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

Par : Madame Delphine IMBERT

Titre du mémoire :
Création d'un modèle d'estimation de la sinistralité des tempêtes en France métropolitaine

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

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

Membres présents du jury de l'Institut des Actuaires signature

Entreprise :

Nom : ALLIANZ

Signature :

Directeur de mémoire en entreprise :

Nom : Stéphane KOLASA

Signature :

Membres présents du jury de la filière

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Mémoire présenté en vu de l'obtention du diplôme d'actuariat de l'ISUP et de l'admission à l'Institut des Actuaires

❧

Prédiction de la sinistralité des tempêtes en France métropolitaine

Par : **Delphine IMBERT**

Sous la direction de :
Stéphane KOLASA
et Anna BEN-HAMOU

Date de rendu : 23 novembre 2018
Date de soutenance : 31 janvier 2019

Remerciements

Je tiens à remercier l'ensemble de l'équipe Pilotage et Études Techniques Indemnisation pour leur accueil et pour le très bon environnement de travail qui m'ont permis de profiter pleinement de cette expérience professionnelle et de rédiger mon mémoire dans les meilleures conditions.

Je souhaite tout particulièrement remercier mon tuteur de stage, M. Stéphane KOLASA, pour le temps et les précieux conseils qu'il m'a accordé tout au long de mon mémoire.

Je tiens également à remercier Mme. Anna BEN-HAMOU, tutrice de mémoire académique, pour son aide lors de la rédaction de ce mémoire.

Résumé

Prédire la sinistralité des tempêtes est un enjeu crucial pour les compagnies d'assurance qui ont tendance à les sous-estimer. En effet, les tempêtes n'étant pas considérées comme des catastrophes naturelles, l'Etat n'est pas réassureur de ces évènements et une bonne prédiction apporte à la compagnie un avantage concurrentiel indéniable du fait de sa capacité à gérer au mieux ces événements.

Le but de ce mémoire est donc de créer un modèle permettant l'estimation de la sinistralité des tempêtes en France métropolitaine. Ce modèle aura une contrainte majeure : il devra être en mesure de rendre une estimation fiable trois jours après la survenance d'une tempête. L'ensemble des sinistres n'étant pas connus à J+3, il sera essentiel de trouver des données externes fiables permettant d'apporter un complément d'informations aux données portefeuilles présentes dans les bases d'Allianz.

Le modèle mis en place repose sur plusieurs sous-parties interconnectées, chacune étant analysée séparément avec plusieurs méthodes distinctes pour une meilleure prédiction. Notre modèle s'articule autour des axes suivants : récupérations et traitements des données relatives aux tempêtes pour la création d'une base de données fiable, prédiction des tempêtes de références, estimation du nombre de sinistres total ainsi que le coût global de la tempête. Ce coût sera subdivisé en deux : le coût des sinistres graves et celui des sinistres attritionnels.

Mots clefs : Tempête, France métropolitaine, Classification, Imputation multiple, GLM, GAMLSS, Théorie des valeurs extrêmes, Apprentissage supervisée, Perceptron multicouches.

Pour des raisons de confidentialité, les valeurs numériques résultant de l'activité d'Allianz France ont été modifiées. Les ordres de grandeur restant inchangés, les conclusions de mon mémoire sont les mêmes que celles présentées à Allianz. Par ailleurs, l'ensemble du traitement des données a été effectué grâce au logiciel SAS et les différentes modélisations ont été réalisées à partir du logiciel libre de statistique R.

Abstract

Predicting the accident severity of storms is a crucial issue for insurance companies that tend to underestimate them. Indeed, storms are not considered natural disasters, the State is not reinsurer of these events and a good prediction provides the company an undeniable competitive advantage because of its ability to better manage these events.

The purpose of this thesis is to create a model allowing for the estimation of the accident severity of storms in metropolitan France. This model will have a major constraint: it will have to be able to make a reliable estimate three days after the occurrence of a storm. As all the claims are not known at D+3, it will be essential to find reliable external data to provide additional information to the portfolio data present in Allianz databases.

The model implemented is based on several interconnected subparts, each being analyzed separately with several distinct methods for better prediction. Our model is based on the following aces: recovery and processing of storm data for the creation of a reliable database, prediction of reference storms, estimation of the total number of claims and the overall cost of the storm. This cost will be subdivided in two parts: the cost of serious and attritional claims.

Keywords: Storm, Metropolitan France, Classification, Multiple Imputation, GLM, GAMLSS, Extreme Value Theory, Supervised Learning, Multilayer Perceptron.

For confidential reasons, the numerical values resulting from the activity of Allianz France have been modified. The orders of magnitude remaining unchanged, the conclusions of my thesis are the same as those presented to Allianz. Moreover, all the data processing was done using the SAS software and the different modelizations were carried out from the free statistical software R.

Note de Synthèse

Contexte et problématique

La majorité des risques assurables dispose d'un historique de sinistres conséquent, permettant aux compagnies d'assurances de modéliser leur sinistralité à venir avec une très grande précision. En revanche, pour l'ensemble des phénomènes naturels catastrophiques tels que les tempêtes, leur occurrence est si faible que l'historique associé ne permet généralement pas d'appliquer les méthodes de statistiques classiques.

L'objectif de ce mémoire est de construire un modèle au sein d'Allianz, robuste et fiable, à l'exécution rapide, permettant d'estimer la sinistralité des tempêtes trois jours après leur survenance en France métropolitaine. Ce modèle a pour but d'être utilisé sur le long terme au sein de l'Indemnisation. Afin qu'il s'adapte aux différents changements, nous avons choisi de faire reposer notre modèle sur trois principes:

- l'humain pour profiter des dires d'expert et des remontées terrains
- les statistiques afin d'obtenir des résultats fiables et compréhensibles
- des modèles auto-apprenants afin qu'ils s'adaptent aux différentes évolutions (climatiques, interne à Allianz, réglementaire) et qu'ils identifient des relations invisibles jusqu'alors.

Les contraintes temporelles et de fiabilité sont fortes puisque les résultats du modèle sont destinés à être communiqués tant en interne qu'en externe pour des prises de décision rapides. C'est pourquoi, il a été décidé dans ce modèle de lier l'humain, les statistiques et le machine learning au lieu de les opposer comme c'est généralement le cas.

Contenu du mémoire

Cadre de l'étude

Avant de débuter la modélisation, il est nécessaire de fixer un cadre à l'étude. Dans ce mémoire, le terme « tempête » désignera l'ensemble des phénomènes climatiques dont le vent est la principale cause de dommages. Dans notre étude, une tempête peut être associée à d'autres phénomènes climatiques tels que des orages, de la grêle ou des inondations. Le modèle prendra en compte l'ensemble de ces phénomènes qui seront distingués grâce à une variable qui précisera le type de la tempête. Nous nous contenterons ici d'estimer la

sinistralité des tempêtes sur les biens personnels (Auto, MRH...) en laissant de côté des dommages corporels qui sont très différents à modéliser et très peu nombreux lors des tempêtes.

Plan du mémoire

Les tempêtes sont un phénomène complexe tant d'un point de vue météorologique qu'assurantiel. En effet, les tempêtes n'étant pas considérées comme des catastrophes naturelles, l'Etat n'est pas réassureur de ces évènements pouvant coûter des millions d'euros. Une mauvaise estimation de la sinistralité peut donc être désastreuse pour une compagnie d'assurance. Afin de modéliser au mieux ce risque, il est nécessaire de bien le comprendre. La première partie est donc consacrée à son étude en profondeur, la seconde partie aux données. Il s'agit dans cette deuxième partie d'identifier les variables explicatives du phénomène, de les traiter et de créer une base de données des tempêtes passées mis en "as if". Enfin, la dernière partie concerne la modélisation a proprement parlée ainsi que l'ensemble des tests qui ont été mis en place pour valider son utilisation et sa robustesse.

Choix du modèle

Le modèle qui va être présenté dans ce mémoire est composé de trois parties interconnectées, qui interagissent les unes avec les autres. Dans chacune de ces parties, l'homme, les statistiques et la data science coexistent pour donner les résultats interprétables et fiables. Le modèle global est le suivant :

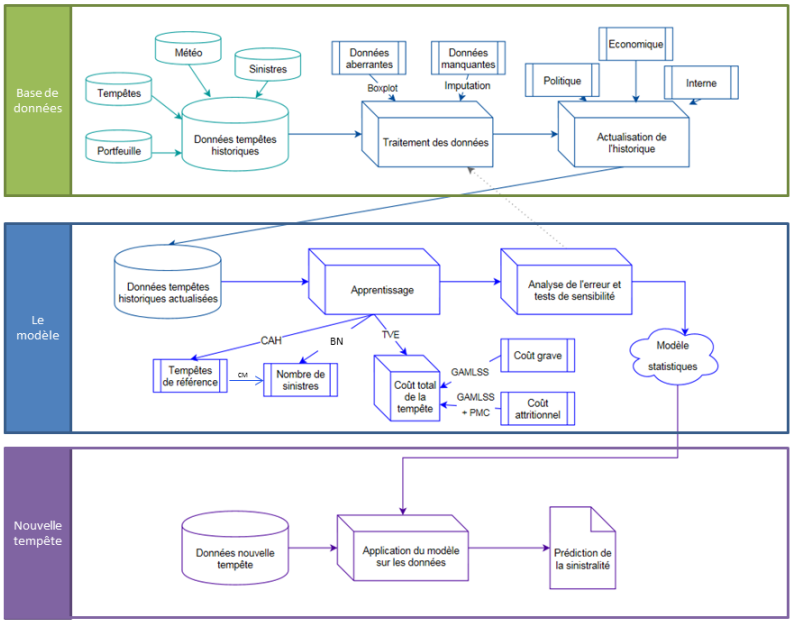


Figure 0.0.1: Schéma du modèle final

Choix, résultats et analyses

Création de la base de données

Les données sont au cœur des modèles donc au cœur du métier d'actuaire. Par conséquent, il est important de leur prêter une attention toute particulière. En effet, si les données de départ sont de mauvaises qualités (données manquantes, aberrantes, non appropriées...) le modèle aussi sophistiqué soit-il sera faussé.

La première partie du mémoire consiste à créer une base de données des tempêtes passées complète et pertinente. Pour cela, un long travail sur leur qualité a été mené, travail que nous résumerons par le schéma suivant :

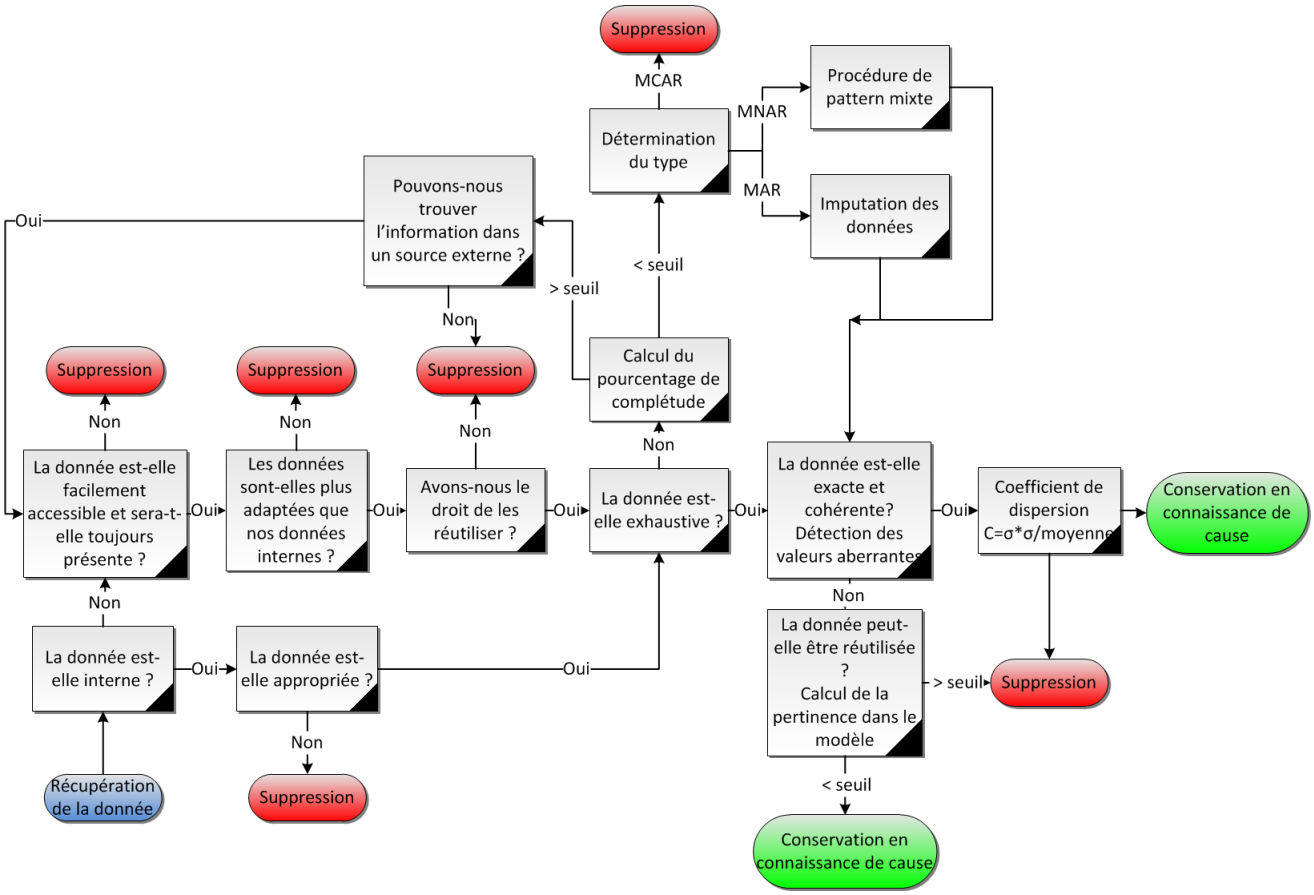


Figure 0.0.2: schéma de la sélection des données

Une fois la base de données constituées, il a été nécessaire de l'actualiser c'est-à-dire modifier le passé pour qu'il ressemble au présent. Pour ce faire, nous avons étudié l'ensemble des modifications internes à Allianz, politiques et économiques, survenues entre 2003 (date de début de notre historique) et 2018.

Le modèle

Une fois les données traitées, nous nous sommes intéressés à la création du modèle. Ce dernier repose sur 3 sous-modules :

- La modélisation des tempêtes de références
- La modélisation du nombre de sinistres
- La modélisation du coût des tempêtes

1. Les tempêtes de références

« Pour bien juger, il faut connaître. Et le seul moyen de connaître est de décomposer l'objet inconnu... »[1]. Or, en associant une tempête qui vient de se produire à des tempêtes passées, on se soustrait à l'inconnu. Ainsi, le phénomène est plus facilement appréhendable et il est plus simple de communiquer dessus, de l'estimer ...

Afin de déterminer les tempêtes de références, une méthode basée sur les dires d'expert ainsi qu'une méthode de classification supervisée (K plus proches voisins) et non-supervisée (Classification Ascendante Hiérarchique) ont été mises en place. La combinaison de ces méthodes permet d'obtenir une connaissance approfondie de la tempête qui vient de survenir.

2. La modélisation du nombre de sinistres

Concernant de la modélisation du nombre de sinistres, deux méthodes ont été mises en place. L'une est basée sur les cadences moyennes des tempêtes de référence de la tempête. L'autre, plus statistique, est une régression binomiale négative.

Ces méthodes ont la particularité d'être complémentaires. Lorsque l'une ne tient compte que de certaines caractéristiques fortes de la tempête, l'autre prend en compte un ensemble de petits paramètres. Il est intéressant de remarquer que les résultats obtenus par les deux méthodes sont quasiment identiques, ce qui permet de faire une double validation de la prédiction.

3. Le coût des tempêtes

En assurance, on distingue les sinistres graves dont l'occurrence est faible et le coût élevé des sinistres attritionnels. Ces deux types de sinistres n'ayant pas la même distribution, il est essentiel de les étudier séparément.

La détermination du seuil des sinistres graves a été réalisée grâce à la théorie des valeurs extrêmes. L'estimateur de Hill désigne le seuil des graves comme étant autour de

[1] Citation de Philippe-Auguste de Sainte-Foy; Mes loisirs, ou pensées diverses (1755)

90000€. Une fois déterminé, nous avons appliqué un GAMLSS Logistic pour réaliser notre prédiction. A cette dernière s'ajoute des remontées terrains qui permettent d'affiner notre estimation.

L'estimation du coût des sinistres attritionnels s'est révélée plus complexe. La distribution de ces sinistres ne semblait correspondre à aucune loi connue. Afin de contourner ce problème, un réseau de neurones a été mis en place et son initialisation a été effectuée à partir des coefficients de régression d'un GAMLSS Gumbel. Loi qui estime correctement mais qui ne prenait pas en compte les spécificités d'une tempête. L'association des deux a permis de rendre le réseau de neurones plus performants et plus rapide d'exécution.

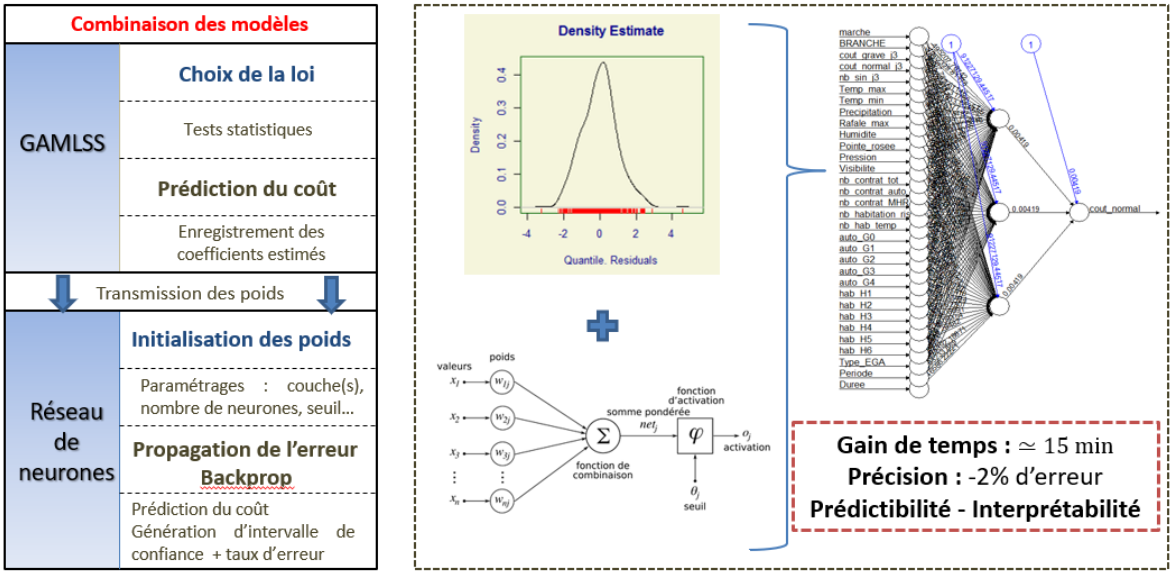


Figure 0.0.3: Schéma de regroupement du Réseau de neurones avec le GLM

Résultats du modèle

Ce modèle a été créé sur le principe de diviser pour mieux régner. Chacun des modules est indépendant mais connecté avec les autres. Ainsi chaque module peut être lancé simultanément, rendant l'exécution beaucoup plus rapide.

Grâce à ses différentes méthodes, le modèle possède en permanence une double validation. Ainsi, si une méthode diffère d'une autre, une analyse approfondie peut-être rapidement menée pour en identifier les causes. Par ailleurs, en liant l'humain, les statistiques et la data science, le modèle est robuste, rapide, fiable et facilement interprétable.

L'un des atouts de ce modèle est son adaptabilité. En effet, il peut être utilisé pour estimer la sinistralité de n'importe quel phénomène climatique si l'on fournit en entrée du modèle l'historique adéquat.

Executive summary

Context and problem

The majority of insurable risks have a significant claims' history, allowing insurance companies to model their future claims with great precision. However, the occurrence of catastrophic natural phenomena such as storms is so low that the associated history does not allow to use conventional statistical methods most of the time.

The objective of this thesis is to build a robust and reliable model within Allianz. This model would execute fast and would allow to estimate the sinistrality of storms three days after their occurrence in metropolitan France. This model is intended for long-term use in Compensation department. In order to adapt to the different changes, we chose to base our model on three principles:

- human vision to benefit from expert reports
- statistics to obtain reliable and understandable results
- self-learning models so that they adapt to the different evolutions (climate changes, evolutions within Allianz, ...) and that they identify relationships that were previously invisible.

As the results of the model are intended to be communicated both internally and externally for fast decision-making, time and reliability constraints are high. Rather than opposing human vision, statistics and machine learning as it is usually done, it was decided in this model to link them.

Thesis' Content

Study's framework

Before starting to model, it is necessary to set a framework for the study. In this thesis, the word "storm" refers to all climatic phenomena for which wind is the main cause of damage. In this study, a storm may be associated with other climatic phenomena such as thunderstorms, hail or floods. The model will take into account all of these phenomena that will be distinguished by a variable that specifies the type of storm. We will only estimate

the storms' sinistrality on personal assets (Auto, MRH ...) and leave aside physical injuries that are very different to model and not very common during storms.

Thesis' plan

Storms are a complex phenomenon from a meteorological and insurance point of view. Because they are not considered as natural disasters, the government is not reinsurer of these events that can cost millions of euros. A poor estimate of the sinistrality can therefore be disastrous for an insurance company. In order to better model this risk, it is necessary to understand it well. Thus, the first part is devoted to its in-depth study and the second part to the data. In this second part, we will try to identify the explanatory variables of the phenomenon, to treat them and to finally create an updated database of the past storms. The last part deals with the actual modeling as well as all the set of tests that allowed to validate its use and its robustness.

Choice of the model

The model that will be presented in this thesis is composed of three interconnected parts, which interact with each other. In each of these parts, human vision, statistics and data science coexist to give interpretable and reliable results. The global model is as follows:

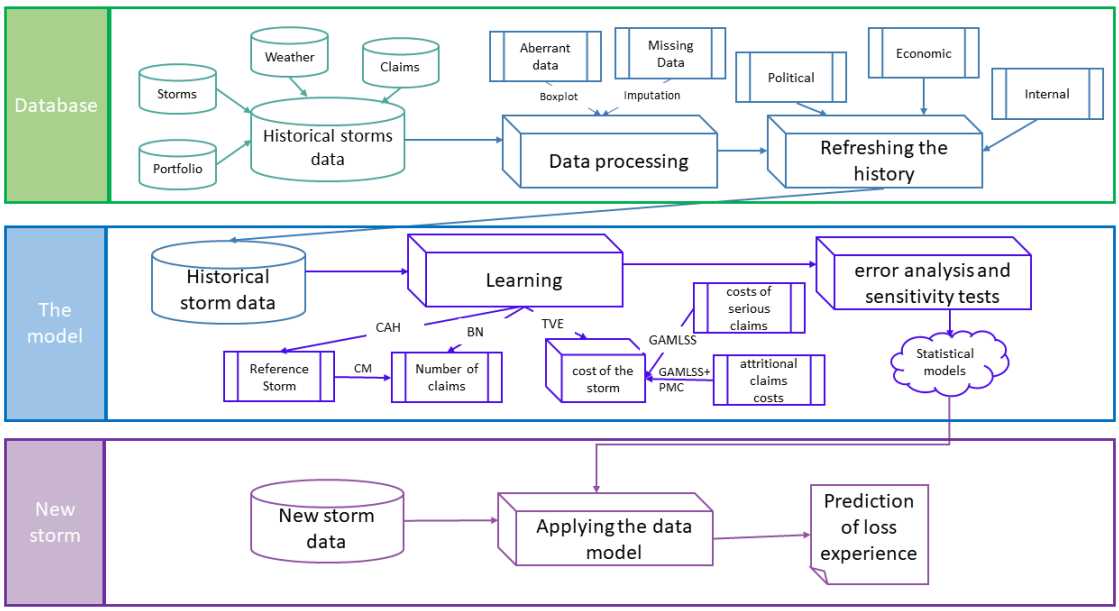


Figure 0.0.4: Final model diagram

Choices, results and analyzes

Database creation

Data is the center of model creation and therefore of the actuarial profession. Thus, it is important to pay special attention to them. Indeed, if the initial data is of bad quality (missing data, aberrant data, inappropriate data ...) the model will be distorted even if it is sophisticated.

The first part of this thesis is to create a complete and relevant past storms' database. To that end, a hard work was carried out on their quality. We will summarize this work thanks to the following diagram:

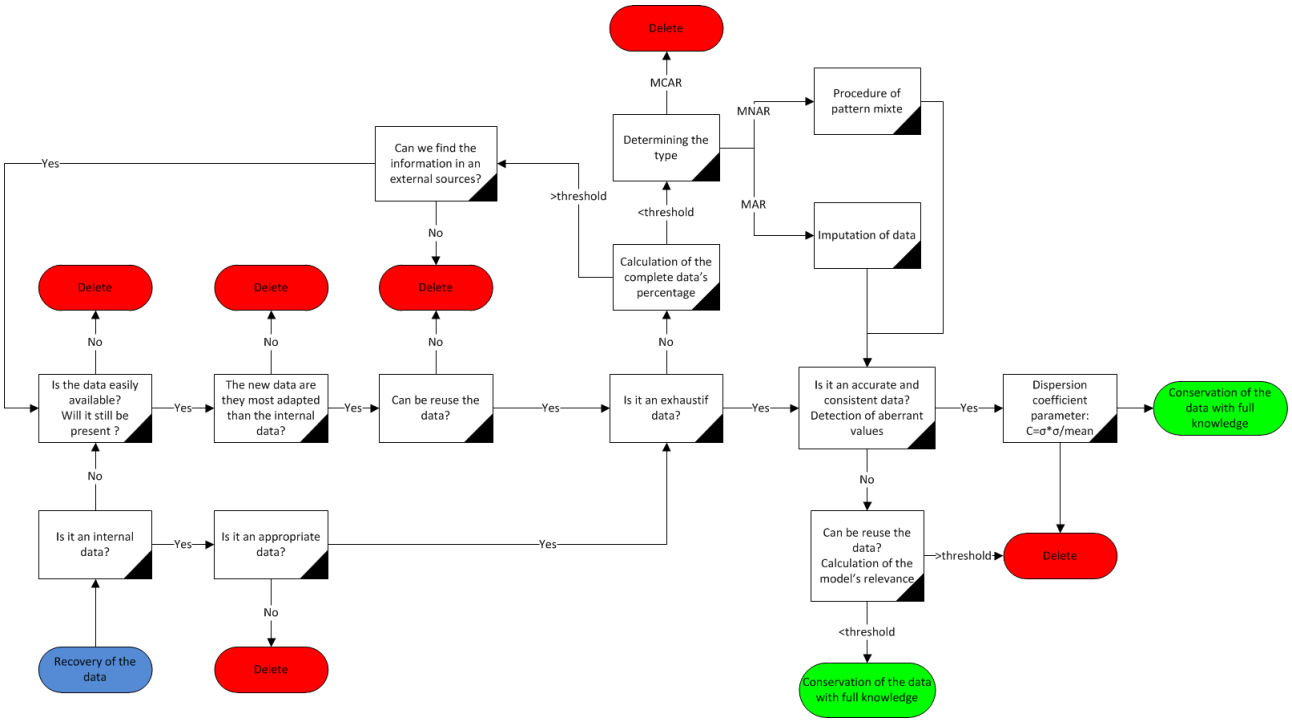


Figure 0.0.5: Schema of data selection

Once the database has been created, it was necessary to update it to integrate the past event into the current situation. To do this, we studied all the internal changes of Allianz as well as political changes and economical changes that occurred between 2003 (beginning of our history) and 2018.

The model

Once the data was processed, we were able to create the model. The latter is based on 3 sub-modules:

- Modeling of the reference storms
- Modeling of the number of claims
- Modeling of the storms' cost

1. Reference storms

"To judge well, one must know. And the only way to know is to break down the unknown object. . . ". By associating a storm that has just occurred with past storms, one evades the unknown. Thus, the phenomenon is more easily apprehendable and it is easier to communicate about it, to estimate it ... In order to determine the reference storms, a method based on expert statements as well as a supervised classification method (K nearest neighbors) and unsupervised (Hierarchical Ascending Classification) have been put in place. The combination of these methods provides in-depth knowledge of the storm that has just occurred.

2. The modeling of the number of claims

Concerning this model, two methods have been used. One of them is based on the average rates of the reference storms associated to the current storm. The other one, more statistical, is a negative binomial regression.

These methods are complementary. When one of them considers only important features of the storm, the other one takes into account a set of small parameters. It is interesting to note that the results obtained by the two methods are almost identical, which allows a double validation of the prediction.

3. The cost of storms

In insurance area, there is a distinction between low-occurrence claims with high cost and attritional claims. Since these two types of claims do not have the same distribution, it is essential to study them separately. The determination of the threshold of severe claims was carried out thanks to the theory of extreme values. Hill's estimator refers to the severe claims' threshold as being around 90000 €. Once determined, we applied a GAMLSS Logistic to make our prediction. In addition, experts' reports allow us to refine our estimate.

Estimating the cost of attritional claims was more complex. The distribution of these claims did not seem to correspond to any known law. In order to circumvent this prob-

lem, a neural network was set up and its initialization was carried out from the regression coefficients of a GAMLSS Gumbel. This law estimated well but did not take into account the specific aspects of a storm. The combination of them has made the neural network faster and more efficient.

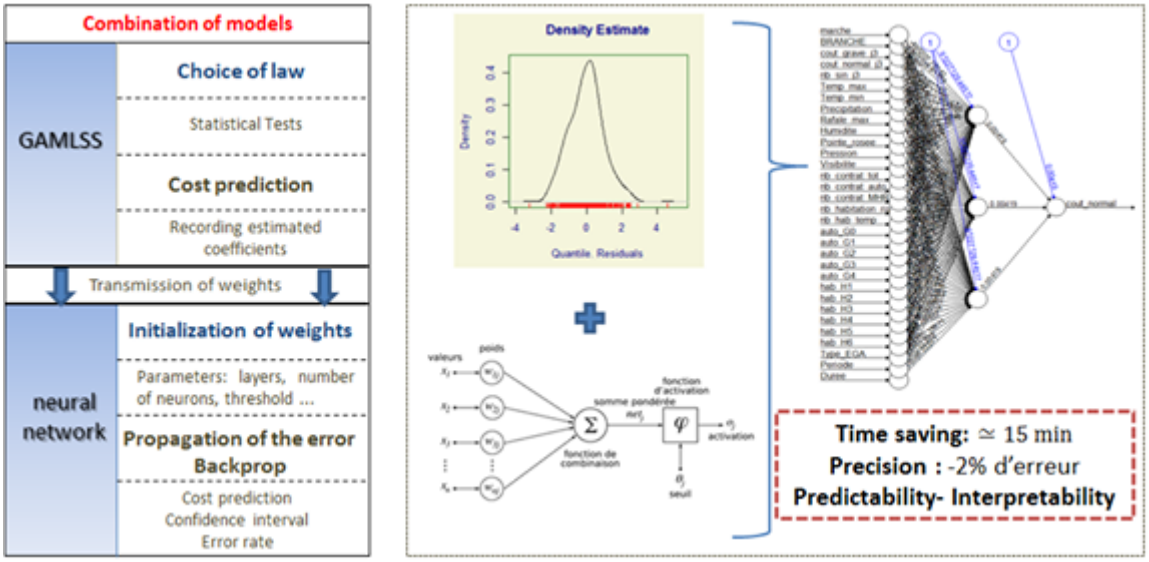


Figure 0.0.6: Diagram of the Neural Network with the GLM

Model results

This model was created on the principle of "divide and rule". Each module is independent but connected with others. Thus each module can be launched simultaneously allowing the execution to be much faster. Thanks to its different methods, the model always has a double validation. If one method differs from another, an in-depth analysis can be quickly conducted to identify the causes. Moreover, by linking human vision, statistics and data science, the model is robust, fast, reliable and easily interpretable. One of the strengths of this model is its adaptability. Indeed, it can be used to estimate the sinistrality of any climatic phenomenon if one provides the adequate history model.

Table des matières

Introduction

En France, une quinzaine de tempête surviennent chaque année avec une intensité plus ou moins forte. Il arrive cependant, que certaines, d'une ampleur exceptionnelle marquent la mémoire collective comme Lothar et Martin ou Xynthia.

Or ces tempêtes ont tendance à être sous-estimées par les compagnies d'assurance ce qui peut avoir des effets catastrophiques, notamment à cause de la non-constance de la fréquence et de l'intensité des tempêtes. Bien qu'aujourd'hui ce risque soit mieux appréhendé par les assureurs qu'en 1999, nombreux sont ceux qui ne disposent pas d'un modèle propre et qui n'ont pas une connaissance approfondie de ce risque.

Pourtant, la prédiction de ce risque est stratégique. Un assureur qui est en mesure de prédire correctement sa future sinistralité a un avantage concurrentiel majeur. L'assureur détenteur de cette information est alors en mesure de réduire l'impact des tempêtes et gagne la sympathie des clients. Il développe ainsi son image de « bon » assureur, et gagne par la même occasion de nouvelles parts de marché.

L'objectif de ce mémoire est de proposer un modèle pour Allianz qui permette d'estimer la sinistralité d'une tempête trois jours après sa survenance en France métropolitaine.

Dans un première temps, nous étudierons le risque tempête en météorologie et en assurance. Ensuite, nous verrons la construction de la base de données et l'ensemble des traitements nécessaires qui ont été appliqués pour que cette dernière soit complète, fiable et exhaustive. Enfin, nous appliquerons ces données à différents modèles afin d'estimer au mieux la sinistralité de nos tempêtes.

Partie I

La modélisation du risque tempête et ses enjeux

Chapitre 1

Le risque tempête en France

1.1 Identification et classification des événements tempêtes

1.1.1 Qu'est-ce qu'une tempête ?

Définition 1.1. Tempête : zone étendue de vents violents générés aux moyennes latitudes par un système de basses pressions (dépression).

En météorologie, le **vent** désigne le **mouvement horizontal** de l'air qui se mesure en prenant en compte deux paramètres : **sa direction et sa vitesse** (sa force). En France, la vitesse du vent est toujours une moyenne sur une période donnée et se mesure en km/h ou m/s. **L'origine des vents est due à une différence de pression.**

En assurance, un **événement tempête** peut être reconnu de deux manières différentes. Soit en demandant un certificat attestant de l'intensité exceptionnelle des vents (plus de 100km/h) par la station météorologique nationale la plus proche du sinistre, soit si le vent a causé des dommages à des bâtiments de bonne construction dans la commune où se trouvent les biens sinistrés ou dans les communes avoisinantes, alors on peut considérer l'événement en question comme une tempête.

1.1.2 Identification et classification des tempêtes

Identification

En France, l'**identification des tempêtes est réalisée par Météo France**. Pour qu'une série de vents violents soit considérée comme une tempête, les conditions suivantes doivent être réunies :

- Les rafales doivent être supérieures à **100km/h**
- La surface des vents maximaux observés doit couvrir au **minimum 2% du territoire**

Pour vérifier les conditions énoncées précédemment, Météo France se sert de capteurs, **fiables à 90%**, qui suivent les normes énoncées par l'**Organisation Météorologique Mondiale** (OMM, en anglais World Meteorological Organization ou WMO). Cette organisation créée en 1950, a pour but d'établir les normes permettant de standardiser les

mesures météorologiques et ainsi faciliter les échanges internationaux.

En 1950, une météorologue allemande a proposé de nommer les tempêtes afin de rendre plus efficace et plus simple la lecture des cartes météorologiques. Dans la suite de ce mémoire, nous garderons la nomenclature proposée par cette météorologue et utilisée par Météo France depuis lors.

Classification

Les **conséquences d'une tempête** (nombre de victimes, coût ...) peuvent être très **différentes suivant les caractéristiques de la tempête** (trajectoire, durée, surface touchée, vitesse...). Il est donc utile de les classer selon différents critères liés à la géographie afin de mieux les analyser.

Il existe différentes approches possibles pour classifier une tempête. Météo France, a privilégié **la méthode Dreveton** pour ranger les tempêtes. Cette classification a été établie en 1997 et permet de classifier les tempêtes en douze types correspondant à des caractéristiques météorologiques issues de données observées. La méthode consiste à regrouper les tempêtes en fonction de 3 critères :

- L'origine de la tempête
- L'origine de la dépression
- La zone touchée

L'ensemble de ces critères permet de déterminer le type de dépression.

