Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires le 8 février 2021

Par : Martin DUBOST
Titre : Utilisation de données non structurées pour modéliser la fréquence des incendies en MRH.

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

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : Sia Partners
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

Directeur de Mémoire en entreprise :
Nom : Fabien CHERANCE
Signature :

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

Secrétariat :

Signature du responsable entreprise

Bibliothèque :

Signature du candidat

Résumé

En assurance multirisque habitation, le risque incendie est l'une des sources principales de coût pour l'assureur. Du fait de sa fréquence relativement faible, la modélisation de ce risque est complexe et dépend fortement de la qualité de l'historique de sinistres de l'assureur. Dans le monde concurrentiel de l'assurance, mieux comprendre un risque permet de gagner en attractivité et, ainsi, d'accroître ses parts de marché. L'essor des données publiques et d'internet offrent de nouvelles sources d'information. L'objectif de l'étude est de présenter diverses techniques de récupération de données publiques (*web scraping*, utilisation d'API, TAL) ainsi qu'un cadre d'utilisation respectueux des sources. Dans un second temps, les données non structurées récupérées sont retraitées pour construire une base de données permettant la modélisation de la fréquence des incendies. Les méthodes classiques de modélisation en assurance sont utilisées (MLG, régression pénalisée) avec une lissage spatial pour construire un zonier sur la France à la maille commune. Cette étude se termine par la production de variables et graphiques permettant de déterminer la valeur de l'information qui peut être extraite de données publiques non structurées.

Mots-clés : *MRH*, *Risque incendie*, *MLG*, *Web-scraping*, *TAL*, *Zonier*, *API*, *RGPD*.

Abstract

One of the most material perils covered by Home Insurance is Fire. Modelling this risk can be particularly difficult and reliant on historical data's quality and depth. In competitive markets, a better understanding of the risk leads to better pricing, healthy business growth being conditional to such a process. The rise of public data and especially the one held on the internet offers new information sources to support this exercise. The dissertation aims at presenting several data-science techniques for data collection (scraping, use of APIs, NLP) as well as the framework to do this in a respectful manner of the data providers. As a second step, the produced unstructured data is pre-processed to build a fire frequency modelling database. Classical modelling methods are used (GLM, penalised regressions) along with spatial smoothing to build scores later displayed on mainland France maps at municipality mesh. This process ends up producing variables and plots that allow us to assess whether some value could be extracted from consolidated web-data sources.

Keywords : *Home insurance* , *Fire risk*, *GLM*, *Web-scraping*, *NLP*, *Zoning*, *API*, *GDPR*.

Note de Synthèse

Contexte de l'étude et problématique

L'assurance multirisque habitation (MRH) est un des différents contrats de l'assurance IARD qui permet aux particuliers de couvrir et protéger leur logement, son contenu ainsi que la responsabilité civile des résidents. Les contrats MRH comportent plusieurs types de garanties dont notamment les dégâts des eaux, l'incendie et le vol.

Cette étude portera plus particulièrement sur la garantie incendie et explosion. Cette garantie est obligatoire dans les polices de MRH. En France, il y a chaque année plus de 260 000 sinistres d'incendie déclarés aux assureurs, soit en moyenne un toutes les deux minutes. Ce nombre a doublé en 20 ans.

La garantie incendie est une garantie MRH complexe à modéliser. Le risque incendie, avec les dégâts des eaux et le vol sont les plus grandes sources de frais en MRH pour l'assureur. Le coût moyen d'un incendie en 2018 est estimé à 7 420 € pour une fréquence de 0,55%. En complément de la garantie dommages sur les biens assurés, le risque incendie peut également activer la garantie responsabilité civile en cas de propagation, représentant une couverture financière potentiellement plus élevée que la valeur du bien assuré.

Malgré un nombre relativement faible de sinistres, la distribution de la fréquence du nombre d'incendies peut être représentée par des lois classiques comme la loi de Poisson. Différents facteurs de risques sont étudiés pour affiner les modèles de tarification comme la densité de population, le niveau de richesse de la zone géographique, l'ancienneté ou les normes en place.

L'objectif de ce mémoire est d'étudier la possibilité et l'intérêt d'utiliser des données publiques non structurées, pour modéliser à l'échelon national sur la fréquence du risque incendie en MRH, indépendamment du portefeuille d'un assureur particulier. Le risque incendie ne dépendant pas uniquement de l'assuré mais de son environnement, l'ajout de données externes peut en effet permettre d'améliorer la compréhension du risque. Pour cela, une base de données a été établie avec la variable réponse (nombre d'incendies) ainsi que différentes variables descriptives expliquant les incendies. La finalité n'est pas de remplacer les modélisations existantes, mais de les compléter avec une vision extérieure.

Construction de la base de données

Pour créer la base de données décrivant la variable réponse « nombre d'incendies par commune », différentes techniques issues de la science des données sont utilisées comme le *web scraping* ou le Traitement Automatique des Langues (TAL). Cette variable réponse est complétée par différentes bases de données publiques. L'ensemble des données récoltées sont utilisées sous le respect des différentes réglementations en place.

Le Règlement Général sur la Protection des Données ou RGPD est une règlementation européenne. L'objectif du RGPD est de définir un texte de référence en matière de gestion et de protection des données personnelles et de faire appliquer des règles identiques aux différents acteurs étatiques ou

privés. L'utilisation des données et la mise en place des modèles sont faites dans le respect des normes de pratique relatives aux modèles actuariels et à l'utilisation et la protection des données massives, des données personnelles et des données de santé à caractère personnel (NPA 2 et NPA 5).

Le ***web scraping*** regroupe les diverses techniques permettant de récupérer des informations sur des pages internet comme par exemple des listes d'articles, des adresses ou des numéros. Les données scrapées peuvent être utilisées seules ou enrichies avec une base déjà disponible.

Trois droits régulent les données issues de *web scraping* : le droit pénal, le droit de la concurence et le droit de la propriété intellectuelle. Pour respecter les conditions d'utilisation d'un site, les pages `robots.txt` détaillent les différentes interdictions et autorisations.

Les sites d'informations sont des sources importantes de données et le *web scraping* permet d'obtenir une estimation du nombre d'incendies à la maille INSEE sur l'ensemble de la France. Ainsi, pour une page donnée de fait divers, l'ensemble des titres et dates des articles sont récupérés. L'ensemble du corps d'un article n'est pas récupéré pour gagner en temps de réalisation.

Le choix des différents journaux retenus s'est fait après un parcours exhaustif des sites régionaux d'actualités pour assurer une couverture uniforme des incendies sur l'ensemble de la France.

Le **Traitement Automatique des Langues**, TAL ou NLP pour *Natural Language Processing* est un mélange de plusieurs domaines techniques comme la linguistique, l'informatique ou l'intelligence artificielle. Son objectif est le développement d'outils permettant l'interaction entre un ordinateur et un humain.

Le TAL peut être décomposé en plusieurs étapes. Les deux principales sont la segmentation et la labellisation. La première permet de découper un texte (une chaîne de caractères) en un ensemble de mot distincts : les tokens. La seconde consiste à identifier pour chaque token son ensemble grammatical.

Le TAL permet d'identifier les articles concernant les incendies. Une base de mot clés permettant l'identification de chaque article est définie après une étude de plusieurs titres. Cette base comprend différents mots comme : « incendié » ou « brûlée ». Si aucun des tokens du titre ne contient un des mots de la base, le titre n'est pas conservé. Sur les articles conservés après vérification du sujet, les tokens catégorisés comme nom propre sont utilisés pour identifier une commune.

Différentes **variables explicatives** sont captées dans plusieurs bases de données publiques pour compléter la variable réponse construite avec le *web scraping*. Deux bases de l'INSEE sont utilisées. La première contient des informations géographiques comme la surface des communes. La seconde, issue du recensement, comptabilise par commune, le nombre de ménages correspondant à certains critères prédéfinis. La base d'artificialisation contient les informations relatives à l'occupation des sols et les évolutions dans le temps ainsi que le taux d'emplois par rapport à la surface urbanisée. L'ensemble des transactions notariées en France sur un historique de cinq ans sont également récupérées dans la base de demande de la valeur foncière.

Une **gestion de l'historique** est nécessaire. La base des incendies récupérés avec le *web scraping* contient ainsi pour chaque ligne un code INSEE et la date de publication de l'article. Pour pouvoir être exploitée cette base doit être agrégée par INSEE et par période. Dans le cas du risque incendie, de probabilité faible, la période sera l'année, qui correspond à la durée classique d'un contrat MRH.

La variable réponse qui comptabilise le nombre d'incendies sur une commune par an est utilisée pour créer la variable explicative historique. Cette variable permet de structurer la sinistralité sur les cinq dernières années. Cette moyenne glissante permet de lisser les années exceptionnelles.

Données utilisées et modélisation

Les différentes bases à l'échelle nationale sont agrégées avec le code INSEE. Une étude des variables explicatives est réalisée sur la base des sinistres scrapés. Cette base représente l'ensemble des codes INSEE de France sur un historique de quatre ans : de 2017 à 2020. L'analyse de la corrélation entre les variables permet de supprimer celles qui ne sont pas significatives et d'en agréger certaines pour réduire la dimension de l'espace de modélisation.

Le nombre d'incendies est modélisé en fonction des variables explicatives selon un modèle linéaire généralisé. La dimension de l'espace de modélisation est également réduite grâce à la régression pénalisée qui permet de limiter la valeur des coefficients et de sélectionner les variables les plus pertinentes.

Zonier

Cette base des incendies scrapés sera utilisée pour mettre en place un zonier. L'objectif d'un zonier est de représenter sur une seule variable l'ensemble des risques liés à l'environnement du bien assuré.

Pour cela, en partant de la base initiale, deux modèles sont utilisés pour estimer le nombre d'incendies par commune. Le modèle complet utilise l'ensemble des variables explicatives, tandis que le modèle interne n'utilise que les variables qui décrivent le bien et pas l'environnement dans lequel ce dernier se trouve. Les résidus, issus de la comparaison des prédictions de ces deux modèles représentent ainsi le risque géographique. Ces résidus sont ensuite lissés pour réduire les écarts entre les communes limitrophes, car le risque incendie est très diffus. Après le lissage, un nombre fini de modalités est conservé à l'aide d'un algorithme de partitionnement en k-moyennes pour consolider le calcul du niveau de risque et permettre son emploi dans un modèle ainsi qu'une représentation visuelle.

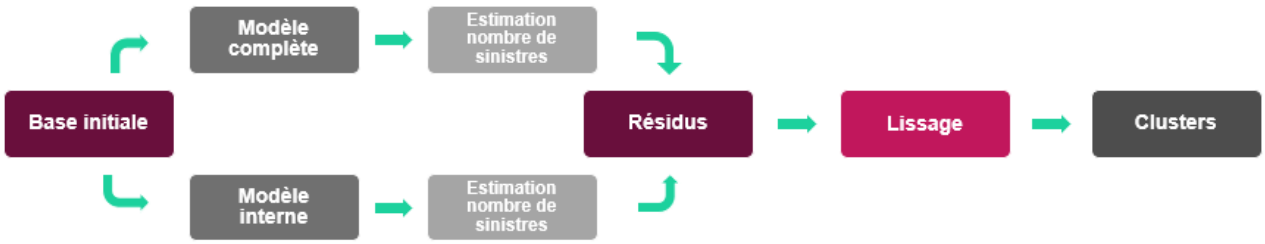


FIGURE 1: Schéma de mise en place d'un zonier sur une base de données

La **théorie de la crédibilité** permet de lisser le risque entre communes. Un portefeuille possède des risques hétérogènes. Pour améliorer la tarification, il est nécessaire de déterminer plusieurs groupes d'assurés. La théorie de la crédibilité permet de pondérer l'historique personnel avec l'expérience de l'ensemble du portefeuille. Pour cela, les probabilités, individuelles et collectives sont prises en compte. Albert Withney a défini en 1918 la prime individuelle par : $P = \alpha X + (1-\alpha)C$, où C et X sont respectivement l'expérience collective et l'expérience individuelle. Le facteur de crédibilité α peut prendre différentes formes.

La théorie de la crédibilité (3.2.5) est utilisée pour lisser les résidus. Soit r_i le résidu d'une commune i, d'après (3.14) alors le résidu lissé r_i^* est défini par la formule :

$$r_i^* = z(e_i)r_i + (1 - z(e_i))\frac{\sum_j e_j r_j f(d_{i,j})}{\sum_j e_j f(d_{i,j})}, \tag{1}$$

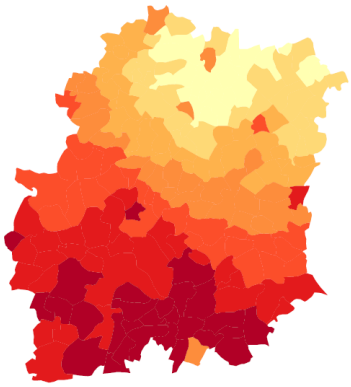
Le risque de chaque commune est alors rationnalisé. Pour chaque commune le risque est multiplié par son exposition rationalisée et comparé à l'ensemble des expositions des autres communes multiplié par l'inverse de la distance. La fonction inverse permet de réduire le lien entre les communes avec leur distance.

La distance $d_{i,j}$ entre deux communes i et j est calculée avec leurs coordonnées géographiques. La distance $d_{i,j}$ est donc définie par la formule suivante :
$d_{i,j} = R \times \arccos\left[\sin\varphi_i \times \sin\varphi_j + \cos\varphi_i \times \cos\varphi_j \times \cos\Delta\lambda\right]$ où $\Delta\lambda = \lambda_i - \lambda_j$, avec $R = 6371$ km, φ_i et λ_i respectivement la latitude et longitude de la commune i.

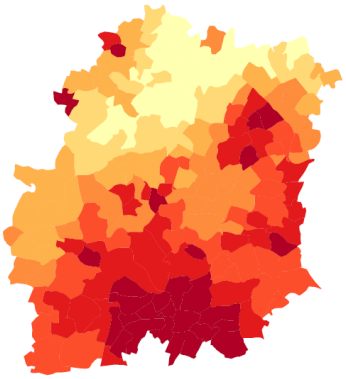
Zonier Essonne

Un premier zonier est réalisé sur le département de l'Essonne. La méthode mise en place pour réaliser un zonier à l'aide d'un modèle complet et d'un modèle interne est reproduite avec la base du SDIS sur les interventions pour incendie en Essonne. Cette table est supposée complète, c'est à dire comprenant l'ensemble de la variable réponse à savoir le nombre d'interventions.

En partant de la base des incendies scrapés réduite au département de l'Essonne, une deuxième base est créée en remplaçant la variable réponse par les valeurs figurant sur la base du SDIS. Sur ces deux bases, deux zoniers du risque de survenance d'incendie par commune sont réalisés avec la même méthode et les mêmes paramètres de lissage. Les deux figures ((6a) et (6b)) représentent, avec sept niveaux de risque (croissant lorsque la couleur tend vers le rouge), le risque incendie en Essonne.



(a) Zonier fréquence incendie base SDIS



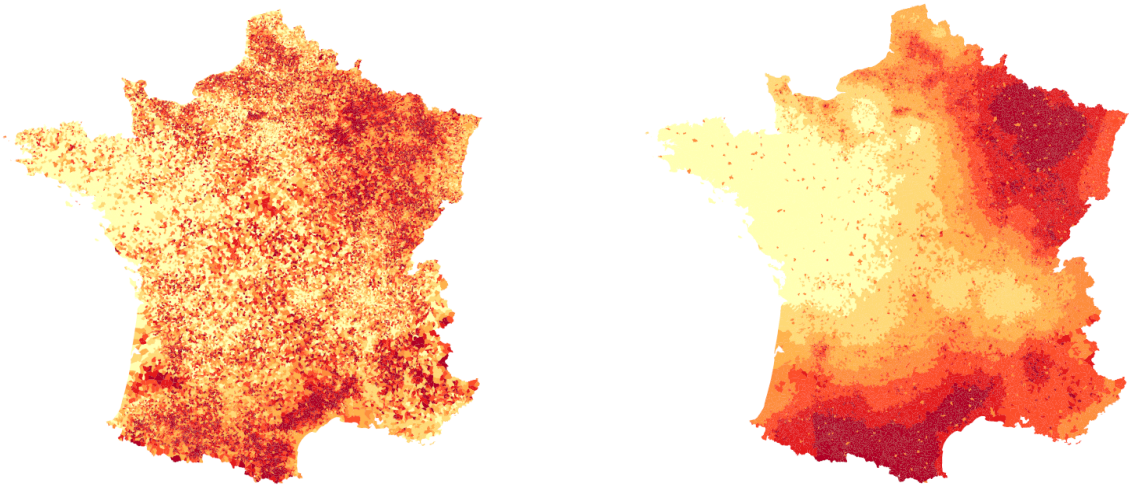
(b) Zonier fréquence incendie données scrapées

FIGURE 2: Zoniers de la fréquence incendie pour l'ensemble des communes en Essonne pour deux bases de données différentes avec le même lissage.

Dans les deux figures, assez semblable, le risque apparait plus important dans le sud du département qui est plus rural que le nord qui est plus proche de Paris. Cependant dans l'est du département, le risque apparaît moins bien modélisé par la base de données d'incendies scrapés.

Zonier France

Un second zonier est mis en place sur l'ensemble de la France avec la base de données scrapées. Ce zonier est de même lissé à l'aide de la théorie de la crédibilité. La figure suivante (7) présente pour deux lissages le niveau de risque de survenance d'incendie. Le dégradé de couleurs correspond au niveau du risque, plus la couleur est rouge, plus la commune est exposée.



(a) Zonier lissé, $a = 2$, $m = 1$

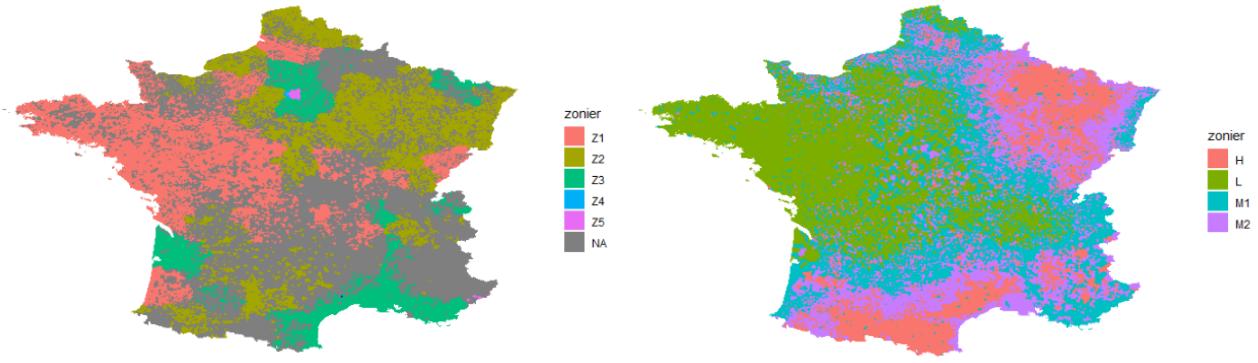
(b) Zonier lissé, $a = 10$, $m = 3$

FIGURE 3: Zoniers de la fréquence incendie pour l'ensemble des communes en France pour différents lissages.

Le lissage est réalisé de manière visuelle pour obtenir un lissage cohérent entre communes voisines c'est à dire ne pas avoir deux codes INSEE géographiquement limitrophes avec un écart de risque trop important. Un test de qualité de prédiction sur le modèle interne aurait pu être réalisé pour chaque lissage. Deux régions ont une survenance forte : l'Occitanie et le Grand-Est. L'ouest de la France porte un risque faible.

Test du zonier incendie

Le zonier scrapé est testé sur un portefeuille d'assureur MRH. Pour cela un modèle interne est simulé avec les différentes variables explicatives présentes dans ce portefeuille. Deux mesures de qualité sont réalisées en changeant la variable zonier incendie du portefeuille par le zonier scrapé. La figure suivante (8) présente le zonier assureur (8a) et le zonier scrapé (8b) avec seulement quatre modalités, suite à l'application de l'algorithme de k-moyennes.



(a) Zonier incendie du portefeuille assureur.

(b) Zonier scrapé à 4 modalités après lissage.

FIGURE 4: Zoniers de la fréquence incendie en France.

La qualité de prédiction est la même pour les deux zoniers. Le zonier réalisé sur les données scrapées permet donc d'obtenir le même niveau d'information qu'avec le zonier incendie de l'assureur.

Table 1: Evalution de la qualité de prédicton des modèles selon le zonier choisit.

Zonier	MAE	RMSE	Déviance
Zonier de l'assureur	0,005063	0.051230	5946,01
Zonier sur les données scrapées	0,005062	0,051232	5945,72

Néanmoins, la fréquence observée dans le portefeuille de l'assureur est bien plus importante que dans la base des incendies scrapés.

Le nombre d'incendies divisé par l'exposition dans le portefeuille d'assureur est le même qu'annoncé par la FFA : 0,55%, alors que dans la base des incendies scrapés, ce chiffre est 500 fois plus faible.

Il apparait finalement que le zonier mis en place ne permet pas un gain significatif dans la compréhension de la fréquence incendie mais permet de maintenir le même niveau de prédiction avec moins d'information.

Application sur le modèle interne d'un assureur

Un assureur historique de la MRH en France a accepté de tester différents résultats produits précédemment. Il s'agit, pour chaque couple (commune,année), de la prédiction du nombre d'incendies ainsi que le zonier scrapé avant regroupements en modalités. Ces deux variables sont ajoutées à leur ensemble de variables.

L'assureur étudie séparément la modélisation des fréquences pour les appartements et les maisons. Pour les appartements, aucune des variables communiquées n'a été conservée lors de la sélection. Pour les maisons, seule la variable zonier a été conservée. Sur les 13 variables conservées le zonier est la sixième plus importante. Trois critères de mesure de qualité sont utilisés par l'assureur : le coefficient de Gini, la déviance et le RMSE. Par rapport au modèle sans la variable zonier, différentes améliorations sont observées : gain de 0,48 point de Gini, amélioration de 0,1% de déviance et le RMSE reste stable.

Le gain sur le coefficient de Gini permet de mieux discriminer les variables en ordonnant le risque, et ainsi faciliter la mise en place de règles de souscriptions. Cette amélioration ne se fait pas au détriment de la qualité de prédiction car le RMSE reste stable. L'apport de la variable zonier au modèle est faible au vu de l'amélioration de la déviance.

Les informations récupérées à l'aide des données non structurées peuvent apporter une amélioration dans la modélisation de la fréquence des incendies de l'assureur. Cependant, ce gain n'est présent que sur les ≪ maisons ≫ qui sont plus exposées aux facteurs géographiques. L'utilisation du TAL pour séparer les informations selon le type de bien est un axe d'amélioration possible.

Limites de l'étude

- Le *web scraping*
  - Le *scraping* a été réalisé sur une période très courte. Une mise en place sur une période plus longue permettrait de réduire l'influence de la gestion des historiques sur les sites scrapés.
  - La date de publication est considérée comme la date d'incendie. Cette hypothèse suppose que la vitesse de publication de chaque journal est la même. De plus il n'est pas possible de différencier un incendie qui aurait fait l'objet d'articles pendant plusieurs jours.
  - Les incendies mineurs ne sont pas systématiquement rapportés dans la rubrique faits divers. Une partie des incendies attritionnels est donc censurée.

- Traitement automatique des langues
  - Le TAL n'a été mis en application seulement sur les titres et les chapeaux d'articles pour permettre un traitement plus rapide. Une utilisation approfondie de ces techniques pourrait permettre de mieux distinguer les incendies.
  - Les incendies détectés ne permettent pas la distinction entre risque pour les contrats MRH ou MRP. L'hypothèse est de considérer qu'un incendie sur un bien en MRP présente un risque pour les biens MRH voisins.
- Modélisation
  - L'ensemble des variables explicatives est fixe avec les années. Seule la variable historique est cadencée par année.
  - L'âge et la profession des occupants d'un bien ont une forte influence sur le risque incendie. Cependant ces variables ne sont pas représentées dans les variables explicatives.
  - Les différentes modélisations sont réalisées avec des GLM. Ces modèles sont utilisés en assurance pour leur transparence. L'utilisation de méthodes moins transparentes pourrait permettre une meilleure modélisation du nombre d'incendie.

Conclusion

Les techniques de science des données permettent de récupérer par différentes manières des données applicables à l'assurance. Néanmoins, ces données non structurées doivent être retraitées pour former une base de données exploitable. Dans le cadre de cette étude, les données ont servi à mettre en place un score de survenance d'incendie sur les logements de particuliers sur l'ensemble des communes de France métropolitaine.

Dans premier temps, un test visuel du score de survenance entre les données issues du *scraping* et les données des services de secours est réalisé et montre une compréhension globale risque « incendies » sur un département.

Ensuite, un test sur un portefeuille d'assureur avec un modèle interne simulé montre que même si le gain en qualité prédictive n'est pas significatif, la mise en place d'un zonier indépendant du portefeuille de l'assureur permet de maintenir le niveau de compréhension du risque malgré une fréquence d'incendie observée plus faible. En effet, la fréquence d'incendie est 500 fois plus faible sur la base de données issues du *scraping* que dans le portefeuille.

Finalement, une application dans le modèle interne actuellement utilisé par un assureur historique de la MRH en France permet de confirmer la présence d'information sur des données non structurées. Même si le zonier n'est significatif que pour la modélisation de la fréquence des « maisons », les résultats sont encourageants. La modélisation des incendies sur les appartements est moins liée aux facteurs environnementaux. Pour un assureur souhaitant étendre sa zone de marché et disposant de peu de données clients locales, ce type d'outil pourrait présenter un intérêt important pour déterminer sa tarification.

L'évolution des techniques des *web scraping*, de Traitement Automatique des Langues et la démocratisation des bases de données publiques devrait permettre d'obtenir avec le temps de plus en plus d'informations et ainsi permettre d'améliorer la qualité des données récupérées et les modèles réalisés pour les représenter.

Synthesis note

Home insurance is one of the various Property and Casualty insurance contracts which allows individuals to cover and protect their home, its contents as well as the civil liability of residents. Home insurance contracts include several types of covers such as water damage, theft or fire and explosion.

This study will focus more particularly on the fire and explosion cover. This cover is mandatory in the policies of home insurance. In France, there are over 260,000 fire claims each year reported to insurers, an average of one every two minutes. This number has doubled in 20 years.

The fire cover is a home insurance cover that is complex to model. The risk of fire, along with water damage and theft are the biggest sources of costs for the home insurer. The average cost of a fire in 2018 is estimated at 7,420€ and a frequency of 0.55 %. In addition to the damage cover on the insured goods, the fire risk can also activate the civil liability cover in the event of propagation, representing a financial cover potentially higher than the value of the insured property.

Despite a relatively small number of claims, the frequency distribution of the number of fires can be represented by classical laws such as Poisson's law. Different risk factors are studied to refine pricing models such as population density, level of wealth in the geographic area, seniority, or the standards in place.

The objective of this dissertation is to study the feasibility of developing, using unstructured public data, a national model on the frequency of fire risk in home insurance, regardless of the portfolio of a particular insurer. As the fire risk does not depend solely on the policyholder but on his environment, the addition of external data can help improve understanding of the risk. To do this, a database was established with the response variable (number of fires) as well as various descriptive variables explaining the fires. The aim is not to replace existing models, but to complete them with an external vision.

Database construction

To create the database describing the response variable "number of fires per municipality", different techniques from data science are used such as *web scraping* or Natural Language Processing. This response variable is supplemented by various public databases. All the data collected is used in compliance with the various regulations in place.

The General Data Protection Regulation or GDPR is a European regulation. The objective of the GDPR is to define a reference text for the management and protection of personal data and to apply identical rules to the various state or private actors. The use of data and the implementation of models are carried out in compliance with the standards of practice relating to actuarial models and the use and protection of big data, personal data, and personal health data (French Actuarial Standards 2 and 5).

Web scraping groups together the various techniques making it possible to retrieve information on internet pages such as lists of articles, addresses or numbers for example. Scraped data can be

used alone or enriched with an already available database.

Three rights regulate the data obtained from *web scraping*: criminal law, competition law and intellectual property law. To comply to conditions of use of a website, the `robots.txt` pages detail the various prohibitions and authorizations.

Information websites are an important source of data and the *web scraping* makes it possible to obtain an estimate of the number of fires at the INSEE network over the whole of France. Thus, for a given news item page, all the titles and dates of the articles are retrieved. The entire body of an article is not recovered to save production time.

The choice of the various newspapers retained was made after a thorough examination of regional news websites to ensure uniform coverage of fires throughout France.

Natural Language Processing, NLP, is a mixture of several fields such as linguistics, computer science or artificial intelligence. Its objective is the development of tools allowing the interaction between a computer and a human.

NLP can be broken down into several steps. The two main ones are segmentation and labeling. The first allows you to split a text (a string of characters) into a set of distinct words: tokens. The second consists in identifying for each token its grammatical set.

The NLP identifies articles relating to fires. A keyword base allowing the identification of each article is defined after a study of several titles. This database includes different words such as: "burned". If none of the tokens of the title contains one of the words in the database, then this title is not kept. On articles kept after checking the subject, tokens categorized as proper names are used to identify a municipality.

Different **explanatory variables** are captured in several public databases to complete the response variable constructed with *web scraping*. Two INSEE databases are used. The first contains geographic information such as the area of municipalities. The second, from the census, counts the number of households by municipality that correspond to different criteria. The artificialization database contains information relating to land use and changes over time as well as the employment rate in relation to the urbanized area. All notarized transactions in France over a five-year history are also collected in the land value demand base.

History management is required. The database of fires recovered with *web scraping* thus contains an INSEE code for each line and the publication date of the article. In order to be able to be used, this database must be aggregated by INSEE and by period. In the case of fire risk, of low probability, the period will be one year, which corresponds to the traditional duration of a home insurance contract.

The response variable which counts the number of fires in a municipality per year is used to create an explanatory variable: the history. This history makes it possible to consider the loss experience over the last five years. This moving average can smooth out exceptional years.

Data used and modeling

The various databases at the national level are aggregated with the INSEE code. A study of the explanatory variables is carried out based on claims scraped. This base represents all the INSEE codes in France over a four-year history: from 2017 to 2020. The analysis of the correlation between the variables makes it possible to remove and aggregate them to reduce the size of the modeling space.

The number of fires is modeled as a function of the explanatory variables according to a generalized linear model. To reduce the dimension, penalized regression is used. It makes it possible to limit the value of the coefficients and to select the most relevant variables.

Zoning

This base of scraped fires will be used to set up a zoning. The objective of a zoning is to represent all the risks linked to the environment of the insured property on a single variable.

To do this, starting from the initial base, two models are used to estimate the number of fires per municipality. The full model uses all the explanatory variables, while the internal model only uses the variables that describe the asset and not the environment in which it is located. The residuals resulting from the comparison of the predictions of these two models thus represent the geographic risk. These residues are then smoothed out to reduce the differences between the neighboring municipalities, because the fire risk is very diffuse. After smoothing, a finite number of modalities is kept using a k-means partitioning algorithm to consolidate the calculation of the risk level and allow its use in a model as well as a visual representation.

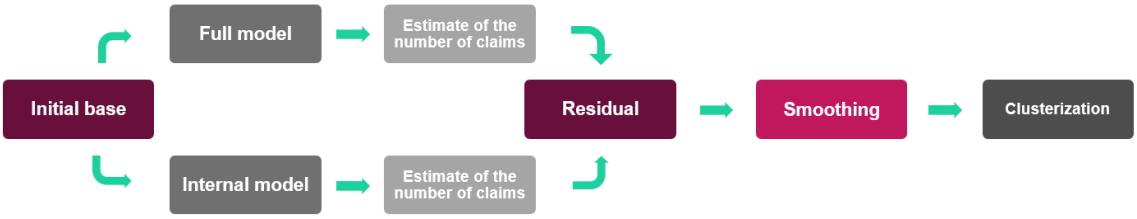


Figure 5: Diagram of the implementation of a zoning on a database

The **credibility theory** makes it possible to smooth the risk between municipalities. A portfolio has heterogeneous risks. To improve pricing, it is necessary to determine several groups of policyholders. Credibility theory allows personal history to be weighed against the experience of the entire portfolio. For this, the probabilities, individual and collective are considered. Albert Withney defined in 1918 the individual premium as: $P = \alpha X + (1-\alpha)C$, where C and X are respectively the collective experience and the individual experience. The credibility factor α can take different forms.

The credibility theory is used to smooth the residuals. Let r_i be the residue of a municipality i, then the smoothed residue r_i^* is defined by the formula:

$$r_i^* = z(e_i)r_i + (1 - z(e_i))\frac{\sum_j e_j r_j f(d_{i,j})}{\sum_j e_j f(d_{i,j})}, \tag{2}$$

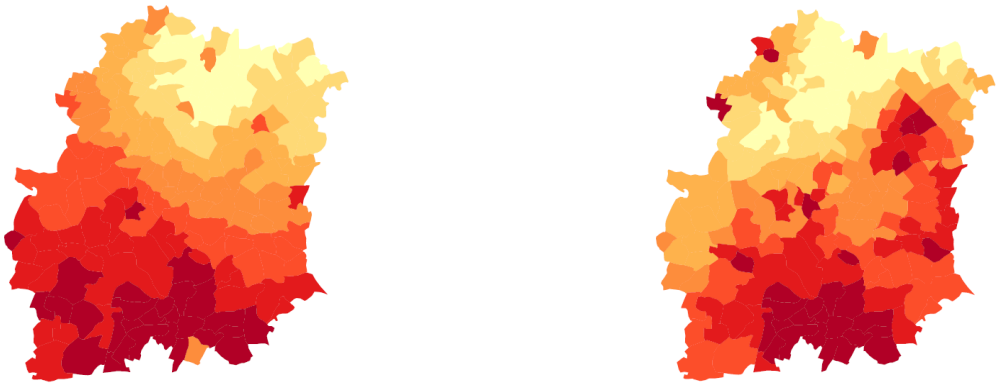
The risk of each municipality is then rationalized. For each municipality, the risk is multiplied by its rationalized exposure and compared to all the exposures of other municipalities multiplied by the inverse of the distance. The inverse function makes it possible to reduce the link between the municipalities with their distance.

The distance $d_{i,j}$ between two municipalities i and j is calculated with their geographic coordinates. The distance $d_{i,j}$ is therefore defined by the following formula :
$d_{i,j} = R \times \arccos\left[\sin\varphi_i \times \sin\varphi_j + \cos\varphi_i \times \cos\varphi_j \times \cos\Delta\lambda\right]$ where $\Delta\lambda = \lambda_i - \lambda_j$, with $R = 6371$ km, φ_i and λ_i respectively the latitude and longitude of the municipality i.

Essonne zoning

A first zoning is carried out in the department of Essonne. The method put in place to create a zoning using a complete model and an internal model is reproduced with the SDIS database on fire interventions in Essonne. This table is assumed to be complete, i.e. comprising all of the response variable, namely the number of interventions.

Starting from the base of scraped fires reduced to the department of Essonne, a second base is created by replacing the response variable by the values appearing based on the SDIS. On these two bases, two zonings of the risk of fire occurrence for each municipality are carried out with the same method and the same smoothing parameters. The two figures ((6a) and (6b)) represent the risk in Essonne, with seven levels of risk (increasing when the color tends towards red).



(a) SDIS base fire frequency zoning

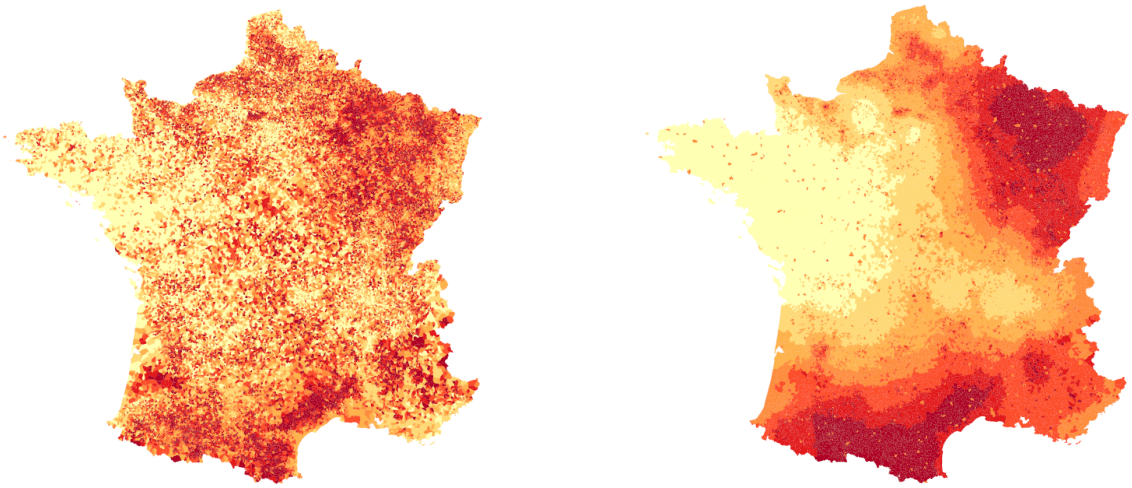
(b) Fire frequency zoning scraped data

Figure 6: Fire frequency zoning for all the communes in Essonne for two different databases with the same smoothing.

In both figures, the risk appears greater in the south of the department, which is more rural than the north, which is closer to Paris. However, in the east of the department, the risk appears to be less well modeled by the database of scraped fires.

France zoning

A second zoning has been set up throughout France with the scraped database. This zoning is also smoothed using the theory of credibility. The following figure (7) shows the risk level of fire occurrence for two smoothing. The color gradient corresponds to the level of risk, the redder the color is, the more the municipality is exposed to home fires.



(a) Smoothed zoning, $a = 2$, $m = 1$

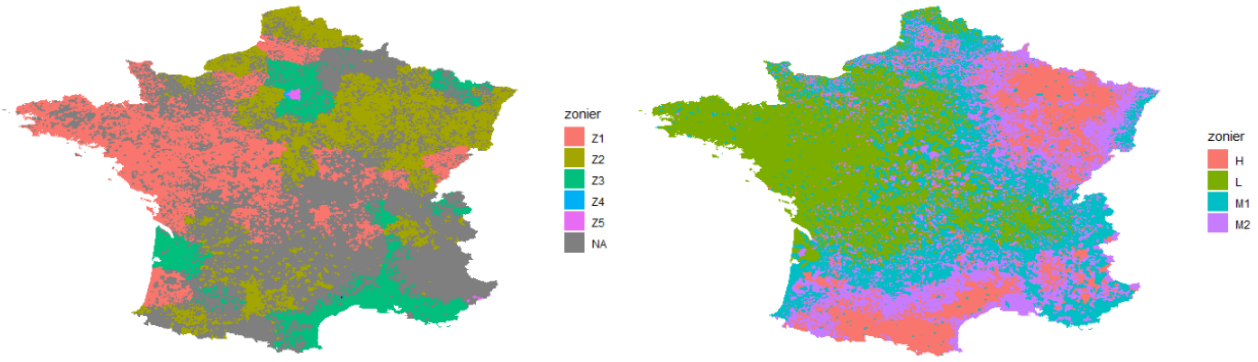
(b) Smoothed zoning, $a = 10$, $m = 3$

Figure 7: Fire frequency zoning for all municipalities in France and for different smoothing operations.

The smoothing is carried out visually to obtain a consistent smoothing between neighboring municipalities, that is to say not to have two geographically adjacent INSEE codes with an excessively large risk gap. A prediction quality test on the internal model could have been carried out for each smoothing. Two regions have a strong presence: Occitanie and Grand-Est. Western France carries a low risk.

Fire zoning test

The scraped zoning is tested on an MRH insurer portfolio. To do this, an internal model is simulated with the various explanatory variables present in this portfolio. Two quality measurements are carried out by changing the fire zone variable of the portfolio by the scraped zone. The following figure (8) presents the insurer zone (8a) and the scraped zone (8b) with only four modalities, following the application of the k-means algorithm.



(a) Insurer portfolio fire zoning

(b) Scraped zoning to 4 modalities after smoothing.

Figure 8: Insurer portfolio fire zoning

Table 2: Assessment of the model prediction quality according to the chosen zone.

Zoning	MAE	RMSE	Deviance
Insurer zoning	0,005063	0.051230	5946,01
Scraped data zoning	0,005062	0,051232	5945,72

The prediction quality is the same for the two zoning. The zoning carried out on the scraped data therefore makes it possible to obtain the same level of information as with the insurer's fire zone. Nevertheless, the frequency observed in the insurer's portfolio is much greater than the one in the base of scraped fires.

The number of fires divided by the exposure in the insurer's portfolio is the same as announced by the FFA: 0.55 %, while in the base of scraped fires, this figure is 500 times lower.

It finally appears that the zoning set up does not allow a gain in the understanding of the fire frequency but makes it possible to maintain the same level of prediction with less information.

Application on the internal model of an insurer

A historical insurer of the MRH in France has agreed to test various results produced previously. For each couple of commune and year, the prediction of the number of fires as well as the area scraped before regrouping into modalities. These two variables are added to their set of variables.

The insurer is studying the frequency modeling for apartments and houses separately. For apartments, none of the variables communicated were kept during the selection of variables. For houses, only the zoning variable has been retained. Of the 13 variables kept, the zoner is the sixth most important. The three criteria for measuring the quality of the insurer are the Gini coefficient, the deviance and the RMSE. Compared to the model without the zoning variable, various improvements are observed: gain of 0.48 Gini point, improvement of 0.1

The Gini gain makes it possible to better discriminate the variables by ordering the risk, and thus facilitate the implementation of underwriting rules. This improvement is not made to the detriment of the quality of prediction because the RMSE remains stable. The contribution of the zoning variable to the model is low in view of the improvement in deviance.

