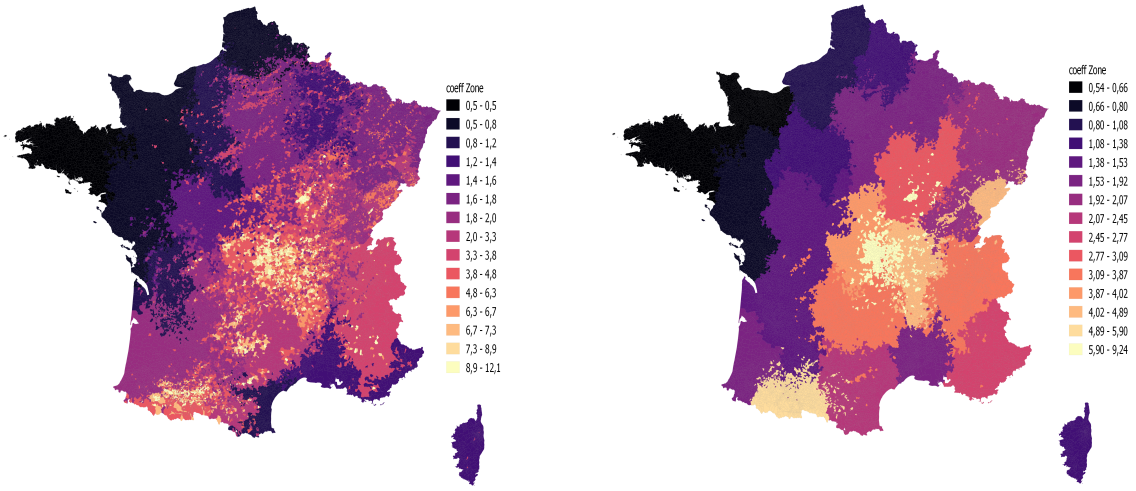


FIGURE 67 – Explicabilité de la proportion d'inertie par les deux distances selon la valeur de α

Selon ce graphique, la valeur α optimale est comprise entre 0.1 et 0.4 car c'est pour ces valeurs que les deux distances contribuent suffisamment. Pour une valeur de α supérieure à 0.4, une forte décroissance de l'exlicabilité par les coefficients (D0) est constatée. A l'inverse, pour un α proche de 0, nous n'utilisons pas suffisamment l'information géographique et restons dans le cas d'un zonier à la maille commune, ce qui n'est pas souhaité. Il difficile de choisir un alpha optimal à partir de ce graphique car nous visualisons mal quelle valeur d'alpha correspond le plus à nos besoins et contraintes.

Compte-tenu du risque grêle très aléatoire et de la volonté de mutualisation, nous allons surpondérer la distance géographique dans le package. Nous allons d'abord tester ci-dessous deux valeurs de α bien distinctes, et ce pour différents nombres de zones [15] car l'influence de α varie selon ce nombre de zones :

i) Visualisation de α avec 25 zones géographiques



(a) $\alpha = 0.3$ (b) $\alpha = 0.9$

FIGURE 68 – Zonier en 25 zones

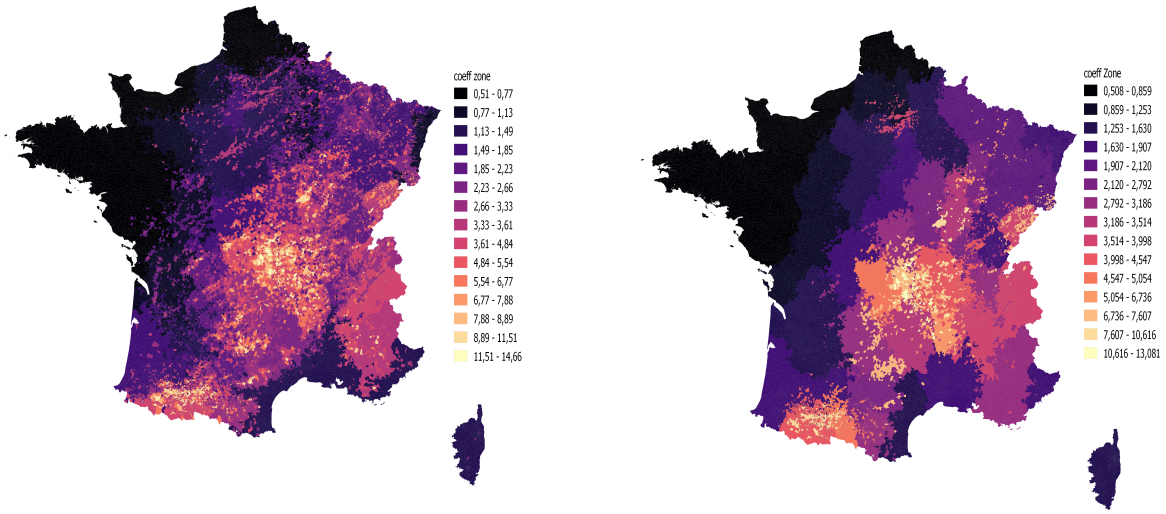
La carte 68a indique que pour un α plus petit, nous avons un maillage très fin, proche de celui à la commune. Le surapprentissage est moindre que pour α nul, mais il semble subsister. Nous distinguons toujours les communes ayant subi une forte tempête durant l'historique.

À l'inverse, la carte 68b, celle représentant le cas accordant beaucoup d'importance à la distance géographique, montre un zonier qui segmente moins les communes risquées. Nous observons à travers celui-ci moins de variance entre communes voisines, ce qui répond à notre défi de mutualisation. Nous retrouvons tout de même des poches de quelques communes risquées dans ce zonier avec un α à 0.9, symbolisant que les communes très sinistrées ont pu être isolées. Toutefois, le zonier visible à travers la carte 68b semble accorder trop d'importance à la géographique : le dessin des zones est tel que chacune d'elles ont une superficie similaire.

15. 15 classes de coefficients sont représentées sur les cartes pour une meilleure visualisation. Une zone correspond à un groupement de communes voisines et ayant des coefficients similaires, alors qu'une classe correspond à un groupement de zones aux coefficients proches moyens

ii) Visualisation de α avec 60 zones géographiques

Pour confirmer nos précédentes analyses, nous visualisons cette fois des zoniers pour 60 zones géographiques créées, ce qui induit un maillage plus fin que dans le cas **i)** (35 zones supplémentaires créées).



(a) $\alpha = 0.3$ (b) $\alpha = 0.9$

FIGURE 69 – Zonier en 60 zones

60 zones sont représentées en Figure 69. Le α optimal semble augmenter avec le nombre de zones créées car une valeur de 0.9 pour le paramètre paraît plus cohérente dans le cas de 60 zones que dans le cas de 25 zones. Ce phénomène est justifié par le fait que l'augmentation du nombre de zones permet de plus facilement différencier géographiquement.

Au final, nous pouvons retenir des quatre figures ci-dessus que choisir un α de l'ordre de 0.3 conduit à une segmentation trop fine du risque qui manquera de stabilité. Nous retiendrons plutôt une valeur proche de 0.9, valeur qui dépendra du nombre de zones retenues.

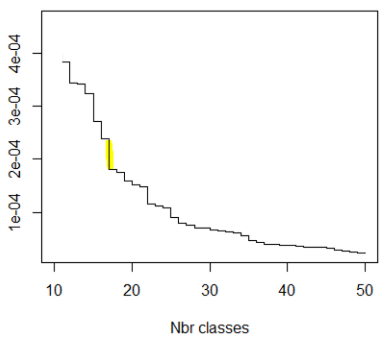
Nous avons maintenant décider d'un ordre de grandeur pour α qui s'adapte à nos contraintes. Pour prendre notre décision, nous analysons à nouveau la Figure 67, qui représentait l'explicabilité de la part d'inertie selon la distance géographique et selon celle des coefficients. Nous remarquons une forte baisse de l'explicabilité par la distance des coefficients (DO) après 0.7. Nous retiendrons donc comme valeur de α 0.7.

b) Choix du nombre de zones

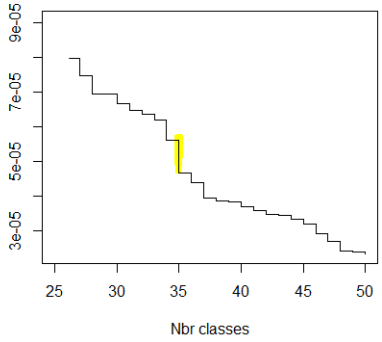
L'optimisation de chacun des deux paramètres étant liée à celle de l'autre, nous avons essayé d'abord de visualiser α dans plusieurs configurations de nombre de zones pour avoir une idée de l'ordre de grandeur du paramètre.

α était difficile à optimiser précisément car nous nous éloignons volontairement de l'optimal théorique, pour éviter le surapprentissage. L'optimisation de k, le nombre de zones retenues, est soumise elle aussi à la contrainte de surapprentissage. Nous fixerons donc une borne supérieure à k pour palier à ce problème. De plus, pour répondre à la contrainte de segmentation suffisante, nous devons fixer une borne minimale.

Nous choisissons une valeur maximale pour k de 100 et une minimale de 10. Ce choix nous permet de segmenter suffisamment le risque sur le territoire et de contrôler la variance en regroupant les communes voisines. En effet, au-delà de 100 zones pour notre cas, nous avons remarqué que la mutualisation du risque devient trop faible et le surapprentissage trop important. Nous allons, à travers les graphiques suivants, essayer de trouver différentes valeurs optimales de k qui nous permettront de visualiser différents regroupements.



(a) Vision globale de l'évolution de l'inertie



(b) Zoom pour k compris entre 25 et 50

FIGURE 70 – Inertie selon le nombre de zones

Sur la Figure 70a, le saut le plus important de l'inertie est visible en jaune, pour un nombre de zones générées de 17. Sur la Figure 70, nous effectuons un zoom pour un nombre de classes compris entre 26 et 50 afin d'observer un k optimal d'un ordre de grandeur différent. Le saut le plus important de l'inertie est également visible en jaune sur cette Figure 70, pour un nombre de zones générées de 35. Enfin, en Figure 71 est représentée la variation d'inertie entre k et k-1. Nous remarquons une forte évolution pour le cas de 59 zones, point le plus haut sur la figure suivante :

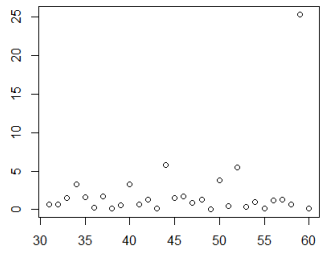
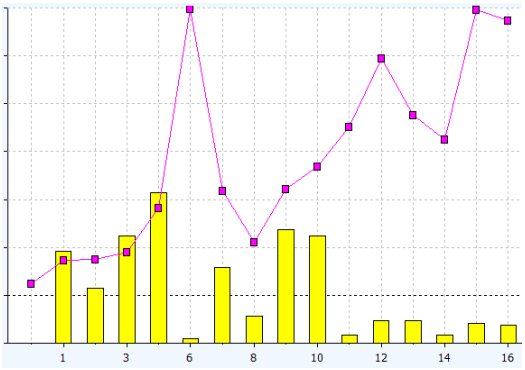


FIGURE 71 – Variation de l'inertie selon k-1 et k zones générées

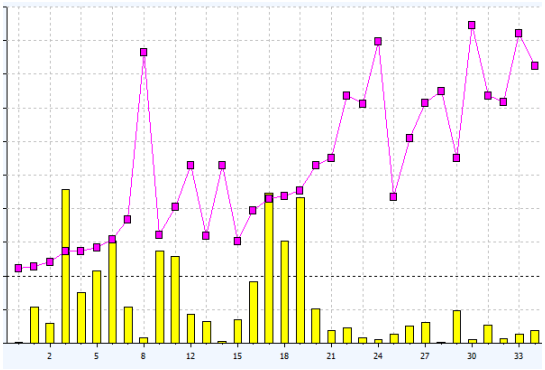
Nous obtenons au final trois valeurs de k optimales, pour trois ordres de grandeur différents. Nous voulions choisir trois valeurs de k bien distinctes et cohérentes afin de comparer les résultats obtenus par différents maillages. Nous retenons donc 17, 35 et 59 comme nombre de zones à analyser dans les modèles suivants.

IV.2.3.2 Regroupement des zones en classes de risque

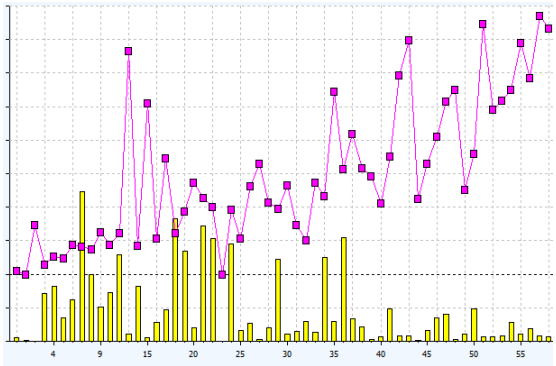
a) Quelle prime pure et quelle exposition par modalité selon le nombre de zones ?



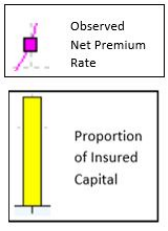
(a) Zonier à 17 zones



(b) Zonier à 35 zones



(c) Zonier à 59 zones



(d) Légende

FIGURE 72 – Valeur du taux de prime pure observé par zone

En Figure 72 est représentée l'exposition en terme de capitaux assurés par zone selon le nombre de classes conservées. Les zones sont classées dans l'ordre croissant de leur niveau de risque. Nous observons que le taux de prime pure observé, en rose, a une tendance croissante du risque. Toutefois, les courbes ne sont pas monotones car les zones n'ont les mêmes caractéristiques.

Sur les trois graphiques en Figure 72, nous remarquons que plusieurs modalités sont très faiblement représentées au sein du portefeuille. Ceci engendre des problèmes de volatilité comme nous pouvons le voir avec la classe 6 en Figure 72a. Aucun des trois zoniers semble offrir une répartition satisfaisante des capitaux assurés. De la même manière que pour le zonier utilisé par Pacifica, nous allons regrouper des zones entre elles pour former des classes de risque afin de réduire la volatilité.

b) Regroupement des zones en classe de risque

Le regroupement en classe se fait manuellement via trois indicateurs :

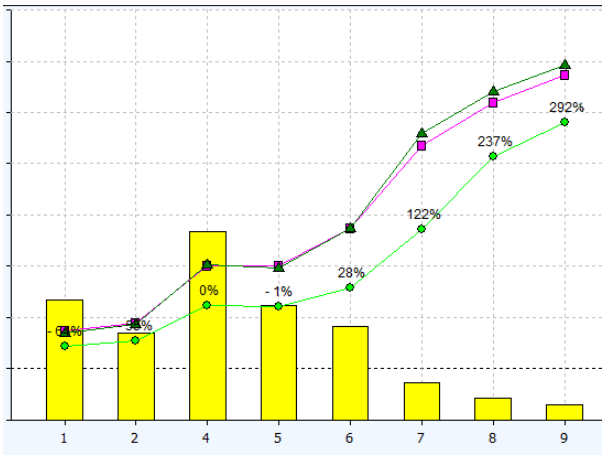
— Les volumes de capitaux assurés d'une classe.

— Les variances réduites intra-zone.

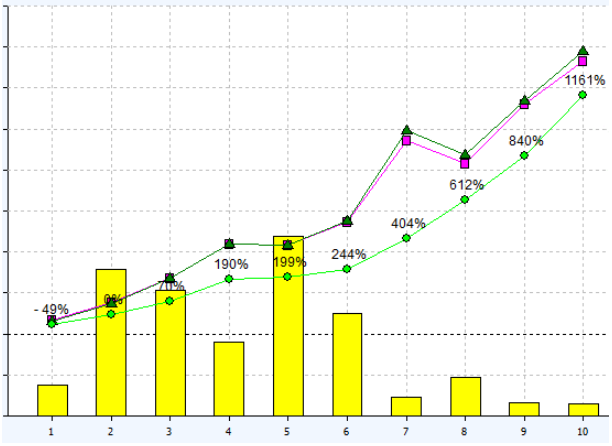
— Les taux de croissance du coefficient moyen entre les classes voisines créées.

Le premier indicateur est utilisé de telle sorte à éviter d'avoir des zones trop faiblement représentées. Nous aurons des classes composées d'au minimum 3% des capitaux assurés en portefeuille. Ce seuil est inférieur à celui précédemment utilisé, 5%, car nos modèles discriminent mieux le risque et nous ne voulons pas perdre trop d'informations à cause de regroupements. Enfin, nous contrôlons le taux de croissance du coefficient d'une classe à l'autre. Nous souhaitons avoir des sauts tarifaires entre classes contrôlés, plutôt qu'une évolution trop irrégulière.