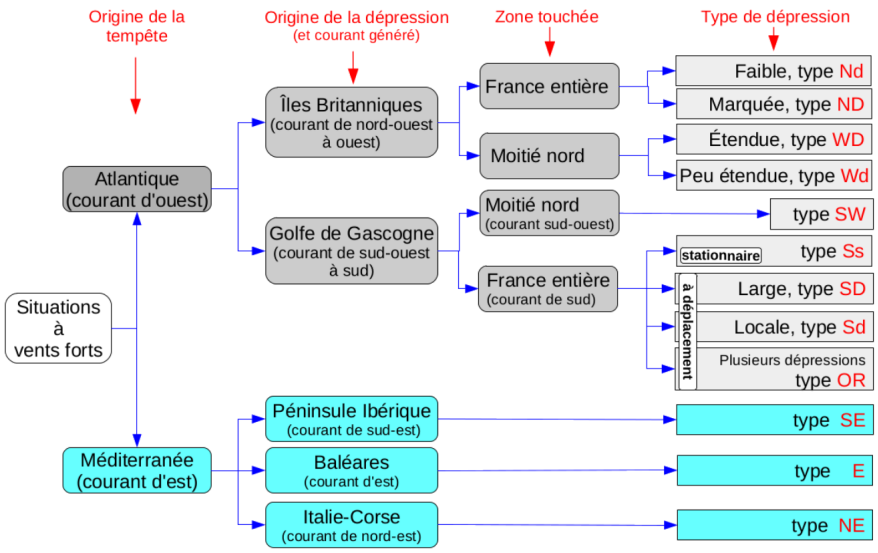


Figure 1.1.1: Schéma de la méthode de classification Dreveton - Source : Météo France

1.1.3 Les phénomènes associés aux tempêtes

Le phénomène des tempêtes s'accompagne généralement de phénomènes climatiques secondaires tels que la grêle, la pluie, de petites tornades, des vagues-submersions ... Ces phénomènes annexes sont bien souvent à l'origine du coût souvent exorbitants des tempêtes.

De plus, les **submersions marines** sont un phénomène récurrent lors de la survenance d'un **EGA**(Événement naturel de grande ampleur) tempête. Elles ont la particularité d'être très localisées, dangereuses et coûteuses. Elles surviennent lorsqu'il y a combinaison de plusieurs événements climatiques :

- **l'intensité de la marée** : plus le coefficient est fort, plus le niveau de la marée haute sera élevé
- **le passage d'une tempête** produisant une surélévation du niveau marin appelée **surcote**

Ces deux phénomènes conjoints provoquent un **déferlement de vagues sur la côte** conduisant à des destructions de biens matériels et à une infiltration des eaux qui aggravent les dégâts d'une tempête.

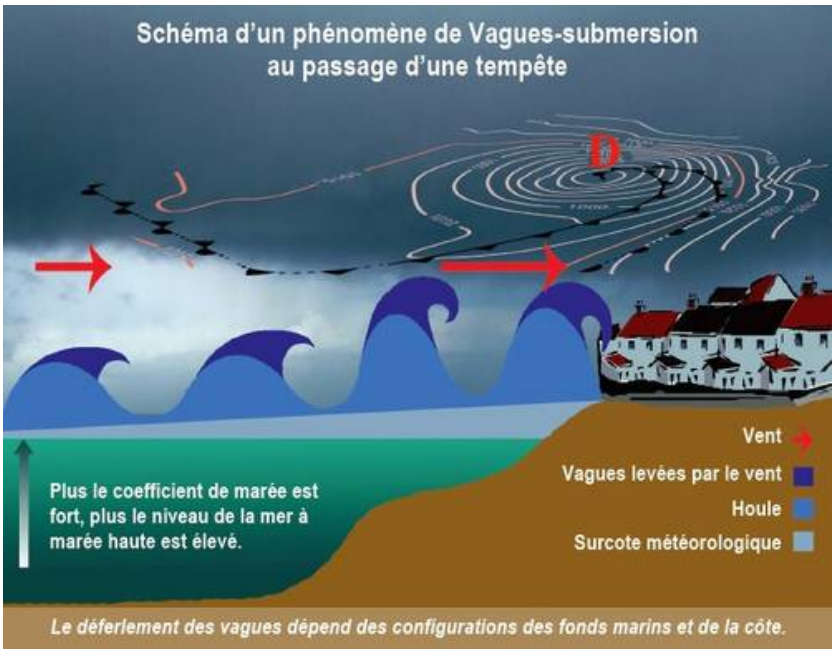


Figure 1.1.2: Schéma d'un phénomènes de vagues-submersion au passage d'une tempête-Source : Météo France

Ce phénomène a été très visible lors de la **tempête Xynthia**. L'eau de la mer est montée par endroits à plus de deux mètres dans les habitations. Cette nuit-là, les conditions atmosphériques ont provoqué une surélévation du niveau de la mer (surcote) de

1.53 mètre à La Rochelle alors que le niveau de la mer était au plus haut. La mer avait alors dépassée de plus d'un mètre le niveau des plus grandes marées.

Dans le cas des tempêtes provoquant des inondations, Allianz découpe l'événement en deux événements distincts (une événement tempête et une inondation). Dans la suite, nous étudierons uniquement les conséquences du vent même si une variable nous permet de prendre en compte la présence de vagues submersions pendant l'EGA.

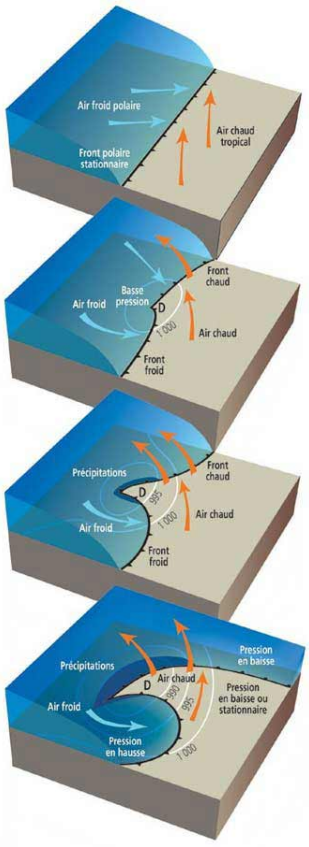
1.1.4 La difficulté à prévoir ses événements

La difficulté de prédiction des tempêtes réside dans l'estimation de leur intensité et de la zone géographique touchée. En effet, Météo France est capable de prédire la survenance de ces dernières deux ou trois jours à l'avance mais connaît des difficultés à déterminer avec précision la trajectoire de la tempête. Or, pour une compagnie d'assurance, la trajectoire est essentielle car elle permet d'anticiper l'impact que cette dernière aura sur le portefeuille. Le portefeuille d'un assureur n'étant pas homogène sur l'ensemble du territoire, la somme assurée diffère en fonction des zones géographiques touchées.

1.2 La France, un terrain propice aux tempêtes

1.2.1 Une inégalité Européenne face aux vents

En Europe, entre les années 1950 et 1990, **25 tempêtes et tornades** ont provoqué **la mort de 3 500 personnes** environ et ont coûté près de 4 milliard d'euros. Plus récemment, quatre tempêtes : Ana, Bruno, Carmen et Eleanor se sont produites en trois semaines (entre le 9 décembre 2017 et le 5 Janvier 2018). Cette **multiplication des tempêtes hivernales** résulte directement de la position géographique de notre continent. La partie nord de l'Europe est la plus fréquemment touchée car elle est située dans l'axe de la trajectoire empruntée par une grande partie des tempêtes d'hiver (axe Sud-Ouest / Nord-Est).



En plus d'avoir une configuration géographique propice aux tempêtes, l'Europe connaît un **adoucissement de ces températures hivernales** ces dernières années. L'hiver 2017 a, par exemple, été l'un des hivers **les plus doux depuis 150 ans**. Or, la formation des tempêtes se fait grâce à une variation de température (voir figure) d'où la multiplicité des tempêtes ces dernières années.

La configuration particulière de la France fait qu'elle est plus fortement exposée aux aléas climatiques de grande ampleur que ses voisins européens. Une étude réalisée par l'International Disaster Database de l'Institut de Louvain a montré que **la France est le 2e pays européen le plus touché par les événements naturels très grave depuis 1900** juste après l'Allemagne.

En effet, en France, les tempêtes sont principalement dues au **courant jet**. Ce courant est un tube de vents forts situés à environ 10 000 mètres d'altitude et de 600 à 100km de largeur. En été, ce courant se trouve très au nord de la France mais en hiver celui-ci se trouve au large de la France. Or, lorsqu'une masse d'air chaud rencontre le côté polaire du jet cela provoque des mouvements verticaux qui s'étirent sur toute la hauteur de la troposphère. **La dépression et la masse nuageuse associées** provoquent des vents violents au sol.

1.2.2 Une inégalité départementale

Le territoire français n'est pas soumis uniformément aux vents. Certaines zones sont plus ventées que d'autres, comme par exemple:

- **Sur les littoraux** : Manche, Atlantiques, Méditerranée
- **En montagne** : à proximité des cols et des crêtes
- **Aux débouchés des vallées**

Les cartes ci-dessous permettent d'étudier l'impact des tempêtes sur les départements français entre 2003 et 2017.

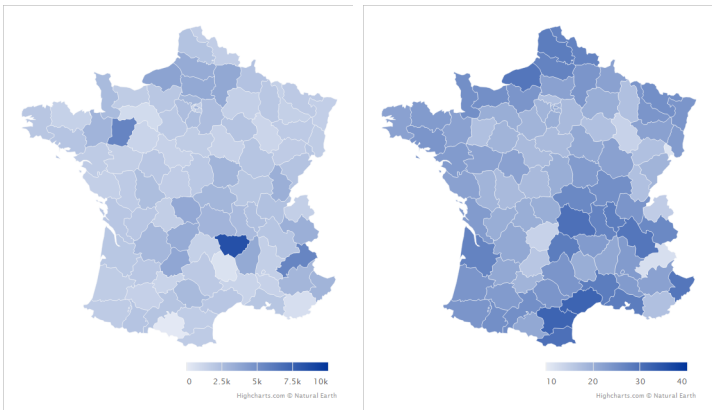


Figure 1.2.3: Cartographique du coût moyen des tempêtes en 2016 et du nombre de tempêtes entre 2003 et 2018 en France métropolitaine

Bien que l'ensemble du territoire soit impacté par les tempêtes, on remarque que certains départements sont plus touchés que d'autres. Lorsque l'on s'intéresse au coût moyen

des tempêtes en 2017, on se rend compte de l'inégalité de la population française face aux tempêtes. Ce constat renforce notre volonté de traiter les départements séparément dans la prédiction de la sinistralité des tempêtes.

1.3 Réchauffement climatique et tempêtes

Définition 1.2. Le **réchauffement climatique** est un phénomène global de transformation du climat caractérisé par une **augmentation générale des températures moyennes**, et qui modifie durablement les équilibres météorologiques et les écosystèmes.

Lorsque l'on parle du réchauffement climatique aujourd'hui, on parle du phénomène d'augmentation des températures qui se produit depuis 100 à 150 ans. Depuis le début de la Révolution Industrielle, les températures moyennes sur terre ont plus ou moins augmenté régulièrement. En 2016, la température moyenne sur la planète terre était environ 1 à 1.5 degrés au dessus des températures moyennes de l'ère pré-industrielle (avant 1850).

1.3.1 Vers une augmentation du risque tempête ?

Depuis la COP21 (COnférences des Parties n°21), les médias évoquent régulièrement l'idée selon laquelle le réchauffement climatique aurait un impact important sur la survenance et l'intensité des événements naturels catastrophiques. Or, ce fait est inexact d'après Météo France : « **aucune tendance climatique ne peut être établie sur l'évolution de l'intensité des tempêtes au cours des dernières décennies** » comme nous le montre le graphique suivant :

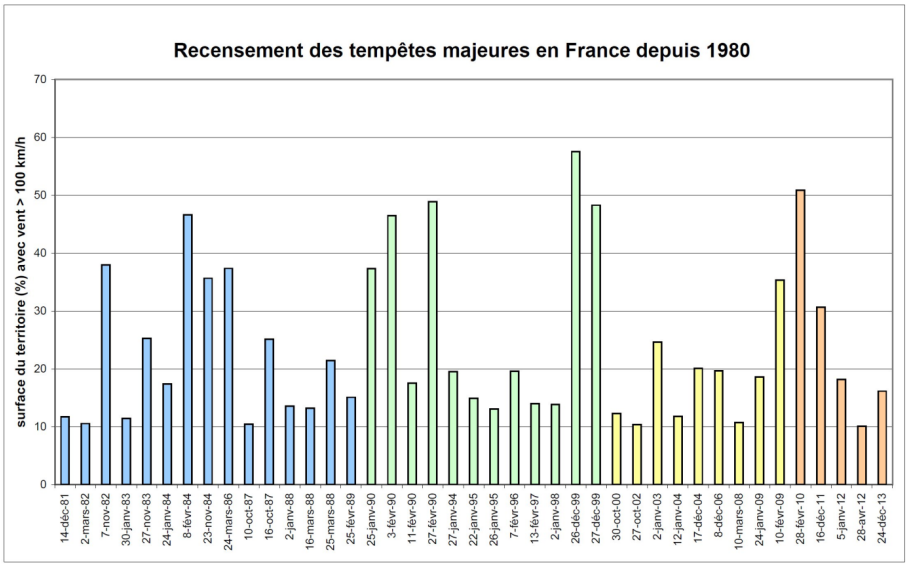


Figure 1.3.4: Recensement des tempêtes majeures en France depuis 1980 - Source: Météo France

Météo France va même plus loin en annonçant qu'en l'état actuel des connaissances, ils ne sont pas en mesure d'affirmer que les tempêtes seront plus nombreuses ou plus violentes en France métropolitaine au cours du 21e siècle.

Par conséquent, il a été décidé de ne pas prendre en compte le réchauffement climatique dans notre modèle.

Cependant, **le projet ANR-SCAMPEI** réalisé par météo France de 2009 à fin 2011 sur l'évolution des vents les plus forts à horizon 2030 à 2080 a montré que le modèle ALADIN-Climat prévoit une faible augmentation des vents forts au Nord et une faible diminution au Sud sur l'ensemble du 21e siècle.

Au vu de ces résultats, nous avons pu valider notre hypothèse selon laquelle le département était un critère clef dans notre analyse.

1.3.2 Mise en place de stratégies pour réduire le risque

D'un point de vue légal

La **loi du 2 février 1995** – communément appelée **loi Barnier** – a mis en place des **plans de prévention des risques (PPR)** permettant de clarifier, de simplifier et de rendre plus opérationnel le dispositif de prévention des risques et notamment les risques naturels. Ces plans réglementent la construction d'habitation (interdiction de construire dans certains zones ou sous certaines conditions). Ils n'empêchent pas la survenance des événements naturels catastrophiques mais **permettent d'en limiter les dégâts** et donc le coût.

Mise en place de nouvelles normes de construction

Afin de protéger les hommes et leurs biens contre les tempêtes, le gouvernement français a mis en place des **normes de construction** visant à minimiser l'impact du vent et de la neige. Ces normes sont spécifiques à chaque région car la France est inégalement soumise aux tempêtes. Dans les zones les plus sensibles comme le littoral ou les vallées, les constructions doivent être adaptées aux vents régionaux (pentes de toits, orientation des ouvertures...). Ces normes sont appelées **Normes de Construction conformité Neige et vent**.

Meilleure signalétique de Météo France et FIR d'EDF

En 1999, Météo France s'est trompée dans sa prévision des tempêtes Lothar et Martin. De plus, les informations transmises à la population comme la vitesse du vent n'étaient pas parlantes, ce qui a amené les gens à minimiser le risque qu'ils encouraient. Afin de

remédier à cela, Météo France a mis en place un **système de cartographie** avec des couleurs (du vert au rouge) **pour mieux informer le public**. Lorsque Météo France met un département en alerte rouge, un système d'alerte déclenche des mesures comme la suspension des services de transports. Ce système a permis en 2009 lors de la tempête Klaus de sauver des vies.

A cela s'ajoute la mise en place d'une **force d'intervention rapide** en 2009 appelée **FIR** de la part d'EDF qui permet en cas de crise de raccorder le plus rapidement possible les foyers et les entreprises au système électrique. Ces interventions rapides ainsi que l'enfouissement des lignes électriques permettent de limiter les coûts dus à la suspension d'activités professionnelles.

Prévention de la part des assureurs

Les assureurs jouent un rôle majeur dans la prévention des risques naturels majeurs. C'est pourquoi ils ont décidé en 2000 de créer l'association **Mission Risques Naturels (MRN)** qui a pour but de favoriser la compréhension des risques naturels et de sensibiliser la population à l'importance de la prévention. Elle développe également des méthodes d'expertise et de conseil à l'usage des assureurs pour les entreprises et les particuliers. Elle teste et valide les techniques permettant aux assureurs d'affiner la connaissance des enjeux économiques, l'évaluation des cumuls de risques et des dommages réels pour chaque société comme pour la profession dans son ensemble. Cette association propose également des conseils de prévention auprès des particuliers pour réduire les conséquences d'une tempête. Par exemple, dans leur lettre d'information n°26 publiée le 8 septembre 2017, ils montrent qu'un élagage régulier des arbres à proximité des maisons, une inspection annuelle des cheminées et l'attache des portails permettraient de réduire les coûts des tempêtes.

De plus, grâce à son application Allianz WeatherSafe, Allianz envoie un message d'alerte pour prévenir d'un risque probable d'événements naturels de forte intensité à ses assurés.

Le rôle de la mémoire

Il n'existe pas à l'heure actuelle de moyen de se prémunir contre les évènements naturels de grande ampleur. Les leçons tirées des précédents événements permettent néanmoins d'en atténuer les causes et les conséquences (nouvelles normes, application plus rapide des lois, informations massives de la population ...).

Les scientifiques s'accordent à dire que **le plus important pour limiter l'impact des catastrophes naturelles en général est de les garder en mémoire.**

Malheureusement, l'homme oublie ces phénomènes rapidement[1]. **C'est donc dans ce rôle de mémoire que les réseaux de neurones peuvent être intéressants.** Ils peuvent remplacer la mémoire humaine, en stockant et traitant des années d'informations. Ainsi, le savoir est conservé, et on peut tirer des leçons du passé.

1.3.3 Les conséquences des tempêtes

Les conséquences des tempêtes sont souvent importantes du fait de la **pluralité** des phénomènes climatiques associés (vent, pluie, vagues...) et des zones géographiques touchées souvent étendues.

Le danger des tempêtes réside dans les rafales de vents qui font s'envoler les toits, cassent des branches voir couchent des arbres et causent de nombreux dégâts tant humains qu'environnementaux. **L'intensité des dégâts est exponentielle à la vitesse du vent.**

Les enjeux humains

Les tempêtes ne font que peu de blessés et de morts contrairement aux sécheresses ou aux inondations et ceux-ci sont souvent dus aux phénomènes annexes de la tempête (blessure causée par la chute d'un arbre, noyade provoquée par une mer déchaînée . . .).

Les enjeux environnementaux

Les répercussions des tempêtes sur la faune et la flore sont nombreuses. Certaines sont directement liées aux tempêtes comme la destruction des forêts tandis que d'autres sont des effets indirects tels que la pollution du littoral. Pour limiter ces impacts, de nombreuses mairies mettent en place des groupes de protection de l'environnement dont la mission est de nettoyer les plages et les différentes sources d'eau des déchets emportés par le vent.

Les enjeux économiques

Les tempêtes sont à l'origine de dégâts matériels, pertes agricoles et perturbations d'activités importantes. L'ensemble de ces pertes conduit à un coût élevé, souvent en millions d'euros, que doivent supporter assurés et assureurs.

[1] Garnier2010

Chapitre 2

Le risque Tempête en assurance IARD

De tout temps, les populations ont cherché à se protéger contre les risques naturels (tempêtes, inondations, sécheresses, raz de marée...). On retrouve dès le 5ème siècle avant JC des traces de pratiques assurantielles visant à lutter contre les désastres naturels en Chine. A partir de cette date, l'assurance contre les phénomènes climatiques n'a cessé de prendre de l'ampleur et est apparue sur presque l'ensemble des continents à des degrés divers et sous des formes différentes.

2.1 Principe de l'assurance et de l'indemnisation

2.1.1 Le concept d'assurabilité

Définition 2.1. Une assurance est un moyen permettant à une personne appelée « l'**assuré** » de gérer les risques et de bénéficier du secours de l'assureur en cas de survenance d'un sinistre. En souscrivant une assurance, l'assuré transfère le coût d'une **perte potentielle** à une compagnie d'assurance en échange d'une somme d'argent appelée « **prime** » ou « cotisation » que l'assuré est tenu de verser selon les conditions et termes du contrat.

2.1.2 Fonctionnement de l'indemnisation

Définition 2.2. Indemnisation : compensation financière destinée à réparer un dommage.

On distingue deux processus différents dans la collecte d'information par un assureur. Dans le cas où le sinistre dépasse une certaine somme, l'assureur fait appel à un expert qui se rend chez l'assuré pour évaluer le montant des dommages au titre de la garantie souscrite. L'indemnisation de l'assuré se base sur cette expertise. Dans le cas contraire, l'assureur fait confiance à son assuré pour lui donner les informations nécessaires à l'estimation du montant des dommages subis. Cela permet un gain de temps et d'argent mais entraîne une perte d'information apportée par l'expert.

Figure 2.1.1: Schéma de la procédure d'indemnisation avec expertise

2.2 Différence entre Catastrophes Naturelles et Tempêtes

2.2.1 Principes de bases

L'indemnisation des conséquences matérielles des risques climatiques et naturels se fait sous certaines conditions : la nature de l'événement climatique mais également selon les biens sinistrés. Elle s'articule autour de 3 régimes spécifiques :

- Tempête
- Catastrophes naturelles
- Calamités agricoles

Les tempêtes et les catastrophes naturelles ne suivent pas le même régime car les catastrophes naturelles ne respectent pas toutes les conditions d'assurabilités qui sont :

1. L'événement doit être aléatoire
2. Il ne doit pas y avoir d'anti-sélection géographique
3. Le type et la gravité de l'événement peuvent être évalués pour un **prix acceptable**

Tempête Grêle Neige (TGN)	Catastrophes naturelles (CATNAT)
• Type d'événement connu • Évaluation possible de sa gravité • Aléa quasi-total • Pas d'anti-sélection géographique • Risque assurable	• Gravité très variable • Risque plus ou moins géographique • Anti-sélection géographique → prix de l'assurance plus élevée dans les zones plus exposées

Tableau 2.2.1: Comparaison des TGN et des CATNAT

Au vu des différents points, **les tempêtes sont assurables et pour un prix acceptable tandis que les catastrophes naturelles ne le sont pas**. Dans ce cas, c'est le principe de solidarité qui s'applique.

En France, il existe deux types de régime :

- **Un régime assurantiel "normal"** contractuel avec une assurance de marché et une réassurance privée pour les dommages considérés comme assurables (dommages causés par la tempête, la grêle ou le poids de la neige)

- **Un système mixte** faisant appel à la fois à l'État et à l'assurance avec l'État réassureur en dernier ressort, dans le cadre du système catnat instauré par la loi du 13 juillet 1982.

Il est donc important de comprendre que la garantie tempête est indépendante de toute notion de catastrophe naturelle. Il n'est pas nécessaire qu'un arrêté de reconnaissance de l'état de catastrophe naturelle soit publié au Journal officiel pour faire jouer la garantie TGN contrairement aux sinistres causés par une catnat.

Pour en savoir plus, une annexe récapitulant la prise en charge des risques naturelles est disponible.

2.2.2 La garantie TGN

En France, le risque naturel le plus fréquent est le risque de tempête. Cette garantie est inscrite dans le code des Assurances grâce à la loi n° 90-509 du 25 juin 1990 article L 122-7 qui stipule que :

« Les contrats d'assurance garantissant les dommages d'incendie à des biens situés en France ainsi qu'aux corps de véhicules terrestres à moteur ouvrent droit à la garantie de l'assuré contre les effets du vent dû aux tempêtes, ouragans ou cyclones, sur les biens faisant l'objet de tels contrats.

Sauf en ce qui concerne les effets du vent dû à un événement cyclonique pour lequel les vents maximaux de surface enregistrés ou estimés sur la zone sinistrée ont atteint ou dépassé 145 km/h en moyenne sur dix minutes ou 215 km/h en rafales, qui relèvent des dispositions des articles L. 125-1 et suivants du présent code.

Sont exclus les contrats garantissant les dommages d'incendie causés aux récoltes non engrangées, aux cultures et au cheptel vif hors bâtiments. Sont également exclus les contrats garantissant les dommages d'incendie causés aux bois sur pied. En outre, si l'assuré est couvert contre les pertes d'exploitation après incendie, cette garantie est étendue aux effets du vent dû aux tempêtes, ouragans ou cyclones. »

Sont notamment exclus de la garantie toutes les causes de dommages qui ne proviennent pas des circonstances précitées, et liés à la chose assurée :

- les constructions non ancrées selon les règles de l'art
- les dommages résultant d'un manque d'entretien indispensable

- les bâtiments non entièrement clos et couverts ou comportant certains matériaux tels carton ou feutre bitumé, etc., non fixés
- les parties vitrées de la construction
- les biens en plein air

2.2.3 Les limites de la garantie : les sinistres graves

Nous avons vu précédemment que le réchauffement climatique n'entraînait pas une augmentation du nombre de tempête cependant, il semblerait que l'évolution du climat tende à aggraver les sinistres majeurs. Cette tendance, qui concerne principalement les tempêtes et inondations, est observée dans le monde entier. Ainsi, les assureurs voient les coûts liés aux événements naturels augmenter graduellement à travers le monde. On peut expliquer cette aggravation par l'évolution de la matière assurée, un accroissement de la densité de la population, ainsi qu'une concentration des valeurs assurées dans des zones plus fortement exposées au risque.

Il est donc important de se concentrer sur ces sinistres graves, qui ont un impact très fort sur le coût global des tempêtes. Pour ce faire, notre modèle sera découpé en plusieurs parties dont l'une se concentrant uniquement sur les sinistres graves.

2.3 Le client au cœur du risque

Afin de limiter l'impact dramatique que pouvait avoir les événements naturels de grande ampleur sur les clients, la direction de l'Indemnisation d'Allianz France a mis en place l'**Unité Mobile d'Intervention (UMI).** Ce camion de 55m^2 aménagé sert en cas d'EGA de support logistique aux Agents Généraux au plus près des clients (accueil client, espace de convivialité, postes de gestion, salle de réunion...).

Il permet également de **coordonner les actions sur les terrains** (assistance, relogement, mesure d'urgence, expertise, réparation) et de mettre en place, heure par heure, des mesures pour venir en aide aux clients sinistrés et les aider dans leur démarches directement après un EGA.

Cette prise en charge rapide a de nombreux avantages : **rassurer les clients**, **les aider dans leur démarche**. Par ailleurs, grâce à cela, Allianz a une **connaissance plus rapidement de l'étendue des sinistres**. Ainsi, la prédiction de la sinistralité est plus juste, le service client meilleur et la connaissance de la nature de nos coûts à venir approfondie.

L'ensemble de ces actions sur le terrain et en interne (mise en place de la solution Simplissimo, équipe d'intervention prête à doubler le nombre de d'agents pour recevoir les appels...) permettent de créer un cercle vertueux autour du client et de **réduire l'impact de ces événements naturels** chez les personnes concernées.

Chapitre 3

La modélisation de ce risque

3.1 Par Météo France

Chez Météo-France, la prévision des tempêtes se fait grâce à plusieurs modèles. Tout d'abord, le **modèle européen CEPMMT** fournit des prévisions jusqu'à 10 jours d'échéance. Ces prévisions sont affinées grâce à des **modèles nationaux Arpège et Aladin**.

Les résultats des modèles sont alors analysés et interprétés par des ingénieurs prévisionnistes. Il s'agit, d'une part, de traduire les résultats numériques sous une forme utilisable et, d'autre part, de soumettre ces résultats à un examen critique afin de discriminer l'information fiable de l'information incertaine et, le cas échéant, de détecter les signaux précurseurs d'événements dangereux. Les vérifications des prévisions de modèles opérationnels font l'objet de scores objectifs qui permettent de suivre l'évolution de leur qualité.

L'ensemble de ces modèles permettent à Météo France d'établir des cartes de vigilance pour la prévention d'événements naturels de grande envergure.

3.2 Par Allianz

3.2.1 Cadre et contraintes

Le modèle d'estimation de la sinistralité des tempêtes a pour but d'estimer le coût moyen et le nombre de sinistres pour d'une tempête 3 jours ouvrés après sa survenance (J1).

Le cadre

Le but de cette estimation est de faire une **communication rapide et compréhensible** à nos interlocuteurs internes et externes. En effet, cette estimation de la sinistralité sert :

- **A la direction financières (actuariat et comptable)** : l'estimation est intégrée au **compte de résultat** sous forme d'**IBNR**(Incurred but not reported).
- **A la direction Indemnisation** : qui prend les décisions d'urgence pour gérer au mieux l'impact des tempêtes
- **Aux interlocuteurs externes** : tels que les médias par exemple ou la FFA

La Direction Indemnisation est en charge de l'estimation de la sinistralité des tempêtes depuis 2009. Seules les tempêtes de plus de 20 millions d'euros donnaient lieu à un chiffrage à J+3. **Depuis 2016, l'ensemble des tempêtes sont étudiées à partir du moment où le coût prévisible est supérieur à 3millions d'euros**.

Les contraintes

Dans ce type de modélisation, les contraintes sont nombreuses mais la plus importante reste celle de **la non-connaissance de nos sinistres**. En effet, un client a jusqu'à 5 jours pour déclarer son sinistre donc lors de notre estimation à J+3, nous ne connaissons pas l'intégralité des dommages subis par nos assurés. De plus, même lorsque les sinistres sont connus, il est rare que les experts aient eu le temps de passer et de rendre une estimation globale des dégâts.

A cela s'ajoute des contraintes de communication. Les résultats obtenus étant communiqués à de nombreux interlocuteurs, il est nécessaire qu'ils soient clairs et facilement interprétables.

3.2.2 Modélisation par Chain Ladder "arrangé"

La méthode de Chain Ladder est la **méthode de provisionnement** la plus répandue sur le marché de l'assurance non-vie, essentiellement pour les sinistres à payer. Cette méthode déterministe permet de projeter des valeurs observées jusqu'à l'ultime (extinction de tous mouvements sur les sinistres).

Cette méthode a été détournée afin d'estimer le coût des tempêtes dans un but d'indemnisation et non de provisionnement. Les lignes ne sont plus des années de survenance mais des tempêtes de références et les colonnes des jours de déclarations.

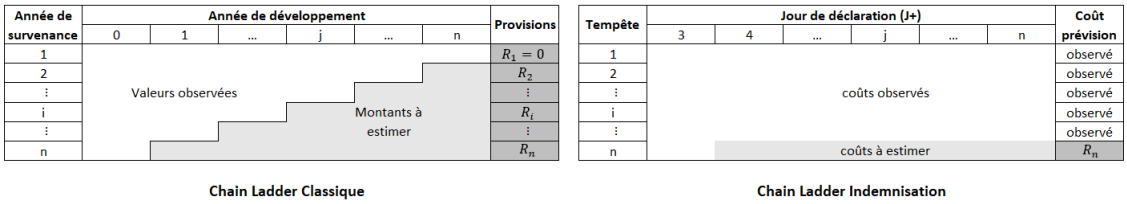


Figure 3.2.1: Schéma Chain Ladde

L'objectif de la méthode est d'estimer le montant global d'une tempête en se servant des tempêtes passées.

- Hypothèses du Chain Ladder

1. **Les tempêtes sont indépendantes entre elles** $\iff$ les vecteurs $C_{y,1}, ..., C_{y,k}$ et $C_{z,1}, ..., C_{z,k}$ avec y$\neq$ z sont indépendants
2. **les facteurs de passage f_k sont stables** par tempête de référence. Pour k = 1 ,..., K, il existe un paramètre f_k tel que le $E(C_{i,k+1}|C_{i,k}) = f_k * C_{i,k}$

- Hypothèse d'Allianz

L'hypothèse d'Allianz repose sur le fait que le Chain Ladder fonctionne si les tempêtes choisies en ligne sont des tempêtes de référence c'est-à-dire des tempêtes qui sont proches de la tempête à étudier d'un point de vue coût et zone géographique touchée.

Ces tempêtes sont choisies de manière arbitraire par la personne en charge de l'estimation. Elle se sert de son expertise ainsi que des remontées terrains faites par les experts pour estimer au mieux les tempêtes de référence.

Notations

- i (en ligne) = la tempête de référence
- j (en colonne) = le nombre de jour écoulé depuis la tempête
- $Y_{i,j}$ = les coûts des sinistres de la tempête i, j jour après sa survenance
- $C_{i,n}$ = les coûts cumulés où $C_{i,j} = \sum_{j=1}^{n-1} Y_{i,j}$
- $N_{i,j}$ = le nombre cumulé de sinistres pour la tempête i vu au bout de j jours

La Méthode

1. On part d'un triangle de données non-cumulées et on le cumule

$$C_{i,n} = \sum_{j=1}^{n-1} Y_{i,j} \tag{3.1}$$

2. On calcule les coefficients de passage à partir du triangle de données cumulées

$$f_{l,k} = \frac{C_{i,n+1}}{C_{i,n}}, 1 \leq i \leq l-1, 1 \leq n \leq k-1 \tag{3.2}$$

Afin de respecter la première hypothèse de chain Ladder, il est possible d'exclure des ratios de passage, surtout s'ils sont atypiques.

3. Calcul des facteurs de développement

$$\hat{f}_k = \frac{\sum_{i=1}^{I-j} C_{i,j} * F_{i,j} \hat{*} 1_{i,j}}{C_{i,j} * 1_{i,j}} \tag{3.3}$$

avec $1 \leq j \leq J-1$

$et\ 1_{i,j} = \begin{cases} 1\ si\ \hat{F_{i,j}}\ n'est\ pas\ exclu \\ 0\ si\ \hat{F_{i,j}}\ est\ exclu \end{cases}$

Les $\hat{f}_k$ sont des estimateurs sans biais des f_k.

Il est ensuite possible d'estimer les valeurs du triangle inférieur grâce aux estimateurs :

$$\hat{C_{i,j}} = C_{i,I-i} * \prod_{k=I-i}^{j-1} \hat{f}_k \tag{3.4}$$

$\forall i, j\ tels\ que\ i+j > I$

Exclusion d'un ratio de passage : Toute exclusion doit pouvoir être justifiée (exemple : présence d'un grand nombre de sinistres tardifs exceptionnels du à un évenement climatique de fin d'année).

4. Projections des données observées.

Pour cela, on utilise les facteurs de développement

$$\hat{C_{i,j+1}} = \begin{cases} C_{i,j} * \hat{f}_j\ pour\ i+j = J+1 \\ \hat{C_{i,j}} * \hat{f}_j\ pour\ i+j > J+1 \end{cases} \tag{3.5}$$

avec $J+1 \leq i+j \leq 2*J$

Ainsi, on obtient une estimation des ultimes

5. Calcul des coûts

- Par jour : $\hat{R}_i = \hat{C_{i,j}} - \hat{C_{i,j-1}}$
- Total : $\hat{R} = \sum_{i=1}^{k} \hat{R}_i$

Les avantages de la méthode

- Elle est facilement paramétrable
- Elle est utilisée à l'international
- Elle s'applique à de nombreux types de données : réglements, charges, recours, nombre de sinistre, coûts moyens, etc...

Les inconvénients

- La première hypothèse sur laquelle repose Chain Ladder suppose l'existence de coefficient de passage entre $C_{i,j}$ et $C_{i,j+1}$. Cela signifie qu'elle ne renvoie pas d'"effet diagonale" des valeurs : la progression de la déclaration des tempêtes est supposée identique, quelque soit l'année de survenance. Cette hypothèse se vérifie grâce à un graphique (les couples $C_{i,j}$ et $C_{i,j+1}$ sont sensiblement alignés sur une droite passant par l'origine).
- L'autre limite réside dans l'estimation des coefficients de passage, notamment le dernier paramètre $\hat{f}_k$ qui n'est estimé qu'à l'aide de deux données. Si le modèle est multiplicatif, l'erreur d'estimation peut présenter une distorsion importante du résultat final.
- Cette méthode ne fait aucune hypothèse quant à la loi que peut suivre les coûts ou leur fréquence.

3.3 Vers une nouvelle approche de modélisation ?

Les **modèles statistiques "classiques"** tels que le GLM(modèle linéaire généralisé) ont l'avantage de fournir une certaine **compréhension des données et du mécanisme qui les a engendrées** à travers une représentation parcimonieuse d'un phénomène aléatoire. De plus, la grande force des statistiques classiques réside dans leur **capacité à évaluer la fiabilité d'un résultat** (par exemple grâce à la p_value). Or, cette notion est absente des modèles de machines learning. Cependant, elles ont tendances à avoir un **biais important** ce qui entraîne des **erreurs dans les prédictions**. A cela s'ajoute l'**obligation de choisir une loi** qui représente les données.

Grâce à la data science, les modèles de "machine learning" c'est-à-dire les **modèles auto-apprenants, s'exemptent d'hypothèse sur les lois, ce qui leur permet d'obtenir des résultats plus précis**. Cependant, cela se fait souvent **au détriment de l'interprétabilité**. On peut résumer les avantages et les inconvénients de chacun grâce au tableau suivant :

Méthodes	Avantages	Inconvénients
Classiques	• Interprétabilité des paramètres • Analyse de l'impact des variables • Communication plus aisée	• Biais important (erreur de modèle) • Choix d'une loi
Machine learning	• Moins de biais • Peu ou pas d'hypothèses • Résultat plus précis	• Interprétabilité complexe • Analyse de l'impact des variables plus compliquée • Souvent phénomène de "boîte noire"

Tableau 3.3.1: comparaison des avantages et des inconvénients des statistiques traditionnelles et des modèles de machine learning

En conclusion, il sera intéressant de lier les deux méthodes pour obtenir des résultats à la fois précis et interprétables.