The information gathered using the unstructured data can provide an improvement in the modeling of the insurer's fire frequency. However, this gain is only present for "houses" which are more exposed to geographic factors. The use of NLP to separate information according to the type of property is a possible area for improvement.

Limitations of the study

- *Web scraping*
  - The *scraping* was done over a very short period of time. Setting it up over a longer period would reduce the influence of historical management on scraped websites.
  - The date of publication is considered the date of fire. This assumption assumes that the publication speed of each newspaper is the same. In addition, it is not possible to differentiate a fire that would have items over several days.
  - Minor fires are not systematically reported in the miscellaneous section. Part of attritional fires is therefore censored.
- Natural Language Processing
  - NLP has only been implemented on article titles and headlines to allow for faster processing. Further use of these techniques could help distinguish between fires better.
  - The fires detected do not allow the distinction between risk for home insurance or professional insurance contracts. The hypothesis is to consider that a fire on an commercial insurance asset presents a risk for neighboring home insurance assets.
- Modelization
  - The set of explanatory variables is fixed over the years. Only the historical variable is clocked by year.
  - The age and occupation of the occupants of a property have a strong influence on the fire risk. However, these variables are not represented in the explanatory variables.
  - The various modelizations are carried out with GLMs. These models are used in insurance for their transparency. The use of less transparent methods could allow better modeling of the number of fires.

Conclusion

Data science techniques allow for the recovery of insurance data in different ways. This unstructured data must be reprocessed to form a database.

The visual test in comparison between the data from the emph scraping and the data from emergency services shows an overall consistency of the understanding of the risk of fire occurring in a department.

The test on an insurer's portfolio with a simulated internal model shows that even if the gain in predictive quality is not significant, the establishment of a zone independent of the insurer's portfolio makes it possible to maintain the level of understanding risk despite a lower observed fire frequency. Indeed, the fire frequency is 500 times lower on the database from the *scraping* than in the portfolio.

The application in the internal model currently used by a historical insurer of the HRM in France makes it possible to confirm the presence of information on unstructured data. Even if the zoning is only significant for modeling the frequency of "houses", the results are encouraging. Modeling of apartment fires is less related to environmental factors.

The evolution of *web scraping* techniques, Automatic Language Processing and the democratization of public databases should make it possible to obtain more and more information over time and thus improve the quality of the data recovered and of models made to represent them.

Remerciements

Mes premiers remerciements sont adressés à Anthony Réveillac pour ses cours à l'INSA Toulouse et pour m'avoir permis de découvrir le monde de l'actuariat. Merci à Christophe Dutang, responsable du Master Actuariat de l'Université Paris Dauphine pour son temps, sa disponibilité et ses retours.

Mes remerciements vont aussi à l'ensemble du corps enseignant de l'INSA Toulouse et du Master Actuariat de l'Université Paris Dauphine pour la qualité de leurs différents enseignements.

Dans un second temps, je souhaite remercier Michaël Donio, Partner du pôle actuariat à Sia Partners de m'avoir accueilli au sein de son équipe. En particulier, je remercie Fabien Chérancé, Manager au sein de l'équipe d'actuariat, et tuteur de ce mémoire, pour son expertise du domaine et plus particulièrement en assurance IARD. Ses conseils et son intérêt m'ont guidé pendant l'ensemble de mon stage.

Pour son aide et le temps passé ensemble, je remercie Céline Houdayer ainsi que l'ensemble des stagiaires en actuariat. Je remercie aussi l'ensemble des actuaires de Sia Partners pour leur gentillesse et mon intégration au sein de l'équipe.

Je remercie Charles Borderie pour sa curiosité et les différents accès à sa technologie.

Finalement, je remercie l'ensemble de ma famille et mes amis pour leur soutien sans faille durant l'ensemble de ma scolarité dans l'enseignement supérieur.

Table des matières

Introduction

Le monde de l'assurance a toujours utilisé les outils mathématiques, puis informatiques, à sa disposition pour analyser et comprendre le comportement des assurés. Distinguer les clients présentant une fréquence et un coût de sinistre faibles, de ceux présentant une fréquence ou un coût élevés, est un enjeu déterminant pour la réussite d'une compagnie. En assurance, la tarification dépend en premier ordre de la nature des risques couverts et de son appropriation par l'assureur. Pour ce faire les compagnies d'assurances peuvent, soit utiliser l'historique de sinistres du client (c'est le principe du bonus-malus pour les assurances automobiles), soit interroger celui-ci sur sa situation propre en amont de la souscription d'un premier contrat (demande, par exemple, de certificats médicaux préalables pour les assurances vie). Ces informations sont essentielles pour définir le montant final de la prime.

Construire des modèles de tarification efficaces avec peu d'informations sur les caractéristiques du client s'avère essentiel. Poser trop de questions à un potentiel client pourrait rendre le processus long et rébarbatif et *in fine* préjudiciable à la souscription. Il faut donc un modèle reposant sur une base de données représentative pour rapidement catégoriser un client sans le soumettre à un questionnaire long et intrusif. Une compagnie d'assurance traditionnelle possède en général un historique de données adapté à son activité, mais ce n'est pas le cas de toutes les compagnies, et particulièrement des sociétés nouvellement créées qui ne disposent pas d'historique important.

Internet est actuellement une des sources les plus vastes de données, elle s'avère toutefois difficilement exploitable sous forme directe. Des techniques, comme notamment le web-scraping, permettent de récupérer des informations d'une page internet et de les consolider dans une base de données. Cet outil peut être un avantage concurrentiel très important pour augmenter la finesse de la modélisation d'un risque.

Le processus pour orienter la collecte et la transformation de sources de données non-structurées doit être appuyé par des études applicatives claires de tarification et d'évaluation du risque consolidées par une approche actuarielle adaptée. Dans le cadre d'une modélisation de l'effet géographique, le but est en général de compléter un modèle interne, construit à partir de données de portefeuille avant enrichissement de celles-ci avec des variables associées à l'emplacement du risque assuré.

Dans la modélisation des risques associés à la MRH, l'enrichissement de données à partir de variables géocodées a fait l'objet de nombres travaux de recherche avec des applications essentiellement tournées vers la création de zoniers pouvant améliorer la calibration de la prime pure.

L'objet du présent document est de formaliser un procédé d'application pour tester l'apport en qualité de discrimination du risque pour un portefeuille d'assureur disposant d'un modèle de fréquence pour la garantie incendie. L'élément différenciant de ce travail est la mise en valeur de la démarche en amont de la création du zonier pour le risque incendie avec l'identification d'un enrichissement pertinent pour ce risque dont les caractéristiques rendent en général sa modélisation complexe (volatile, queue de distribution longue et épaisse versus d'autres risques ce qui nécessite par exemple d'un écrêtage des sinistres graves).

Ces particularités rendent les incendies particulièrement « visibles » dans la sphère publique, et entreprendre d'isoler des traces de leur incidence à partir de web-scraping et requêtage d'APIs publiques semblait donc une voie pertinente à explorer. La relative accessibilité à un ensemble d'informations qui en suivi ex-post constitue une réponse de modèle permet en particulier de générer des variables explicatives d'une façon « informée ». Dans ce cadre, une attention particulière doit toutefois être portée au risque de sur-apprentissage et au risque d'utilisation d'information *a posteriori*, faussant le caractère prédictif du modèle.

Chapitre 1

Contexte de l'étude et présentation de la garantie incendie dans un contrat MRH

L'ensemble des notions présentées dans le premier chapitre sont des rappels sur l'assurance. En plus des différentes sources citées, ce chapitre est principalement construit à l'aide du cours d'introduction à l'assurance IARD d'Adrien Suru pour le Master d'Actuariat de Paris-Dauphine.

1.1 Définition de l'assurance

1.1.1 Inversion du cycle de production

En plus d'être souvent obligatoire, l'assurance est très influente dans les pays développés à économie de marché mais aussi dans les pays émergents. L'économie de marché est un terme représentant les systèmes économiques où la valeur des échanges est déterminée par l'offre et la demande. L'ensemble des primes enregistrées par les assureurs mondiaux dépassent les 5 000 milliards de dollars américains, avec 1500 milliards de primes émises par le marché européen, Willis Towers Watson (2019). En France le chiffre d'affaire des assureurs s'élève à 219 milliards d'euros, Fédération Française de l'assurance (2018). Pour pouvoir comprendre le fonctionnement d'un produit d'assurance, il faut premièrement rappeler la définition de l'assurance qui est : « un contrat par lequel l'assureur s'engage à indemniser l'assuré, moyennant une prime ou une cotisation, de certains risques ou sinistres éventuels » Larousse (2020). Le terme « éventuel » est important car l'assureur ne peut pas connaître avec certitude le nombre et le montant des sinistres de ses assurés, et doit se baser sur des lois statistiques et autres outils mathématiques.

L'inversion du cycle de production est la clef de la tarification en assurance : c'est le fait que le produit soit vendu avant que son coût final ne soit connu. La tarification est donc une décision ex-ante, c'est à dire que l'appréhension de la survenance du sinistre est faite avant qu'il se produise. L'ajustement de la prime après une ou plusieurs années d'historique est un acte ex-ante pour les primes futures, c'est une réaction face au constat de la situation.

Les notions d'a priori et d'a posteriori, habituellement utilisées en droit sont importantes pour la pérennité des compagnies d'assurance. Ce mécanisme est à l'opposé de toutes les autres industries où le coût de fabrication est connu et détermine le prix de vente.

1.1.2 Mutualisation des risques

Le travail d'un assureur consiste à traiter globalement tous les clients exposés aux mêmes aléas, c'est le principe de mutualisation des risques. Ce principe repose sur le fait que tous les clients sont indépendants entre eux, ce qui permet de diminuer la variance du sinistre moyen. Si X_i est le risque porté par l'assuré i, l'ensemble des risques est représenté par la variable réelle $(X_1, X_2, ..., X_n)$ et la variance de cette variable est définie par :

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x}), \text{ avec } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ le risque moyen.} \tag{1.1}$$

Par indépendance des X_i, la variance de la moyenne empirique du risque est donc :

$$V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\frac{1}{n^2}V(X_i) \ + 2\sum_{1\leq i\leq j\leq n}\frac{1}{n^2}Cov(X_i, X_j) = \sum_{i=1}^{n}\frac{1}{n^2}V(X_i) \tag{1.2}$$

La covariance est un terme positif, la mutualisation d'assurés indépendants réduit donc la variance du risque moyen, par rapport à une mutualisation entre des assurés avec des risques corrélés, en annulant la covariance, Charpentier et al. (2015).

Cette notion permet de mettre en commun toutes les primes des assurés pour couvrir les sinistres de seulement un petit nombre de sinistrés. C'est le principe de l'assurance par répartition : l'assureur redistribue seulement ses primes aux victimes d'un sinistre. L'assuré accepte de payer un montant forfaitaire qu'il soit sinistré ou non en échange de faire porter son risque à l'assureur. La valeur maximum d'une prime que l'assuré peut accepter de payer ou non dépend de son aversion au risque.

L'aversion au risque est un comportement récurrent en économie, De Palma (2008). Les investisseurs, préfèrent s'assurer un gain faible avec une grande probabilité plutôt qu'un gain important mais avec une probabilité faible. Si une personne a 100€, et qu'elle a le choix de les conserver ou bien de les investir sur un produit P pour 100€ pouvant rapporter 1000€ avec une probabilité de 15% ou bien de perdre la mise avec une probabilité de 85%, la réponse majoritaire est souvent de conserver la mise initiale alors que l'espérance est supérieure à la mise :

$$\mathbb{E}[P] = 0,85 \times 0 + 0,15 \times 1000 = 150 \geq 100 \tag{1.3}$$

Ce constat est décrit dans la théorie des fonctions d'utilité établie par John Von Neumann et Oskar Morgenstern. La décision de prendre un risque ou non dépend de l'utilité, c'est-à-dire la satisfaction ou le gain ressenti que peut en tirer l'investisseur. L'investisseur cherche à maximiser son utilité selon des critères qui lui sont propres. Dans l'exemple précédent : (1.3), si u est la fonction d'utilité alors si une personne refuse le produit P :

$$\mathbb{E}[P] = 0,85 \times u(0) + 0,15 \times u(1000) \leq u(100) \tag{1.4}$$

En souscrivant un contrat d'assurance, l'assuré est dans la situation de l'investisseur mais dans le cadre d'une perte certaine. Pour maximiser son utilité, il préfère s'assurer une petite perte annuelle en payant la prime d'assurance plutôt que de payer un sinistre de coût plus important avec une incertitude sur sa fréquence. Le montant qu'un assuré est prêt à payer pour s'enlever un risque dépend bien de son aversion à ce risque.

Pour qu'une société d'assurance soit solvable, il faut que la somme des primes payées par ses assurés soit supérieure à la somme des sinistres et des frais de l'entreprise, c'est la théorie de la ruine, Biard (2010). Le niveau de réserve d'une compagnie augmente avec les cotisations mais diminue ponctuellement lors du règlement des sinistres. L'évolution du niveau de réserve d'une entreprise d'assurance peut être représentée par le modèle de Cramer-Lundberg. Soit $R(t)$ le niveau de réserve alors :

$$R(t) = u + ct - \sum_{i=0}^{N(t)} X_i. \tag{1.5}$$

Avec

- u les réserves initiales et c le taux de cotisation demandé aux assurés, supposé ici constant.
- N(t) un processus de Poisson homogène de loi de Poisson d'intensité λ.
- X_i la variable aléatoire qui représente le coût du i-ème sinistre qui a lieu au temps t_i.

Le niveau de réserve peut être visualisé de manière simple avec la figure (1.1) suivante :

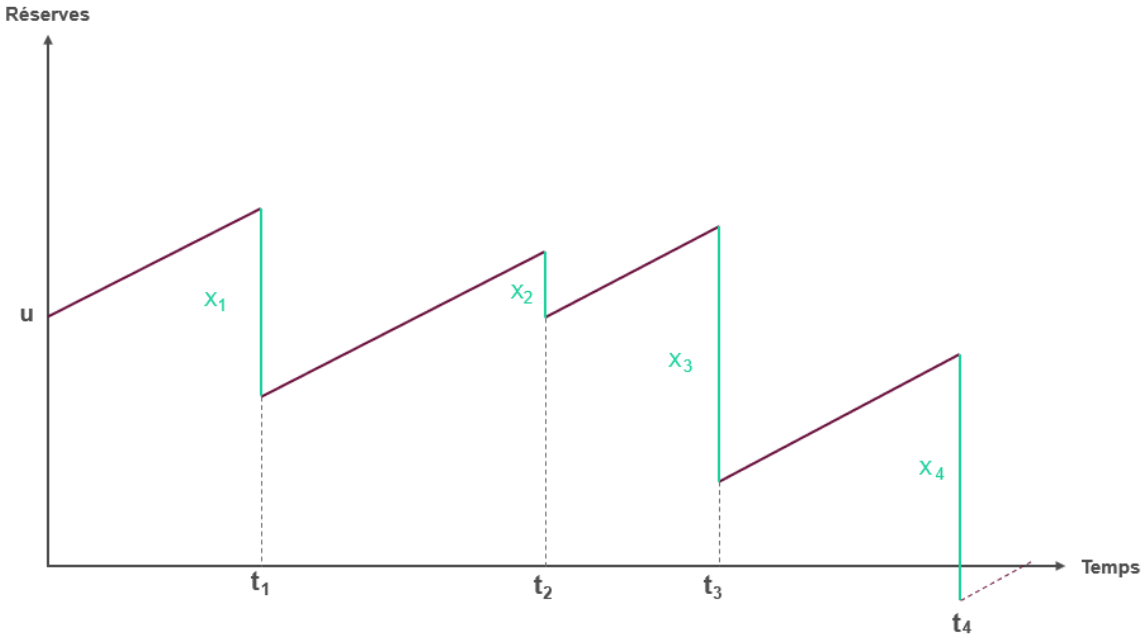


Figure 1.1: Théorie de la ruine de Cramer-Lundberg.

Dans cet exemple la ruine de l'assureur intervient lors du règlement du quatrième sinistre au temps t_4.

Le principe d'assurance par mutualisation est l'inverse de l'assurance par capitalisation où le montant final perçu par le bénéficiaire dépend principalement des primes versées par l'assuré et s'effectue sur des temps longs (plusieurs années).

1.1.3 L'assurance IARD

Les contrats d'assurance se divisent en deux grandes catégories : les assurances vies et l'assurance non-vie. L'assurance non-vie concerne les contrats sur les biens et leurs occupants et l'ensemble des produits couverts sont très variés. La dénomination anglaise est P&C (*Property and Casualty*) ou *General Insurance*.

L'assurance IARD offre aux particuliers comme aux entreprises un moyen de prévoyance et de protection de biens contre des risques et leurs éventuelles conséquences. L'article $\text{L}121-6$ du code des assurances définit simplement l'assurance dommage : "Toute personne ayant intérêt à la conservation d'une chose peut la faire assurer. Tout intérêt direct ou indirect à la non réalisation d'un risque peut faire l'objet d'une assurance". L'indemnité ne peut cependant pas excéder le montant du bien assuré, c'est le principe de non-enrichissement.

L'assurance non-vie est fortement régulée pour protéger l'assuré qui souhaite préserver du patrimoine financier. Pour construire un produit d'assurance il faut lister les garanties, les conditions d'exclusion et de prévention, définir les valeurs assurées et affecter le produit aux différentes branches d'agrément. En assurance IARD, il existe 18 branches, et pour pouvoir mettre sur le marché un produit, l'assureur doit obtenir de l'Autorité de Contrôle Prudentiel et Résolution (ACPR) un agrément pour chaque branche concernée, ACPR (2020) et Code des Assurances (2020b).

La France est le cinquième marché mondial de l'assurance IARD, derrière respectivement, les Etats-Unis, la Chine, le Japon et le Royaume-Uni, ABI (2019).

L'inversion du cycle de production en assurance demande une attention particulière aux différents principes de tarification.

1.2 Principes de tarification

La tarification en assurance et particulièrement en IARD repose, en Europe principalement, sur la séparation de la prime en plusieurs couches : la prime pure, la prime technique et la prime commerciale. Chaque couche prend en compte la précédente et améliore la tarification. Dans d'autres pays où la réglementation est différente, les méthodes le sont aussi. Le calcul est effectué pour chaque garantie du contrat, séparant ainsi chaque risque quand ils peuvent être supposés indépendants.

La prime pure est la partie du montant payée par le souscripteur du contrat qui correspond seulement à la couverture du risque. Elle est calculée à l'aide de diverses variables relatives à la nature et l'occurrence du risque. Ces variables sont souvent des informations sur les produits assurés ou bien leurs propriétaires : comme l'âge, la catégorie socio-professionnelle, le secteur d'activité d'une entreprise, la surface du logement ou la puissance fiscale d'un véhicule. Certaines variables ne peuvent

pas être prises en compte par les assureurs pour éviter les discriminations, comme le sexe pour les assureurs européens depuis la directive du 21 décembre 2012, Institut des Actuaires (2020) . Les principales méthodes de modélisation sont :

- Détermination de la fréquence et la sévérité.
- Définition d'un ratio de sinistres sur primes.
- Séparation des sinistres graves, des sinistres attritionnels.
- Utilisation de la théorie de la crédibilité.
- Création d'une nouvelle variable représentant le risque par zone géographique : zonier.

La prime pure est le montant à payer dans un monde idéal sans friction et sans recherche d'enrichissement de l'assureur.

La prime technique prend en compte les frais de l'assureur dont la réassurance et le coût du capital. La réassurance est l'opération qui permet à un assureur de faire couvrir tout ou partie des risques qu'il a accepté de couvrir à un autre assureur. La prime technique permet de consolider la prime pure et d'établir un coût de revient.

La prime commerciale comprend la prime technique plus une étude de la demande du marché pour déterminer la marge sur chaque contrat. Cette opération est possible grâce à une optimisation commerciale. L'élasticité-prix est le centre de cette étude, elle permet de comprendre la sensibilité d'un potentiel client au prix. La valeur du client tout au long de sa vie sera aussi analysée. En considérant les affaires nouvelles (taux de transformation) et les affaires en portefeuille (taux de résiliation), un modèle de demande en assurance peut être créé afin d'ajuster la prime commerciale. La prime commerciale permet de définir l'appétit du client au produit, son comportement face au prix et sa potentielle rétention . La prime est étudiée pour prendre en compte le risque de perte de valeur comme la perte d'affaire nouvelles sur les bons risques, en s'assurant de rester attractif.

Un montant maximal et minimal de la prime sont établis. Si pour une catégorie de client le produit est déjà le plus attractif du marché, il ne sert à rien de baisser la prime. La valeur maximale permet à l'assureur de ne pas être mal classé lors des études de marchés de tarifications, même si certains mauvais risques pourraient avoir leurs primes majorées. Une fois cet intervalle défini, les différents distributeurs du contrat peuvent accrocher des prospects avec des escomptes et autres gestes commerciaux sans prises de risques pour la solvabilité de l'assureur. La figure (1.2) résume le processus de tarification d'un produit d'assurance.

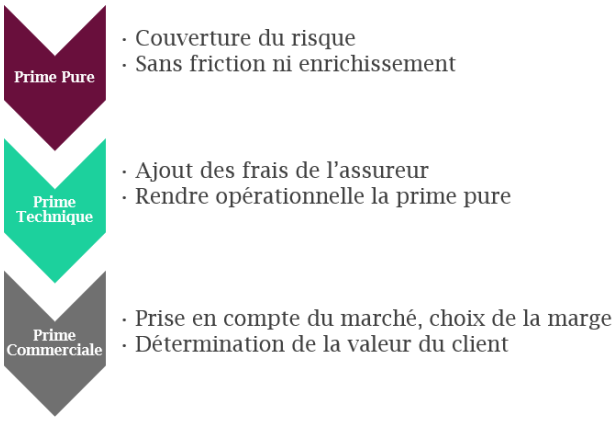


FIGURE 1.2: Schéma du processus de tarification en assurance.

La tarification repose principalement sur le portefeuille de l'assureur, qui par nature lui est personnel, mais elle peut également être influencée par sa stratégie commerciale ainsi que ses méthodes de souscription. Un assureur peut ainsi avoir tendance à refuser certains risques dans son portefeuille, sans pour autant avoir de certitude quant à leur éviction. Pour pouvoir améliorer la tarification, les assureurs peuvent compléter leurs modèles par des expériences extérieures. Ce mécanisme est courant lorsque, par exemple, un assureur accepte d'assurer une entreprise, mais qu'il n'a pas accès aux informations personnelles sur chaque employé de cette entreprise et doit utiliser son expérience qui est externe à la situation pour couvrir une expérience propre.

1.3 Modélisation du risque

Le calcul du montant de la prime pure dépend de la valeur totale des sinistres et de leur occurrence. Cette valeur peut être représentée par deux variables aléatoires indépendantes : la sévérité et la fréquence. L'objectif est de définir l'espérance et la variance des sinistres, WERNER et MODLIN (2010). Une seconde partie détaillera la modélisation du risque incendie dans la suite de ce chapitre (1.4.5).

Soit X la variable aléatoire représentant le risque alors :

$$X = \begin{cases} \sum_{k=1}^{N} B_k & \text{si } N > 0 \\ 0 & \text{si } N = 0. \end{cases} \tag{1.6}$$

La variable aléatoire discrète positive N est le nombre de sinistre sur une période. N représente la fréquence. B_k le montant du sinistre numéro k, cette variable aléatoire est continue positive. Les B_k représente la sévérité, elles sont supposées indépendantes et identiquement distribuées. La fréquence et la sévérité sont supposés indépendantes, le montant du sinistre n'est pas influencé par sa récurrence. La variable X est une loi composée de B et N, DENUIT et CHARPENTIER (2005).

Modèle sévérité

La variable sévérité représente le montant moyen de chaque sinistre. Dans le cas d'un historique de sinistre l'espérance de cette variable peut être estimée par :

$$\mathbb{E}[Sévérité] = \frac{Valeur\ totale\ des\ sinistres}{Nombre\ de\ sinistres}. \tag{1.7}$$

Modèle fréquence

La variable fréquence représente la probabilité de sinistre pour chaque contrat. Dans le cadre d'un historique de sinistre la variable peut être estimée par :

$$\mathbb{E}[Fréquence] = \frac{Nombre\ de\ sinistres}{Exposition}. \tag{1.8}$$

Un contrat n'est pas forcement souscrit au 1^{er} janvier. L'occurrence d'un sinistre est généralement étudiée par année et ne peut donc pas simplement dépendre du nombre de contrats souscrit et rompu dans l'année. Pour prendre en compte la période réelle concernée, une nouvelle variable est définie : l'exposition. L'exposition est la durée où le contrat est soumis au risque sur l'année étudiée par rapport à sa date de souscription, par exemple 0.5 correspond à un contrat souscrit depuis 6 mois. L'exposition dans la formule (1.8) est la somme des expositions de chaque contrat du portefeuille.

1.3.1 Segmentation des sinistres selon leur sévérité

En assurance, les sinistres peuvent être séparés en trois types différents selon leur sévérité :

- Les sinistres attritionnels. Le coût moyen d'un sinistre attritionnel est faible mais sa fréquence est élevée. Il est possible de modéliser ces sinistres avec des loi statistiques.
- Les sinistres graves. Les sinistres graves sont rares ou très volatiles et ont des coûts exceptionnels pouvant affecter la solvabilité d'un assureur. Le nombre limité de sinistres graves dans les historiques d'assureurs dus à leur fréquence rend leur modélisation plus difficile.
- Les sinistres liés à une catastrophe naturelle exceptionnelle. Les sinistres sont situés dans une même zone géographique et provoqués par un unique événement naturel sur une période donnée. Les montants des sinistres sont très variables mais leur fréquence est élevée dans la zone touchée.

1.3.2 Prime

Par hypothèse d'indépendance entre la fréquence et la sévérité, la prime pure peut s'écrire de deux façons :

- D'après l'équation (1.6) :
$$\mathbb{E}[X] = \mathbb{E}[N] \times \mathbb{E}[B]. \tag{1.9}$$
- D'après les équations (1.7) et (1.8) :
$$PP = \mathbb{E}[S] \times \mathbb{E}[F] = \frac{Valeur\ totale\ sinistres}{Exposition},\ \text{avec } S \text{ la sévérité.} \tag{1.10}$$

L'équation (1.9) est obtenue à l'aide de l'espérance conditionnelle et la propriété de distribution des B_k, Denuit et Charpentier (2005).

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\sum_{k=1}^{N} B_k\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{N} B_k|N\right]\right] \mathbb{E}\left[\mathbb{E}\sum_{k=1}^{N}[B_k|N]\right] = \mathbb{E}\left[\mathbb{E}\sum_{k=1}^{N}[B_k]\right] \mathbb{E}\left[\mathbb{E}\sum_{k=1}^{N}[B]\right] \\ &= \mathbb{E}[N] \times \mathbb{E}[B]\end{aligned} \tag{1.11}$$

L'utilisation de ce modèle illustre le problème de tarification en assurance lié à l'inversion du cycle de production : sur un nouveau produit aucun historique de fréquence ou de coût n'est connu. Pour un contrat classique qui possède plusieurs options et risques couverts, si ces derniers sont supposés indépendants, le calcul de la prime précédent peut être repris pour chaque risque. Par exemple, pour un contrat MRH qui couvre le risque incendie, le vol et les dégâts des eaux, la prime pure peut être décomposée de la manière suivante :

$$\mathbb{E}[S] = \mathbb{E}[S_{incendie}] + \mathbb{E}[S_{vol}] + \mathbb{E}[S_{dégâts\ des\ eaux}],\ \text{avec } S \text{ la sinistralité.}$$

Pour chaque garantie le calcul est séparé entre les sinistre graves et attritionels :

$$\mathbb{E}[S_{incendie}] = \mathbb{E}[S_{grave\ incendie}] + \mathbb{E}[S_{attritionel\ incendie}].$$

Et finalement, la modélisation est séparée en coût et fréquence

$$\mathbb{E}[S_{incendie}] = \underbrace{\mathbb{E}[N_{grave\ incendie}]}_{fréquence\ grave\ incendie} \times \underbrace{\mathbb{E}[B_{grave\ incendie}]}_{coût\ grave\ incendie} + \mathbb{E}[N_{attritionel\ incendie}] \times \mathbb{E}[B_{attritionel\ incendie}].$$

Le prime technique est alors déterminée avec cette prime et l'ajout des charges de l'entreprise .

Echantillonnage

Les calculs précédents reposent sur un portefeuille de taille suffisante mais il est impossible d'étudier un comportement sur la population totale. La vision d'un assureur est limitée à l'expérience de son portefeuille. Les estimateurs calculés pour définir le risque sont biaisés et leurs variances élevées. Pour pallier ce problème les bases de données utilisées doivent répondre à certains critères.

- La pertinence et la qualité doivent être vérifiées, il faut ainsi pouvoir contrôler si les informations données par les contrats sont justes. Par exemple, les questions posées à la souscription peuvent être orientées car un assuré aura tendance à minimiser son exposition à certains risques.
- La base doit être représentative de la population étudiée : les caractéristiques ainsi que les comportements doivent être semblables. Une taille minimum de l'échantillon ou de la base est donc requise, cette valeur peut être déterminée à l'aide de la loi des grands nombres en fonction de la précision voulue.

1.3.3 Provisions

La tarification d'une police d'assurance est importante pour être en mesure de couvrir les indemnisations des futurs sinistres. La gestion financière d'une entreprise se fait par années comptables, alors qu'un contrat peut porter sur plusieurs années et donc dépasser l'année comptable de souscription. Pour résoudre ce problème, les compagnies d'assurances sont obligées de provisionner, c'est-à-dire, lors de la souscription de réserver des fonds pour les sinistres qui pourraient potentiellement avoir lieu les années prochaines. Ce mécanisme permet de compenser le fait que la prime sera comptabilisée en tant que chiffre d'affaire sur une seule année comptable si celle-ci est payée entièrement à la signature du contrat (dans la cadre d'une prime mensuelle le mécanisme est identique, mais les provisions prennent en comptes les futurs paiements de l'assuré). La durée de vie d'un sinistre peut être très longue. C'est notamment le cas pour les incendies où les montants engagés sont très importants et nécessitent l'intervention d'experts. Le schéma suivant (1.3) est un exemple de la durée de vie d'un contrat sinistré.

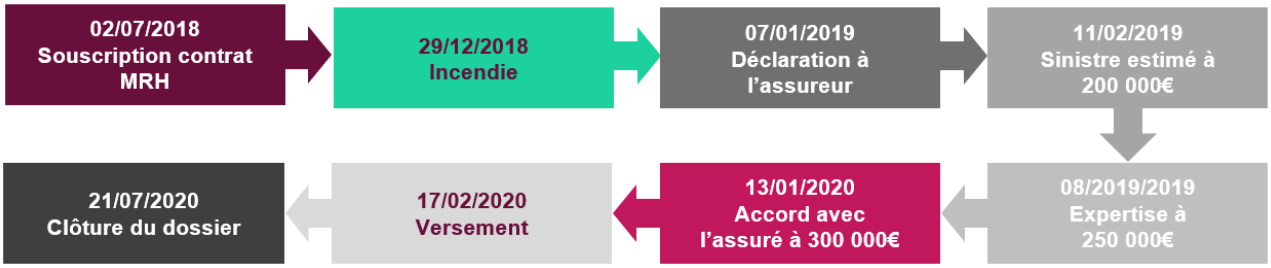


FIGURE 1.3: Schéma de la durée de vie d'un sinistre.

Ces cycles longs et aléatoires ont des impacts sur la gestion de la trésorerie d'un assureur. Ce dernier doit être capable d'estimer le montant à payer et le moment du versement. Les provisions sont des stocks qui sont amenés à évoluer. Les provisions principales en assurance IARD sont :

- Provision pour Prime Non Acquise (PPNA) : part de la prime au *prorata temporis* de l'exercice comptable de souscription.
- Provision Pour Risque en Cours (PREC) : compensation pour contrat sous tarifé.
- Provision pour Sinistre à Payer (PSAP) : montant qu'un assureur doit mettre en réserve pour indemniser tous les sinistres survenus durant la période de couverture passée. C'est la provision principale en assurance IARD.
- Incured But Not Reported (IBNR) : la provision pour les sinistres déjà survenus mais non communiqués à l'assureur.
  - Incured But Not enough Reported (IBNE) : couverture d'éventuelle variation du montant estimé tant que le dossier n'est pas clos.
  - Incured But Not Yet reported (IBNY) : Couverture des sinistres réalisés mais non encore déclarés à l'assureur.

La garantie incendie est une branche d'assurance longue, c'est-à-dire que le temps entre le sinistre et son remboursement intégral est important (de l'ordre de plusieurs années). Le versement du montant se fait rarement d'un seul coup, par exemple, en incendie les frais de relogement sont versés très

rapidement, puis après expertise le sinistre est remboursé ou non. Les années durant lesquelles les différents règlements ont lieu s'appellent les années de développement. La garantie vol, par exemple, est quant à elle une branche courte car le remboursement peut se faire rapidement.

1.4 Les contrats MRH

L'assurance multirisque habitation (MRH) est un des contrats possibles de l'assurance IARD. Elle permet aux particuliers de couvrir et protéger leurs logements (maison, appartement, grenier, cave), le contenu du logement (meubles et objets personnels) ainsi que la responsabilité civile des résidents qu'ils soient propriétaires ou locataires. Cette assurance est obligatoire pour une personne titulaire d'un bail locatif et permet de protéger le bien du propriétaire. Son absence peut conduire à une résiliation du contrat locatif. Un logement vacant peut être protégé avec l'assurance propriétaire non occupant (PNO). Cette assurance n'est pas obligatoire, sauf dans le cas d'une copropriété pour la garantie responsabilité civile : Article 58, LOI N 2014-366 (2014). Les contrats MRH ont plusieurs types de garanties dont les plus courantes sont :

- Responsabilité civile : couvre les dommages corporels ou matériels occasionnés accidentellement à des tiers. Elle s'applique au souscripteur et les personnes à sa charge.
- Dégâts des eaux : couvre les dommages liés à l'eau, c.-à-d. les fuites d'eaux, infiltrations, ruptures de canalisation chez l'assuré ainsi qu'aux éventuels biens limitrophes.
- Bris de glace : couvre les coûts de réparations des éléments de séparations avec l'extérieurs comme les portes ou fenêtres.
- Vol et vandalisme : couvre les pertes matérielles liées à une effraction, cependant les conditions de validités sont très variables selon les contrats (porte forcée, par exemple).
- Catastrophe naturelle : couvre les dommages d'un événement naturel quand celui-ci est déclaré catastrophe naturelle par le ministre de l'intérieur.
- Acte de terrorismes ou attentat.
- Catastrophe technologique : couvre les dégâts liés à un accident sur une installation technologique classée (sites dits « seveso ») en dehors de tout accident lié au risque atomique.
- Protection juridique : couvre les éventuels frais de justice entre l'assuré et un tier.
- Incendies et explosion : couvre les dégâts liés au feu ou à la fumée, ainsi que les dégâts éventuels liés à l'intervention des pompiers.
- Assistance : Service d'aide matérielle pour un assuré sinistré.

D'autres options ou des compléments des garanties déjà cités sont possibles (dommages électrique, assurance scolaire), pour pouvoir proposer la meilleure solution d'assurance pour chaque client selon son besoin, CODE DES ASSURANCES (2020c).

Les limites de garanties

Chaque assureur défini des limites de garanties par sinistres. Pour un sinistre donné, les montants d'indemnisation possibles sont plafonnés et il existe des critères pour que le sinistre soit pris en charge. Chaque assureur est libre de définir ses limites. Par exemple, AXA limitait en 2012 le montant de la garantie incendie à la valeur de reconstruction à neuf (y compris frais de démolition et de déblaiement), AXA (2012). Par principe de non-enrichissement, l'assureur ne rembourse jamais un montant supérieur à la valeur du bien sinistré. Si c'était le cas, un assuré pourrait être tenté d'augmenter sa prise de risque voire même de provoquer délibérément le sinistre, c'est l'aléa moral. Pour s'assurer du non enrichissement de l'assuré, des critères sont établis comme une décoté liée à la vétusté pour prendre en compte l'usure lié au temps du bien assuré. Une franchise peut être mise en place pour empêcher l'enrichissement de l'assuré.

1.4.1 Le marché de l'assurance MRH

Malgré le fait que la sinistralité climatique augmente significativement depuis plusieurs années, cf. l'étude Optimind (2019), le marché de l'assurance MRH en France reste un marché à fort potentiel et hautement stratégique. C'est un marché très disputé : les clients restent assez fidèles à leurs assureurs, il est donc compliqué d'augmenter sa clientèle. La progression du nombre de logements en France métropolitaine est constante depuis 35 ans. Avec une croissance de 1,1% par an, le parc immobilier français est passé de 24 millions en 1984 à plus de 35 millions de logements en 2019. Le marché de l'assurance MRH augmente donc aussi parallèlement au nombre de logements. En 2019, en moyenne 82% des logements étaient des résidences principales, 10% des résidences secondaires et 8% des logements vacants. Ces proportions sont quant à elles constantes depuis 35 ans. Depuis la mise en place de la loi Hamon en 2014, Code des Assurances (2020a), permettant aux assurés de changer à tout moment leurs contrats d'assurance après un an d'ancienneté, les bancassureurs en ont profité pour essayer d'augmenter leurs parts de marché, lors de la souscription d'un crédit ils proposent leurs produits MRH. Dans le même temps, les assureurs MRH en ligne se sont développés permettant de grosses réductions de coûts avec moins de personnels et de locaux. Cependant, la souscription MRH se fait encore majoritairement, à 61%, dans le réseau d'agence d'assurance historique, 17% des souscriptions sur internet fixe et 7% sur mobile, INSEE (2019a) et Argus de l'assurance (2019).

En 2018, le nombre de contrats MRH était de 41,9 millions (les logements locatifs peuvent être assurés par les occupants et les propriétaires), il a ainsi progressé de 2,0% par rapport à 2017, avec une augmentation de 1,6% pour les contrats occupants et 4,2% pour les contrats non occupants. L'ensemble des cotisations est estimé à 10,5 milliards d'euros. Les primes ont augmenté en moyenne de 1,6%, ce qui est inférieur à l'augmentation de 2,3% de l'indice de la fédération du bâtiment, qui permet d'estimer le coût de construction. La fréquence des sinistres des contrats MRH a augmenté de 7,9% par rapport à l'année précédente. L'ensemble de ces chiffres proviennent de FFA (2020).

Le marché français est détenu en grande majorité par une petite dizaine de leaders :

TABLE 1.1: Classement 2020 des assureurs MRH (chiffre 2019, en M €).

RANG	ASSUREUR	CA 2019	CA 2018	VARIATION 2019/2018	PART MRH DANS LE CA	NOMBRE DE CONTRATS EN 2019	VARIATION 2019/2018
1	Covéa	1861,0	1808,0	2,9%	12,1%	8 084 992	+39 480
2	Groupama	1165,3	1140,5	2,2%	NC	3 808 463	-359
3	Crédit agricole Assurance	1 129,4	1 039,6	8,6%	NC	4 396 973	+219 247
4	Axa	1 010,0	1 022,0	-1,2%	3,7%	3 494 483	-90 267
5	Macif	904,4	877,8	3,0%	27,3%	4 303 074	+45 951
6	Groupe Maif	864,0	838,0	3,1%	30,0%	3 305 599	+7 309
7	Allianz	630,2	627,4	0,4%	NC	2 228 955	-5 904
8	Groupe des Assurances du Crédit mutuel	610,0	575,0	6,1%	5,0%	2 632 783	+108 559
9	Natixis Assurances	504,1	473,1	6,6%	32,0%	2 198 070	+112 156
10	Matmut	451,6	434,4	4,0%	20,0%	2 250 858	+45 882
11	Generali	362,0	351,0	3,1%	13,0%	1 253 233	+22 914
12	Aviva	195,7	194,4	0,7%	NC	604 171	+771
13	La Banque postale Assurances IARD	166,1	164,2	1,2%	46,2%	748 945	+39 387
14	Société générale Assurances	164,0	156,0	5,1%	24,4%	729 547	+29 120
15	Suravenir Assurances	103,7	98,5	5,3%	25,4%	489 479	+14 987
16	BNP Paribas Cardif	100,0	101,0	-1,0%	0,8%	496 256	+38 085
17	Mutuelle de Poitiers Assurances	89,1	85,4	4,3%	22,6%	400 950	+6 591
18	Thélem Assurances	80,5	76,0	5,9%	22,4%	298 831	+3 625
19	Groupe MACSF	69,7	66,4	5,0%	10,8%	293 663	+6 637

Le tableau ci-dessus représente les 19 plus gros assureurs MRH en 2020 avec les chiffres de 2019. Source : ARGUS DE L'ASSURANCE (2020a). CA : Chiffre d'affaire, NC : non communiqué.

1.4.2 L'effet du COVID-19 sur le marché MRH

Certaines branches de l'assurances IARD ont profité aux assureurs avec une baisse de la sinistralité lors du confinement lié au COVID-19. Les contrats d'assurance automobile principalement, avec une réduction de 75% du kilométrage moyen, Argus de L'assurance (2020c). Le marché de l'assurance MRH a lui subi une augmentation des sinistres incendies jusqu'à 40% et une augmentation des accidents domestiques par rapport aux tendances habituelles des années précédentes. Pendant le confinement, les utilisations plus intensives d'appareils électriques liées à l'allongement des temps passés à domicile ont déclenché plus d'incendies, Crawford (2019). Mais la baisse du nombre de cambriolages et de dégâts des eaux va permettre de compenser le coût des sinistres d'incendies et générer un gain sur l'ensemble des contrats MRH sur la période du confinement, Argus de L'assurance (2020b).