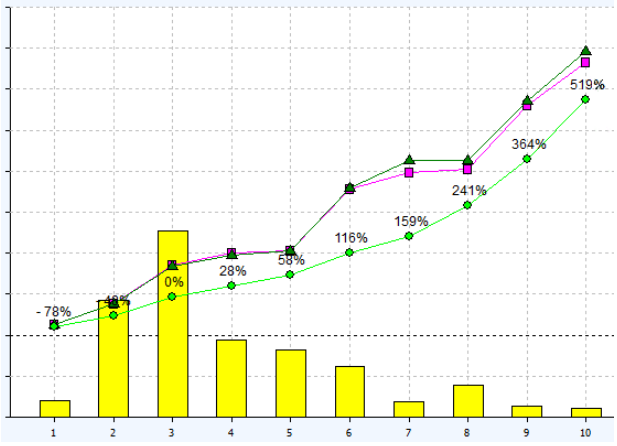
Nous avons effectué les regroupements pour dix classes car ce nombre s'adapte le mieux aux processus de regroupement évoqué. A noter que, pour le cas avec 17 zones où le regroupement en neuf classes s'adapte mieux. Nous avons ensuite ajouté cette variable créée, que nous nommons *Zonier*. Nous obtenons les résultats ci-dessous pour les trois différentes valeurs de k :



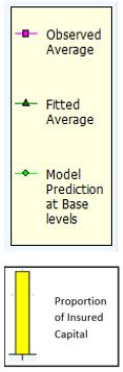
(a) 17 zones



(b) 35 zones



(c) 59 zones



(d) Légende

FIGURE 73 – Ajustement du taux de prime pure selon le nombre de zones du zonier

Nous remarquons tout d'abord sur les trois graphiques en Figure 73 une pente de risque croissante. La croissance est la plus régulière dans le cas du zonier à 59 zones, visible en Figure 73c. À l'inverse, pour le zonier en 17 zones, Figure 73a, nous constatons que les classes 4 et 5 se confondent, témoin d'un risque sûrement mal différencié à cause d'un faible découpage géographique. La même tendance est visible pour le cas des 35 zones. Dans ces deux cas, il semblerait que le modèle ait des difficultés à différencier certaines classes de risque. La différenciation entre les groupes n'est alors pas suffisante car la segmentation géographique est trop faible.

Nous retenons donc un zonier avec un alpha à 0.7 et un nombre de zones à 59 car ce découpage offre une plus grande finesse et une stabilité maîtrisée. Par la suite, nous allons regarder si d'autres méthode ou d'autres données peuvent permettre d'améliorer la segmentation du risque tout en contrôlant suffisamment la volatilité.

IV.3 Utilisation des résidus pour améliorer le zonier

Le précédent modèle construit utilise plusieurs variables explicatives afin d'améliorer le taux de prime pure. Toutefois, nous avons constaté plusieurs freins à ce bon ajustement :

— La grêle est un phénomène dont la localisation n'est pas entièrement expliquée par les caractéristiques du territoire.

— Les variables explicatives retenues manquent parfois de précision.

— Notre ensemble de variables pour expliquer la grêle n'est pas exhaustif.

L'utilisation des résidus face à ces constats peut être une solution afin d'améliorer notre connaissance de la sinistralité grêle des communes du territoire. Nous allons regarder dans un premier temps si les résidus sont répartis aléatoirement sur le territoire, et ensuite voir si nous pouvons les utiliser pour mieux ajuster notre zonier. Il faudra être précautionneux dans le traitement de ces résidus car ils peuvent transmettre une information erronée à cause du bruit qu'ils comportent.

IV.3.1 Choix du type de résidus

Les résidus ont pour objectif d'apporter une information supplémentaire non transmise par les variables explicatives retenues. L'enjeu sera de différencier l'information apportée par les résidus entre l'aléa sur la zone et les tempêtes de grêle récurrentes. Ces tempêtes peuvent s'apparenter à la détection d'un couloir de grêle non expliqué par l'ensemble de variables explicatives utilisées actuellement.

L'important dans un premier temps est de recueillir un maximum d'information sur la sinistralité, non captée par le modèle précédent. Pour obtenir cette information de la meilleure façon possible, nous avons le choix entre deux types de résidus : relatifs et additifs. Pour chaque commune, le résidu est obtenu ainsi [16] :

— Le Résidu Relatif d'une observation i, nommé RR_i :$RR_i = \frac{TxPPobs_i}{TxPPaj_i}$

— Le Résidu Additif d'une observation i, nommé RA_i : $RA_i = TxPPobs_i - TxPPaj_i$

Pour des résidus relatifs, l'objectif est d'obtenir une valeur proche de 1, témoin que l'ajusté est proche de l'observé. Les résidus relatifs engendrent un problème dans notre cas : l'information captée est erronée lorsque la charge des sinistres est nulle sur l'historique. Nous avons alors des résidus relatifs automatiquement nuls, quelque soit le taux ajusté, à cause d'une prime pure

16. $TxPPobs_i$: Taux de prime pure observée sur l'observation i
$TxPPaj_i$: Taux de prime ajustée sur l'observation i

observée nulle. Ces cas sont très présents avec le risque grêle. Nous ne retiendrons donc pas les résidus relatifs comme variable explicative supplémentaire de notre modèle d'ajustement.

Les résidus additifs sont en théorie moins adaptés dans notre cas car nous sommes dans le cadre d'un modèle multiplicatif. Les résidus normés pourraient être une solution.

Le territoire est tel que les taux de prime pure n'ont pas le même ordre de grandeur d'une zone à l'autre. Il pourrait être judicieux de normer les résidus additifs avec le taux de prime pure ajusté, toujours différent de zéro lui. Ces résidus, nommés RN_i, sont présentés dans la formule ci-dessous :

$$RN_i = \frac{(TxPPobs_i - TxPPaj_i)}{TxPPaj_i} \tag{25}$$

Toutefois, nous observons un problème avec cette méthode car l'information des résidus est modifiée. L'information décrivant la forte exposition au risque grêle sur la zone tend à disparaître puisque l'ordre de grandeur devient similaire pour toutes les observations. Ceci est illustré à travers les deux exemples chiffrés suivants. Nous calculons les résidus normés, RN, pour les communes i, faiblement risquées, et j, fortement risquées :

Cas où i et j surestimées

$$RN_i = \frac{0 - 0.0008}{0.0008} = RN_j = \frac{0 - 0.028}{0.028} = -1 \tag{26}$$

Cas où i et j sous-estimées

$$RN_i = \frac{0.0016 - 0.0008}{0.0008} = RN_j = \frac{0.056 - 0.028}{0.028} = 1 \tag{27}$$

Dans ces exemples, la normalisation des résidus additifs empêche de capter l'importante erreur d'estimation réalisée sur les observations j. Cette dernière est considérée similaire à l'observation i suite à la normalisation. Pourtant, l'observation i dans les deux cas correspond à un ajustement satisfaisant, alors que pour les observations j non. Les résidus additifs dans les deux cas ne déforment pas l'information et montrent bien l'erreur d'ajustement sur les observations j.

A noter que, les résidus additifs considèrent les deux cas suivants comme de bons ajustements :

$$RA_i = 0.002 - 0.001 = RA_j = 0.036\text{–}0.037 \tag{28}$$

Ces deux applications sont conformes à nos besoins. Au final, nous retenons les résidus additifs pour capter l'information non contenues dans les variables explicatives.

IV.3.2 Analyse des résidus additifs

Les résidus additifs du modèle d'ajustement construit précédemment sont représentés en Figure 74 :

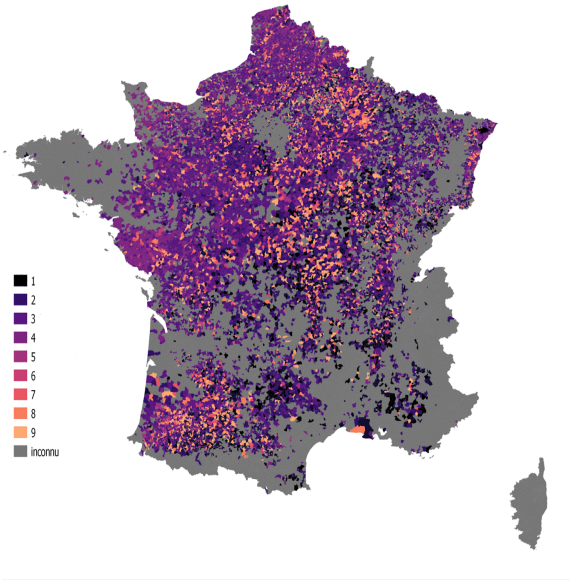


FIGURE 74 – Valeur des résidus additifs, issus du GLM précédent, sur les communes représentées dans le portefeuille

Nous remarquons des départements avec de forts résidus sur un nombre significatif de communes, et particulièrement dans le Gers et la Nièvre. Le premier est un département déjà évoqué pour son caractère fortement grêligène, alors que le second ne présente pas à première vue de risque. Pour la Nièvre, de forts résidus sont également visibles sur des zones limitrophes : des morceaux du Cher, de l'Allier et de l'Yonne. Nous notons une petite poche dans les Ardennes avec des résidus importants.

Nous pourrions, par l'intermédiaire d'un lissage spatial sur la zone, créer une variable *Grandes zones de résidus*. Celle-ci nous permettrait alors de résoudre le problème de sous-estimation du risque pour ces départements, ou bien de surestimation comme c'est le cas pour les Alpes de Haute-Provence, l'Aveyron et les Vosges. Ce lissage large ne nous intéresse pas dans cette partie car nous souhaitons, à travers la variable résidus, détecter les couloirs de grêle, qui sont à une maille plus fine que la maille départementale.

En plus de ces départements, nous observons en Figure 74 des dizaines de communes dont les résidus sont significativement positifs. Ces communes peuvent représenter des couloirs de grêles et

nous perdrions cette information avec un lissage trop important.

De plus, pour reprendre les départements évoqués précédemment, nous notons des poches avec de très faibles résidus dans la zone du Gers et celle de la Nièvre, voisines de communes avec des forts résidus. Sommes-nous face à des ensembles de communes voisines aux risques grêles très hétérogènes ? Ou bien cette différence est-elle due à l'évolution de notre portefeuille durant l'historique ?

Nous remarquons enfin dans la partie Nord-Ouest du territoire des résidus qui appartiennent à une classe intermédiaire, témoin de la bonne appréhension du risque grêle par le modèle sur le territoire. Cette partie du territoire est plus facile à appréhender pour deux raisons : son volume de capitaux assurés est très important sur l'historique et l'incidence de la grêle à cet endroit est faible.

A partir de ces observations, nous considérons qu'une certaine autocorrélation géographique des résidus additifs existe. Nous nommerons à présent cette variable : *Résidus*. Cette autocorrélation est suffisante pour qu'un lissage soit intéressant à implémenter. L'inconvénient de la variable *Résidus* est qu'elle ne peut être construite que pour les communes dont nous avons des contrats en portefeuille. Or, comme nous l'avons présenté, le produit grêle est représenté seulement sur une partie d'une territoire. Le lissage peut permettre de compléter cette variable. Il permettra également d'effacer une partie du bruit porté par les forts résidus.

IV.3.3 Lissage des résidus

IV.3.3.1 Descriptif de la méthode retenue pour lisser

a) Krigeage

Nous avons déjà présenté les bénéfices du krigeage, mais nous ne le retiendrons pas comme méthode de lissage spatial des résidus pour plusieurs raisons :

Premièrement, l'algorithme de krigeage est complexe. Nous souhaitons effectuer un lissage dont l'analyse de l'influence de chacun des paramètres sera aisément identifiable. Nous ne pourrons pas, avec le krigeage, optimiser simplement ces paramètres en fonction de nos contraintes et justifier nos choix. Notamment, nous voulons maîtriser le degré de lissage, pour ne pas trop déformer l'information en lissant.

Deuxièmement, les résultats d'un lissage par krigeage ne sont pas optimaux car l'algorithme a du mal à s'adapter à l'hétérogénéité du nombre de voisins selon les observations. En effet, notre

portefeuille est représenté non uniformément sur le territoire : certaines communes ont des voisins limitrophes fortement représentés sur l'historique, alors que certaines zones n'ont aucun voisin sur un rayon de plusieurs kilomètres. Ce problème rejoint le premier, il est difficile de paramétrer un modèle de krigeage s'adaptant aux différents cas.

Troisièmement, l'ensemble d'apprentissage passe de quelques centaines de communes dans le cas des données météorologiques, à plusieurs milliers de communes dans le cas des résidus. Ceci vient complexifier le modèle et ainsi augmente considérablement le temps de calcul.

b) Interpolation linéaire

Le portefeuille est non uniformément réparti sur le territoire. La grêle quant à elle est sujette à un fort aléa. Afin de pallier à ces deux problèmes, nous allons lisser les résidus. Un lissage trop important fera disparaître à la fois une grande partie du bruit, mais également disparaître les informations pertinentes non captées par le modèle. Un lissage faible ne réduira pas suffisamment le bruit. Il nous faut alors optimiser notre lissage, s'apparentant à répondre aux questions suivantes :

— Quelle taille de rayon devons-nous choisir pour lisser ?

— Quel poids donner aux communes voisines ? Poids proportionnel à leur exposition dans le portefeuille ? à leur distance ? une combinaison des deux ?

— Quelle importance donner à l'information donnée par la commune que nous estimons ?

Nous allons lisser les résidus par le biais d'une moyenne pondérée comprenant plusieurs paramètres. Cette moyenne pondérée permet une bonne compréhension des paramètres, facilitant l'optimisation. Nous utilisons la formule suivante pour calculer ces résidus lissés RL_j :

$$RL_j = \frac{RA_j * k_j * P + \sum_{i=1}^{N}(RA_i * k_i)}{k_j * P + \sum_{i=1}^{N}(k_i)} \tag{29}$$

Où :

— j est la commune dont nous souhaitons calculer le résidu lissé RL_j. Nous utilisons le résidu additif de j, RA_j, et sa pondération, k_j.

— i représente une commune parmi les N communes voisines sélectionnées, possédant un résidu additif, RA_i, et une pondération, k_i.

— P correspond au poids accordé à la commune d'intérêt.

Nous allons donc chercher à optimiser le nombre de voisins retenus de la commune j, le poids P accordé à la commune j et choisir la bonne variable de pondération accordée aux communes. Il faudra également trouver le bon nombre de classes de résidus à retenir, afin d'éviter de construire une variable qui surapprenne et qui soit sujette à une importante volatilité.

IV.3.3.2 Optimisation de la sélection des voisins, N

La distance de sélection des voisins est le paramètre indiquant combien de voisins et quels voisins nous allons prendre en compte dans le lissage de chaque observation. Pour effectuer ce choix, nous recensons nos besoins et contraintes quant à cette variable. Le premier est de conserver l'information locale mise en évidence par les résidus. La grêle est un phénomène pouvant toucher une commune alors que ses communes voisines sont épargnées. Nous faisons donc le choix de conserver seulement les voisins limitrophes lors du lissage. Le bruit pourra rester conséquent pour certaines observations, pouvant occasionner un fort apprentissage. Nous ne pouvons pas différencier la pertinence de l'information des voisins limitrophes afin de n'en sélectionner que certains pour le lissage. Nous prenons donc tous les voisins limitrophes présents en portefeuille pour effectuer le lissage.

De plus, dans le cas d'une commune non présente en portefeuille, mais dont au moins un voisin y figure, celle-ci prendra l'information contenue dans les résidus de ses voisins limitrophes.

IV.3.3.3 Optimisation du poids accordé à l'information de chaque commune du lissage, k

Chaque observation retenue pour effectuer le lissage d'une observation, que ce soit le voisin ou la commune d'intérêt, ne transmet pas la même information. Certaines observations sont présentes en portefeuille seulement quelques années, certaines totalisent peu de capitaux assurés sur l'historique. De la même manière que nous utilisons le taux de prime pure, nous pourrions effectuer une pondération par le volume de capitaux assurés sur la commune. Si nous considérons qu'une information dépend plutôt du nombre d'année de présence en portefeuille, alors le lissage pourrait s'effectuer par le nombre d'exercices d'exposition de la commune. Une solution qui combine les deux pondérations pourrait être également mise en place.