Partie II

L'essor de la DataScience en Actuariat dans l'analyse explorative des données

Chapitre 1

La Data science et l'actuariat

1.1 Définition

Les données sont au cœur du métier d'actuaire. Avant l'essor de la datascience de ces dernières années, l'actuaire n'était généralement confronté qu'aux données de son entreprise et à quelques indices de marché. Désormais, il est assailli de **données diverses** et doit trouver un moyen de les utiliser à bon escient.

L'enjeu pour l'actuaire est alors de savoir **combiner son savoir traditionnel** avec les **méthodes émergentes** afin de trouver des **approches innovantes** permettant de répondre à des problèmes toujours plus complexes et d'utiliser des données jusqu'à présent inexploitables. En effet, le niveau de sophistication des modèles de machine learning permet parfois de donner une explication plus fine de la sinistralité. Cela est principalement dû à l'absence d'hypothèse (ou au peu d'hypothèse) de ces modèles.

La data science permet donc un **gain d'information** car elle autorise l'ajout d'un nombre important de **variables de différents types** (structurée, non structurée, continue, catégorielle...) ce qui n'est pas le cas des méthodes traditionnelles.

1.2 Le fléau de la dimension

Le terme **fléau de la dimension** (ou curse of dimensionality en anglais) a été inventé par Richard Bellman en 1961 pour parler du **problème de l'augmentation explosive du volume de données** associé à l'ajout de dimensions supplémentaires dans un espace mathématique.

Ce phénomène aussi appelé **malédiction de la dimension** est un obstacle majeur dans les algorithmes d'apprentissage et encore plus dans les réseaux de neurones. En effet, lorsque la dimension de l'espace des variables augmente c'est-à-dire lorsque le nombre de variables augmente, **les données deviennent éparses et éloignées**, ce qui **biaise et fausse les résultats**. De plus, l'augmentation du nombre de variables rend l'interprétation des modèles de plus en plus compliquée. Augmenter le nombre de variables permet donc de **gagner de l'information et de la précision** mais **rend les modèles moins lisibles**.

"Leo Breiman donne l'exemple de 100 observations couvrant l'intervalle unidimensionnel [0,1] dans les réels : il est possible de dresser un histogramme des résultats et d'en tirer des inférences. En revanche, dans l'espace correspondant à 10 dimensions $[0,1]^{10}$, les 100 observations sont des points isolés dans un vaste espace vide, et ne permettent pas l'analyse statistique. Pour réaliser dans $[0,1]^{10}$ une couverture équivalente à celle des 100 points dans [0,1], il ne faut pas moins de 10^{20} observations – entreprise gigantesque et souvent impraticable." - exemple issu de Wikipédia.

On retrouve ce fléau dans les réseaux de neurones. Pour qu'ils soient efficients, le nombre de données nécessaires est égal à : μ^p avec :

- μ :le nombre moyen de neurone par couche
- p : la profondeur du réseau = Nombre de couche * nombre de nœuds par couche

Donc si on a un réseau avec une unique couche cachée de 3 nœuds, on a besoin de $\mu^p = 3^3 = 27$ données alors que si on le réseau passe à 10 nœuds on a besoin de 10^{10} données.

Pour parer à cette malédiction de la dimension, il est possible de faire appel à des méthodes de réductions de dimension. Il est par exemple fréquent d'utiliser des méthodes par extractions de variables telle que l'ACP pour réduire la dimension tout en conservant le maximum d'information.

1.3 Le dilemme prédictibilité / interprétabilité

Ce sempiternel dilemme s'est renforcé avec les méthodes de machine learning et leur « boîte noire ». Il s'agit de faire un **compromis entre un modèle de prédiction précis avec une interprétabilité limitée et un modèle moins performant mais plus transparent et plus exploitable.**

Plusieurs méthodes peuvent être envisagées pour appréhender ce phénomène. Ce mémoire s'est axé sur :

- L'utilisation la Datavisualisation pour avoir des rendus plus interprétables
- La combinaison des méthodes traditionnelles aux réseaux de neurones pour gagner en interprétabilité et en précision.

1.4 Les limites de la datascience : le risque de surapprentissage

Les modèles auto-apprenants ont deux inconvénients majeurs : la difficulté d'interprétabilité et le risque de sur-apprentissage.

Le **sur-apprentissage** ou **overfitting** est l'une des causes principales des mauvaises performances des modèles prédictifs générés par les algorithmes de machine learning avec le **sous-apprentissage** (underfitting). En statistique, le surapprentissage ou surajustement est une analyse statistique qui correspond trop étroitement ou exactement à un ensemble particulier de données et qui peut donc ne pas correspondre à des données supplémentaires ou ne pas prévoir de manière fiable les observations futures. **Un modèle surajusté est un modèle statistique qui contient plus de paramètres que ne peuvent le justifier les données.**

De par sa trop grande capacité à stocker des informations, **une structure dans une situation de surapprentissage aura de la peine à généraliser les caractéristiques des données.** Elle se comporte alors comme une table contenant tous les échantillons utilisés lors de l'apprentissage et perd ses pouvoirs de prédiction sur de nouveaux échantillons. Le surapprentissage s'interprète comme un apprentissage « par cœur » des données. Il résulte souvent d'une trop grande liberté dans le choix du modèle (mauvais paramétrage du modèle).

Pour limiter ce type de phénomène dans le cas des réseaux de neurones, il est nécessaire de veiller à utiliser un nombre adéquat de neurones et de couches cachées. Cependant, ces paramètres sont difficiles à déterminer. Il est donc possible de mettre en place des méthodes qui permettent de contrôler l'apprentissage pour éviter le sur-apprentissage telles que :

La validation croisée

Pour détecter un surapprentissage, on sépare les données en deux sous-ensembles : **l'ensemble d'apprentissage qui correspond à 70% de nos données et l'ensemble de validation (qui correspond aux 30% de données restantes)** L'ensemble d'apprentissage permet de faire évoluer les poids du réseau de neurones avec par exemple une rétropropagation. L'ensemble de validation (exclu de l'apprentissage) permet de vérifier la pertinence du réseau avec sur un échantillon de données jusqu'alors inconnu. On peut vraisemblablement parler de surapprentissage si l'erreur de prédiction du réseau sur l'ensemble d'apprentissage diminue alors que l'erreur sur la validation augmente de manière significative. Cela signifie que le réseau continue à améliorer ses performances sur les échantillons

d'apprentissage mais perd son pouvoir de prédiction sur ceux provenant de la validation. Pour avoir un réseau qui généralise bien, on arrête l'apprentissage dès que l'on observe cette divergence entre les deux courbes. On peut aussi diminuer la taille du réseau et recommencer l'apprentissage.

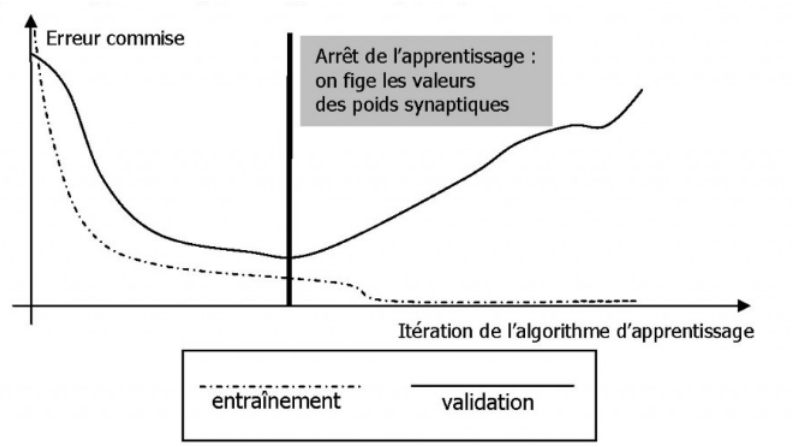


Figure 1.4.1: Exemple de régularisation

Le tracé en pointillé représente le coût d'erreur global du modèle à travers le temps sur l'échantillon d'entraînement. Le tracé complet représente la même information mais sur l'échantillon de validation. La ligne verticale quant à elle représente le point optimal pour le modèle prédictif. A droite de ce point on tombe dans le Overfitting, à gauche dans l'Underfitting.

La Régularisation

Une autre méthode permettant d'éviter le surapprentissage est d'utiliser une forme de régularisation. Durant l'apprentissage, on pénalise les valeurs extrêmes des paramètres, car ces valeurs correspondent souvent à un surapprentissage. Il faut cependant faire attention à ne pas trop les pénaliser pour ne pas risquer le sous-apprentissage. L'une des méthodes de régularisation les plus utilisées en réseau de neurones est la méthode du weight decay.

1.5 Choix des données grâce à Twitter

1.5.1 Twitter : un reflet de la pensée populaire

Twitter est devenu un **instrument de communication incontournable** ces dernières années. Avec ses tweets de 140 caractères, le site au plus de **300 millions d'utilisateurs dans le monde**, se démarque par sa capacité à toucher un **public très large**.

L'ensemble des "citoyens français" est concerné par le phénomène Twitter. Les hommes politiques, les sportifs, les dirigeants d'entreprise, les mères aux foyers l'utilisent pour

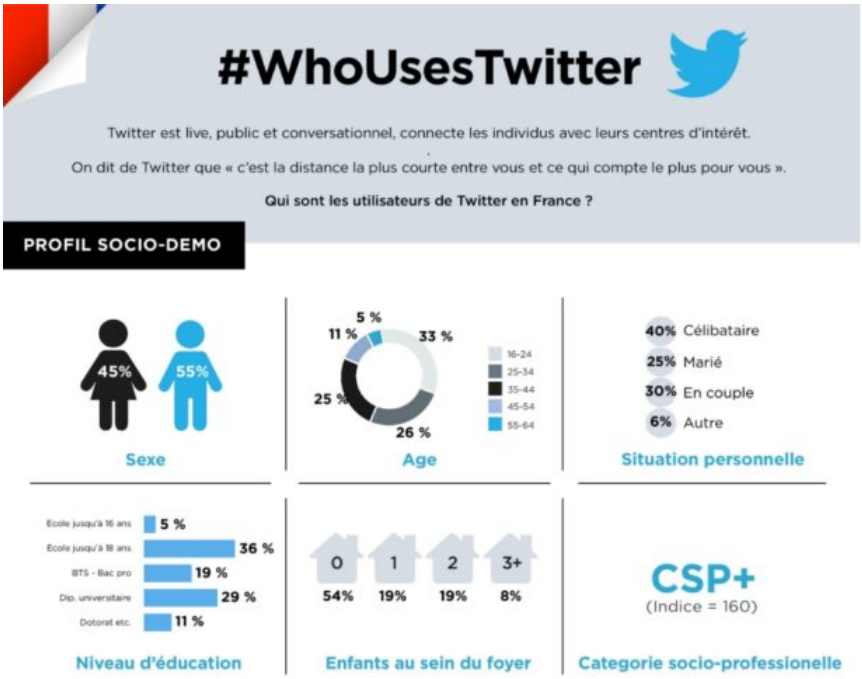


Figure 1.5.2: Répartition des usagers de Twitter - source : Twitter

évoquer l'actualité, leur opinons, leurs décisions futures. Chacun échange avec l'ensemble de la communauté, mettant ainsi à disposition une **multitude d'informations**.

Allianz revendique la volonté de placer le client au cœur de son activité. C'est pourquoi, nous avons décidé d'analyser les tweets à l'aide de méthode de text mining pour **savoir ce que pensaient les français des tempêtes**. Cette analyse de tweet va nous permettre de déterminer les mots/idées associés à celui de "tempête".

1.5.2 L'API Twitter et Text mining

Twitter dispose de plusieurs API (Application Programming interface) permettant de collecter les tweets et de les analyser. En effet, ces derniers étant publics, ils sont accessibles à tous et chacun peut les étudier à volonté.

Une API est une série de méthodes mise à disposition du programmeur permettant d'accéder ou d'utiliser certaines fonctionnalités du site depuis un programme informatique. Je me suis servie de cette API pour **récupérer tous les tweets français, publiés depuis le 1 janvier 2003 contenant le mots Tempête** (soit dans le texte du tweet soit dans un hashtag). Le but est de trouver les mots que les français associent à la tempête. Et comme une image vaut mieux que mille mots d'après Confucius, nous avons créé un nuage de points afin de visualiser les associations. Le résultat est le suivant :

Figure 1.5.3: Nuage des associations de mots Twitter

Pour obtenir un tel nuage, un travail de retraitement a été nécessaire. Il a fallu enlever les tweets inappropriés (vulgarité, éléments non pertinents...), éliminer les doublons, retirer les URL, harmoniser la case et l'orthographe de certains mots en enlevant les accents par exemple.

1.5.3 Résultat de l'analyse

Cette analyse nous a permis de tirer les conclusions suivantes :

- les français associent la tempête à des **événements climatiques**. On retrouve les mots vents, orages, pluie, marée ... **C'est pourquoi nous allons utiliser l'ensemble de ces phénomènes climatiques associés aux tempêtes dans notre modèle**
- **la prévention et la compréhension des enjeux climatiques** évoqués dans la partie précédente semble fonctionner. On remarque que le mot risque, changement climatique et réchauffement sont très présents.
- les français sont désireux d'être bien informés. Le **#infofiable** est apparu plus de 50 fois lors de l'analyse des tweets. Cela donne de l'importance à notre travail. **Un bon modèle permettra de transmettre une information fiable et rapide aux clients.**

Chapitre 2

Qualité des données

Mesurer la qualité de ses données, c'est connaître le taux d'erreur et les risques qui en découlent.

Les données sont au cœur du travail des compagnies d'assurance. Il est donc essentiel d'étudier la qualité de ces dernières avant de les utiliser. En effet, des données de qualités permettent d'entretenir de bonnes relations avec ses clients et de prendre des décisions rapidement. Afin de les aider, la commission du 10 octobre 2014 a publié un règlement délégué 2015/25, précisant les exigences en matière de qualité des données. Ces dernières s'articulent autour de trois critères (**l'exhaustivité, l'exactitude et le caractère approprié des données**) et sont décrites dans différents articles de loi (**en annexe**).

La pertinence de l'utilisation d'une donnée se détermine grâce à une analyse de cette dernière. On pourrait résumer l'ensemble du processus de validation d'une donnée de la façon suivante :

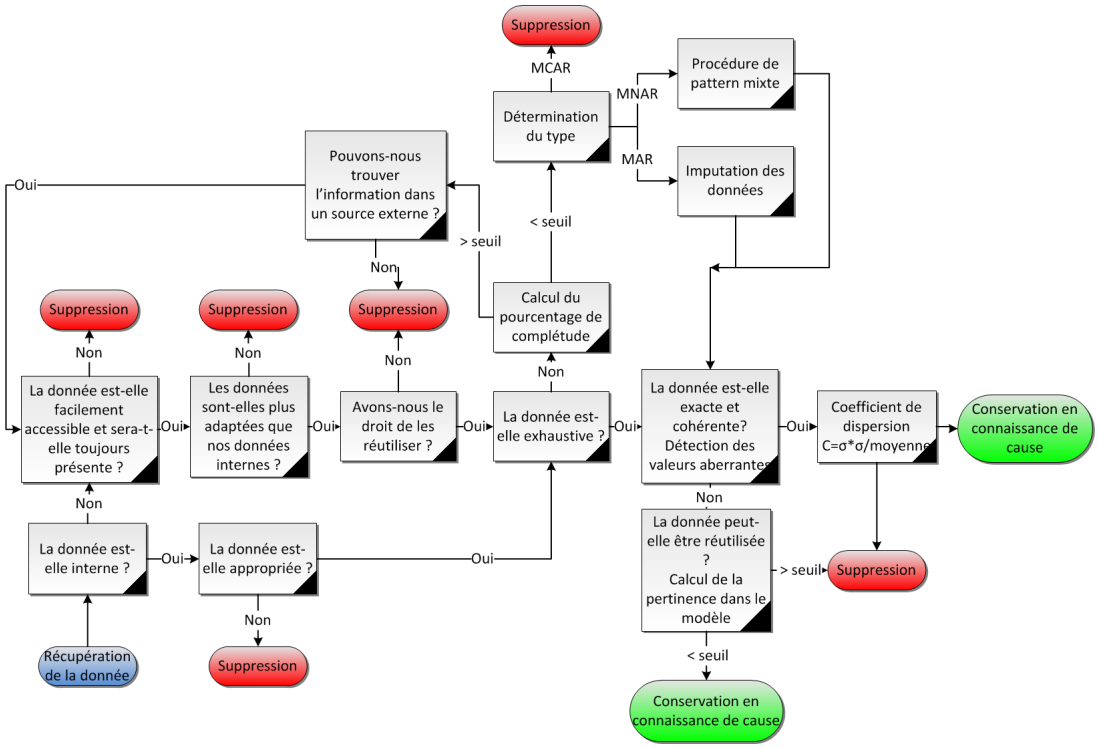


Figure 2.0.1: Schéma de validation d'une donnée

De plus, la directive **solvabilité II** précise l'importance d'une bonne qualité des données en plaçant ce concept au cœur de ces trois piliers.

2.1 Données internes, données externes, quelles différences ?

2.1.1 Qu'est-ce qu'une donnée ?

Définition 2.1. Une donnée est une description élémentaire d'une réalité. C'est par exemple une observation ou une mesure.

2.1.2 Les données internes

Les données internes sont l'ensemble des données internes à une entreprise. Il est fréquent que ces données proviennent d'une multitude de sources d'informations, ce qui peut occasionner des erreurs. Afin d'apprécier la qualité de ces données, il est possible de mettre en place plusieurs contrôles. Les résultats de ces contrôles doivent être analysés avec les experts métiers afin de voir s'ils sont « **acceptables** ». La qualité des données se décompose en 6 aspects : **la standardisation, le nettoyage, le dédoublonnage, le profilage (profiling), la surveillance et l'enrichissement**.

2.1.3 Les données externes

Les données externes (issues des réseaux sociaux, des smart phones, ou de « l'open data »), sont soumises à de nombreuses conditions supplémentaires par rapport aux données internes de l'entreprise. En effet, il est obligatoire de démontrer que chaque donnée externe utilisée est plus adaptée que les données disponibles en interne, que l'on peut mesurer leurs origines et disposer sur ces données d'une piste d'audit semblable à celle faite pour les données internes. **A ces contraintes s'ajoutent les questions sous-jacentes de droit, de protection des données et de sécurité de l'information.**

2.2 Définitions des critères relatifs aux données

Pour qu'une donnée puisse être utilisée dans un modèle, elle doit répondre à plusieurs critères qui sont :

- l'exhaustivité
- l'exactitude
- le caractère approprié des données

2.2.1 L'exhaustivité

L'exhaustivité consiste à avoir « suffisamment d'informations historiques pour qu'il soit possible d'apprécier les caractéristiques des risques sous-jacents et de dégager des tendances d'évolution des risques » (Article 19 du code des Assurance).

L'exhaustivité se divise en deux notions :

1. La notion de complétude
2. la notion d'accessibilité

Remarque : Lorsqu'une donnée pertinente vient à être exclue, il est impératif de justifier ce choix.

2.2.2 L'exactitude

Une donnée exacte est une donnée :

- Exempte d'erreurs importantes
- Cohérente (à un instant t et sur la durée)
- Enregistrée en temps utile

Lorsqu'une donnée semble inexacte, une **réconciliation** des données à partir de différentes sources de données ainsi qu'un **nettoyage** de ces dernières sont nécessaires.

Afin de valider la qualité des données, il existe des **tests d'exactitudes des données** (ceux-ci sont décrits dans la partie suivante).

L'exactitude peut également provenir de l'émetteur des données en cas de données externes. Dans notre cas, les données proviennent principalement de Météo France, les données météo qu'elle fournit sont jugées exactes, donc l'assureur n'a pas besoin de justifier de leur qualité.

2.2.3 Le caractère approprié des données

Une donnée est appropriée lorsqu'elle est adaptée aux besoins de l'utilisateur.
Par exemple : faire une étude du délai de clôture sur une branche longue n'est pas appropriée si l'on ne dispose que d'un petit historique.

2.3 Exhaustivité ou présence de donnée manquante

2.3.1 Qu'est-ce qu'une Valeur manquante ?

Définition 2.2. Soit une variable aléatoire X quelconque. Une **donnée manquante** (DM) x_m est une donnée pour laquelle la valeur $X = x$ est inconnue. On ne dispose pas de la valeur de X pour le sujet i.

Définition 2.3. La probabilité d'absence correspond à la probabilité qu'une donnée soit manquante suivant son type. Le mécanisme de données manquantes est caractérisé par la distribution conditionnelle de M sachant Y donnée par $\rho(M|Y)$

Il existe deux grandes catégories de valeurs manquantes : Les valeurs pour lesquelles on ne dispose d'aucune information sur un individu (non-réponse totale) et celles où l'information est incomplète (non-réponse partielle). Ces données peuvent être manquantes pour de nombreuses raisons: défaillance d'un capteur, mesure impossible, données perdues, données non disponible ...

La perte d'information entraîne un **biais** dans l'analyse et dans la cas présent dans la prédiction. Il est donc important de comprendre d'où proviennent ces pertes et de les traiter au mieux. Ce traitement est d'autant plus important que bon nombre de modèles statistiques n'acceptent pas les données manquantes.

2.3.2 Structure des données manquantes

Il existe trois types de répartition des données manquantes :

1. **Univariées** : Pour une variable Y_k, si une observation Y_{ki} est manquante, alors il n'y a plus d'observation de cette variable.
2. **Monotones** : L'ensemble des variables sont manquantes dès lors qu'une est manquante : $Y_{k,k>j}$ manquante si Y_j manquante.
3. **Arbitraires (non monotones)** : on définit la matrice de valeurs manquantes par $M = (m_{ij})$ avec $m_{ij} = 1$ si y_{ij} est manquant et zéro sinon.

2.3.3 La classification de Little et Rubin

Les données manquantes ne sont pas toutes de même **nature**. Elles sont classifiées selon les formes des mécanismes qui génèrent ces données. On utilise la **classification proposée par Little et Rubin** (1987) qui répartie les données manquantes en trois catégories :

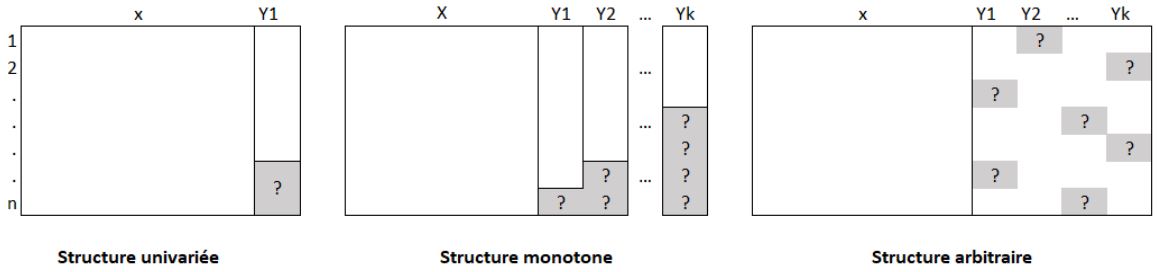


Figure 2.3.2: Répartitions des données manquantes (a)univariées, (b)monotones, (c) arbitraires

1. **Missing Completely At Random (MCAR)** : La probabilité qu'une observation soit manquante est constante c'est à dire qu'elle ne dépend ni des variables observées ni des variables non-observées. La probabilité d'absence est donc :

$$\forall Y, \rho(M|Y) = \rho(M)$$

   Ce type de valeur manquante est très rare. Cependant, lorsque cela arrive, il est possible de les ignorer sans que cela ne biaise l'analyse si leur nombre est restreint. Cependant, cela sera au détriment de la précision.

2. **Missing At Random (MAR)** : La probabilité qu'une observation soit manquante ne dépend que des valeurs observées. La probabilité d'absence est donc :

$$\forall Y^{miss}, \rho(M|Y) = \rho(M|Y^{obs})$$

   Ce type de données entraîne un biais conséquent dans l'analyse. Afin d'éviter cela, des méthodes d'imputations peuvent être mise en place. Nous verrons dans la partie suivante **les algorithmes d'imputations des données**.

3. **Missing Not At Random (MNAR)** : La probabilité qu'une observation soit manquante dépend des valeurs non observées c'est-à-dire si la distribution de M dépend aussi de Y^{miss}. Les données MNAR induisent une perte de précision ainsi qu'un biais ce qui nécessite d'avoir recours à une analyse de sensibilité pour l'analyse du modèle (par exemple utilisation de la procédure du pattern-mixture model).

2.3.4 Théorie de l'imputation des données

Il est fréquent dans les entreprises que les données proviennent d'une multitude de sources d'informations, ce qui peut occasionner des erreurs. Chez Allianz par exemple, on distingue trois systèmes d'informations différents dus aux nombreuses fusions et acquisitions survenues au cours du temps. Or les données étant la base des modèles, il est nécessaire

de retravailler ses données afin qu'elles soient le plus justes possible et exemptes d'erreurs.

L'erreur la plus fréquente est l'erreur d'échantillonnage c'est-à-dire l'erreur liée au processus de sélection d'un échantillon. Dans notre cas, cette erreur est restreinte car nous sommes dans un cadre bien précis. L'erreur suivante provient des valeurs manquantes. Ces dernières sont fréquentes voire inévitables dans une base de données, mais rares sont ceux qui consacrent le temps nécessaire à leur étude.

Plusieurs solutions sont envisageables pour gérer les données manquantes:

Suppression des individus ayant des données manquantes

Bon nombre d'ouvrages traitant des valeurs manquantes énoncent le fait que la façon la plus «propre» de gérer les données manquantes est de supprimer de l'analyse les individus dont l'information manque sur l'une des caractéristiques pertinentes. Ainsi aucun biais n'est introduit. Cette théorie est renforcée par la loi forte des grands nombres selon laquelle :

Théorème 2.4. *Soit $(X_n)_{n\leq 0}$ une suite de variables aléatoires indépendantes, identiquement distribuées. Posons*

$$S_n = X_1 + ... + X_n$$

et supposons $E[|X_1|] < \infty$. Alors $\frac{S_n}{n}$ converge presque sûrement vers $E[X_1]$ [1]

Dans notre cas cette solution n'est pas envisageable car cela reviendrait à perdre des individus (des tempêtes) qui sont déterminantes dans notre analyse. En effet, au vu du petit nombre de tempêtes présentes dans notre historique, une suppression reviendrait à perdre une quantité phénoménale d'informations. Il faut donc nécessaire de trouver une autre méthode.

Maintien des valeurs non renseignées

Certaines méthodes d'apprentissage statistiques ont la particularité de pouvoir composer avec les données manquantes. C'est le cas par exemple des arbres de régression CART qui s'appuient sur le **principe de divisions de substitutions**. Dans ce cas, une donnée manquante n'est pas considérée comme candidate potentielle pour la division optimale d'un nœud.

Au début de notre réflexion, cette méthode a été envisagée. Cependant, en l'optant, nous renoncions à la comparer aux méthodes statistiques classiques. Nous avons donc décidé de faire un autre choix : réaliser une imputation des données manquantes, comparer

[1] démonstration en annexe

les différents modèles pour choisir celui qui donne la meilleure prédiction mais également comparer les prédictions d'un arbre CART avec et sans imputation pour comprendre l'erreur qu'engendre une telle reconstitution.

Imputation des données manquantes

Cette méthode consiste à remplacer les données manquantes par des valeurs plausibles. Il existe de nombreuses méthodes d'imputations : remplacer la valeur manquante par la moyenne des données, par la valeur la plus récurrente ...

Ces approches ont été critiquées notamment dans l'article de Schafer et Graham (2002) pour le manque de précision qu'entraînent ces règles. Cependant les méthodes d'imputations restent une méthode assez robuste et efficace avec peu d'imputations. De plus, elles ont l'avantage de refléter l'incertitude due aux données manquantes dans les résultats. Ces deux avantages nous confortent dans l'idée d'utiliser une méthode d'imputation des données.

Choix de la méthode d'imputation : Méthode d'imputations multiples par équations chaînées[2]

2.4 Exactitude et détection des valeurs aberrantes

2.4.1 Qu'est-ce qu'une Valeur aberrante ?

Bien qu'il n'existe pas de définition mathématique claire sur ce qu'est une donnée aberrante, on pourrait définir une **donnée aberrante** comme étant une **valeur ou une observation qui est « distante » des autres observations effectuées sur le même phénomène** c'est-à-dire contrastant grandement avec les valeurs « normalement » mesurées.

Ces données aberrantes sont un problème car l'ensemble des **calculs statistiques** utilisant les propriétés de la loi normale sont **très sensibles à leur présence**. De plus, leur présence dans un modèle de machine learning peut rendre **la phase d'apprentissage** très longue et entraînera bien souvent un **sur-apprentissage** comme nous l'avons vu dans la partie précédente.

Il existe différentes méthodes pour détecter ces dernières et les traiter.

[2]voir détail de la méthode en annexe

2.4.2 Identification des valeurs aberrantes

La méthode la plus fréquemment utilisée est la **méthode graphique** de la boîte à moustache qui permet de visualiser la distribution d'une seule variable.

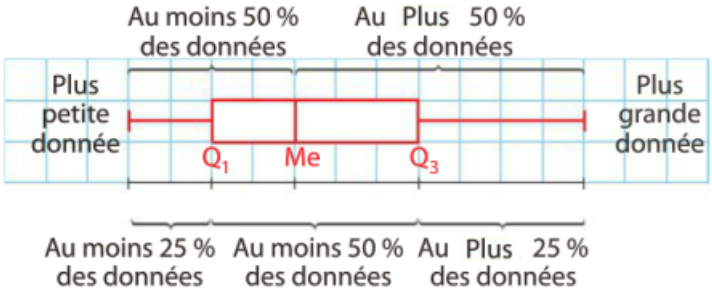


Figure 2.4.3: Boîte à moustache

Une valeur est considérée comme aberrante si elle se situe à l'extérieur de l'intervalle :

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$

avec :

- k une constante positive souvent égale à 1.5
- Q1 le premier quartile
- Q3 le troisième quartile

2.4.3 Traitement des valeurs aberrantes

Le choix de **conserver ou non la donnée aberrante dépend de la cause de cette dernière**. La suppression des valeurs aberrantes est une pratique controversée dans le milieu scientifique car il n'existe pas de méthode objective et quantitative permettant le rejet de certaines valeurs. Cependant, **le rejet d'une donnée est acceptable si la distribution des erreurs de mesure est connue**.

L'exclusion des données peut se faire par deux approches : **la césure** ou **la méthode de Winsorising**. La césure élimine les données aberrantes alors que le Winsorising remplace les données aberrantes par les valeurs « non suspectes » les plus proches.

En régression, une autre approche consiste à exclure uniquement les valeurs qui présentent un haut degré d'influence sur les coefficients estimés, notamment en utilisant une mesure telle que la distance de Cook. Cette méthode évite le sur-apprentissage dans les modèles d'apprentissage.

2.5 Les limites du traitement des données

2.5.1 Des infrastructures coûteuses

L'une des composantes clefs pour réussir à utiliser correctement le big data est l'infrastructure. En effet, le volume des données générées et stockées par les entreprises a explosé avec l'essor du big data et les entreprises se voient confrontées à la problématique du stockage de ces dernières. Elles doivent investir dans des moyens de stockage toujours plus grands (disque dur, cloud, serveur ....) et toujours plus coûteux (du fait de la réglementation, de la demande croissante ...), dans des ordinateurs plus puissants (pour pouvoir traiter rapidement les montages de données). La conséquence de ce phénomène est importante puisque beaucoup d'entreprises renoncent à utiliser le big data au quotidien pour des raisons logistiques et financières.

2.5.2 Des données de différents types

Les données proviennent de partout et peuvent être de structure très différentes. Il est possible de distinguer deux types : les données dites structurées ou non structurées. Les données structurées sont les données indexées, celles présentent dans les bases de données. Ce sont les données les plus faciles à utiliser. Les données non structurées sont beaucoup plus complexes. Ce sont les images, les vidéos, les textes... Leur traitement est souvent long et fastidieux mais apporte des informations uniques et souvent impossibles à obtenir autrement.

2.5.3 Une réglementation exigeante

« Lorsque les données ne satisfont pas aux dispositions de l'article 19, les entreprises d'assurance et de réassurance documentent de **manière appropriée les limites de ces données**, y compris en indiquant si et comment il y sera remédié et en précisant quelles fonctions de leur système de gouvernance seront responsables de ce processus. Les données sont enregistrées et stockées de manière appropriée avant de faire l'objet d'ajustements destinés à remédier à leurs limites » - Article 20 de la réglementation

2.5.4 La cyber-insécurité des données

En France, le nombre d'incidents de sécurité a augmenté de 34% entre 2013 et 2014. Les pertes financières liées à ces incidents sont estimées en moyenne à 2,9 millions de dollars en France et par entreprise. Si les nouvelles technologies offrent des opportunités, elles constituent également des risques qu'il convient d'intégrer en amont. Malheureusement, ces risques sont encore méconnus, et il est difficile de les estimer correctement.

Chapitre 3

Création de la base de données d'apprentissage

Choisir les données qui vont alimenter le modèle est compliqué. En choisissant trop peu de données, on fausse le modèle car on ne lui fournit pas assez d'informations pour qu'il estime et prédise correctement. En lui en donnant trop, on prend le risque de rendre notre modèle très long et de le biaiser avec des valeurs qui n'ont pas lieu d'être.

C'est pour cette raison que le travail préalable sur les données est si important. La sélection fait partie intégrante du modèle car elle l'alimente. Un bon modèle deviendra mauvais s'il n'a pas les données adéquates en entrée. Pour éviter ce type de problème, nous avons décidé de travailler de la façon suivante : prendre un maximum de variables et les éliminer au fur et à mesure en fonction de leur pertinence.

3.1 Récupération des données

3.1.1 Le cadre

Notre base de données est constituée de tempêtes passées :

- survenues entre 2003 et début 2018
- dont le coût est supérieur à 3 millions d'euros
- survenues en France métropolitaine
- exclusion du corporel et de la grêle agricole

Ce cadre étant différent de celui précédemment mis en place, quelques travaux ont du être menés. Auparavant, nous ne nous intéressions qu'aux tempêtes de plus de 20 millions d'euros survenues après 2009. Il a donc fallu rechercher les tempêtes survenues entre 2003 et 2009 ainsi que celles survenues après 2009 dont le coût était compris entre 3 et 20 millions d'euros.
Pour ce faire, plusieurs méthodes ont été utilisées :

- Recherche en interne dans les différentes documentations et programmes pour voir si certaines tempêtes n'étaient pas pré-analysées.

- Recherche de tempêtes historiques sur le site de Météo France.
- Recherche sur internet grâce aux différentes publications médiatiques
- Et enfin, les souvenirs des uns et des autres ont permis de reconstruire les dernières tempêtes manquantes.

On peut récapituler l'ensemble des données à notre disposition dans le tableau suivant :

	Nom de la variable	Type	Unité	Si catégorielle valeur	Donnée calculée
données tempête	Nom de la tempête	qualitatif			non
	Date de début	date	jj/mm/aaaa		non
	Date de fin	date	jj/mm/aaaa		non
	Durée	quantitatif	en jours		non
	Période	catégorielle		automne, hiver, printemps, été	non
	Liste de département	Liste			non
	Nb de mort	quantitatif			non
	Nb de blessés	quantitatif			non
	nb de foyers privé d'électricité	quantitatif			non
	Origine	catégorielle		E, N, NE, NO, O, S, SE, SO, NI	non
	Dreveton	catégorielle		Nd,ND,WD,Wd, SW, Ss, SD, Sd,OR,SE,E,NE,NI	oui
	Tempête avec submersion	catégorielle		oui,non	non
	Surface touchée	quantitatif	%		oui
Données Météo - Par département - 1er jour de survenance de la tempête	Température maximale	quantitatif	°C		non
	Température minimal	quantitatif	°C		non
	Précipitation	quantitatif	mm		non
	rafale maximale	quantitatif	km/h		non
	Humdité	quantitatif	%		non
	Pointe de rosée	quantitatif			non
	Pression	quantitatif	hPa		non
	Visibilité	quantitatif	m		non
	PPRN	quantitatif			non
	Coefficient de marée	quantitatif			non
	Hauteur de l'eau	quantitatif	m		non
Données portefeuille - Par département - 1er jour de survenance de la tempête	Nombre total de contrat dans notre	quantitatif			oui
	Nombre de contrat Auto	quantitatif			oui
	Nombre de contrat MRH	quantitatif			oui
	Nombre d'habitation à risque	quantitatif			oui
	Nombre de tempête déjà subi par	quantitatif			oui
	Classe du véhicule	catégorielle		G0,G1,G2,G3,G4	oui
	Classe de l'habitation	catégorielle		H1,H2,H3,H4,H5,H6	oui
Données sinistres	branche ega	catégorielle		Auto, DAB	oui
	marché	catégorielle		Part, Pro, Ent, Flotte, Construction, Autres	oui
	Top_MRH	catégorielle		0, 1	oui
	nouveau_cout	quantitatif			oui
	cout_j+3	quantitatif			oui
	nombre sinistre j+3	quantitatif			oui
	Nature de l'évenement	catégorielle			non

Figure 3.1.1: Descriptif des données utilisées

L'idée de départ était d'inclure la liste des départements d'une tempête dans le modèle. Cependant, après analyse, nous nous sommes rendu compte que cela alourdissait le modèle (donc le ralentissait) sans que l'information supplémentaire apportée soit significative. En effet, les départements touchés sont assez similaires du point de vue du coût moyen comme nous avons pu le voir dans la 1ère partie de ce mémoire. De plus, lorsque l'on étudie le nombre de contrat par département, on remarque qu'il est presque identique dans chaque département, excepté en Ile de France et dans le Nord où il est plus important. Fort de ce constat, nous avons décidé de ne pas inclure directement les départements dans le réseau de neurones. Cependant, cette information est apportée indirectement puisque chaque donnée (température maximale, nombre de contrat total ...) est calculée en fonction de ceux-ci.