1.4.3 Le risque Incendie

C'est le risque incendie qui sera au cœur de ce mémoire. C'est un risque couvert pour tous les contrats de MRH souscrits. Les polices d'assurance couvrent les dommages causés par un incendie, une explosion ou implosion. Les dommages induits liés à la foudre, la fumée et l'intervention des secours sont inclus dans la couverture. En plus des risques légaux précédents, d'autres risques peuvent être souscrits en option comme le risque électrique, les dommages causés par un excès de chaleur (brûlure sur un plan de travail en bois) ou encore les dommages indirects (perte de loyer, frais de relogement, propriétaire non occupant). La définition est donnée par l'article Ł122 − 1 du Code des Assurances (2020d).

Chaque contrat contient un plafond de couverture ainsi qu'une franchise (montant du sinistre restant à la charge de l'assuré). Pour que le contrat soit valable, différentes conditions sont obligatoires. Depuis 2010, dans tous les logements, un détecteur de fumée doit être présent et en état de marche. Son entretien est à la charge de l'occupant ou le propriétaire selon l'usage du logement LOI n 2010-238 (2010). Son objectif est de sauver des vies, si ce dernier est absent dans un logement non locatif assuré contre les incendies, l'assureur peut appliquer une franchise mais ne peut pas refuser d'indemniser le sinistre. Si un conduit de fumée est présent dans le logement (cheminée, chaudière à gaz, poêle), deux ramonages par an par un professionnel agréé sont obligatoires au risque de voir le montant maximum de la garantie fortement diminué. Pour un contrat incluant un terrain, un débroussaillage autours des zones d'habitation (variable selon les communes) est obligatoire, Code des Assurances (2020d)

1.4.4 Les causes d'incendies

Les causes d'un incendie peuvent être diverses et souvent imprévisibles. Ce qui est couramment appelé un incendie est le résultat d'une réaction chimique : la combustion. Pour qu'une combustion se produise, il faut que trois éléments se retrouvent conjugués : le combustible, le comburant et une énergie d'activation. Le combustible est le composé chimique susceptible de brûler et peut se présenter sous les trois états de la matière : solide, liquide ou gazeux. Des combustibles potentiels sont par exemple le plastique, le bois, l'essence ou bien le propane. L'oxygène présent dans l'air est le comburant le plus commun. L'énergie d'activation est une énergie thermique permettant d'atteindre un point d'inflammation local, elle est souvent simplement appelée chaleur. L'absence ou la disparition d'un de ses trois éléments empêche ou met fin à la combustion. C'est la notion du triangle du feu. Pour obtenir une combustion avec des flammes, il faut en plus des radicaux libres : ce sont des atomes ou

molécules instables. C'est le tétraèdre du feu. Sans les radicaux libres la combustion est par exemple un feu de braises de charbon, INRS (2020).

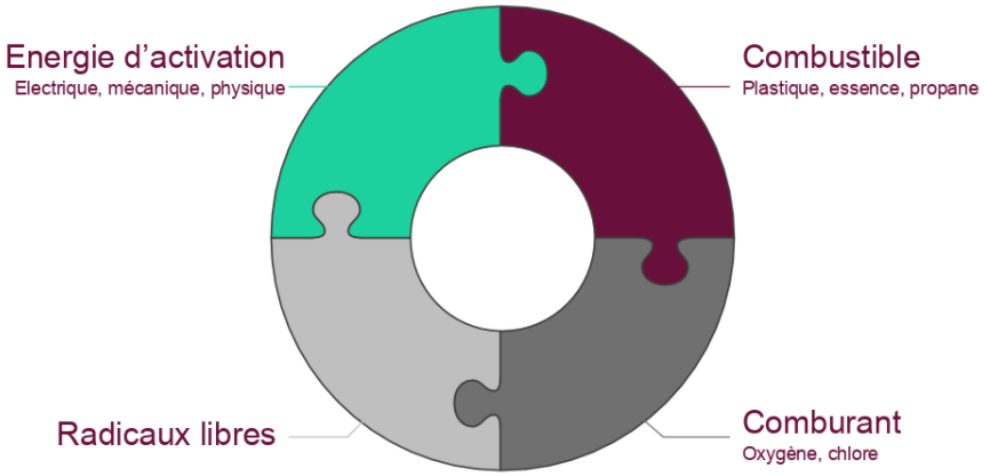


FIGURE 1.4: Représentation du tétraèdre du feu.

Les sources de chaleurs et donc d'inflammations sont nombreuses dans un logement ou local. Une installation électrique vétuste, plus aux normes, peut générer des échauffements et des étincelles. Les surfaces chaudes de cuisine ou les appareils de chauffages sont également par leur usage des sources potentielles. L'inattention et la négligence sont aussi des initiateurs fréquents, un mégot de cigarette peut ainsi atteindre plus de 700 degrés. La malveillance est enfin à l'origine de nombreux incendies.

Le dégagement de chaleur lors d'une combustion va permettre la propagation du feu. La chaleur se diffuse selon quatre modes différents : la conduction, la convection, le rayonnement et le transport. La conduction est la propagation au sein du même milieu, de proche en proche. C'est le cas quand un combustible est en contact avec des flammes ou de la chaleur et qu'il s'embrase progressivement. La convection est l'ensemble des mouvements d'un fluide. Les gaz dégagés par la combustion peuvent transmettre la chaleur. Le rayonnement est une onde électromagnétique émis par le foyer de l'incendie qui propage de la chaleur aux comburants à proximité. Le transport est la projection d'une source de chaleur. Une explosion ou des vents peuvent être des sources de transport en projetant une source de chaleur, ENASIS (2020).

En France, il y a plus de 260 000 sinistres d'incendies qui sont déclarés aux assureurs chaque année, en moyenne un toutes les deux minutes. Ce nombre a doublé en 20 ans. 25% des incendies sont dus à une défaillance de l'installation électrique du logement. La réactivité pour prévenir les secours est un facteur déterminant du montant du préjudice : en 3 minutes une pièce en proie aux flammes peut atteindre plus de 600 degrés. Le risque n'est pas seulement dans la présence de flammes mais aussi le dégagement de monoxyde de carbone lors de la combustion. En moyenne, une centaine de décès et plus de 10 000 hospitalisations sont dénombrés chaque année. Les incendies domestiques représentent la deuxième cause de mortalité chez les enfants de moins de 5 ans. Plus de 70% des sinistres mortels se produisent la nuit alors que 70% des incendies ont lieu de jour. Un français sur trois est victime d'un incendie une fois dans sa vie, MINISTÈRE DE L'INTÉRIEUR (2020) et PLANETOSCOPE (2020).

1.4.5 Modélisation du risque incendie

La garantie incendie est une garantie MRH complexe à modéliser : la fréquence de sinistre est faible avec moins de 10 contrats impactés pour 1000, mais les coûts des dégâts sont souvent très élevés. Le risque vol peut être plus simple à modéliser car les objets assurés peuvent être identifiés et le montant maximum de remboursement plafonné. Le risque incendie, avec les dégâts des eaux et le vol sont les plus grandes sources de frais MRH pour l'assureur. Le coût moyen d'un incendie en 2018 est estimé à 7 420 € pour une fréquence de 0,55%, Fédération Française de l'assurance (2019). Le risque incendie active la garantie incendie dommages sur les biens assurés mais peut également activer la garantie responsabilité civile. De fait, la responsabilité civile peut atteindre un montant plus élevé que la valeur du bien assuré. Par exemple, un sinistre sur une cabane de bois qui cause des dégâts sur un bien voisin. La garantie d'une habitation pour le risque incendie est calibré avec deux facteurs clefs : la valeur du bien et la valeur du contenu assuré. Les contenus assurés ont souvent des limites complexes avec une distinction si les objets sont considérés dans leur ensemble ou séparés (par exemple, une limite pour l'objet le plus cher, une limite pour l'ensemble, une limite pour une catégorie d'objet).

La modélisation du coût des sinistres d'incendie est assez spécifique. L'étude commence par l'analyse de la variable réponse, comme pour les autres risques de MRH, ici : le coût des incendies. L'adéquation à une loi est assez souvent compliquée. La modélisation peut se faire avec ou sans les sinistres graves. Le choix du seuil à partir duquel les sinistres sont considérés comme graves est une problématique à lui seul et peut être le sujet d'un mémoire, Naji (2016). Dans un cas simple, ce seuil peut être choisi à l'aide d'outils graphiques comme un QQ-plot. Une franchise doit être déterminée pour le seuil de petite perte. Une franchise trop faible ou inexistante peut générer beaucoup de charges de faibles valeurs. Prises indépendamment, celles-ci seraient négligeables, mais rapportées à l'échelle d'un portefeuille complet, ce risque devient significatif. La franchise peut être déterminée à l'aide d'un modèle avec une distribution permettant de capturer une sur-représentation des observations à valeurs faibles.

Le coût du risque incendie subit une inflation avec les années. En effet, la valeur d'un logement ou des biens qu'il contient augmente. Un exemple est le développement d'appareils électriques onéreux dans les logements, augmentant le montant de chaque sinistre. Le niveau de finition et la qualité des matériaux utilisés s'est aussi amélioré, augmentant la valeur des biens assurés. De plus, la tendance actuelle dans la construction ou rénovation de logement est de cacher le plus possible les différents fils ou tuyaux réduisant les possibilités d'entretien. La construction de logement neuf étant très encadrée par différentes normes, l'âge du logement est un facteur important de la valeur du bien. Par exemple, l'augmentation des normes liées à une installation électrique peut augmenter son coût : augmentation du nombre d'interrupteurs différentiels ou de disjoncteurs divisionnaires. L'inflation des coûts est un phénomène à traiter en particulier en considérant l'écrêtage réalisé précédemment sur les coûts. Le nombre de sinistrés est aussi en augmentation avec le temps, ce qui participe à l'inflation quand les coûts de construction suivent l'inflation nationale. La modélisation de l'inflation du risque incendie dépend du temps, mais des études pré et post modélisation sont nécessaires pour s'assurer de la stabilité du modèle, Actuaris (2017).

L'indice de la fédération française du coût de la construction, ICC FFB, est un indicateur du prix de revient d'un immeuble moyen à Paris émis de manière trimestrielle depuis le 1^{er} janvier 1941. L'indice prend en compte tous les coûts de construction comme les matériaux, la main d'oeuvre, les taxes ou frais. La valeur du terrain n'entre pas en compte dans le calcul. Cet indice est principalement utilisé par les assureurs pour revaloriser les montants des primes chaque année. L'ICC est passé de 574,8 au 1^{er} janvier 2000, à 988,2 au dernier trimestre de 2018, FFB (2020). Le prix de construction a donc augmenté de 3,05% par an en moyenne sur les dix-huit dernières années. L'augmentation moyenne de l'inflation en France sur cette même période est de seulement 1,4%, INSEE (2020d).

Un deuxième indicateur du coût de la construction est réalisé par l'INSEE depuis 1953. Cet indicateur est spécialisé dans les immeubles à usage d'habitation et sert généralement à la révision des baux commerciaux ou professionnel. Cet indice était de 1071 en janvier 2000 et de 1703 au dernier trimestre 2018, INSEE (2020b). Son augmentation moyenne sur la période est de 2,6%. Les coûts de sinistres augmentent plus vite que l'inflation.

L'étude de la distribution de la fréquence du nombre de sinistres en incendie est assez classique avec une adéquation à une loi malgré le faible nombre de sinistres. Différents facteurs de risques sont étudiés pour affiner les modèles de tarification comme la densité de population, le niveau de richesse de la zone géographique, l'ancienneté ou les normes en place. Idéalement, le facteur humain est utilisé à l'échelle individuelle : c'est-à-dire que seul le comportement de l'assuré et son historique sont pris en compte dans les modèles. Les comportements des populations d'une zone géographique donnée sont d'un ordre secondaire.

Nouvelle approche de la modélisation

L'objectif de ce mémoire est d'essayer à l'aide de données publiques non structurées, c.-à-d. sans format particulier, d'avoir une approche nationale sur la modélisation de la fréquence du risque incendie en MRH. L'idée est de chercher à évaluer cette fréquence sur l'ensemble du territoire métropolitain (hors Corse) en restant indépendant du portefeuille d'un assureur particulier. Le risque incendie ne dépend pas seulement de l'assuré mais de son environnement. L'assureur ne connaît pas exactement le contenu ou l'état de chaque bien assuré. L'ajout de données externes peut permettre d'améliorer la compréhension de ce risque. La création d'une base de données, contenant une composante de variable réponse (le nombre d'incendies) ainsi que différentes variables descriptives expliquant les incendies, constituera l'élément différenciant de ce mémoire. Pour cela différentes techniques originellement utilisées en science des données seront appliquées dans un contexte actuariel. La finalité n'est pas de remplacer les modélisations déjà en place, mais plutôt de les compléter avec une vision extérieure. Cependant, le risque pour chaque commune pourrait être intégralement expliqué par le portefeuille d'assureur. Un modèle correct, construit uniquement sur la base d'informations externes, peut avoir un gain marginal selon la qualité du modèle interne d'assureur.

Chapitre 2

Utilisation de données publiques

Chaque année le nombre de données échangées ou stockées est en augmentation et parallèlement à cette croissance, les techniques d'attaques informatiques se font toujours plus efficaces. Cette corrélation créée pour le détenteur de données un enjeu de taille : leur protection.

Le volume de données actuellement stocké est déjà conséquent avec environ 33 zettaoctet de données soit 33×10^{12} Go de données. Mais ce nombre devrait être multiplié au moins par 5 dans les années à venir. La France a une croissance de données plus importante que la moyenne des autres pays avec une génération de données de 701% depuis 2016 pour seulement 569% dans le monde sur la même période. Mais cette croissance n'empêche pas les pertes de données qui sont elles aussi en augmentation (augmentation de 50% en 3 ans du nombre d'entreprises ayant subi des pertes de données). Plus d'une entreprise sur deux ne pensent pas avoir de système de protection suffisant pour prévenir une attaque sur leurs données.

L'ensemble des données publiques disponibles sur internet ne possèdent pas de format prédéfini. Ces données non structurées représentent un défis pour leur collecte, traitement et analyse. La récupération de données est un enjeu commercial important depuis l'essor d'internet. Ces dernières permettent de comprendre le comportement d'un client dans un monde devenu toujours plus concurrentiel grâce à des canaux de distributions dématérialisés. Les assureurs se sont également intéressés à la possibilité d'étudier le comportement de leurs assurés pour améliorer leurs modèles de tarification ou bien d'automatiser la détection de fraudes par exemple. Cette démarche est récente : en 1995 moins de 1% de la population mondiale utilisait internet. Ces données sont essentiellement à caractère personnel, et afin d'assurer une protection et une transparence aux consommateurs, des règlementations ont été mises en place pour structurer et uniformiser le cadre de collecte, de stockage et d'emploi de ces données par les entreprises concernées, Journal du Net (2019) et Le Big Data (2018).

2.1 La réglementation

2.1.1 Données personnelles

Au sens de la loi, une donnée personnelle est une information qui permet de manière directe ou indirecte, d'identifier une personne physique. Une personne peut être identifiée de manière directe : par son nom, ou de façon indirecte : par un numéro, une plaque d'immatriculation ou même par un enregistrement de voix, CNIL (2020b).

Certaines informations sont qualifiées de "sensibles". Elles concernent les origines raciales ou ethniques, l'état de santé, les opinions et croyances personnelles, l'appartenance à un syndicat, l'orientation sexuelle. Cependant, toutes les informations concernant une entreprise ne sont pas considérées comme des données personnelles. Ainsi, les adresses, numéros de téléphone ou encore différentes adresses mails ne sont pas soumises à la même protection, CNIL (2020c).

Table 2.1: Exemple de données personelles.

Santé	Données Biométriques	Etat civil	Situation financière	Comportement	Opinions	Biographie
Maladie	**Données génétiques**	Nom, Prénom	Carte crédit	Préférences	**Opinions politiques, religieuses**	Diplôme
Handicap	**Données biométriques**	Sexe	Domiciliation bancaire	Cookies		Profession
Taux invalidité	**ADN**	Date de naissance	Données patrimoniales	Géolocalisation	**Appartenance syndicale**	**Passif pénal**
Etat physique		Taille	financière	Relations sociales		**Orientation sexuelle**

Un exemple de différentes données personnelles, avec en gras les données dites "sensibles".

Ainsi une base de données ne comprenant pas de nom ou prénom mais des informations précises comme : goût et habitudes d'achat, une localisation, un âge, est considérée comme contenant des données personnelles car il est possible d'identifier par recoupement une personne grâce aux différentes informations.

2.1.2 Règlement Général sur la Protection des Données

Le Règlement Général sur la Protection des Données ou RGPD est un règlement promulgué par l'Union Européenne. L'objectif du RGPD est de définir un texte de référence en matière de gestion et de protection des données personnelles et de faire appliquer des règles identiques aux différents acteurs étatiques ou privés. Une uniformisation globale du marché Européen permet de concurrencer les deux géants d'internet que sont les Etats-Unis et la Chine. Cela permet de regagner en souveraineté sur internet et développer des produits et services qui sont orientés autour du droit des personnes, CNIL (2020a).

Le RGPD est la dernière règlementation mise en place en matière de protection des données personnelles. Il fait suite à diverses réglementations présentées chronologiquement dans le schéma suivant (2.1).

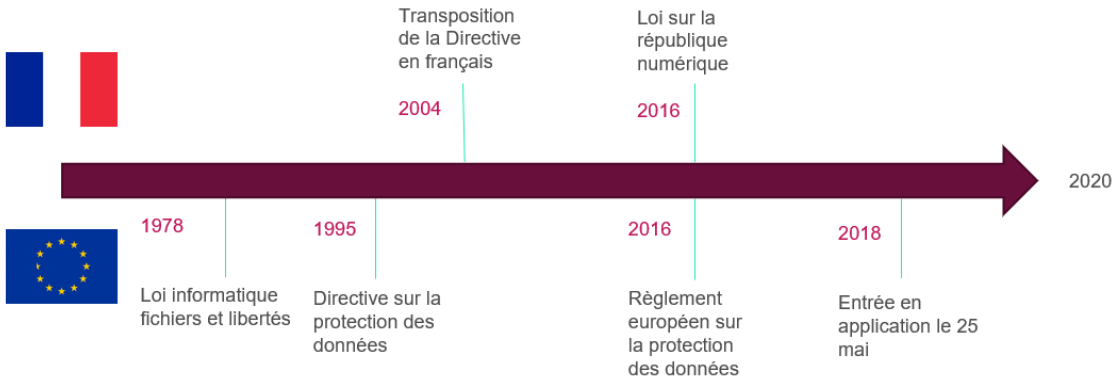


FIGURE 2.1: Evolution des réglementations sur les données en France et en Europe.

Application

Les entités, indépendamment de leur taille et de leur localisation peuvent être potentiellement concernées par le RGPD du moment qu'ils soient installés sur le territoire de l'Union Européenne ou bien que les données appartiennent à des résidents européens. Par exemple, une entreprise allemande qui a des clients situés au Moyen-Orient est obligée de respecter le RGPD. De même, une entreprise américaine commerçant en ligne avec une version française de son site est obligée de respecter le RGPD pour les clients français. Les sous-traitants qui utilisent des données personnelles pour le compte de leurs contractants sont aussi soumis au RGPD. La figure (2.2) présente les conditions d'application de la réglementation.

Les compagnies d'assurance sont concernées par le RGPD du fait de la profondeur des données dont elles disposent sur leurs clients. Elles peuvent donc être sanctionnées en cas de non-respect et de négligence de la réglementation. Différents risques sont liés à ces sanctions :

- Risque financier : Le montant de l'amende peut atteindre jusqu'à 2 à 4% du chiffre d'affaire, en proportion de l'importance des failles. Un dédommagement des assurés pour une atteinte à la protection de leurs données peut être possible.
- Risque opérationnel : L'activité d'assurance peut être arrêtée jusqu'à la mise aux normes, et atteindre le niveau de protection attendu.
- Risque d'image : Un scandale médiatique peut ternir la réputation d'un assureur pendant plusieurs années

En 2019, la société Active Assurance, un courtier spécialisé dans les contrats automobiles, a été condamné à une amende de 180 000 € pour ne pas avoir assez protégé les données personnelles de ses clients. Les données des clients étaient disponibles depuis le site de l'assureur. CNIL (2019).

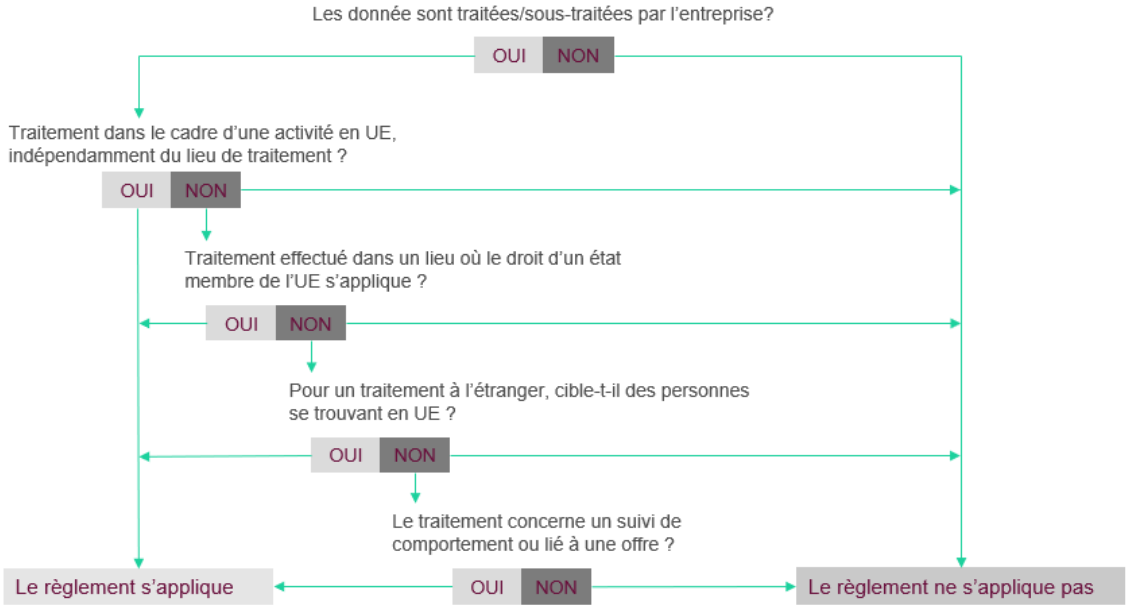


FIGURE 2.2: Graphique résumant l'application ou non du RGPD.

Traitement des données

Dans la suite du mémoire, le traitement de données représente l'ensemble des opérations effectuées sur des données personnelles. Enregistrer, organiser, trier ou croiser avec d'autres données par exemple. Mais cela ne concerne pas seulement les bases de données, un fichier ou un tableur mais aussi toutes les données liées à la mise en place de différents services comme une vidéo surveillance, un paiement par carte bancaire ou bien d'une application pour smartphone peuvent être considérées comme du traitement de données. En complément à l'écrasante majorité de données disponibles sur les supports informatiques, tous documents physiques (formulaires par exemple) organisés selon un critère peuvent aussi être considérés comme un traitement de données.

Le responsable des traitements est l'organisme, public ou privé qui a commandité les actions.

Un sous-traitant qui travaille sur des données personnelles n'est pas soumis à toutes les obligations du RGPD, seulement les suivantes :

- Obligations spécifiques en matière de sécurité.
- Obligation de conseil : informer les responsables de certains traitements à effectuer pour maintenir le niveau de conformité attendue.
- Obligation de tenir un registre et de définir un responsable de la protection de données.

2.1.3 Principes Fondateurs

Le périmètre du RGPD est très large et couvre un grand nombre de problématiques qu'il est possible de regrouper en quatre objectifs principaux.

- Renforcement des droits des personnes : les droits d'accès, de modification ou d'oubli, protection des mineurs, demande de consentement, portabilité.
- Principe de responsabilité : La responsabilité de l'entreprise est engagée en cas de faute et elle doit définir une liste d'obligations pour permettre le respect du règlement.
- Notification de la perte de données : Prévenir les autorités de contrôle du pays concerné (en France : la CNIL).
- Renforcement du pouvoir de contrôle et de sanction : La CNIL peut réaliser des contrôles et sanctionner plus sévèrement.

Renforcement du droit des Personnes

Le RGPD oblige qu'une série d'informations soit fournie à la personne qui renseigne des données personnelles. Ainsi, l'identité et les coordonnées du responsable des traitements, la finalité des données, la durée de conservation, la possibilité du droit de modification et d'oubli doivent être détaillées. Pour un assureur, cela se traduit par l'établissement de politiques internes sur la gestion des données. Les moyens de communication ont dû être réformés et les clauses des contrats modifiées pour spécifier les informations obligatoires.

Le profilage est défini par l'article 4 du RGPD. Il consiste à analyser les données d'une personne afin de prédire son comportement. Ces prévisions peuvent aboutir à des prédictions inexactes et se traduire par un refus de différents services, perpétuer des stéréotypes et enfermer des personnes dans leurs choix. Le profilage est couramment utilisé pour réaliser des décisions automatisées, c'est-à-dire qu'un choix est décidé à l'aide des données personnelles d'un individu sans qu'un humain intervienne. Ces décisions sont très courantes et peuvent affecter la personne concernée. Par exemple, un crédit peut être refusé après qu'un algorithme ait déterminé que les critères renseignés n'étaient pas suffisants.

Des règles structurent les décisions automatisées (Article 22). Ainsi, les individus ont le droit de ne pas faire l'objet d'une décision entièrement automatisée. Des obligations de transparence sont nécessaires pour pouvoir informer la personne concernée. Finalement, un droit à une intervention humaine leur est aussi autorisé.

Principe de responsabilité

Un registre ou RoPA (*Register of Processing Activities*) est un document contenant toutes les caractéristiques des données traitées par l'organisme comme leurs finalités, la catégorie des destinataires ou encore les délais de conservations. Ce document est à remettre à l'autorité de contrôle et remplace les déclarations qui étaient faites auparavant. Les sous-traitants doivent établir un registre détaillant les catégories de traitements effectués pour tous les clients.

Les entreprises de moins de 250 employés ne sont pas obligées de réaliser ce document si les données ne sont pas sensibles et que leurs traitements sont occasionnels.

Ce registre est un outil de pilotage grâce à sa vision globale des opérations effectuées. Pour un assureur cela représente une illustration complète des processus utilisés, leurs nombres doivent être cohérents avec l'activité. Ce registre peut être mis à jour de manière automatisée. D'abord réalisés sur des fichier texte ou tableur, les registres s'uniformisent avec le développement d'outils spécifiques à cet usage.

La maîtrise du cycle de vie des données est obligatoire pour assurer leur protection. Les données ne peuvent pas être conservées éternellement. Une durée de conservation doit être établie dans une politique de gestion des données. La loi peut disposer des durées de conservations, par exemple trois ans pour les données de prospection commerciale. Quand la loi est silencieuse, c'est le responsable du traitement des données qui définit le temps de conservation.

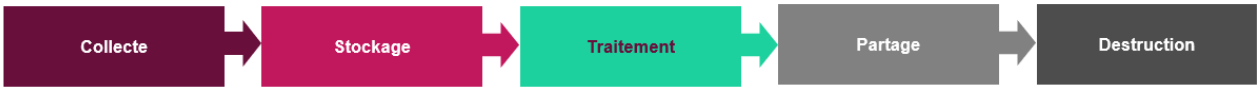


FIGURE 2.3: Cycle de vie des données.

2.1.4 Les données dans le monde de l'assurance

Les données d'assurances par leur précision et sensibilité sont directement impactées par le RGPD. Depuis l'application de la directive européenne Solvabilité II, la qualité des données est associée à un meilleur contrôle de ces dernières. La qualité est axée selon 4 chantiers complémentaires :

- Gouvernance : Mise en place d'une gestion de la qualité des données.
- Documentation : Informer le régulateur du travail de gestion mit en place dans l'entreprise.
- Mesure de la qualité : Liste de critères sur la pertinence, précision et exhaustivité des données.
- Amélioration de la qualité : Mise en place d'un cercle vertueux pour rechercher à toujours accroître la qualité.

Ces quatre points sont surveillés par le régulateur qui s'assure de leur mise en place et maintien dans le temps.

La qualité des données utilisées en assurance est contrôlée par différents indicateurs. L'exhaustivité permet d'assurer que les données contiennent assez d'information pour pouvoir comprendre le risque et déterminer des tendances. L'exactitude de la donnée est le deuxième indicateur : il ne doit pas y avoir d'erreur importante dans la base. Les données qui viennent de périodes de temps différentes doivent être concaténées et mises en cohérence entre elles avec dans des formats d'enregistrements semblables. Le dernier critère est l'adéquation des données à l'utilisation voulue. La nature et le volume des données doivent fournir des estimations sans biais important sur les risques et comportements

étudiés avec les techniques statistiques et actuarielles utilisées. La démarche entière : la collecte, le traitement et les finalités, doit être transparente, justifiée et documentée.

Lors de la conférence de l'ACPR du 21 juin 2019 lors de la présentation de l'actualité de la supervision en assurance les trois points suivants ont été présenté résumant la situation. Source : ACPR (2019).

- L'amélioration de la qualité des données doit encore être envisagée comme un processus d'amélioration permanente, exercice après exercice.
- Au-delà des contrôles bloquants embarqués par la taxonomie, la validité, la cohérence et la pertinence des données mérite d'être envisagée au niveau de l'input (donnée), de l'agrégat (template), de la collecte (annuel, trimestriel, etc.) et globalement (vérifications croisées des reportings quantitatifs et narratifs).
- L'option de ne pas reporter tel ou tel input ou agrégat de données et le choix d'utiliser telle ou telle simplification doivent, en général, être explicitement justifié (le plus souvent dans le RSR).

Les données utilisées par les assureurs peuvent être séparées en deux parties : les données internes et les données externes.

Données internes

Les données internes sont propres à chaque compagnie d'assurance. Elles proviennent de leurs historiques de souscription et de risque. Ces données sont souvent sous-exploitées à causes de formats différents, le coût d'harmonisation peut être très élevé mais significatif dans la compréhension d'un portefeuille de polices d'assurance. Le développement de nouvelles méthodes de souscription souvent dématérialisées permet d'augmenter le nombre de données récoltées. Les données doivent souvent subir plusieurs traitements avant d'être exploitées comme des anonymisations par exemple. Elles sont réparties en plusieurs catégories comme le montre la table suivante avec différents exemples de données pour chaque catégorie.

Table 2.2: Répartition d'exemples de données internes.

Données personnelles	Données contractuelles	Données sinistres	Données d'entreprise
Age	Nombre de polices	Prestations et réserves	Portefeuilles d'actifs
Statut matrimonial	Garanties	Coûts moyens	Structures de frais
Variables de segmentation	Sommes assurées	Tables d'expériences	Taux d'imposition

Données externes

Les données externes sont toutes les données qu'un assureur peut récupérer en dehors de son historique personnel. Ces données servent à enrichir les données internes, permettant d'avoir une vision plus large du contexte dans lequel l'assureur se trouve comme l'environnement macro-économique. Certaines données externes sont obligatoires à la réalisation de documents comptables comme la courbe de taux sans risque ou le taux moyen d'emprunt de l'Etat. Cependant ces données doivent respecter les législations en vigueur lors de leur récupération (RGPD, droit d'accès, licences) et être actualisées régulièrement.

Table 2.3: Répartition d'exemples de données externes.

Données financières	Données concurrentielles	Données de référence
Courbe taux sans risque	Tarifs	Table de mortalité
Inflation	Taux revalorisation annuels	Barèmes indemnisation
Probabilités de défaut	Nouvelles garanties	Etudes et analyses

Utilisation de la donnée externe

Les données en accès libre ou *open data* sont en augmentation avec l'essor d'internet. Un classement des pays fournissant le plus d'open data gouvernemental classe la France comme un des pays les plus riches en nombre et volumes de bases de données publiques, Open knowledge Foundation (2020). Ces données sont disponibles sur les sites comme l'INSEE, data gouv ou encore open data Paris. Au niveau mondial, des données sont disponibles sur différents site comme Open Street Map, Office for National Statistics, data.nasa.gov.

Les données externes offrent un volume et une variété à forte valeur commerciale mais le coût de traitement peut être important. En effet, la qualité est variable d'une base à l'autre et les méthodes utilisées sont aussi diverses rendant leur utilisation de manière automatique compliquée. La réforme réglementaire Solvabilité II encadre la qualité des données utilisées pour des calculs de provisions techniques.

- Les données doivent être appropriées : elles doivent être représentatives du portefeuille et des cash-flows futurs.
- Les données doivent être complètes : les niveaux de précision et d'historique doivent être suffisants pour comprendre toutes les tendances de risques.
- Enfin les données doivent être exactes : elles doivent être sans erreur et cohérentes entre les périodes de collectes.

Les données doivent être anonymisées, c'est-à-dire qu'il n'est pas possible d'identifier la personne à l'origine de chaque donnée. Pour mesurer l'efficacité d'un traitement d'anonymisation trois critères sont utilisés :

- l'individualisation : est-il possible de distinguer un individu ?
- la corrélation : y a-t-il un lien entre différentes données disjointes pour une personne ?
- l'inférence : est-il possible de de tirer des informations sur un individu de la base ?

Si au moins un des trois critères est réalisable alors la base ne peut pas être considérée comme anonyme.

Variables interdites

Depuis plusieurs années des améliorations ont été apportées dans le monde de l'assurance pour éviter les discriminations. Par exemple, depuis le 21 décembre 2012 la tarification d'un produit d'assurance ne peut plus dépendre du sexe du bénéficiaire. Pour éviter les décisions liées à l'état de santé

d'un assuré la mise en place du s'Assurer et Emprunter avec un Risque Aggravé de Santé (AERAS) permet une étude individuelle de la situation.

Cependant, certains critères subjectifs sont toujours présents et forment une discrimination légale.

- ***L'intuitu personae*** autorise un assureur à utiliser son intuition pour juger s'il souhaite assurer ou non une personne. Ainsi un assureur peut rompre un contrat déjà souscrit en cas de doutes sur la bonne foi d'un client ayant subi un nombre important de sinistres.
- **Le défaut d'aléa** est une méthode légale de refus de souscription d'une police d'assurance. Si un assureur considère qu'un risque n'est pas aléatoire il est en droit de ne pas le protéger.

2.2 Les outils de récupération de données

Pour récupérer des données externes divers outils existent. Cette section présente l'ensemble des outils utilisés dans ce mémoire.

2.2.1 Application Programming Interface

Définition

Les *Application Programming Interface* ou API sont des interfaces de programmation qui permettent à deux logiciels d'échanger des données. L'avantage de ses applications est de faciliter la mise en œuvre des échanges en cachant les détails techniques des transferts. Une API est donc une façade donnant accès à diverses fonctionnalités. Les avantages à utiliser une API sont nombreux : les données sont directement accessibles avec une requête R ou Python, ce qui est facilement implémentable dans un code, ou des modélisations déjà existantes. Les données peuvent être extraites sous le même format, simplifiant ainsi leur manipulation après extraction. Il y a cependant des limites aux API : la qualité des données dépend entièrement de l'hébergeur qui les possède. De plus, la disponibilité des API, bien que très variée, est variable selon le type d'informations et de données recherchées. Le nombres de requêtes peut aussi être limité : par exemple 20 requêtes par seconde et adresse IP.

De nombreuses API sont disponibles gratuitement. Les états, en plus de mettre à disposition différentes bases de données, proposent des API régulièrement misent à jour. En France ces API sont disponibles sur le site api.gouv.fr . Différentes API sont disponibles comme la base ***Sirene*** qui permet d'obtenir des informations d'immatriculations des entreprises. Pour ce mémoire, plusieurs API gouvernementales sont utilisées afin de retrouver des codes INSEE, ainsi qu'une API privée.

La majorité des API sont cependant payantes. En effet, le marché de la donnée a explosé au cours des dernières années. Depuis le 11 Juin 2018 Google a par exemple rendu payante son API qui permet de récupérer de nombreuses informations présentes dans le service gratuit Google Maps. Lettria, est une startup qui propose au travers de ses API payantes différentes offres de NLP (2.2.3) spécialisées dans la langue française, comme la conformité de textes et commentaires aux nouvelles normes du RGPD.

Les applications d'utilisation d'API pour l'assurance sont très nombreuses. Permettant d'améliorer différentes modélisations, de simplifier les traitements de données ou encore la souscription, les assureurs utilisent déjà cette technologie. Allianz, Axa ou encore La Maif par exemple ont développé leurs propres API. Ces solutions font évoluer les canaux de distributions des produits d'assurance mais aussi les outils informatiques. La mise en place d'une API a cependant un coût de développement important et ne doit pas se faire au détriment de la sécurité.

Les API sont principalement majoritaires chez les assurtechs (start-up d'assurance basée sur les nouvelles technologies), qui n'ont pas révolutionné le marché depuis leur mise en place mais ont défini de nouvelles normes tout en gagnant des parts de marché aux assureurs historiques, Argus de l'assurance (2020).

2.2.2 Web Scraping

Avec plus de quatre milliards d'utilisateurs, soit plus de la moitié de la population mondiale, internet contient des dizaines de milliards de pages différentes et offre un grand nombre de possibilités et d'opportunités. Le *web scraping* regroupe les diverses techniques permettant de récupérer des informations sur des pages internet comme des listes d'articles, adresses ou numéro par exemple. Les données récupérées peuvent être réutilisées dans des contextes différents que ceux d'origine et analysées pour être valorisées. Le *web scraping* est différent du *web crawling*. Le *web crawling* est une technique de navigation qui permet de se déplacer automatiquement entre différents hyperliens, Macapinlac (2018).

Pourquoi avoir recours aux différentes techniques de *scraping* ? Les données sont le premier maillon dans le processus de modélisation de différents phénomènes et comportements. Ainsi les données scrapées peuvent avoir vocation à devenir une source autonome de données pour une étude particulière, ou à compléter et enrichir une base de données déjà existante. L'enrichissement peut soit servir à consolider des valeurs manquantes ou bien de pouvoir présenter une ou plusieurs variables complémentaires pertinentes. En assurance, le *web scraping* peut également servir à la veille concurrentielle, c.-à-d. surveiller automatiquement les prix appliqués par la concurrence au travers de leurs offres de devis en ligne par exemple.

Réglementation

Les processus de scraping ne sont pas directement répréhensibles. Il n'existe pas de définition stricte dans la réglementation ainsi, si l'extraction est faite dans le respect des conditions d'utilisations de la page internet alors le scraping est légal. Mais la réutilisation de données avec ou sans modification, peut être condamnable. En France, trois différents droits interviennent dans le cadre de données provenant de scraping.

- **Droit Pénal** : L'article 323-3 du Code pénal dispose que ≪ le fait d'introduire frauduleusement des données dans un système de traitement automatisé, d'extraire, de détenir, de reproduire, de transmettre, de supprimer ou de modifier frauduleusement les données qu'il contient est puni de cinq ans d'emprisonnement et de 150 000 d'amende. ≫.

- **Droit de la Concurrence** : L'utilisation de techniques de *web scraping* s'assimilerait à un acte de parasitisme ou de concurrence déloyale. En effet, une personne morale ou physique, peut à l'aide de *web scraping* tirer profit des efforts du propriétaire du site web ainsi que son savoir-faire et ce sans rien dépenser. La concurrence déloyale est définie par l'article L121-1 du code de la consommation.

- **Droit de la propriété intellectuelle** :

  – L'article L. 341-1 du Code de la propriété intellectuelle dispose que « Le producteur d'une base de données, entendu comme la personne qui prend l'initiative et le risque des investissements correspondants, bénéficie d'une protection du contenu de la base lorsque la constitution, la vérification ou la présentation de celui-ci atteste d'un investissement financier, matériel ou humain substantiel. Cette protection est indépendante et s'exerce sans préjudice de celles résultant du droit d'auteur ou d'un autre droit sur la base de données ou un de ses éléments constitutifs ».

  – L'article L. 342-1 du Code de la propriété intellectuelle dispose que « Le producteur de base de données a le droit d'interdire l'extraction par transfert permanent ou temporaire de la totalité ou d'une partie qualitativement ou quantitativement substantielle du contenu d'une base de données sur un autre support, par tout moyen et sous toute forme que ce soit. ».

La récupération de données avec le *web scraping* peut donc enfreindre principalement les deux articles du code de la concurrence en portant atteinte aux producteurs des différentes bases de données. Le droit pénal n'intervient que dans le cas d'une attaque volontaire d'un site. Pour éviter que le scraping soit interprété comme une attaque, il existe différentes bonnes pratiques à prendre en compte avant de scraper des données sur internet. Les conditions d'utilisation du site web doivent être lues, et le détenteur des données doit être clairement identifié. Il est préférable de s'identifier dans les requêtes envoyées au site, et respecter les limites des hébergeurs.

Un grand nombre de pages internet fournissent une page `robots.txt` qui renseigne sur ce qui est permit de scraper ou de crawler. L'exemple suivant présente la page `robots.txt` du journal Les Echos, disponible à l'adresse https://www.lesechos.fr/robots.txt :

```
User-agent: *
Disallow: /internal
Disallow: /recherche
Disallow: /connexion?redirect
Disallow: /signup?redirect

Sitemap: https://sitemap.lesechos.fr/sitemap_index.xml
Sitemap: https://www.lesechos.fr/sitemap_news.xml.
```

La page se lit de la manière suivante : Les Echos interdit à tous les *crawler* (`User-agent: *`) d'acceder aux URLs comportants les termes : `internal`, `recherche`, `connexion?redirect` et `signup?redirect`.