Pour ce paramètre, il est difficile de connaître la bonne valeur de la pondération à travers une fonction d'optimisation. Chaque commune porte une information, plus ou moins bruitée. Il est souvent impossible de savoir quel voisin contient l'information la plus juste. Nous pouvons comme pour le paramètre précédent décider selon nos contraintes et besoins. Le problème de la pondération

par le volume de capitaux assurés est que nous accorderons un poids trop important aux cultures à fortes valeurs. À partir d'un volume important de capitaux assurés, nous pouvons considérer le volume de capitaux assurés suffisant pour une information de qualité. Par exemple, prenons une commune de 10 millions de capitaux assurés et une de 20 millions, nous pouvons considérer le volume suffisant dans les deux cas. Toutefois, la pondération classique conduit à ce que l'information de la commune avec 20 millions de capitaux soit considérée, lors du lissage, deux fois plus importante que l'autre. Nous retenons donc une pondération selon le nombre d'années d'exposition du voisin car chaque année supplémentaire en portefeuille apporte une information supplémentaire proportionnellement, sans que nous devions fixer un seuil arbitraire.

IV.3.3.4 Optimisation du poids accordé à l'information de la commune d'intérêt, P

Le lissage a pour intérêt de capter l'information des voisins, notamment lorsque la commune d'intérêt manque d'exposition durant l'historique. Toutefois, la variable ***Résidus***, dans le cadre de notre modèle, a vocation d'être une variable très différenciante géographiquement nous permettant de localiser les couloirs de grêle. Il faut donc conserver au maximum l'information disponible sur la commune et ne pas trop donner d'importance aux voisins.

Deux types de poids, dont les valeurs seront à déterminer, peuvent être utilisés pour donner une importance supplémentaire à la commune d'intérêt par rapport à ses voisins :

— Un poids constant, P_f, appliqué à la commune que nous cherchons à lisser.

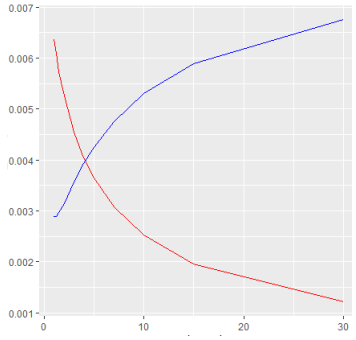
— Un poids variable, P_v, fonction du nombre de voisins que la commune d'intérêt possède et qui sont présents en portefeuille. Il est calculé ainsi : $P_v = max(w * Nbrvoisins, 1)$

Nous allons tester les deux poids. L'objectif sera de minimiser la variance intra-groupe sous contrainte de ne pas faire exploser le biais, ce qui arriverait si nous ne donnons pas assez d'importance à la commune lissée. Nous allons donc chercher à minimiser le carré de l'erreur quadratique :

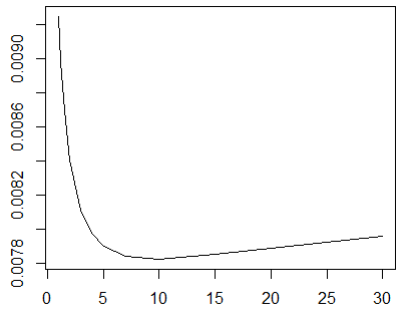
$$MSE = E[(\hat{\theta}_s - \theta)^2] = Biais^2(\hat{\theta}_s, \theta) + Var(\hat{\theta}_s) \tag{30}$$

Où θ_s est le résidu brut et $\hat{\theta}_s$ est le résidu lissé.

Sur les graphiques présents en Figures 75 et 76 est représenté l'évolution du biais, de la variance et du carré de l'erreur quadratique de différents lissages des résidus. Sur ces graphiques, nous représentons ces indicateurs selon différents P potentiels, c'est-à-dire différentes valeurs de pondération par poids fixes et par poids variables :

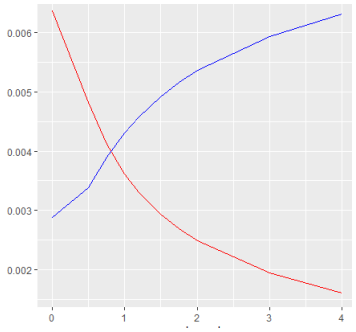


(a) Évolution du biais de lissage (rouge) et de la variance intra-groupe (bleue)

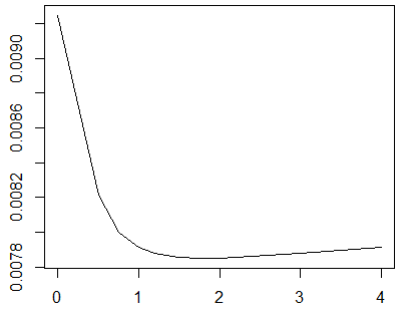


(b) Évolution du carré de l'erreur quadratique

FIGURE 75 – Analyse de la pondération selon la valeur du poids fixe, P_f



(a) Évolution du biais de lissage (rouge) et de la variance intra-groupe (bleue)



(b) Évolution du carré de l'erreur quadratique

FIGURE 76 – Analyse de la pondération selon la valeur du paramètre w dans le cas du poids variable, P_v

Pour les deux types de poids représentés, la fonction est minimisée pour une valeur d'indice de 0,783%, obtenue pour une valeur de w valant 1,5 dans le cas variable et un poids de 7 pour le cas fixe. La variance intra-groupe permet de montrer quel lissage limite le mieux la volatilité. Un poids constant de 7 nous donne la variance intra-groupe la plus faible, 0,47%. Dans le cas variable, la variance intra-groupe est de 0,48% pour un poids équivalent à 1,5 fois le nombre de voisins, valeur minimisant le carré de l'erreur quadratique.

Ces poids offrent des performances quasi similaires. Malgré une variance intra-groupe légèrement inférieur pour le cas variable, nous décidons au final de choisir un lissage via un poids P de 1,5 fois le nombre de voisins. Cette méthode semble mieux s'adapter au voisinage de chaque commune.

Cette méthode nous permet d'effectuer un lissage dont nous maîtrisons la déformation de l'information. Elle s'adapte également au caractère incomplet et non uniforme du portefeuille sur le territoire.

A noter que, l'indice est construit de telle sorte que nous accordons autant d'importance au biais

du résidu lissé et de l'écart-type intra-groupe de voisins. Dans une étude cas plus approfondie, nous pourrions peut-être accorder plus d'importance à l'un qu'à l'autre via la mise en place de poids à l'intérieur de la formule du carré de l'erreur quadratique, selon ce que nous voulons contrôler.

IV.3.4 Utilisation des résidus lissés dans le modèle

Nous venons de créer la variable *Résidus Lissés*, que nous pouvons ajouter dans le modèle d'ajustement construit auparavant. Nous effectuons un dernier retraitement sur les classes 9 et 10 de cette variable *Résidus Lissés*, classes avec les résidus les plus forts. Nous forçons la valeur des coefficients de cette modalité en deçà de l'ajustement indiqué par le modèle. Face à l'instabilité remarquée, nous voulons éviter d'effectuer un apprentissage à partir d'une information bruitée. Nous diminuons donc volontairement l'influence de la variable *Résidus lissés* sur cette modalité. Sur la Figure 77 et la Figure 78 qui suivent nous avons représenté l'ajustement de notre modèle sur le taux de prime pure avec et sans l'utilisation de la variable *Résidus lissés* afin d'observer l'information apportée par celle-ci.

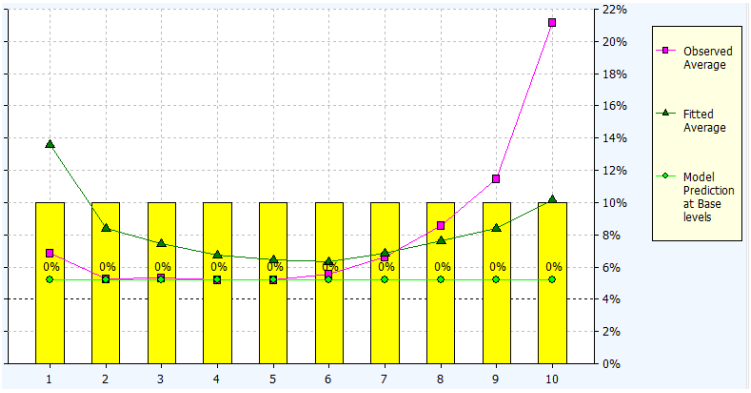


FIGURE 77 – Ajustement du taux de prime pure par un modèle n'utilisant pas la variable *Résidus Lissés*

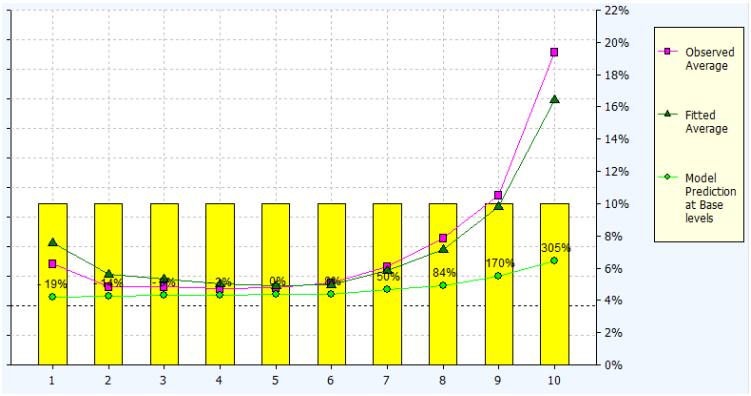


FIGURE 78 – Ajustement du taux de prime pure par un modèle utilisant la variable *Résidus Lissés*

Nous remarquons d'abord que les écarts entre le taux prime pure ajusté sans les *Résidus Lissés* et le taux de prime pure observé est très importante. Une surestimation est nette au niveau des classes avec de faibles résidus. A l'inverse, une très importante sous-estimation est à signaler sur la dernière classe.

En Figure 78, nous remarquons qu'un multiplicateur de trois entre la première et la dernière classe est obtenu sur la pente des coefficients, ce qui est très important. Nous augmentons donc l'amplitude tarifaire entre certaines commune d'un multiplicateur de trois. L'utilisation de cette variable semble apportée une information très différenciante. Nous représentons également en Figure 78 la valeur moyenne de la prime pure pour chaque modalité par exercice, pour juger de la stabilité :

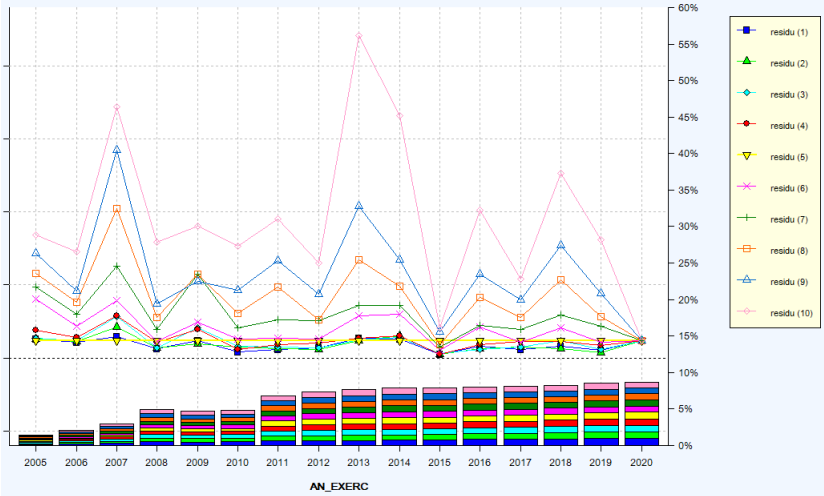


FIGURE 79 – Stabilité de la variable résidus pour chaque exercice

L'évolution de la prime pure pour chaque groupe de *Résidus lissés* est stable au cours du temps, c'est-à-dire que la hiérarchie entre les classes est conservée et ce pour la majorité des exercices. Il y a très peu de croisements entre les courbes pour les classes à forts résidus, notamment une hiérarchie respectée pour les années fortement sinistrées.

Enfin, nous répétons le processus de classification géographique avec le package HClustGeo pour créer la variable *Zonier résidus*. Cette fois le coefficient tarifaire global pour chaque commune, en entrée de la matrice des distances, est obtenu avec le produit du coefficient des variables explicatives classiques ainsi que de celui de la variables *Résidus lissés*.

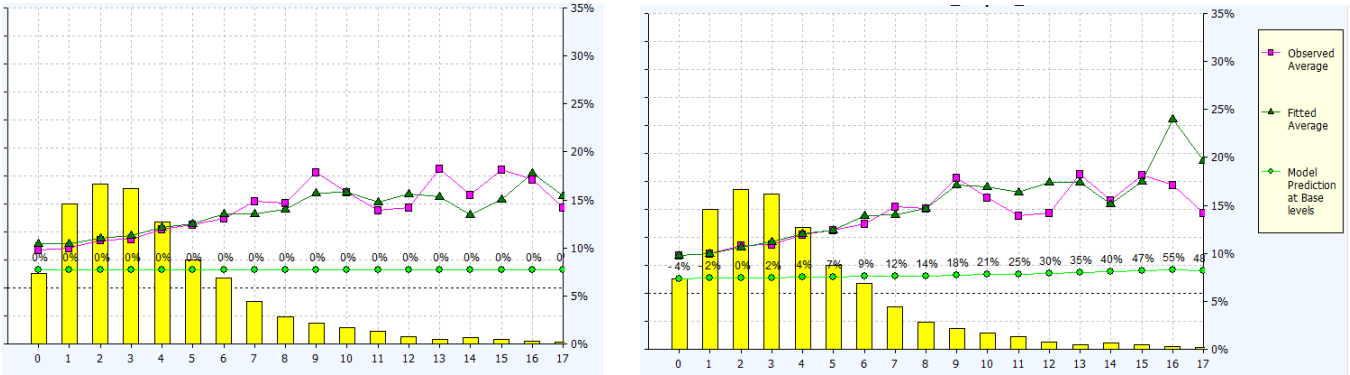
IV.4 Construction d'un zonier avec l'utilisation de données externes privées

A présent nous allons réaliser un zonier à partir de données privées recensant les tempêtes de grêle. Ces données, déjà présentées, ont pour but d'apporter des informations sur la sinistralité grêle en France. Toutefois, ces données ont un coût. Nous allons donc construire un zonier avec celles-ci afin d'observer si leur achat est pertinent et offre des résultats significativement meilleurs que pour les précédents zoniers obtenus.

Pour construire ce zonier, nous allons effectuer le même processus que lors de la construction du zonier sans l'utilisation de données externes privées. Lors de la première étape, nous allons sélectionner les variables pertinentes pour effectuer l'ajustement par un modèle linéaire généralisé. Ensuite, nous construirons le zonier à partir des variables pertinentes sélectionnées.