Ainsi chaque tempête est découpée de la façon suivante en branches (Auto ou DAB) et marchés (Particulier, entreprise, professionnel, construction, flottes et groupement et

autres).

Ensuite, il est nécessaire d'étudier la corrélation entre variables. En effet, une corrélation trop forte (de plus de 80%, biaise l'analyse surtout dans le cas des réseaux de neurones). L'étude des corrélations donne le résultat suivant :

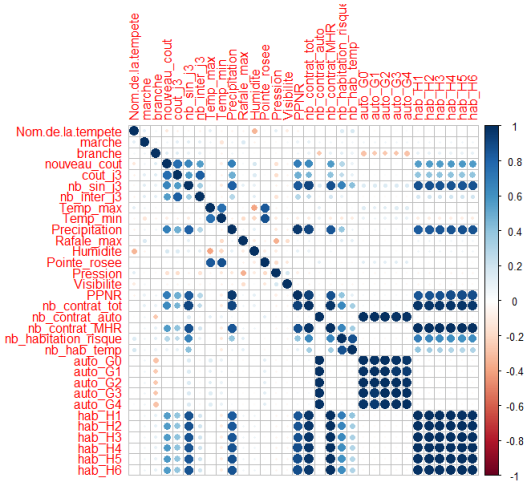


Figure 3.1.2: Graphique des corrélations

On remarque que les classes automobiles et habitations sont très corrélées entre elles. Afin de ne pas biaiser la sortie du réseau de neurones, les classes ont été revues afin que ces dernières soient moins corrélées.

3.1.2 Les contraintes

Dans une entreprise aussi importante qu'Allianz, les contraintes sont nombreuses.

La contrainte la plus importante concerne l'information (la donnée). Allianz dispose de plusieurs systèmes d'informations, ayant chacune des particularités. Un travail de mise en commun est donc nécessaire avant de pouvoir traiter l'intégralité de l'information.

La deuxième contrainte importante repose sur le temps d'exécution des programmes. Les modèles mis en place doivent être efficients et rapides. Le délai de rendu des résultats étant très courts (3jours), il est nécessaire que les programmes soient rapides dans l'exécution pour laisser du temps à l'analyse.

3.1.3 Les données utilisées

Nos variables sont de quatre catégories différentes :

- Les variables propres à une tempête : comme son nom, sa date de début et de fin, la liste des départements touchés ...
- Les variables météorologiques : on regarde pour chaque département les données météorologiques telles que la hauteur des précipitations ou encore les températures, provenant des informations fournies par Météo France
- Les variables relatives aux portefeuilles : pour chaque département touché, on regarde le nombre de contrat MRH et automobiles, provenant de nos bases de données portefeuille
- Enfin, les variables relatives aux sinistres, comme le nombre de sinistres déclarés ou encore le coût à l'ultime, provenant de nos bases de données indemnisation

Une fois les données récupérées, il est nécessaire de les analyser pour ne garder que celles qui sont cohérentes.

3.2 Qualité de nos données

3.2.1 Identification des données manquantes

La plupart des logiciels de modélisation statistiques n'acceptent pas les données manquantes. Deux choix se sont donc présentés à nous : soit n'utiliser que des méthodes auto-apprenantes acceptant les données manquantes (comme les random forest par exemple) soit reconstruire les données manquantes à l'aide d'une méthode d'imputation de données, pour pouvoir tester l'ensemble des méthodes et les comparer. Nous avons décidé de reconstruire nos données afin d'être en mesure de choisir le meilleur modèle possible.

Nos données manquantes concernant uniquement les données météorologiques, nous les avons récapitulées dans le tableau suivant :

	Humidité	Pointe de rosée	Pression	Visibilité
Nombre de valeurs manquantes	665	628	880	697
Pourcentage de valeurs manquantes	5.5%	5.2%	7.3%	5.8%

Tableau 3.2.1: Tableau de valeurs manquantes

On remarque que l'ensemble des valeurs manquantes est inférieur à 10%. On peut donc envisager une reconstitution de ces dernières. On pourrait s'interroger sur l'utilité de reconstruire la variable pression dont plus de 7% des données sont manquantes. Après création des différents modèles, nous nous sommes rendu compte que la pression était une variable d'influence donc qu'il était important de la conserver.

Une fois les variables manquantes identifiées, il est nécessaire de faire une analyse pour identifier le type de données manquantes auquel on est soumis (MAR, MCAR,MNAR). Pour cela, deux méthodes peuvent être utilisées :

- Le test MCAR[1]
- La méthode graphique

Dans les deux cas, on trouve que nos données sont MAR donc qu'il est possible d'appliquer une imputation des données :
Le test MAR, réalisé avec R, nous indique que les données manquantes sont présentes sous 9 patterns différents. La statistique de test donne une p_value à 0.0001 pour un χ^2 de 571.98 de 32 degré de liberté. La p_value étant inférieure à 0.05, on rejette l'hypothèse selon laquelle les données sont MCAR. Par conséquent, nous pouvons supposer que nos données sont MAR.

La méthode graphique est plus parlante. Elle nous permet de vérifier que les données manquantes sont réparties de manière aléatoire. Nous remarquons ici que c'est le cas donc les données sont MAR.

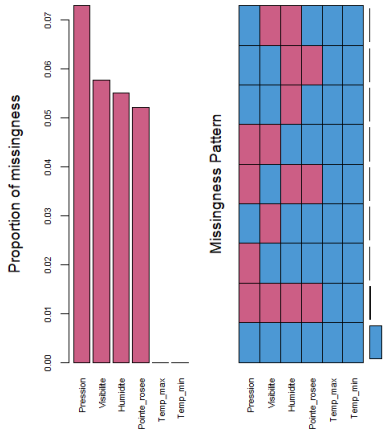


Figure 3.2.3: Schéma des données manquantes

Nos données étant MAR, il est possible de les reconstruire grâce à une méthode par imputation.

3.2.2 Détection des valeurs aberrantes

La détection des valeurs aberrantes a été effectuée pour chaque variable grâce aux boîtes à moustaches. Le graphique suivant est un exemple de trois boîtes à moustache obtenus.

Dans le premier cas, la température moyenne ne connaît pas de valeurs aberrantes (représentées par de petits cercles extérieurs à la chandelle). Le cas de la pression est

[1]décrit en annexe

intéressant, car il en ressort des valeurs aberrantes situées au niveau des faibles pressions. Or en France, les baromètres indiquent un risque de tempête si la pression est inférieure à 980hPa. Les valeurs aberrantes présentées ici sont donc les pressions atmosphériques qui correspondent à une tempête. Ce graphique nous indique que la pression est une donnée importante car c'est un indicateur précieux du type de tempête auquel nous sommes confrontés.

Le dernier cas concerne la proportion de sinistres déclarés en fonction du nombre de contrats. On remarque qu'il existe un nombre de sinistres dont la proportion est anormalement élevée. Après analyse du portefeuille, nous nous sommes rendu compte que cette proportion correspondait au marché entreprise. En effet, une entreprise possède un unique contrat pour l'ensemble de ces succursales/filières. Cette étude nous a poussés à découper notre sinistralité en fonction du marché.

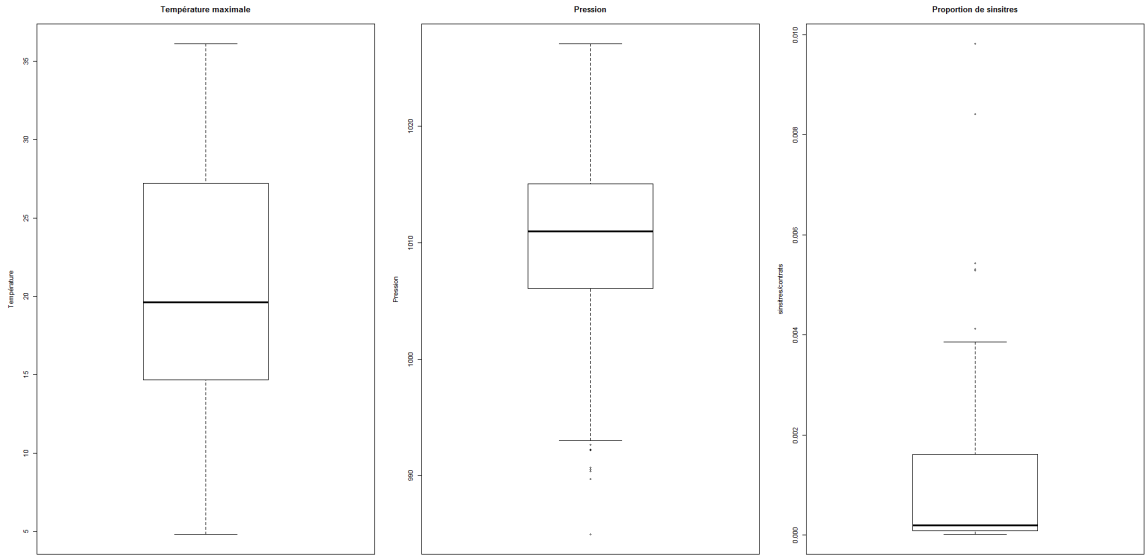


Figure 3.2.4: Exemple de boîte à moustaches

3.3 Nettoyage de nos données et reconstruction

3.3.1 Imputation des données manquantes

L'imputation des données a été réalisée grâce au package MICE (Multivariate Imputation by Chained Equations) de R qui suit la méthode décrite par Van Buuren et Groothuis-Oudshoorn en 2011. Chaque variable est imputée selon son propre modèle.

L'imputation de nos données manquantes peut se faire car nous sommes en présence de données MAR. Pour reconstruire les données, on utilise non seulement les données observées de la variable à reconstruire mais également les variables qui peuvent être en lien. On parle alors de prédicteurs.

Sélection rapide des prédicteurs

Pour déterminer les prédicteurs nécessaires, nous utilisons la corrélation (calcul des corrélations mutuelles entre les données et les indicateurs de réponse). La proportion de cas utilisables mesure le nombre de cas où des données manquantes ont réellement un impact sur la variable cible. Cette proportion sera faible si la cible et le prédicteur sont manquants dans les mêmes cas.

	Mois	DEPT	Temp_max	Temp_min	Precipitation	Rafale_max	Humidite	Pointe_rosee	Pression	Visibilite
Mois	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
DEPT	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Temp_max	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Temp_min	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Precipitation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rafale_max	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Humidite	0.008	-0.086	0.071	0.044	-0.069	0.403	NA	0.010	0.013	0.074
Pointe_rosee	0.007	-0.059	0.073	0.039	-0.069	0.402	NA	NA	-0.009	0.071
Pression	0.005	-0.050	0.087	0.084	-0.053	0.361	0.087	0.111	NA	0.061
Visibilite	-0.002	-0.111	0.097	0.066	-0.062	0.377	-0.062	0.009	0.027	NA

Figure 3.3.5: Tableaux de corrélations

Une méthode plus complexe de sélection des prédicteurs existe pour les données volumineuses. En règle générale, l'utilisation de chaque information disponible produit des imputations multiples qui ont un biais minimal et une certitude maximale (Merg 1995, Collins et al 2001). Ce principe implique que le nombre de prédicteurs doit être aussi grand que possible. Selon Von Buuren et al 1999, le mieux est de choisir un sous-ensemble de 4 à 25 variables.

On lance l'algorithme sur l'ensemble de nos données. Pour savoir si l'algorithme a convergé, une étude de la convergence est nécessaire. Dans le cas contraire, il faut augmenter le nombre d'itérations, le nombre de jeux de tests ou encore changer de méthode d'imputation.

Étude de la convergence

L'étude de la convergence se fait souvent de manière graphique. Prenons l'exemple de la variable Visibilité, on trouve le schéma de converge suivant :

On remarque que les différents flux sont entremêlés les uns avec les autres sans distinction de tendance. De plus, la variance entre chaque séquence étant inférieure à la variance de chaque séquence, on peut en conclure qu'il y a convergence. Une fois que l'algorithme a convergé, il faut examiner les imputations créées. Une bonne valeur imputée est une valeur qui aurait pu être observée si elle n'avait pas été manquante.

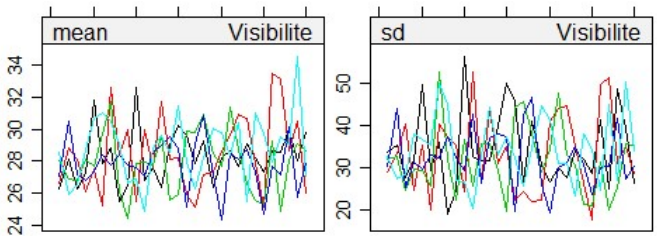


Figure 3.3.6: Validation des choix d'imputation

Choix de la méthode d'imputation

Nous avons vu qu'il existait plusieurs méthodes pour imputer nos données (CART, Random Forest, Predictive mean matching, Moyenne et Norme). Pour trouver quelle méthode convient le mieux, nous avons utilisé : La méthode graphique et le test de Kolmogorov. Prenons par exemple le cas de la reconstruction de la variable avec les méthodes PMM, CART et RF.

- **Méthode graphique**

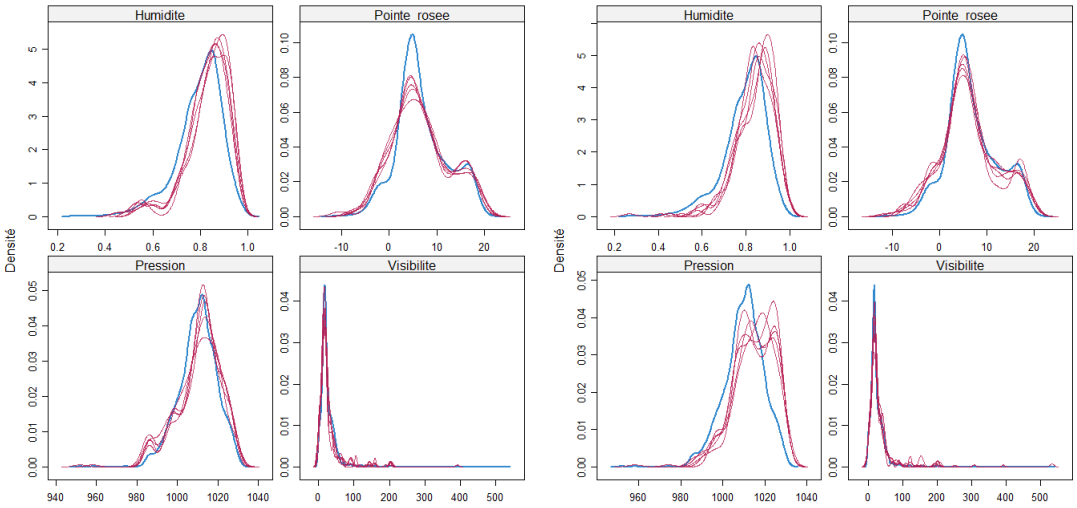


Figure 3.3.7: Visualisation des résultats de l'imputation - CART&PMM

- **Test de Kolmogorov**

Définition 3.1. Soit H0 l'hypothèse nulle : les deux groupes proviennent d'une même distribution. Cette dernière est rejetée si la statistique de test

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}} = 0.0296$$

Où

- $Dn, m = \sup_x |F_{1,n} - F_{2,n}|$ avec $F_{1,n}$ et $F_{2,n}$ les distributions du 1er et 2e échantillon.
- $\alpha = 0.05$
- $c(\alpha) = 1.36$

En appliquant ce test sur l'ensemble de nos variables pour les différentes méthodes, nous obtenons le résultat suivant :

	Humidité	Pointe de Rosée	Pression	visibilité
CART	D=0.08	D=0.01	D=0.009	D=0.007
RF	D=0.08	D=0.009	D=0.02	D=0.006
PMM	D = 0.01	D = 0.007	D= 0.04	D=0.01
Mean	D = 0.04	D=0.04	D=0.05	D=0.06
Norm	D = 0.1	D=0.01	D=0.03	D=0.03

Tableau 3.3.2: Résultats du test de Kolmogorov pour différents types d'imputation

Étude des résidus

Sous l'hypothèse MAR, la répartition des résidus doit être similaire à la répartition des données imputées de sorte que les distributions se chevauchent. Prenons l'exemple de la variable Pression, qui a été imputée avec la méthode PMM et CART. Nous avons vu lors de l'analyse précédente, que la méthode CART était meilleure. L'analyse les résidus (images ci-dessous), nous confirme le choix de la méthode CART.

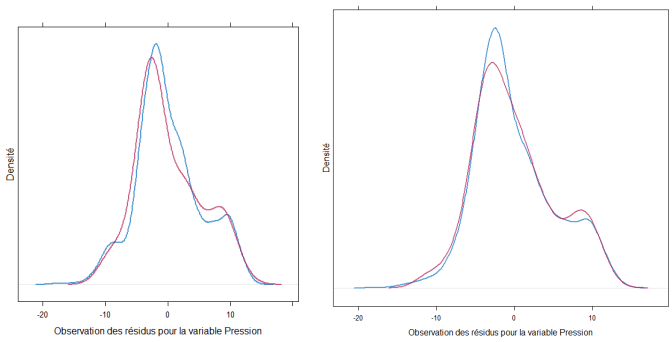


Figure 3.3.8: Visualisation des résidus

Au final, lorsqu'on applique à chaque variable, la méthode qui lui convient, on obtient les distributions suivantes :

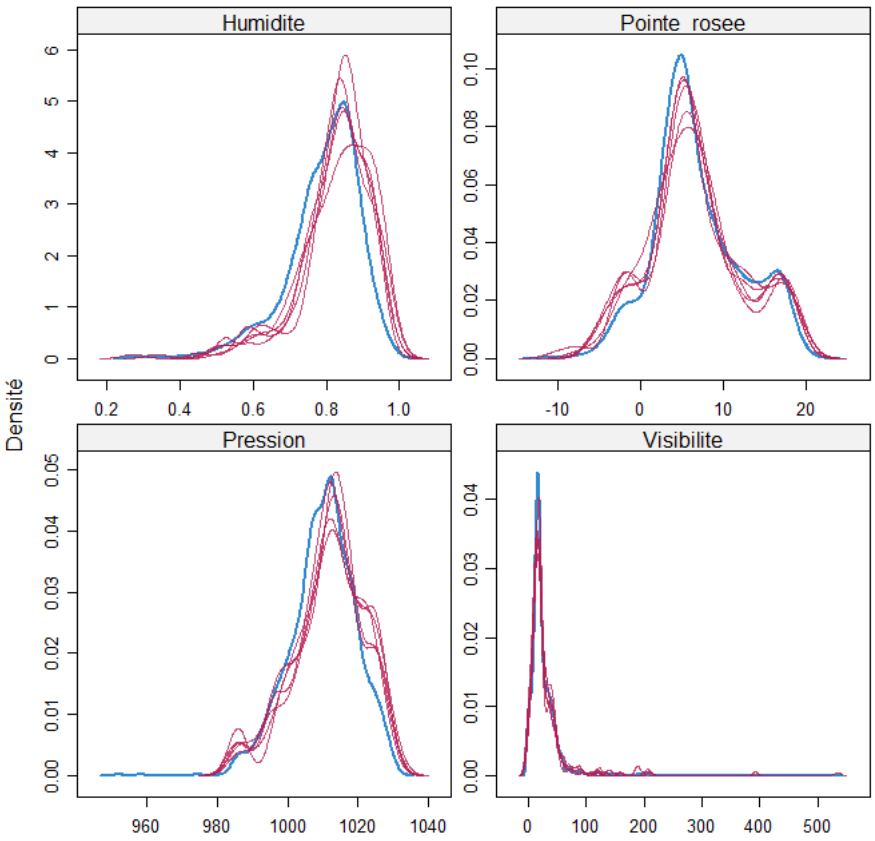


Figure 3.3.9: Distributions optimales - imputations des données

Vérification des résultats obtenus

Une étape importante consiste à évaluer si les imputations sont plausibles, c'est-à-dire que les valeurs imputées auraient pu être observées si elles n'avaient pas été manquantes. Si les valeurs sont clairement impossibles (vent négatif par exemple), elles doivent être supprimées. Les imputations doivent respecter les relations entre les variables et respecter le degré d'incertitude quant à leur vraie valeur.

Les vérifications de diagnostique sur les données imputées permettent de vérifier la vraisemblance des imputations.

L'imputation est correcte si les valeurs imputées (points rouges) suivent raisonnablement bien les valeurs observées (point bleus), y compris les trous de distribution.

Sélection des variables imputées

Maintenant que l'imputation est faite, l'étape suivante consiste à analyser les multiples données imputées par le modèle. Rubin a élaboré en 1987 un ensemble de règles pour

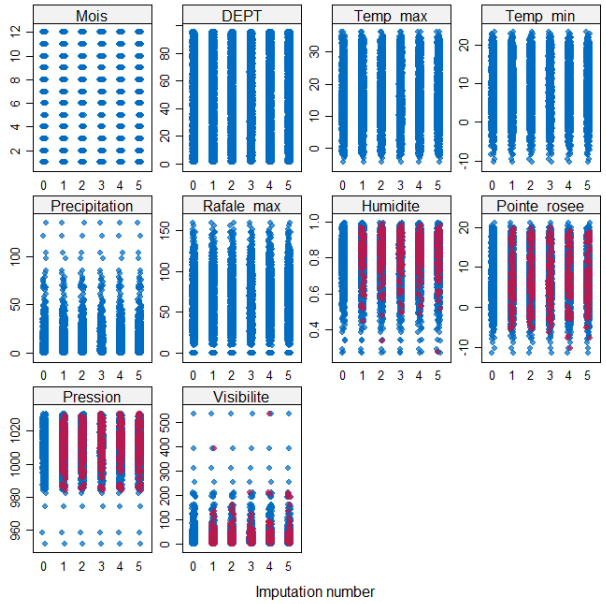


Figure 3.3.10: Vérification des imputations

combiner les estimations séparées et les erreurs-types de chaque m ensembles de données imputées en une estimation globale avec erreur type, intervalle de confiance et p_value. Ces règles sont basées sur la théorie asymptotique de la distribution normale.

Voici le résultat obtenu pour la variable Pression :

	est	se	t	df	Pr(>|t|)	lo 95	hi 95	nmis	fmi	lambda
(Intercept)	1026.93350750	3.519234465	291.8059361	94.06520	0.0000000000	1019.94604812	1.033921e+03	NA	0.21720828	0.20073976
DEPT	-0.01799999	0.006604484	-2.7254194	1988.82017	0.0064782658	-0.03095242	-5.047555e-03	0	0.01531527	0.01432554
Temp_max	0.14202211	0.081461231	1.7434319	156.54130	0.0832214946	-0.01888289	3.029271e-01	0	0.16375121	0.15313496
Temp_min	0.05461282	0.101024002	0.5405925	282.31399	0.5892147055	-0.14424307	2.534687e-01	0	0.11654417	0.11030758
Humidite	-11.92742918	3.818693011	-3.1234323	57.41902	0.0028001937	-19.57302084	-4.281838e+00	157	0.28386232	0.25934504
Rafale_max	-0.10390334	0.006428074	-16.1639939	120.11385	0.0000000000	-0.11663036	-9.117633e-02	0	0.18984411	0.17646569
Visibilite	0.02942987	0.007459240	3.9454249	70.86458	0.0001852917	0.01455608	4.430367e-02	183	0.25342679	0.23264957
Pointe_rosee	-0.45560397	0.160216674	-2.8436739	72.31233	0.0057945792	-0.77496651	-1.362414e-01	146	0.25066614	0.23022390

Figure 3.3.11: Vérification des imputations

La proportion de la variance totale attribuable aux données manquantes (lambda) vaut 23 % pour la variable pointe de rosée, m = 5 donc l'efficience relative de la variable est de 96%.

3.3.2 Normalisation des données quantitatives

Dans notre base de données, on remarque que les ordres de grandeur des différentes variables quantitatives ne sont pas les mêmes (vent en km/h, coût des sinistres en milliers ...). Or, des écarts importants peuvent rendre les algorithmes d'apprentissage inefficaces. Pour éviter cela, il est nécessaire d'effectuer un prétraitement sur les données en les normalisant c'est-à-dire en transformant les variables en variables centrées réduites.

Ainsi, chaque variable x suit une transformation de la forme :

$$x_i \to \frac{x_i - \overline{x}}{\sigma(x)}$$

avec :

- $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- $\sigma(x) = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2}$

3.3.3 Transformation des données en classe

En statistique, il est fréquent de regrouper les données en classe pour faciliter leur lecture. Après une analyse rapide, deux variables pouvaient être regroupées : les voitures et les habitations.

Concernant les contrats auto, nous avons décidé de les regrouper en fonction de la marque de la voiture. 5 classes (G0 à G4) ont ainsi été créées avec G0 les voitures bas de gamme et G4 les voitures très haut de gamme. Ce regroupement a été effectué sur la base du classement SRA. Les contrats habitations ont été regroupés en fonction du montant de la prime. Pour les habitations, 6 classes ont été trouvées de G1 à G6 avec G1 les primes faibles et G6 les primes les plus élevées.

3.3.4 Transformation des données catégorielles

Les variables catégorielles ne subissent aucune normalisation mais elles sont transformées en dummy variable.

En statistique et notamment en analyse de régression, une variable fictive (dummy variable en anglais ou variable de conception) prend la valeur 0 ou 1 pour indiquer l'absence ou la présence d'un effet catégorie susceptible de modifier le résultat. Ces variables sont très importantes puisqu'elles sont des substitutions numériques pour les variables qualitatives. Dans l'analyse de régression, les variables dépendantes peuvent être influencées non seulement par les variables quantitatives (coût d'un sinistre, force du vent...) mais également par des variables qualitatives (nature de l'événement, département...). Cette transformation permet donc de les convertir en variable quantitative et donc de pouvoir aisément les utiliser.

3.4 Actualisation : mise en "as if"

Lorsque les observations se rapportent à une longue période, il convient de corriger les montants de l'effet de l'inflation et de prendre en compte les modifications légales ou

réglementaires. C'est pour cette raison que nous avons actualisé notre historique.

3.4.1 Changements économiques : l'inflation

Définition 3.2. L'**inflation** est un accroissement excessif des instruments de paiement (billets de banque, capitaux) entraînant une hausse des prix et une dépréciation de la monnaie (s'oppose à déflation).

Cette augmentation des prix se traduit par une augmentation du coût des tempêtes. Il est donc nécessaire de la prendre en compte pour ne pas sous-estimer le coût.

Indice FFB

L'indice FFB (appelé auparavant indice FNB) représente le coût de la construction selon la Fédération Française du Bâtiment. En assurance, il sert de base pour la fixation des garanties dommages et des franchises pour les contrats d'assurance habitation, les multirisques commerces et bureaux, ainsi que l'assurance des immeubles en copropriété. Cet indice est publié trimestriellement par la Fédération Française du Bâtiment depuis le 1er janvier 1941.

Exemple de calcul: Pour un contrat d'assurance habitation dont la cotisation de 300 euros est indexée sur l'indice FFB avec pour base l'indice du 2ème trimestre 2013, le nouveau montant en 2014 sera de $\frac{300*925}{915,80} = 303€$.

Indice RI

"L'indice RI est l'indice sur lequel sont indexés tous les contrats d'assurance dommages des entreprises dont le contenu à assurer (matériel et/ou marchandises) a une valeur supérieure à 150 fois la valeur en euros de l'indice RI" - le Traité des Risques d'Entreprises- FFSA.

Il se calcule de la façon suivante :

$$I = 45 + 2,26A + 19,43B + 5,65C + 8,37D$$

avec

- A : Indice FFB du coût de la construction
- B : Indice mensuel du coût horaire du travail révisé
- C : Indice de prix de production de l'industrie française pour le marché français - Produits métallurgiques
- D : Indice de prix de production de l'industrie française pour le marché français - Biens intermédiaires

3.4.2 Changements politiques

Les initiatives politiques ont été nombreuses depuis 2003. Certaines ont eu des impacts plus ou moins notables. Nous en avons étudié 3 :

- **l'interdiction de construire dans certaines zones inondables**

Les zones inondables sont réparties en trois catégories : les zones rouges, zones où le risque est important (zone inconstructible), les zones bleues correspondent à un risque jugé moyen et les zones blanches sont jugées sans risques.

Depuis la tempête Xynthia, deux zones complémentaires se sont ajoutées. Les zones noires où il est interdit d'habiter et les zones jaunes où le risque peut-être limité grâce à des aménagements spécifiques.

De ce fait, il est impossible de souscrire une assurance habitation en zone noire et très difficile en zone rouge. Voici une carte des zones noires :

Figure 3.4.12: Carte des zones noires

Nous avons utilisé cette information pour supprimer certains contrats/sinistre de notre historique. Cela permet de réduire le coût de certains départements et donc de ne pas faire de sur-estimation.

- **Plan de prévention des risques naturels**

Afin de minimiser l'impact des évènements naturels catastrophiques, les communes qui le désirent peuvent mettre en place des PPRN (plan de prévention des risques naturels). Ces aménagements du territoires, ont connu un essor depuis 1995 comme le montre la courbe ci-dessous.

Ne sachant pas comment l'incorporer dans notre historique, nous avons décidé de simplement ajouter une variable qui donne pour chaque département touché par une tempête le nombre de PPRN en place.

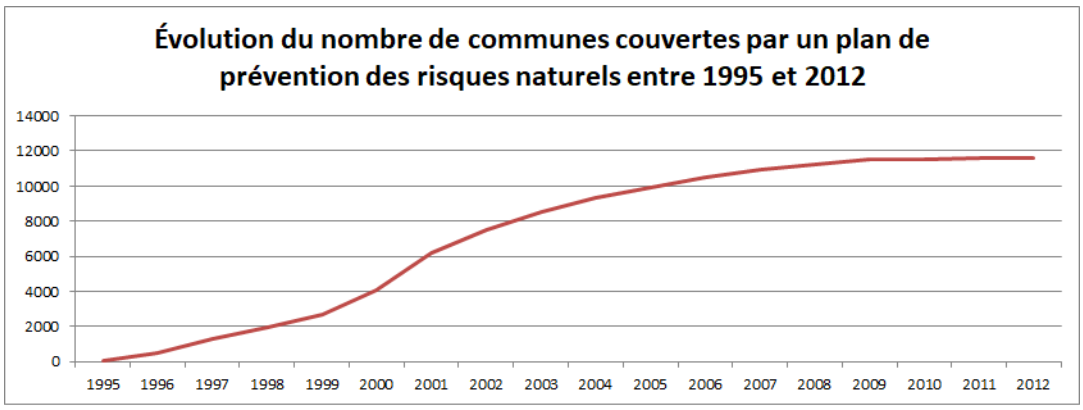


Figure 3.4.13: Évolution du nombre de communes couvertes par un PPRN entre 1995 et 2012

- **Mise en place de nouvelles normes**

De nombreuses normes ont été mises en place pour limiter l'impact des tempêtes : par exemple, la norme vent et neige ou encore la création de crochet anti-tempête pour attacher les tuiles.
Il est difficile de connaître l'ensemble des normes mise en place et encore plus de quantifier leur impact. Par conséquent, aucune actualisation n'a été faite par rapport à la mise en place des nouvelles normes.

3.4.3 Changements indemnisation

Depuis 2003, de nombreux changements ont eu lieu au sein de la direction indemnisation. L'un des changement qui a eu le plus d'impact est la montée des agents en plateforme mis en place à partir de 2011, ce qui a joué un rôle considérable dans les cadences de déclaration des sinistres :

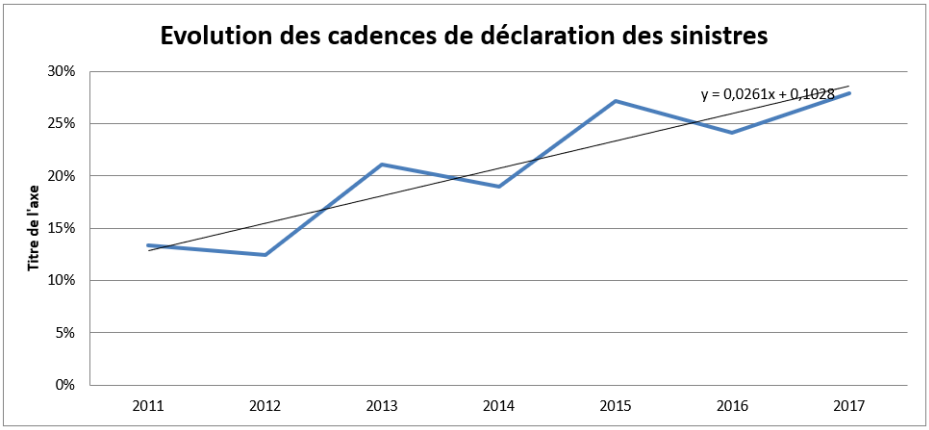


Figure 3.4.14: Evolution des cadences de déclaration entre 2011 et 2017

On remarque qu'entre 2011 et 2018, la déclaration des sinistres à J+3 a évolué de 15% soit environ 2.5% par an. Ainsi, pour ne pas fausser les prédictions, il est nécessaire d'actualiser l'historique. Si nous ne le faisons pas, l'estimation du nombre total de sinistres d'une tempête risquerait d'être sur-estimé. Pour que cela ne se produise pas, nous avons appliqué les coefficients suivants :

2011	2012	2013	2014	2015	2016	2017
115%	112.5%	110%	107.5 %	105%	102.5%	100%

Tableau 3.4.3: Coefficients de l'évolution des cadences de déclarations

Pour chaque tempête, on multiplie le nombre de sinistres par le coefficient associé. Par exemple, si une tempête survenue en 2012 a eu 100 sinistres déclarés à J+3, l'actualisation nous permet de dire que si cette même tempête avait eu lieu en 2017, 100*112.5% = 113 sinistres aurait été déclarés contre 100 en 2012.

Partie III

Le modèle

Chapitre 1

Les tempêtes de référence

Le phénomène des tempêtes est facilement compréhensible par l'homme. Cependant, ce dernier a souvent du mal à associer une tempête en cours à des tempêtes passées. C'est le phénomène d'oubli que nous avons abordé précédemment. Afin de mieux appréhender les résultats prédits par le modèle, il est important de faire référence à des tempêtes passées. Ainsi, on contourne quelque peu le phénomène de "boîte noire" des modèles auto-apprenants.

1.1 Méthode basée sur l'expérience

Cette méthode ne repose pas sur des hypothèses mathématiques strictes mais sur l'expérience. L'expert détermine les tempêtes de référence en fonction des remontées terrain, des dires d'experts et de son ressenti.

Les résultats de cette méthode sont difficiles à vérifier au moment de la survenance d'une tempête car ils ne reposent pas sur des tests statistiques permettant d'estimer leur robustesse. Cependant, l'expérience nous a appris que les experts étaient très doués pour identifier les moyennes et grosses tempêtes. Les tempêtes de petites envergures sont plus difficiles à jauger car le coût est susceptible de passer du simple au double (de 3 à 6 millions d'euros) en raison d'une mauvaise expertise de certaines zones (en raison d'un manque de temps pour se rendre sur place par exemple) ou d'une mauvaise estimation du montant des graves.

Le but des méthodes qui vont suivre est de trouver des tempêtes de référence à partir des informations en notre possession à J+3 sans avoir recours aux experts. Nous comparons ensuite les tempêtes trouvées avec celle proposées par les experts et nous analyserons les similitudes et les différences pour essayer d'en extraire des règles ou une méthode permettant d'optimiser le choix de nos tempêtes de références.

1.2 Méthode du clustering

Le but des méthodes de clustering (classification) est de diviser un ensemble d'individus en groupes, appelés clusters, homogènes partageant des caractéristiques communes.

Il existe deux grands types de classification : supervisée et non supervisée. En classification non supervisée, l'appartenance des observations à l'une de K classes (groupes / populations) n'est pas connue. C'est justement cette appartenance qu'il s'agit de retrouver à partir des k descripteurs disponibles. En classification supervisée au contraire, l'appartenance des n observations aux différentes classes est connue et l'objectif est de construire une règle de classement pour prédire la population d'appartenance des nouvelles observations.

Que la méthode choisie soit supervisée ou non, le but est toujours de trouver un bon partitionnement qui permet de garantir :

- une grande similarité entre les groupes les plus homogènes possibles grâce à la minimisation de l'inertie intra-classe
- des classes bien différenciées entre elles grâce à la maximisation de l'inertie inter-classe.

Définition 1.1. Une dissimilarité est une fonction d qui a tout couple (x_1, x_2) associe une valeur dans $\mathbf{R}^+$ telle que :

- $d(i,j) \geq 0$
- $d(i,i) = 0$
- $d(i,j) = d(j,i)$
- $d(i,j) \leq d(i,k) + d(k,j)$

Cette définition s'interprète de la façon suivante : plus les observations x_1 et x_2 se ressemblent plus leur score est élevé.