En 2015, l'entreprise 3Taps Inc a été condamnée à verser 1 000 000 € à l'entreprise Craiglist Inc. 3Taps pour avoir créé une interface permettant de visionner les annonces de Craiglist après avoir récupéré leurs données. Malgré une lettre de cessation et le blocage de l'adresse IP de l'entreprise, 3Taps avait continué de scraper le site de Craiglist avec différentes adresses IP.

Des packages `R` sont disponibles pour réaliser différentes opérations de requêtes necessaire au *web scraping*. Le package `Rvest` est celui utilisé dans ce mémoire, R Core Team (2019) et Wickham (2020).

Le *web scraping* possède des limites. Chaque site possède une architecture et une construction différentes et les appellations des différents objets et balises sont variables. Ainsi, la portabilité d'un algorithme de *scraping* entre sites, c'est à dire sa capacité à fonctionner pour deux sites différents, est faible. Une automatisation à grande échelle n'est pas possible dans le cadre de la rédaction d'un mémoire d'actuariat. Certains sites refusent le *crawling* et le *scraping* pour protéger leurs informations. De plus, le *scraping* est très sensible aux évolutions de réglementations. Le service de collecte d'informations Google News qui rassemble des articles de journaux à l'aide de méthodes de *scraping*, a décidé d'arrêter son service en Espagne suite à un changement de législation, Monde (2014).

2.2.3 Traitement automatique des langues

Le traitement automatique des langues TAL ou NLP pour *Natural Language Processing* est un mélange de plusieurs domaines comme la linguistique, l'informatique, l'intelligence artificielle. Son objectif est le développement d'outils permettant l'interaction entre un ordinateur et un humain. Son origine remonte aux années 50 avec la publication de Turing (1950) sur la capacité d'une machine, ici un ordinateur, à penser. Un fichier informatique contenant du texte est interprété par un ordinateur pour pouvoir le lire ou le modifier, le NLP développe des méthodes pour qu'un ordinateur comprenne le sens de ce qui est écrit dans un fichier. Les applications du NLP sont très courantes aujourd'hui avec par exemple : traducteurs de texte, correcteurs orthographique, analyses d'opinion, prédiction du prochain mot lors de la rédaction d'un message sur un smartphone. Originellement le NLP a été développé pour la langue anglaise, mais depuis les méthodes ont été implémentées pour d'autres langues comme le français.

Le TAL peut être décomposé en plusieurs étapes. Les deux principales sont la segmentation et la labellisation, Lettria (2020).

- La segmentation permet de découper un texte (une chaîne de caractères) en un ensemble de mot distincts. Ce processus permet de régulariser un langage naturel qui par nature ne l'est pas et évolue en permanence. La segmentation commence par l'analyse des espaces entre chaque mot. Mais en français plusieurs problèmes existent comme les tirets, les points qui servent d'abréviation (M. abréviation de monsieur) ou encore les locutions nominales comme arc-en-ciel. Chaque élément est appelé token.
- La labellisation consiste à identifier pour chaque token l'ensemble grammatical auquel il appartient (nom commun, adjectif, interjection, ...). Cette étape est aussi appelée *pos-tagger*. La connaissance seule des règles grammaticales du français n'est pas suffisante. Différentes techniques d'intelligence artificielle sont utilisées pour déterminer la nature d'un token avec l'étude des tokens qui l'entourent.

Des entreprises, comme Lettria, proposent des API payantes permettant pour une phrase donnée de récupérer les différents tokens, ainsi que leur classification à l'aide de requêtes. Certains de ces services utilisent le NLP pour pouvoir analyser des commentaires laissés sur internet et ainsi déterminer le taux de satisfaction ou le niveau de réputation en ligne.

2.3 Création d'une base de données

2.3.1 Récupération directe de données en lien avec les incendies

L'objectif de ce mémoire est d'essayer de créer un modèle permettant de comprendre la fréquence incendie en France à l'aide de différentes données externes. La recherche d'informations commence par la consultation de bases de données publiques. Deux bases officielles émises par le service départemental d'incendie et de secours (SDIS) contiennent des données sur le nombre d'incendies. La première, appelée base nationale des interventions de pompiers, détaille pour chaque département de France métropolitaine, le nombre d'interventions selon leur nature (accident de circulation, incendie de local industriel, ...), 64 catégories d'interventions sont répertoriées. L'exemple (A.8) présente une utilisation de cette base sur la variable feux d'habitations et bureaux. La deuxième base, appelée base Essonne, compte pour chaque commune de l'Essonne le nombre d'intervention par semaine de 2010 à 2017 selon leurs natures. La nature de l'intervention peut appartenir à cinq classes différentes : secours d'urgence à personne, incendie naturel, incendie urbain, accident de la route ou autre. L'exemple (A.9) présente une utilisation de cette base avec la probabilité d'intervention pour incendie par logement pour chaque commune du département.

Pour étudier un phénomène et pouvoir définir des zones de risque, une échelle de taille doit être déterminée. En France, plusieurs possibilités de tailles existent, elles sont résumées dans la figure (2.4).

Le code INSEE est un identifiant à cinq chiffres équivalent au code postal mais unique à chaque commune. Il n'existe que 6300 codes postaux pour les 36 000 communes françaises, DATA.GOUV.FR (2020). Le découpage IRIS est un découpage des communes de plus de 10 000 habitants. Ce découpage est la définition la plus proche de quartier. L'ordre de grandeur de la population est d'approximativement 2000 habitants par IRIS. En France il y a environ 49 000 IRIS. X et Y sont les coordonnées GPS, c'est la localisation la plus précise possible pour situer un événement, INSEE (2020c).

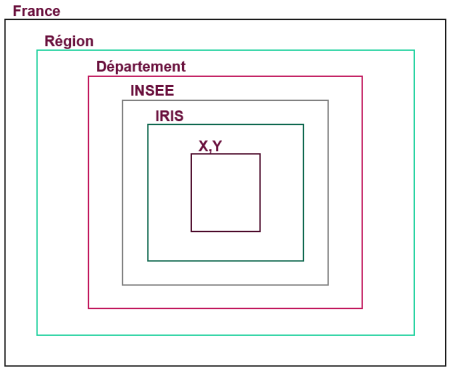


FIGURE 2.4: Les différentes échelles de zonier possible en France.

Bases officielles des services de secours

La base nationale des interventions de pompier est à la maille départementale. Cette maille n'est pas assez fine pour être utilisée directement dans des modèles mais sa couverture de l'ensemble du territoire et la précision des informations sont utiles. La maille de la base Essonne est la commune, ce qui est déjà plus précis pour étudier le risque incendie.

Cependant n'avoir qu'un seul département en France en fait une source d'information fragile. Cette base peut être utilisée pour vérifier des modèles réalisés à la maille commune sur l'ensemble de la France.

Base des feux de fôret

Les feux de forêts sont une source de risque pour les habitations. Avec 16,9 millions d'hectares, la forêt occupe 31% du territoire français. C'est l'occupation des sols majoritaire après l'agriculture, IGN (2017). La base de données publiques, sur les incendies de forêts en France, conçue et développée par l'institut national de l'information géographique et forestière, comptabilise l'ensemble des données disponibles sur les feux de forêts depuis 2006. La maille de cette base est l'INSEE, avec pour chaque incendie : la date, l'heure et la surface brulée selon sa nature (forêt, boisée, artificialisée), IGN (2020a). Il existe aussi la base ***Prométhée***, qui reprend les mêmes incendies mais ne couvre que 15 départements du sud de la France.

Variables explicatives

Connaître le nombre d'incendie permet de définir la variable réponse du problème, mais pour mettre en place différents modèles, des variables expliquant les incendies doivent être trouvées. Il y a donc une distinction entre la variable réponse et les variables explicatives, mais les deux bases sont construites simultanément. Chaque nouvelle information sera ajoutée aux différentes bases en utilisant le code INSEE.

Dans la recherche de variables explicatives des informations géographiques de plusieurs bases concernant toutes les communes de France comme la population, la surface ou l'altitude sont récupérées. Les incendies étant liés à l'urbanisme de la ville étudiée, la base de recensement des logements issue de l'INSEE, qui détaille la nature des biens de chaque commune, est ajoutée à l'ensemble des variables explicatives. La base de logement est très complète avec 238 variables qui précisent le nombre de logements selon différents critères, par exemple le nombre de maisons construites avant l'année 1990. Cette base possède des valeurs pour différentes années : 2007, 2012 et 2017. Une seule version de chaque variable sera conservée, la plus adaptée selon la période de modélisation.

La base d'artificialisation, mesure la transformation des sols naturels et la consommation de l'espace par des aménagements réalisés par l'homme. Cette base est produite par le ***Cerema***. Permettant de connaître la surface urbanisée de chaque commune, la base possède un historique d'évolution de 2009 à 2018. Les différentes variables représentent la typologie de la commune, les différents flux d'urbanisation ainsi que les variations de population, de ménage, d'emplois. Cette base publique est à la maille INSEE.

La base de données demande de valeur foncière

L'ensemble des bases précédentes ne comportent pas de données personnelles. Toutes les informations référencées concernent seulement l'agrégat "communes". Le RGPD ne s'applique donc pas sur ces bases. Depuis le décret n 2018-1350 du 28 décembre 2018 Légifrance (2018), les valeurs foncières déclarées à l'occasion des mutations immobilières sont disponibles avec un historique de cinq ans. Mais ces données sont à caractère personnel et sensible car elles détaillent : la date et l'heure de l'acte, son

prix, l'adresse du bien, les différentes références cadastrales, et des éléments descriptifs du bien comme la surface, le type de local ou le nombre de pièces. Les conditions générales d'utilisations (DATAGOUV (2020)) définissent les conditions de réutilisation : le RGPD doit être respecté et la réidentification des personnes est interdit.

L'agrégation des valeurs moyennes des différentes variables pour chaque commune respecte bien les conditions d'utilisation. L'ensemble des cinq ans d'historique représente plus de 14 millions de transactions, présentent sur 33 000 communes de France. Le calcul du prix moyen des biens vendus sur chaque commune est une des réutilisations principales de la base de demande de valeur foncière pour des outils d'estimation de prix d'un bien. Cette base est une seconde source sur les biens présents sur chaque commune. La valeur médiane de chaque variable est imputée aux communes qui n'ont pas de transaction enregistrée. La valeur médiane est préférée car elle est plus représentative que la moyenne face aux valeurs extrêmes.

Assemblage

Les différentes bases à l'échelle nationale sont toutes agrégées avec la variable INSEE, cependant cette opération n'est pas exempte de problème. La fusion de deux communes crée des modifications d'INSEE et potentiellement des valeurs manquantes dans la base finale. Par exemple en 2018, 627 communes ont connu des fusions, INSEE (2020a). Ces cas sont à traiter individuellement et cette étape est difficilement automatisable. L'objectif final est d'obtenir une base comportant l'ensemble des codes INSEE de France métropolitaine.

Pour les bases avec une maille moins fine que l'INSEE, les valeurs sont répétées pour toutes les communes du même sous-ensemble. Par exemple, pour la base nationale d'incendie, chaque commune d'un département a le même nombre d'interventions.

2.3.2 Récupération de données en lien avec les incendies à l'aide d'une API

La NASA met à disposition sur le site worldview.earthdata.nasa.gov de nombreuses informations météorologiques extraites de ses différents satellites. Parmi toutes ces informations le FIRM *Fire Information for Ressource Management System* est une carte interactive où les incendies détectés par satellites sont représentés. Les historiques français d'incendie sont téléchargeables avec des coordonnées géographiques et une date de détection.

Les données disponibles proviennent de plusieurs satellites de technologies différentes. Les plus anciennes sources, à partir 2000, sont issus des deux satellites Terra et Aqua respectivement lancés en 1999 et 2002. Ils utilisent tous les deux la technologie MODIS (*Moderate Resolution Imaging Spectroradiomer*) où chaque pixel représente une surface de 1 km^2, MODIS (2020). Un nouvel outil de détection composé d'un ensemble de radiomètres pour imageurs opérant dans la bande infrarouge et visible appelé VIIRS permet d'obtenir une seconde source. Le VIIRS est présent sur une nouvelle génération de satellite qui sont sur la même orbite : le Suomi NPP et le NOAA-20 lancés respectivement en 2011 et 2017. Un pixel du VIIRS fait 140 625 m^2, EARTHDATA (2020).

L'historique le plus important commence en 2011, avec 116472 détections en France et provient des satellites équipés avec le VIIRS. L'historique le plus ancien commence en 2000 mais ne répertorie que 44 172 détections à cause d'une technologie moins précise. Les données sont vérifiées avec une source de mesure indépendante : le capteur japonais ASTER (*Advanced Spaceborne Thermal Emission and Reflection Radiometer*) présent dans le satellite Terra. L'utilisation d'un deuxième outil de mesure permet d'éviter les faux positifs, Csiszar et al. (2006).

L'API gouvernementale geo.api.gouv.fr permet de récupérer des données à la maille INSEE. Cette API permet en donnant deux coordonnées (longitude et latitude) d'obtenir l'INSEE de la commune où le point se trouve. Cette méthode est longue à exécuter face au grand nombre de données, ici les deux historiques de satellite utilisés représentent plus de 150 000 requêtes. De plus, des détections successives pour le même incendie existe avec une variation infime des coordonnées. Pour régler ce problème, deux signaux consécutifs adressant un même code INSEE sont considérés comme correspondant au même incendie. Cette hypothèse forte réduit le nombre d'incendies d'environ 25%. Les deux historiques disponibles sont utilisés, en supprimant les doublons lorsque les périodes de mesures se superposent (à partir de 2011). Le nombre d'incendies différents est finalement de 90 005.

L'étude des données montre quelques erreurs manifestes. Après agrégation du nombre d'incendie détectés par code INSEE, certaines communes ont plus de 10 000 incendies en 20 ans sur leur territoire. Par exemple l'INSEE 13039, qui correspond à Fos-sur Mer, comptabilise 12458 incendies. Cette ville du bassin méditerranéen est une des plus industrialisée de France avec des raffineries, des hauts fourneaux et incinérateurs. Ces sources de fumées et de flammes sont identifiées par les satellites comme des incendies.

Pour nettoyer ces données aberrantes de la base, un écrêtement est réalisé. Le nombre d'incendies est compté pour chaque commune, puis trié par ordre croissant. Toutes les communes de la base agrégée dont le nombre d'incendies est supérieur au quantile à 99% sont identifiées avec leur code INSEE. Finalement, tous les incendies qui ont eu lieu sur une des communes identifiées sont supprimés de l'historique de détection. Pour corriger les vrais incendies supprimés, lors de la modélisation, des processus de lissages spatiales seront mis en place pour prendre en compte la valeurs du risque des communes voisines. Le nombre de communes retirées est de 528 ce qui donne une fois le nettoyage effectué, plus de 36 000 incendies différents. L'hypothèse prise est de considérer que le risque incendie d'une habitation n'est pas liée à la présence de site seveso sur la commune.

Les données issues des satellites de la NASA possèdent des limites. Les algorithmes de détection et de surveillance de feu ont été conçus pour des équipements et capteurs non prévus initialement à cet effet. Un incendie n'est pas détecté s'il commence et finit entre deux révolutions du satellite. Et si un incendie est trop froid ou trop petit (moins de 50 m^2) alors il peut ne pas être détecté. Les radiomètres sont aussi sensibles à la météo et la présence de nuage, Giglio et al. (2018). La distinction entre incendie d'espace urbanisé et espace naturel n'est pas possible, même si un incendie de végétation a plus de chance d'être détecté car il se détache plus du fond.

2.3.3 Récupération de données en lien avec les incendies à l'aide de web scraping

Le web scraping est une autre manière d'obtenir le nombre d'incendie à la maille INSEE sur l'ensemble de la France. Les sites d'informations, en particulier les articles dans la section fait divers sont en effet une source importante de données sur les incendies. La mise en place d'un algorithme de web scraping permet, en étudiant la nature de la page internet, de récupérer les informations précises. Ainsi, les titres de tous les articles d'une page et les dates sont récupérés et quand ils sont disponibles les chapeaux. Le chapeau est un résumé concis de l'article disponible sous le titre.

La récupération et le traitement de tous les articles complet serait trop longue, la navigation entre les différents URLs est compliquée : la structure est très variable entre deux URLs, et les traitements de TAL n'en serait que plus complexe. Une solution possible est l'utilisation de `Javascript` qui permet de simuler une navigation sur des pages internet.

Sur l'ensemble des sites scrapés, les conditions générales d'utilisations ainsi que des temps minimums entre deux requêtes ont été respectés. La contrainte la plus importante a été un temps de 10 seconde d'attente. Pour ne pas surcharger les serveurs, le temps minimum entre chaque requête a été fixé à une seconde. Le choix des différents journaux retenus s'est fait après un parcours le plus exhaustif possible des sites d'actualités françaises pour prendre tous ceux qui avaient les différentes informations souhaitées et qui autorisaient le *scraping*.

Utilisation du TAL

Même si les articles ont été récupérés sur la catégorie faits divers incendies, certains articles ne parlent pas d'un incendie mais des conditions météorologiques à risques ou du déplacement de pompiers en prévention des feux d'été par exemples. Le TAL permet de contrôler si un article concerne bien un incendie. Une base de mot clés permettant l'identification de chaque article est définie après une étude de plusieurs titres. Cette base comprend différents mots comme : « incendié » ou « brûlée ». Si un des tokens du titre ne contient pas un des mots de la base alors ce titre n'est pas conservé. Cette méthode empirique garantie un contrôle du contenu de chaque article, mais possède une limite : chaque mot doit être écrit dans l'ensemble de ses formes grammaticales.

Dans un premier temps la recherche de commune pour trouver un INSEE avec le TAL se fait seulement sur le titre. Chaque titre est transformé en token grâce au package `spacy` de python et la base française de mot `fr_core_news_lg`, et chaque token identifié comme un nom propre permet de chercher un INSEE dans une base de données contenant le nom et le code de l'ensemble des communes françaises. L'exemple suivant présente l'utilisation du TAL avec en vert le token permettant de vérifier la nature du titre et en gris le token de localisation.

FIGURE 2.5: Exemple de tokenisation sur un titre de fait divers.

L'utilisation de l'API de Lettria sur les titres qui n'ont pas permis l'identification de la commune concernée permet d'augmenter le taux d'identification. Cette API retourne aussi chaque titre sous forme de token, mais leurs formes sont légèrement différentes de la méthode précédente. Ces différences, notamment sur la gestion des noms de communes composées permettent d'augmenter le taux d'identification d'INSEE par article.

Pour les sites qui présentent des chapeaux, le traitement est identique que celui des titres avec une double vérification augmentant ainsi le taux d'identification. Si un journal parle plusieurs fois du même incendie, les méthodes utilisées ici ne permettent pas de détecter les répétitions. Cette hypothèse sur les doublons est modérée lors de l'assemblage entre les différentes bases de journaux différents, en supprimant les articles qui identifient la même commune à la même date.

L'ensemble des données récoltées, les sites utilisés et les statistiques de performance sont résumées dans les tables de la section chiffres clés (2.4.3).

Assemblage des données scrapées

Après recherche des codes INSEE, les différentes bases d'incendie récupérées sont sous le même format : un code INSEE et une date. Certains sites d'informations sont spécialisés sur des régions de France. Plusieurs sources doivent être utilisés pour couvrir l'ensemble du territoire. Cependant, les sources peuvent se superposer car il arrive que différents journaux parlent du même incendie. Pour résoudre ce problème, un filtrage temporel est réalisé pour ne conserver qu'un seul incendie lors de l'assemblage. Cette hypothèse forte part du principe que chaque journal essaye de parler le plus rapidement d'un incendie, et que la vitesse de publication est sensiblement la même pour les différents sites, soit moins de 24 heures.

Gestion des dates et historiques de sinistres

La base des incendies scrapés contient alors pour chaque ligne un code INSEE et la date de publication de l'article. Pour pouvoir être exploité cette base doit être agrégée par INSEE et par période, c'est à dire compter le nombre d'incendie pour chaque commune sur le temps d'exposition choisit. Dans le cas du risque incendie, de probabilité faible, deux échelles ressortent : la semaine et l'année. Finalement l'année sera conservée, en plus de correspondre à la durée classique d'un contrat de MRH, elle permet de ne pas faire exploser la taille de base finale. En effet, avec plus de 36 000 communes, 52 semaines et un historique de seulement 10 ans, la base dépasserait les 18 millions de lignes (l'historique le plus profond est de 14 ans).

En plus du nombre de sinistres par années et par communes, quatre historiques sont établis. Les trois premiers historiques comptent le nombre d'incendies dans la commune sur, respectivement l'année précédente, les trois et cinq années précédentes. Regarder le risque seulement par année en oubliant le passé n'est pas pertinent en assurance. Le risque incendie évolue peu entre chaque année, et le renouvellement des logements est lent. Les historiques permettent de prendre en compte les mauvaises années mais aussi de pouvoir les oublier si la sinistralité diminue. L'historique stabilise la distribution du nombre d'incendie chaque année.

Ces historiques sont l'équivalent du bonus/malus en assurance automobile. Le bonus/malus est un coefficient de majoration ou de réduction de prime dépendant du nombre de sinistres sur une période antérieure.

Le principal problème des données récupérées à l'aide du web scraping est l'absence de distinction entre les incendies urbains ou naturels. Un approfondissement des techniques de TAL pourrait permettre en étudiant l'ensemble du corps de chaque article de définir la nature exacte de chaque incendie. Le risque défini avec ces données est donc très général et peu manquer de finesse sur certaines zones géographique. Une étude des saisonnalité (2.11a), (2.11b) permet de s'assurer d'une majorité d'incendie urbain dans les incendies scrapés.

La table suivante résume l'ensemble des bases utilisées et les informations quelles contiennent.

Table 2.4: Table résumant les différentes sources utilisées.

Base	Informations
Incendies scrapés	Nombre d'incendies par commune et année et historique de sinistres
SDIS Essonne	Nombre d'incendies urbain par commune et semaine
Détection satellitaire	Nombre d'incendies par commune et année
SDIS France	Nombre d'incendies par nature et département
Base INSEE	Variables socio-géographiques par commune
DVF	Variable sur la nature des biens par commune
Base urbanisation	Variables sur l'évolution des surfaces naturelles de chaque commune
Base logement	Variables sommant le nombre de logements répondant à différents critères par commune

2.3.4 Présentation des bases finales

Les variables explicatives

La base des variables explicatives fait donc 36 253 lignes qui représentent l'ensemble des INSEE du territoire français. Cette base possède 130 colonnes qui représentent chacune une variable explicative. Chaque commune possède pour son année d'étude le nombre d'incendie, ses différents historiques, et un ensemble de variable qui décrivent la nature de la commune et sa composition. L'ensemble des variables ne sera pas conservé. Certaines seront agrégées ou supprimées lors de la modélisation. Seules les variables historiques évoluent avec les années, les autres variables explicatives sont fixes et ne change pas selon l'année étudiée.

La base issue des données scrapées

Une première base est composée uniquement des données récupérées dans les rubriques faits divers. L'objectif de cette base est d'étudier le risque incendie en considérant que que les données scrapées sont plus orientées sur les incendies urbains. L'historique de cette base est de 9 années de 2012 à 2020. L'historique entier ne sera pas utilisé, la profondeur maximale sera étudiée lors de la réalisation des différents modèles de projections.

La base des feux de végétation

Une deuxième base est construite avec seulement les incendies de végétation. Ces incendies représentent aussi un risque pour les habitations, cette base va permettre de comparer les résultats avec la base des données scrapées. L'historique de cette base est 21 années de 2000 à 2020. Cette base sera étudiée dans un second temps après la base des données scrapées.

L'essor des données publiques et les améliorations technologiques facilitant la collecte et le stockage de données permet d'obtenir des bases de données comprenant beaucoup plus de variables que précédemment et même d'avoir plus de variables que d'individus observés. Quand le nombre de dimensions augmente, la taille de l'espace qui contient ces points augmente très rapidement, ainsi que la distance entre chaque point. L'espace est alors très creux et les observations sont isolées. Ce problème est appelé le fléau de la dimension et a été introduit par Bellman (1961). L'objectif est de réduire au maximum le nombre de variables lors de la mise en place des différents modèles pour réduire ce problème.

2.3.5 Contrôle de qualité

En utilisant différentes bases de données pour créer une base adaptée à un besoin spécifique, certaines données peuvent être redondantes. Ces informations en double peuvent être un problème dans les modèles statistiques appliqués sur la nouvelle base et doivent donc être identifiées mais elles peuvent aussi servir à un contrôle de qualité. En comparant deux variables de deux bases indépendantes qui décrivent le même aspect d'un même individu ou objet, la pertinence et l'exactitude des bases utilisées peuvent être mesurés.

Description des biens immobiliers

Dans le cadre de ce mémoire une comparaison va porter sur deux bases de données utilisées pour décrire les biens immobiliers présents dans une commune. Comprendre la composition du parc de logement d'une commune est un critère essentiel à l'établissement d'un modèle d'assurance pour la garantie incendie MRH. Les bases avec des descriptifs sur les biens immobiliers sont celles du recensement de la population par l'INSEE en 2017 décrivant les logements et la base gouvernementale sur les demandes de valeurs foncières avec des valeurs moyennes sur les historiques de 2019 à 2015. Les deux bases permettent d'obtenir le taux de maison et d'appartement par rapport au nombre total de logements dans chaque commune. La base de valeurs foncières est un échantillon forcément moins précis que le recensement car elle ne considère que les biens vendus, ce qui peut être problématique dans les petites communes où le nombre de transactions est réduit. Mais le recensement n'est pas une mesure sans erreur, d'après la source suivante : INSEE (2017), sa qualité dépend de différents facteurs mais surtout de la qualité de la collecte. Les communes de moins de 10000 habitants sont enquêtées exhaustivement, mais pour les communes de 10000 habitants ou plus, seulement environ 40% des ménages sont interrogés. Il y a donc une marge d'incertitude lors de l'étude principale sur les résultats appelée le coefficient de variation, noté CV. Ce coefficient est le rapport de l'écart type à la moyenne, il est utilisé dans la construction de l'intervalle de confiance de l'estimations à 95% des valeurs trouvées.

Si VR est la valeur recensée, son intervalle de confiance est le suivant :

$$[VR \times (1 - 2 \times CV); VR \times (1 + 2 \times CV)]. \tag{2.1}$$

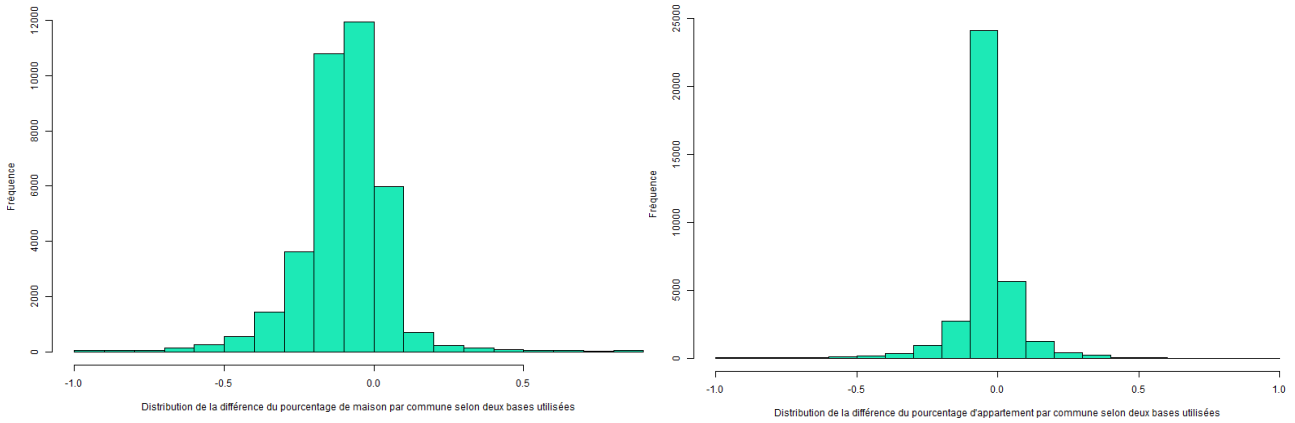
L'imprécision varie donc entre chaque commune, selon le nombre d'habitants et le taux de ménages interrogés. Des études complémentaires, permettant d'obtenir plus d'informations, sont aussi menées sur 20% des ménages pour les communes de moins de 10000 habitants et sur 100% des ménages interrogés pour les communes de plus de 10000 habitants, donc environ 40% des ménages des communes. Ainsi, même pour les communes de moins de 10000 habitants réputées justes un intervalle de confiance doit être déterminé pour les informations récupérées lors des études complémentaires. Soit e l'effectif estimé d'une variable et τ le taux d'incertitude (20 pour les communes de moins de 10000 habitants), l'écart type est estimé par $\sqrt{\frac{e}{\tau}}$, l'intervalle de confiance à 95 est donc :

$$\left[e - 2\sqrt{\frac{e}{\tau}}; e + 2\sqrt{\frac{e}{\tau}}\right]. \tag{2.2}$$

Pour chaque INSEE, deux variables de bases différentes décrivent le taux de maison et d'appartement. Les valeurs ne sont pas identiques comme attendue. Les écarts entre les taux de maison et les taux d'appartements des deux bases de données sont calculés respectivement pour chaque commune et la moyenne de ces différences est ensuite mesurée. Les résultats sont présentés dans le tableau suivant (**??**). Les figures suivantes (2.6a) et (2.6b) présentent la distribution des différences du taux d'appartement et du taux de maison entre les deux variables de bases différentes.

TABLE 2.5: Différence logements entre bases de données en pourcentage

	MAISON	APPARTEMENT
Différence moyenne (%)	12,94	6,36



(a) Erreur sur le pourcentage de maison (b) Erreur sur le pourcentage d'appartement

FIGURE 2.6: Exemple de distribution de la différence de pourcentage entre deux variables de bases différentes

Le nombre de maisons par commune est deux fois moins bien identifié que le nombre d'appartement. La queue de distribution de l'erreur sur le taux de maison est plus épaisse que pour le taux d'appartement. L'étude ligne à ligne relève aussi des erreurs : dans l'exemple suivant une commune ne possède pas d'appartement d'après l'INSEE mais des biens ont fait l'objet d'actes notariés sous

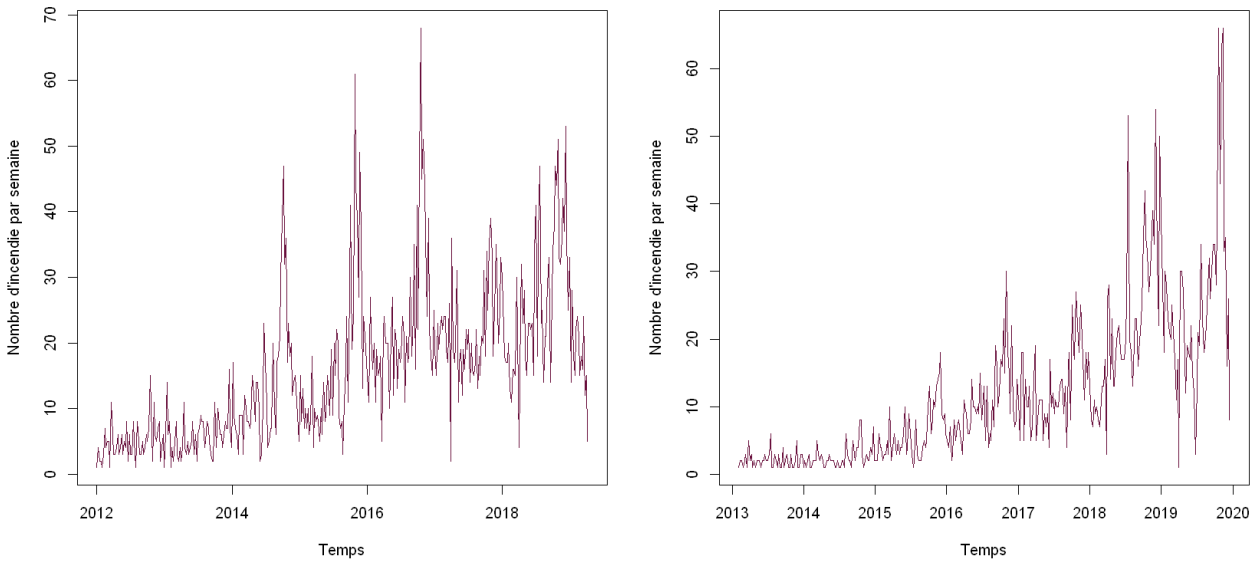
l'appellation appartement. Ce type d'aberration peut s'expliquer par l'incertitude du recensement mais aussi par la définition d'appartement prit dans les deux bases. Une maison transformée en deux logements indépendants peut, en effet, être considérée comme deux maisons ou deux appartements. Les bases ne détaillent pas leurs critères de choix d'appellation. Cette source de biais est à prendre en compte avant de développer un modèle et lors de l'étape de sélection de variable.

Etude des séries temporelles

L'étude des séries temporelles du nombre d'incendie sur les différentes bases permet de vérifier la cohérence des données utilisées. Les bases de données provenant des services de secours, la base nationale des interventions à la maille départementale et la base Essonne, servent respectivement de référence pour la base des feux de végétation et la base des incendies scrapés.

Le nombre d'interventions pour incendie n'est pas en augmentation avec les années. Les statistiques des services d'incendie et de secours comptabilisent 361 240 interventions pour incendies en 2002 et 305 500 en 2018, soit une baisse d'environ 15%, Ministère de l'intérieur (2020). La population française sur la même période est passée de 61 385 070 à 66 883 761 habitants, soit une augmentation d'environ 9%, INSEE (2019b) ce qui signifie que le nombre de logement a aussi augmenté. La tendance du nombre d'incendie identifiés à l'aide de techniques de web scraping est donc inversée par rapport à la tendance réelle sur le territoire français.

Données scrapées (2.7a) et (2.7b) Une étude du nombre d'incendie par semaine pour les données scrapées montre une tendance croissante avec le temps. L'augmentation du nombre d'incendies identifiés est liée à la qualité d'archivage des sites internet. Les articles des années 2012, 2013, sont moins complets et précis que les derniers articles de 2020. La tendance se réduit fortement à partir de 2017.



(a) Données site d'information numéro 1 (b) Données site d'information numéro 2

FIGURE 2.7: Séries temporelles du nombre d'incendies pour les données incendie scrapées selon le site d'information.

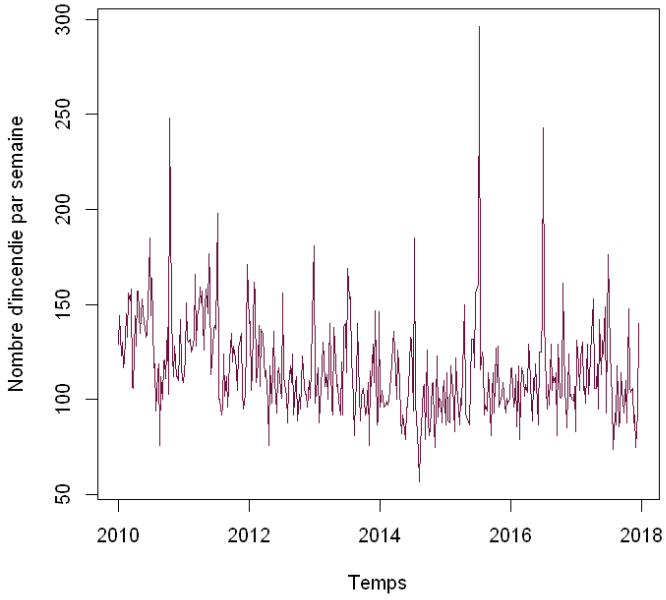
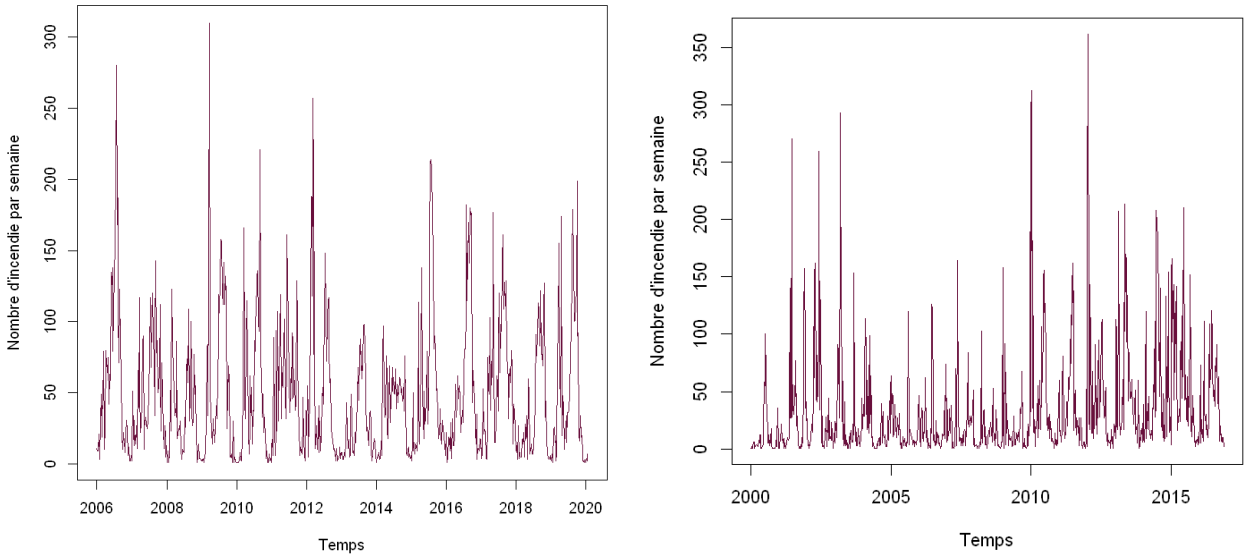


FIGURE 2.8: Série temporelle du nombre d'incendies en Essonne (base du SDIS).

Données incendie en Essonne (2.8) L'étude de la série temporelle du nombre d'incendie en Essonne confirme la tendance décroissante du nombre d'incendie par année, malgré l'augmentation du nombre de logement.



(a) Feux de forêts de la base BDIFF

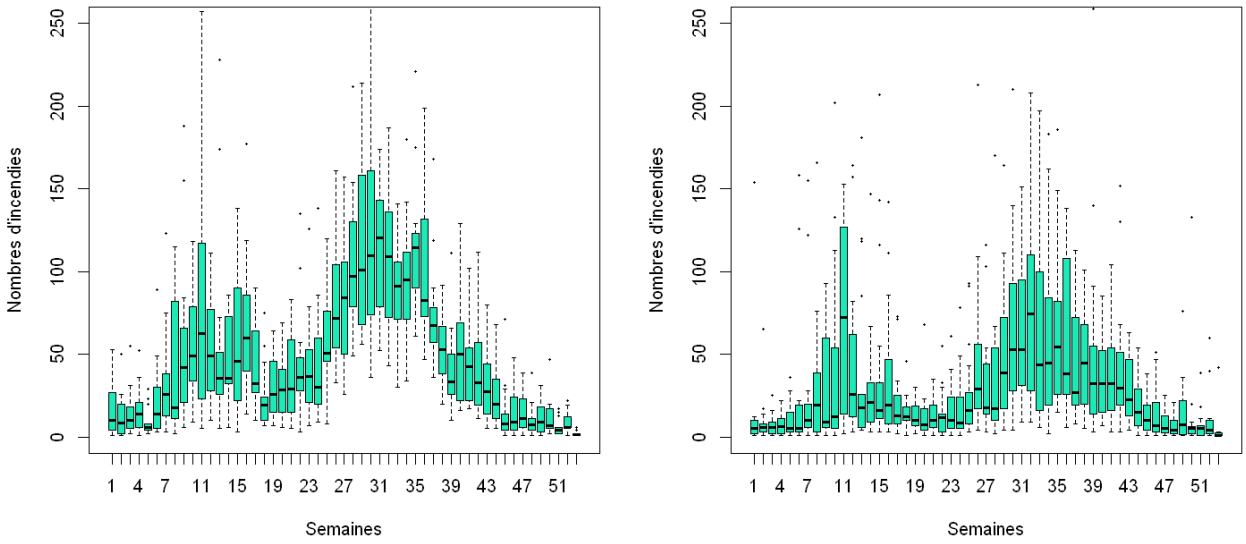
(b) Feux détectés par les satellites de la NASA

FIGURE 2.9: Series temporelles du nombre d'incendies pour les différentes bases de données sur les feux de forêts

Données NASA (2.9b) La série temporelle des incendies détectés par la NASA après la suppression des incendies dans les communes industrialisées montre une saisonnalité annuelle avec une tendance nulle malgré l'arrivée de satellites de nouvelles générations.