IV.4.1 Variables explicatives retenues et ajustement

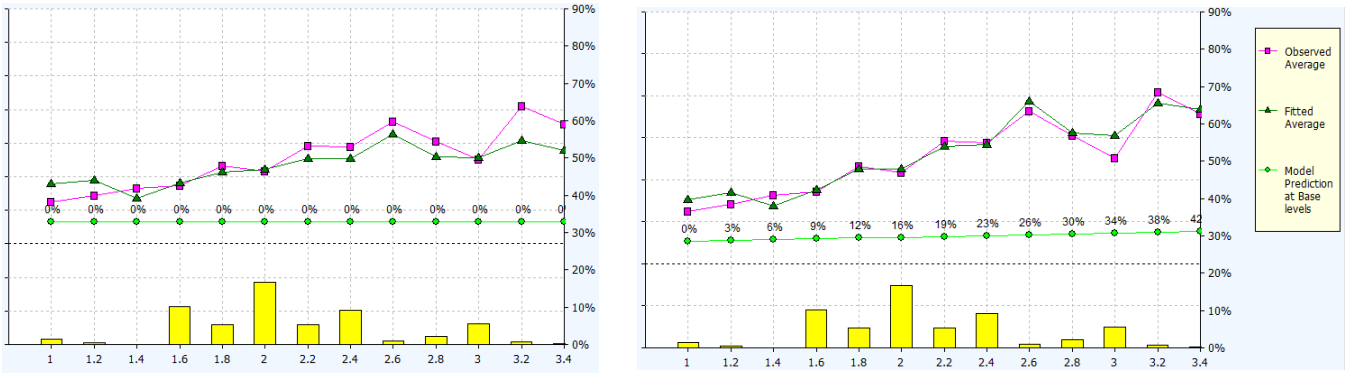
De la même manière que pour la construction du zonier sans les données privées, nous allons analyser l'ajustement du modèle de prime pure selon l'utilisation ou non de différentes variables. Pour ce modèle, nous allons observer seulement des variables construites à partir de la base de données grêle. Le modèle de base est celui présenté précédemment, sans la variable résidus lissés. Nous analyserons si les variables de la base de données apportent une information supplémentaire. Nous avons représenté en Figure 80 et en Figure 81 l'ajustement des modèles selon plusieurs cas d'utilisation des variables issues de la base de données privée :



(a) Non utilisée dans le modèle

(b) Utilisée dans le modèle

FIGURE 80 – Ajustement de la variable nombre de tempêtes de grêle totales



(a) Non utilisée dans le modèle

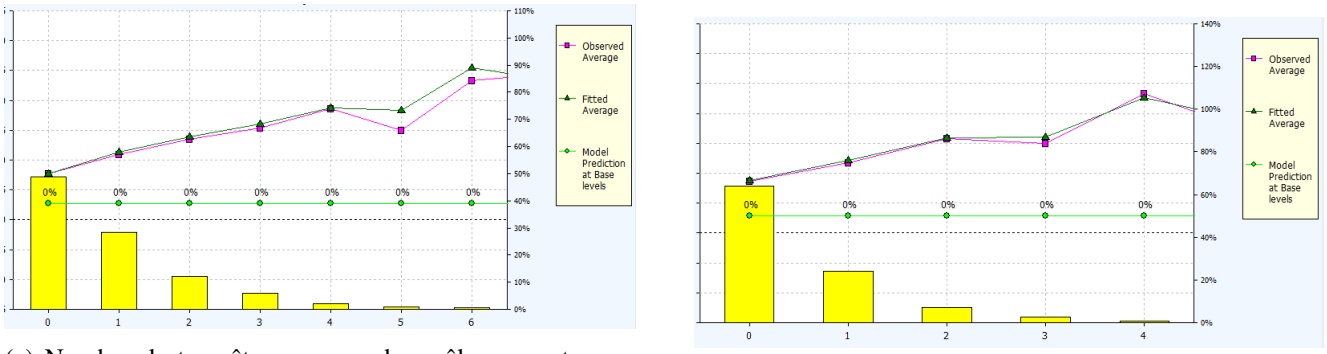
(b) Utilisée dans le modèle

FIGURE 81 – Ajustement de la variable diamètre maximal moyen du grêlon

Les deux variables apportent une information supplémentaire, même si celle-ci semble minime. Sur les deux graphiques en Figures 80a et 81b, nous remarquons une légère surestimation du taux de prime pure pour les communes où le nombre de tempêtes recensées est faible et pour celles dont les grêlons observés sont de petites tailles. A l'inverse, une sous-estimation pour les communes souvent sinistrées par la grêle ou bien celles sinistrées par de gros grêlons.

La base de données grêle nous permet donc de mieux discriminer chaque commune selon le risque grêle qu'elle porte. Toutefois, cette meilleure segmentation a ses limites. Premièrement, 40% des capitaux assurés n'ont pas connu de tempêtes de grêles. Cette modalité n'est pas représentée en Figure 81, mais elle est consécutive de la modalité 0 en Figure 80. Ces communes non sinistrées implique un manque d'information car nous ne connaissons pas la taille maximale des grêlons moyens sur celles-ci.

De plus, concernant la variable nombre de tempêtes, un peu plus de 80% du portefeuille est concentré sur les sept premières modalités, alors que le multiplicateur entre la première modalité, celle peu risquée, et la septième n'est que de 10%. L'utilité de ces deux variables existe bien, mais comme nous pouvons le voir elle demeure faible.



(a) Nombre de tempêtes avec un des grêlons ayant une taille de plus de 2cm

(b) Grêlons de type A2

FIGURE 82 – Variables ajustées mais non utilisées

Le graphique en Figure 82b représente l'ajustement du modèle par rapport au nombre de tempêtes de type A2 [17]. En Figure 82a, l'ajustement est par rapport au seuil de deux centimètres de taille des grêlons. Les graphiques illustrent l'ajustement lorsque ces variables ne sont pas utilisées dans l'implémentation du modèle. Nous remarquons que les variations de cette variable sont déjà prises en compte par le modèle.

IV.4.2 Utilisation des nouvelles variables dans le modèle

Nous venons de sélectionner les variables qui seront utilisées dans le modèle. Nous implémentons alors un modèle linéaire généralisé avec ces dernières afin d'obtenir un coefficient géographique pour chaque commune. Le zonier final construit avec les données externes sera obtenu après l'utilisation du package HClustGeo pour créer les zones tarifaires. Nous reprenons les paramètres précédents lors de l'implémentation, c'est-à-dire 0.7 pour α et 59 pour k.

Nous choisissons de ne pas travailler les résidus dans ce cas. Nous souhaitons seulement connaître l'apport de la base de données privées et le modèle utilisant uniquement les données complémentaires est un meilleur étalon.

Pour conclure cette partie, nous avons construit cinq modèles différents. Certains modèles tentent de réaliser un ajustement très précis du taux de prime pure observé, au détriment d'un potentiel surapprentissage. Certains modèles vont chercher à maîtriser cette volatilité, au prix d'un ajustement plus biaisé. Nous allons dans la suite de ce mémoire chercher à comparer ces modèles, analyser quels sont ceux qui s'adaptent le mieux aux différents scénarios, quels sont ceux qui segmentent le plus les communes selon leur risque grêle, quels sont ceux qui ont la plus grande cohérence de mutualisation.

17. A2 : taille de grêlon à partir de laquelle les cultures sont sinistrées par la grêle

Chapitre V Comparaison des zoniers et axes d'amélioration

V.1 Comparaison des différents zoniers

Précédemment, nous avons construit cinq zoniers différents :

— Un premier à partir de l'équation tarifaire en vigueur à Pacifica.

— Un deuxième reclassant les zones du premier à partir de l'historique des sinistres.

— Un troisième à partir des données complémentaires gratuites disponibles.

— Un quatrième reprenant le troisième zonier et qui utilise l'information des résidus.

— Un cinquième ayant pour socle le troisième et qui utilise des données grêles privées.

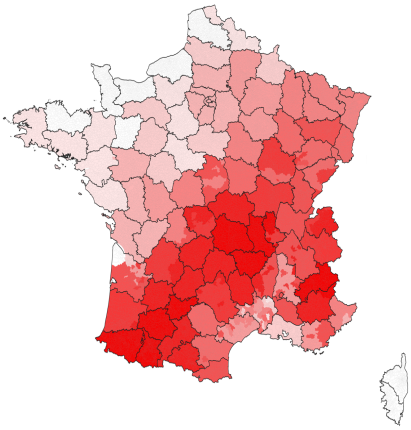
Certains zoniers maîtrisent le surapprentissage, au détriment d'un biais important. Celui utilisant les résidus prend en compte une grande partie de l'information disponible pour minimiser le biais, au risque de surapprendre. Ces différentes méthodes, nous allons les comparer. Nous essaierons de classer les zoniers en fonction de nos objectifs et contraintes. Ainsi, nous effectuerons plusieurs comparaisons :

1. Une comparaison du découpage géographique, à travers une analyse cartographique des différents zoniers.
2. Une comparaison de la segmentation. Nous analyserons la courbe des coefficients calculés à partir des GLM, ainsi que la variance inter-classes.
3. Une comparaison de la volatilité, à travers les résidus par classe et la variance intra-classe des taux de prime pure ajustés et des résidus.
4. Une comparaison du surapprentissage, à travers la validation croisée.

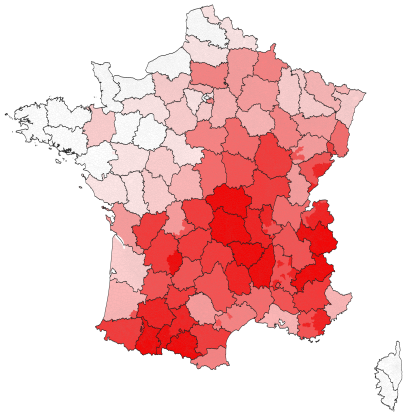
Nous souhaitons, à la fois, un zonier capable de s'adapter aux nouveaux scénarios de sinistralité, et un zonier capable de discriminer au mieux les zones entre elles.

V.1.1 Comparaison des découpages

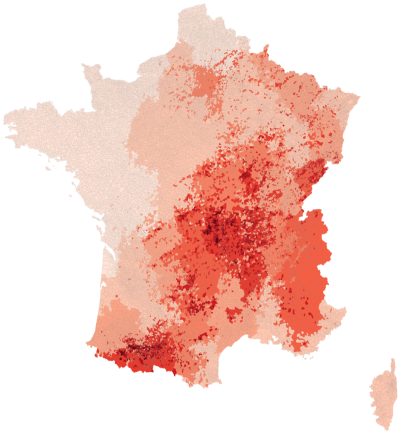
Le découpage du zonier est un des enjeux principaux. Le zonier doit être caractérisé par un maillage suffisamment fin, et doit expliquer la différence tarifaire entre deux proches voisins. De plus, le zonier doit être en mesure de tarifer l'ensemble du territoire.



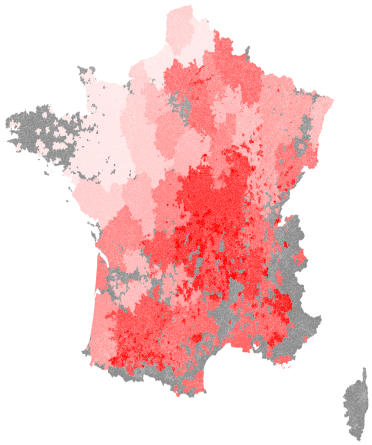
(a) Zonier actuel utilisé par Pacifica



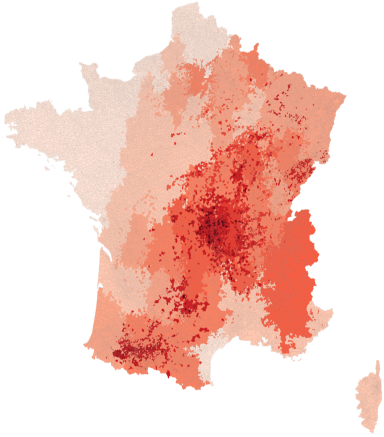
(b) Zonier actuel optimisé



(c) Zonier utilisant les données complémentaires disponibles



(d) Zonier utilisant les résidus



(e) Zonier utilisant les données grêles privées

FIGURE 83 – Représentation cartographique de chaque zonier

Les cartes de la Figure 83 illustrent chacun des découpages : plus l'intensité du rouge est forte,

plus la zone est exposée au risque grêle. La tendance des cinq zoniers est similaire à celle remarquée dans la littérature. Nous retrouvons la diagonale de grêle du Nord-Est au Sud-Ouest. Elle est surtout marquée par le Sud-Ouest et le Centre qui sont les deux zones les plus grêligènes du territoire.

Les deux zoniers construits à partir de l'équation tarifaire de Pacifica, visibles en Figures 83a et 83b, sont majoritairement à la maille départementale. Quelques communes d'exception sont également représentées. Les trois zoniers que nous avons intégralement construits, visibles en Figures 83c, 83e et 83d, sont à la maille communale. Nous remarquons que les zoniers établis à partir des données complémentaires ont une segmentation du risque plus précise géographiquement.

Concernant les zoniers construits à partir du découpage de Pacifica, des disparités existent au sein des départements, mais cette disparité n'est pas suffisamment prise en compte par les communes d'exception. Le découpage départemental a l'avantage d'être plus simple à commercialiser, mais son inconvénient est que les frontières départementales ne sont pas celles du risque grêle. Nous préférerons donc une segmentation au maillage plus fin, où des communes voisines au risque similaire formeront une zone. Les frontières des zones ne sont pas fixes, contrairement aux départements, et s'adaptent aux données et à nos besoins.

Nous avons l'objectif d'associer un coefficient tarifaire à chaque commune. Toutefois, sur la Figure 83d, nous remarquons que le zonier n'est pas appliqué à l'ensemble du territoire. Les communes qui ne sont pas en portefeuille et qui n'ont pas de voisins dans celui-ci n'ont pas de *résidus lissés*. L'utilisation de ces *résidus lissés* génère une problématique : que faire des communes non présentes en portefeuille ?

Nous avons plusieurs options : devons-nous attribuer la modalité mode des *Résidus Lissés* sur le département ? Devons-nous augmenter la taille du lissage pour propager le résidus de voisins à plus de communes ? Ces communes non présentes en portefeuille représentent-elles un enjeu majeur ? Chacune des réponses évoquées génèrent un problème d'approximation. Finalement, nous faisons le choix de retirer les communes non présentes en portefeuille et n'ayant aucun voisin limitrophes présents en portefeuille. La variable *résidus lissés* est très précise géographiquement et est très discriminante d'un point de vu tarifaire, nous ne souhaitons donc pas diffuser son information sur un rayon trop important. L'analyse des découpages indique que le zonier utilisant les résidus n'est pas optimal car incomplet.

Les deux zoniers qui utilisent seulement les données complémentaires offrent une segmentations très précises et un coefficient tarifaire pour l'ensemble des communes du territoire. Ces deux zoniers semblent s'adapter correctement à nos contraintes. Le zonier utilisant les données grêles privées, illustré en Figure 83e, semble légèrement moins sensible aux couloirs de tempêtes

constatés sur l'historique que nous retrouvons légèrement en Figure 83c. Au final, si nous choisissons le zonier optimal, selon la caractéristique du découpage géographique, nous retenons celui utilisant les données grêles privées en plus des données complémentaires disponibles.

V.1.2 Comparaison de la segmentation du risque

Le zonier, pour répondre à l'antisélection, doit être capable de segmenter suffisamment le risque entre les communes du territoire. Nous allons analyser, dans cette partie, la capacité de chacun des zoniers à différencier les communes les unes des autres.

V.1.2.1 Pentes de risque

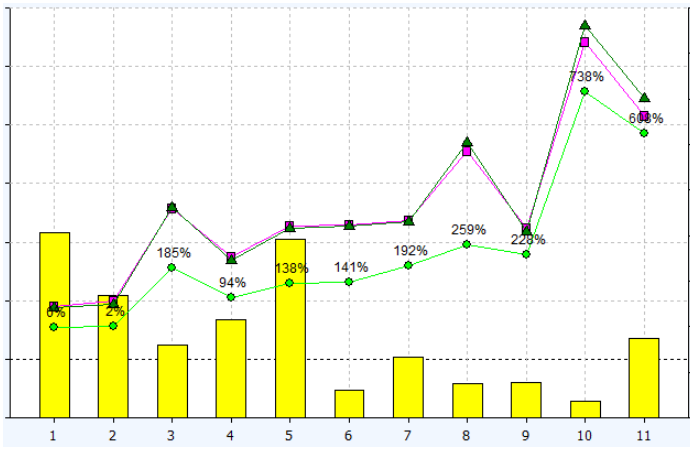
La capacité de segmentation d'un zonier est mesurée par l'amplitude des coefficients constituant la variable *zonier*. Cette amplitude symbolise la capacité d'un zonier à segmenter la commune la moins risquée du portefeuille par rapport à celle la plus risquée. Graphiquement, l'amplitude est visible à travers la pente de coefficients représentée en vert clair sur les graphiques. Sur les graphiques en Figure 84, nous pouvons voir lequel des zoniers construits offre la plus grande pente de coefficients. Pour juger de cette amplitude, nous regarderons le multiplicateur nécessaire pour passer de la première classe, celle moins risquée, à la dernière classe, la plus risquée. Ce multiplicateur est calculé comme suit :

$$M = \frac{coef_{sup} + 1}{coef_{inf} + 1} \tag{31}$$

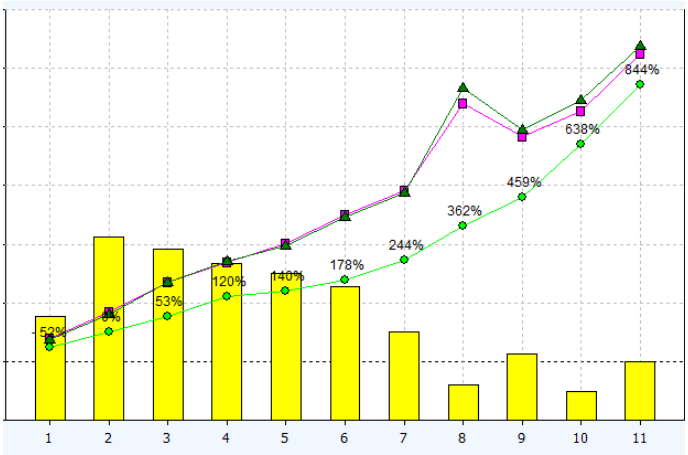
où M est le multiplicateur, $coef_{sup}$ le coefficient le plus grand du zonier et $coef_{inf}$ le coefficient le plus faible du zonier.