Définition 1.2. Une similarité est une fonction s qui a tout couple (x_1, x_2) associe une valeur dans $\mathbf{R}^+$ telle que :

- $s(x_1, x_2) = s(x_2, x_1)$
- $s(x_1, x_1) \geq s(x_1, x_2)$

Le but de cette partie est d'essayer de trouver des tempêtes de référence par une méthode supervisée et une méthode non-supervisée et de voir laquelle des deux conviendrait le mieux ou de trouver un moyen de combiner les résultats des deux méthodes.

1.2.1 KNN : méthode par apprentissage supervisée

L'algorithme des k plus proche voisin est l'un des algorithmes les plus simples de classification supervisé. C'est un algorithme à moyennage locale. Son but consiste à trouver pour

un individu, les individus qui sont les plus proches de lui, pour déterminer à quel groupe il appartient. Prenons un exemple : l'individu x est une tempête dont on cherche à connaître son intensité (faible, moyenne ou exceptionnelle). Si ses deux plus proches voisins sont KLAUS et Xynthia alors l'algorithme conclura que la tempête en question est d'intensité exceptionnelle. D'un point de vue mathématique, cette idée se traduit comme suit :

Cadre : Soit d_n notre ensemble de données. L'ensemble des couples iid (X_i, Y_i) avec $X_i \in Y_i, Y_i \in R$ suivent une loi (X,Y). L'objectif est de trouver une fonction de régression dans R^d telle que :

$$r(x) = E[Y|X = x]$$

On suppose que $E[Y^2]$ est fini. On cherche à estimer M(x) par moyennisation :

Définition 1.3. Estimateur par moyennisation locale
Soit $W_{n,i}, 1 \leq i \leq n$ une famille de poids positifs tels que pour tout $n \geq 1, X, x_1, ..., x_n \in \mathcal{X}, \sum_{i=1}^{n} W_{n,i}(x, x_1, ..., x_n) = 1$.
Un algorithme de moyennage local est un algorithme défini , pour

$$d_1^n = (x_1, y_1), ..., (x_n, y_n)$$

par :

- $\eta_{d_1^n} : x \in \mathcal{X} \to \sum_{i=1}^{n} W_{n,i}(x, x_1, ..., x_n) y_i$ en régression réelle
- $\Phi_{\eta_{d_1^n}} : x \in \mathcal{X} \to signe(\sum_{i=1}^{n} W_{n,i}(x, x_1, ..., x_n) y_i)$ en discrimination binaire (règle de discrimination plug in)
- $\sum W_{n,i} = 1$

Grâce à ces poids, notre estimateur s'écrit désormais :

$$r_n(x) = \sum_{i=1}^{n} W_{n,i} * Y_i$$

L'estimateur pour un nouveau voisin, se trouve en appliquant la règle de discrimination (voir ci-dessous) :

Définition 1.4. On appelle règle de discrimination plug in toute fonction Φ_n de $\mathcal{F}$ telle que :

$$\forall x \in \mathcal{X}, \Phi_n(x) = signe(\eta(x))$$

avec η une règle de régression.

L'algorithme d'apprentissage est alors de la forme :

$$\hat{\eta} : x \leftarrow \sum_{i=1}^{n} W_i(x) Y_i$$

où $W_i(x)$ sont les poids réels des fonctions de x,n,$X_1 ... X_n$

Définition 1.5. On appelle algorithme des k plus proches voisins un algorithme par moyennage local dont les poids vérifient :

$$W_{n,i}(x, x_1, ..., x_n) = \begin{cases} \frac{1}{k} \text{ si } X_i \text{ fait partie des k-p.p.v. de x dans } X_1, ..., X_n \\ 0 \text{ sinon} \end{cases}$$

en cas d'égalité, on procède à un tirage aléatoire.

Théorème 1.6. ***Lemme de Stone - 1977*** *Si $\mathcal{X} = R^d, (k_n)_{n\geq 1}$ est une suite d'entiers tels que $k_n \to +\infty$ et $k_n/n \to 0$, alors l'algorithme des k plus proches voisins pour une distance associée à une norme quelconque de R^d est universellement consistant.*

Théorème 1.7. ***Cover et Hart 1967***
L'algorithme du plus proche voisin (k=1) n'est pas universellement consistant

1.2.2 Inconvénients des KNN

L'inconvénient majeur de la méthode des k plus proches voisins réside dans sa longueur d'exécution. En effet, la méthode effectue le calcul des distances entre x (l'élément que l'on étudie) et chaque élément de l'ensemble. Plus l'ensemble est grand, plus le temps d'exécution de la méthode est long. C'est pourquoi, de nombreuses méthodes d'optimisation ont été développées dont la méthode CAH que nous allons étudier dans la partie suivante.

1.2.3 CAH : méthode par apprentissage non-supervisée

L'objectif de la Classification Ascendante Hiérarchique est de classer les individus ayant des comportements similaires sur un ensemble de variable. Cette algorithme est itératif et ne fonctionne que sur les variables quantitatives.

Le principe est le suivant :

1. Chaque individu constitue une classe
2. On calcule les distances 2 à 2 entre individus et les deux individus les plus proches sont réunis en une classe
3. la distance entre la nouvelle classe et les n-2 individus restants est ensuite calculée et à nouveau les 2 éléments (individus ou classes) les plus proches sont réunis.

On réitère le processus jusqu'à ce qu'il ne reste qu'une unique classe constituée de tous les individus.

On a donc deux distances :

- La distance entre individus (vu précédemment)
- La distance entre classes, appelée critère d'agrégation

La distance de Ward

Il existe de nombreux critères d'agrégation mais la plus connue et la plus utilisée est le distance de Ward.

Définition 1.8. La distance de Ward entre deux classes (C_j, C_l) de barycentre respectif x_{c_j} et x_{c_l} est définie par :

$$D_w^2(C_j, C_l) = \frac{n_j * n_l}{n_j + n_l} ||x_{c_j} - x_{c_l}||^2$$

Lorsque l'on utilise la distance de Ward dans l'algorithme de classification ascendante hiérarchique, on réalise à chaque étape la fusion optimale au sens de la concentration de l'inertie intra-classe. Cette optimisation locale à chaque étape ne garantie pas l'optimum global c'est-à-dire l'identification de la partition optimale en k classes mais permet de donner une bonne partition.

Qualité de la partition

Définition 1.9. L'inertie totale d'une partition est la somme entre l'inertie intra-classe et inter-classe. Elle se calcule de la façon suivante :

$$\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk} - \overline{x_k})^2 = \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk} - \overline{x_{qk}})^2}_{I_w} + \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(\overline{x_{qk}} - \overline{x_k})^2}_{I_B}$$

avec :

- k = nombre de classes
- q = sous-classe de la classe k
- i = nombre d'individus par classe
- I_w = inertie intra-classe
- I_B = inertie inter-classe
- $\overline{x_k}$: moyenne des x_k
- $\overline{x_{qk}}$: moyenne de x_k de la classe q

Définition 1.10. La qualité d'une partition est mesurée par :

$$O \leq \frac{\text{Inertie Inter}}{\text{Inertie totale}} \leq 1$$

- Si $= 0, \forall k \forall q, \overline{x_{qk}} = \overline{x_k}$: les classes ont les mêmes moyennes. Dans ce cas de figure, il est impossible de classifier les variables.
- Si $= 1, \forall k \forall q \forall i, x_{iqk} = \overline{x_{qk}}$: les individus d'une même classe sont identiques. On a une classification optimale.

Définition 1.11. La partition optimale C_k^* des observations en k classes est définie par :

$$C_k^* = \underset{C \in C_k}{\arg\min} \sum_{K}^{k=1} \sum_{i \in C_k} d^2(x_i, x_{c_k})$$

avec C_k l'ensemble des partitions possibles de n observations en K classes

Trouver une partition optimale est presque impossible car cela reviendrait à tester toutes les partitions possibles.

Méthode de Ward

La méthode de Ward consiste à dire que l'inertie entre deux classes vaut :

$$Inertie(a) + Inertie(b) = Inertie(a \cup b) - \underbrace{\frac{n_a * n_b}{n_a + n_b} d^2(a, b)}_{\text{à minimiser}}$$

avec d la distance de Ward

Le dendogramme

Dans la classification ascendante hiérarchique, les partitions successivement obtenues sont hiérarchisées : elles sont emboîtées les unes dans les autres. De ce fait, il est possible de représenter l'historique des différentes étapes de l'algorithme à l'aide d'une arborescence appelée dendogramme.

En bas de l'arbre, on retrouve les individus jouant le rôle de feuille. La fusion de deux éléments est représenté par une branche reliant ces deux éléments entre eux dont la hauteur est proportionnelle à la distance entre les deux éléments fusionnés. Plus on remonte dans l'arbre, plus les classes sont hétérogènes et les branches s'allongent. Les hauteurs des branches sont proportionnelles à l'inertie perdue c'est-à-dire :

$$\frac{I_{intra}^{k+1} - I_{intra}^{k}}{I_{totale}}$$

Cette représentation donne une vision globale de la topologie des observations et permet d'identifier les classes.

CAH en grande dimension

Lorsqu'on est en grande dimension, il est possible de faire quelques ajustements de la méthode pour que cette dernière soit plus rapide :

- Lorsque nos données ont beaucoup de variables : faire une ACP et ne conserver que les premières dimensions sur lesquelles on applique un CAH
- Lorsque nos données ont beaucoup d'individus : faire une partition grâce à la méthode des k-means par exemple, et constuire la CAH à partir des ces classes.

1.3 Résultats et choix

Afin de comparer les résultats, nous avons choisi de regarder les tempêtes de références prédites pour les tempêtes survenues en 2017.

- **Méthode actuellement en place** : La méthode actuellement en place repose sur l'expertise humaine. Un expert propose une série de tempête de référence qui sont réutilisées ensuite pour l'ensemble des tempêtes de l'année. Pour l'année 2017, les tempêtes de référence étaient : Ulla, Dirk et Joachim.
- **La méthode des KNN** : Pour estimer les tempêtes de références, nous avons utilisé la méthode des k plus proches voisins en créant trois groupes : les petites, moyennes et grosses tempêtes. Notre but ici est non seulement de prédire le groupe auquel appartient notre nouvelle tempête mais également de savoir quelles tempêtes historiques sont les plus proches de la tempête qui vient de survenir.

  Nous commençons par créer une matrice de distance permettant pour chaque individu de connaître ses plus proches voisins :

	Calvann	Dagmar	Renate	Vera	Kyrill	Emma	Britta	Andrea	Qendresa	Henri	Petrine	Nannette	KLAUS	QUINTEN	XYNTHIA	JOACHIM	CHRISTIAN	DIRK	ULLA	Anton	EGON	KLM	ZEUS	ANA	Bruno	CARMEN
Dagmar	2460450																									
Renate	24035171	22012522																								
Vera	2284851	224660	22237173																							
Kyrill	1939861	2514934	24497208	2293279																						
Emma	7800063	5617498	16410502	5841601	8087372																					
Britta	4063550	1672443	20402257	1889875	4096227	3991771																				
Andrea	11194903	9079402	12942782	9303857	11555318	3467976	7459691																			
Qendresa	8383510	6210308	15819275	6434428	8679329	592864,9	4583373	2876494																		
Henri	2013455	2607826	24583390	2386735	99920,1	8173128	4181512	11641100	8764873																	
Petrine	3200301	754404	21323434	963917	3174320	4915073	933158	8382515	5507523	3261682																
Nannette	3969211	5492448	27492661	5268319	2995821	11083153	7092015	14551071	11675144	2911009	6169256															
KLAUS	6999955	5245692	17058386	5461612	7746391	1965225	3926128	4449263	2240229	7841989	4696888	10674184														
QUINTEN	2283207	4110406	26091208	3888173	1861272	9721870	5773164	13175657	10314202	1817532	4840492	1704189	9145180													
XYNTHIA	23042189	21068718	1238433	21293279	23565661	15490460	19477411	12031278	14902396	23653509	20391851	26558085	16049147	25127950												
JOACHIM	3232029	4950480	26960150	4726062	2536780	10567052	6591193	14029874	11159908	2467149	5659307	798084,6	10057088	948854,5	26007987											
CHRISTIAN	2724693	365692	21971542	450245	2525671	5561987	1572566	9029842	6154134	2612387	650103	5521257	5320757	4217849	21041440	5021066										
DIRK	5080940	2894278	19118341	3118935	5392798	2738522	1358461	6189049	3329450	5482153	2226701	8382443	2599961	6987322	18175913	7842893	2876112									
ULLA	3581947	5066816	27068104	4842673	2572005	10658973	6668180	14126822	11251065	2488050	5744677	425663,6	10250521	1350335	26132794	569790,4	5096990	7957052								
Anton	4077184	5578696	27574041	5354791	3076945	11163891	7172282	14631866	11755649	2990776	6251211	112416,2	10767394	1814987	26641760	910281,4	5602581	8466838	517266,6							
EGON	15316250	13260362	8752785	13484999	15744802	7659217	11650582	4193256	7068805	15831221	12570826	18740082	8391939	17344566	7841913	18209060	13219139	10366434	18315452	18821716						
KLM	6083095	3855593	18119439	4120079	6381072	1728207	2299356	5183836	2319496	6468751	3207094	9374896	1993391	7995428	17184762	8846045	3856786	1010368	8949937	9457657	9366698					
ZEUS	21141613	19138794	2900909	19363437	21630535	13548902	17538780	10085170	12959441	21717681	18456223	24624660	14160090	23208756	1963238	24082915	19105295	16244749	24199647	24707337	5892371	15249779				
ANA	2164184	814940	22764761	613618	1733658	6356104	2367760	9823730	6948426	1822176	1441333	4727958	6056475	3435703	21832007	4227217	795221	3660104	4303346	4810199	14012134	4647498	19897140			
Bruno	5243075	6867334	28867503	6643109	4370309	12457595	8466111	15925564	13049439	4284620	7544260	1375402	12038107	2961842	27933487	2014431	6895962	9757813	1800775	1294085	20115030	10750182	25999961	6103034		
CARMEN	7505541	5338447	16680768	5562822	7816645	300799,6	3723974	3741393	878873,5	7903235	4642668	10811914	1776742	9439067	15753596	10288880	5291005	2452442	10387337	10893579	7928183	1443729	13814977	6083993	12186848	
ELEANOR	3956815	1620753	20406706	1843524	4092996	4001808	223697	7467619	4594615	4180969	920175	7086786	3827068	5731146	19472676	6565928	1570273	1308278	6661894	7169483	11653930	2288179	17537938	2359365	8462040	3726527

Figure 1.3.1: Matrice des distances

Cette matrice nous permet de trouver les tempêtes les plus proches. L'algorithme donne automatiquement les 3 tempêtes de référence les plus proches. Par ailleurs, nous avons créé une représentation graphique pour visualiser le résultat de la matrice:

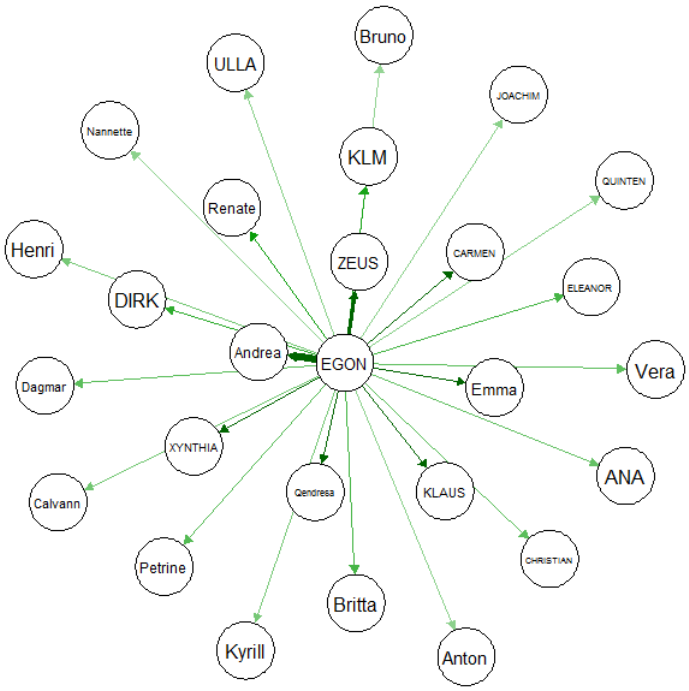


Figure 1.3.2: Visualisation des plus proches voisins d'une tempête

Ainsi, pour chaque tempête, on trouve le groupe auquel elle appartient ainsi que sa distance avec les autres tempêtes.

- **La méthode CAH** :
  L'objectif de la classification Ascendante hiérarchique est de classer les individus ayant des comportements similaires sur un ensemble de variable. Cet algorithme est itératif et ne fonctionne que sur les variables quantitatives.

  Rappelons le principe de la CAH :

  1. Chaque individu constitue une classe
  2. On calcule les distances 2 à 2 entre individus et les deux individus les plus proches sont réunis en une classe
  3. la distance entre la nouvelle classe et les n-2 individus restants est ensuite calculée et à nouveau les 2 éléments (individus ou classes) les plus proches sont réunis.

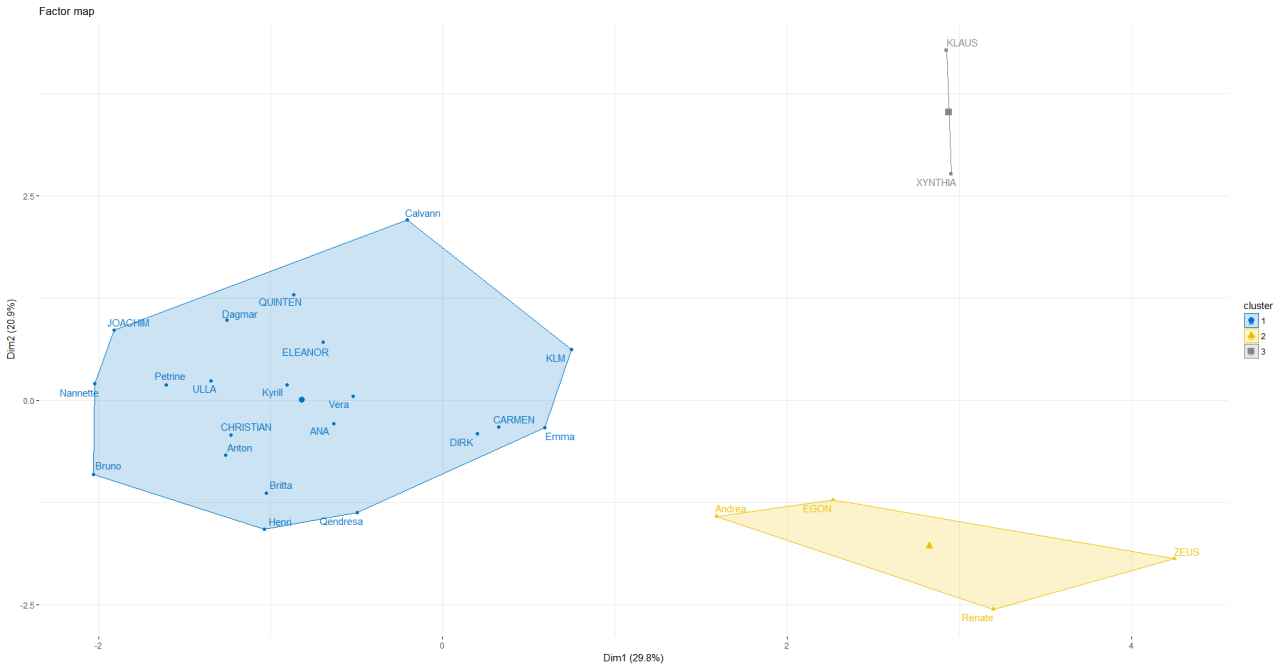


Figure 1.3.3: Regroupement des tempêtes en cluster

On réitère le processus jusqu'à ce qu'il ne reste qu'une unique classe constituée de tous les individus. Une fois, ce travail effectué, il est possible de visualiser le résultat sous forme de dendogramme :

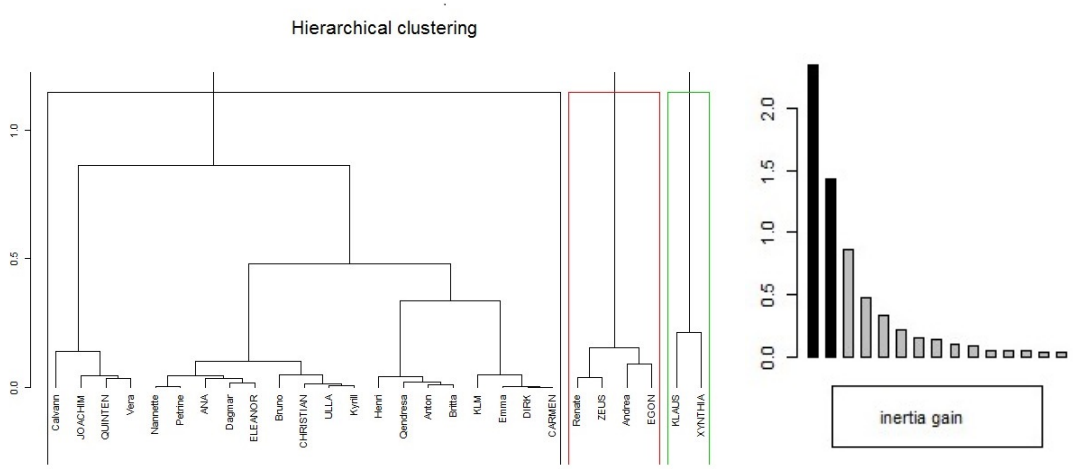


Figure 1.3.4: Dendogramme des individus obtenus par CAH

Grâce à ce dendogramme, les similitudes entre les tempêtes sont mises en avant permettant le choix de nos tempêtes de références.

De plus, pour chaque groupe il est possible de connaitre les tempêtes les plus représentatives :

```
Cluster: 1
     ULLA    Kyrill     Anton CHRISTIAN   ELEANOR
0.6715521 0.7023721 0.8976907 0.9383544 1.2301192
------------------------------------------------------------
Cluster: 2
   Renate      EGON      ZEUS    Andrea
0.9042144 1.2474114 1.5100468 1.7265074
------------------------------------------------------------
Cluster: 3
   KLAUS  XYNTHIA
1.695028 1.695028
```

Figure 1.3.5: Visualisation des individus représentatifs de chaque groupe

Ces tempêtes peuvent être utiles si on cherche à avoir des compléments d'informations sur la tempête en cours.

Une fois l'ensemble de ces simulations effectué, on peut regrouper les tempêtes de références dans un tableau pour mieux analyser les différentes méthodes :

	1	2	3	1	2	3	1	2	3
Tempête1	Ulla	Dirk	Joachim	Anton	Nannette	Ulla	Andrea	Renate	X
Tempête2	Ulla	Dirk	Joachim	Renate	Xynthia	Anton	Dirk	Carmen	X
Tempête3	Ulla	Dirk	Joachim	Anton	Nannette	Ulla	EGON	Andrea	Renate

Tableau 1.3.1: Tableau d'estimation des tempêtes de référence par différentes méthodes de classification

On remarque que les tempêtes de références suggérées sont très différentes suivant la méthode utilisée. Il est donc difficile de conclure sur la méthode la plus appropriée sur une simple suggestion. Afin de la déterminer, nous allons tester ces tempêtes de références en tant que prédicteurs du nombre de sinistres déclarés.

La seule analyse que nous pouvons apporter pour le moment concerne le temps d'exécution et la facilité à récupérer les résultats. La méthode par dire d'expert à l'avantage d'être rapide et s'effectue en début d'année donc le temps d'attente est minime. Cependant, elle est difficile à justifier car elle ne repose sur aucune méthode statistique. La méthode des K plus proches voisins est quant à elle plus longue, mais le résultat repose sur une théorie mathématique connue, dont on peut évaluer la fiabilité. De plus, cette méthode est très visuelle, ce qui facilite l'explication des résultats à autrui. Enfin, la méthode CAH a la particularité d'être très rapide à tourner, de sortir des résultats visuels intéressants. Elle pourrait sembler optimale, cependant, les résultats de ce modèle sont plus difficiles à appréhender.

Dans la partie suivante, nous tâcherons de déterminer quelle méthode convient le mieux à notre problématique en les testant dans le modèle prédictif.

Chapitre 2

Modélisation du nombre de sinistres

La modélisation du nombre de sinistres repose sur les tempêtes de références. Dans la partie qui suit, nous allons tester différentes méthodes de régression pour essayer de déterminer la façon optimale de prédire le nombre de sinistres d'une tempête.

2.1 Méthode des cadences moyennes

Déroulé de la méthode :

1. Pour chaque tempête de référence, on regarde le pourcentage de déclarations à J+3
2. On fait la moyenne des pourcentages de déclarations
3. Pour la tempête étudiée on calcule son nombre de sinistre total grâce à la formule :

$$\text{nombre de sinistres total} = \frac{\text{nombre de sinistres déclarés à J+3}}{\text{cadences moyennes des déclarations}}$$

Cette méthode est performante car elle est adaptative. Elle peut être modifiée pour prendre en compte les dires d'experts.

Cette méthode met en avant l'importance de la sélection des tempêtes de référence et de l'actualisation de son historique. La déclaration des sinistres se faisant de plus en plus rapidement chez Allianz[1], il est nécessaire d'actualiser son historique pour ne pas sous-évaluer le nombre de sinistres total.

2.2 Modèles de régression

2.2.1 Régression Poisson

La loi de Poisson a été présentée pour la première fois en 1838 par Simon Denis Poisson dans "Recherches sur la probabilité des jugements en matière criminelle et en matière civile". Elle est utilisée pour modéliser un nombre d'événements.

[1] voir partie 3.4.3 Changements Indemnisation

Définition 2.1. La Loi de Poisson
Soit λ un réel et Y une variable aléatoire réelle, $Y \sim P(\lambda)$ si et seulement si :

$$\forall k, P(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

En conséquence, on a : $E[Y] = V(Y) = \lambda$

Définition 2.2. Modèle de régression de Poisson
Le modèle de régression Poisson est de la forme :

$$ln(y) = \alpha + \beta_1 x_1 + ... + \beta_i x_i + ... + \beta_k x_k$$

où

- y réalisation de Y variable endogène suivant un loi de Poisson
- α l'ordonnée à l'origine
- β_i le coefficient associé à la ième variable explicative x_i

L'estimation du modèle s'effectue en cherchant les paramètres α et les coefficients β_i de la formule précédente.

Pour cela, on utilise la méthode du maximum de vraisemblance :

$$L = \prod_{i=1}^{n} P(Y_i = k_i) = \prod_{i=1}^{n} e^{-\lambda_i}\frac{\lambda_i^{k_i}}{k_i!}$$

où

- n est le nombre d'observations
- $\lambda_i = e^{\alpha+\beta x_i}$ avec $x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ ... \\ x_{ij} \end{pmatrix}$ et $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_j \end{pmatrix}$

Le logarithme de la vraisemblance est :

$$ln(L) = \sum_{i=1}^{n}[k_i ln(k_i) - \lambda_i]$$

On obtient la maximisation grâce à la dérivée :

$$s(\alpha, \beta) = \begin{pmatrix} \frac{\partial ln(L)}{\partial \alpha} \\ \frac{\partial ln(L)}{\partial \beta} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n}(y_i - \lambda_i) \\ \sum_{i=1}^{n} x_i(y_i - \lambda_i) \end{pmatrix} = \sum_{i=1}^{n}(y_i - \lambda_i)\begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

En utilisant l'algorithme de Newton-Raphson, on obtient :

$$\begin{pmatrix}\alpha_{k+1}\\ \beta_{k+1}\end{pmatrix} = \begin{pmatrix}\alpha_k\\ \beta_k\end{pmatrix} + I^{-1}(\alpha_k, \beta_k)s(\alpha_k, \beta_k)$$

où I^{-1} est la matrice de variance-covariance. On arrête l'algorithme lorsque :

$$\begin{pmatrix}\alpha_{k+1}\\ \beta_{k+1}\end{pmatrix} = \begin{pmatrix}\alpha_k\\ \beta_k\end{pmatrix}$$

Une fois les paramètres calculés, on mesure la qualité de l'ajustement du modèle grâce à la déviance. Le principe est de comparer la vraisemblance obtenue avec celle du modèle de référence grâce à la formule :

$$Déviance = 2[L_s - L_c] = 2\sum_i (y_i log(\frac{y_i}{\hat{\lambda_i}}) - (y_i - \hat{\lambda_i}))$$

avec :

- L_s est la valeur max de la log-vraisemblance modèle complet
- L_c est la valeur max de la log-vraisemblance modèle estimé

2.2.2 La sur-dispersion

Dans un modèle de Poisson, la sur-dispersion se produit lorsque la variance de l'échantillon est supérieure à la moyenne. C'est-à-dire lorsqu'il existe une corrélation positive entre les variables ou une variation excessive entre les variables. La sur-dispersion survient également lorsqu'il y a une violation sur les hypothèses de distribution des variables, par exemple lorsque les données sont regroupées et violent l'hypothèse de l'indépendance des observations.

La sur-dispersion est un problème car elle peut entraîner une sous-estimation des erreurs-types lors de l'estimation, c'est-à-dire qu'une variable peut apparaître comme étant un prédicteur significatif alors qu'en réalité il est insignifiant.

Il est possible de mesurer la dispersion grâce au quotient de dispersion $= \frac{\text{valeur de Pearson}}{\text{degré de liberté}}$. Si ce coefficient est supérieur à 1, il y a une sur-dispersion.

L'équidispersion est une propriété de la loi de Poisson : si $Y \sim P(\lambda), E[Y] = Var(Y) = \lambda$ Si on suppose que $Y|\Theta \sim P((\lambda\Theta)$ où Θ suit une loi gamma de paramètre α de sorte

que $E[\Theta] = 1$, on obtient la loi binomiale négative :

$$P(Y = k) = \frac{\Gamma(k + \alpha^{-1})}{\Gamma(k+1)\Gamma(\alpha^{-1})} \left(\frac{1}{1+\lambda/\alpha}\right)^{\alpha^{-1}} \left(1 - \frac{1}{1+\lambda/\alpha}\right)^{k}, \forall k \in N$$

On pose : $r = \alpha^{-1}$, $p = \frac{1}{1+\lambda/\alpha}$, $b(\theta) = -rlog(p)$ et $\alpha(\phi) = 1$
La moyenne est alors :

$$E(Y) = b'(\theta) = \frac{\partial b}{\partial p}\frac{\partial p}{\partial \theta} = \frac{r(1-p)}{p} = \lambda$$

et la variance vaut :

$$Var(Y) = b''(\theta) = \frac{\partial^2 b}{\partial p^2}\left(\frac{\partial p}{\partial \theta}\right)^2 + \frac{\partial b}{\partial p}\frac{\partial^2 p}{\partial \theta^2} = \lambda + \alpha\lambda^2$$

2.2.3 Régression Binomiale Négative

La régression binomiale négative est une généralisation de la régression de Poisson qui assouplit l'hypothèse restrictive selon laquelle la variance est égale à la moyenne faite par le modèle de Poisson. Le modèle de régression binomiale négative traditionnel, communément appelé NB2, est basé sur la distribution du mélange Poisson-gamma. Cette formulation est populaire car elle permet de modéliser l'hétérogénéité de Poisson en utilisant une distribution gamma, c'est-à-dire que la régression binomiale négative s'utilise lorsque les données de comptage sont très dispersées : lorsque la variance conditionnelle dépasse la moyenne conditionnelle.

2.2.4 La distribution binomiale négative

La distribution de Poisson peut être généralisée en incluant une variable de bruit gamma avec une moyenne de 1 et une paramètre d'échelle v. La distribution binomiale négative qui en résulte est :

$$Pr(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}$$

avec :

- $\mu_i = t_i\mu$
- $\alpha = 1/\nu$
- μ : taux d'incidence moyen de y par unité d'exposition (temps, distance, volume, taille de la population ...)

- t_i : l'exposition pour une observation particulière. Si aucune exposition : $t_i = 1$

Les résultats ci-dessus utilisent la relation suivante dérivée de la définition de la fonction gamma :

$$\ln\left(\frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})}\right) = \sum_{j=0}^{y_i-1} \ln(j + \alpha^{-1})$$

2.2.5 Le modèle de régression binomiale négative

Dans la régression binomiale négative, la moyenne de y est déterminée par le temps d'exposition t et un ensemble de k variables régresseurs (les x). L'expression reliant ces quantités est

$$\mu_i = \exp(\ln(t_i) + \beta_1 x_{1i} + ... + \beta_k x_{ki})$$

Souvent, $x_1 \equiv 1$, auquel cas β_1 est appelé *Intercept*. Les coefficients de régression $\beta_1, ..., \beta_k$ sont des paramètres inconnus qui sont estimés à partir d'un ensemble de données. Leurs estimations sont symbolisées par $\hat{\beta}_1, ..., \hat{\beta}_k$. En utilisant cette notation, le modèle de régression binomiale négative fondamentale pour une observation i s'écrit :

$$Pr(Y = y_i|_m u_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i}$$

Les coefficients de régression sont estimés en utilisant la méthode du maximum de vraisemblance. Cameron (2013, page 81) donne le logarithme de la fonction de vraisemblance :

$$\mathcal{L} = \sum_{i=1}^{n} (\ln[\Gamma(y_i + \alpha^{-1})] - \ln[\Gamma(\alpha^{-1})] - \alpha^{-1}\ln(1 + \alpha\mu_i) - y_i \ln(1 + \alpha\mu_i) + y_i \ln(\alpha) + y_i \ln(\mu_i))$$

$$= \sum_{i=1}^{n} \left\{ \left(\sum_{j=0}^{y_i-1} \ln(j + \alpha^{-1}) \right) - \ln(\Gamma(y_i + 1)) - (y_i + \alpha^{-1})\ln(1 + \alpha\mu_i) + y_i \ln(\mu_i) + y_i \ln(\alpha) \right\}$$

Les premiers dérivés de $\mathcal{L}$ ont été donnés par Cameron (2013) et Lawless (1987)

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1 + \alpha\mu_i}, j = 1, 2, ..., k$$

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_{i=1}^{n} \left\{ \alpha^{-2} \left(\ln(1 + \alpha\mu_i) - \sum_{j=0}^{y_i-1} \frac{1}{j + \alpha^{-1}} \right) + \frac{y_i - \mu_i}{\alpha(1 + \alpha\mu_i)} \right\}$$

$$\frac{-\partial^2 \mathcal{L}}{\partial \beta_r \partial \beta_s} = \sum_{i=1}^{n} \frac{\mu_i(1 + \alpha y_i)x_{ir}x_{is}}{(1 + \alpha\mu_i)^2}, r, s = 1, 2, ..., k$$

$$\frac{-\partial^2 \mathcal{L}}{\partial \beta_r \partial \alpha} = \sum_{i=1}^{n} \frac{\mu_i (y_i - \mu_i) x_{ir}}{(1+\alpha\mu_i)^2}, r = 1, 2, ..., k$$

$$\frac{-\partial^2 \mathcal{L}}{\partial \alpha^2} = \sum_{i=1}^{n} \left\{ \sum_{j=0}^{y_i - 1} \left(\frac{j}{1+\alpha j} \right)^2 + 2\alpha^{-3} \ln(1+\alpha\mu_i) - \frac{2\alpha^{-2}\mu_i}{1+\alpha\mu_i} - \frac{(y_i + \alpha^{-1})\mu_i^2}{(1+\alpha\mu_i)^2} \right\}$$

L'équation des gradients à zéro donne l'ensemble suivant d'équations de vraisemblance

$$\sum_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1+\alpha\mu_i} = 0, j = 1, 2, ..., k$$

$$\sum_{i=1}^{n} \left\{ \alpha^{-2} \left(\ln(1+\alpha\mu_i) - \sum_{j=0}^{y_i - 1} \frac{1}{j + \alpha^{-1}} \right) + \frac{y_i - \mu_i}{\alpha(1+\alpha\mu_i)} \right\} = 0$$

Cameron (2013) donne la distribution asymptotique des estimations du maximum de vraisemblance comme normales multivariées comme suit :

$$\begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} \sim N \left[\begin{array}{c|cc} \beta & V(\hat{\beta}) & Cov(\hat{\beta}, \hat{\alpha}) \\ \alpha & Cov(\hat{\beta}, \hat{\alpha}) & V(\hat{\alpha}) \end{array} \right]$$

où

$$V(\hat{\beta}) = \left[\sum_{i=1}^{n} \frac{\mu_i}{1+\alpha\mu_i} x_i x_i' \right]^{-1}$$

$$V(\hat{\alpha}) = \sum_{i=1}^{n} \left\{ \alpha^{-4} \left(\ln(1+\alpha\mu_i) - \sum_{j=0}^{y_i - 1} \frac{1}{j + \alpha^{-1}} \right)^2 + \frac{\mu_i}{\alpha^2(1+\alpha\mu_i)} \right\}^{-1}$$

$$Cov(\hat{\beta}, \hat{\alpha}) = 0$$

La déviance

La déviance est égale à deux fois la différence entre la log-vraisemblance maximum réalisable et la log-vraisemblance du modèle ajusté. Dans la régression multiple sous l'hypothèse de normalité, la déviance est la somme résiduelle des carrés. Dans le cas de la régression binomiale négative, la déviance est une généralisation de la somme des carrés. La vraisemblance logarithmique maximale possible est calculée en remplaçant μ_i par y_i dans la formule de vraisemblance. Ainsi, nous avons :

$$D = 2[\mathcal{L}(y_i) - \mathcal{L}(\mu_i)] = 2 \sum_{i=1} n \left\{ y_i \ln(\frac{y_i}{\mu_i}) - (y_i + \alpha^{-1}) \ln(\frac{1+\alpha y_i}{1+\alpha\mu_i}) \right\}$$

Critère AIC

L'ajustement d'un modèle peut également être calculé par le critère d'information Akaike (AIC).Il est de la forme :

$$AIC = -2[\mathcal{L} - k]$$

Analyse des résidus

Le résidu brut est la différence entre la réponse réelle et la valeur estimée par le modèle. La formule pour le résidu brut est :

$$r_i = y_i - \hat{\mu}_i$$

2.3 Applications des méthodes

2.3.1 Cadences Moyennes

En appliquant l'ensemble de ces méthodes sur nos données, on obtient les résultats suivants:

Nom de la tempête étudiée	**Nombre de sinistres observés**	**Nombre de sinistres déclarés à J+3**		**Nombre de sinistres prédits**	
		Sans actualisation	avec actualisation	Sans actualisation	Avec actualisation
Tempête1	5738	1225	1255	6000	6126
Tempête2	4095	970	994	7000	7089
Tempête3	8051	2269	2325	8500	8554

Tableau 2.3.1: Prédiction du nombre de sinistres par la méthode des Cadences moyennes et des dires d'experts

Nom de la tempête	**Nombre de sinistres observés**	**Nombre de sinistres déclarés à J+3**		**cadence**	**Nombre de sinistres prédits**	
		Sans actualisation	avec actualisation		Sans actualisation	Avec actualisation
Tempête1	5738	1225	1255	23%	5326	5460
Tempête2	4095	970	994	20%	4850	4971
Tempête3	8051	2269	2325	26%	8616	8831

Tableau 2.3.2: Prédiction du nombre de sinistres par la méthode des Cadences moyennes et des KNN

L'actualisation notre historique provoque une sur-estimation du nombre de sinistres. Il convient donc de se demander si celle-ci est un choix judicieux. En analysant les résultats, on remarque que l'actualisation est intéressante si elle est faite en amont c'est-à-dire au moment de la détermination des tempêtes de référence. L'estimation du nombre de sinistres étant acceptable sans l'actualisation, il a été décidé de ne pas la faire pour gagner en rapidité.