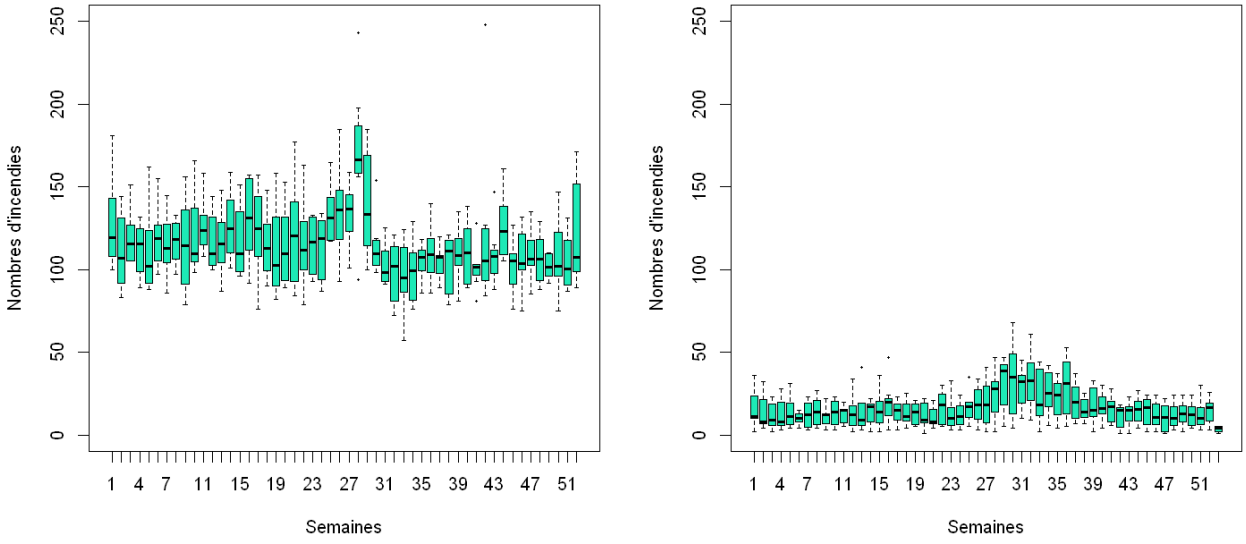
Données feux de forêts (2.9a) L'étude de la série temporelle du nombre de feux de forêts par semaine depuis 2006 jusqu'à fin 2019, montre une saisonnalité annuelle, qui est attendue et qui correspond à la période chaude de l'année. Le nombre d'incendies n'est cependant pas en augmentation sur la période observée.

Les différentes séries temporelles précédentes n'ont pas la même profondeur d'historique, ce qui rend difficile leur comparaison. Pour pouvoir comparer leurs saisonnalités annuelles, les boîtes à moustaches, ou *box plot* en anglais, du nombre d'incendie par semaine sont réalisées. Les *box plots* permettent de représenter différents indicateurs comme les quantiles ou la moyenne et comparer des populations de tailles différentes. Certaines bases ont des historiques faibles : 8 ans au minimum. Réaliser un boxplot sur un nombre si réduit de valeurs n'est pas la méthode la plus représentative. Mais cette démarche permet d'afficher tous les boxplots à la même échelle pour faire apparaitre la saisonnalité et comparer les amplitudes.



(a) Feux de forêts (b) Feux NASA

FIGURE 2.10: Boxplot du nombre d'incendies par semaine pour les différentes bases de données utilisées



(a) Feux urbain en essonne (b) Feux scrapés

FIGURE 2.11: Boxplot du nombre d'incendies par semaine pour les différentes bases de données utilisées

La base des feux de forêts (2.10a) montre deux pics d'incendie : pendant les mois d'été et à la sortie de l'hiver. En début de printemps, la végétation gelée pendant l'hiver est très inflammable. Ces feux de débuts d'années sont majoritairement présents dans les massifs landais et les régions montagneuses de basse altitude, GOUVERNEMENT (2020). La base de données issue de la NASA, (2.10b), montre la même saisonnalité annuelle, ce qui confirme la première intuition sur la présence majoritaire d'incendies de végétation dans cette base. L'ordre de grandeur du nombre d'incendie est proche entre les deux bases.

La base Essonne présente (2.11a) un seul pic d'incendie en été, plus particulièrement la $28^{ème}$ semaine, qui correspond habituellement au 14 juillet. Cette augmentation saisonnière est faible par rapport au reste de l'année. La base des incendies scrapés présente aussi une augmentation en été mais qui dure plus longtemps : sur cette période les articles parlent aussi des feux de forêts. L'ordre de grandeur du nombre d'incendies scrapés est inférieur à l'ensemble des opérations en Essonne.

L'étude des séries temporelles justifie ainsi la séparation des incendies en deux bases de données selon leurs sources.

2.4 Bonnes pratiques

2.4.1 Pre-processing

Avant de commencer tout projet de modélisation à l'aide de bases de données, une étape de vérification doit être réalisée, en anglais *pre-processing*. Même si au premier abord une base de données semble propre et utilisable immédiatement, ne pas prendre le temps d'étudier la base peut en faire perdre beaucoup plus dans la suite de l'étude lors de la conceptions de modèle. Il est notamment indispensable de rechercher des valeurs aberrantes (un âge négatif, les combinaisons impossibles comme par exemple un individu masculin et enceinte) ou les valeurs manquantes.

2.4.2 Dictionnaire de variables

Lors de la réalisation d'une base de données contenant un grand nombre de variables ou d'individus, les informations pertinentes peuvent être difficiles à retrouver. Pouvoir identifier facilement et rapidement la nature des éléments de la base est essentielle à son utilisation et sa réutilisation. Une base propre et facilement exploitable peut faire gagner du temps aux personnes qui l'utilisent, comme l'équipe de tarification d'un assureur par exemple. Un dictionnaire de variable est un outil facile à mettre en place et qui peut être de forme simple comme un tableur. Ce dictionnaire doit contenir toutes les variables de la base de données à laquelle il fait référence. Chaque variable peut être décrite, de manière exhaustive ou partielle. Pour une variable donnée, les informations peuvent être le nombre de modalités différentes et la granularité (rapport entre le nombre de modalités différentes de la variable et d'individus dans la base). Les deux sources d'informations précédentes peuvent être facilement automatisées pour remplir le dictionnaire. Pour une variable quantitative, les différents quantiles, la valeur moyenne et pour une variable qualitative le mode (modalité majoritaire de la variable), sont des informations qui peuvent être ajoutées dans le dictionnaire et servir à compléter des valeurs manquantes.

Le dictionnaire de variables est d'autant plus important en actuariat que la taille des données

utilisées pour tarifer les différents produits est importante. Le rythme de renouvellement (*turnover*) des équipes au sein de la profession est élevé. Le dictionnaire de variables permet de faciliter la réutilisation de la base par une nouvelle personne. Un dictionnaire sert aussi dans la création de modèles de prédiction en offrant une vision globale des données à disposition, particulièrement lors de la sélection de variables. De plus, le dictionnaire peut améliorer la gestion de la base avec un historique de modifications par exemple, tout en permettant de protéger les droits, qu'ils soient de propriété intellectuelle ou d'accès. Le dictionnaire de variables permet de respecter le point 5.4 de la Norme de Pratique relative aux Modèles actuariels qui porte sur la traçabilité, le jugement et la responsabilité des rapports émis, Institut des actuaires (2016).

Un dictionnaire de variables est de la métadonnée, du préfixe grec *meta* qui peut signifier : englober un objet. La métadonnée est de la donnée sur des données. Le dictionnaire de variable possède bien des informations sur sa base de données associée. L'ensemble des métadonnées sont soumises à l'application du RGDP, Ministère de l'éducation nationale et de la jeunesse (2020)

La création du dictionnaire de variable rentre dans le cadre de la mise en place d'un RoPA.

2.4.3 Chiffres clés

Les chiffres clés sont des indicateurs, leurs nature peuvent être très variées comme le nombre de modalité d'une variable. Conserver les différents chiffres clés sur les données lors de leur manipulation permet d'obtenir un moyen d'évaluer la progression d'un projet. En plus de permettre un contrôle de qualité, avoir un indicateur de temps est essentiel pour s'assurer de l'avancement d'un projet. Lors de la mise en place des techniques de scraping et NPL, un seul site a été utilisé pour calibrer les fonctions. La table (2.6) présente l'évolutions de la performances des techniques de scraping et d'identification d'incendie sur un unique site d'information.

Table 2.6: Table des chiffres clés de l'amélioration de la récupération et de l'identification des données scrapés.

Date	Articles scrapés	Identification incendie	Identification INSEE	Taux de correspondance
30-06-2020	7968	6730	3362	0,49
27-07-2020	7968	7178	4811	0,67
06-08-2020	9891	8974	5955	0,66
24-08-2020	9986	9025	6776	0,75

Une fois les méthodes bien définies, elles ont été reproduites sur différents sites. La table (2.7) contient les informations sur l'ensemble des sites scrapés une fois les méthodes de scraping et d'identification fixées.

Table 2.7: Table des chiffres clés sur la récupération de données scrapés sur différents sites.

Date	Articles scrapés	Identification incendie	Article avec un INSEE	Taux de correspondance	Journal
24-08-2020	9986	9025	6776	0,75	France 3
14-09-2020	8528	6483	4182	0,65	France Bleu
19-09-2020	15718	1890	1312	0,69	Midi libre
20-09-2020	1900	216	129	0,6	Le Dauphiné
20-09-2020	1920	222	112	0,50	Le Progrès
20-09-2020	1960	134	50	0,37	Est Républicain
20-09-2020	1960	235	146	0,62	Bien Public
20-09-2020	1940	298	198	0,66	Le JSL
20-09-2020	1960	202	111	0,55	Vosges
22-09-2020	1940	187	88	0,47	L'alsace
22-09-2020	1714	146	98	0,67	Republicain Lorrain
22-09-2020	1695	108	52	0,48	Actu
22-09-2020	20400	862	294	0,34	La provence

Chapitre 3

Analyse exploratoire des données et modélisation

3.1 Statistiques descriptives

Avant de développer la théorie et l'application des différentes modélisations, une étude statistique de la base utilisée est réalisée pour comprendre les différentes problèmatiques présentes.

3.1.1 Corrélations

La base utilisée contient 135 variables, provenant de cinq bases de données publiques différentes ainsi que du web-scraping. L'ensemble de ces variables est détaillé dans les tables A.1, A.2, A.3, A.4, A.5, A.6 situées en annexes. Les informations présentes peuvent être redondantes ou liées entre elles. Pour contrôler cette hypothèse, une matrice de corrélations des variables est établie avec un corrélogramme. La figure suivante (3.1) représente les différentes corrélations (3.4) pour un échantillon de variables (l'ensemble du corrélogramme pour les 135 variables ne peut pas être représenté en bonne qualité sur une page A4).

Il y a peu de variables fortement corrélées négativement entre elles, seules les variables selon la nature du bien le sont. Cependant, beaucoup le sont positivement. En parcourant la diagonale de haut en bas plusieurs blocs de variables apparaissent fortement corrélées :

- Les différents historiques et le nombre d'interventions.
- Les différents types incendie par départements.
- La nature des biens.
- Les variables d'évolution de la population, des ménages et de l'emploi.
- L'ensemble des variables de la base logement.

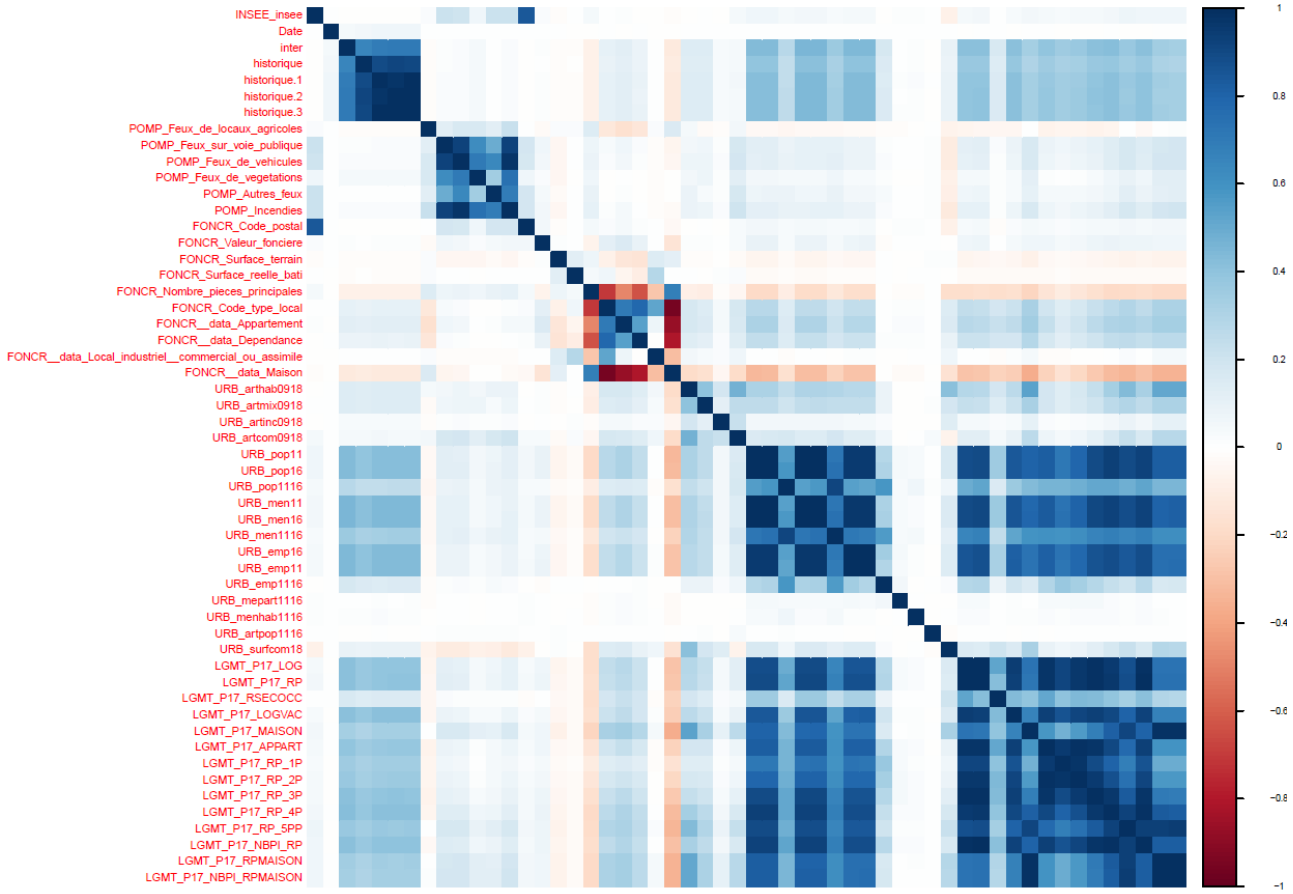


FIGURE 3.1: Corrélogramme d'un échantillon des variables de la base des incendies scrapés

3.1.2 Cluster de variables

Le package ClustOfVar a été développé pour ordonner les variables, qualitatives et quantitatives, redondantes dans une base de données. L'algorithme de classification ascendante hiérarchique agrège ainsi à chaque étape les variables qui minimisent la perte d'homogénéité. La théorie des modèles utilisés n'est pas développée ici mais peut être trouvé dans la documentation associée au package, CHAVENT et al. (2017) et CHAVENT et al. (2012).

La figure en annexe (A.7) présente les différents clusters pour le même échantillon de variables que dans la figure (3.1). Les différents groupes de variables les plus proches sont les mêmes qu'identifiés avec le corrélogramme : les historiques et nombre d'intervention, les types d'incendies par département, les variables sur la nature des biens, celles sur les évolutions démographiques, et celles de la base logement.

3.1.3 Réduction de la dimension

Le corrélogramme et les clusters de variables montrent la possibilité de réduire le nombre de variables dans la base étudiée, et ainsi réduire la dimension de l'espace d'apprentissage sans perte d'information.

Un seul historique sera conservé et divisé par le nombre de logements pour être décorrélé de la variable intervention. Les variables de la base de données sur les logements seront agrégées avec un nombre limité de modalités, par exemple : le nombre de logement construit avant ou après 1990 plutôt que 8 variables pour différentes dates. Différentes variables non explicatives sont supprimées.

La base contient une profondeur de sinistre de 9 années : de 2012 à 2020, cependant l'ensemble des années ne sera pas utilisé, permettant à la variable historique d'être pertinente.

3.1.4 Chiffres clés

La table (3.1) présente pour chaque source d'incendie la fréquence de survenance (nombre de sinistre divisé par l'exposition) relativement à celle de la FFA. La fréquence annoncée par la FFA pour l'année 2018 (0,55%) est considérée comme valeur de référence et la colonne coefficient représente le rapport entre la fréquence étudiée et celle de référence. La fréquence d'incendie observée sur la base des incendies scrapés est de 0,001%, ce qui est environ 500 fois plus faible que la moyenne du marché annoncée par la FFA. Ce déficit d'information montre les limites du *web scraping*, qui par nature n'est pas exhaustif. Bien que la valeur de la FFA ne soit pas non plus exhaustive des incendies (tous les sinistres nécessitant l'intervention des pompiers ne sont pas forcément déclarés à l'assureur), elle sera utilisée comme référence car elle est plus représentative du risque porté par l'assureur. Le portefeuille d'assureur sera présenté dans la suite de ce chapitre, (3.4).

TABLE 3.1: Fréquence d'incendie relative par rapport à la FFA.

SOURCE	COEFFICIENT
SDIS France	1,5
SDIS Essonne	1,8
FFA	1
Incendies scrapés	0,02
Portefeuille assureur	0,98

3.1.5 Cartes choroplèthes

L'objectif de l'étude étant d'améliorer la compréhension du risque incendie sur l'ensemble de la France métropolitaine, les différents résultats seront représentés sur une carte de France avec une granulométrie communale. La figure (3.2) présente ainsi pour chaque commune la fréquence d'incendie par logement avec l'ensemble des incendies scrapés.

La densité des incendies scrapés est assez homogène sur le territoire. La fréquence relativement plus faible dans le Nord de la France, s'explique par la taille des communes qui sont plus petites. A fréquence d'incendie comparable, ces communes sont donc moins visible. L'exposition de chaque commune au risque incendie est le nombre de logements dans la commune. La figure suivante (3.3) représente la distribution de l'exposition en sept niveau pour chaque commune. Plus la couleur se rapproche du rouge, plus la commune compte de logement.

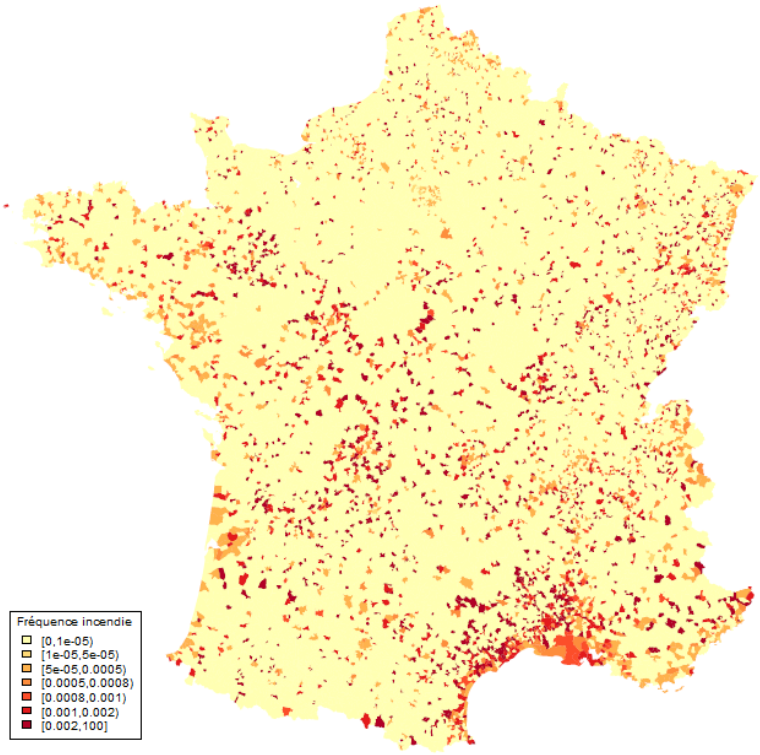


FIGURE 3.2: Carte choroplèthe de la fréquence des incendies par commune

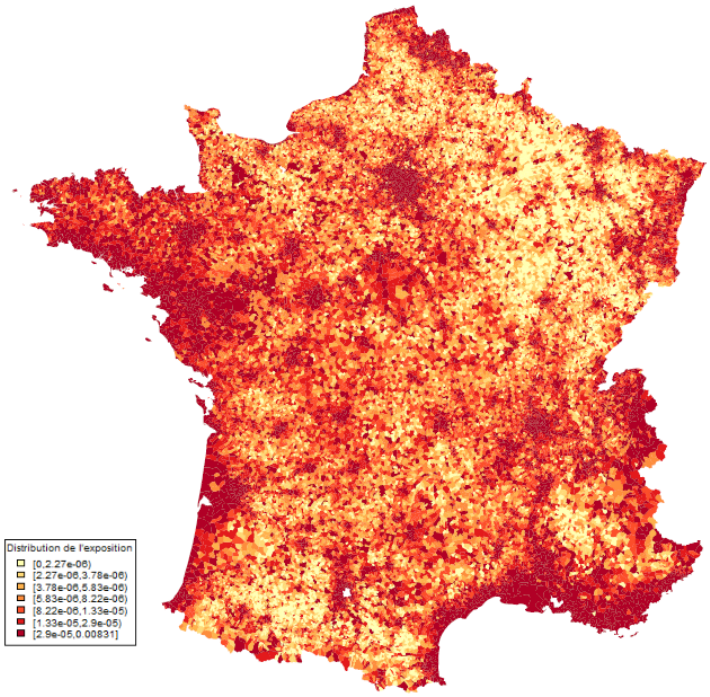


FIGURE 3.3: Carte choroplèthe de la distribution des logements par communes

Trois régions ont un faible nombre de logement : les Pyrénées, les Alpes et le grand Est. Les côtes possèdent une forte concentration de logement. La répartition des logements fait apparaître la diagonale des faibles densités de population.

3.2 Les différents modèles

Pour déterminer la prime pure, il faut pouvoir estimer le coût et la fréquence des sinistres en fonction des différentes informations connues sur l'assuré. Cela nécessite de disposer d'une fonction entre la sinistralité de l'assuré et les différentes variables qui le définissent. Différents modèles permettant d'identifier ce lien sont présentés dans ce chapitre.

3.2.1 Modèle linéaire généralisé

Les modèles linéaires généralisés ou GLM (*Generalized Linear Models*) sont les modèles les plus utilisés en assurance, et particulièrement en IARD et en santé prévoyance. Ces modèles sont des généralisations souples de la méthode de régression par moindres carrés (Wikistat (2020e) et P McCullagh (1989)), qui cherchent à définir un lien entre une variable réponse, notée Y et différentes variables explicatives notés $(X_1, X_2, \cdots, X_N)$. Pour le risque incendie en MRH, les GLM peuvent être utilisés pour définir la fréquence ou le coût. Par exemple Y peut représenter le nombre d'incendie sur une année et $(X_1, X_2, \cdots, X_N)$ les différentes variables décrivant l'assuré et le bien : âge, catégorie socioprofessionnelle, localisation, ancienneté du bien assuré...

Un GLM peut être décomposé en trois éléments : la variable réponse que le modèle cherche à reproduire, les variables explicatives qui constituent le cœur de la modélisation et la fonction de lien qui fait la liaison entre les deux premiers points.

La variable réponse La variable réponse Y est la variable que le GLM cherche à estimer. Pour chaque individu, l'espérance de Y est déterminée. La base de données doit contenir K observations de la variable Y, $(y_1, y_2, \cdots, y_K)$. La loi de la variable réponse doit appartenir à la famille des lois exponentielles, c'est-à-dire que la densité de Y, notée $f(y)$ peut s'écrire de la manière suivante :

$$f_{\theta,\phi}(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{3.1}$$

où $a(.)$, $b(.)$ et $c(.)$ sont des fonctions. Le paramètre θ est le paramètre d'intérêt ou encore appelé paramètre naturel. ϕ est le paramètre de nuisance.

Par exemple, la loi de Poisson de paramètre λ est une loi qui appartient à la famille exponentielle, sa densité est :

$$f(y) = \exp(-\lambda)\frac{\lambda^y}{y!} = \exp\left(y\log(\lambda) - \lambda - \log(y!)\right),\ y \in \mathbb{N}. \tag{3.2}$$

Par identification : $\theta = \log(\lambda)$, $\phi = 1$, $a(\phi) = 1$, $b(\theta) = \exp(\theta) = \lambda$ et $c(y,\phi) = -\log(y!)$.

Les lois gaussiennes, de Bernoulli, binomiale, gamma ou binomiale négatives sont aussi des lois exponentielles.

Pour une variable Y dont la densité peut s'écrire de manière exponentielle , son espérance et sa variance peuvent s'écrire respectivement comme :

$$\mathbb{E}[Y] = b^{'}(\theta) = \mu \text{ et } V(Y) = b^{''}(\theta)a(\phi). \tag{3.3}$$

Les variables explicatives

Déterminer quelles variables choisir pour modéliser la variable Y est complexe. Le nombre de variables conservées doit être assez grand pour pouvoir singulariser chaque individu. Mais il ne doit pas être trop important pour ne pas rendre l'algorithme trop lourd et trop individualisé. Les variables conservées doivent être celles dont la puissance explicative est la plus significative. Pour ce faire, différents critères mesurent la qualité d'un modèle statistique comme les critères :

- AIC $= 2k - 2\ln(L)$ où L est le maximum de vraisemblance du modèle et k est le nombre de paramètres à estimer. Plus la valeur de l'AIC sera faible, meilleur sera le modèle, Akaike (1974).
- BIC $= -2\ln(L) + k\ln(N)$ où L et k sont les mêmes que pour le critère AIC et N le nombre d'observations de l'échantillon. Comme pour l'AIC, le modèle conservé sera celui qui réduira le plus le BIC, Schwarz et al. (1978).

Il existe des méthodes visuelles qui permettent de déterminer si plusieurs variables sont corrélées. C'est le cas des corrélogrammes qui représentent les corrélations entre plusieurs variables quantitatives. La corrélation entre chaque variable est calculée avec la formule suivante :

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}. \tag{3.4}$$

Pour les variables qualitatives, Harald Cramer a défini en 1946 une mesure, le V de Cramer, permettant de noter entre 0 et 1 le lien entre deux variables qualitatives X et Y, avec respectivement r et k modalités, qui s'écrit de la manière suivante :

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}} \text{ où } \chi^2 = \sum_{i,j} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}. \tag{3.5}$$

Avec, n est le nombre total d'observations, k le nombre de colonnes, r le nombre de rangs, n_{ij} le nombre de fois où le couple de modalité (X_i, Y_j) est observé dans l'échantillon, $i = 1, \cdots, r$ et $j = 1, \cdots, k$, Cramér (1946).

Pour une recherche de variables plus approfondie différentes méthodes sont possibles :

- Recherche exhaustive : cette méthode consiste à tester selon le critère de qualité choisie (AIC, BIC, ou autre) toutes les combinaisons possibles de variables. Le principal problème de cette méthode est le temps de calcul car le nombre de combinaison est 2^N avec N le nombre de variable.
- Recherche descendante : cette méthode itérative commence par le modèle complet, avec toutes les variables et à chaque étape, la variable la moins pertinente après un test de Student est retirée du modèle. Une fois qu'une variable a été enlevé elle ne peut plus être réintroduite.
- Recherche ascendante : cette méthode est l'inverse de la recherche descendante, en partant du modèle vide sans variable, un test de Student est réalisé sur toutes les variables séparément, la variable expliquant le plus le modèle est intégré. Cette étape est répétée jusqu'à ce qu'aucune des variables encore sélectionnées ne respecte le test de Student. Une fois une variable introduite, elle ne peut plus être retirée.
- Recherche step-wise : à chaque étape de la sélection, les variables déjà choisies précédemment sont prises en compte dans l'évaluation. Ainsi une variable introduite peut être retirée si elle ne participe plus à expliquer le modèle lorsque d'autres variables ont été introduites.

Régression pénalisées

Si les méthodes précédentes s'avèrent trop lentes du fait d'un nombre de variables trop important, des sélections par pénalisation sont possibles. L'ajout d'une pénalité lors de l'optimisation des paramètres permet de contrôler l'ampleur des coefficients calibrés. La forme de cette pénalité ou contrainte peut être différente selon les modèles (variation de la valeur α). Dans le cadre d'une modélisation par une loi de Poisson la valeur estimée de la variable réponse y par les variables explicatives x :

$$\hat{y} = e^{x^T\beta+\beta_0}, \tag{3.6}$$

alors le problème est estimé en maximisant la log-vraisemblance (3.9) pénalisée :

$$\max_{\beta,\beta_0} \frac{1}{N}\sum_{i=1}^{N}\Big[y_i(x_i^T\beta+\beta_0) - e^{x_i^T\beta+\beta_0}\Big] - \lambda\Big[\alpha||\beta||_1 + \frac{1}{2}(1-\alpha)||\beta||_2^2\Big]. \tag{3.7}$$

La méthode **Lasso** introduite par Tibshirani (1996) est une pénalisation par la norme $\mathbb{L}^1$: le problème 3.7 est résolu sous la contrainte suivante : $\sum_{j=1}^{p}|\beta_j| \leq t$, soit $\alpha = 0$ dans (3.7). Si plusieurs variables sont corrélées, cette méthode en retient une partialement en ôtant les autres.

La méthode **Ridge** est une pénalisation par la norme $\mathbb{L}^2$, le problème 3.7 est résolu sous la contrainte : $\sum_{j=1}^{p}|\beta_j|^2 \leq t$, soit $\alpha = 0$ dans (3.7). (Régularisation de tikhonov). Cette méthode a tendance à réduire les coefficients des variables corrélés de manière simultanés.

La méthode **Elastic Net** est un mélange des deux méthodes précédentes. La contrainte est de la forme : $\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} |\beta_j|^2 \leq t$. Le choix de α doit se faire avec prudence. Une méthode de détermination est la validation croisée. Les cas extrêmes où $\alpha = 1$ et $\alpha = 0$ correspondent respectivement à la méthode Lasso et la méthode Ridge.

La réécriture d'un problème sous contraintes avec les conditions de Karush-Kuhn-Tucker fait apparaître un terme λ, le multiplicateur de Lagrange, qui doit lui aussi être optimisé. Ce terme définit le poids attribué à la pénalité. Plus sa valeur est grande, plus la pénalisation sera forte et les coefficients estimés seront proche de zéro, si $\lambda = 0$, la régression sera linéaire. Il peut être estimé par validation croisée. Les packages R qui réalisent les calculs de la valeur λ proposent généralement deux valeurs : une valeur minimale et une valeur avec une erreur. Le deuxième λ est souvent plus parcimonieux en sélectionnant moins de variables, particulièrement pour la méthode Lasso.

Le choix entre ces trois modèles est complexe. En étant une combinaison des deux premières, la méthode d'Elastic Net apparaît alors comme la meilleure solution. La méthode de Ridge avantage la sélection de variables pertinentes alors que la méthode de Lasso cherche à réduire l'ensemble des variables présentes ce qui peut être intéressant en très grande dimension, Guillot (2015). La méthode de Lasso sera préférée pour sa sélection parcimonieuse et la pertinence de la sélection : elle permet de sélectionner les variables d'intérêt avant les autres. L'utilisation d'Elastic Net est très coûteux en temps de calcul avec la recherche de λ et α en validation croisée sur la base *train* (voir 3.2.6 et (3.5)). Cette méthode ne sera pas utilisée dans ce mémoire.

Fonction de lien

La fonction de lien est la fonction qui associe les variables explicatives à la variable réponse. Cette fonction est notée g, elle doit être réelle, monotone et différentiable. La fonction de lien doit être choisie pour que les valeurs prédites soient de même nature que la variable Y. Par exemple dans le cas où Y est une variable de comptage, le modèle ne doit retourner que des valeurs positives. Dans le cas où plusieurs fonctions de liens semblent donner de bons résultats, celle minimisant l'erreur sera préférée.

$$g\left(\mathbb{E}[Y]\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{3.8}$$

Les fonctions de liens les plus courantes sont :

- Fonction de lien identité : $g = Id$, c'est le modèle normal, pour des valeurs comprises dans $\mathbb{R}$.
- Fonction de lien log : $g = \ln$, c'est le modèle multiplicatif, qui permet d'obtenir seulement des valeurs positives. La structure du modèle permet de séparer les effets de chaque variable.
- Fonction de lien logit : $g(x) = \ln(\frac{x}{1-x})$. Par écriture inverse, les valeurs générées seront comprises entre 0 et 1 inclus.

Pour l'actuariat, la fonction de lien log est utilisée pour avoir un tarif multiplicatif, la loi de Poisson ou la loi binomiale négative sont utilisées pour la fréquence, et la loi gamma ou log-normale sont utilisées pour le coût, Planchet (2017).

L'estimation des différents coefficients β_i avec $(i = 1, \cdots, k)$, est réalisée à l'aide de la méthode du maximum de vraisemblance. Par hypothèse d'indépendance des Y_i avec $i = 1, \cdots, n$, la log-vraisemblance est donc :

$$\log L_{y,\beta} = \sum_{i=1}^{n} \ln(f_{\theta,\phi}(y_i)) = \sum_{i=1}^{n} \frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi). \tag{3.9}$$

Dans l'équation (3.9), les différents coefficients β_i ne semblent pas être explicite, mais leur présence s'explique par la réécriture de (3.8) sous forme matricielle :

$$\mathbb{E}[Y_i] = g^{-1}(\mathbb{X}_i^T \beta). \tag{3.10}$$

Le maximum de vraisemblance est estimé par des méthodes d'optimisation itératives. L'ensemble des démonstrations sur les différentes conditions d'optimalités ne seront pas décrites dans ce mémoire, elles sont présentées dans le mémoire de Pariente (2017).

Soit $\hat{\mu}_i$ la valeur estimée par un GLM pour l'observation i, l'erreur de prédiction, ou résidus est $\hat{\epsilon}_i = Y_i - \hat{\mu}_i$. Mais l'hétéroscédasticité des modèles ne permet pas une étude approfondie de ces erreurs. Pour cela deux autres résidus sont plus couramment utilisés :

- Résidus de Pearson : $\hat{\epsilon}_{P,i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}}$.
- Résidus de déviance : Ils mesurent la participation de chaque individu à la déviance du modèle saturé.

L'utilisation du GLM possède des limites. Il faut que l'échantillon d'étude soit suffisamment grand, et il faut un nombre de variables inférieur à la taille de l'échantillon. Différentes techniques ont été mises en place pour pouvoir utiliser un nombre de variables élevé comme la régression pénalisée ou l'agrégation de modèles. Wikistat (2020d).

3.2.2 Modèles additifs généralisés

Dans la modélisation de la fréquence du risque incendie, le lien entre les variables explicatives n'est pas forcément linéaire. C'est couramment le cas de la variable âge de l'assuré. Les GLM, avec la fonction de lien, permettent une modélisation plus précise que les régressions linéaires mais ne permettent pas d'identifier les effets non monotones. Les modèles additifs généralisés ou GAM pour *Generalized Additive Model* sont des prolongements des modèles linéaires. Ils sont plus souples dans leur modélisation en s'émancipant de la contrainte de linéarité, mais restent facilement interprétables en préservant l'additivité, Hastie et Tibshirani (2014). Les modèles sont de la même forme que les GLM avec une fonction de lien g pour estimer l'espérance de la variable réponse (3.8) :

$$g\left(\mathbb{E}[Y]\right) = \beta_0 + f_1(X_1) + \cdots + f_N(X_N) \text{ où } f_i,\ (i = 1, \cdots, N) \text{ sont des fonctions non linéaires.} \quad (3.11)$$

Le choix des fonctions f_i peux se faire parmi les fonctions polynomiales, en escalier ou les splines, Guillot (2015). L'exemple (A.10) présente l'intérêt du GAM dans le cas où le lien entre la variable réponse et la variable d'étude est non linéaire : $f(x) = \sin(2(4x-2)) + 2e^{-(16^2)(x-0.5)^2} + \epsilon$ avec $\epsilon \sim N(0, .3^2)$.

3.2.3 Arbres binaires de décision

Les arbres binaires de décision modélisant une discrimination ou une régression sont couramment appelés arbres CART (*Classification and Regression Trees*), ces arbres sont construits par itérations successives des subdivisions d'un échantillon d'individus. Les méthodes de partitionnement récursif ont commencé à être développées dans les années 60, et une formalisation a été proposée par Breiman et al. (1984). A chaque étape, les séparations sont choisies pour obtenir des sous-populations avec le plus de différences possibles. La figure suivante (3.4) est un exemple d'arbre de décision pour décider le niveau de risque d'un logement donné selon son prix et sa surface.

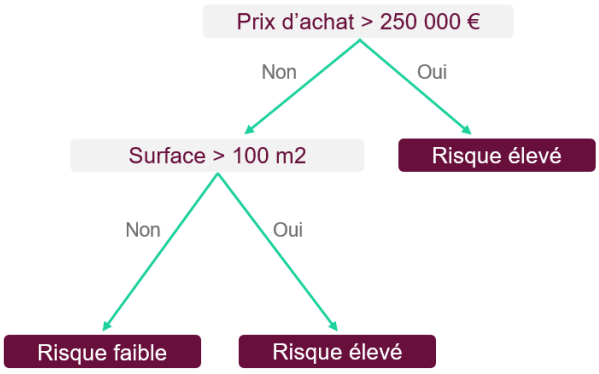


Figure 3.4: Exemple d'arbre binaire de décision.

Soit Y la variable réelle aléatoire à expliciter qui peut être qualitative ou quantitative. Si la variable Y est qualitative, l'arbre doit définir à quelle classe appartient l'observation Y_i : c'est un arbre de **classification**. Si la variable Y est quantitative, l'arbre doit permettre de préciser la valeur de Y_i : c'est un arbre de **régression**.

Soit $(X_1, X_2, \cdots, X_N)$ les variables explicatives. Les variables explicatives peuvent être qualitative ou quantitative et contenir des valeurs aberrantes, ces valeurs seront isolées dans certains nœuds. Cela est possible car les arbres CART sont des méthodes non paramétriques. Les algorithmes d'arbre CART peuvent gérer un grand nombre de variables sans pré-sélection préalable.

Dans sa recherche de variables pour obtenir une division avec le plus de différences sur la population, l'algorithme aura donc tendance à préférer la sélection sur les variables explicatives qualitatives nominales avec un grand nombre de modalités car elles offrent plus de souplesse dans la construction de deux sous populations différentes. Il y a $2^{m-1}-1$ manières de diviser une telle variable à m modalités en deux groupes. Si la variable est ordinale à m modalités, le nombre de division est $(m-1)$. Le cas des variables quantitatives est identique à celui des variables ordinales. Ces variables sont donc à utiliser avec précaution, Wikistat (2020b).

Lors de la première étape, appelé racine, l'arbre reprend toutes les observations de la base d'apprentissage. A chaque étape τ, $\tau \in \mathbb{R}_*^+$, une division est réalisée créant deux sous ensemble appelés fils gauche et fils droite, notés respectivement τ_G et τ_D. Pour choisir la variable sur laquelle effectuer la division, l'algorithme cherche à réduire l'hétérogénéité, notée D, des nœuds fils. Cela revient à effectuer, à chaque étape τ, le calcul suivant :

$$\max_{\text{divisions selon } X_j,\ j=1,\cdots,N} D_\tau - (D_{\tau_G} + D_{\tau_D}). \tag{3.12}$$

Le développement de l'arbre s'arrête si le nœud est homogène, si le nœud n'admet plus de division possible ou bien quand le nombre d'individus dans le nœud est inférieur à un seuil prédéterminé, généralement de 5. Le nœud devient alors une feuille.

Le calcul de l'hétérogénéité dépend de la nature de l'arbre. Si Y est quantitative alors l'hétérogénéité du nœud J est définie par :

$$D_\tau = \frac{1}{|\tau|}\sum_{i\in J}(y_i - \bar{y}_\tau)^2, \text{ avec } |\tau| \text{ l'effectif du noeud } \tau. \tag{3.13}$$

Si Y est qualitative avec m modalités plusieurs fonctions permettent de calculer l'hétérogénéité :

- Entropie : L'hétérogénéité est définie par : $D_\tau = -2\sum_{l=1}^m |\tau| p_\tau^l \log(p_\tau^l)$, avec la convention $0\log(0) = 0$ et p_τ^l la proportion de la modalité l dans le noeud τ.
- Concentration de Gini : L'hétérogénéité est définie par : $D_\tau = \sum_{l=1}^m p_\tau^l(1-p_\tau^l)$.

L'arbre ainsi construit est l'arbre maximal, il possède une grande variance et un biais faible, avoir trop de précision peut altérer la prédiction avec, par exemple, des zones de niches. Pour rendre la prédiction plus efficace, l'arbre peut être simplifié, c'est le principe d'élagage. L'objectif est de trouver un compromis entre l'arbre avec une seule feuille, dont la variance est faible mais avec un biais élevé, et l'arbre maximal.

L'interprétabilité est un des points fort de la méthode des arbres CART. Pour un nouvel individu donné, il est facile de déterminer sa valeur prédite en suivant les différents nœuds. Cet outil permet également de mettre en évidence heuristiquement les variables les plus utiles : plus le nœud est proche de la racine plus la réduction de l'hétérogénéité est importante. Mais cette approche n'est pas la plus pertinente et d'autres méthodes plus performantes seront utilisées pour chercher l'importance des variables. Cette méthode est simple à programmer, grace à un nombre réduit d'opérations élémentaires que doit réaliser un ordinateur pour obtenir un résultat : de l'ordre de $\circ(nN\log(n))$ pour un arbre équilibré à $\circ(Nn^2)$ dans le cas pire, avec N et n respectivement le nombre de variable explicative et le nombre d'individus de l'échantillon. Wikistat (2020c).

3.2.4 Forêt d'arbres décisionnels

Dans ses travaux sur les arbres CART, Breiman (2001) présente une nouvelle méthode : les forêts d'arbres décisionnels ou forêts aléatoires de l'anglais *randoms forests*. Une forêt aléatoire est l'agrégation d'un ensemble d'arbres de décisions réalisée à l'aide d'une méthode de rééchantillonnage et de sélections aléatoires de variables.