Nous souhaitons également que les classes de risque créées soient cohérentes entre elles. Une classe jugée plus risquée qu'une autre doit avoir un coefficient tarifaire significativement supérieur. Cette caractéristique est matérialisée par une stricte monotonie de la courbe verte claire sur les graphiques et par une certaine régularité de l'augmentation des coefficients tarifaires entre classes.

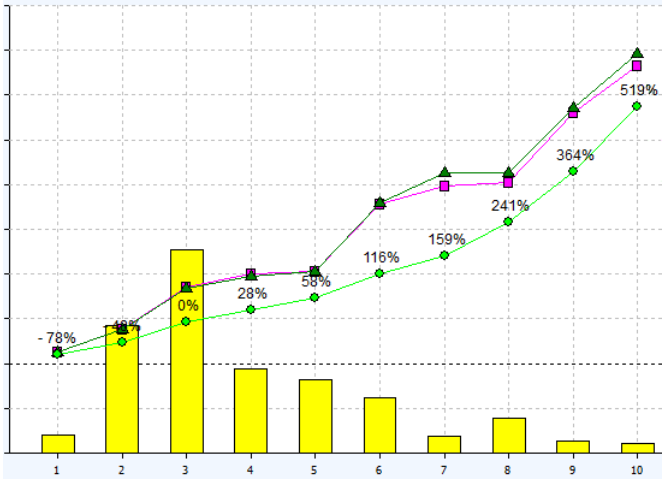
Les graphiques en Figure 84, font apparaître les diagrammes bâtons d'exposition pour chaque classe en terme de capitaux assurés. Nous regardons également si tous les zoniers offrent une répartition suffisante des capitaux assurés afin de limiter la volatilité pour certaines classes.



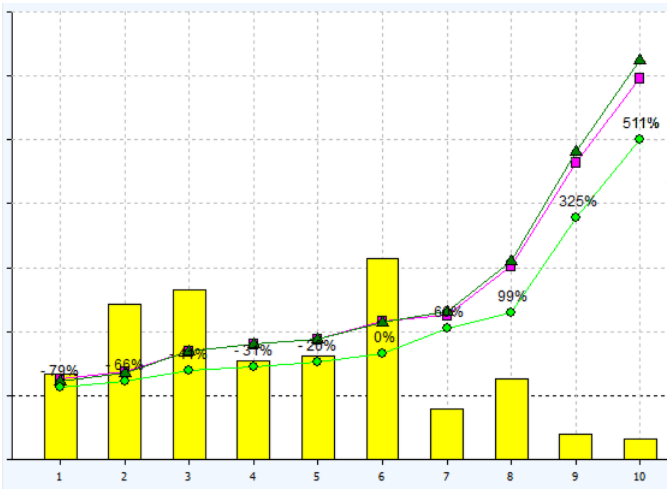
(a) Zonier actuel utilisé par Pacifica



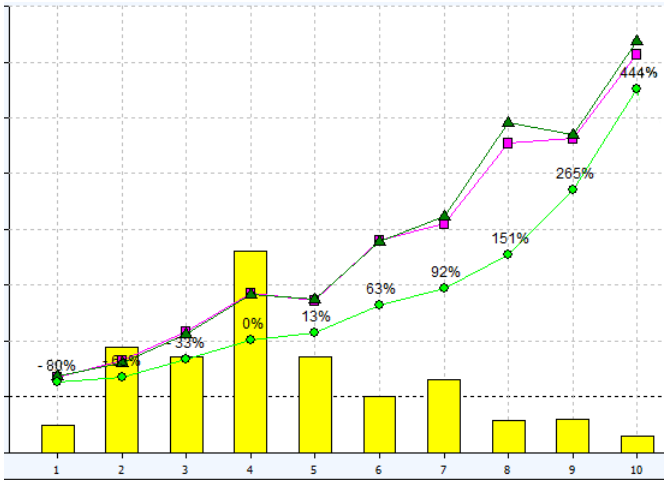
(b) Zonier actuel optimisé



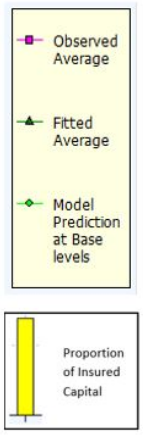
(c) Zonier utilisant les données complémentaires disponibles



(d) Zonier utilisant les résidus



(e) Zonier utilisant les données grêles privées



(f) Légende

FIGURE 84 – Ajustement du taux de prime pure et courbe des coefficients tarifaires de chaque zonier

Sur la Figure 84a est représentée la pente tarifaire du zonier actuellement utilisé par Pacifica. Comme nous l'avons déjà vu, cette courbe n'est pas strictement monotone. La différence des coefficients entre les classes 3 et 7 est faible et les classes 8 et 9 se confondent. Nous remarquons également que le multiplicateur tarifaire entre la classe la moins risquée et celle la plus risquée est seulement de 8,4, alors qu'il est de 19,6 dans le cas du zonier actuel optimisé comme nous le voyons à travers le Tableau 6 . La segmentation du modèle actuellement utilisé par Pacifica a nettement été améliorée par le reclassement.

Type de zonier	**Coefficient inf**	**Coefficient sup**	**Multiplicateur**
Actuel utilisé par Pacifica	0	738%	8,4
Actuel optimisé	-52%	844%	19,6
Utilisant les données complémentaires disponibles	-78%	519%	28,1
Utilisant les résidus	-79%	511%	29,1
Utilisant les données grêles privées	-80%	444%	27,2

TABLE 6 – Statistiques sur l'amplitude de la pente de risque par zonier

Les pentes tarifaires des zoniers qui utilisent les données complémentaires et du zonier actuel optimisé, visibles respectivement en Figures 84c, 84e et 84b, ont des pentes tarifaires strictement croissantes et régulières. Chaque classe de risque est bien différenciée de sa classe voisine pour ces zoniers.

Le zonier optimisé par les reclassements, dont le multiplicateur est de 19,6, segmente moins bien que les deux autres, dont les multiplicateurs sont de l'ordre de 27. Nous ne retenons donc pas le zonier optimisé à partir d'un reclassement sur le zonier actuel car il segmente moins bien les individus.

Les trois zoniers qui utilisent des données complémentaires ont un multiplicateur plus important que les deux premiers cités dans le Tableau 6. A noter que le zonier faisant un reclassement des zones à partir du zonier actuel discrimine un plus grand nombre de capitaux assurés car les classes 9, 10 et 11 sont fortement représentées. Le multiplicateur est le plus grand pour le zonier utilisant les résidus. Ce zonier est donc celui que nous retenons, si nous devions décider seulement en terme de segmentation des communes.

Par ailleurs le modèle utilisant les données complémentaires disponibles segmente mieux le risque que celui utilisant en plus les données grêles privées. Le multiplicateur est légèrement plus important pour le premier zonier et nous remarquons une meilleure segmentation pour les autres classes de risque.

Nous observons notamment sur la Figure 84c que certaines classes de risque sont faiblement représentées en terme de capitaux assuré, pouvant engendrer une certaine volatilité. De plus, cette représentation des capitaux assurés est plus égalitaire pour les zoniers utilisant les données grêles privées et les résidus, visibles respectivement en Figure 84e et Figure 84d. Cette meilleure répartition indique que ces zoniers vont segmenter un plus grand nombre d'observations car les classes 9 et 10, très différenciantes, comprennent un plus grand nombre de capitaux assurés. Ceci vient nous confirmer que le zonier utilisant les résidus est celui qui segmente le mieux les communes.

V.1.2.2 Distance inter-classes

Plus la segmentation est importante, plus les taux de prime pure moyen des classes sont différents les uns des autres. Une dispersion importante entre les classes va mettre en évidence l'hétérogénéité entre les classes de risque créées pour chaque zonier. Un indicateur mesurant cette dispersion va permettre de prendre en compte à la fois la différence des coefficients entre les classes ainsi que le volume de capitaux de celles-ci.

a) Définition

Pour mesurer la distance inter-classes du zonier z, nous utilisons D_z. Ce dernier est calculé comme suit :

$$D_z = \frac{\frac{1}{K}\sum_{i=1}^{N_z} K_{zi} * |Coef_{zi} - \frac{1}{K}\sum_{j=1}^{N_z} K_{zj} * Coef_{zj}|}{\frac{1}{K}\sum_{i=1}^{N_z} K_{zi} * Coef_{zi}} \tag{32}$$

Où N_z correspond au nombre de classes dans le zonier z. $Coef_{zi}$ correspond au coefficient calculé par le modèle pour la classe i du zonier z. K représente le volume de capitaux assurés total et K_{zi} le volume de capitaux assurés de la classe i du zonier z.

Cette formule correspond à la moyenne des écarts, en valeur absolue, des coefficients par classe par rapport à la moyenne. Nous effectuons une pondération par les capitaux de chaque classe afin de prendre en compte les inégalités en terme de capitaux assurés de celles-ci. De plus, nous réduisons D_z par la moyenne pondérée des coefficients calculés car chaque zonier n'a pas la même classe de référence. Les distances ne seraient pas comparables sans cette réduction.

Plus D_z sera grand, plus la segmentation du zonier sera importante.

b) Comparaison de D_z

Le tableau 7 synthétise les écarts-types calculés pour les cinq zoniers construits :

Type de zonier z	D_z
Actuel utilisé par Pacifica	48,6%
Actuel optimisé	52,4%
Utilisant les données complémentaires disponibles	52,1%
Utilisant les résidus	56,4%
Utilisant les données grêles privées	61,9%

TABLE 7 – Dispersion des coefficients par zonier

La distance la plus grande est obtenue à partir du zonier utilisant les données complémentaires disponibles ainsi que la base données grêle privée. Nous remarquons également que le faible volume de capitaux assurés présents dans les classes 8, 9 et 10 du zonier utilisant seulement les données complémentaires conduit à une faible segmentation du risque. En effet, la segmentation pour ce zonier est du même niveau que zonier optimisé via un reclassement.

V.1.3 Comparaison de la volatilité au sein des classes

V.1.3.1 Présentation des indicateurs

Nous cherchons à maximiser la segmentation du risque tout en maîtrisant le bruit capté par nos modèle. Le bruit apparaît quand le modèle effectue un trop grand ajustement de l'échantillon d'apprentissage. Il sera alors incapable de s'ajuster convenablement à un scénario de sinistralité futur dont les caractéristiques sont inconnues et différentes de celles constatées en historique.

Certaines zones peuvent avoir leurs résultats portés en grande partie par certaines années exceptionnelles. Ces années exceptionnelles correspondent parfois à de l'aléa et nous ne cherchons pas à nous ajuster très précisément sur celles-ci. Un tel zonier, qui cherchera à être ajusté précisément sur ces observations exceptionnelles, sera sujet à une forte variance. Pour mesurer cette volatilité nous utilisons deux indicateurs :

— Un premier mesurant l'erreur de prédiction au sein de chaque classe de risque.

— Un second mesurant l'homogénéité au sein des classes.

a) La valeur absolue moyenne des résidus par classe

Cet indicateur, nommé I_{zc}, est calculé comme suit pour chaque classe c d'un zonier z :

$$I_{zc} = \frac{\frac{1}{K_{zi}} \sum_{j=1}^{N_{zi}} K_{zij} * |TxPPobs_{zij} - TxPPpred_{zij}|}{\frac{1}{K_{zi}} \sum_{j=1}^{N_{zi}} K_{zij} * TxPPobs_{zij}} \tag{33}$$

Où N_{zi} correspond au nombre d'observations dans la classe i du zonier z. Res_{zij} correspond au résidus observé sur l'observation j de la classe i du zonier z. $TxPPobs_{zij}$ et $TxPPpred_{zij}$ correspondent au taux de prime pure observé et celui prédit sur l'observation j de la classe i du zonier z. Leur soustraction correspond aux résidus. K représente toujours le volume de capitaux assurés total et K_zij le volume de capitaux assurés de l'observation j de la classe i du zonier z.

Cette formule correspond à la division de la moyenne des valeurs absolues des résidus observés sur une classe par la moyenne des taux de prime pure observés sur cette même classe. Diviser par le taux de prime pure permet de rendre comparables les résidus car les classes de risque ont des niveaux de taux de prime très différents. La moyenne en valeur absolue des résidus permet de voir si le résidu moyen de la classe est élevé.

Nous souhaitons des classes composées de communes au risque grêle homogène. Donc si les résidus sont élevés en valeur absolue, nous pouvons considérer que les classes ne sont pas suffisamment homogènes. Nous allons analyser pour quels zoniers cet indicateur est le plus faible.

L'indicateur présenté peut également être agrégé et former un indicateur, I_z, unique pour chaque zonier :

$$I_z = \frac{1}{K} \sum_{i=1}^{N_z} K_{zi} * \frac{\frac{1}{K_{zi}} \sum_{j=1}^{N_{zi}} K_{zij} * |TxPPobs_{zij} - TxPPpred_{zij}|}{\frac{1}{K_{zi}} \sum_{j=1}^{N_{zi}} K_{zij} * TxPPobs_{zij}} \tag{34}$$

Où l'agrégation d'un zonier se fait à travers une pondération par le volume de capitaux assurés de chaque classe, K_{zi}. N_z correspond au nombre de classes de risque de z. Enfin, K représente le volume de capitaux assurés total en portefeuille.

b) Écart-type intra-classe

Nous allons regarder également l'homogénéité des observations regroupées au sein de chaque classe. Nous cherchons à obtenir des classes dont les primes pures observées pour chacune des

observations sont le plus proches possibles les unes des autres. Pour analyser cela, nous allons calculer, pour chaque classe, les écarts de prédiction par rapport à la moyenne du taux de prime observé sur la classe. La formule de cet indicateur, σ_{intra_z}, est calculée comme suit :

$$\sigma_{intra\_z} = \frac{1}{K}\sum_{i=1}^{N_z} K_{zi} * \frac{\sqrt{\frac{1}{(K_{zi})^2}\sum_{j=1}^{N_{zi}}(K_{zij} * TxPPobs_{zij} - \overline{TxPPobs_{zi}})^2}}{\overline{TxPPobs_{zi}}} \tag{35}$$

Où : $\overline{TxPPobs_{zi}} = \frac{1}{K_{zi}}\sum_{j=1}^{N_{zi}} K_{zij} * TxPPobs_{zij}$, c'est-à-dire la moyenne du taux de prime pure observé pondérée par les capitaux assurés de la classe i du zonier z.

Les classes de risque qui auront une valeur de σ_{intra_z} faible seront des classes homogènes. Une classe homogène regroupe des observations au taux de prime pure similaire. L'utilisation du taux de prime pure observé permet surtout de constater si les regroupements ou les lissages géographiques ne créent pas de zones de risque composées de communes aux risques hétérogènes.

V.1.3.2 Analyse de l'indicateur selon les zoniers

La Figure 85 représente la valeur de l'indicateur I_cz, c'est-à-dire la moyenne en valeur absolue des résidus pour chaque classe c du zonier z :

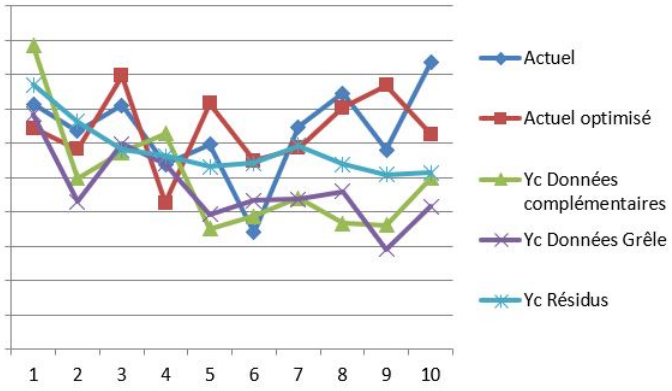


FIGURE 85 – Analyse de l'erreur de prédiction moyenne, réduite, par classe pour chaque zonier

Les courbes violette et verte, qui représentent respectivement le zonier utilisant les données grêles privées et celui utilisant les données complémentaires disponibles seulement, sont majoritairement en dessous des autres. C'est le cas surtout pour les classes 7, 8, 9 et 10 où l'erreur de prédiction moyenne est la plus faible pour ces deux zoniers que les autres. Ceci vient montrer la qualité d'ajustement de ces deux zoniers pour les classes fortement exposées au risque grêle.