Il faut cependant remarquer que ce choix entrainera un biais à la longue puisque plus une tempête sera ancienne, moins sa probabilité d'apparition en temps que tempête de

référence sera grande. On se retrouve en présence d'un cas de "perte de mémoire". Pour éviter ce phénomène, le modèle possédera une option permettant d'obtenir la prédiction avec et sans actualisation.

Étudions maintenant le cas de la prédiction des sinistres à partir des tempêtes de référence trouvées grâce au CAH. Nous trouvons les résultats suivants :

Nom de la tempête	**Nombre de sinistres observés**	**Nombre de sinistres déclarés à J+3**		**Nombre de sinistres prédit**	
		Sans actualisation	avec actualisation	Sans actualisation	Avec actualisation
Tempête1	5738	1225	1255	5326	5456
Tempête2	4095	970	994	4409	4518
Tempête3	8051	2269	2325	10159	10410

Tableau 2.3.3: Prédiction du nombre de sinistres par la méthode des Cadences moyennes et CAH

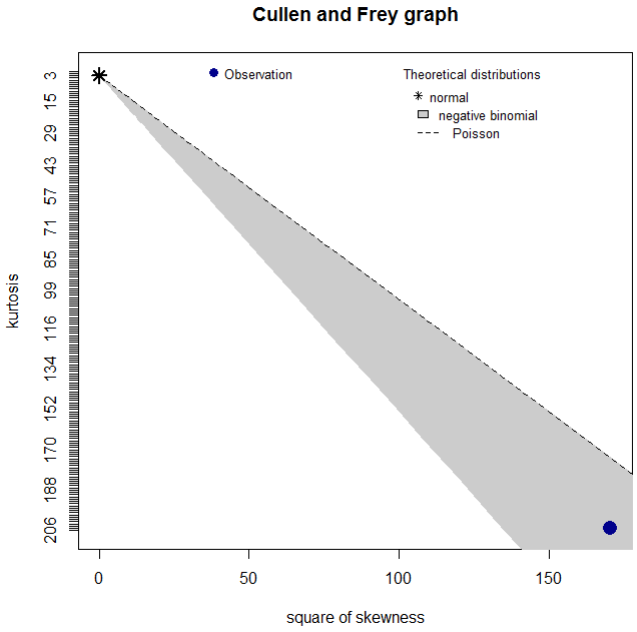
On peut remarquer que la prédiction du nombre de sinistres est très similaire entre l'utilisation des tempêtes de références prédites par les KNN et celles prédites par les CAH pour Tempête1 et Tempête2. On remarque également que quelque soit la méthode utilisée, le nombre de sinistres prédit est surestimé pour la tempête Tempête3. L'hypothèse émise pour expliquer cette surestimation est la suivante : Tempête3 étant la 3e tempête en moins de trois semaine, les assurés connaissaient la marche à suivre pour la déclaration de leur sinistre et ont donc déclaré plus rapidement. Une variable indiquant le nombre de tempête survenu le mois précédent va donc être ajouté pour éviter ce type d'erreur.

2.3.2 Régression Binomiale Négative

Les résultats proposés précédemment reposent sur une sélection de tempête et non sur l'ensemble des tempêtes. Il y a donc une perte d'information importante. On pourrait appliquer un pourcentage de cadences moyennes pour chaque tempête mais on perdrait alors les caractéristiques spécifiques d'une tempête (information fournie par la sélection de tempête de référence).

Pour palier à ces problèmes, nous avons utilisé une méthode de régression pour compter le nombre de sinistre. Après une rapide analyse de notre jeu de données, on trouve une moyenne de 3110 sinistres, et une variance de 5 585 132 ainsi qu'un quotient de dispersion était de 2.5, excluant ainsi automatiquement la loi Poisson. Nous avons donc choisi la loi Binomiale négative pour gérer ce phénomène de sur-dispersion.

Afin de vérifier notre hypothèse, nous avons décidé de faire un Cullen and Frey Graph. Ce graphique permet d'estimer la loi la plus appropriée en fonction de la valeur du skweness et du kurtosis. Sur ce dernier, les valeurs des coefficients des données d'observations sont symbolisées par un point bleu. Son emplacement permet de déterminer la loi la plus proche. Sur le graphique suivant, nous pouvons remarquer que la loi la plus appropriée est la binomiale négative, ce qui confirme notre intuition précédente.



Une fois la loi trouvée, il s'agit de déterminer le GML le plus approprié. Pour cela, des tests d'adéquations ont été réalisés. Le fonctionnement est le suivant :

1. On crée le glm le plus complet (avec l'ensemble des variables)
2. On étudie sa déviance. La déviance résiduelle doit être inférieure au nombre de degrés de liberté résiduels pour que le GLM soit cohérent
3. On calcule l'AIC
4. On teste la qualité de l'ajustement des résidus de Pearson et leur homoscédaticité. Si ces deux hypothèses sont validées, on supprime la variable avec le plus petit coefficient de régression et on recommence à l'étape 2.

Le modèle choisi est celui dont le coefficient AIC est le plus petit mais qui valide l'ensemble des hypothèses. Cette méthode nous a ainsi permis de supprimer 5 variables qui alourdissaient notre modèle.

Nom de la tempête	**Nombre de sinistres observés**	**Nombre de sinistres déclarés à J+3**		**cadence**	**Nombre de sinistres prédits**	
		Sans actualisation	avec actualisation		Sans actualisation	Avec actualisation
Tempête1	5738	1225	1255	23%	5717	5178
Tempête2	4095	970	994	22%	3866	3151
Tempête3	8051	2269	2325	22%	8621	10438

Tableau 2.3.4: Prédiction du nombre de sinistres par régression Binomiale Négative

On remarque qu'une fois de plus l'actualisation fausse le résultat final. En effet, l'actualisation doit se faire en amont de l'application des modèles et non en aval.

2.4 Analyses et choix de la méthode

Récapitulons les résultats obtenus :

	Cadence Moyenne KNN	Cadence Moyenne CAH	BN - KNN	BN - CAH
Tempête1	5%	-7%	-7%	-0.4%
Tempête2	73%	18%	8%	-5.6%
Tempête3	6%	7%	26%	7%

Tableau 2.4.5: Récapitulatif de l'erreur générée par les différentes méthodes de prédictions du nombre de sinistres

A première vue, le pourcentage d'erreur des sinistres prédits peut sembler important mais en analysant de plus près les résultats, on se rend compte qu'une erreur de 5% équivaut à une différence de moins de 200 sinistres entre les résultats observés et ceux prédits. Différence largement acceptable dans le cas présent puisqu'on s'intéresse une estimation arrondie du nombre de sinistres.

Après réflexion, il a été décidé de garder la méthode classique, méthode avec l'estimation du nombre des sinistres par l'utilisation de la cadence moyenne. Celle-ci donne des résultats intéressants et permet une communication plus aisée. De plus, ce chiffre est affiné par l'expert ce qui donne de meilleur résultat.

Néanmoins en cas de besoin, l'expert pourra consulter le nombre de sinistres prédit grâce à la méthode binomiale négative. Cette estimation sera une option facultative du modèle, que l'expert pourra utiliser au besoin.

En ce qui concerne la méthode de détermination des tempêtes de référence, la méthode des k plus proches voisins a été retenue en raison de sa facilité d'interprétation et de sa

lisibilité. Par ailleurs, il sera possible dans l'outil final d'entrer manuellement les tempêtes de références proposées par l'expert afin de combiner les résultats.

Chapitre 3

Détermination du seuil des graves : Théorie des valeurs extrêmes

---

En assurance, il existe différents types de sinistres, les sinistres « normaux » ou « attritionnels » et les sinistres graves ou extrêmes. Lorsque l'on modélise ces deux types de sinistres ensembles, on fausse la modélisation car ces derniers n'ont pas les mêmes propriétés, caractéristiques. Afin d'améliorer la prédiction, il convient de séparer les sinistres attritionnels des graves. Un sinistre grave est un sinistre avec une occurrence faible et un coût élevé. Ces sinistres sont peu nombreux et ont une réelle influence sur le coût total d'une tempête. Par exemple, les sinistres graves représentent 12.21% du coût total des sinistres professionnels mais uniquement 0.34% des sinistres déclarés. En moyenne, les graves représentent 13.12% du coût global des tempêtes pour moins de 0.11% des sinistres déclarés mais peuvent représenter jusqu'à 80,7% du coût lors des tornades. Pour ne pas fausser les analyses, il est nécessaire de les traiter séparément.

Il existe au sein d'Allianz, une règle stipulant qu'un sinistre grave est un sinistre dont le montant des règlements est supérieur à 150 milles euros. Cependant, lorsque l'on fait une analyse de ce seuil, on remarque qu'il n'est pas adapté aux tempêtes. Dans cette partie, nous allons donc revoir la définition des sinistres graves afin de les étudier de manières séparées.

Le montant seuil se détermine en étudiant les queues de distribution et repose sur un compromis biais-variance.

Une fois ce seuil déterminé, deux modèles distincts seront créés : l'un pour les graves, l'autre pour sinistres « normaux ».

La théorie des valeurs extrêmes repose sur deux grands théorèmes débouchant sur deux types d'approches distinctes :

- la méthode des maxima
- la méthode des seuils

La différence entre ces deux théorèmes repose sur les données fournies en entrée. Le premier est basé sur l'ensemble des simulations tandis que le deuxième est établi seulement pour

les simulations dépassant un certain seuil.

3.1 Estimation du paramètre de queue

La théorie des valeurs extrêmes apparait en 1928, avec l'énonciation du théorème de Fisher et Tippett. Elle est basée sur la loi du maximum :

Définition 3.1. Soit $M_n = max(X_1, ..., X_n)$ alors

$$P(M_n \leq x) = P = (X_1 \leq x, ..., X_n \leq x) = P(X_1 \leq x)...P(X_n \leq x) = [F(x)]^n$$

Notons x^F le point extrémal de F noté : $x^F = sup(x|F(x) < 1)$. Le support de F peut-être borné ($x^F < \infty$) ou infini. Par conséquent, la distribution asymptotique de M_n est dégénérée : $M_n \rightarrow^P_{n \leftarrow \infty} x^F$

Le théorème de Fisher-Tippett précise les lois asymptotiques que peut suivre le maximum normalisé d'une suite de variable IID.

3.1.1 Théorème de Fisher - Tippett

Théorème 3.2. *S'il existe des suites de réels a_n et b_n telles que quand $n \rightarrow \infty$*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = [F(x * a_n + b_n)]^n \rightarrow G(x)$$

pour une distribution non dégénérée G, alors G peut-être l'une des ces trois distributions:

$$Fréchet(\alpha > 0) : \phi_\alpha(x) = \begin{cases} 0 & si\ x \leq 0, \\ exp(-x^{-\alpha}) & si\ x > 0 \end{cases}$$

$$Weibull(\alpha > 0) : \Psi_\alpha(x) = \begin{cases} 0 & si\ x \leq 0, \\ exp(-(-x^{-\alpha})) & si\ x > 0 \end{cases}$$

$$Gumbel : \Lambda_\alpha(x) = exp(-e^{-x})\ x \in R$$

Afin de déterminer G, Von Mises(1954) et Jenkins(1955) ont introduit la loi d'extremum généralisée (notée GEV pour Generalized Extreme Value).

3.1.2 Loi Généralisée et domaine d'attraction

Définition 3.3. Loi d'extremum généralisée
La loi d'extremum généralisée $GEV(\mu, \sigma, \xi)$ est définie par la fonction de répartition :

$$G(x) = \begin{cases} exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right) & \text{si } \xi \neq 0, \\ exp\left(-exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right) & \text{si } \xi = 0 \end{cases}$$

avec ξ le paramètre de forme (paramètre de queue), μ le paramètre de position et σ le paramètre d'échelle.

Définition 3.4. Domaine d'attraction
F appartient au domaine d'attraction de G ($F \in D(G)$) s'il existe deux suites (a_n) et (b_n) telles que la convergence ait lieu.

La correspondance entre la GEV et les lois limites est déterminée par le paramètre de queue :

- Si $\xi = 0$, la distribution possède une queue fine : distribution de Gumbel
- Si $\xi > 0$, la distribution possède une queue lourde : distribution de Fréchet
- Si $\xi < 0$, la distribution est à support borné : distribution de Weibull

La GEV permet donc de faire l'approximation suivante :

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \approx GEV(\mu, \sigma, \xi)$$

3.2 Application

Avant d'effectuer une modélisation GEV pour estimer le montant des sinistres graves, il est nécessaire d'estimer son paramètre de forme afin d'appliquer les méthodes de détermination de seuil.

Pour commencer, il est nécessaire de s'interroger sur l'indépendance de nos données. Si un immeuble est touché par une tempête, quelle est la probabilité que les immeubles voisins soit impactés eux aussi ? Si cette probabilité est élevée, alors on doit rejeter l'hypothèse iid des variables et la théorie des valeurs extrêmes ne peut s'appliquer. Dans le cas des tempêtes, les sinistres sont principalement dus aux dégâts provoqués par la chute d'arbres ou d'objets. Or, il est rare qu'un arbre tombe sur deux maisons côte à côte, on peut donc supposer que les sinistres sont indépendants.

L'estimation des paramètres de la $GEV(\mu, \alpha, \xi)$ s'effectue grâce à la méthode des « Maxima par bloc ». Pour cela, nous subdiviserons l'échantillon de n observations en K sous-ensemble de taille n/K, nous permettant ainsi de disposer d'un échantillon de K maxima.

La GEV étant une loi limite pour le maximum, il est important de sélectionner correctement le nombre de blocs. Ceux-ci ne doivent pas être trop grand pour ne pas perdre trop d'information : phénomène de superposition de maxmima mais suffisamment grand pour que les maxima soient élevés.

3.2.1 Estimation du nombre de bloc

La méthode du Bloc Maxima consiste à regrouper les données en bloc de même taille et à déterminer le maximum de chaque bloc. Il n'existe pas de méthodes statistiques permettant de déterminer le nombre optimal de bloc et leur taille. Il faut donc faire attention à ce que la taille des blocs soit assez grande pour appliquer le théorème de Fisher-Tippett mais assez petite pour avoir un bon nombre de maxima.

Nous avons en notre possession 181 894 sinistres pour 56 tempêtes. Ce qui représente en moyenne 3 249 sinistres/tempête. Créons 3 blocs par tempête soit = 56*3 = 168 blocs et appliquons la méthode du block Maxima pour trouver la valeur maximal de chaque bloc.

3.2.2 Generalized Extreme Value

Une fois nos maximums trouvés, nous regardons si nos données fitte correctement avec une GEV. Pour cela, nous utilisons la méthode des moments.

Nous obtenons les informations suivantes :

On obtient un paramètre de queue de $0.69 > 0$ donc on est dans le domaine de Fréchet de paramètre $\alpha = 1/\xi = 1/0.69 = 1.45$

3.2.3 Estimation du seuil par l'estimateur de Hill

L'estimateur de Hill ne se fait que lorsque l'on est dans le domaine de Fréchet, ce qui est le cas ici. Cette méthode, entièrement graphique, est la méthode la plus fréquemment utilisée en théorie des valeurs extrêmes lorsque $\xi > 0$ car il assure un bon compromis

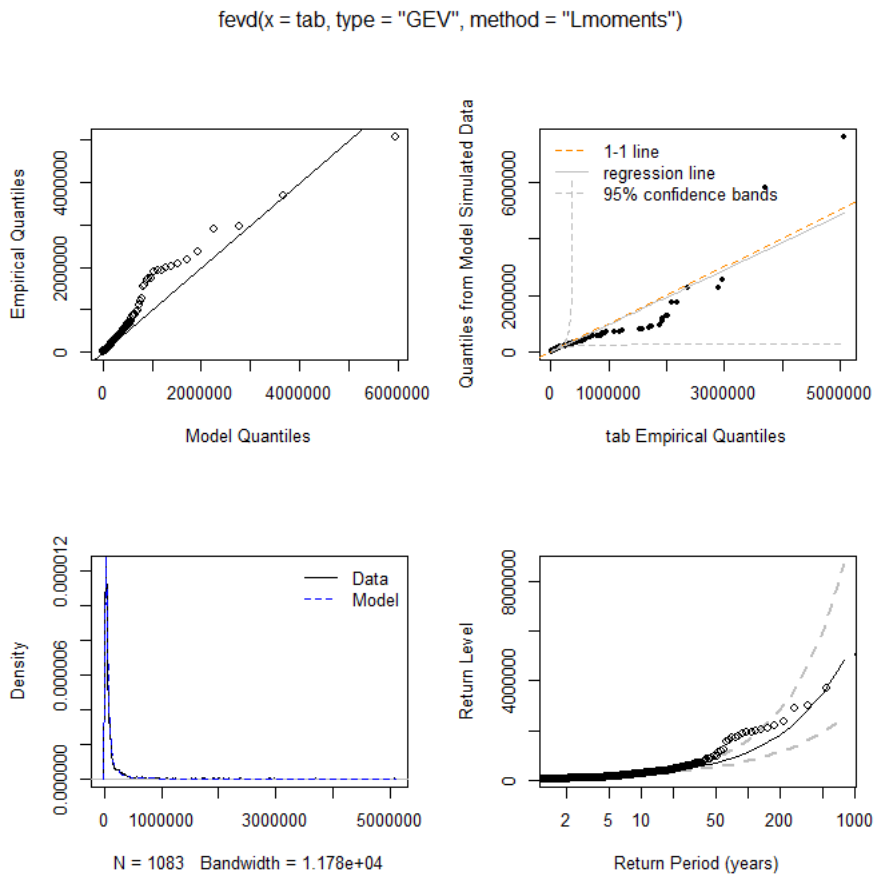


Figure 3.2.1: Fit GEV

biais-variance [Dress1998]. L'estimateur de Hill s'écrit :

$$\xi_{k,n}^{Hill} = \frac{1}{n}\sum_{i=1}^{k} ln(X_i) - ln(X_k)$$

avec :

- $X_1 > ... > X_n$ la statistique d'ordre
- n le nombre d'observation
- k un entier inférieur ou égale à n

Le graphique de l'estimateur de Hill consiste à représenter la valeur de l'estimateur en fonction de l'indice k de la statistique d'ordre, c'est-à-dire l'estimateur construit sur les observations supérieures ou égales au seuil X_k. Les lignes rouges représentent l'intervalle de confiance construit de la façon suivante:

$$IC_{95\%}(\xi) = \left[\hat{\xi}^{Hill} - 1.96 * \frac{\hat{\xi}^{Hill}}{\sqrt{k}}; \hat{\xi}^{Hill} + 1.96 * \frac{\hat{\xi}^{Hill}}{\sqrt{k}}\right]$$

En haut du graphique se trouvent les seuils de X_k. L'estimateur de Hill est volatile

lorsque k est faible puis se stabilise. Le seuil des graves est déterminé à partir du moment où il y a stabilisation.

Sur le graphique ci-dessous, on remarque que la courbe se stabilise autour de 90 000 euros.

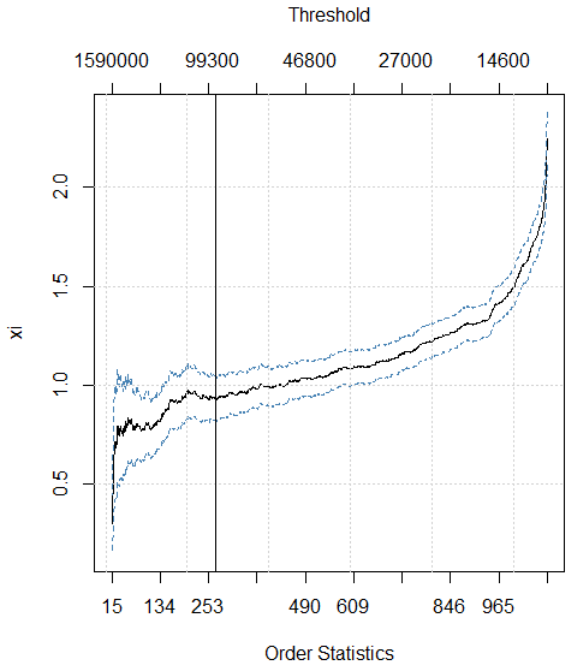


Figure 3.2.2: Hill Plot

Par conséquent, nous créons deux bases : une base comportant pour chaque tempête les sinistres de moins de 90 000 et une autre les sinistres de plus de 90 000.

3.3 Conclusion

La détermination du seuil à 90 000€ nous permet de créer deux modèles distincts : un modèle pour les petits sinistres et un autre pour les graves. Cette subdivision est nécessaire comme nous pouvons le voir sur l'image ci-dessous.

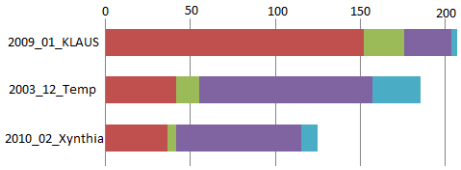


Figure 3.3.3: Répartition du coût des sinistres pour une tempête

On remarque que les graves ont plus ou moins d'importance dans le rôle des tempêtes. Or comme leur nombre est très faible, si on ne les divise pas, ces sinistres sont très mal estimés, ce qui fausse le résultat prédit. Nous verrons dans la partie suivante que leur estimation ainsi que l'actualisation du coût des sinistres permet d'estimer au mieux la sinistralité de nos tempêtes.

Chapitre 4

Modélisation GLM / GAMLSS

4.1 GLM

4.1.1 Définitions et Propriétés

Définition 4.1. Le modèle linéaire généralisé est un modèle statistique de la forme :

$$E[Y] = \mu = g^{-1}(X * \beta)$$

où

- Y est la matrice des variables de réponses
- X est la matrice des variables explicatives
- β est la matrice des coefficients à estimer

Les modèles linéaires généralisés sont caractérisés par trois éléments :

1. Le prédicteur $\eta = X * \beta$
2. L'appartenance de la loi Y a une famille exponentielle
3. La fonction de lien g

Définition 4.2. Famille exponentielle
La famille exponentielle est l'ensemble des lois à deux paramètres θ et ϕ dont la densité est de la forme :

$$f(y|\theta,\phi) = exp(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)) \text{ avec } y \in R \text{ ou } y \in N$$

avec :

- θ le paramètre naturel
- ϕ le paramètre de dispersion

D'après la formule précédente, on peut écrire :

$$E[Y] = \mu = b'(\theta) \text{ et } var(Y) = \phi\beta''(\theta)$$

avec $\beta''(\theta)$ la fonction de variance.

Théorème 4.3. *Pour les fonctions de liens canoniques, les lois appartenant à la famille exponentielle sont convexes dans les coefficients à estimer du modèle (Semenovich, 2013).*

Cette propriété permet de prouver l'existence et l'unicité de l'estimateur de maximum de vraisemblance pour les coefficients Beta.

La loi Y est choisie en fonction des connaissances a priori sur Y. Par exemple : son type (binaire, continue, de comptage...), sa dispersion ...

Définition 4.4. Fonction de lien
Une fonction de lien est une fonction inversible, dérivable et définie sur l'univers de Y. elle lie la moyenne μ au prédicteur η

Le choix de la fonction de lien est très important puisqu'il a un impact direct sur la bonne prédiction du modèle.

4.1.2 Estimation des paramètres

L'estimation des paramètres inconnus (β) peut être faire par l'estimation du maximum de vraisemblance. La vraisemblance du modèle s'écrit :

$$L(y_1, ..., y_n; \beta) = \prod_{i=1}^{n} exp(\frac{y_i\theta_i - b(\theta_i)}{\phi_i/w_i} + c(y_i, \phi_i))$$

L'estimateur β^* du maximum de vraisemblance se définit comme suit :

$$\beta^\star = argmax_\beta(L(y_1, ..., y_n; \beta))$$

La solution de cette équation se trouve en remplaçant la vraisemble du modèle par la log vraisemblance

$$l(y_1, ..., y_n; \beta) = log(L(y_1, ..., y_n; \beta)) = \sum_{i=1}^{n} exp(\frac{y_i\theta_i - b(\theta_i)}{\phi_i/w_i} + c(y_i, \phi_i))$$

On cherche alors β^* tel que :

$$\forall j \frac{\partial l(y_1, ..., y_n; \beta)}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = 0$$

avec :

- $\mu_i = E[Y|X = x_i]$
- $\eta_i = \sum_j x_{ij} * \beta_j$

La concavité de la fonction log vraisemblance au voisinage des $\hat{\beta}_j$ permet de s'assurer que l'on est en présence d'un maximum. Cette dernière est convexe si sa dérivée seconde est négative aux points critiques $\hat{\beta}_j$.

Cette équation se résout généralement de façon numérique car l'obtention d'une formule explicite est souvent difficile à trouver.

4.1.3 Sélection du modèle et vérification des hypothèses

Une fois l'estimation des paramètres effectuée, il est nécessaire de vérifier que le modèle construit est performant. Les vérifications à faire sont les suivantes :

- Évaluer la qualité de l'ajustement
- Vérifier les hypothèses de départ
- Évaluer la complexité du modèle : la suppression ou l'ajout d'une variable est-elle nécessaire?

Ces vérifications se font en calculant la déviance :

$$D(y, \mu) = \phi D^*(y, \mu)$$

ou la déviance standardisée :

$$D^*(y, \mu) = 2 * (logL(y, \beta) - logL(\mu, \beta))$$

La déviance standardisée suit approximativement une loi du χ^2 à $(n - p - 1)$ degrés de liberté. L'objectif est de minimiser D lors de l'ajustement d'un GLM.

4.1.4 Prédiction

Après avoir estimé les paramètres du modèles (les β), il est possible de prédire la valeur de $\hat{Y}_i$ au nouveau point x_i grâce à la formule :

$$\hat{Y}_i = \hat{\mu}_i = g^{-1}(x_i * \hat{\beta})$$

4.1.5 Avantages et inconvénients des GLM

- **Avantages du GLM**

1. Modèle très rapide à faire tourner
2. Les résultats sont facilement exploitables
3. Les variables explicatives du modèle GLM peuvent être quantitatives ou qualitatives

- **Inconvénients du GLM**

1. Les prédictions du modèle dépendent fortement des hypothèses prises lors de sa construction (la loi Y et la fonction de lien)
2. Le GLM ne fournit pas de coefficient permettant de modéliser l'interaction entre les variables.

4.2 GAMLSS

GAMLSS pour Generalized Additive Models for Location, Scale and Shape (modèles additifs généralisés pour la localisation, l'échelle et la forme) est un cadre général pour ajuster des modèles de régression semi-paramétrique où la distribution de la variable de réponse n'appartient pas à la famille exponentielle. Il permet de modéliser tous les paramètres de la distribution de la variable réponse sous la forme de fonctions linéaires ou non linéaires des variables explicatives.

GAMLSS est donc particulièrement adapté à la modélisation d'une variable de réponse qui ne suit pas une distribution exponentielle de la famille (par exemple, leptokurtique ou platykurtique et / ou des données de réponse asymétrique positive ou négative, ou des comptages surdispersés) ou qui présentent une hétérogénéité (par ex. ou forme de la distribution de la variable de réponse change avec des variables explicatives).

Un modèle GAMLSS suppose l'indépendance des variables y_i pour tout i=1,...,n avec une fonction de probabilité égale à $f(y_i|\theta^i)$ conditionnellement à $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$. θ^i représente les paramètres de distribution.

Posons les notations suivantes :

- n : le nombre d'observations indépendantes
- $Y = (Y_i)_{i=1,..,n}$ la variable à expliquer
- $\theta^i = (\mu_i, \sigma_i, \nu_i, \tau_i)$ le vecteur des paramètres de la distribution de Y tel que : $Y_i|\theta^i \sim D(\theta^i)$ avec :
  - μ_i : le paramètre de localisation de moment d'ordre 1
  - σ_i : le paramètre d'échelle de moment d'ordre 2
  - (ν_i, τ_i) : les paramètres de forme de moment d'ordre 3 et 4

Rigby et Stasinopoulos (2005) définissent la formulation d'un modèle GAMLSS comme suit :

Définition 4.5. Soit $Y^T = (y_1, ..., y_n)$ le vecteur de longueur n de la variable de réponse. Pour k = 1,2,3,4, $g_k(.)$ sont des fonctions de liaisons monotones telles que :

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} Z_{jk}\gamma_{jk}$$

où :

- μ, σ, ν, τ sont des vecteurs de taille n
- β_k^T correspond au vecteur de taille J_k des coefficients à estimer
- X_k correspond à la matrice des variables explicatives du modèle
- $\sum_{j=1}^{J_k} Z_{jk}\gamma_{jk}$ une partie aléatoire permettant d'introduire du bruit avec les variables explicatives.

Si l'on omet la partie aléatoire, alors le modèle GAMLSS paramètrique devient : $g_1(\mu) = X_1\beta_1, g_2(\sigma) = X_2\beta_2, g_3(\nu) = X_3\beta_3, g_4(\tau) = X_4\beta_4$. Ce modèle permet à l'utilisateur de modéliser chaque paramètre de distribution comme une fonction linéaire des variables explicatives.

L'analogie avec les GLM est forte. Les paramètres sont également estimés grâce au maximum de vraisemblance. De plus, tout comme les GLM, on calcule l'AIC et le coefficient de déviance afin de sélectionner le meilleur modèle.

4.3 Prédictions du coût attritionnels des tempêtes

4.3.1 Choix de la loi

En assurance, il est fréquent d'appliquer un GLM gamma pour modéliser le coût des sinistres. C'est donc cette loi que nous allons utiliser pour le GLM. Concernant le GAMLSS, le choix de la loi est plus compliqué. En effet, comme l'appartenance à une famille exponentielle n'est plus obligatoire, bon nombre de lois sont à notre disposition pour modéliser le coût global des tempêtes. Le choix de la loi étant déterminant pour faire une bonne estimation, il est nécessaire de la choisir avec soin. Pour cela, six lois ont été étudiées : la loi de Weibull, la loi Gamma, la loi Log Normale, la loi Normale, la loi de Cauchy, la loi de Gumbel.

Afin de déterminer la loi la plus adaptée, il est possible de comparer les distributions théoriques avec la distribution des données observées. On obtient les résultats suivants :

Le graphique précédent nous permet de déterminer qui s'ajuste le mieux à notre distribution. Intéressons-nous plus particulièrement au QQ-Plot et PP-Plot. Le QQ-Plot pour

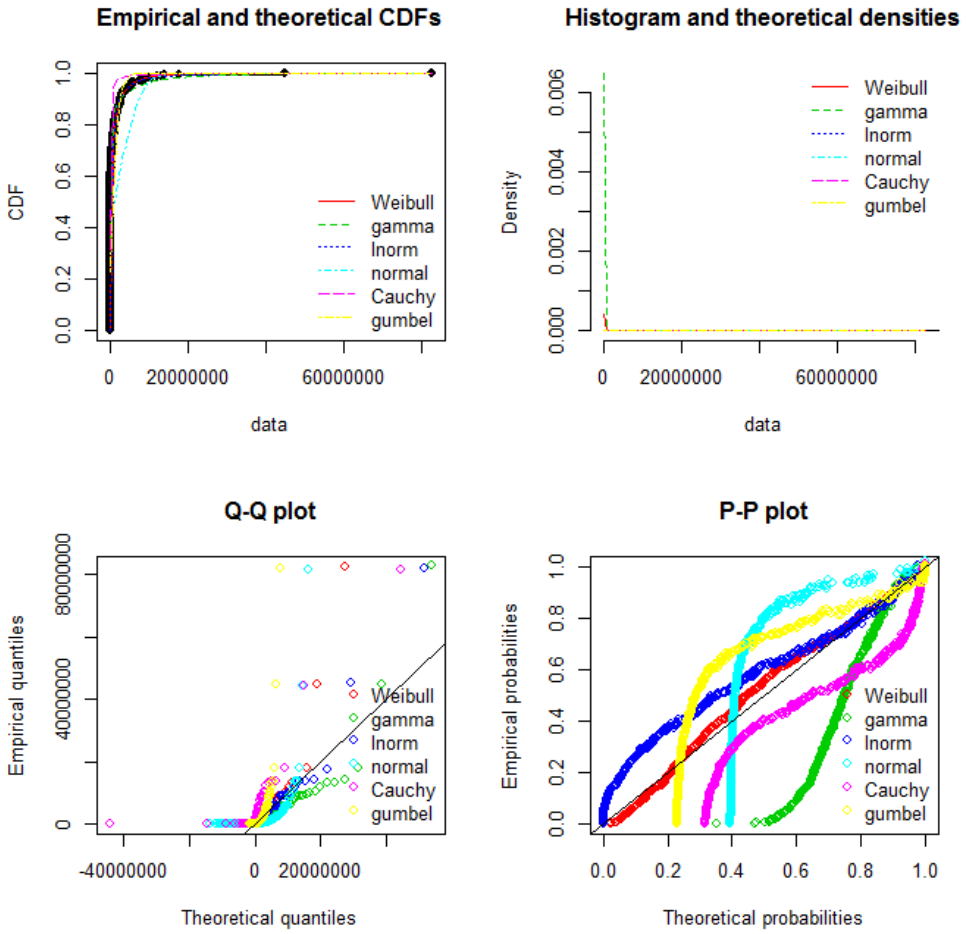


Figure 4.3.1: Visualisation des différentes lois possibles pour la modélisation des sinistres attritionnels

dirgramme quantile-quantile, permet d'évaluer la pertinence de l'ajustement d'une distribution donnée par rapport à une distribution théorique. Lorsque l'alignement avec la bissectrice est parfait, cela indique la présence d'une identité de loi. Dans notre cas, la lecture du graphique est complexe et les points semblent dispersés à partir de quantile 200000000.

Afin de faciliter la lecture, il est possible d'utiliser le PP-Plot (probabilité – Probabilité Graph). Ce dernier est utilisé pour évaluer l'asymétrie d'une distribution. L'avantage du PP-Plot par rapport au QQ-Plot est qu'il est discriminant dans les régions à forte densité de probabilité puisque dans ses régions, les distributions cumulatives empiriques et théoriques changent plus rapidement que dans les régions à faible densité de probabilité.

Par conséquent, l'analyse de ces deux graphiques nous permet de dire que la loi de Weibull semble s'ajuster le mieux.

Afin de valider les résultats énoncés précédemment, une série de test d'ajustement ont été mis en place : le test de Kolmogorov Smirnov et celui d'Anderson Darling (tests décrits en annexe).

```
Goodness-of-fit statistics
                                 Weibull          gamma          lnorm         normal         Cauchy         gumbel
Kolmogorov-Smirnov statistic 0.06424715608   0.5190262471  0.1737604332  0.3953978671  0.3159138362  0.2745425822
Cramer-von Mises statistic   0.40356325176  43.5609647612  5.1233239484 21.3315734351 13.3336482125 11.7465791206
Anderson-Darling statistic   2.45236605232 195.4700434294 43.4616581155           Inf 95.7851845952           Inf

Goodness-of-fit criteria
                                 Weibull       gamma       lnorm      normal      Cauchy      gumbel
Akaike's Information Criterion 13070.91455 13761.65483 13251.17746 15366.87833 13714.36503 14347.67038
Bayesian Information Criterion 13079.16391 13769.90420 13259.42682 15375.12769 13722.61440 14355.91975
```

Figure 4.3.2: Résultats statistiques des visualisations précédentes

Dans notre cas, on choisit un alpha = 0.05. Notre loi théorique est en adéquation avec notre loi empirique si la statistique de test est inférieure à 0.752 pour le test d'Anderson Darling et 0.07 pour le test de Kolmogorov Smirnov.