Ré-échantillonnage

Le rééchantillonnage, ou *bagging* (pour *bootstrap aggregation*) est une méthode introduite par Breiman (1996), permettant d'augmenter le taux de réussite de différentes méthodes de classification et régression des arbres CART. La principale critique des arbres CART était leur sensibilité aux échantillons de données sur lesquels ils réalisaient leurs apprentissages. A partir d'une base de données, cette méthode génére un nombre m d'échantillons tirés aux hasards avec remise sur la base d'origine. Sur chaque échantillon un arbre de décision est réalisé. Une fois tous les arbres obtenus, la prédiction finale est agrégée : soit par moyenne dans le cas des arbres de régression, soit par vote à la majorité dans le cas des arbres de classification. La variance est réduite en obtenant une combinaison de différents estimateurs indépendants, mais chaque arbre est construit sur un nombre plus faible de données. Cette méthode est parfois appelée bagging d'arbre CART.

Sélection aléatoire de variable

Dans sa version de 2001, Breiman a ajouté un choix aléatoire de variables pour chaque nœud de chaque arbre pour améliorer les méthodes de bagging d'arbre CART. Ainsi sur N variables explicatives, seulement ρ sont sélectionnées, avec $\rho \in [\![1, N]\!]$. Les arbres développés ne sont pas élagués et leur agrégation se fait comme en rééchantillonnage.

Dans le cas où $\rho = N$, la forêt aléatoire est simplement un bagging d'arbre CART et si $\rho = 1$ le choix de variable à chaque nœud est aléatoire. Ainsi dans les forêts aléatoires deux sources d'aléas sont présentes : dans le choix des sous échantillons et lors du choix des variables. Un choix courant est $\rho = \sqrt{N}$ en classification ou $\rho = N/3$ en régression.

Estimation de l'erreur de prédiction

Pour estimer l'erreur de prédiction sur les forêts aléatoires, la procédure *Out of Bag* (OOB) est souvent utilisée. Soit $\mathbb{X}_i$ les variables explicatives associées à l'observation Y_i. Pour chaque arbre ne contenant pas l'individu $(Y_i, \mathbb{X}_i)$, la procédure calcule la prédiction de la valeur de Y_i. La prédiction moyenne est notée $\hat{Y}_i$. L'erreur out of bag est alors :

- en régression : $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$.
- en classification : $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\hat{Y}_i \neq Y_i}$.

L'avantage de la méthode OOB par rapport à d'autres estimateurs est que cette celle-ci ne nécessite pas de découpage de l'échantillon d'apprentissage. Pour chaque individu testé, les arbres agrégés sont différents. L'erreur est donc une estimation de l'erreur de généralisation d'une forêt, elle n'utilise pas les prédictions de la forêt entière seulement une petite partie de la forêt.

3.2.5 Théorie de la crédibilité

Sur l'ensemble d'un portefeuille d'un assureur pour un risque donné, tous les assurés n'ont pas le même comportement. Certains correspondent à des bons risques : ils n'ont pas ou peu de sinistres et ceux-ci sont de coûts limités. Au contraire d'autres assurés sont des mauvais risques avec un nombre de sinistres supérieur à la moyenne, générant potentiellement chacun des coûts plus importants. Un portefeuille est toujours hétérogène, pour affiner la tarification il est nécessaire de déterminer plusieurs groupes d'assurés. Les modèles classiques ne sont toutefois pas applicables aux groupes réduits à quelques assurés (souvent les groupes extrêmes, bon ou mauvais). La tarification d'un contrat ne peut pas être construite uniquement sur l'historique de sinistres. Le principe de mutualisation ne serait pas respecté. La théorie de la crédibilité permet de pondérer l'historique personnel avec l'expérience de l'ensemble du portefeuille.

Si en 2018 la probabilité que les pompiers interviennent dans un logement est de 0,0084 (305 500 interventions pour incendie, Ministère de l'intérieur (2019) et 36 300 000 logements, INSEE (2019)). La probabilité pour un assuré donné d'avoir une intervention de pompier pour incendie dans l'année peut être plus élevée ou plus faible que la moyenne nationale selon son comportement. La théorie de la crédibilité permet de prendre en compte ces deux probabilités, individuelle et collective. Albert Withney a défini en 1918 la prime individuelle par :

$$P = \alpha X + (1 - \alpha)C, \tag{3.14}$$

où C et X sont respectivement l'expérience collective (construite sur l'historique de l'assureur) et l'expérience individuelle (construite sur l'historique de l'assuré). Le facteur de crédibilité α peut prendre différentes formes, Suru (2019).

3.2.6 Partitionnement

Pour valider les différents modèles qui seront mis en place, les bases de données doivent être partitionnées. Une première partie va servir à l'apprentissage afin de permettre au modèle de calculer les différents paramètres et une seconde partie va servir de test pour comparer les prédictions faites par le modèle. Utiliser des observations qui ne servent pas à paramétrer le modèle pour calculer leurs prédictions par ce dernier permet de se protéger du sur-apprentissage. Si l'erreur se réduit pour la base d'apprentissage mais augmente sur la base de validation, le modèle «sur-apprend» et perd en efficacité. En effet, si un modèle possède trop de liberté, il peut décrire parfaitement les données d'apprentissage mais n'a pas la capacité d'extrapoler les informations fondamentales. , Wikistat (2020a) et Robin Genuer (2017).

Pour les bases suffisamment grandes, une troisième partie de calibrage peut être utilisée pour déterminer les paramètres optimaux de certains critères comme le λ des régressions pénalisées. La répartition de la base dans le cas d'une division en trois groupes est de 70% pour la base d'apprentissage, 20% pour la base de calibration et 10% pour la base de test. Pour les bases de petites dimensions, seulement deux partitions sont réalisées : la base d'apprentissage et celle de test. L'avantage est de pouvoir conserver un plus grand nombre d'observations pour la base d'apprentissage. La répartition est alors de 80% pour la base d'apprentissage et 20% pour la base de test.

La validation croisée, souvent appelée *cross-validation* dans la littérature, est une méthode permettant de contrôler les capacités prédictives d'un modèle. Cette méthode est particulièrement adaptée aux bases de petites dimensions mais peut aussi être utilisée pour des bases de grande taille. L'échantillon original est divisé en k échantillons distincts, un des k échantillons est ensuite choisi pour être l'ensemble de validation et les autres les échantillons d'apprentissage. La table (3.2) représente les différents blocs pour un exemple où la base est divisée en trois. Le nombre de divisions est compris entre 1 et n, où n est le nombre d'individus dans l'échantillon initial. Si $k = n$, le test ne se fait que sur une seule observation.

Table 3.2: Répartion des données pour un validation croisée avec $k = 3$.

K	bloc 1	bloc 2	bloc 3
1	validation	apprentissage	apprentissage
2	apprentissage	validation	apprentissage
3	apprentissage	apprentissage	validation

Choix du partitionnement

Pour les bases d'incendies scrapés et de feux de végétation, un premier partitionnement 80/20 est réalisé pour obtenir une base de *design* et de *test*. Un second partitionnement 80/20 est réalisé sur la base de *design* pour obtenir une base de *train* et de *validation*. La base de *validation* est le symétrique de la base de *test* mais permet de paramétrer les modèles. La figure (3.5) représente la séparation de la base complète et les différents partitionnements.

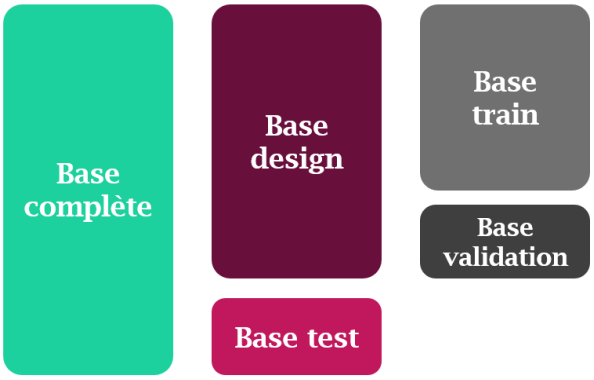


FIGURE 3.5: Représentation des différents partitionnements

Contrôle du partitionnement

Le taux de présence de la variable réponse, le nombre moyen d'incendie par commune, est étudié pour contrôler le partitionnement. Ce taux doit être le même dans la base test et dans une seconde base test tirée sur la base d'apprentissage avec la même méthodologie. Les valeurs obtenues sont présentées dans la table (3.13).

TABLE 3.3: Table des taux de fréquence observées de la variable réponse pour la base des incendies scrapés.

BASE	FRÉQUENCE ANNUELLE (%)
Base totale	0.001
Base design	0.001
Base test	0.001
Base train	0.001
Base validation	0.001

L'ensemble des différentes bases possèdent bien le même ordre de grandeur sur les différents taux de fréquence observés.

3.2.7 Contrôle de qualité

Pour évaluer la qualité de prédiction d'un modèle plusieurs critères existent. Pour obtenir un contrôle homogène et optimal, la même base de test doit être utilisée sur chaque modèle. Les prédictions alors réalisées sont comparées aux valeurs originales selon différentes formules, WIKISTAT (2020f).

- La racine de l'erreur quadratique moyenne un indicateur couramment utilisé. L'abréviation est $RMSE$ pour *Root Mean Square Deviation* :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widetilde{y}_i)^2}, \quad \text{où } \widetilde{y}_i \text{ est la valeur prédite pour l'observation } y_i. \tag{3.15}$$

- L'erreur absolue moyenne ou MAE pour *Mean Absolute Error* mesure la qualité de prédiction avec la moyenne arithmétique de la valeur absolue des écarts entre les prédictions et les valeurs initiales.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\widetilde{y_i} - y_i|, \text{ où } \widetilde{y_i} \text{ est la valeur prédite de } y_i \text{ par le modèle.} \tag{3.16}$$

- La déviance permet de mesurer l'apport de variables explicatives entre deux modèle i et j $(i \geq j)$:

$$D_i = 2(\mathcal{L}_i - \mathcal{L}_{sat}) \text{ où } \mathcal{L} \text{ est la vraisemblance et } sat \text{ le modèle saturé.} \tag{3.17}$$

- Le rapport de vraisemblance ou différence de déviance permet de mesurer l'apport de variables explicatives entre deux modèle i et j $(i \geq j)$:

$$D_i - D_j = 2(\mathcal{L}_i - \mathcal{L}_j) \text{ où } \mathcal{L} \text{ est la vraisemblance et } i \text{ et } j \text{ les modèles.} \tag{3.18}$$

3.3 Création d'un zonier

En assurance IARD, l'environnement du bien assuré est un critère important pour déterminer le montant de la prime. Pour les assurances de dommages, l'assuré n'est pas toujours responsable des dégâts qui peuvent être de sources variées. Pour un immeuble, le risque incendie ne dépend ainsi pas seulement d'un appartement mais de l'ensemble du bâtiment, Suru (2019).

Pour pouvoir déterminer un niveau de risque par zone géographique, une pratique courante est d'établir un zonier. L'objectif du zonier est double : améliorer la tarification et définir le risque géographique sur une variable unique (intégrée dans un GLM par exemple). Les données internes d'un assureur peuvent expliquer la sinistralité qu'il observe, malgré la présence éventuelle de quelques biais opérationnels :

- Exposition non uniforme : le portefeuille d'un assureur peut être concentré sur certaines zones géographiques et il peut alors manquer d'information sur d'autres régions.
- Manque de précision : pour pouvoir gagner des parts de marché, des assureurs produisent des zoniers peu précis pour couvrir l'ensemble du territoire et ainsi produire des estimations de tarif rapidement.
- Robustesse : ces zoniers de substitution sont peu détaillés et par leur faible exposition aux risques n'ont pas tendance à s'adapter aux évolutions de l'environnement qu'ils représentent.

3.3.1 Méthodologie

Un zonier mit en place sur un portefeuille d'assuré est réalisé en séparant les facteurs internes des facteurs géographiques. Les facteurs internes sont les données que l'assureur possède sur l'assuré, comme l'âge, la profession, la nature du bien couvert... Les facteurs géographiques sont l'ensembles des données sur l'environnement du bien, comme le nombre d'habitant ou l'altitude de la commune ou ce dernier est situé. Dans ce mémoire, deux modélisations du nombre d'incendies par commune sont réalisées. La première avec l'ensemble des variables disponibles (internes et externes) et la seconde avec seulement les variables internes. Les résidus sont la différence de prédiction entre les deux modèles. Ils sont supposés liés aux différentes variables géographiques.

Les résidus absolus (la différence entre la prédiction et l'observation) sont utilisés plutôt que les résidus relatifs (le rapport entre l'observation et la prédiction) pour prendre en compte la valeur de la différence entre les valeurs prédites et observées. En effet, avec une base comportant un grand nombre de valeurs nulles, diviser les valeurs observées par la prédiction n'a que peu de sens et une erreur même proche de zéro ne serait pas garante d'un modèle prédictif de qualité. Les résidus sont agrégés pour chaque commune (plusieurs métriques sont possibles : la moyenne, la somme) puis lissés.

Fonction de lissage

Dans son portefeuille, l'assureur ne dispose pas obligatoirement de contrat dans toutes les villes d'une zone étudiée, le lissage permet de prendre en compte la structure spatiale des données avec les valeurs des communes aux alentours. L'objectif du lissage est triple : déterminer une valeur de risque aux villes non exposées, réduire les imperfections des prédictions initiales et augmenter la robustesse du zonier en diminuant la dépendance avec le portefeuille initial.

La théorie de la crédibilité (3.2.5) est utilisée pour lisser les résidus. Soit r_i le résidu d'une commune i, d'après (3.14) alors le résidu lissé r_i^* est défini par la formule :

$$r_i^* = z(e_i)r_i + (1 - z(e_i))\frac{\sum_j e_j r_j f(d_{i,j})}{\sum_j e_j f(d_{i,j})}, \tag{3.19}$$

avec e_i et r_i respectivement l'exposition totale et les résidus initiaux moyen de la commune i.

La fonction f représente l'effet de la distance entre deux communes. Le risque incendie d'une commune est proche de celui des communes limitrophes (dans l'hypothèse d'un libre échange entre elles), néanmoins les effets entre communes se réduisent avec la distance. La fonction f est une fonction décroissante, les exemples suivants présentent différentes formes possible : $f(d_{i,j}) = \frac{1}{1+d_{i,j}}$,$f(d_{i,j}) = \frac{1}{b^n+d_{i,j}^n}$ ou $f(d_{i,j}) = \exp(-nd_{i,j})$.

Et $z(e_i) = \left(\frac{e_i}{e_i+a}\right)^m$ la fonction de crédibilité, qui est croissante avec l'exposition de la commune. Le paramètre m, $m > 0$, est la puissance de la crédibilité et a la paramètre d'ajustement de crédibilité.

Cette fonction de lissage par distance est facilement interprétable, avec une mise en œuvre facile, même si le coût de calcul peut être important. Deux problèmes principaux ressortent de l'utilisation de cette méthode :

- Les obstacles naturels comme les montagnes, rivières ou artificiels comme les ponts ne sont pas pris en compte. Seule la distance intercommunale intervient, une île sera ainsi influencée par les villes de la côte la plus proche.
- La nature des communes (rurale ou urbaine) n'est pas prise en compte dans le lissage.

L'utilisation de table de continuité est une option de lissage possible. Ces tables représentent les obstacles entre chaque commune et les moyens de communications disponibles.

Calcul de la distance $d_{i,j}$ La distance $d_{i,j}$ entre deux communes i et j est calculée avec leurs coordonnées géographiques. Chaque ville est située par sa latitude et sa longitude. L'étude étant réalisé seulement sur la France métropolitaine, le rayon de la terre est supposé constant. Avec cette hypothèse sa valeur est de 6371 kilomètres. Le site data.gouv.fr fournit pour chaque commune, un couple de coordonnées représentant le "centre" géographique de la commune, c'est à dire, le barycentre par rapport aux limites définies par le découpage administratif. Ce point ne correspond pas forcement au centre-ville, surtout dans les communes rurales étendues. La figure (3.6) représente la distance entre deux points dans un système de coordonnées sphériques où chaque point est représenté par deux angles et une distance à l'origine. La distance $d_{i,j}$ est donc définie par la formule suivante :

$$d_{i,j} = R \times \arccos\left[\sin\varphi_i \times \sin\varphi_j + \cos\varphi_i \times \cos\varphi_j \times \cos\Delta\lambda\right] \text{ où } \Delta\lambda = \lambda_i - \lambda_j, \tag{3.20}$$

avec $R = 6371$ km, φ_i et λ_i respectivement la latitude et longitude de la commune i.

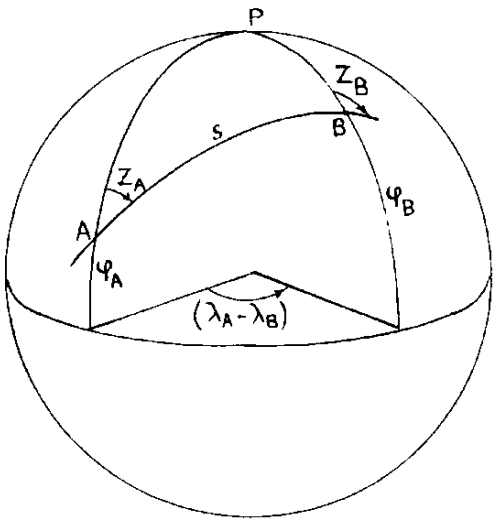


FIGURE 3.6: Représentation du calcul de la distance entre deux communes. Source image : IGN (2020b)

Les annexes (A.6a) et (A.6b) montre l'intérêt d'appliquer un lissage sur des valeurs prédites. Ces annexes représentent le niveau de risque (7 classes) estimé pour les communes de l'Essonne avec un lissage faible et un lissage important. Les frontières entre modalités sont plus nettes avec le second lissage, ce qui permet en tarification de ne pas avoir d'écart important entre deux communes limitrophes.

Classification des résultats

Une fois les résidus lissés, la valeur du risque pour chaque commune est une variable continue. La classification permet de regrouper l'ensemble des valeurs en un nombre fini de modalités et ainsi de rendre implémentable la variable en production. Réduire la granularité permet de rendre le zonier plus robuste, diminuer la complexité algorithmique et permettre un affichage visuel avec une carte choroplèthe. Un algorithme de clustering est appliqué aux valeurs du risque lissées pour les regrouper en différentes classes. L'algorithme du *k-means* est adapté à la problématique de séparation d'un vecteur en différentes classes.

K-means Les algorithmes d'agrégation autour de centres mobiles permettent de regrouper les individus qui sont semblables en un nombre de classes déterminé *a priori*. L'algorithme est initialisé avec une liste de centres (choisit aléatoirement ou non) et à chaque itération, tous les individus sont affectés au le centre le plus proche selon une distance déterminée pour former une classe. Le centre des nouvelles classes est alors déterminé et la distance peut être calculée de différentes manières :

- Distance euclidienne $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.
- Distance de Manhattan $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$.
- Distance de Tchebychev $d(x, y) = \max_i |x_i - y_i|$.

La variance totale expliquée par la clusterisation permet d'estimer la qualité de la clusterisation. Quand la variance totale expliquée atteint 100%, le nombre de classe est égal au nombre de modalités initiales. La méthode du coude peut être appliquée pour déterminer différentes possibilités de clusterisation.

Chaque choix de clusterisation sera testé sur la base de validation avec le modèle interne pour déterminer le nombre final de clusters. Une fois le nombre de modalités fixé, la variable continue expliquant le risque géographique est transformée en une variable qualitative à plusieurs facteurs représentant le risque.

La figure suivante (3.7) résume le processus de mise en place du zonier :

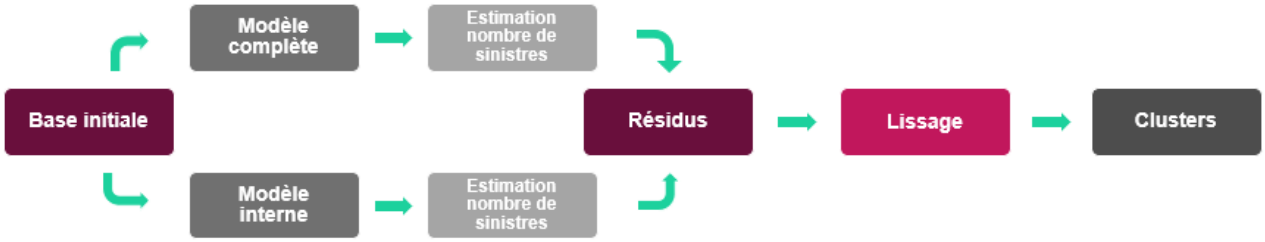


FIGURE 3.7: Schéma de mise en place d'un zonier sur une base de données

3.3.2 Contrôle des modélisations

Cette partie présente l'utilisation des différentes méthodes de la section (3.2). Les techniques de sélection de variables et de contrôle de prédiction sont des outils que les équipes de tarification utilisent pour déterminer les facteurs révélateurs de la sinistralité. Les équipes opérationnelles doivent pouvoir identifier les éléments pouvant impacter la sinistralité d'un contrat.

Les différents modèles cherchent à estimer le nombre d'incendies pour chaque commune de France. Ce nombre dépend de l'exposition de la commune au risque, et dans le cas de cette étude, l'exposition est représentée par le nombre de logements. Le risque de chaque commune est le rapport entre le nombre d'incendies et le nombre de logements. Sinon une commune urbanisée risque d'avoir plus de logements incendiés qu'une commune rurale. Pour corriger ce problème, le paramètre `offset` permet de forcer le coefficient β_i associé au logarithme de la variable exposition à 1, (3.8). Avec une fonction de lien log, les prédictions du nombre d'incendie peuvent être divisées par l'exposition, CHARPENTIER (2013).

L'historique entier des données scrapées depuis 2012 n'est pas complètement exploitable. L'étude des séries temporelles (2.7a),(2.7), montre en effet que le nombre d'incendies se stabilise après 2017. Le choix de cette année permet aussi à la variable historique 5 ans d'être utilisable. Cette hypothèse permet de ne pas prendre en compte la variable année dans le modèle, et de pouvoir échantillonner la base de test.

L'ensemble des variables descriptives ne peut pas être utilisé car il est trop important. Une première réduction est réalisée manuellement en supprimant les variables redondantes et en regroupant certaines variables quand cela est possible. Par exemple, les six variables renseignant le nombre de logements selon l'année de construction sont regroupées en deux variables (respectivement le nombre de logements construits avant et après 1990). Les variables surface et population sont transformées en une seule variable densité de population. Le nombre de variables est ainsi presque réduit par deux avec cette première démarche, mais reste importante pour un GLM.

Parmi les différents historiques possible (1,3 et 5 ans), seul l'historique 5 ans est conservé. Cette variable, qui pour une commune donnée, somme le nombre d'incendie dans la commune est corrélée avec le nombre d'interventions dans l'année. La variable historique est donc divisée par le nombre de logements (l'exposition).

L'étape de modélisation permet de contrôler les différentes opérations effectuées précédemment sur la base de données qui peut contenir des raccourcis ou des erreurs. L'objectif est d'analyser les variables principales des modèles pour s'interroger sur leurs importances relatives. Pour cela des indicateurs et des sorties graphiques sont édités et analysés.

Modèle complet

La régression pénalisée est testée sur l'ensemble des variables de la base pour déterminer le modèle discriminant le moins de variable. Le choix de la valeur du paramètre d'apprentissage λ est un sujet complexe. L'utilisation de la validation croisée permet d'obtenir deux valeurs. La première, λ_{min}, est la valeur de lambda permettant d'obtenir la déviance moyenne estimée la plus faible, et la deuxième valeur, λ_{1se}, représente le modèle avec une erreur standard sur la déviance.

Pour chaque régression pénalisée, et chaque valeur de λ, la qualité de prédiction des modèles est testée sur la base de validation. Les différents critères de contrôle sont résumés dans la table suivante.

Table 3.4: Table des critères de sélection du meilleur modèle pour le zonier direct.

Méthode	RMSE	MAE	Déviance
Lasso, λ_{min}	0,4649	0,089	4941
Lasso, $\lambda_{1.se}$	0,4998	0,083	6228

Le Lasso avec le λ_{1se} est radical en ne conservant que 2 variables : la population de la commune et le nombre de pièces des résidences principales. Après comparaison des déviances respectives, le modèle est moins bon que dans le cas du λ_{1se}, la sélection est réalisée avec λ_{min}. Les variables principales de ce modèle sont listées dans la table suivante (3.5), avec la valeur du coefficient de régression pénalisée.

Table 3.5: Table des variables les plus importantes du modèle complet.

Variable	Coefficient
Le pourcentage d'appartements construit avant 1990	0,51
L'historique d'incendie de la commune	0,11
Le pourcentage d'appartements	0,06
Le pourcentage de résidences principales avec le chauffage central	0,06
Le rapport du nombre de ménages sur la population de la commune	0,03
Le pourcentage de résidences principales HLM dans la commune	0,02
Le nombre de ménages ayant emménagé depuis plus de 2 ans dans leurs logements	0,01
Pourcentage de la surface de la commune converti en surface artificialisée	0,01

Les pourcentages précédents (3.5) sont par rapport au nombre total de logements dans la commune (hors pourcentage de surface artificialisée).

Le chauffage central est un mode de chauffage où une seule source de chaleur permet de chauffer les différentes pièces du logement. Il peut être individuel ou collectif (généralement pour un immeuble). Pour les variables les plus importantes, un contrôle des prédictions est réalisé à l'aide de sorties graphiques qui représentent pour chaque modalité d'une variable donnée, la fréquence d'incendie observée et celle prédite. Les variables étant continues, pour permettre une lecture plus évidente, les différentes modalités sont regroupées en un nombre réduit selon la distribution de ces modalités par rapport à leur exposition.

Sur ces sorties, différents problèmes sont identifiés, dont notamment la capacité de prédiction qui présente du sous et du sur apprentissage. Ces difficultés nécessitent une étude approfondie des variables.

Pour contrôler le choix des variables internes, un *random forest* est appliqué. Les variables qui participent le plus à la réduction du carré de l'erreur moyenne sont représentées dans la figure suivante (3.8).

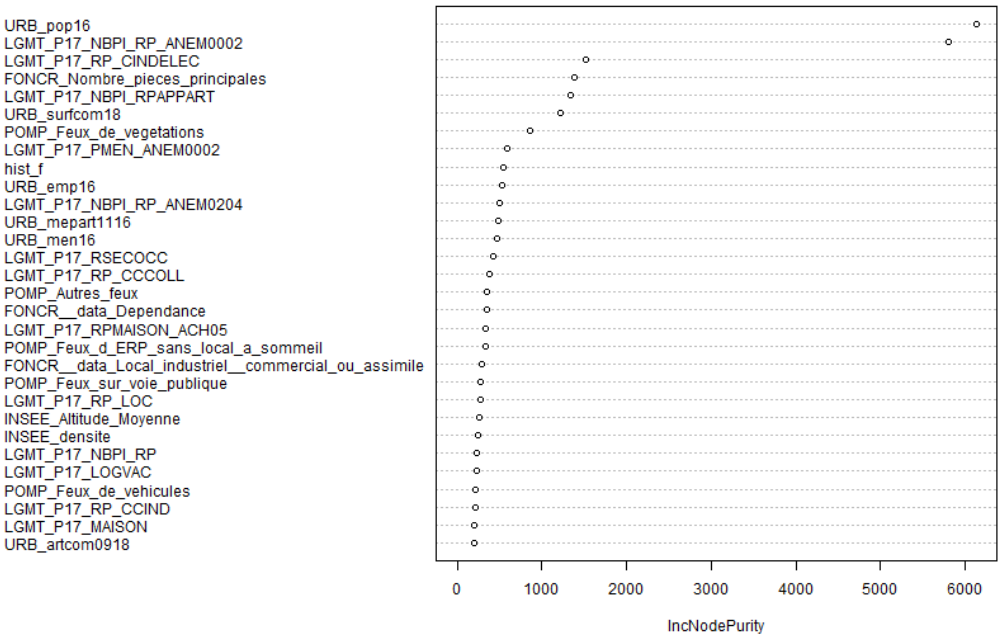


FIGURE 3.8: Sélection de variables avec *random forest*

L'axe des abscisses *IncNodePurity* représente l'importance de la variable selon l'indice de diversité de Gini. Plus la valeur de décroissance moyenne de la précision est grande plus la variable est importante pour le modèle.

Les variables d'importances ne sont pas les mêmes qu'avec la sélection par Lasso. La population de la commune et le nombre de ménages ayant aménagé depuis moins de 2 ans sont les variables les plus importantes. Néanmoins quatre autre variables se détachent sans atteindre le niveau d'importance des deux premières, ce sont le pourcentage de logements avec le chauffage central et le nombre de pièces principales, le nombre de pièces des appartements ainsi que la surface artificialisée. L'historique de la commune n'est que la $9^{ème}$ variable la plus significative au modèle de prédiction.

Même si la sélection de variable ne se fera pas avec les prédictions du *random forest*, pouvoir comparer différents modèles est indispensable pour s'assurer de ne pas oublier une variable. Même si les variables sélectionnées ne sont pas identiques, le type d'informations conservé est proche (chauffage centrale, urbanisation ...). Les GLM ne prennent pas en compte tous les effets locaux des variables. L'apprentissage local du *random forest* permet de prendre en compte ces variations locales, et ainsi de rendre plus robuste la sélection de variables.

Modèle interne

Pour pouvoir réaliser un zonier sur les résidus comme présenté en (3.3.1), l'ensemble des variables explicatives de la base doit être séparé des facteurs géographiques. Les sites de devis en ligne de contrat MRH permettent de souscrire un contrat rapidement en renseignant différentes informations. Ces interrogations représentent les variables internes. Pour un assureur en ligne, les différentes questions posées concernent :

- la nature du bien : maison, appartement, autre (mobil-home, grande demeure, ...) ;
- le nombre de pièces, le nombre de pièces principales et combien font plus de 40 m^2 ;
- l'âge du bien : plus ou moins de cinq ans ;
- la surface exacte ;
- la présence de cheminés, de voitures ;
- la nature de l'occupant : propriétaire ou locataire ;
- la nature de l'utilisation : résidence principale ou secondaire et le taux d'occupation : inférieur ou supérieur à 75% ;
- l'adresse du bien, l'étage ;
- une description de l'occcupant : nom, prénom, âge, profession ;
- le nombre de sinistres déclarés dans les deux dernières années ;
- les caractéristiques du contrat actuel.

Les variables connues de l'assureur représentent le logement et son occupant. L'ensemble des facteurs de la base de données logement de l'INSEE sont donc supposés comme les données internes. Le nombre d'incendies, et les historiques des communes sont aussi conservés dans le modèle interne pour représenter le passif de l'assuré. L'âge et la profession de l'occupant sont des variables non représentées dans la base de données. Les caractéristiques du contrat actuel sont supposées servir à une veille concurrentielle et n'interviennent pas dans le processus de tarification.

La modélisation du nombre d'incendies par commune est réalisée seulement avec les variables internes. L'utilisation de la régression pénalisée permet de réduire le nombre de variables pour déterminer le meilleur modèle. Les caractéristiques des différents modèles sont résumées dans la table suivante.

Comme pour le modèle complet, la sélection par Lasso avec $\lambda_{1.se}$ est trop restreinte en ne sélectionnant qu'une seule variable : le nombre de pièces des résidences principales occupées depuis moins de 2 ans.

TABLE 3.6: Table des critères de sélection du meilleur modèle pour le zonier sur les résidus.

MÉTHODE	RMSE	MAE	DÉVIANCE
Lasso, λ_{min}	0,4643	0,08	5078
Lasso, $\lambda_{1.se}$	0,4903	0,08	6115

Le modèle conservé est celui réalisé avec λ_{min}, dont la qualité de prédiction est meilleure sur l'ensemble des critères utilisés. Les variables principales de ce modèle sont résumées dans la table suivante, avec la valeur du coefficient de la régression pénalisée.

TABLE 3.7: Table des variables les plus importantes du modèle interne.

VARIABLE	COEFFICIENT
Pourcentage de résidences principales qui possèdent au moins une voiture	0,30
Pourcentage d'appartements dans la commune	0,20
L'historique d'incendie de la commune	0,14
Le pourcentage de logements vacant dans la commune	0,12
Le pourcentage de dépendances dans la commune	0,07

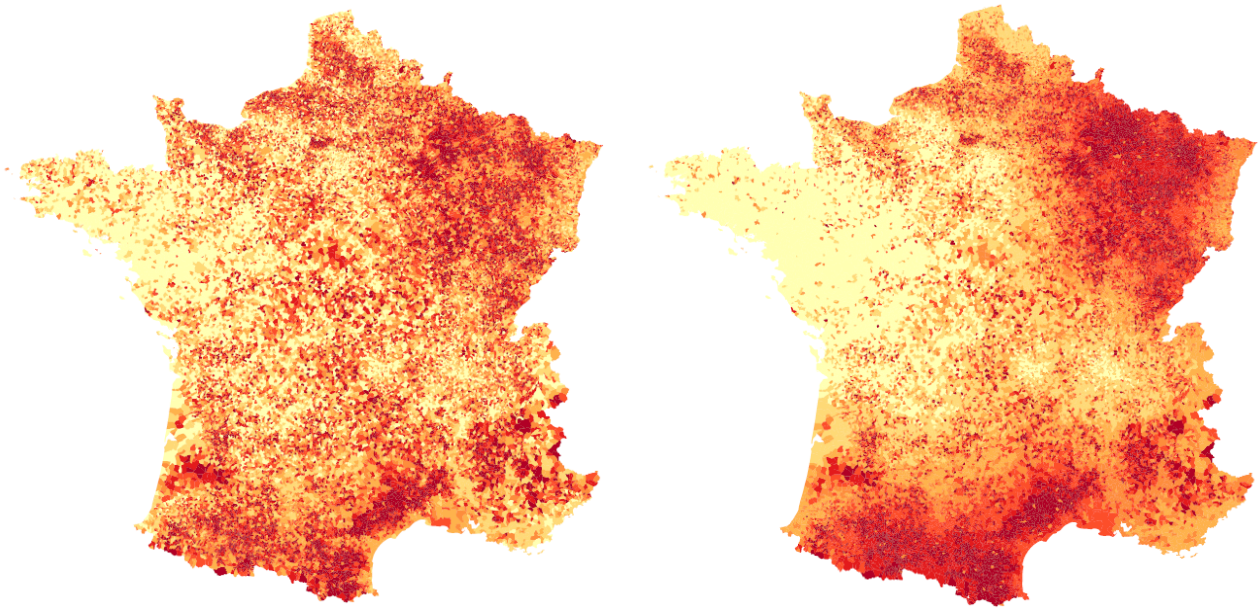
L'ordre de grandeur du coefficient d'importance des variables varie peu. Parmi les variables sélectionnées aucune ne se démarque significativement des autres.

3.3.3 Zonier

Une fois les prédictions réalisées avec le modèle complet et le modèle interne, une différence entre les prédictions peut être calculée pour chaque couple de valeur commune/année. Le résidu moyen est ensuite calculé pour chaque commune. Le résultat obtenu est ensuite lissé comme présenté dans l'équation (3.19).

Présentation des résultats

L'ensemble des figures suivantes présentent des cartes du risque de survenance d'un incendie pour chaque commune de France pour différents lissage. Le niveau de risque est réparti selon 7 niveaux de couleur. Plus la couleur est claire plus le risque est faible et quand la couleur tend vers le rouge, le risque est important. Toutes les cartes sont représentées avec le même nombre de niveau de risque pour analyser visuellement le lissage.



(a) Zonier lissé, $a = 2$, $m = 1$

(b) Zonier lissé, $a = 5$, $m = 2$

FIGURE 3.9: Zoniers de la fréquence incendie pour l'ensemble des communes en France pour différents lissages.

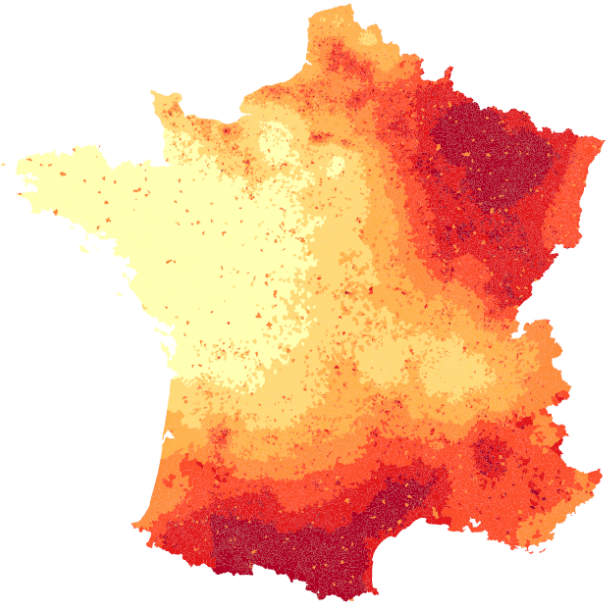


FIGURE 3.10: Zonier lissé, $a = 10$, $m = 3$

Les régions montagneuses possèdent un fort risque incendie, ainsi que le Nord de la France. La région Grand Est est assez disparate au niveau de son risque, du fait de sa faible densité de logement (3.3).

Le lissage n'est pas optimal, certaines communes à fort risque sont limitrophes de communes à faible risque. La procédure de lissage étant empirique et assez longue en temps de calcul, le choix a été pris de conserver une version satisfaisante (3.10) mais non parfaite pour continuer la démarche. Le contrôle a été réalisé de manière visuel, mais une méthode optimale est de mesurer pour chaque zonier lissé, en plus du contrôle visuel, la qualité de prédiction de ce dernier sur un portefeuille d'assureur, et d'arreter lorsque le lissage est cohérents entre communes voisines et n'apporte plus de gain significatif sur les prédictions.

Un zonier a aussi un intérêt dans la tarification des contrats, il ne faut pas que l'écart de prix entre deux communes limitrophes soit trop important. Dans les zones de fort risque, il reste des communes à risque moyen qui pourraient bénéficier d'un passage manuel au risque le plus fort pour améliorer la continuité du risque.

Classification

Pour déterminer le nombre de cluster la méthode du coude est appliquée sur le pourcentage de la variance déterminé par la clusterisation avec la méthode des *k-means*. Considérant un seul vecteur de risque lissé, les différentes formules de calcul de la distance permettant d'évaluer les différentes classes donne un même résultat. La figure (3.11) présente le pourcentage de la variance expliquée pour chaque cluster possible.

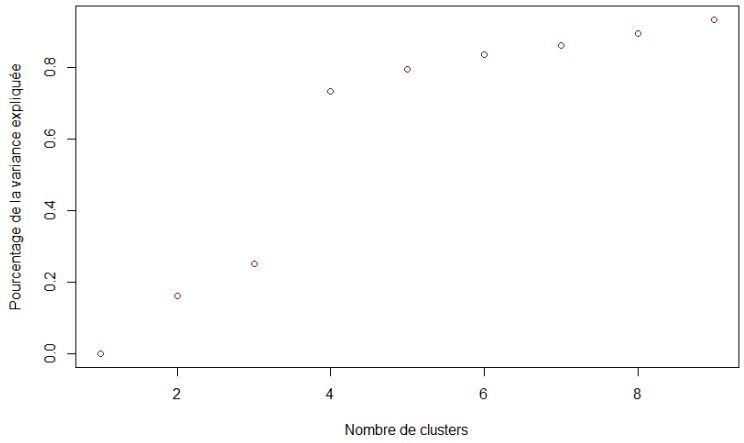


FIGURE 3.11: Le pourcentage de la variance expliquée selon le nombre de cluster

Il est nécessaire de retenir au moins de 4 clusters pour avoir un gain significatif dans le pourcentage de la variance expliquée. Les différentes possibilités : 3,4,5,6,7 sont testées sur la base de validation avec le modèle interne et les valeurs de prédictions sont résumées dans la table (3.8).

TABLE 3.8: Comparatif de la qualité de prédiction sur la base de validation des données.

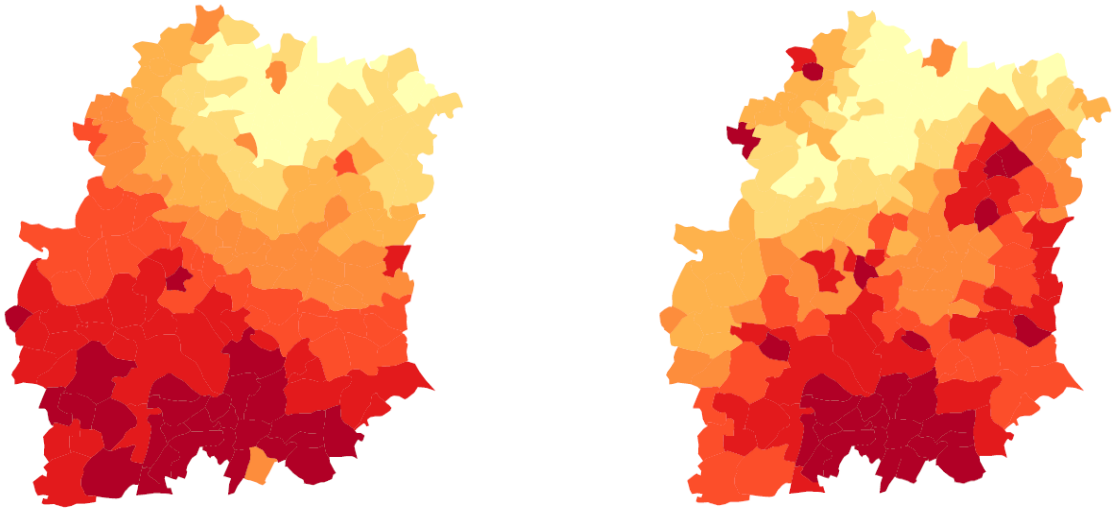
NOMBRE DE CLUSTER	RMSE	DEVIANCE
3	0.5650	5754,16
4	0,5618	5752,43
5	0,5634	5752,54
6	0,5618	5752,96
7	0,5652	5753,05

L'ordre de grandeur de variation entre les différents critères de contrôle de la qualité de prédiction est faible. L'étude de la table (3.8) montre que la prédiction la moins bonne a lieu lorsque le nombre de clusters est de 3. Pour un nombre de clusters supérieur au coude (3.11), la déviance est monotone avec la granularité du zonier. Ainsi, le nombre de 4 clusters est retenu.