Le Tableau 8 récapitule pour chaque zonier la valeur absolue moyenne des résidus par classe et l'écart-type intra-classe moyen :

Type de zonier z	I_z	σ_{intra_z}
Actuel utilisé par Pacifica	53,4%	39,8%
Actuel optimisé	52,9%	37,1%
Utilisant les données complémentaires disponibles	47,8%	31,4%
Utilisant les résidus	49,8%	31,5%
Utilisant les données grêles privées	45,9%	31,1%

TABLE 8 – Statistiques sur l'homogénéité des classes selon les zoniers

I_z et σ_{intra_z} sont les plus élevés pour les deux zoniers construits à partir du zonier actuel de Pacifica. Ceux-ci n'ont donc pas des classes suffisamment homogènes. Nous pouvons expliquer ce constat par le fait que ces zoniers sont contraints par un découpage à la maille départementale. Nous constatons également que le reclassement permet seulement une légère amélioration de l'homogénéité des écarts-types intra-classe.

Les trois zoniers construits à partir d'un découpage variable ont un écart-type intra-classe similaire. Ces zoniers ont donc une meilleure homogénéité au sein des classes. L'indicateur I_z est les plus faible pour le zonier utilisant les données complémentaires, disponibles et privées.

V.1.4 Comparaison de la stabilité des coefficients

Afin d'analyser la stabilité des coefficients, nous allons mesurer la capacité du zonier à rester cohérent pour de nouveaux scénarios de sinistralité. Pour ce faire, nous allons mettre en place un processus de validation croisée.

V.1.4.1 Description de la validation croisée

Nous allons séparer l'historique de données en sous-ensembles : un ensemble de test et un ensemble d'entraînement. La procédure est représentée en Figure 86. Ce procédé va nous permettre notamment de juger du surapprentissage.

FIGURE 86 – Diagramme validation croisée à k-blocs

La division de notre base de données totale, en un ensemble de test et un ensemble d'entraînement, fait face à une contrainte importante : les deux ensembles créées doivent avoir des structures similaires.

Premièrement, une séparation aléatoire du portefeuille revient à mélanger des scénarios de sinistralité, ce qui biaiserait le test. Nous ne pouvons également pas scinder notre base selon la variable géographique. Il n'est pas pertinent, dans notre cas, de s'entraîner sur une base composée d'une partie des communes du portefeuille et de tester à partir de communes différentes de l'entraînement. En effet, notre objectif principal est de connaître l'exposition au risque grêle du portefeuille par commune au cours du temps. Pour tester la stabilité, nous avons donc intérêt de scinder la base selon les scénarios de sinistralité, c'est-à-dire selon les exercices.

Deuxièmement, nous avons évoqué que la structure du portefeuille évoluait au cours du temps. Dans l'objectif d'avoir des exercices de test comparables à ceux d'entraînement, nous ne pouvons pas nous entraîner sur les premières années d'historique car ils ne représentent pas bien le portefeuille actuel.

Troisièmement, nous avons souligné la volatilité de la grêle et sa faible fréquence sur une large partie du territoire. L'ensemble de test doit donc être suffisant en terme d'exercices pour éviter une incidence trop importante de l'aléa sur les résultats. Face à ces contraintes, nous scindons finalement la base d'entraînement en deux parties égales en terme de nombre d'exercices (huit années) :

— Une base d'entraînement constituée des observations des années impaires de l'historique.

— Une base d'entraînement constituée des observations des années paires de l'historique.

Pour juger de la qualité des modèles en terme de sur-apprentissage, nous allons comparer les coefficients calculés sur la variable *zonier*, entre le modèle ajusté sur les données d'entraînement et le modèle ajusté sur celles de test.

Pour réaliser cette comparaison, plusieurs étapes, sont effectuées :

1. La variable *zonier Entraînement* est construite à partir de l'ensemble d'entraînement, selon les processus déjà définis pour chacun des zoniers. Le reclassement est donc réalisé à partir des données de sinistres des années impaires. Les zoniers utilisant la classification hiérarchique ascendante effectuent également le dessin des zones à partir des années impaires.

2. Deux modèles linéaires généralisés sont implémentés pour chaque zonier à analyser. Un zonier à partir de l'ensemble d'entraînement et un autre à partir de l'ensemble de test. Dans ces modèles, le taux de prime pure est ajusté par la variable *zonier entraînement*, construite dans l'étape précédente, ainsi que les autres variables explicatives non géographiques.

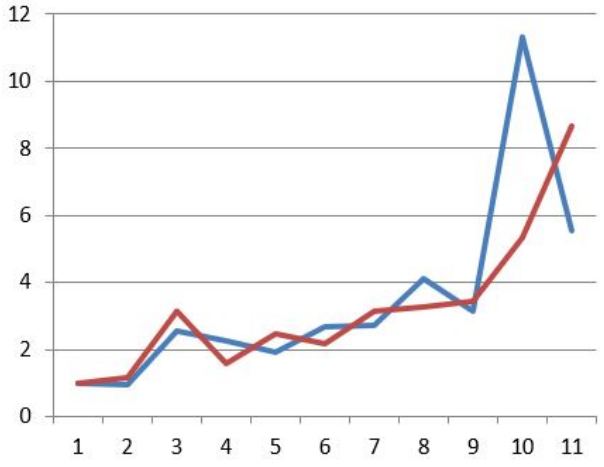
3. Les coefficients calculés pour la variable *zonier entraînement* sont comparés par classe de risques, entre le modèle d'entraînement et celui de test.

Un modèle stable correspond à un modèle dont ses coefficients par classe de risque sont similaires entre l'ensemble de test et celui d'entraînement. Lorsque ceux-ci auront des résultats éloignés, nous pourrons conclure que le modèle sera dans l'incapacité de s'adapter à de nouveaux scénarios de sinistralité, différents de ceux présents en historique. Le modèle dans ce cas surapprend de l'ensemble de données à disposition.

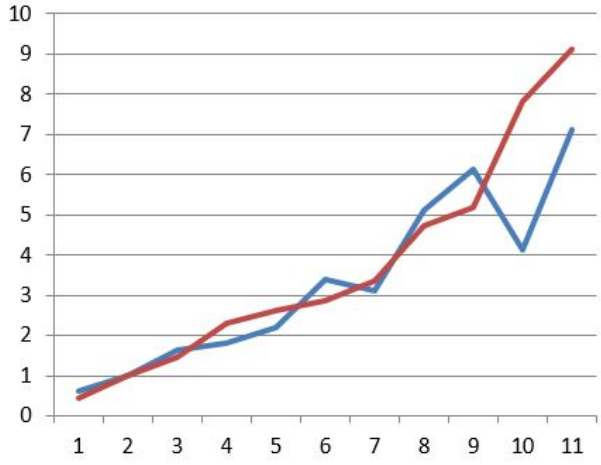
A noter que, la validation par le biais d'une base d'entraînement n'est pas possible pour le zonier actuel puisque celui-ci est déjà construit. Pour ce zonier, nous comparons seulement les coefficients calculés sur la base des années paires à ceux calculés sur la base de données impaires.

V.1.4.2 Analyse des résultats de l'ensemble de test

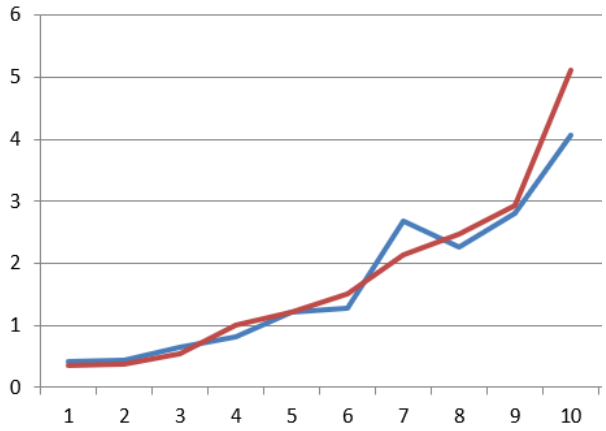
La Figure 87 fait apparaître plusieurs graphiques, un par zonier. Ils comparent, dans chaque cas, la pente des coefficients obtenus à partir d'un modèle ajusté sur les données d'entraînement et celle obtenue à partir d'un modèle ajusté sur les données de test :



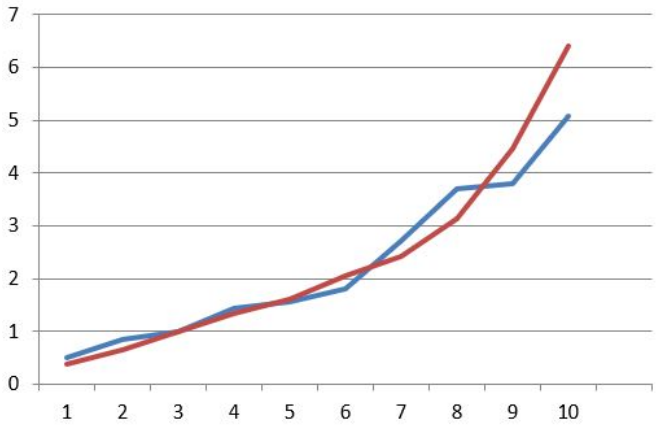
(a) Zonier actuel



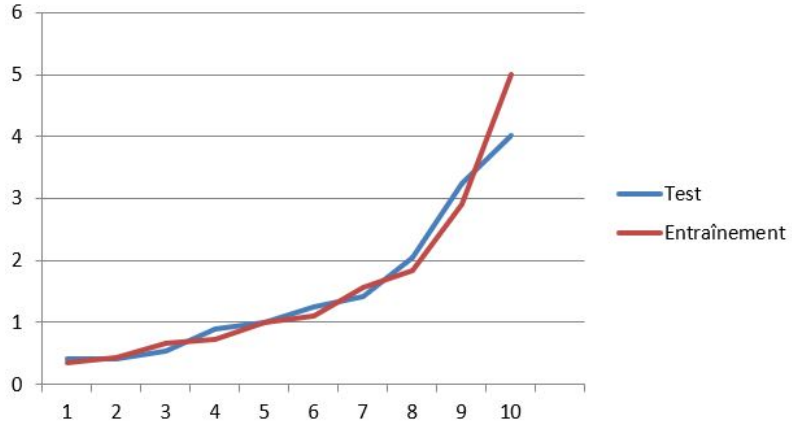
(b) Zonier actuel optimisé



(c) Zonier utilisant les données complémentaires disponibles



(d) Zonier utilisant les résidus



(e) Zonier utilisant les données grêles privées

FIGURE 87 – Représentation des pentes de coefficients pour l'ensemble d'entraînement (en rouge) et celui de test (en bleu). Les abscisses correspondent aux classes du zonier et les ordonnées à la valeur du coefficient de chaque classe.

Les deux zoniers actualisés ne sont pas suffisamment stables. Le zonier actuel fait apparaître une forte différence d'estimation des coefficients entre les années paires et celles impaires pour toutes les classes risquées. Cette différence est surtout visible pour les classes 10 et 11 en Figure 87a. Le zonier reclassé a du mal à s'adapter à de nouveaux scénarios de sinistralité. Nous constatons notamment que la pente des coefficients obtenus à partir de l'ensemble de test n'est pas strictement croissante. Le surapprentissage est important également pour les classes 10 et 11.

Le zonier utilisant les résidus semble surapprendre pour les classes faiblement sinistrées et fortement sinistrées dans l'ensemble d'entraînement. Nous observons en Figure 87d que les classes au risque faible sont légèrement sous-estimées dans l'entraînement, et inversement surestimées pour les classes au risque grêle important. Logiquement, le modèle utilisant les résidus capte très bien ces informations de classes extrêmes, mais ceci implique des difficultés pour ce dernier à s'adapter à des scénarios de sinistralité différents.

Enfin, concernant le zonier utilisant les données complémentaires disponibles et celui utilisant en plus les données privées, nous remarquons une meilleure stabilité que pour les autres zoniers. Les cinq premières classes de risques sont notamment très stables entre l'estimation sur la base d'entraînement et celle de test. Toutefois, la stabilité est plus importante pour le zonier utilisant les données grêles privées. Comme nous l'avions présumé dans les précédentes analyses, le zonier utilisant seulement les données complémentaires gratuites est exposé à la volatilité avec des classes de risque parfois faiblement représentées en capitaux assurés. C'est le cas des classes 7 et 10 qui sont faiblement représentées, visible en Figure 84c et où nous constatons une forte instabilité sur la Figure 87c.

Au final, le zonier grêle utilisant les données complémentaires disponibles et les données grêles privées est le meilleur zonier du point de vu de la stabilité. C'est également le zonier qui offre la meilleure homogénéité intra-classe. De plus, ce zonier offre un découpage adapté à nos besoins et une segmentation du risque entre les communes très satisfaisant.

Toutefois, cette base de données est coûteuse et le bénéfice est-il suffisant pour compenser ce coût ? Cette base de données nécessite potentiellement des coûts futurs, notamment si nous souhaitons obtenir les données récentes ainsi qu'observer les effets sur le long terme de la nouvelle technologie de détection. Nous remarquons qu'en terme de découpage et segmentation du risque, le zonier utilisant les données complémentaires offre des résultats similaires. La stabilité doit cependant être améliorée pour pouvoir se passer avec certitude de cette base de données. Nous allons voir, à travers les axes d'amélioration, qu'il est possible d'améliorer les résultats.

V.2 Axes d'amélioration

V.2.1 Réduction du bruit contenu dans la base d'apprentissage

a) Patience

Nous avons ainsi construit différents modèles. Ceux-ci pourront simplement être implémentés, avec une paramétrisation adaptée, à partir de l'historique futur. Cet historique sera plus conséquent que ce soit en terme de scénarios de sinistralité grêle qu'en terme de communes représentées.

L'occurrence des tempêtes de grêle est très faible pour une majorité des communes. La période de retour est donc pour ces communes de plusieurs années. La volatilité du risque grêle est également importante. Il est nécessaire d'avoir à disposition un historique de données conséquent, de plusieurs décennies, afin de maîtriser la volatilité et d'observer la période de retour selon les zones. Une telle base de données permettrait de mettre en application la loi des grands nombre, d'observer un taux de prime pure grêle empirique sur l'historique proche de celui théorique pour chaque commune.

L'historique des sinistres est également incomplet géographiquement. La sinistralité est observée seulement pour les communes présentes en portefeuilles. De plus, pour un nombre important de communes, le nombre d'années de présence en portefeuille est de quelques années seulement.

L'étude empirique de la grêle est récente sur les différentes bases que nous avons à disposition. Le début de ces dernières correspond à 2005. De plus, pour chacune des bases il y a une évolution importante de la structure conduisant à un certain biais. C'est pourquoi, plusieurs années d'étude supplémentaires des tempêtes de grêle dans ces bases amélioraient significativement le volume d'information à disposition sur le risque grêle.

Toutefois, des solutions existent pour contourner ce déficit d'historique. Pour réduire le bruit ou compléter l'information géographique, il est possible d'approfondir les méthodes statistiques évoquées dans ce mémoire ou d'en développer d'autres.

Il faudra également veiller à vérifier si le changement climatique ne vient pas modifier nos certitudes quant aux zones de risques sur le territoire français.

b) Utilisation plus approfondie des outils de lissage

Nous avons effectué des lissages sur les variables explicatives. Ceci nous a permis d'obtenir un ensemble de variables explicatives complet sur l'ensemble du territoire. Nous n'avons cependant pas réussi à implémenter un modèle de lissage des résidus satisfaisant. Les lissages effectués n'ar-

rivaient pas à gérer l'arbitrage entre un lissage suffisant pour réduire le bruit et un lissage limité pour conserver l'information contenue dans les résidus.

Une meilleure compréhension des modèles de lissage, tel que le krigeage, pourrait permettre d'utiliser l'information contenue dans les résidus. Nous avons pu voir que cette variable, *résidus lissés*, était prometteuse, mais que le retraitement était essentiel pour conserver un modèle stable. Une paramétrisation du krigeage adaptée pourrait permettre une réduction du bruit, tout en limitant la perte d'information engendrée par le lissage.

c) Techniques de régularisation

Il pourrait être intéressant d'implémenter un plus grand nombre de variables dans le modèle, quitte à ce que certaines variables aient un faible apport dans l'ajustement. Ces variables viendrait apporter des informations et permettrait à notre zonier de gagner en qualité.