D'après le tableau ci-dessus, la seule loi dont on ne peut pas rejette l'hypothèse d'adéquation est la loi de Weibull, ce qui confirme l'analyse du QQ-Plot.

La loi la plus adaptée semble être la loi de Weibull. Il est nécessaire de tester notre hypothèse.

4.3.2 Prédiction du coût des tempêtes

Testons l'ensemble de ces lois sur nos données et comparons les résultats fournis : La

Nom de la tempête	Coût Observé	GAMLSS Weibull	GML Gamma	GAMLSS Gumbel
Tempête1	15.8	17.2	17.6	19.4
Tempête2	9.7	4.9	5	9.1
Tempête3	20.1	14.1	18	20.7
Tempête4	5.6	3.8	4	5.7

Tableau 4.3.1: Prédiction du coût des tempêtes par GLM et GAMLSS

loi de Gumbel semble la loi la plus appropriée pour modéliser nos sinistres. Ce résultat est surprenant car il diffère de la théorie énoncée précédemment. Analysons les résultats obtenus pour comprendre d'où peut provenir cette différence.

4.3.3 Différence entre la théorie et la pratique

On remarque que la théorie et la pratique se contredisent dans notre simulation précédente. En effet, théoriquement, la loi de Weibull est meilleure que la loi de Gumbel mais en pratique, le contraire s'applique. Il convient donc de se demander d'où peut provenir cette erreur.

L'erreur semble provenir du fait que la distribution de nos échantillons de test est différente de la distribution de notre échantillon d'apprentissage. En effet, habituellement des sinistres sont tirés de façon aléatoire pour constituer l'échantillon test. Dans notre cas, la sélection aléatoire n'était pas pertinente, car on souhaite vérifier que notre modèle modélise correctement les événements en cours et non les tempêtes passées.

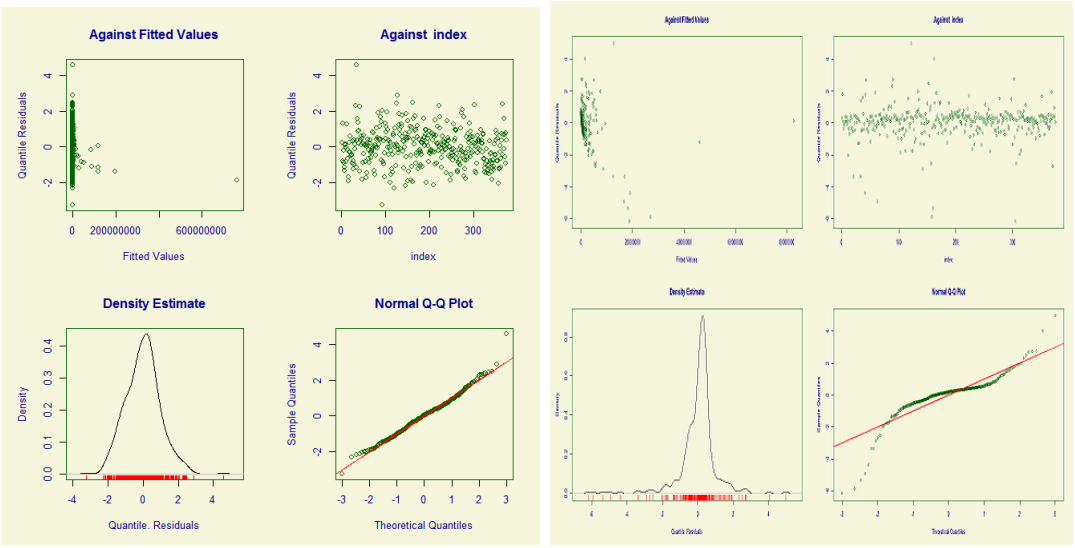


Figure 4.3.3: Résultat de l'estimation Gumbel et Gamma

Lorsque l'on compare les résultats d'un point de vue des distributions, on remarque que la GLM Gamma est plus appropriée que la Gumbel. En effet, le QQplot est meilleur et les résidus sont moins dispersés. Cette analyse montre que dans notre cas, plusieurs paramètres sont à prendre en compte et pas uniquement l'AIC ou la déviance.

4.3.4 Conclusion

La méthode par GAMLSS Gumbel est la méthode qui prédit le mieux. C'est cette méthode qui sera comparée à la prédiction par réseau de neurones étudiée dans la partie suivante.

4.4 Prédiction du coût grave des tempêtes

4.4.1 Choix de la loi

Comme précédemment, nous cherchons à déterminer la loi la plus adéquate à utiliser dans notre modèle GAMLSS. Pour cela, nous faisons appel au graphique de Cullen et Frey qui permet d'estimer quelle loi conviendrait le mieux.

Le graphique semble suggérer que l'utilisation de la loi bêta conviendrait parfaitement pour notre modélisation. Cependant, en effectuant les tests présentés précédemment, la loi qui semble convenir le mieux est la loi log normale.

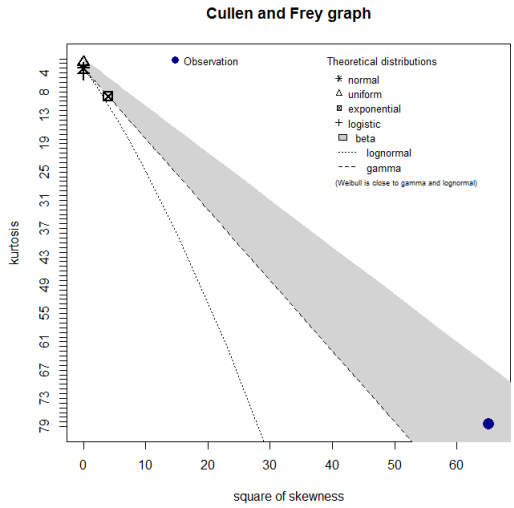


Figure 4.4.4: Graphique de Cullen et Frey

```
Goodness-of-fit statistics
                           Weibull      gamma     lnorm    normal     Cauchy    gumbel
Kolmogorov-Smirnov statistic 0.1824562  0.6611295 0.1218542 0.3934976  0.2816315 0.2837742
Cramer-von Mises statistic   0.8795435 17.3766602 0.3320156 7.7535758  2.7492182 4.2799963
Anderson-Darling statistic   5.8658839 77.4533920 2.5335104       Inf 17.7991998       Inf

Goodness-of-fit criteria
                              Weibull    gamma    lnorm   normal   Cauchy   gumbel
Akaike's Information Criterion 4565.478 4845.426 4513.594 5188.186 4627.330 4861.071
Bayesian Information Criterion 4571.500 4851.447 4519.615 5194.207 4633.351 4867.092
```

Figure 4.4.5: Choix de la loi pour modéliser les graves

4.4.2 Prédiction du coût grave des tempêtes

Appliquons la loi log Normale sur l'ensemble de nos tempêtes afin de prédire le coût grave de ces dernières. Nous obtenons les résultats suivants :

Nom de la tempête	Coût Observé	Coût Prédit
Tempête1	3.1	3.9
Tempête2	0.8	0.8
Tempête3	3.8	3.5
Tempête4	0.7	0.6

Tableau 4.4.2: Prédiction du coût grave des tempête par GAMLSS Log normale

On remarque que les résultats prédits sont très bons. En effet, l'erreur de prédiction est très faible. L'utilisation de la loi log Normale est validée et celle-ci sera utilisée à l'avenir pour modéliser le coût grave de chaque tempête.

Nom de la tempête	normaux	grave	Total	normaux	grave	Total
Tempête1	15.8	3.1	18.9	19.4	3.9	22.8
Tempête2	9.7	0.8	10.5	9.2	0.8	10
Tempête3	21	3.1	24.1	20.7	3.5	24.2
Tempête4	5.7	0.6	6.3	5.6	0.7	6.3

Tableau 4.5.3: Tableau récapitulatif des estimations

4.5 Conclusions de la partie

4.5.1 Récapitulatif des estimations

Les résultats obtenus répondent aux attentes de prédictions. Nous étions donc tentés de valider le modèle créé, c'est à dire utilisé un GAMLSS Gumbel pour les sinistres normaux et une logNormale pour les sinistres graves. Cependant, au moment de rédiger nos conclusions, une grêle et des orages ont frappé la France. Nous avons donc voulu vérifier les estimations de notre modèle dans un cas de figure pour ce type de phénomènes très liés au tempête.

4.5.2 Cas des Grêles et Orages

Nous avons gardé la même base de données tempêtes, en ajoutant simplement les tempêtes de grêle et les orages avec vents violents survenus entre 2003 et 2018. Bien que ceux-ci soient peu nombreux, nous avons avons relancer la prédiction avec les mêmes modèles. Les résultats obtenus sont les suivants :

Nom de la tempête	normaux	grave	Total	normaux	grave	Total
Grêle	1.5	0.1	1.6	4.1	0.1	4.3
Orage 1	5.3	1.4	6.7	8.3	0.9	9.2
Orage 2	6.4	2.1	8.5	7.5	2.3	9.3
Orage 3	3.9	0.7	4.6	6.4	0.4	6.8

Tableau 4.5.4: Estimation du coûts des graves par la méthodes GAMLSS Gumbel

On se rend compte que le modèle des graves restent très performants, ce qui valide une nouvelle fois son utilisation. Cependant, l'utilisation du GAMLSS Gumbel surestime fortement le coût des sinistres attritionnels. Cette différence de prédiction peut s'expliquer de la façon suivante : les grêles et les orages sont peu représentés dans la base de données, leurs spécificités sont donc noyées dans la masse des tempêtes. C'est à ce moment là que les réseaux de neurones peuvent être intéressants. Avec leur particularité d'estimateur universel, et l'ajustement des poids, les tempêtes particulières peuvent être prise en compte. Nous testerons cette hypothèse dans la partie suivante.

Chapitre 5

Modélisation du coût des sinistres "normaux" grâce aux PMC

5.1 Présentation succincte d'un réseau de neurone

5.1.1 Définitions et caractéristiques d'un neurone

Définition 5.1. Un réseau de neurones est un graphe orienté pondéré où chaque noeud est appelé neurone formel.

C'est un ensemble de processeurs élémentaires, les neurones, connectés les uns aux autres et qui sont capables d'échanger des informations au moyen des connexions qui les relient. Les connexions sont directionnelles et à chacune d'elle est associé un réel appelé poids de la connexion. L'information est ainsi transmise de manière unidirectionnelle du neurone j vers le neurone i, affectée du coefficient pondérateur (un poids) $w_{i,j}$. Un neurone calcule son état à partir d'informations venues de l'extérieur, ou bien il détermine son entrée à partir des neurones auxquels il est connecté et calcule son état comme une transformation souvent non linéaire de son entrée. Il transmet à son tour son état vers d'autres neurones ou vers l'environnement extérieur.

Un neurone est donc défini par trois caractéristiques:

- son état
- ses connexions avec les autres neurones
- sa fonction de transfert

Un réseau de neurones typique comprend :

- Une couche d'entrée: couche qui prend des entrées en fonction des données existantes
- des couches masquées: couches qui utilisent la rétropropagation pour optimiser les poids des variables d'entrées afin d'améliorer la puissance prédictive du modèle
- Une ou plusieurs couches de sortie: Sortie de prédictions basées sur les données des couches d'entrée et cachées

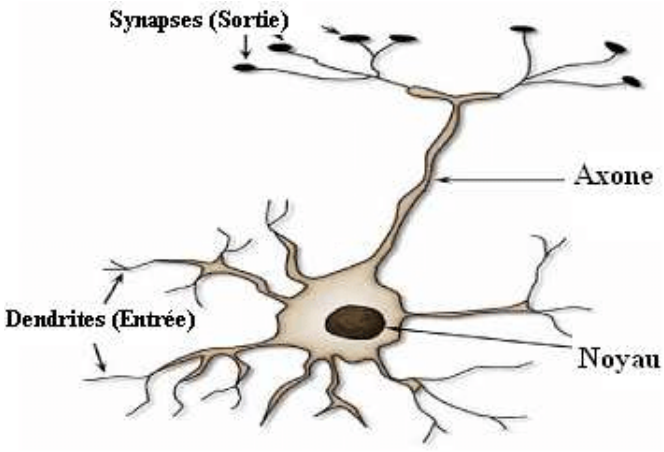


Figure 5.1.1: Exemple d'un réseau de neurones biologique

5.1.2 Fonction d'entrée et d'activation

Les fonctions d'entrée h et d'activation f sont généralement identiques pour les neurones d'une même couche. **Fonction d'entrée**
La fonction d'entrée permet d'obtenir l'activation pondérée $e_i = h(i, a, W)$. Cette fonction est souvent une fonction linéaire et de la forme :

$$h(i) = h(i, a, W) = \sum_{j=1}^{N} (w_{i,j} a_j)$$

Fonction d'activation
La fonction d'activation f s'applique alors à l'activation pondérée. L'activation à l'instant t est alors de la forme :

$$a_i(t) = f(e_i(t-1), \theta_i)$$

avec θ_i est un seuil.

Il existe différentes fonctions d'activation usuelles : linéaire, sigmoïde exponentielle, sigmoïde tangentielle,... avant de choisir la fonction d'activation, il est important de regarder son intervalle de définition.

5.1.3 Apprentissage supervisé d'un neurone

L'apprentissage des réseaux de neurones peut être supervisé ou non supervisé. Cependant dans cette partie nous n'étudierons que le cas du supervisé.

L'objectif est de calculer la matrice de poids optimale au sens d'une fonction d'erreur préalablement définie. Cette fonction d'erreur est souvent l'erreur quadratique.

Les grandes étapes de l'apprentissage sont :

- Initialiser la matrice des poids $W(0)$
- Mettre des données en entrée et propager l'activation de manière à calculer la sortie s
- Calculer l'erreur entre s et la sortie désirée d. Cette étape est possible en supervisé car nous avons les valeurs attendues en sortie.
- En déduire une nouvelle matrice W(t+1) à l'aide d'une règle d'apprentissage.

5.1.4 Compromis biais/variance

Définition 5.2. L'erreur de généralisation est l'erreur commise en appliquant le modèle sur un nouveau jeu de données.

Définition 5.3. Soient la prédiction g au point x paramétrée par w notée $g(x, w)$ et $G(x, W)$ la variable aléatoire correspondante.
L'erreur de prédiction théorique notée P^2 estimant l'erreur de généralisation en un point x est donnée par la relation :

$$P^2 = \underbrace{\sigma^2}_{\text{bruit}} + \underbrace{var(G(x,W))}_{\text{variance}} + \underbrace{(E(f(x) - G(x,W)))^2}_{\text{biais}}$$

avec :

- **le biais** = l'ajustement du modèle sur les données d'apprentissage
- **la variance** = la variabilité en appliquant le modèle sur les nouvelles données

Le biais et la variance varient en sens inverse de la complexité du modèle. Plus un modèle est complexe plus le biais est faible et la variance importante et inversement. Il est donc important de trouver un bon **compromis entre l'ajustement et la robustesse du modèle**. Ce compris s'appelle **le compromis biais/variance**.

5.2 Théorie du PMC

5.2.1 Définition du réseau

Les PMC (multi-layer perceptron) sont des « approximateurs universels de fonctions». Ils sont composés d'une couche d'entrée, d'une ou plusieurs couches cachées et d'une couche de sortie.

Les neurones de la couche i sont reliés à tous les neurones de la couche i+1. Les neurones d'une même couche ne sont pas interconnectés. Un neurone ne peut donc transmettre son état qu'à un neurone de la souche postérieure à la sienne. On parle alors de neurones « à propagation avant » (**feedforward neural network**) : une couche ne peut utiliser que les sorties des couches précédentes.

Les fonctions d'entrée sont linéaires pour les neurones i des couches cachées et de la couche de sortie et sont de la forme :

$$h(i) = h(i, a, W) = \sum_{j=1}^{N} (w_{i,j} * a_j)$$

Pour les fonctions d'activations :

- La fonction d'activation de la couche de sortie est la fonction d'identité
- Celle de la couche cachée est une fonction dérivable. On choisit souvent une **sigmoïde exponentielle** ou tangentielle afin de représenter les phénomènes non linéaires. La propriété de dérivation est essentielle pour la méthode de rétropropagation du gradient qui permet l'ajustement des poids. La fonction d'activation de type "sigmoïde" ou logistique est de la forme : $s = \frac{1}{1+exp^{-\alpha}}$

5.2.2 Notations

- i : indice de la cellule
- p : nombre de cellule de la couche de sortie
- d_i : la sortie désirée pour une cellule de sortie
- w_{ij} : le poids synaptique de la cellule j vers la cellule i
- Pred(i) : l'ensemble des cellules en entrée de la cellule i
- Succ(i) : l'ensemble des cellules en sortie de la cellule i
- y_i : l'entrée totale de la cellule i. $y_i = \sum_{j \in Pred(i)} w_{ij} x_{ij}$
- a_i : la sortie de la cellule i obtenue. $a_i = \phi(y_i)$
- $\phi(x)$: la fonction d'activation (fonction sigmoïde).
  $\phi(x) = \frac{1}{1+e^{-x}}$ et $\phi(x)' = \phi(x)(1 - \phi(x))$
- E : l'erreur d'apprentissage
- n : le nombre d'entrées
- p : le nombre de sorties
- $\overrightarrow{w}$: le vecteur des poids
- S : la couche de sortie
- C : la couche cachée

5.2.3 Initialisation des poids

Avant de lancer un algorithme d'apprentissage, il est important de le paramétrer et cela commence par l'initialisation des poids et des biais. Les poids jouent un rôle très important dans l'apprentissage. De ce fait, s'ils ne sont pas correctement initialisés cela aura un

gros impact sur la généralisation et la vitesse d'apprentissage de l'algorithme.

Pour cela, plusieurs méthodes de paramétrages existent. La plus courante est la méthode décrite par Smieja [SMI91] qui propose d'initialiser les poids de la couche cachée vers la couche de sortie à 0 et les poids de la couche d'entrée vers la couche cachée par un tirage aléatoire selon une loi normale centrée réduite d'intervalle : $\left[\frac{-2}{\sqrt{d_{in}}}, \frac{2}{\sqrt{d_{in}}}\right]$ avec d_{in} est le nombre de connexions entrantes.

5.2.4 Apprentissage

Qu'est ce que l'apprentissage?

Définition 5.4. Apprentissage automatique - Mitchell 1997
Soit E l'ensemble de toutes les tâches possibles et S un système. Soit T $\in E$ un ensemble de tâches appelé training set. Soit P:SxE $\rightarrow R$ une mesure de performance d'un système sur des tâches. Alors un système S apprend lors des expériences si la performance de S sur les tâches T mesurée par P s'améliore c'est-à-dire si :

$$P(S_{\text{avant expérience}}, T) \leq P(S_{\text{après expérience}}, T)$$

Le but de l'apprentissage est de trouver les poids optimaux pour bien approcher les données.

Soit un PMC avec une couche cachée. Notions d_i les sorties désirées et a_i les sorties obtenues avec notre réseau avec $i \in S$ (S la couche de sortie). Le but de cet algorithme d'apprentissage est de faire converger le modèle vers un minimum d'erreur appelée fonction de coût entre les valeurs attendues et celles calculées. Pour chaque neurone du réseau, le gradient de l'erreur est calculé de manière à mesurer la contribution de chacun des poids synaptiques à l'erreur commise. Par ce biais, les corrections sont effectuées au fur et à mesure que les exemples d'apprentissage sont présentés. Comme la modification d'un poids influe sur tous ceux des neurones des couches suivantes, les corrections d'erreur doivent être propagées de la dernière couche vers la première.

Sur un ensemble d'échantillon, l'erreur du PMC est définie de la façon suivante :

$$E(\overrightarrow{w}) = \sum_{s \in S} \frac{1}{2} \sum_{k=1}^{p} (d_k^s - a_k^s)^2$$

Ainsi, en réduisant à un seul exemple, on obtient la fonction d'erreur suivante :

$$E_s(\overrightarrow{w}) = \frac{1}{2} \sum_{i \in S} (a_i - d_i)^2$$

Il convient ensuite d'appliquer la règle de modification des poids (règle DELTA):

$$\Delta w_{i,j} = -\lambda \frac{\partial E}{\partial w_{i,j}}$$

et de minimiser l'erreur sur chaque exemple (et non sur l'erreur globale).

En appliquant la règle DELTA, on obtient la règle de modification des poids suivante[1]:

$$\begin{aligned} \forall i \in S : \Delta w_{i,j} &= \lambda a_j (d_i - a_i) f'(h(i)) \\ \forall i \in C : \Delta w_{i,j} &= f'(h(i)) \sum_{k \in S} (w_{k,i}(d_k - a_k) f'(h(k)) \end{aligned}$$

Avec λ le pas d'apprentissage. Il permet d'adapter une équation d'un système dynamique dans laquelle la variable temps t varie de façon continue sur $\mathbf{R}^+$ à un cas où t varie de manière discrète dans $\mathbf{N}$. Il existe différentes manières de déterminer la valeur du pas d'apprentissage, mais nous n'étudierons pas ce point ici.

On remarque donc que pour la couche cachée l'erreur est une pondération des erreurs sur les couches suivantes. Il s'agit de la rétropropagation du gradient.

Les différentes stratégies d'apprentissage

Deux principes fondamentaux guident les différentes stratégies employées pour entraîner des PMC :

- Entraîner aussi efficacement que possible, c'est-à-dire faire baisser l'erreur d'entraînement aussi vite que possible, éviter d'être bloqué dans une vallée étroite ou un minimum local de la fonction de coût
- Contrôler la capacité, de manière à éviter le sur-apprentissage, afin de minimiser l'erreur de généralisation

L'optimisation du critère d'apprentissage dans les réseaux de neurones multi-couches est difficile car il y a de nombreux minima locaux. Cependant, dans de nombreux cas, on peut se limiter à un «bon» minimum local.

5.2.5 La rétropropagation du gradient

L'algorithme de retropropagation consiste à faire un calcul récursif du gradient sur l'ensemble des unités du PMC.

Le critère d'arrêt

Les étapes de calcul des δ_i et la mise à jour des poids sont à faire jusqu'à ce que le critère d'arrêt soit satisfait. Il existe plusieurs critères d'arrêt possible :

[1]Démonstration en annexe

- **Nombre maximum d'itération** : permet de contrôler le temps d'exécution de l'algorithme.
- **Early stopping** : moment où l'erreur quadratique moyenne (EQM ou MSE en anglais : Mean Square Error) devient inférieure à un certain seuil. Permet d'éviter le surapprentissage.

C'est cette méthode qui va être utilisée dans notre algorithme de réseau de neurones.

Early stopping Cette technique permet d'anticiper le phénomène de surapprentissage en arrêtant prématurément la présentation des exemples. Un premier cycle est réalisé en présentant aléatoirement tous les exemples de la base d'apprentissage. Un second est réalisé sur la base de validation. Le processus s'arrête dès lors que l'erreur simulée sur la base de validation a atteint son minimum.

L'EQM est la moyenne arithmétique des carrés des écarts entre les prévisions et les observations. Elle se calcule de la façon suivante :

$$EQM(i) = \frac{1}{N}\sum_{k=1}^{N}(y_k - g(x_k))^2$$

avec :

- N : le nombre d'exemples dans la base d'apprentissage
- y_k : la k-ième valeur attendue
- $f(x_k)$: la k-ième valeur prédite

Dans le cadre d'une régression, on cherche à minimiser cette valeur. Cette méthode est fondée sur le théorème de König : on cherche à minimiser la variance résiduelle.

Théorème 5.5. ***Théorème de König***
Pour toute variable aléatoire réelle X qui admet un moment d'ordre 2, on a :

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

5.2.6 Comment choisir l'architecture du réseau ?

Choisir l'architecture d'un PMC revient à choisir :

- Le nombre de couche cachée
- Le nombre de neurones par couche
- La nature des différentes connexions entre les neurones

La couche d'entrée

Le nombre de neurones de la couche d'entrée est égale au nombre d'entités c'est-à-dire de colonnes dans les données étudiées (nombre de variables). Dans certaines configurations de réseaux de neurones (ou certains packages) un neurone supplémentaire est ajouté pour le biais.

Il existe un mythe urbain selon lequel le fait d'ajouter des données améliorerait l'estimation du réseau de neurones. Le théorème d'approximation parcimonieuse contredit cette idée.

Théorème 5.6. ***Théorème d'approximation parcimonieuse [HOR94]***
Si le résultat de l'approximation est une fonction non linéaire des paramètres ajustables, elle est plus parcimonieuse que si elle était une fonction linéaire de ses paramètres. De plus, pour des réseaux de neurones à fonction d'activation sigmoïdale, l'erreur commise dans l'approximation varie comme l'inverse de neurones cachés, et est indépendante du nombre de variables de la fonction à approcher. Par conséquent, pour une précision fixée (un nombre de neurones donné), le nombre de paramètres du réseau est proportionnel au nombre de variable de la fonction à approcher.

Ce théorème signifie que les réseaux de neurones à une couche cachée étant des approximateurs parcimonieux, ils sont capables d'approximer correctement d'avantage de fonctions que des polynômes pour un même nombre de paramètres. Ce nombre de paramètres croît de façon linéaire avec le nombre de variables (et non pas exponentiellement comme avec les approximateurs linéaires). Par conséquent, il est possible de limiter le nombre d'exemples permettant d'arriver à une estimation de la fonction de régression.

Ainsi, nous pouvons justifier le fait de n'utiliser qu'une cinquantaine de tempêtes en entrée et non des milliers.

La couche de sortie

Concernant la couche de sortie, une chose est à retenir. Si on est face à un PMC régresseur, la couche de sortie possède un neurone tandis qu'il en possédera plusieurs dans le cas d'une classification.

Nombre de couches cachées Il n'existe pas de règle précise pour trouver le nombre optimal de couche cachée et de neurones par couche. Cependant, le théorème d'approximation universelle nous donne une piste sur le nombre de couche :

Théorème 5.7. ***Approximation universelle [CYB89]]***
Toute fonction bornée suffisamment régulière peut être approchée uniformément, avec une précision arbitraire, dans un domaine fini de l'espace des variables par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire.

Pour en savoir plus sur cette propriété et les différentes approches de démonstration, vous pouvez vous référer à l'article de P.Common[2] , sur la classification supervisée par réseaux multicouches.

Nous nous sommes appuyés sur ce théorème pour la création d'un PMC à une couche. Ce choix a également été validé par un consensus de personnes utilisant les réseaux de neurones. Selon eux, l'ajout de couche cachée n'améliore pas forcément la performance du réseau. Au contraire, dans la majorité des cas, une seule couche cachée est suffisante.

Dans son livre intitulé « Introduction to Neural Networks for Java, Section 2 », Jeff Heaton crée un tableau d'aide à la décision pour le nombre de couche :

- 0 - Seulement capable de représenter des fonctions ou des décisions linéaires séparables.
- 1 - Peut approximer n'importe quelle fonction qui contient une cartographie continue d'un espace fini à l'autre.
- 2 - Peut représenter une limite de décision arbitraire à une précision arbitraire avec des fonctions d'activation rationnelle et peut approximer tout mappage à toute précision.

L'ensemble de ces éléments nous ont amené à faire le choix de la création d'un réseau de neurone avec une unique couche cachée.

Nombre de neurones par couche cachée Déterminer le nombre de couche cachée n'est pas un problème bien complexe à résoudre. Ce qui est le plus important est de réussir à déterminer correctement le nombre de neurones dans la ou les couches cachées. En effet, bien que ces couches n'interagissent pas directement avec l'environnement externe, elles ont une influence considérable sur la sortie finale. Utiliser trop peu de neurones dans la couche cachée entraînera de l'**underfitting**. Les phénomènes complexes risquent donc d'être laissés de côté. De l'autre côté, l'utilisation de trop de neurones peut entraîner de nombreux problèmes tels que : le **surapprentissage**, **un temps d'entraînement du réseau très important**. Il faut donc trouver un compromis entre trop peu et trop de neurones. Il n'existe pas de méthodes précises pour trouver ce nombre de neurones optimales, cependant, de nombreuses règles empiriques existent pour orienter dans le choix de ce nombre telles que :

- Le nombre de neurones cachés doit être compris entre la taille de la couche d'entrée et la taille de la couche de sortie.
- Le nombre de neurones cachés doit être égal à 2/3 de la taille de la couche d'entrée, plus la taille de la couche de sortie.

[2][COMMON1991]

- Le nombre de neurones cachés doit être inférieur à deux fois la taille de la couche d'entrée.

Cependant, il est toujours préférable de créer un réseau trop important puis de supprimer des neurones grâce des méthodes telles que l'élagage (i.e réduire ou contrôler le nombre de paramètre non-nuls), la régularisation (i.e modifier le problème d'apprentissage de sorte que l'optimisation est susceptible de trouver un réseau de neurones avec un petit nombre de paramètres) plutôt que d'en oublier.

Toutefois, il est possible de borner le nombre de neurones maximum à utiliser grâce au théorème de Vapnik-Chervonenkis.

Théorème 5.8. ***Théorème de Vapnik-Chervonenkis*** *Si la VC-dimension notée d_{vc} est finie, alors $\forall f \in F$, avec une probabilité au moins égale à $1-\delta$, pour $n > d_{vc}$:*

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{d_{vc}(\log(\frac{2n}{d_{vc}}) + 1) - log(\frac{\delta}{4})}{n}}$$

où:

- $R(f) = \int (y - f(x,\Theta))^2 dP(x,y)$ *est le risque réel avec :*
  - $f(x,\Theta)$ *la valeur prédite*
  - y *la valeur attendue*
  - dP *la loi de probabilité jointe*
- $R_{emp} = \sum_{i=1}^{n} (y_i - f(x_i,\Theta))^2$ *est le risque empirique*
- $R_{emp} \leq R(f)$

Karpinski et Macintyre[3] ont publié en 1995 des travaux dans lesquels ils ont montré que le **nombre de cellules cachées notées M_n est inférieur à $O(\sqrt[4]{n})$.** Cette borne théorique est souvent trop large en pratique mais elle présente l'avantage d'être universelle et ne nécessite aucune hypothèse sur les données.

5.2.7 La Cross-Validation

La validation croisée est une autre étape très importante de la construction de modèles prédictifs. Bien qu'il existe différents types de méthodes de validation croisée, l'idée de base consiste à répéter le processus suivant un certain nombre de fois:

- Division du jeu de données en deux échantillons l'un de test l'autre d'apprentissage
- Monter le modèle sur l'échantillion d'apprentissage

[3] [KAR95]

- Tester le modèle sur l'échantillon de test
- Calculer l'erreur de prédiction
- Répéter le processus K fois

Ensuite, en calculant l'erreur moyenne, nous pouvons avoir une idée de la façon dont le modèle fonctionne.

Dans le cas du neuralnet, la cross validation peut prendre un long moment. Il est donc conseillé de créer une barre de progression afin de visualiser l'avancement. Cette dernière peut facilement se faire grâce au package « plyr » .

Il est possible de visualiser l'erreur engendrée grâce à bloxplot. Cela permet de plus facilement interpréter l'erreur et enlève « un peu » le côté boite noire.

La cross-validation permet d'avoir une estimation plus précise sur l'erreur. Cela est très important dans le cas d'une régression afin de savoir où l'on en est.

5.2.8 Intervalle de confiance, élagage et pertinence

Les poids d'un réseau de neurones suivent une distribution normale multivariée si le réseau est identifié (White 1989 => Neural Computation 1:425-464)

Un réseau de neurone est identifié s'il ne comprend aucun neurone non pertinent ni dans la couche d'entrée ni dans la couche de sortie. Un neurone non pertinent dans la couchée d'entrée peut par exemple être une covariable qui n'a aucun effet ou qui est une combinaison linéaire des autres variables explicatives. Si cette condition est remplie et si la fonction d'erreur est égale à la log vraisemblance négative, alors un intervalle de confiance peut être calculé pour chaque poids.

Pour identifier un réseau de neurone, il est possible d'utiliser des méthodes d'élagage. Pour chaque neurone, on calcule sa pertinence et coupe la connexion lorsque la pertinence associée est faible.

Méthode d'élagage - Algorithme de Garson

La technique dite d'élagage ou de pruning en anglais a pour objectif de limiter le surapprentissage en réduisant la complexité du modèle.Le principe est de supprimer, une fois l'apprentissage terminé, les connexions du modèle ayant la plus faible influence sur l'erreur de sortie du réseau.

L'algorithme de Garson, créé en 1912, permet d'identifier l'importance relative des variables explicatives pour des variables de réponses spécifiques dans un réseau de neurone supervisé en déconstruisant les poids du modèle.

En effet, les poids qui relient les variables dans un réseau de neurones sont partiellement analogues aux coefficients de régression et peuvent être utilisés pour décrire les relations entre les variables. Les pondérations dictent l'influence relative des informations traitées dans le réseau de sorte que les variables d'entrées qui ne sont pas pertinentes dans leur corrélation avec la variable de réponse soient supprimées.

L'algorithme de Garson calcule l'importance d'une variable de la façon suivante :

$$\text{importance relative} = \frac{Q_{ik}}{\sum_{i=1}^{N}\sum_{j=1}^{L}(|w_{rj} * w_j|/\sum_{r=1}^{N}|w_{rj}|)}$$

où Q_{ik} est le pourcentage d'influence de la variable d'entrée sur la sortie et est égale à :

$$Q_{ik} = \frac{\sum_{j=1}^{L}|w_{rj} * w_j|)}{\sum_{r=1}^{N}|w_{rj}|}$$

avec:

- w_{ij} la connexion entre le neurone d'entrée i et le neurone caché j
- w_j le poids du neurone caché j à la sortie
- $\sum_{r=1}^{N}|w_{rj}|$: la somme de la connexion des poids entre les N neurones d'entrée et le neurone caché j

L'utilisation de l'algorithme de Garson permet donc de ne garder que les variables les plus pertinentes, allégeant ainsi l'exécution. Les conséquences sont nombreuses : réduction du temps d'exécution, meilleure lisibilité ...

5.2.9 Les avantages et inconvénients du PMC

- **Les avantages du PMC sont :**

  1. Sa grande capacité à traiter les bases "irrégulières" : les variables peuvent être de natures différentes (quantitative, qualitative, catégorielle...). De plus, grâce à son étape de normalisation obligatoire, aucun effet d'échelle n'est présent et chaque donnée est traitée de manière équivalente.
  2. Les réseaux de neurones sont d'une grande précision
  3. Les réseaux de neurones fonctionnent correctement en présence de données bruitées.

- **Les inconvénients du PMC sont :**

  1. Le paramétrage d'un réseau de neurone est très long et compliqué. Les paramètres à définir sont nombreux et dépendent de beaucoup de facteurs. Or, les écarts

d'estimation sont très importants suivant les paramètres choisis comme nous avons pu le voir dans l'application.

2. Il est très compliqué de comprendre le fonctionnement interne d'un réseau de neurone du fait de son statut de "boîte noire".
3. Il est très difficile de sélectionner les variables explicatives parmi les variables candidates. En effet, le réseau de neurones fonctionne même en présence de données non significatives en entrée. Un travail supplémentaire en plus du paramétrage doit être effectué pour choisir les données pertinentes.

5.3 Application du PMC sur nos données

5.3.1 Choix de l'algorithme d'apprentissage

L'apprentissage repose sur la modification des poids. La prise en compte de cette modification peut être plus ou moins longue suivant la technique utilisée. Dans cette partie, on ne s'intéresse pas à l'amélioration de la prédiction mais à la vitesse d'exécution de l'algorithme. Ce critère souvent négligé est essentiel dans notre cas. En effet, la contrainte de temps est très importante. Il convient donc de trouver un modèle à la fois fiable et rapide.

Backprop : Back propagation

L'algorithme Backprop est l'un de plus connu et utilisé pour calculer un gradient qui est nécessaire dans le calcul des poids. C'est une technique d'apprentissage du 1er ordre. Les algorithmes présentés ensuite ne seront que des modifications et améliorations de cet algorithme.