Zoniers Essonne

La méthode mise en place pour réaliser un zonier à l'aide d'un modèle complet et d'un modèle interne est reproduite avec la base du SDIS sur les interventions pour incendie en Essonne. Cette table est supposée complète, c'est à dire comprenant l'ensemble de la variable réponse, le nombre d'intervention. Cette opération permet de contrôler les zoniers à l'échelle nationale, en comparant le département de l'Essonne. Cependant la fréquence d'incendie selon l'exposition dans la base du SDIS (nombre de logements) est de 1%, ce qui est plus important, les résultats attendus ne seront donc analysés que visuellement.

Les différentes tables qui résument les critères de sélection sont présentent sur l'annexe (A.2). Pour que le zonier en issu des données scrapées ne soit pas influencé par l'ensemble des communes exterieure à l'Essonne, un lissage est réalisé seulement entre les communes du département comme pour les valeurs du zonier réalisé avec la base du SDIS.



(a) Zonier fréquence incendie base SDIS

(b) Zonier fréquence incendie données scrapées

FIGURE 3.12: Zoniers de la fréquence incendie pour l'ensemble des communes en Essonne pour deux bases de données différentes avec le même lissage.

Les deux figures représentent avec sept niveaux de risque (croissant lorsque la couleur tend vers le rouge) le risque incendie en Essonne. Dans les deux figures, le risque est plus important dans le sud du département qui est plus rural que le nord qui est plus proche de Paris. Cependant dans l'est du département le risque apparaît moins bien modélisé par la base de données d'incendies scrapés.

Réussir à établir un zonier sur la fréquence incendie sans donnée d'un assureur est une première étape, mais ne renseigne pas sur la qualité des informations qu'il contient. L'utilisation de données issues d'un portefeuille d'assureur permet d'évaluer le gain prédictif de ce dernier sur la base de données concrètes.

3.4 Application du zonier sur un portefeuille d'assureur

Pour tester le zonier, celui-ci sera comparé avec un zonier déjà existant dans un portefeuille MRH d'assureur. Cette base de données de plus de 1 200 000 contrats sur quatre années détaille les caractéristiques des habitations assurées ainsi que le nombre de sinistres associés.

Ce portefeuille contient sur une seule base, l'ensemble des polices et des sinistres. Les sinistres étant séparés selon leur nature (vol, dégât des eaux, responsabilité civile, incendie), cette base permet l'étude de la fréquence des incendies. La fréquence d'incendies est de 0,0054 soit du même ordre de grandeur que la valeur du marché FFA.

3.4.1 Méthodologie

La base a été partitionnée en 4 comme présenté dans la figure (3.5). Pour évaluer le zonier construit sur la base des incendies scrapés sa qualité prédictive sera comparée au zonier déjà présent dans le portefeuille. Pour cela, deux GLM sont appliqués sur la base en différenciant uniquement la variable zonier.

Les variables descriptives conservées de cette base sont :

- l'exposition,
- la superficie,
- le nombre d'occupants du logement,
- le nombre de pièces du logement,
- l'âge du souscripteur,
- l'année de construction du logement,
- les antécédents de sinistre incendie,
- l'utilisation du bien : résidence secondaire ou principale,
- l'étage du logement,
- la nature de l'occupant : propriétaire ou locataire,
- le zonier incendie.

Seules les variables pouvant expliquer la fréquence sont conservées. Les différentes modalités sur les montants ne sont pas étudiées dans ce mémoire.

La figure suivante présente la distribution de l'exposition dans le portefeuille. Des contrats sont souscrit et rompus tous les jours de l'année. Néanmoins, l'exposition est croissante avec le temps, le portefeuille est en période de croissance avec un progression de son nombre de contrats.

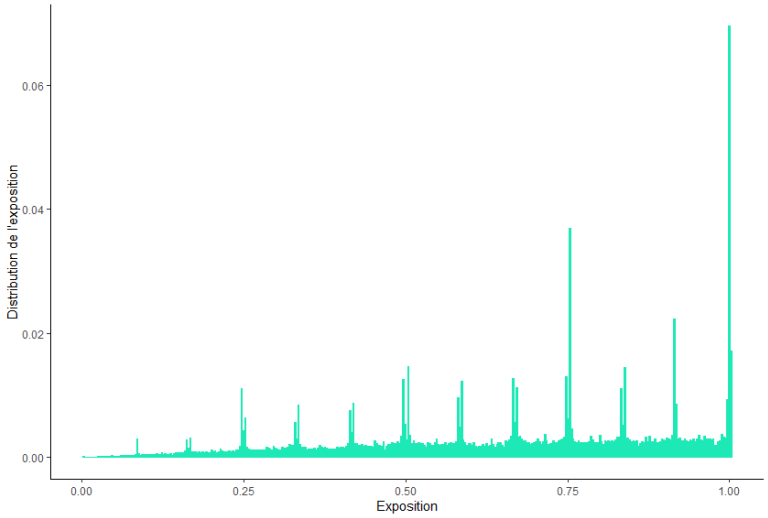


FIGURE 3.13: Distribution de l'exposition pour la base d'assureur

Pour des raisons de confidentialité l'étude statistique descriptive de l'ensemble des variables ne sera pas décrite dans ce mémoire. Le portefeuille contient majoritairement des logements construites avant 1980. Le profil du souscripteur moyen est une personne de 46 ans vivant en couple dans une habitation de 4 pièces et d'une surface de 100m^2.

La variable du nombre de sinistres a une espérance de **0.0025** et une variance de **0.0026** . L'utilisation d'une loi de Poisson dans le GLM et la sélection de variable sont donc pertinentes.

Le zonier incendie transmis par l'assureur contient 5 zones. La table suivante détaille pour chaque modalité du zonier, le pourcentage d'exposition total représenté par cette dernière.

TABLE 3.9: Répartition de l'exposition dans les modalités du zonier.

MODALITÉ	P.C DU NOMBRE DE CONTRATS	P.C DE L'EXPOSITION TOTALE	P.C DU NOMBRE DE SINISTRES
Z1	27	28	42
Z2	37	37	31
Z3	25	25	23
Z4	2.5	2.5	1
Z5	8.5	7.5	3

P.c : pourcentage.

Les trois premières modalités sont majoritaires et représentent presque 90% de l'exposition et du nombre de contrats et 96% des sinistres. La répartition entre le nombre de contrats et l'exposition sont équivalentes. La signification des modalités n'est pas spécifiée. La figure suivante représente les différentes modalités du zonier en France métropolitaine pour l'ensemble des contrats du portefeuille.

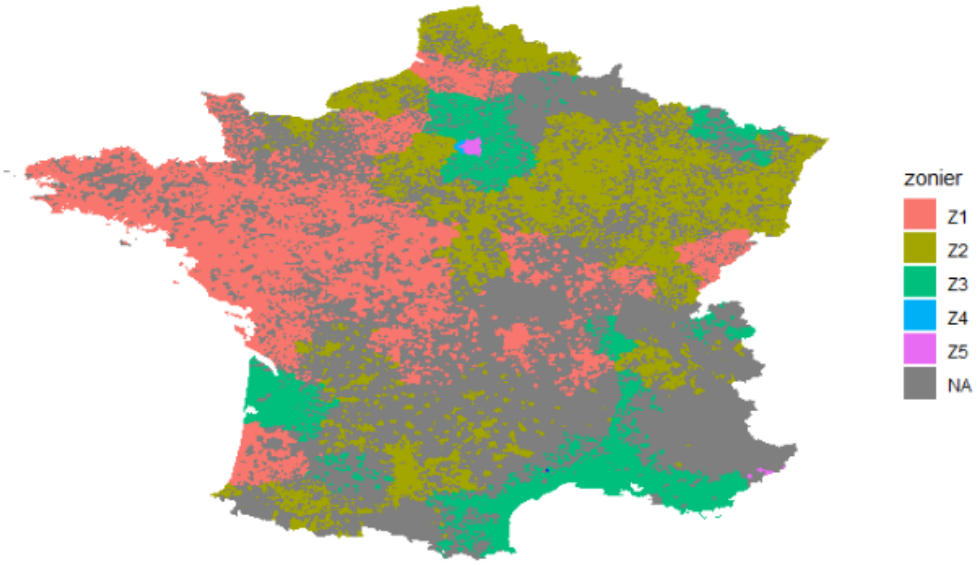


FIGURE 3.14: Zonier incendie présent dans le portefeuille d'assureur

Les modalités sont globales et graphiquement distinctes, il y a des frontières nettes entres elles. L'ensemble du territoire n'est pas couvert dans le portefeuille comme une partie des Alpes ou des Pyrénées. C'est dans cette situation qu'un zonier réalisé sur l'ensemble du territoire permet de gagner des parts de marché sur les régions avec peu d'exposition en maitrisant le risque de fréquence sur la garantie incendie.

Un contrôle du partitionnement est ensuite effectué comme sur la base des données scrapés. Les valeurs sont résumées dans la table suivante.

TABLE 3.10: Table des taux de réponse observées sur la base assureur relatifs à celui de la base totale.

BASE	COEFFICIENT
Base totale	1
Base design	1
Base test	1
Base train	0,98
Base validation	1,02

La figure suivante représente le zonier à quatres modalités pour le risque incendie qui sera opposé au zonier présenté dans la figure (3.15). Les différentes modalités représentent le niveau de risque par ordre croissant :

- L : risque faible,
- M1 : risque moyen 1,
- M2 : risque moyen 2,
- H : risque élevé.

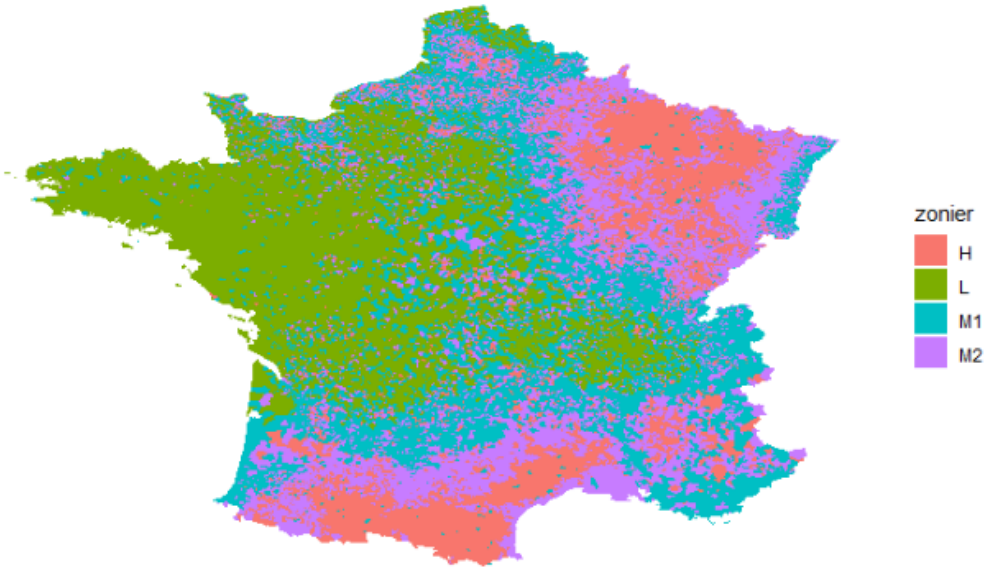


FIGURE 3.15: Zonier incendie à 4 modalités après lissage

Les régions Grand Est et Occitanie sont les régions de France qui présentent le risque de survenance d'un incendie par logement le plus fort. L'ensemble de l'ouest de la France est quant à elle la région présentant la fréquence la plus faible.

Pour les deux zoniers, une sélection de variables par Lasso est paramétrée sur la base *train* et validée sur la base *validation*. Un GLM est paramétré sur les variables les plus importantes ainsi que la variable représentant le risque géographique. Les prédictions sont réalisées sur la base *test* puis comparées avec les valeurs observées.

3.4.2 Présentation des résultats

La sélection de variables est réalisée avec la méthode de Lasso, le choix de la valeur de λ est résumé dans la table suivante (3.11).

TABLE 3.11: Table des critères de sélection du meilleur modèle sur le portefeuille d'assureur.

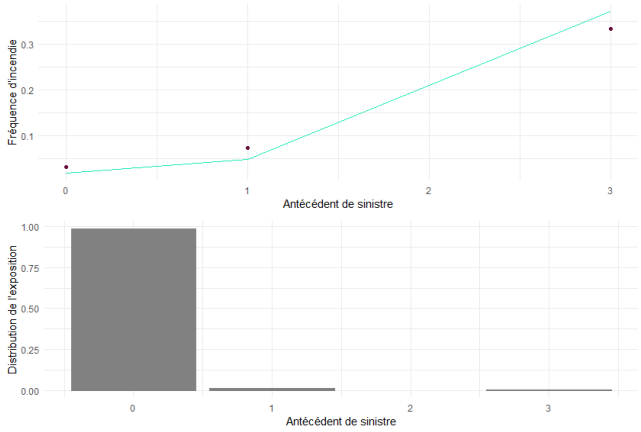
MÉTHODE	RMSE	MAE	DÉVIANCE
Lasso, λ_{min}	0,05	0,005	5948
Lasso, $\lambda_{1.se}$	0,05	0,005	6100

Le modèle conservé est la sélection avec λ_{min} pour la réduction de variance plus importante. L'ensemble des variables sont sélectionnées quelque soit la valeur de λ. La table suivante (3.12) présente les plus importantes.

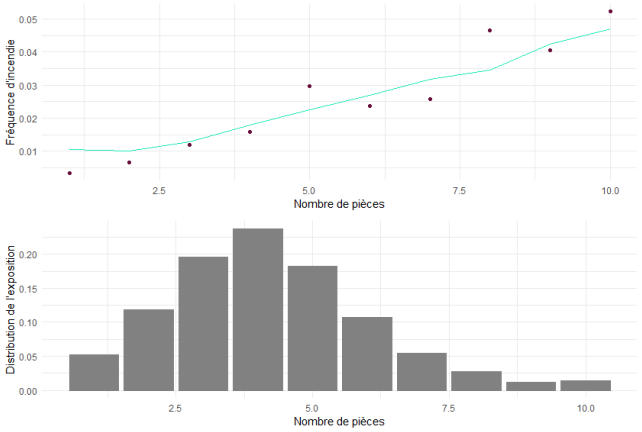
TABLE 3.12: Table des variables les plus importantes du portefeuille d'assureur

VARIABLE	COEFFICIENT
Antécédent de sinistre dans le logement	0,67
Nombre de pièces dans le logement	0,10
Nombre d'occupants du logement	0,06
Superficie du logement	0,01

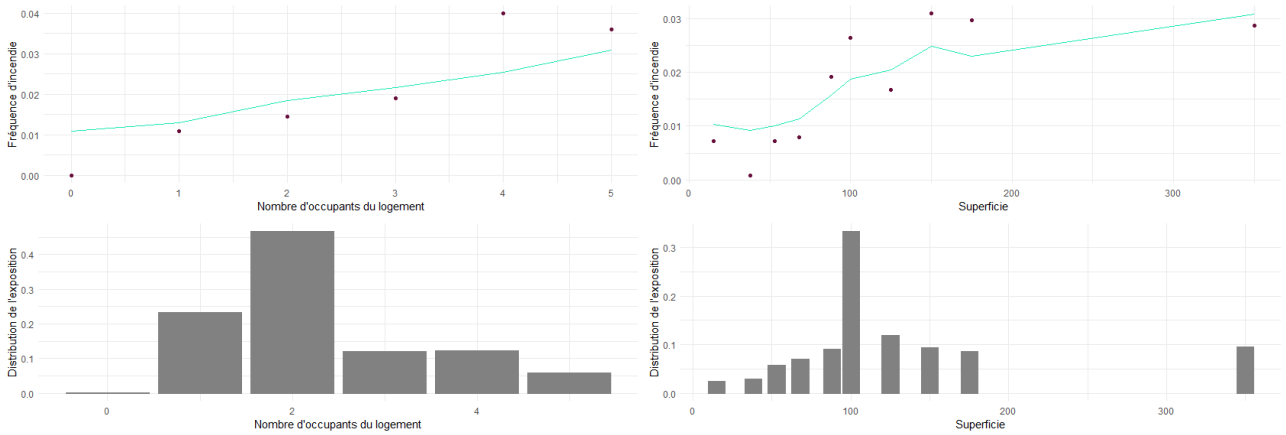
Pour les variables les plus importantes les qualités de prédiction du modèle sont contrôlées. Ainsi pour toutes les modalités de variable, la fréquence de sinistre observée sur la base de test (point bordeau) est comparée à la la fréquence prédite (ligne verte). Les prédictions sont accompagnées de la distribution des modalités de la variable



(a) Historique de sinistre de l'assuré



(b) nombre de pièces du logement



(a) nombre d'occupants du logement (b) superficie du logement

La majorité des contrats n'ont pas d'antécédent de sinistre incendie. La fréquence d'incendie augmente si le nombre d'antécédent est important. Cette variable étant quantitative la modalité 3 représente tous les contrats ayant plus d'un sinistre dans le passé.

La fréquence d'incendie augmente avec le nombre de pièces et malgré l'exposition plus faible des logements avec beaucoup de pièces la prédiction est bonne. Les conclusions sont les mêmes pour la variable nombre d'occupants du logement mais les prédictions sont moins bonnes avec une sur-estimation du risque sur les premières modalités et une sous-estimation sur les dernières.

La fréquence d'incendie n'est pas linéaire avec la surface du logement, les valeurs estimées sont assez mauvaise. Le GLM n'est pas l'outil idéal pour modéliser ce lien, une source d'amélioration serait l'application d'un GAM.

Comparaison des deux zoniers

La table suivante présente les critères de qualités des prédictions sur la base de test pour les deux zoniers.

TABLE 3.13: Evalution de la qualité de prédicton des modèles selon le zonier choisit.

ZONIER	MAE	RMSE	DÉVIANCE
Zonier de l'assureur	0,005063	0.051230	5946,01
Zonier sur les données scrapées	0,005062	0,051232	5945,72

Le zonier sur les données scrapées est théoriquement meilleur mais le gain de prédiction n'est pas significatif avec une réduction de 0,005% de la déviance. Les différences de valeurs rentrent dans la marge d'erreur, si les résultats sont refaits avec une nouvelle graine aléatoire, le zonier sur les données scrapées ne serait peut être plus le meilleur théoriquement.

La qualité de ces résultats peut être expliquée par les différentes limites présentes dans la démarche utilisée dans ce mémoire.

Utilisation des feux de végétations

La même méthodoligie que présentée dans la section (3.4) pour mettre en place un zonier et tester sa qualité est réalisée sur la base de données des feux de végétation. L'ensemble des étapes sont détaillées dans l'annexe (A.3). Dans cette base la variable réponse étudiée et prédite n'est plus les incendies urbains scrapés mais les feux de végétations. La fréquence observée est 10 fois plus importante pour les feux de végétation que pour les incendies scrapés. Les feux de végétation représentent un risque pour les logements.

Les résultats obtenus sur cette deuxième base sont les mêmes que pour les incendies scrapés malgré des différences dans la répartition des niveaux de risque sur les communes de France. En effet, avec une fréquence plus faible qu'observé dans le portefeuille d'assureur la qualité de prédiction reste la même quelque soit le zonier utilisé.

3.5 Application du zonier dans le modèle interne d'un assureur

Une compagnie d'assurance qui a accepté de tester les méthodes et outils développés dans le cadre de cette étude. Il s'agit d'un acteur majeur et historique du marché de la MRH en France, il est désigné comme « l'assureur » dans cette section du mémoire. Pour des raisons de confidentialité, le modèle interne assureur n'a pas été accessible. Néanmoins, l'évaluation d'une inclusion de variables générées par l'étude est reçue. Ainsi, les différents retours obtenus ne sont pas auditables et leur format non modifiable. Cette évaluation a été réalisée sur une version récente du modèle analysée dans sur les années d'accidents 2017 et 2018 en France.

3.5.1 Contexte de la collaboration

Dans la modélisation des risques associés à la MRH, l'enrichissement de données à partir de variables géocodées a fait l'objet de nombres de travaux de recherche avec des applications essentiellement tournées vers la création de zoniers pouvant améliorer la calibration de la prime pure. L'objectif de cette collaboration est double. Pour l'assureur elle permet d'évaluer l'intérêt de variables géographiques externes pour calibrer son risque. Pour cette étude, elle permet de compléter l'analyse de valorisation de la donnée en réalisant une application en tarification MRH IARD avec un assureur expérimenté et bien positionné dans le marché français. Pour assurer une cohérence des bases de données, l'assureur a fourni l'ensemble des codes INSEE utilisés dans son modèle interne. Les valeurs manquantes ont été complétées avec les valeurs des communes voisines. Pour protéger la qualité d'information issus des données non structurées, les informations fournies à l'assureur sont dégradées. Pour cela, la base de données d'entrainement du modèle a été dégradée de l'ordre de 40%.

Deux variables sont communiquées à l'assureur par commune et par année : la prédiction du nombre annuel d'incendies par commune est réalisée à l'aide de GLM et du zonier incendie. Pour rester compatible avec les outils de l'assureur, le zonier n'est pas réduit en un nombre de clusters fini. La variable continue est réduite à 25 modalités équitablement réparties en quantiles.

Démarche d'intégration et d'évaluation

Le test est réalisé par l'assureur dans une base contenant les sinistres graves, car la construction du zonier réalisée dans cette étude ne fait pas de distinction entre les sinistres attritionnels et les graves. La granularité du zonier partagé n'est, de plus, pas optimisée pour la base de l'assureur. Les exercices de calibrations sont réalisés en tenant compte de la stratégie de partitionnement définie le modélisateur interne de l'assureur. Des échantillons de test et de validation sont utilisés pour communiquer les métriques d'adéquation, et sorties de diagnostic de la qualité du modèle obtenu comparativement au modèle interne initial.

Seul les années 2017 et 2018 sont utilisées pour tester les deux variables. Pour la garantie incendie, l'assureur fait la distinction entre les appartements et les maisons. Ainsi deux modélisations de la fréquence sont réalisées. Ces modélisations sont réalisées avec des GLM ainsi que l'aide d'un logiciel de modélisation tiers spécialisé dans le *machine learning* appliqué à l'assurance.

L'assureur utilise trois indicateurs pour comparer les modèles avec ou sans la variable zonier : le Gini normalisé, la déviance, et le RMSE. Le coefficient de Gini permet de mesurer la répartition d'une variable au sein d'une population. Une sortie graphique clusterisation de la variable ≪ zonier ≫ ainsi que les coefficients appliqués aux différentes modalités de la variable est établie pour présenter la qualité des résultats.

3.5.2 Application

Appartement Sur les appartements, aucune des deux variables ne ressort comme variable importante pour la modélisation de la fréquence.

Maison Sur les maisons, le meilleur modèle sélectionne seulement 13 variables d'importance. Sur ces 13 variables, le zonier incendie intervient en $6^{ème}$. Une comparaison est réalisée par rapport à un modèle sans la variable zonier et les différents critères fournissent les résultats sont présentés dans la table suivante (3.14) :

TABLE 3.14: Table de l'évolution des critères de mesure de la qualité de prédiction par rapport à un modèle sans la variable zonier.

	GINI	DÉVIANCE	RMSE
GAIN	0,48 point	0,1%	Reste stable

Le gain de 0.48 point de gini, montre que l'utilisation de la variable ≪ zonier ≫ permet effectivement de discriminer les variables. Cette discrimination ne se fait pas au détriment de l'erreur qui reste stable comme le montre le RMSE. Pouvoir ordonner les risques est important pour mettre en place les règles de souscription. La méthode permet donc d'améliorer la connaissance du risque.

La tendance des coefficients sur la variable zonier a été communiqué par l'assureur, figure (3.18).

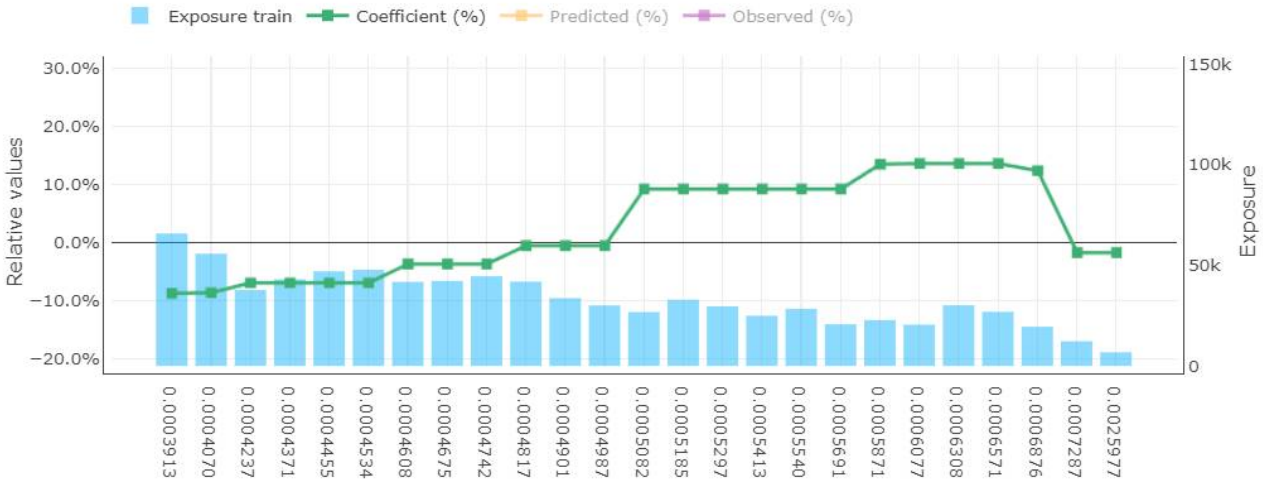


FIGURE 3.18: Coefficient et exposition des modalités de la variable zonier

Sur cette figure, la variable, a été segmenté en huit niveaux de risque. Cette valeur est deux fois supérieure à la segmentation réalisée avec les algorithmes de moyennes mobiles lors du test dans la section (3.4). L'exposition des niveaux de risque est décroissante. Cela peut signifier que l'assureur évite de souscrire des contrats dans des zones à risques élevé.

Les coefficients représentent le comportement du risque pour les modalités. Le caractère du zonier généralement monotone est rassurant même si la chute aux deux derniers niveaux déroge à cette impression. Le gain de gini est intéressant car la valeur ajoutée de la variable est surtout au niveau de la discrimination.

Synthèse

Ce test a permis de constater qu'un zonier construit uniquement à l'aide de données publiques non structurées et donc de manière indépendante par rapport au portefeuille de l'assureur permet d'améliorer la modélisation de la fréquence des incendies sur les maisons. Les appartements sont principalement situés dans des zones à forte densité de population et sont ainsi beaucoup moins sensibles aux données géographiques communales.

3.5.3 Limites de l'étude

Tout au long de la réalisation de ce mémoire, différentes hypothèses ont été prises pour pouvoir mener à bien l'étude. Cependant, certains de ces postulats peuvent constituer des contraintes qui limitent la représentativité des résultats.

Le *web scraping*

Lors de la récupération d'information sur les incendies, le nombre de sites parcourus a été limité, de plus le *scraping* a été réalisé de manière statique en étant sensible aux différentes qualités d'archivages. Une mise en place régulière ou automatique du scraping sur une période plus longue permettrait d'enrichir le nombre d'informations récoltées. La date exacte du sinistre n'est pas disponible facilement et la date de publication de l'article est supposée être celle du sinistre. La commune est la maille la plus fine utilisée, l'idéal serait l'adresse. De nombreux sites interdisent le scraping réduisant la collecte d'information. Les doublons pour un même incendie ne sont pas détectés et altère la base des incendies. La fréquence d'incendie est beaucoup plus faible que la tendance observée par ailleurs.

Traitement automatique des langues

L'utilisation du TAL ne s'est faite que sur les titres d'articles et les chapeaux quant ils étaient disponibles et non sur l'ensemble du contenue de l'article. Dans le cadre de cette étude le TAL a servi à déterminer si un article parlait d'un incendie, mais la distinction entre incendie de logements, ou de végétations n'a pas été mise en place. De plus la distinction entre un sinistre de MRH ou MRP (multirisque professionnelle) n'est pas possible. Néanmoins, un incendie sur un bien couvert avec une contrat de MRP est un risque pour les différents biens couvert par un des contrats de MRH voisins. L'étude du risque étant à la maille commune, cette hypothèse est justifiée, mais une amélioration du TAL pourrait améliorer les résultats.

Modélisation

Les variables géographiques ne sont pas cadencées avec les années. L'âge des occupants par commune n'est pas présent dans les variables explicatives. Les observations du risque d'une commune peuvent être intégralement expliquées par la distribution des caractéristiques d'individus connues dans le portefeuille d'assureur : l'effet mix. Un assureur cherche dans l'étude de l'effet géographique à identifier un effet orthogonal à cet effet de mix. Ainsi l'apport du zonier peut être marginal à un assureur disposant d'un bon modèle interne.

Les prédictions utilisées dans cette étude sont différentes régressions linéaires généralisées. Ces modèles classiques sont utilisés en assurance grace à leur transparence. Cependant, d'autre méthodes plus performantes mais moins transparentes peuvent être utilisées pour essayer de prédire la fréquence des incendies. Une modélisation à l'aide d'une méthode de gradient boosting est possible.

Conclusion

L'objectif de cette étude était d'évaluer la possibilité et l'intérêt d'utiliser des données externes, et en particulier les données non structurées disponibles sur des pages internet, pour améliorer la compréhension de la fréquence du risque incendie sur l'ensemble de la France métropolitaine.

Il s'agissait de mettre en place dans un temps réduit, une méthode agile sur une problématique actuarielle afin de dégager rapidement les axes d'intérêt et d'amélioration. La mise en place des outils spécifiques nécessaires à la constitution des bases de test a été particulièrement chronophage et toutes les potentialités identifiées n'ont pu être développées.

En conformité avec l'ensemble de la réglementation relative au domaine, des outils de *web scraping* ont pu être paramétrés pour récupérer un volume important de données à l'échelon national. Les données récupérées ont été traitées avec des outils de traitement automatique du langage pour être agrégées à la maille de la commune. Une dizaine de bases de données publiques ont ainsi été utilisées pour construire un modèle prédictif de la variable réponse «nombre d'incendies annualisé par commune pondérés par l'exposition» afin de réaliser un zonier pour le risque incendie.

L'étude montre qu'il possible en quelques mois de développer un outil de veille automatique des incendies sur les sites d'informations, moyennant une rapide phase de paramétrage, propre à chaque site.

A l'échelle du département de l'Essonne qui dispose d'une base de données détaillée des interventions de pompier pour incendie, la fréquence prédictive des incendies par commune calculée par le modèle s'avère représentative. Il y a donc une forte probabilité que les modèles générés à l'aide de données scrapées soient transposables à l'échelle nationale.

Bien que la fréquence d'apparition de la variable réponse zonier soit 500 fois plus faible avec les données scrapées que dans le portefeuille de l'assureur, les fréquences d'incendies, les qualités de prédiction de le fréquence incendie entre le zonier sur les données scrapées et le zonier de l'assureur sont semblables. Néanmoins, la fréquence d'apparition de la variable réponse est 500 fois plus faible avec les données scrapées. Un unique portefeuille d'assureur a été utilisé comme objet de comparaison, et il n'est pas encore possible de garantir l'efficacité de l'outil développé.

Une application dans le modèle interne actuellement utilisé par un assureur historique de la MRH en France permet de confirmer la présence d'informations utiles sur des données non structurées. Même si le zonier n'est significatif que pour la modélisation de la fréquence des « maisons », les résultats sont encourageants. La modélisation des incendies sur les appartements étant moins liée aux facteurs environnementaux, les résultats ne sont, par contre, pas concluants sur ces polices à ce stade.

L'évolution des techniques de *web scraping*, de Traitement Automatique des Langues et la démocratisation des bases de données publiques devraient permettre d'obtenir avec le temps de plus en plus d'informations. Ce développement permettra d'améliorer la qualité des données récupérées et les modèles réalisés pour les représenter.

Bibliographie

ABI (2019). UK Insurance and long-term saving. The state of the market 2019.

ACPR (2019). Actualité de la supervision en assurance. https://acpr.banque-france.fr/publications/conferences-et-seminaires/conferences-de-lacpr.

ACPR (2020). Agrément administratif. https://acpr.banque-france.fr/autoriser/procedures-secteur-assurance/regime-administratif/agrement-administratif/principes.

Actuaris (2017). Assurance non-vie : quels tarifs pour 2017. *Actuaris*.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control* 19.6, p. 716-723.

Argus de l'assurance (2019). Classement Auto-MRH 2019 : les assureurs en (re)conquête. https://www.argusdelassurance.com/a-la-une/classement-auto-mrh-2019-les-assureurs-en-re-conquete.146750.

Argus de l'assurance (2020). API : la tech qui bouscule l'assurance. https://www.argusdelassurance.com/tech/api-la-tech-qui-bouscule-l-assurance.161681.

Argus de L'assurance (2020a). Classement Auto et MRH 2020. https://www.argusdelassurance.com/classements/classement-auto-et-mrh-2020.165481.

Argus de L'assurance (2020b). Coronavirus : un gain de près de 2 Md sur l'assurance auto et habitation. https://www.argusdelassurance.com/assurance-dommages/coronavirus-un-gain-de-pres-de-2-md-sur-l-assurance-auto-et-habitation.163746.

Argus de L'assurance (2020c). Covid-19 : les assureurs ont aussi fait des économies sur les flottes auto. https://www.argusdelassurance.com/les-assureurs/covid-19-les-assureurs-ont-aussi-fait-des-economies-sur-les-flottes-auto.167451.

AXA (2012). Conditions générales, Assurance Habitation.

Bellman, R. E. (1961). Adaptive control processes: a guided tour. T. 2045. Princeton university press.

Biard, R. (2010). Dependence and extreme events in ruin theory: univariate and multivariate study, optimal allocation problems. Thèse de doct.

Breiman, L. (1996). Bagging predictors. *Machine learning* 24.2, p. 123-140.

Breiman, L. (2001). Random forests. *Machine learning* 45.1, p. 5-32.

Breiman, L., Friedman, J., Stone, C. J. et Olshen, R. A. (1984). Classification and regression trees. CRC press.

Charpentier, A. (2013). Actuariat IARD - ACT2040 Partie 3 - régression Poissonienne et biais minimal (Y N).

Charpentier, A., Denuit, M., Elie, R. et al. (2015). Segmentation et Mutualisation, les deux faces d'une même pièce ? *Risques* 103, p. 19-23.

Chavent, M., Kuentz, V., Liquet, B. et Saracco, J. (2012). ClustOfVar: An R Package for the Clustering of Variables.

Chavent, M., Kuentz, V., Liquet, B. et Saracco, J. (2017). ClustOfVar: Clustering of Variables. url : https://CRAN.R-project.org/package=ClustOfVar.

CNIL (2019). ACTIVE ASSURANCES : sanction de 180 000 euros pour atteinte à la sécurité des données des clients. https://www.cnil.fr/fr/active-assurances-sanction-de-180-000-euros-pour-atteinte-la-securite-des-donnees-des-clients#:~:text=donn%C3%A9es%20des%20clients-,ACTIVE%20ASSURANCES%20%3A%20sanction%20de%20180%20000%20euros%20pour%20atteinte%20%C3%A0,s%C3%A9curit%C3%A9%20des%20donn%C3%A9es%20des%20clients&text=La%20formation%20restreinte%20de%20la,utilisateurs%20de%20son%20site%20web..

CNIL (2020a). Comprendre le RGPD. https://www.cnil.fr/fr/comprendre-le-rgpd.

CNIL (2020b). Donnée personnelle. https://www.cnil.fr/fr/definition/donnee-personnelle.

CNIL (2020c). Donnée sensible. https://www.cnil.fr/fr/definition/donnee-sensible.

Code des Assurances (2020a). *Code des assurances* Article L113-15-2.

Code des Assurances (2020b). Agrément administratif. *Code des assurances* Article R321-1.

Code des Assurances (2020c). Dispositions générales. *Code des assurances* Articles L121.

Code des Assurances (2020d). Les assurances contre l'incendie. *Code des assurances* Articles L122.

Cramér, H. (1946). Mathematical methods of statistics. *Princeton U. Press, Princeton* 500.

Crawford (2019). Data driven Fire.

Csiszar, I. A., Morisette, J. T. et Giglio, L. (2006). Validation of active fire detection from moderate-resolution satellite sensors: The MODIS example in Northern Eurasia. *IEEE Transactions on Geoscience and Remote Sensing* 44.7, p. 1757-1764.

datagouv (2020). Conditions générales d'utilisation : Demandes de valeurs foncières. url : https://www.data.gouv.fr/fr/datasets/r/99549bdd-91f1-4a99-ac00-855b9a14e5f6.

data.gouv.fr (2020). Base officielle des codes postaux. url : https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux/.

De Palma, A. (2008). Rationalité, aversion au risque et enjeu sociétal majeur. Rapp. tech. OECD/ITF Joint Transport Research Centre Discussion Paper.

Denuit, M. et Charpentier, A. (2005). Mathematiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement.

EARTHDATA (2020). FIRMS Frequently Asked Questions. https://earthdata.nasa.gov/faq/firms-faq#ed-modis-fire-size.

enasis (2020). Annexe 1 - Notion de combustion.

FFA (2020). Etude et chiffres clés assurance habitation 2018. https://www.ffa-assurance.fr/etudes-et-chiffres-cles/assurance-habitation-en-2018.

FFB (2020). INDICE FFB DU COÛT DE LA CONSTRUCTION. url : https://www.ffbatiment.fr/federation-francaise-du-batiment/le-batiment-et-vous/en_chiffres/indices-index/Chiffres_Index_FFB_Construction.html.

Fédération Française de l'assurance (2018). Les incendies domestiques. https://www.ffa-assurance.fr/etudes-et-chiffres-cles?f%5B0%5D=field_categorie_chiffre_cle%253Aparents_all%3A42.

Fédération Française de l'assurance (2019). L'assurance habitation en 2018.

Giglio, Schroeder, Hall et Justice (2018). MODIS Collection 6 Active Fire Product User's Guide Revision B.

Gouvernement (2020). Feux de forêts. https://www.gouvernement.fr/risques/feux-de-forets.

Guillot, A. (2015). Apprentissage statistique en tarification non-vie : quel avantage opérationnel ? Mém. de mast.

Hastie, T. et Tibshirani, R. (2014). Generalized Additive Models.

IGN (2017). Le mémento inventaire forestier.

IGN (2020a). Base de données sur les incendies de forêts en France. url : https://bdiff.agriculture.gouv.fr/.

IGN (2020b). Comment obtenir la distance entre deux points connus en longitude et latitude sur la sphère ? https://geodesie.ign.fr/contenu/fichiers/Distance_longitude_latitude.pdf.

INRS (2020). Incendie sur le lieu de travail. http://www.inrs.fr/risques/incendie-lieu-travail/conditions-survenue.html.

INSEE (2017). fiche-precision. https://www.insee.fr/fr/statistiques/fichier/2383177/fiche-precision.pdf.

INSEE (2019a). 36,6 millions de logements en France au 1 janvier 2019. https://www.insee.fr/fr/statistiques/4263935.

INSEE (2019b). Bilan démographique 2019. https://www.insee.fr/fr/statistiques/1892117?sommaire=1912926.

INSEE (2019). Tableaux de l'économie française. URL : https://www.insee.fr/fr/statistiques/3676693?sommaire=3696937.

INSEE (2020a). Communes nouvelles. https://www.insee.fr/fr/information/2549968.

INSEE (2020b). Indice du coût de la construction des immeubles à usage d'habitation (ICC). https://www.insee.fr/fr/statistiques/3532708?sommaire=3530678.

INSEE (2020c). IRIS definition. URL : https://www.insee.fr/en/metadonnees/definition/c1523.

INSEE (2020d). taux d'inflation. https://www.insee.fr/fr/statistiques/2122401#tableau-figure1.

Institut des actuaires (2016). Norme de Pratique relative aux Modèles actuariels - (NPA2).

Institut des Actuaires (2020). La Gender Directive un an après. https://www.institutdesactuaires.com/magazine/article/la-gender-directive-un-an-apres/2023.

Institut National de l'Information Géographique et Forestière (2017). Le mémento inventaire forestier.

Journal du Net (2019). Le volume de données mondial sera multiplié par 45 entre 2020 et 2035. https://www.journaldunet.com/solutions/dsi/1424245-le-volume-de-donnees-mondial-sera-multiplie-par-45-entre-2020-et-2035-selon-statista/.

Larousse (2020). Dictionnaire. https://www.larousse.fr/dictionnaires/francais/assurance/5915.

Le Big Data (2018). Big Data : le volume de données mondial multiplié par 5 d'ici 2025. https://www.lebigdata.fr/big-data-2025-idc.

Lettria (2020). Une boîte à outils au services des développeurs. https://lettria.com/fr/dev.

LOI n 2010-238 (2010). LOI n 2010-238 du 9 mars 2010 visant à rendre obligatoire l'installation de détecteurs de fumée dans tous les lieux d'habitation(1). https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000021943918/.

LOI n 2014-366 (2014). LOI n 2014-366 du 24 mars 2014 pour l'accès au logement et un urbanisme rénové (1). https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000028772256&categorieLien=id.