Des méthodes pour maîtriser l'ajout important de variables dans les modèles linéaires généralisés existent. Les techniques de régularisation notamment sont des méthodes couramment utilisées en machine learning pour améliorer la performance des modèles en réduisant leur complexité et en évitant le surapprentissage. Différentes méthodes de régularisation existent, nous en recensons quelques unes parmi les plus courantes :

— *La régularisation L1* (ou régularisation de Lasso) : elle consiste à ajouter une pénalité à la fonction de coût du modèle. Cette pénalité est proportionnelle à la valeur absolue des poids des paramètres du modèle. Cette technique pousse les poids des paramètres vers zéro et conduit à des modèles moins complexes en terme de paramètres et variables.

— *La régularisation L2* (ou régularisation de Ridge) : comme la régularisation L1, elle pénalise la fonction de coût du modèle, mais cette fois la pénalité est proportionnelle au carré des poids des paramètres du modèle.

— *Elastic Ne*t : cette technique correspond à une combinaison de régularisation Ridge et Lasso. Il utilise à la fois la pénalité L1 et L2 pour réduire la complexité du modèle et encourager la sélection automatique des caractéristiques.

Les techniques de régularisation peuvent améliorer la performance des modèles grâce à la réduction de leur complexité et en évitant le surapprentissage.

La méthode la plus efficace pour lutter contre le bruit et le surapprentissage est de répéter, selon différents jeux de paramètres, les opérations de validation croisée. Nous l'effectuons déjà, mais seulement sur deux ensembles.

d) Autres bases de données grêle disponibles

Nous avons également la possibilité d'améliorer notre connaissance du risque grêle dès maintenant. Il est possible de se rapprocher de Météo France dans l'objectif d'obtenir une base de données. Toutefois, cette base sera plus coûteuse que celle de CatNat et le bénéfice d'information n'est pas certain. Il est également possible de mutualiser les historiques de sinistralité, avec d'autres acteurs du marché assurantiel, comme il est suggéré dans le projet de loi à travers le système de pool.

V.2.2 Modélisation des couloirs de grêle

Nous avons construit plusieurs modèles, et ceux-ci traitent l'entièreté du territoire. Toutefois, du fait des différences de sinistralité importantes, les enjeux et contraintes sont très différents d'une zone à l'autre. Certaines zones, moins risquées, ont une mutualisation importante. En effet, ces zones ont un risque grêle plutôt bien estimé et les tempêtes constatées correspondent à l'aléa. Les zones fortement exposées au risque grêle, comme la commune de Cheyenne présentée précédemment, doivent être captées par le zonier. Ces zones aux risques très différents pourraient être traiter très différemment.

Des couloirs grêligènes semblent exister sur le territoire. Ces couloirs sont des zones propices à d'importantes tempêtes de grêle. Ils sont difficiles à déceler. La survenance d'une ou deux tempêtes sur une zone peut être aussi bien interprétée comme un couloir ou bien un aléa. C'est pourquoi, il peut être intéressant de construire un modèle uniquement sur les zones risquées, communes dont le traitement du risque diffèrent des autres. Dès lors, l'analyse pourra être plus fine géographiquement et la paramétrisation des modèles sera plus adaptée à ce cas spécifique.

Les couloirs de grêle ont fait l'objet d'études comme par exemple celle de Mickael Gruber « modélisation du risque grêle en France »[8]. Ce dernier énonce un processus de simulation d'une année de trajectoire de grêle. Ce travail pourrait nous permettre d'isoler certaines communes, ou groupes de communes très grêligènes. Des modélisations de dizaines ou centaines d'exercices peuvent être effectuées avec ce type de modélisation, sous contrainte de données disponibles.

La modélisation précise de la grêle d'un point géographique, comme il est fait dans l'étude de Mickael Gruber, pourrait être un apport complémentaire des modèles construits dans ce mémoire. Toutefois, de tels modèles pourraient se heurter au problème d'insuffisance de l'historique de données disponible. A noter que, la modélisation de Mickael Gruber porte sur le risque grêle en Habitation. Ce risque est différent de celui agricole, la vulnérabilité des cultures est plus forte que celle des habitations. De plus, les données disponibles sur le produit Habitation sont plus conséquentes.

V.2.3 Meilleure paramétrisation des modèles

V.2.3.1 Tester un ensemble plus grand de paramètres

Tout au long de la construction des zoniers, nous avons essayé d'optimiser un nombre important de paramètres :

— Pour la classification hiérarchique ascendante, nous avons cherché à optimiser le nombre de zones créées et l'importance accordée à la distance géographique par rapport à celle des coefficients.

— Pour le lissage des résidus, il nous a fallu choisir la méthode, le nombre de voisins et la pondération des voisins.

— Pour la variable *zonier* décider du nombre de classes de risque à créer.

Lors du paramétrage du nombre de classes de risque, nous avons retenu 10 ou 11 classes dans ce mémoire, à cause de la contrainte de volatilité. En effet, les classes de risque fortement exposées à la grêle (7 à 11) sont faiblement représentées en terme de capitaux assurés. Il est alors difficile de les scinder sans perdre en stabilité. Il pourrait être intéressant de réaliser un processus afin d'optimiser cette variable *classes de risque* pour différentes paramétrisation des zoniers. Toutefois, cette tâche est redondante et coûteuse en temps de calcul. De plus, les autres paramètres sont à tester et ceux-ci influencent directement l'optimisation de cette variable.

Nous pouvons entrevoir l'automatisation d'une procédure par le biais du logiciel R. Ce programme mettrait en application les étapes suivantes :

1. Implémentation d'un premier GLM qui va ajuster le taux de prime pure à partir des variables explicatives retenues. Nous considérons que le meilleur modèle est choisi en amont (choix des variables et paramétrisation des variables). Cette étape va donc comprendre qu'une seule itération.
2. Lissage des résidus. Nous allons effectuer pour cette étape autant d'itérations qu'il y aura de combinaisons possibles des paramètres de lissages. Nous avons vu que ceux-ci sont nombreux, cette étape peut être coûteuse en temps de calcul.
3. Construction des zones de risques. Nous allons implémenter le package HClustGeo pour créer les zones géographiques ayant une exposition similaire au risque grêle. Pour cette étape, le nombre d'itérations correspond au produit du nombre de combinaisons possibles pour l'en-

semble de paramètres k [18], α [19] et du nombre d'itérations de l'étape précédente.

4. Construction de la variable *zonier*. Nous allons implémenter un GLM qui va ajuster le taux de prime à partir des variables explicatives non géographiques retenues ainsi qu'à partir de la variable *Zonier*. Le nombre d'itérations que nous souhaitons tester correspond au produit du nombre de classes de risque et du nombre d'itérations réalisées à l'étape précédente.

5. Comparaison des zoniers. Il faudra enfin calculer pour chaque zonier les différents indicateurs de segmentation du risque, de stabilité et d'homogénéité des classes pour juger lesquels offrent les meilleurs performances. Ces résultats devront être centraliser dans une même base pour faciliter l'analyse. Un indicateur sera plus compliqué à obtenir, celui de la stabilité. En effet, la validation croisée vient alourdir le processus et augmenter une nouvelle fois le temps de calcul.

Au final, ce processus permet de balayer l'ensemble des paramètres qui nous intéressent et offre un large éventail de modèles différents à comparer. Toutefois, l'implémentation du package HClustGeo est très lourde à cause de la taille des matrices des distances. Le temps de lissage peut également être important compte tenu du nombre de communes à lisser. Ces modèles nécessitent donc un ordinateur avec une mémoire disponible conséquente. Le temps de calcul se comptera en journées pour un ensemble de paramètres conséquents car comme nous l'avons décrit le nombre d'itérations augmente considérablement avec l'ajout de paramètres à tester.

V.2.3.2 Modifier les étapes de construction

Nous avons constaté lors des comparaisons que le zonier utilisant les résidus offrait une segmentation du risque grêle similaire, voire légèrement inférieure, aux deux autres zoniers construits avec des données complémentaires. Nous pouvions nous attendre, logiquement, à ce que le zonier utilisant les résidus offre une segmentation très nettement améliorée au prix d'un important surapprentissage. Nous n'observons pas cela car les étapes successives de lissages et de classifications viennent diminuer l'information transmise par les résidus. Nous avons fait ce choix car nous voulions éviter d'apprendre à partir d'une information bruitée.

Il serait intéressant de voir si en changeant notre processus de construction, comme représenté en Figure 88, la stabilité demeure maîtrisée.

18. Nombre de zones
19. Importance accordée à la distance géographique

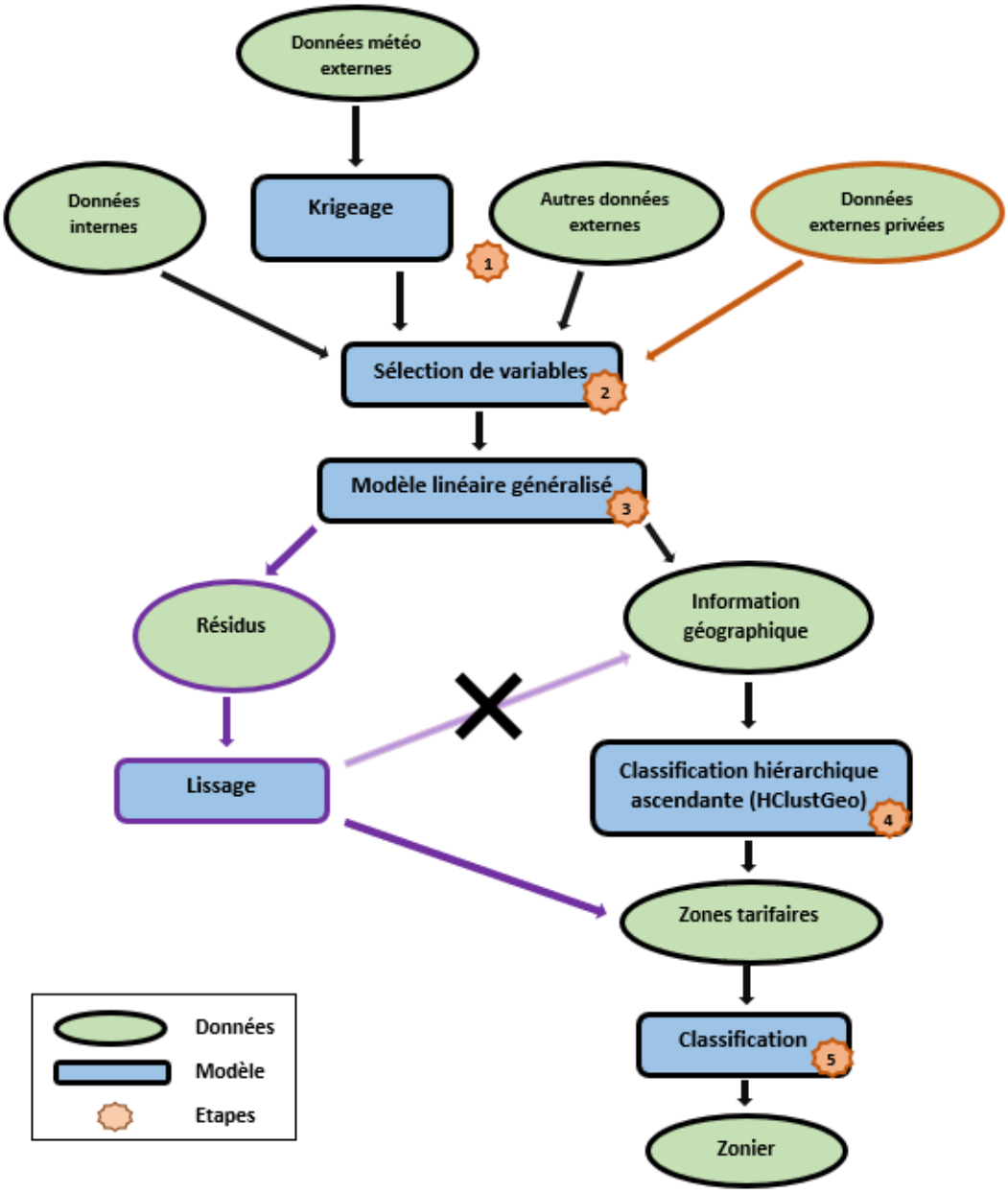


FIGURE 88 – Étapes de construction des zoniers modifiée

À noter qu'une segmentation du risque nettement améliorée réduirait la mutualisation. Cette mutualisation était permise par la classification hiérarchique, en modifiant l'ordre des étapes nos réduisons son impact.

De plus, l'ordre de grandeur des coefficients calculés à partir de l'ajustement sur la variable *Résidus lissés* peut être très important. Leur ajout direct dans le calcul de la variable *Zonier* viendrait grandement déformer la classification effectuée à l'étape précédente.

V.2.4 Utilisation d'algorithmes de machine learning plus sophistiqués

V.2.4.1 Dans un cadre exploratoire

Plusieurs variables explicatives sont utilisées pour expliquer le niveau d'exposition au risque grêle d'une commune. Pour construire cet ensemble, nous avons essayé de nous appuyer sur les références scientifiques portant sur le sujet. Nous nous sommes aperçus que cet ensemble était incomplet et que d'autres variables, inconnues, pourraient venir expliquer la sinistralité grêle de chacune des communes.

Il peut être pertinent, dans un cadre exploratoire, d'utiliser les arbres de décision. Nous pourrons observer quelles variables contribuent le plus dans l'ajustement du modèle. Dans le cadre d'un modèle linéaire généralisé, il peut être compliqué d'implémenter un modèle avec des centaines de variables, surtout si celles-ci sont quantitatives. L'utilisation de l'importance à travers l'implantation d'arbres de décisions nécessiterait un important temps de calcul, mais pourra permettre de juger de la pertinence d'un ensemble conséquent de variables explicatives.

V.2.4.2 Dans un objectif d'améliorer les ajustements

La régression linéaire est un modèle de machine learning simple qui peut être utilisé pour prédire les sinistres d'un contrat en fonction d'un ensemble de caractéristiques des communes, des cultures et du contrat en lui-même. La régression linéaire est facile à interpréter et permet d'identifier les effets des variables explicatives sur la variable ajustée. Cependant, la régression linéaire peut être limitée dans sa capacité à modéliser les relations non linéaires entre les variables. Nous allons regarder si d'autres modèles peuvent pallier à cette limite et améliorer nos performances d'ajustement, sans pour autant engendrer une importante régression sur les bénéficies de ce type de modèles.

Il est important de noter que chaque modèle a ses avantages et inconvénients, et qu'il est important de choisir un modèle qui s'adapte à notre ensemble de données spécifique et à notre problématique.

Réseaux de neurones

Les réseaux de neurones sont des modèles de machine learning complexes qui peuvent détecter des relations non linéaires entre les variables d'entrée et de sortie. Cela pourrait être particulièrement utile si la relation entre les caractéristiques des communes et l'exposition au risque grêle est com-

plexe et difficile à modéliser. Cependant, les réseaux de neurones peuvent être très sensibles aux données bruyantes ou insuffisantes, ce qui peut être un problème avec un historique limité de données sur les événements grêle.

Arbres de décision

Les arbres de décision sont des modèles simples et interprétables pouvant être utilisés pour classer ou prédire des événements en fonction d'un ensemble de variables d'entrée. Cependant, les arbres de décision peuvent être sensibles aux données bruyantes ou manquantes. Ils peuvent ainsi avoir tendance à sur-ajuster les données d'entraînement.

Machines à vecteurs de support (SVM)

Les machines à vecteurs de support sont un ensemble de techniques d'apprentissage supervisé utilisées pour la classification ou la régression. Ils peuvent être utilisés pour séparer les communes à faible risque en utilisant des données historiques. Ils ont les avantages d'être efficaces avec les données de petite taille et peuvent gérer les données bruyantes. En revanche, ils peuvent être difficile à interpréter.

Au final, l'utilisation des réseaux de neurones n'est pas adaptée dans notre cas car nous avons des données bruitées et en faible volume. De plus, les réseaux de neurones sont complexes, alors que notre modèle comprend peu de variables explicatives. Les arbres de décisions ne semblent également pas s'adapter à notre ensemble de données bruité. Enfin, les SVM pourraient être une solution, mais leur utilisation nécessitera un travail important pour comprendre les effets de chaque variable sur le taux de prime pure.

Conclusion

L'objectif de ce mémoire était de créer un zonier destiné au produit Grêle. Dans le cadre de la construction de celui-ci, nous avons tout d'abord dû faire face à la volatilité des tempêtes de grêle, pouvant générer une information bruitée. Nous avons également été confrontés au fait que l'historique de sinistralité sur le produit Grêle est limité, restreignant ainsi le volume d'information disponible sur l'incidence des tempêtes de grêle sur le territoire français sur les récoltes assurées. Ce volume d'information limité et bruité a nécessité le recours à d'autres bases de données et des outils mathématiques afin d'en améliorer l'ajustement.