Légifrance (2018). Décret n 2018-1350 du 28 décembre 2018 relatif à la publication sous forme électronique des informations portant sur les valeurs foncières déclarées à l'occasion des mutations immobilières. URL : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000037884472&categorieLien=id.

Macapinlac, T. (2018). The Legality of Web Scraping: A Proposal. *Fed. Comm. LJ* 71, p. 399.

Ministère de l'intérieur (2019). Les statistiques des services d'incendie et de secours. URL : https://www.interieur.gouv.fr/Publications/Statistiques/Securite-civile/2018.

Ministère de l'intérieur (2020). Statistiques Sécurité civile. https://www.interieur.gouv.fr/Publications/Statistiques/Securite-civile.

Ministère de l'éducation nationale et de la jeunesse (2020). Indexation de ressources. https://eduscol.education.fr/numerique/dossier/archives/metadata/metadonnees.

MODIS (2020). About. https://modis.gsfc.nasa.gov/about/.

Monde, L. (2014). Fermeture de Google News en Espagne. URL : https://www.lemonde.fr/economie/article/2014/12/11/fermeture-de-google-news-en-espagne_4538263_3234.html.

Naji, F.-Z. (2016). Mém. de mast. Nouvelle modélisation du risque extrême dans la tarification de la garantie incendie en assurance multirisques habitation.

Open knowledge Foundation (2020). Global open data index. https://index.okfn.org/place/.

Optimind (2019). Impact du changement climatique sur l'assurance IARD.

P McCullagh, J. N. (1989). Generalized Linear Models, second edition.

Pariente, J. (2017). Modélisation du risque géographique en assurance habitation. Mém. de mast.

Planchet, M. (2017). Tarification IARD, Introduction aux techniques avancées.

Planetoscope (2020). Les incendies domestiques. https://www.planetoscope.com/habitat/502-nombre-d-incendies-domestiques-en-france.html.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL : https://www.R-project.org/.

Robin Genuer, J.-M. P. (2017). Arbres CART et Forêts aléatoires, Importance et sélection de variables.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics* 6.2, p. 461-464.

Suru, A. (2019). Introduction à l'assurance IARD, Université Paris Dauphine, Année scolaire 2019-2020.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, p. 267-288.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, Vol.59(236)*, p. 433-460.

Werner, G. et Modlin, C. (2010). Basic ratemaking. *Casualty Actuarial Society*. T. 4, p. 1-320.

Wickham, H. (2020). rvest: Easily Harvest (Scrape) Web Pages. URL : https://CRAN.R-project.org/package=rvest.

Wikistat (2020a). Agrégation de modèles — Wikistat. URL : http://wikistat.fr/pdf/st-m-app-agreg.pdf.

Wikistat (2020c). Arbres binaires de décision — Wikistat. URL : http://wikistat.fr/pdf/st-m-app-cart.pdf.

Wikistat (2020b). Arbres binaires de décision. URL : http://wikistat.fr/pdf/st-m-app-cart.pdf.

Wikistat (2020d). Introduction au modèle linéaire général — Wikistat. URL : http://wikistat.fr/pdf/st-m-modlin-mlg.pdf.

Wikistat (2020e). Modèle gaussien : régression linéaire multiple — Wikistat. URL : http://wikistat.fr/pdf/st-m-modlin-regmult.pdf.

Wikistat (2020f). Régression linéaire multiple ou modèle gaussien — Wikistat. URL : http://wikistat.fr/pdf/st-m-modlin-regmult.pdf.

Willis Towers Watson (2019). Les marchés de l'assurance en 2020.

Annexe A

Annexes

A.1 Variables de la base de données

Table A.1: Variable issues du web scraping.

Web scraping		
Date	Intervention	historique
historque.2	historique.3	historique.4

Table A.2: Variables issues de la base INSEE.

INSEE			
Insee	Altitude.Moyenne	Superficie	Population
Code.Canton	Code.Arrondissement	Code.Région	

Table A.3: Variable issues de la base du SDIS.

SDIS		
Feux.d.habitations.bureaux	dont.feux.de.cheminées	Feux.d.ERP.avec.local.à.sommeil
Feux.d.ERP.sans.local.à.sommeil	Feux.de.locaux.industriels	Feux.de.locaux.artisanaux
Feux.de.locaux.agricoles	Feux.sur.voie.publique	Feux.de.véhicules
Feux.de.végétations	Autres.feux	Incendies
Code.Canton	Code.Arrondissement	Code.Région

Table A.4: Variable issues de la base Demande de Valeur Foncière.

DVF		
Code.postal	Valeur.fonciere	Surface.terrain
Surface.reelle.bati	Nombre.pieces.principales	Code.type.local
.data_Appartement	.data_Dépendance	.data_Local.industriel..commercial.ou.assimilé
.data_Maison	.data_Chef.lieu.canton	.data_Commune.simple
.data_Préfecture	.data_Préfecture.de.région	.data_Sous.préfecture

Table A.5: Variable issues de la base d'urbanisation.

Urbanisation			
typau	typpopau10	nafart0918	artact0918
arthab0918	artmix0918	artinc0918	artcom0918
pop11	pop16	.pop1116	men11
men16	men1116	emp16	emp11
emp1116	mepart1116	menhab1116	artpop1116
surfcom18			

Table A.6: Variables issues de la base logement de l'INSEE.

Logement INSEE			
P17_LOG	P17_RP	P17_RSECOCC	P17_LOGVAC
P17_MAISON	P17_RP_4P	P17_RPAPPART	P17_RP_ACH70
P17_APPART	P17_RP_1P	P17_RP_2P	P17_RP_3P
P17_RP_5PP	P17_NBPI_RP	P17_RPMAISON	P17_NBPI_RPMAISON
P17_NBPI_RPAPPART	P17_RP_ACHTOT	P17_RP_ACH19	
P17_RP_ACH45	P17_RPMAISON_ACH45	P17_RPAPPART_ACH19	P17_RPAPPART_ACH14
P17_RP_ACH90	P17_RP_ACH05	P17_RP_ACH14	P17_RPMAISON_ACH19
P17_RPMAISON_ACH70	P17_RPMAISON_ACH90	P17_RPMAISON_ACH05	P17_RPMAISON_ACH14
P17_RPAPPART_ACH45	P17_RPAPPART_ACH70	P17_RPAPPART_ACH90	P17_RPAPPART_ACH05
P17_MEN	P17_MEN_ANEM0002	P17_MEN_ANEM0204	P17_MEN_ANEM0509
P17_MEN_ANEM10P	P17_NPER_RP	P17_ANEM_RP	P17_RP_SDB
P17_RP_PROP	P17_RP_LOC	P17_RP_LOCHLMV	P17_RP_GRAT
P17_NPER_RP_PROP	P17_NPER_RP_LOC	P17_NPER_RP_LOCHLMV	P17_NPER_RP_GRAT
P17_ANEM_RP_PROP	P17_ANEM_RP_LOC	P17_ANEM_RP_LOCHLMV	P17_ANEM_RP_GRAT
P17_RP_SDB	P17_RP_CCCOLL	P17_RP_CCIND	P17_RP_CINDELEC
P17_RP_GARL	P17_RP_VOIT1P	P17_RP_VOIT1	P17_RP_VOIT2P

A.2 Essonne

Table A.7: Table des critères de sélection du meilleur modèle complet pour le zonier sur les résidus.

Méthode	RMSE	MAE	Déviance
Lasso, λ_{min}	10,72	6,48	309
Lasso, $\lambda_{1.se}$	10,63	6,46	225

Le modèle retenu est la régression pénalisée par Lasso avec le $\lambda_{1.se}$. La qualité de prédiction sur la base de validation selon le RMSE et MAE sont proche pour les deux modèle mais la déviance est meilleure pour $\lambda_{1.se}$, sa selection de variable est donc conservée. La table suivante précise les variables les plus importantes de ce modèle et leurs coefficients de pénalisation.

Table A.8: Table des variables les plus importantes du modèle complet.

Variable	Coefficient
Nombre de pièce dans les maisons qui sont des résidences principales	961 288
Pourcentage de logement vacant dans la commune	133 647
Pourcentage de résidence principale locative	65 295
Le pourcentage de résidence principale avec le chauffage centrale indivuduel	1 007
La densité de population	26

Table A.9: Table des critères de sélection du meilleur modèle interne pour le zonier sur les résidus.

Méthode	RMSE	MAE	Déviance
Lasso, λ_{min}	18,84	12,17	885
Lasso, $\lambda_{1.se}$	18,20	12,14	783

De même que pour le modèle complet, la sélection de variable avec $\lambda_{1.se}$ est conservé suite au gain significatif sur la déviance. La table suivante présente les variables les plus importantes du modèle et leurs coefficients de pénalisation.

Table A.10: Table des variables les plus importantes du modèle interne.

Variable	Coefficient
Population des ménages par commune	63 087 410
Population des ménages ayant aménagée depuis plus de 2 ans	10 286 250
Pourcentage de résidence secondaire dans la commune	560 550
Pourcentage de résidence principale construite après 1990	309 511
Pourcentage de résidence principale avec une pièce principale	188 429

A.3 Feux de végétation

Une deuxième modélisation est réalisée sur la base des feux de végétation. Cette base est composée de l'historique des feux de forêt en France et des détections par satellite. Si un incendie est détecté le même jour dans les deux bases, celui-ci n'est comptabilisé qu'une seule fois.

La figure suivante (A.1) présente le nombre d'incendies par commune divisé par le nombre de logements de cette commune. Pour les feux de végétation, il est attendu d'avoir plus d'incendie dans le sud de le France. La fréquence est représentée avec sept modalités réparties selon des quantiles équivalent. Plus la couleur est foncée, plus la fréquence est élevée. La fréquence moyenne sur l'ensemble de la France est de 0,01%, ce qui est 55 fois inférieur au taux de référence de la FFA mais 10 fois plus important que la fréquence observée dans la base d'incendies scrapés.

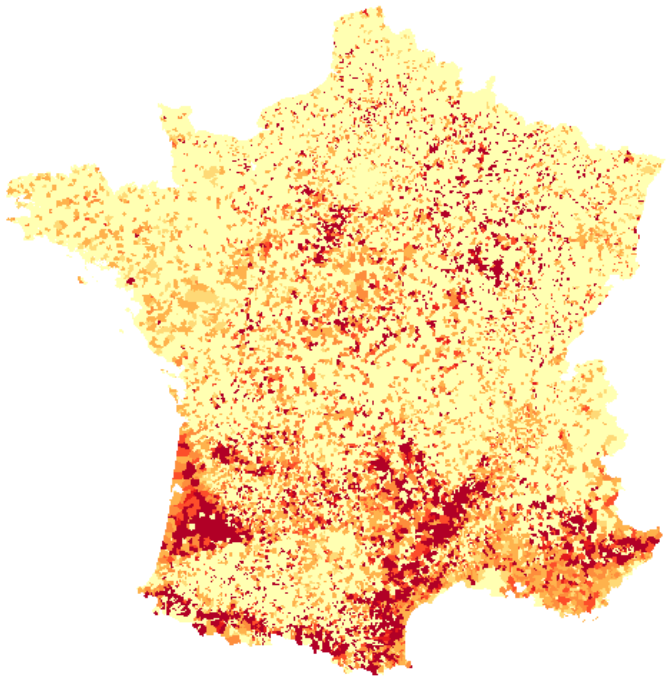


FIGURE A.1: Fréquence des feux de végétation par logement pour l'ensemble des communes de France métropolitaine.

Comme prévu, les communes situées en dessous de l'axe La Rochelle-Besançon sont fortement exposées à des feux de végétations. Les région PACA, Occitanie et le sud de la Nouvelle-Aquitaine possèdent de fort rassemblement de communes limitrophes avec une fréquence élevée. Mais dans la partie nord de la diagonale des faibles densités, plusieurs communes non limitrophes possèdent une fréquence d'incendie élevée. Cette fréquence est importante dû au nombre réduit de logements dans cette région. De plus, le Grand-Est est deuxième région de France avec le plus gros volume de bois sur pied à l'hectare après la Bourgogne-Franche-Comté (INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE (2017)), et les évolutions climatiques font de cette région une des plus touchée par la sécheresse (cf. figure (A.2)). Ainsi, il est cohérent de remarquer des communes dans le nord de la France avec une fréquence d'incendie élevée. Néanmoins, ces communes avec un fort risque sont plus espacées sur le territoire que dans le sud de la France.

A.3.1 Zonier fréquence incendie de végétation

Un zonier de la fréquence incendie pour les feux de végétation est réalisé sur l'ensemble de la France à la maille commune. Pour cela, les deux modèles : le complet et l'interne sont testés et mis en

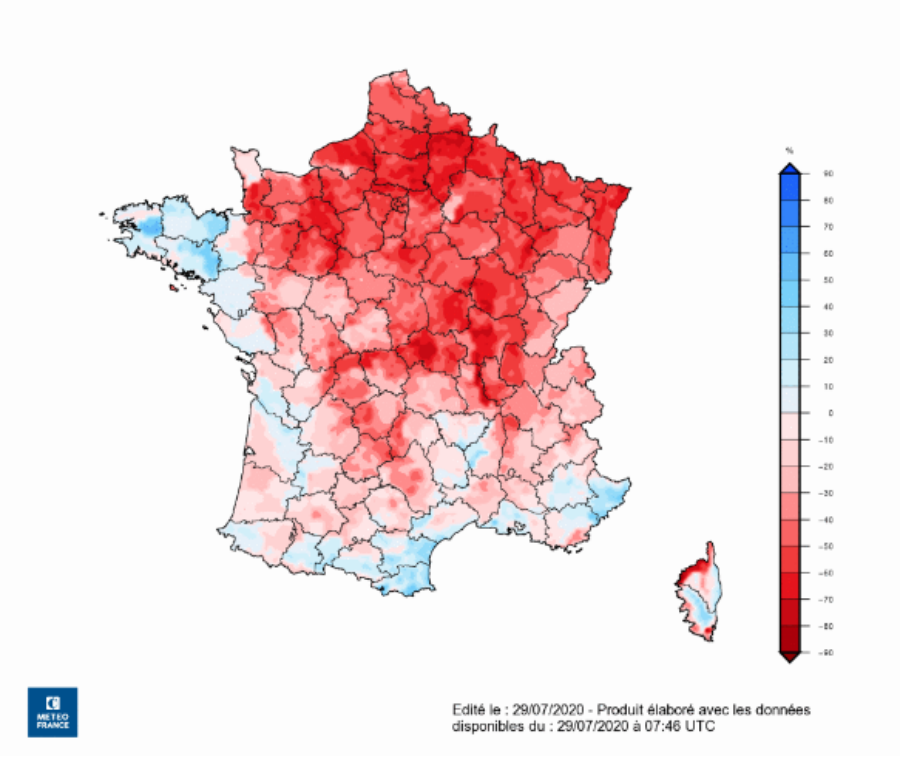


FIGURE A.2: Ecart pondéré à la moyenne quotidienne de référence 1981-2010 de l'indice d'humidité des sols - 28 juillet 2020. Source Météo France.

place.

Modèle complet

Avant de réaliser une modélisation du nombre d'incendies par commune avec le modèle complet, une sélection de variables par la méthode de Lasso est réalisée. La qualité de prédiction entre les différents λ est testée sur la base de validation. Les différents critères permettant de mesurer cette qualité sont résumés dans la table suivante (A.11).

TABLE A.11: Table des critères de sélection du meilleur modèle complet pour le zonier sur les résidus.

MÉTHODE	RMSE	MAE	DÉVIANCE
Lasso, λ_{min}	0,5481	0,1947	10 284,7
Lasso, $\lambda_{1.se}$	0,5679	0,1926	13 409,0

La selection avec λ_{min} est préférée. En effet, le gain en déviance est significatif pour λ_{min}. Les variables les plus importantes pour ce modèle sont résumées dans la table suivante (A.12).

Une fois la sélection effectuée, pour chaque commune et année, le nombre d'incendies est prédit avec la variable représentant le nombre de logement en `offset` pour pouvoir diviser les valeurs prédites par cette variable et ainsi obtenir une fréquence.

Table A.12: Table des variables les plus importantes du modèle complet.

Variable	Coefficient
Pourcentage d'appartement construit avec 1990	0,56
Historique de sinistre dans la commune	0,12
Pourcentage d'appartement dans la commune	0,07
Pourcentage de logement avec le chauffage centrale individuel	0,06
Pourcentage d'habitation à loyer modéré dans la commune	0,06

Modèle interne

Avant de réaliser une modélisation du nombre d'incendies par commune avec le modèle interne, une sélection de variables par la méthode de Lasso est réalisée. La qualité de prédiction entre les différents λ est testée sur la base de validation. Les différents critères permettant de mesurer cette qualité sont résumés dans la table suivante (A.13).

Table A.13: Table des critères de sélection du meilleur modèle interne pour le zonier sur les résidus.

Méthode	RMSE	MAE	Déviance
Lasso, λ_{min}	0,5574	0,1958	11 498,1
Lasso, $\lambda_{1.se}$	0,5757	0,1964	14 792,1

La sélection avec λ_{min} est préférée, la réduction de déviance est plus importante avec cette dernière. Les variables les plus importantes pour ce modèle sont résumées dans la table suivante (A.14).

Table A.14: Table des variables les plus importantes du modèle interne.

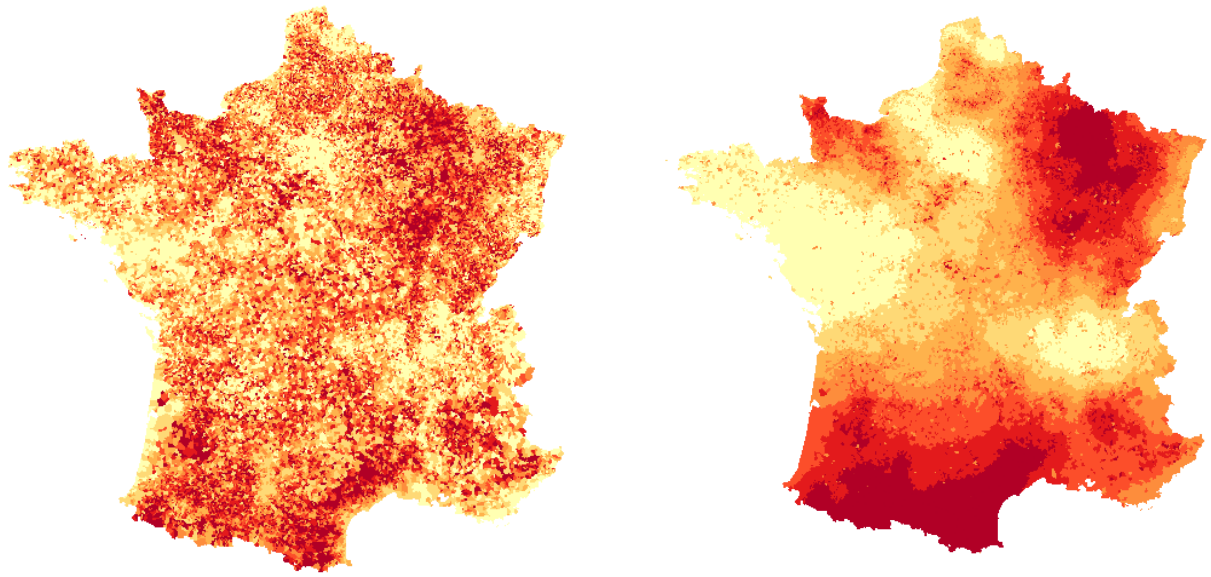
Variable	Coefficient
Pourcentage de résidence principale avec au moins une voiture	0,27
Pourcentage d'appartement dans la commune	0,19
Historique de sinistre dnas la commune	0,14
Pourcentage de logement vaccant dans la commune	0,14

Une fois la sélection effectuée, pour chaque commune et année, le nombre d'incendies est prédit avec la variable représentant le nombre de logement en `offset` pour pouvoir diviser les valeurs prédites par cette variable et ainsi obtenir une fréquence.

Zonier

Une fois les deux modèles de simulation : interne et complet, mis en place, leurs prédictions sont comparées entre elles. Les résidus issus de la comparaison entre le modèle interne et modèle complet sont ensuite lissées à l'aide de la théorie de la crédibilité. La figure suivante présente la fréquence prédite sans lissage et le lissage le plus faible permettant d'obtenir un résultat visuel satisfaisant c.-à-d. réduire les différences de risque trop fortes entre deux communes. Le niveau de risque est représenté avec sept modalités : plus la couleur est foncée plus le risque est important.

La répartition des risques est proche de celle obtenue avec les incendies scrapées, cependant le



(a) Zonier non lissé

(b) Zonier lissé

FIGURE A.3: Zonier de la fréquence des feux de végétation pour l'ensemble des communes en France pour différents lissage

risque est beaucoup plus faible pour l'île de France et les agglomérations Lyonnaise et Lilloise. L'ensemble du sud de la France possède un fort risque avec la région Grand-Est.

Comme pour la base des incendies scrapées, le nombre de clusters est déterminé en utilisant un algorithme de *k-means* sur les fréquences prédites. L'étude de l'évolution de la variance expliquée en fonction du nombre de clusters est présentée dans la figure (A.4).

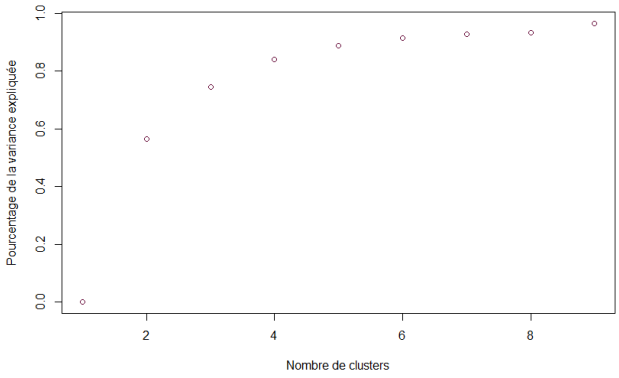


FIGURE A.4: Variance expliquée en fonction du nombre de clusters.

La méthode du coude n'est pas explicite dans ce genre de situation. Le gain de variance expliquée est de forme logarithmique, il n'y a pas de variation forte qui forme un «coude». La qualité de

prédiction de chaque zonier est testée sur la base de validation avec le modèle interne. Les critères d'évaluation de la qualité de prédiction sont résumés dans la table (A.15).

TABLE A.15: Comparatif de la qualité de prédiction sur la base de validation des données.

NOMBRE DE CLUSTER	RMSE	DEVIANCE
2	0.5811	39 747,3
3	0.5794	39 559,2
4	0,5768	39 277,1
5	0,5781	39 301,4
6	0,5782	39 246,7
7	0,5776	39 220,6

La qualité de prédiction augmente avec le nombre de clusters. Cependant pour cinq clusters les prédictions sont moins bonnes que pour quatre clusters. Ce minimum local dans la déviance permet de choisir 4 comme nombre optimal de clusters. La figure suivante (A.5) présente le zonier du risque de survenance d'incendie par commune pour quatre niveaux de risque croissant :

- L : risque faible,
- M1 : risque moyen 1,
- M2 : risque moyen 2,
- H : risque élevé.

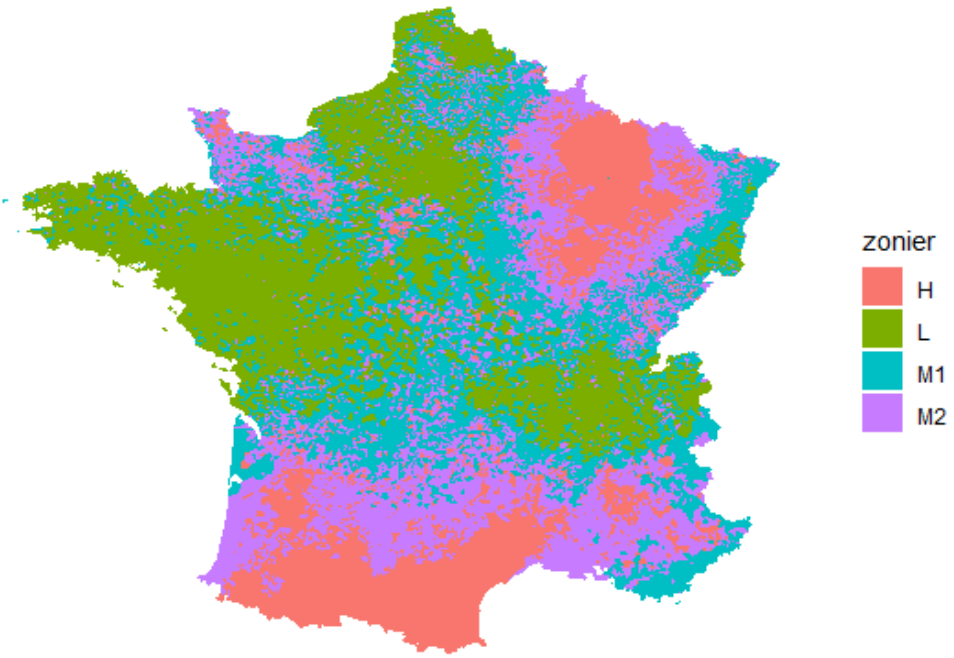


FIGURE A.5: Zonier à 4 modalités.

A.3.2 Test du zonier sur le portefeuille d'assureur MRH

Une fois le nombre de clusters choisit, un test sur le portefeuille d'assureur est réalisé pour comparer la qualité de prédiction sur la base de test selon le zonier. Les différents critères de mesure de la qualité de prédiction sont résumés dans la table suivante (A.16).

Table A.16: Evalution de la qualité de prédicton des modèles selon le zonier choisit.

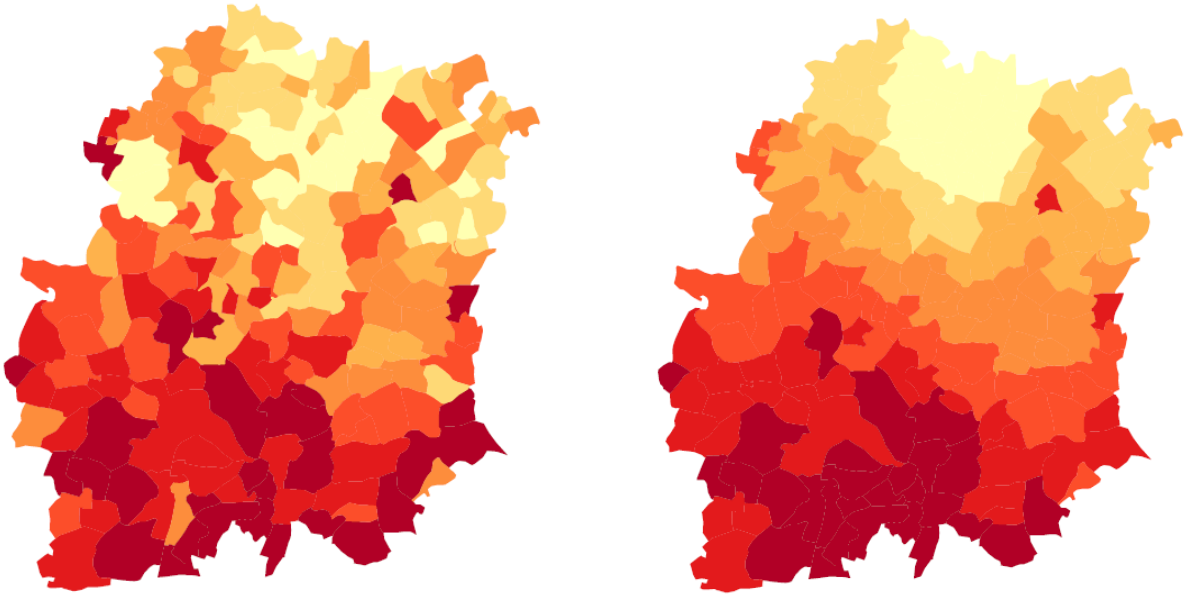
Zonier	MAE	RMSE	Déviance
Zonier de l'assureur	0,005063	0.051230	5946,01
Zonier sur les données scrapées	0,005062	0,051232	5945,72

Conclusion feux de végétations

Les feux de végétation sont un risque pour les logements mais ce risque n'est pas réparti équitablement sur l'ensemble des logements d'une commune. En effet, si une habitation est éloignée d'une zone fortement urbanisée, elle est plus exposée. De même l'ensemble des appartements semble moins concerné par ce risque. Une tarification à l'adresse de ce type de risque pourrait être plus significative.

Les différentes limites de l'études sont les mêmes que celles présentées pour le zonier sur les données scrapéees. Néanmoins, les données issues des satellites de la NASA ne sont pas une source sans erreurs et peuvent présenter certain sinistre de MRH ou de MRP.

A.4 Graphiques



(a) Lissage faible

(b) Lissage élevé

FIGURE A.6: Risque incendie pour l'ensemble des communes de l'Essonne avec différents lissage

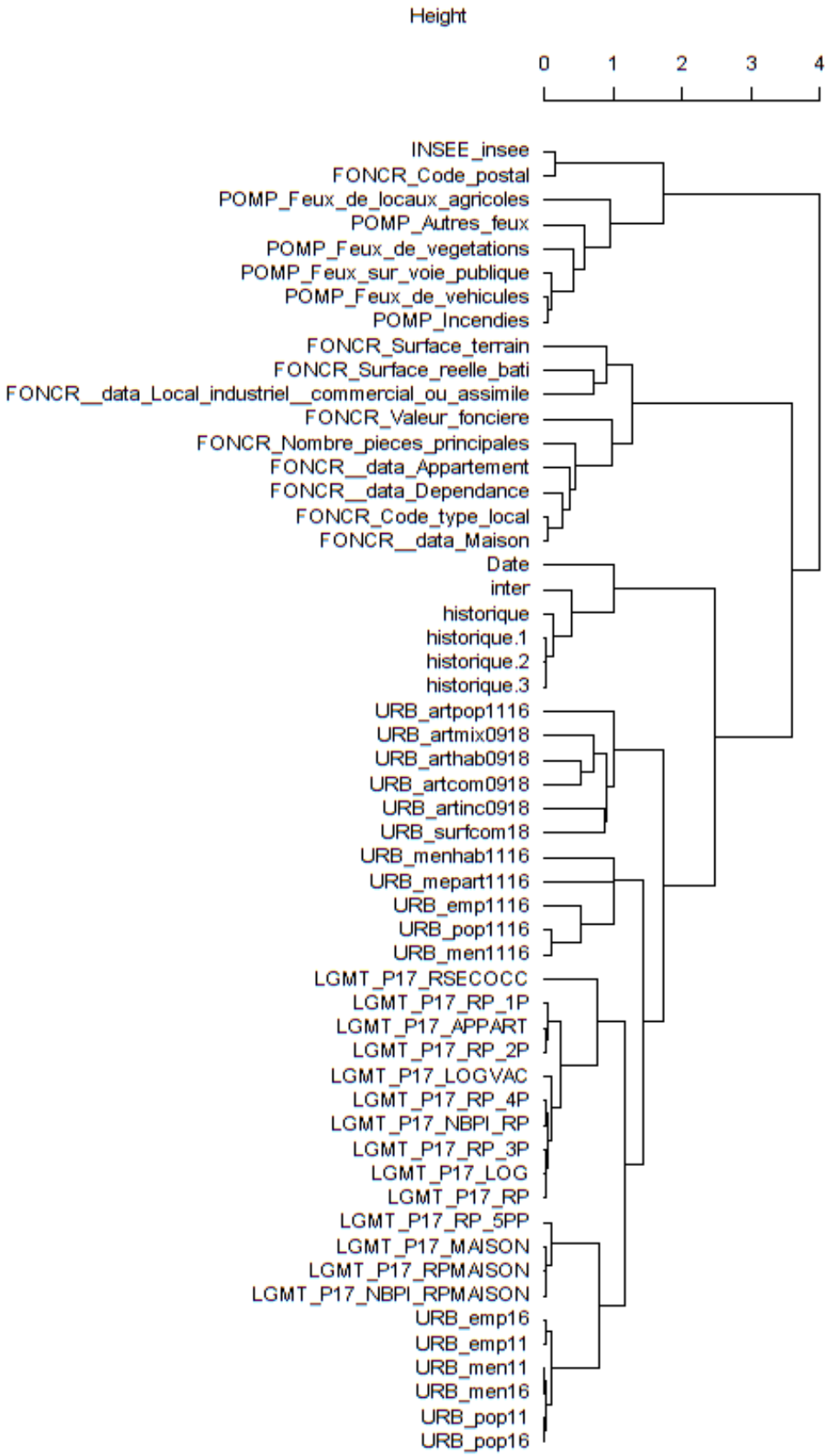


FIGURE A.7: Cluster de variable sur un echantillon de la base d'incendie scrapés

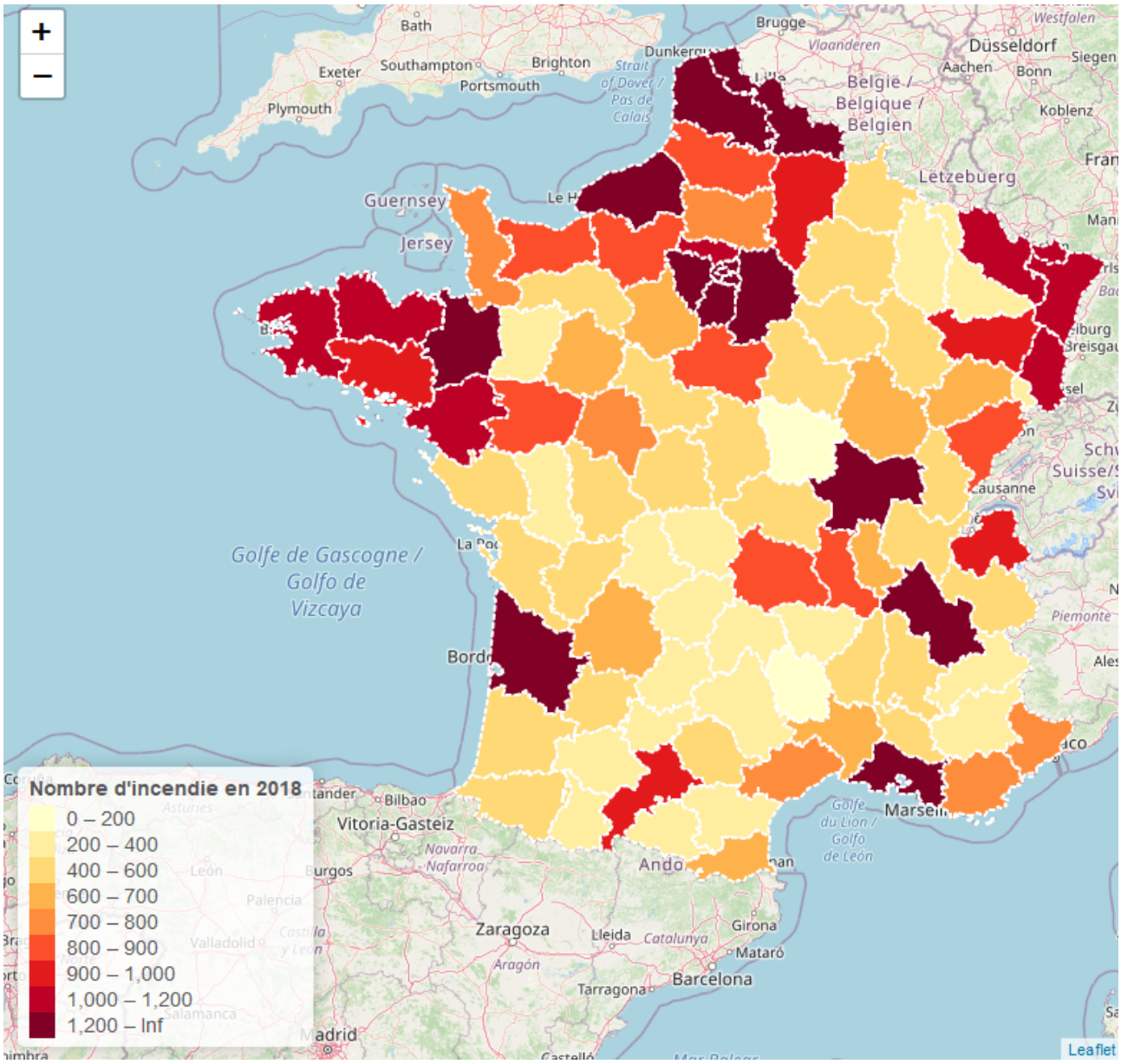


FIGURE A.8: Utilisation de la base de données nationale interventions des pompiers pour feux d'habitations et bureaux.

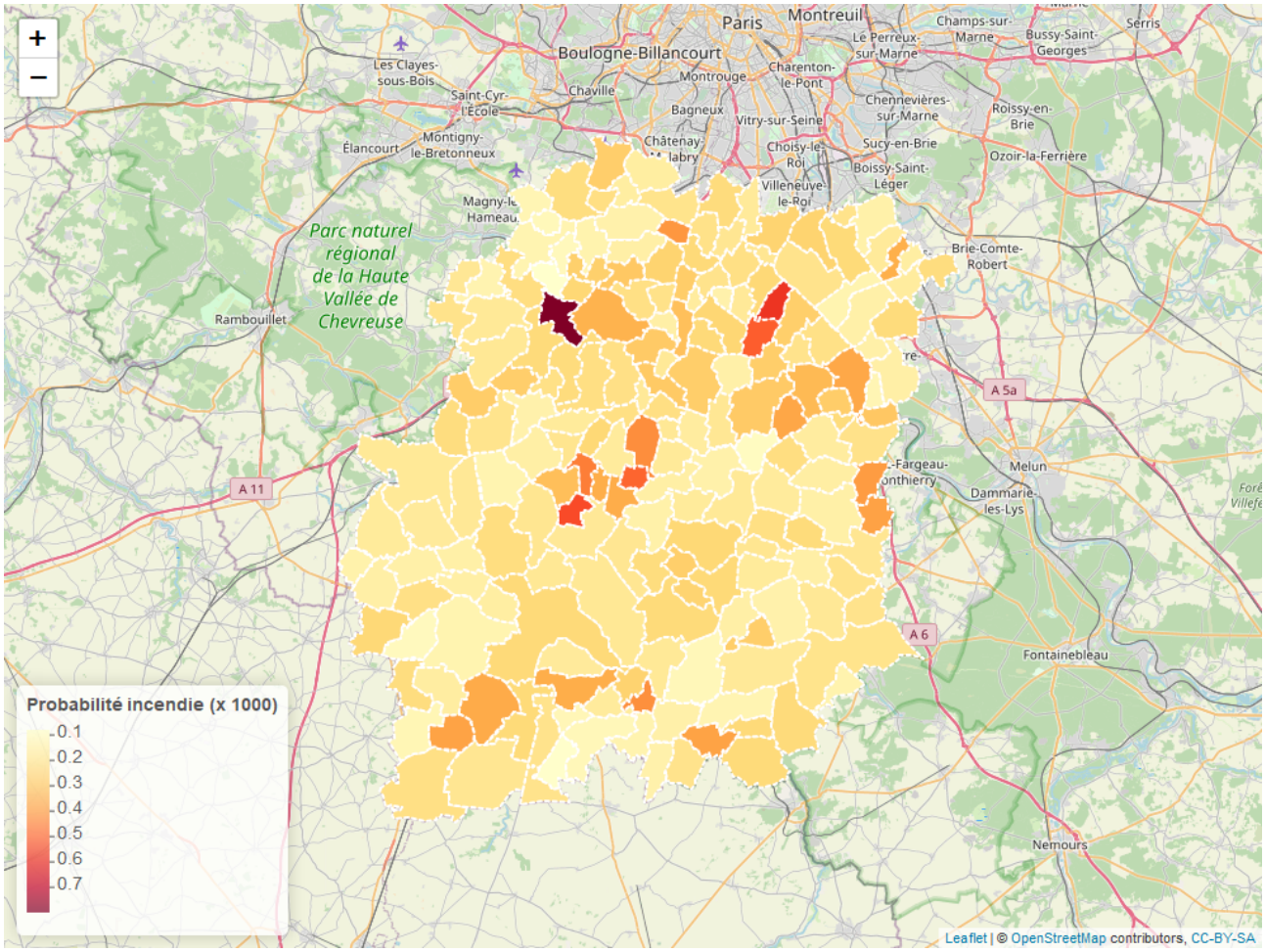


FIGURE A.9: Utilisation de la base de données des interventions de pompiers en Essonne.

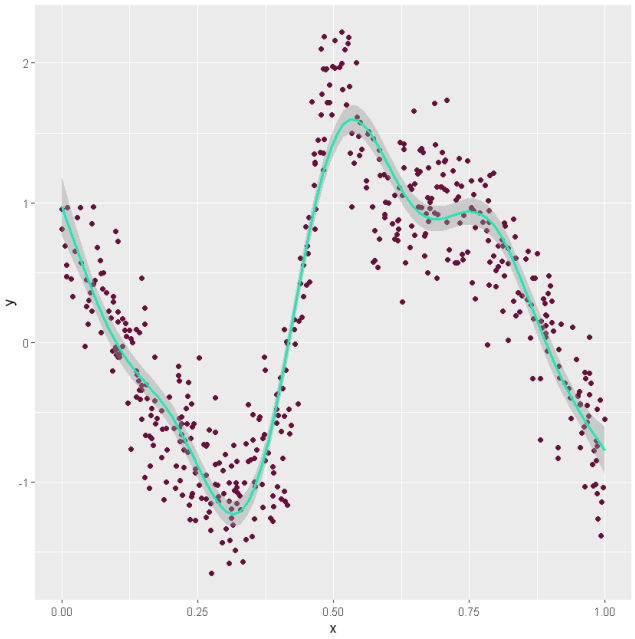


FIGURE A.10: Exemple d'utilisation du GAM.