Nous avons utilisé des données complémentaires internes à travers les produits Habitation et Multirisques Climatiques sur cultures. Nous avons également eu recours à des sources de données externes disponibles telles qu'Infloclimat, l'Insee ou CORINE Land Cover, ainsi que des données externes privées qui nous ont transmis des informations sur les tempêtes de grêle. Le krigeage nous a permis de diffuser l'information des variables météorologiques, alors limitée à certaines communes, à l'ensemble du territoire. Les modèles linéaires généralisés ont contribué à quantifier les relations entre le taux de prime pure observé et les variables explicatives. Enfin, la classification ascendante hiérarchique a concouru à limiter le bruit et à mutualiser le risque, en regroupant des communes voisines ayant le même niveau de risque.

Nous avons pu construire, dans ce mémoire, différents zoniers qui permettent d'améliorer celui actuellement utilisé par Pacifica. Parmi les zoniers construits, nous avons essayé notamment de travailler l'information contenue dans les résidus. Ce travail consiste en un lissage géographique afin de retirer le bruit contenu dans les résidus. Pour ce zonier, nous n'avons pas obtenu de gains significatifs par rapport à ceux utilisant des données complémentaires. Il serait intéressant de revoir les paramètres de classification et l'enchaînement des étapes pour l'améliorer. Nous avons également construit un zonier utilisant une base de données de données grêle achetée, qui nous a permis d'obtenir les meilleurs résultats. Toutefois, ceux-ci ne surpassent pas significativement ceux des autres zoniers construits avec des données complémentaires. Nous pouvons donc conclure que l'achat d'une base de données recensant les tempêtes de grêle n'est pas impératif pour construire un zonier grêle satisfaisant.

En revanche, la paramétrisation du modèle de classification utilisé, ici HClustGeo, est primordiale. Celle-ci permet de répondre à nos besoins en termes de segmentation et de mutualisation du risque entre les communes. Ce mémoire a contribué à décrire un processus ainsi que des indicateurs qui permettent de trouver un ensemble de paramètres qui satisfasse les contraintes de construction.

Dans notre cas, nous étions face à un risque volatile et hétérogène sur le territoire, notre portefeuille était quant à lui représenté de manière hétérogène sur le territoire. Nous avons pu constater que le jeu de paramètre optimal, répondant à chaque contrainte, n'existe pas, mais nous nous sommes efforcés de répondre au mieux à l'arbitrage entre une bonne stabilité et une bonne segmentation.

Nous avons constaté dans ce mémoire qu'une tarification à la maille communale du produit Grêle en assurance climatique est trop ambitieuse. Nous avons ainsi fait le choix de créer des zones tarifaires afin de gagner en stabilité. À l'avenir, il pourrait être possible de tarifer finement par le biais d'un historique plus conséquent et également grâce à l'achat renouvelé de la base de données grêle qui offrira des relevés plus précis grâce au bond technologique.

Enfin, ce zonier répond aux besoins de Pacifica et est en mesure d'être mis en commercialisation. Il permet une segmentation beaucoup plus précise du risque Grêle sur le territoire. Une adaptation sera toutefois nécessaire pour répondre aux diverses exigences commerciales et à la potentielle volonté de mutualiser d'avantage le risque.

Table des figures

Liste des tableaux

Bibliographie

[1] Antonio AFALCO. “Classification ascendante hiérarchique (CAH)”. In : *GitHub* (). URL : https://afalco.github.io/analyse-R/classification-ascendante-hierarchique.html.

[2] Laurent BATISSE et Yann MERCUZOT. “Étude de l’assurance des rendements agricoles face aux risques climatiques”. Centre d’études actuarielles, 2009.

[3] Gennaro CAPPELLUTI. “A global hail climatology using the UK Met Office convection diagnosis procedure and model analyses”. In : *Meteorological Applications* 18 (2011), p. 446-458.

[4] M. CHAVENT et al. “ClustGeo : an R package for hierarchical clustering with spatial constraints”. In : *R-CRAN* (2018). URL : https://cran.r-project.org/web/packages/ClustGeo/vignettes/intro_ClustGeo.html.

[5] Alexis COUTURES. “Grêle sur les cultures : Encore trop peu d’agriculteurs assurés”. In : *Challenges* (2022). URL : //www.challenges.fr/economie/agriculture/grele-sur-les-cultures-pourquoi-si-peu-d-agriculteurs-sont-assures_816079.

[6] Thomas DOWNING, Alexandre OLSTHOORN et Richard TOL. *Climat, change et risk* . Routledge, 1999.

[7] L. GENEBES, A. RENAUD et F. SÉMÉRCURBE. *Lissage spatial*. INSEE. URL : https://www.insee.fr/fr/statistiques/fichier/3635442/imet131-l-chapitre-8.pdf.

[8] Mickael GRUBER. “Modélisation du risque grêle en France”. L’Institut de Statitisque de l’Université de Paris, 2017.

[9] M. HENNEQUI. “Spatialisation des données de modélisation par Krigeage”. Université de Strasboug, 2010. URL : https://dumas.ccsd.cnrs.fr/dumas-00520260/document.

[10] L'AFP. "Sécheresse, gel, grêle... le Parlement adopte la réforme de l'assurance récolte". In : *Challenges* (2022). URL : https://www.challenges.fr/politique/secheresse-gel-grele-derniere-ligne-droite-pour-la-reforme-de-l-assurance-recolte_802071.

[11] Fédération Française de L'ASSURANCE. "L'assurance agricole en 2020". In : (2021). URL : https://www.challenges.fr/politique/secheresse-gel-grele-derniere-ligne-droite-pour-la-reforme-de-l-assurance-recolte_802071.

[12] Joseph LARMARANGE. "Classification ascendante hiérarchique (CAH)". In : *GitHub* (). URL : https://larmarange.github.io/analyse-R/classification-ascendante-hierarchique.html.

[13] Le SÉNAT. "Projet de loi portant réforme des outils de gestion des risques climatiques en agriculture". In : (2022). URL : http://www.senat.fr/rap/a21-386/a21-3861.html.

[14] "Severe storm". In : *Géoscience Australie* (2013).

[15] A. T. "Orages : comment se forme la grêle ?" In : *Sud-Ouest* (2022). URL : https://www.sudouest.fr/environnement/meteo/orages-comment-se-forme-la-grele-11372516.php.

[16] Fredy VINET et patrick PIGEON. "La question du risque climatique en agriculture : le cas de la grêle en France". In : (2002).

Annexe A : Description de la base Infoclimat

Nous avons pu obtenir et agréger les données suivantes pour 719 stations :

- TXN :Température maximum absolue sur le mois (en °C)
- TNN :Température minimum absolue sur le mois (en °C)
- TXM : Moyenne des températures maximum quotidiennes sur le mois (en °C)
- TNM : Moyenne des températures minimum quotidiennes sur le mois (en °C)
- TMM : Moyenne entre TXM et TNM (en °C)
- TXX : Température maximum absolue sur le mois (en °C)
- RR : Cumul des précipitations sur le mois (en mm)
- Rafale : Vitesse maximale du vent sur le mois (en km/h)

Certaines stations n'ont pas de relevés complets sur la période qui nous intéresse. Nous retenons finalement les relevés de 383 stations pour une période de 14 années.

Annexe B : Extrait base de données CatNat.net

Date	Indicatif	Annee / Yea	Mois / Month	Jour / Da	Commune / Town	INSEE / Town code	Département
30/07/2017	39420	2017	07	30	PIMORIN	39420	39
30/07/2017	42005	2017	07	30	ANDREZIEUX-BOUTHEON	42005	42
30/07/2017	42010	2017	07	30	AVEIZIEUX	42010	42
30/07/2017	42011	2017	07	30	BALBIGNY	42011	42
30/07/2017	42021	2017	07	30	BOISSET-SAINT-PRIEST	42021	42
30/07/2017	42022	2017	07	30	BONSON	42022	42
30/07/2017	42029	2017	07	30	BUSSIERES	42029	42
30/07/2017	42030	2017	07	30	BUSSY-ALBIEUX	42030	42
30/07/2017	42039	2017	07	30	CHALMAZEL	42039	42
30/07/2017	42041	2017	07	30	CHAMBEON	42041	42
30/07/2017	42062	2017	07	30	CHEVRIERES	42062	42
30/07/2017	42088	2017	07	30	EPERCIEUX-SAINT-PAUL	42088	42
30/07/2017	42096	2017	07	30	FONTANES	42096	42
30/07/2017	42097	2017	07	30	FOUILLOUSE	42097	42
30/07/2017	42100	2017	07	30	GIMOND	42100	42
30/07/2017	42102	2017	07	30	GRAMMOND	42102	42
30/07/2017	42133	2017	07	30	MARCENOD	42133	42
30/07/2017	42155	2017	07	30	NERVIEUX	42155	42
30/07/2017	42189	2017	07	30	ROCHE-LA-MOLIERE	42189	42
30/07/2017	42206	2017	07	30	SAINT-BONNET-LES-OULES	42206	42
30/07/2017	42207	2017	07	30	SAINT-CHAMOND	42207	42

FIGURE 89 – Extrait n°1 de la base de données CatNat.net

Nom departement / Departement name	Population	Region	Longitude	Latitude	Classe de diamètre / Diameter clas	Classe simple / Simple class	Diamètre min / Min diamete	Diamètre max / Max diamete
Jura	195	Bourgogne-Franche-Comté	5.502778053	46.50388718	A1 (1 à 1.9 cm)	A1	1	1.9
Loire	10061	Auvergne-Rhône-Alpes	4.25694418	45.52694321	A4 (4 à 4.9 cm)	A4	4	4.9
Loire	1601	Auvergne-Rhône-Alpes	4.371943951	45.56611252	A3 (3 à 3.9 cm)	A3	3	3.9
Loire	2989	Auvergne-Rhône-Alpes	4.185555935	45.8172226	A2 (2 à 2.9 cm)	A2	2	2.9
Loire	1209	Auvergne-Rhône-Alpes	4.103332996	45.51166534	A1 (1 à 1.9 cm)	A1	1	1.9
Loire	3765	Auvergne-Rhône-Alpes	4.229166985	45.52000046	A1 (1 à 1.9 cm)	A1	1	1.9
Loire	1612	Auvergne-Rhône-Alpes	4.268610954	45.83610916	A2 (2 à 2.9 cm)	A2	2	2.9
Loire	515	Auvergne-Rhône-Alpes	4.036388874	45.79527664	A1 (1 à 1.9 cm)	A1	1	1.9
Loire	472	Auvergne-Rhône-Alpes	3.850277901	45.70360947	A3 (3 à 3.9 cm)	A3	3	3.9
Loire	531	Auvergne-Rhône-Alpes	4.175000191	45.69722366	A3 (3 à 3.9 cm)	A3	3	3.9
Loire	1092	Auvergne-Rhône-Alpes	4.40194416	45.58916855	A1 (1 à 1.9 cm)	A1	1	1.9
Loire	705	Auvergne-Rhône-Alpes	4.213611126	45.79249954	A3 (3 à 3.9 cm)	A3	3	3.9
Loire	676	Auvergne-Rhône-Alpes	4.435555935	45.54722214	A3 (3 à 3.9 cm)	A3	3	3.9
Loire	4414	Auvergne-Rhône-Alpes	4.320000172	45.50222015	A3 (3 à 3.9 cm)	A3	3	3.9
Loire	292	Auvergne-Rhône-Alpes	4.41055584	45.55833435	A3 (3 à 3.9 cm)	A3	3	3.9
Loire	907	Auvergne-Rhône-Alpes	4.44166708	45.56638718	A2 (2 à 2.9 cm)	A2	2	2.9
Loire	700	Auvergne-Rhône-Alpes	4.483333111	45.57361221	A3 (3 à 3.9 cm)	A3	3	3.9
Loire	969	Auvergne-Rhône-Alpes	4.155832767	45.80666733	A2 (2 à 2.9 cm)	A2	2	2.9
Loire	10240	Auvergne-Rhône-Alpes	4.323888779	45.4347229	A1 (1 à 1.9 cm)	A1	1	1.9
Loire	1602	Auvergne-Rhône-Alpes	4.329721928	45.54444504	A3 (3 à 3.9 cm)	A3	3	3.9
Loire	35933	Auvergne-Rhône-Alpes	4.507500172	45.47277832	A2 (2 à 2.9 cm)	A2	2	2.9

FIGURE 90 – Extrait n°2 de la base de données CatNat.net

Annexe C : Présentation dendrogramme

Exemple d'un dendrogramme obtenu à partir d'une classification ascendante hiérarchique utilisant la distance de Ward :

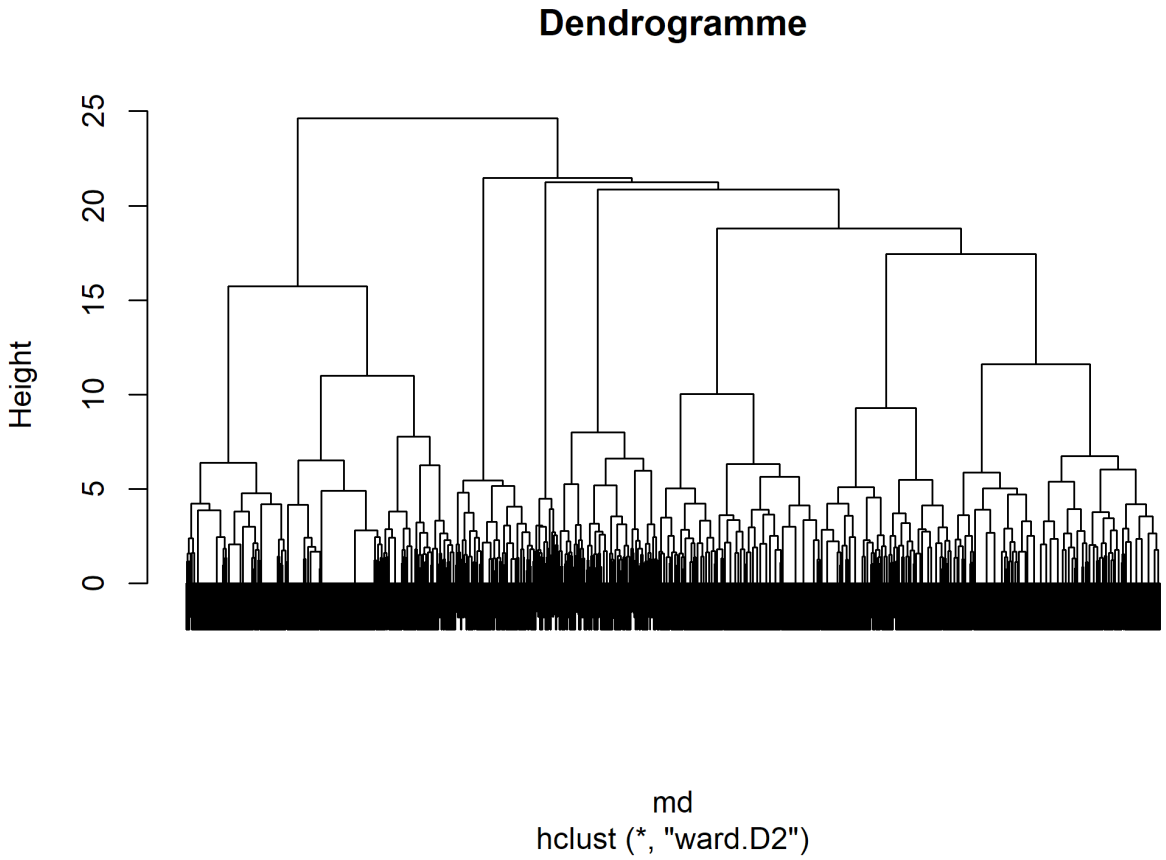


FIGURE 91 – Dendrogramme (source : Classification ascendante hiérarchique (CAH) de J.Larmarange [12]

Pour segmenter la population, il est possible de séparer le dendrogramme à une certaine hauteur. Le choix de la hauteur dépend des contraintes et des objectifs de l'utilisateur. Il faut également regarder les noeuds séparant le plus d'observations. Dans l'exemple en Figure 91, une forte séparation se trouve au premier et au cinquième noeud. Afin de faciliter notre démarche, il est envisageable de tracer les sauts d'inertie du dendrogramme en fonction du nombre de groupes sélectionnés. Nous avons préféré cette démarche dans le mémoire.