Algorithm 1: Algorithme de rétropropagation du gradient

Data: S : l'échantillon d'entrée;
C_0 :la couche d'entrée;
C_q : la couchée de sortie;
$q-1$: nombre de couche cachée;

Initialisation des poids $w_{i,j}$;

Calcul de la sortie o_i en propageant les entrées;

while *Le critère d'arrêt n'est pas satisfait* **do**
  for *Chaque neurone de la couche de sortie* **do**
    $\delta_i = o_i(1-o_i)(c_i - o_i)$
  for *Chaque couche cachée* **do**
    for *Pour chaque neurone de la couche cachée étudiée* **do**
      $\delta_i = o_i(1-o_i)\sum_{k \in Succ(i)} \delta_k w_{ki}$
  for *Chaque poids* w_{ij} **do**
    $w_{ij} = w_{ij} + \epsilon \delta_i x_{ij}$

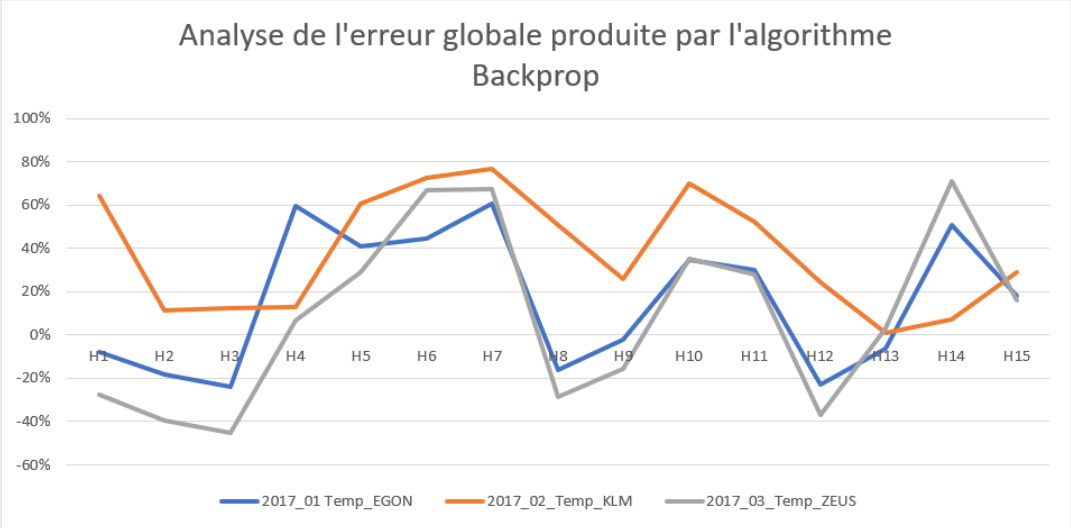


Figure 5.3.2: Estimation de l'erreur en fonction du nombre de neurones backprop

Rprop : Resilient Propagation

La technique Rprop est une technique du premier ordre inventée par Richard et Braun en 1993 [RIE93]. Cet algorithme est l'un des algorithmes les plus rapides. Le principe de la règle de mise à jour des poids est la suivante :

$$\Delta w_{i,j} = -sign(\frac{\partial E}{\partial w_{i,j}}) * \epsilon_{i,j}$$

où $\epsilon_{i,j}$ (l'update value) représente le pas d'apprentissage.

Le principe est le suivant : Le pas d'apprentissage s'adapte en fonction de la direction du gradient par rapport au gradient précédent. Si le gradient ne change pas, alors le pas d'apprentissage augmente sinon il diminue. Lorsqu'il y a un changement de signe, cela signifie que le minimum local a été manqué. Il faut donc revenir en arrière pour l'atteindre et ralentir le rythme pour ne pas passer à côté. Au contraire, tant qu'il n'y a pas de changement de signe, on est sur la bonne voie, il est donc possible d'accélérer le pas d'apprentissage pour accélérer l'apprentissage.

A la $k^{ème}$ itération, le pas d'apprentissage $\epsilon_{i,j}(t)$ est :

$$\epsilon_{i,j}(k+1) = \begin{cases} min(\epsilon_{i,j}(k) * up, \epsilon_{max}) si \frac{\partial E}{\partial w_{i,j}}(k) * \frac{\partial E}{\partial w_{i,j}}(k-1) > 0 \\ max(\epsilon_{i,j}(k) * down, \epsilon_{max}) si \frac{\partial E}{\partial w_{i,j}}(k) * \frac{\partial E}{\partial w_{i,j}}(k-1) < 0 \\ \epsilon_{i,j}(k) sinon \end{cases}$$

où $0 < d < 1 < u$

La mise à jour des poids devient alors :

$$\Delta w_{i,j}(k+1) = \begin{cases} -\epsilon_{i,j}(k) sign(\frac{\partial E}{\partial w_{i,j}}) si \frac{\partial E}{\partial w_{i,j}}(k) * \frac{\partial E}{\partial w_{i,j}}(k-1) \leq 0 \\ 0 sinon \end{cases}$$

Il existe différentes variantes de l'algorithme Rprop :

- **Rprop+** : Rprop avec retour en arrière des poids
  Après avoir ajusté le pas d'apprentissage, les poids sont ajustés :

  $$si \frac{\partial E^{(t-1)}}{\partial w_{i,j}} * \frac{\partial E^{(t)}}{\partial w_{i,j}} \leq 0 \text{ alors } \Delta w_{i,j}^{(t)} := -sign(\frac{\partial E^{(t)}}{\partial w_{i,j}}) * \Delta_{i,j}^{(t)}$$

  En cas de changement de signe de la dérivée partielle, la mise à jour des poids précédente est annulée :

  $$si \frac{\partial E^{(t-1)}}{\partial w_{i,j}} . \frac{\partial E^{(t)}}{\partial w_{i,j}} < 0 \text{ alors } \Delta w_{i,j}^{(t)} := -\Delta w_{i,j}^{(t-1)} \text{ et } \frac{\partial E^{(t)}}{\partial w_{i,j}} := 0$$

  La modification des poids devient alors :

  $$w_{i,j}^{(t+1)} := w_{i,j}^{(t)} + \Delta w_{i,j}^{(t)}$$

- **Rprop-** : Rprop sans retour en arrière des poids
  Avec l'algorithme Rprop-, il n'y a pas de retour en arrière, il n'est pas donc pas nécessaire de stocker la mise à jour des poids précédents. Cela rend l'algorithme encore plus rapide.

En effectuant les différentes simulations, on obtient les résultats suivants :

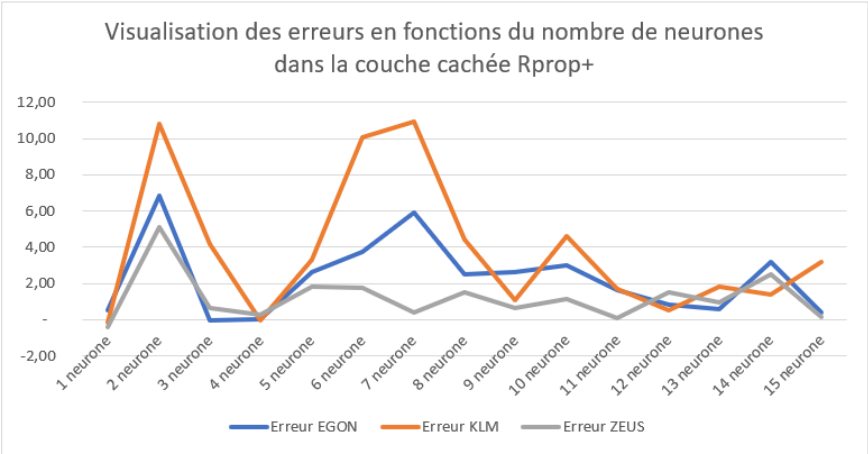


Figure 5.3.3: Estimation de l'erreur en fonction du nombre de neurones Rprop+

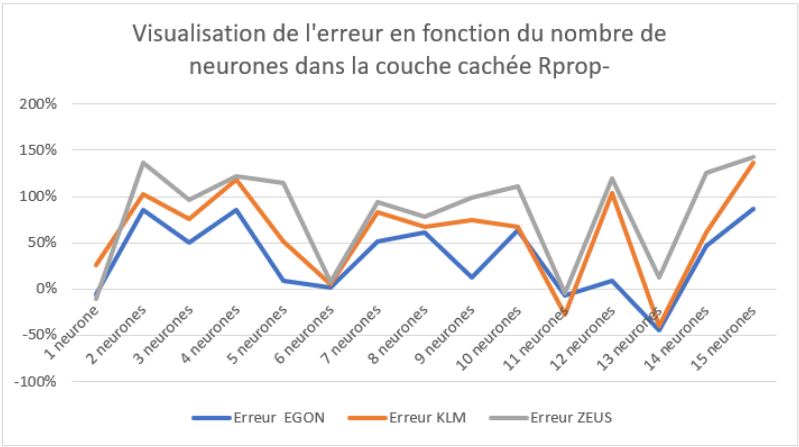


Figure 5.3.4: Estimation de l'erreur en fonction du nombre de neurones Rprop-

L'ensemble des différentes simulations nous on permis de faire les conclusions suivantes:

- rprop+ meilleure que rprop-. Cela peut s'expliquer par le faire que l'algorithme ne met pas à jour les poids, ce qui rend l'algorithme plus rapide à s'exécuter.
- L'algorithme backprop met beaucoup plus de temps à s'exécuter que les algorithmes rprop. La pratique confirme donc la théorie.
- les prédictions effectuées avec l'algorithme backprop sont meilleures qu'avec les autres algorithmes.

5.3.2 La cross-validation

Le résultat des réseaux de neurones dépend de l'initialisation des paramètres. Deux choix sont donc possibles : trouver une moyen d'initialiser les poids de façon optimale ou réaliser

une cross-validation sur notre PMC et trouver la prédiction en moyennant les résultats obtenus.

Nous avons décidé d'utiliser la méthode de la cross-validation car cette dernière permet une meilleur prédiction. En effet, comme il n'existe pas de méthode connue permettant de trouver une initialisation optimale.

En appliquant la cross-validation sur l'ensemble des algorithmes d'apprentissage nous obtenons les résultats suivants :

Algorithme	Nombre de neurones	Tempête1	Tempête2	Tempête3	temps exécution
Backprop	13	15.1	11.4	21.2	21.7 min
Rprop +	4	16.1	13.5	19.1	3.2 min
Rprop -	6	16.6	13.8	17.2	2.5 min

Tableau 5.3.1: Synthèse des meilleurs résultats obtenus par algorithmes d'apprentissage

Il est possible de visualiser les différentes erreurs générées par les différents réseaux de neurones grâce à un boxplot et le temps d'exécution. Prenons le cas de la prédiction de la tempête Tempête1. Nous simulons 15 créations de réseaux de neurones avec une initialisation différente et l'algorithme d'apprentissage Backprop à 13 neurones cachés. Suivant l'initialisation de départ, l'estimation et le temps d'exécution peuvent différer. La cross-validation permet donc de moyenner les résultats obtenus et d'obtenir un résultat satisfaisant. Les résultats obtenus sont les suivants :

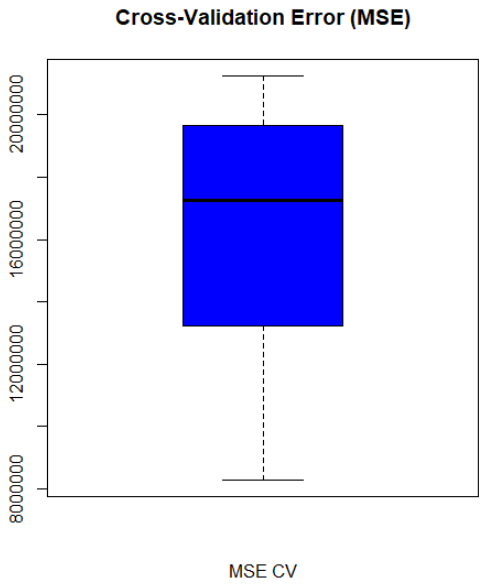


Figure 5.3.5: Cross Validation de l'erreur d'estimation d'un PMC

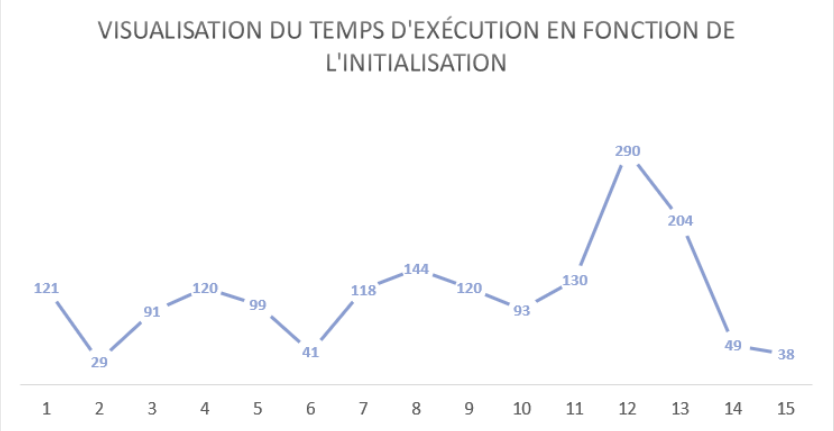


Figure 5.3.6: Visualisation du temps d'exécution en fonction de l'initialisation du PMC

La figure 5.2 est très intéressante car elle nous permet de visualiser l'importance de l'initialisation. En effet, en cas de mauvaise initialisation (comme c'est le cas dans le test numéro 12), le temps peut-être multiplié par 10 par rapport à une bonne initialisation (cas numéro 2). La cross-validation est donc un bon moyen de résoudre le problème de la mauvaise initialisation.

5.3.3 Choix du seuil

Le choix du seuil comme celui du nombre maximal d'itération est très important. En effet, celui-ci conditionne non seulement le résultat mais également le temps d'exécution du réseau de neurones.

On pourrait supposer qu'un seuil plus grand entrainerait systématiquement une meilleure prédiction. Cependant, cela n'est pas toujours le cas comme nous avons pu le constater lors des différentes simulations.

En effet, le seuil et le pas d'apprentissage sont très liés. Le pas d'apprentissage devant toujours être inférieur au seuil, lorsque ce dernier diminue, le pas d'apprentissage diminue également. Or, si ce dernier est trop petit, on prend le risque de rester coincé dans un mauvais minimum local et de ne pas pouvoir en sortir.

L'ensemble des différents tests que nous avons mené, nous ont permis de conclure que le seuil le plus adapté était de 0.01 et le pas d'apprentissage de 0.001.

5.3.4 Sélection des variables explicatives

Le PMC prend en compte l'intégralité des variables, même celles qui ne sont pas pertinentes. Cette capacité permet aux PMC de tirer toute l'information disponible mais

ralentit grandement l'apprentissage et son exécution. Afin de gagner en rapidité sans perdre en précision, il est possible d'effectuer une sélection des variables explicatives en ne gardant que celles qui sont les plus pertinentes.

Simulons l'algorithme de Garson sur notre PMC. On obtient le graphique suivant :

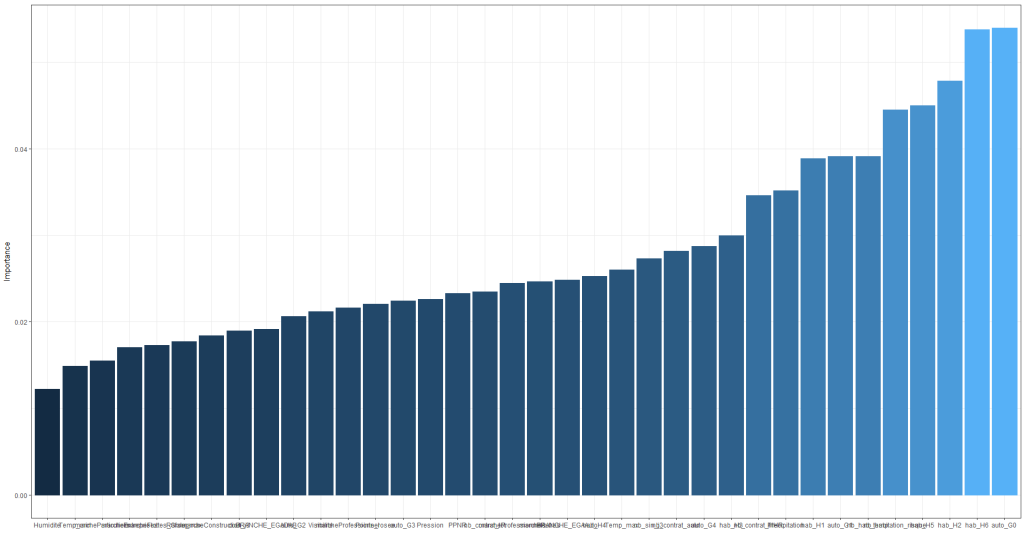


Figure 5.3.7: Résultat de l'algorithme de Garson sur notre PMC

L'algorithme de Garson est un algorithme itératif. A chaque suppression de variable, il est nécessaire de le relancer pour vérifier la pertinence de chaque variable. En effet, une variable peut devenir pertinente à la suppression d'une autre variable. Ce travail est long et fastidieux mais permet de gagner de précieuses minutes dans l'exécution du PMC et permet de gagner en lisibilité une fois.

L'un des inconvénients majeurs de l'élagage est qu'à chaque variable supprimée, il faut refaire l'intégralité du travail énoncé précédemment. Le choix des paramètres optimaux est alors un travail fastidieux. Le PMC optimal prend du temps à tourner, et un mauvais paramétrage rend le réseau de neurones moins performant dans sa capacité à prédire.

5.3.5 Le cas des Grêles et des Orages

Maintenant que nous avons étudié le cas des tempêtes tests, intéressons-nous aux grêles et orages. Pour rappel, ces derniers étaient mal estimés avec le modèle GAMLSS. Qu'en est-il avec le réseau de neurones ?

On remarque que lorsqu'il s'agit d'événements spéciaux, le réseau de neurones estime très bien et bien mieux que le GAMLSS. Cette différence provient du fait que le réseau de neurones est un estimateur universel. Il crée donc une loi capable de prendre en compte les cas extrêmes, cas qui différent, contrairement aux GLM où ils sont noyés dans la masse.

Nom de la tempête	sinistres normaux	grave	Total	sinistres normaux	grave	Total
Grêle	1.5	0.1	1.6	1.5	0.1	1.6
Orage 1	5.3	1.4	6.7	5.6	0.9	6.5
Orage 2	6.4	2.1	8.5	6.7	2.3	9.1
Orage 3	3.9	0.7	4.6	4.3	0.4	4.7

Tableau 5.3.2: Génération des estimations pour les grêles et orages par la méthode du PMC

5.3.6 Visualisation du PMC final

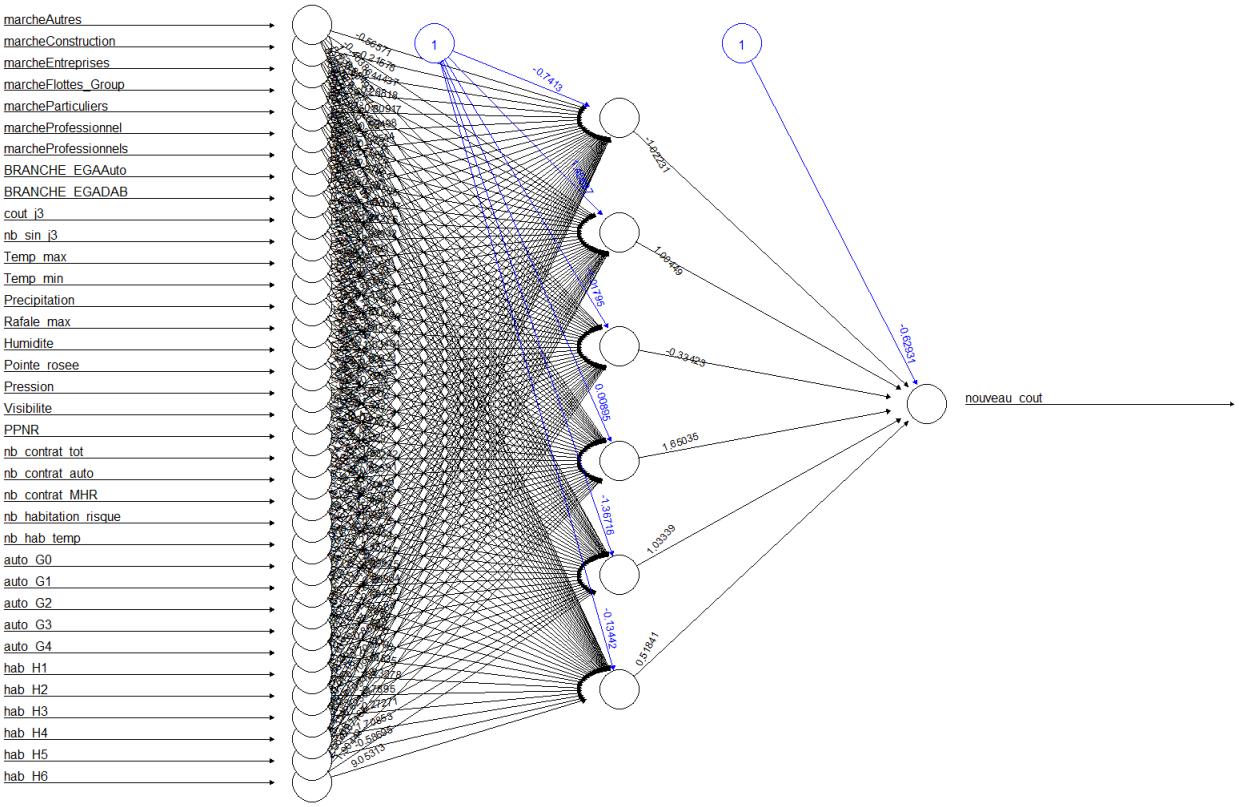


Figure 5.3.8: Visualisation de notre PMC final

Chapitre 6

Modèle final

6.1 Récapitulatif des résultats

La prédiction de la sinistralité des catastrophes naturelles repose sur de nombreuses hypothèses dont une fondamentale : les tempêtes futures ressembleront aux tempêtes passées. Pour que cette assertion soit vraie, il est nécessaire de retravailler les tempêtes historiques pour qu'elles s'adaptent au présent.

Une fois le travail de reconstitution de l'historique effectué, nous avons tester plusieurs modèles pour prédire le coût et le nombre des tempêtes. Dans cette partie, nous récapitulerons les résultats obtenus et ferons nos choix définitifs.

6.1.1 Choix des tempêtes de référence

Nous avons vu lors de ce mémoire que le choix des tempêtes de référence était fondamental. Un mauvais choix rendra la modélisation future biaisée. Alors que choisir ? Classification supervisée (KNN), non supervisée (CAH) ou dire d'expert? Choisir la méthode la plus adaptée n'a pas été facile car chacune des méthodes a ses particularités, ses avantages et ses inconvénients.

Le savoir humain est très important car il repose sur une expertise basée sur l'expérience et sur le ressenti. Les modèles traditionnels se servent des événements passés pour prédire le futur. Les ordinateurs toujours plus puissants permettent aux modèles traditionnels d'engranger de nombreux événements passés pour affiner l'apprentissage par expérience. Ils deviennent alors plus performants que l'humain. Cependant, lorsque certains événements sortent de l'ordinaire, la prédiction de l'humain peut surpasser celle de l'ordinateur, en se servant de son intuition. Capacité dont l'ordinateur est dépourvue.

Fort de ce constat, nous avons décidé de garder le dire d'expert dans notre modèle. Il sera cependant lié à la méthode des K plus proches voisins pour affiner l'estimation.

6.1.2 Modélisation du nombre de sinistres d'une tempête

L'estimation du nombre de sinistres pour une tempête peut-être faite de deux façons différentes : soit en utilisant les tempêtes de références prédites précédemment soit en

utilisant l'ensemble des tempêtes. La première méthode permet d'obtenir une estimation plus fine si les tempêtes de référence sont correctement choisies. Cependant, en cas d'erreur dans l'estimation des tempêtes de références, la modélisation peut être grandement faussée. Afin d'éviter que cela n'arrive, deux méthodes vont être combinées : l'estimation du nombre par une régression Poisson sur l'ensemble des données et l'estimation par la méthode des cadences moyennes sur les tempêtes de référence.

Lorsque les résultats seront similaires ou presque (à 200 près), la méthode des cadences moyennes sera privilégiée. Cependant, en cas de gros écarts, l'expert se servira de son intuition pour trancher entre les deux méthodes de prédiction. Des tests sur l'année 2018 et 2019, devraient nous permettre de déterminer la méthode la plus adaptée et de supprimer l'autre du modèle afin de gagner du temps.

6.1.3 Modélisation du coût d'une tempête

La modélisation du coût global d'une tempête a été la partie la plus délicate. De prime abord, la modélisation par GAMLSS Gumbel semblait la modélisation la plus adéquate. Mais les grêles et orages de début d'année nous ont fait nous interroger sur la pertinence de ce modèle.

Les réseaux de neurones semblaient donc être la solution puisqu'ils sont à la fois robustes et performants. De plus, comme ce sont des prédicteurs universels, ils sont en mesure d'estimer les phénomènes rares tels que les orages ou les grêles. Cependant, les tests de la partie précédente nous ont montré que les réseaux de neurones lorsqu'ils sont mal configurés sont de très mauvais prédicteurs d'où l'importance de passer par des étapes intermédiaires pour trouver les paramètres optimaux. Et leur paramétrage est si complexe que pour des résultats rapides et lisibles, il est souvent préférable de garder les méthodes traditionnelles.

Alors que choisir ? Le modèle traditionnel, meilleur prédicteur des tempêtes pures (sans phénomène climatique associé tel que la grêle ou l'orage), plus rapide et lisible mais très mauvais prédicteur des cas rares ou les réseaux de neurones, longs et complexes mais avec une erreur d'estimation constante ? Les deux méthodes ayant chacune leurs avantages et leurs inconvénients, il semble difficile de choisir entre elles. Cependant, si au lieu de les comparer, on pourrait les assembler pour profiter des qualités des uns et des autres.

Le principe est le suivant : utiliser les coefficients de régression estimés grâce au GAMLSS Gumbel et les utiliser comme poids initiaux dans le réseau de neurones. Ainsi, le réseau de neurones devient beaucoup plus rapide puisque l'initialisation des poids n'est

plus aléatoire. En effet, l'objectif du réseau de neurones est de mettre à jour les poids de façon à ce que la sortie prédite soit proche de la sortie désirée. Lors de l'initialisation aléatoire, si les poids de départ sont très éloignés des poids finaux, le temps d'apprentissage devient très long. Grâce à notre initialisation des poids à partir des coefficients de régression, les poids initiaux sont proches des poids finaux, rendant le réseau de neurones beaucoup plus rapide.

Ce gain de temps permet par ailleurs un gain de précision en modifiant les paramètres et notamment le pas d'apprentissage. Le pas d'apprentissage détermine la vitesse de modifications des poids. Plus la sortie désirée est loin de la sortie estimée, plus grand est le pas d'apprentissage, et se réduit à lorsqu'on s'en approche. Ce pas permet d'éviter de tomber dans un minimum local. Cependant, dans notre cas, l'initialisation des poids nous permet de savoir que nous sommes proches de la sortie désirée, il est donc possible de réduire ce dernier afin d'estimer au plus juste.

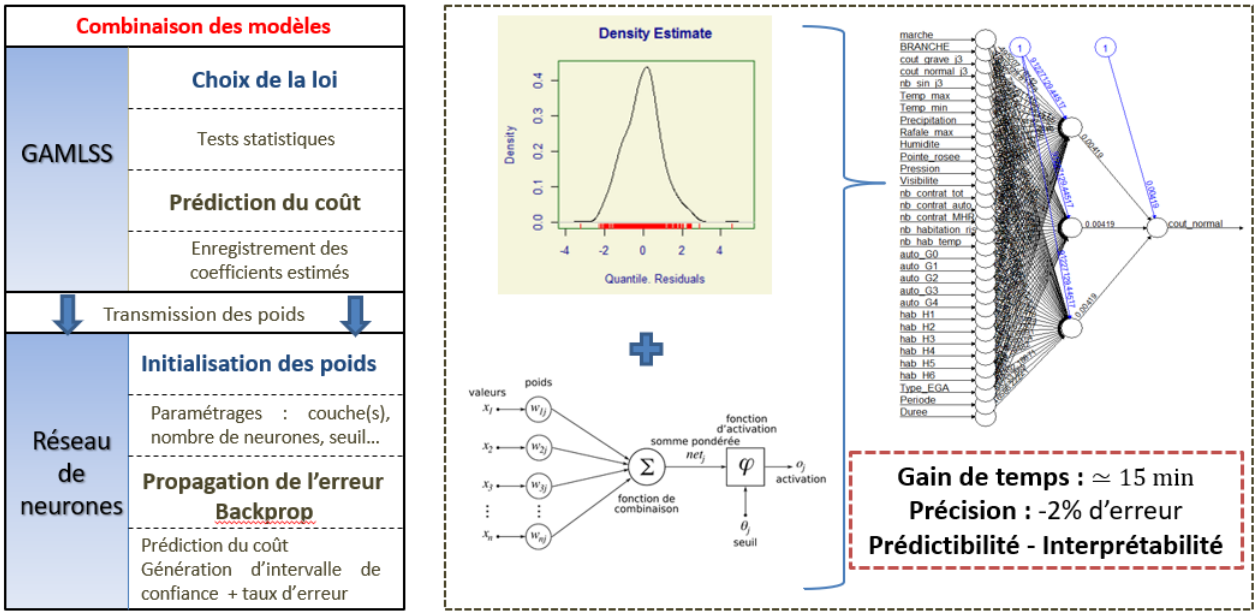


Figure 6.1.1: Regroupement du GLM et du réseau de neurones

6.1.4 Visualisation du modèle final

Notre modèle peut donc se résumer ainsi :

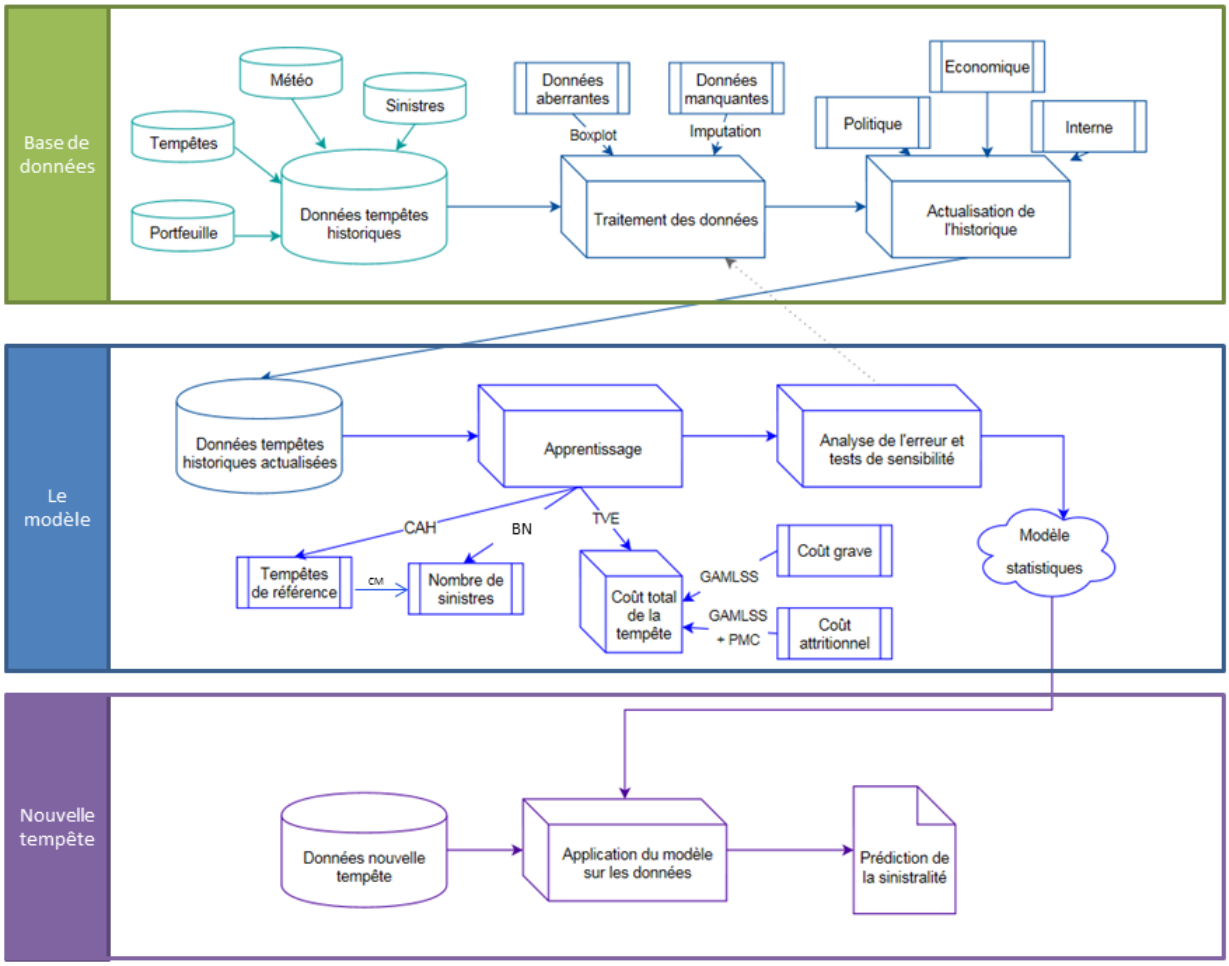


Figure 6.1.2: Schéma du modèle final

6.2 Limites et améliorations du modèle

La principale limite de notre modèle est la complexité de ce dernier. Il est composé de différents modules inter-connectés. Cette connexion rend le modèle long à l'exécution et ses différentes propositions d'estimation demande un travail d'analyse très approfondi de la part de l'actuaire en charge du modèle pour comprendre les écarts et essayer de trouver l'estimation la plus juste possible.

Pour résoudre ces problèmes, plusieurs améliorations peuvent être envisagées :

- **Ajouter des tempêtes de références**
  Retrouver les tempêtes passées est un travail fastidieux car il n'existe pas une base

de données les récapitulant toutes. Il est nécessaire de faire un travail compliqué d'exploration d'archive pour les retrouver. La mise en place d'un algorithme de textmining qui analyserait le contenu des sites internet et archives historiques pourrait permettre de reconstituer notre historique. Plus de tempêtes permettraient à nos modèles d'engranger plus d'informations et donc de mieux prédire.

- **Ajouter de l'information**
  Nous avons vu que l'une des limites de nos modèles résidait dans l'estimation des cas rares tels que les grêles ou les orages. L'ajout d'une ou de variables permettant de distinguer ce type de phénomènes pourrait améliorer la prédiction. Une réflexion sur le type de variables à créer et le moyen de le faire devra donc être menée.

Conclusion

Le but de ce mémoire était de créer un outil pour Allianz permettant l'estimation de la sinistralité du risque tempête. Le deuxième objectif était de caractériser le risque tempêtes et de tirer le plus d'information possible de ce phénomène.

Pour cela, nous avons commencé par définir ce qu'est le risque tempête et l'ensemble des phénomènes qui y sont liés. Cela nous a permis de déterminer que l'apport de variables explicatives externes telles que les données météorologiques étaient essentielles pour obtenir un modèle complet et performant. Suite à cela, nous avons travaillé sur la qualité de ces données afin qu'elles soient fiables. En effet, peu importe la performance du modèle, si les entrées de départ sont fausses, le modèle sera faussé. Ainsi, nous atteignons l'objectif d'ajouter des variables explicatives externes dont la qualité est maîtrisée grâce à une série de tests automatisés.

Une fois les données traitées, nous avons exploité trois pistes de recherche pour créer notre outil : le dire d'expert, les statistiques classiques et le machine learning. Car l'objectif de ce mémoire n'était pas uniquement de créer un modèle mais également de s'interroger sur l'essor de la datascience en actuariat et de savoir si son utilisation était pertinente dans tous les domaines. Ce travail a permis de montrer que contrairement aux idées reçues, les méthodes auto-apprenantes n'étaient pas toujours les méthodes les plus performantes et qu'au lieu d'opposer les méthodes traditionnelles et modernes, les combiner permet d'obtenir de très bons résultats, interprétable, fiable, robuste et rapide.

Pour créer cet outil quatre modèles interconnectés ont été créés. Pour une tempête venant de se produire, le premier modèle détermine ses tempêtes de référence grâce à la méthode des k plus proches voisins. Le second estime le nombre de sinistres liés à cette tempête grâce à une régression Binomiale Négative et à une méthode des cadences moyennes basée sur le premier module. Enfin les deux derniers modules permettent de déterminer le coût global d'une tempête en distinguant le coût produit par les sinistres attritionnels et celui des sinistres graves.

Les objectifs de ce mémoire ont été remplis. Cependant, certaines améliorations pourraient être apportées pour affiner les résultats et rendre le modèle plus ergonomique.

List of Figures

List of Tables

Liste des acronymes

- **MRH :** Multi Risque Habitation
- **EGA :** Événement de grande ampleur
- **KNN :** K-nearest neighbors (k plus proches voisins)
- **CAH :** Classification Ascendante Hiérarchique
- **PMC :** Perceptron MultiCouche
- **GAMLSS :** Generalized Additive Models for Location, Scale and Shape
- **GLM :** Generalized Linear Model
- **TGN :** Tempête, Grêle, Neige
- **IARD :** Incendie, Accidents et risques divers
- **IID :** Indépendant et Identiquement Distribuées
- **GEV :** Generalized Extreme Value

Annexes

A.1. Prise en charge des risques naturels en France

Types de dommages		Phénomènes naturels couverts	
		Tempête, Grêle, Neige (TGN)	Inondations, sécheresses, MVT, séismes, cyclone, ...
Personnes		Assurances de personnes (assurances des dommages corporels et/ou assurance sur la vie)	
Biens des particuliers		Garantie TGN	Extension de garantie CatNat
Biens des professionnels	Entreprises et ACPS		
	Agricoles (bâtiments)		
	Agricoles (cultures)	Assurance grêle	Multirisques climatique récoltes ou Fonds National de Garantie des Calamités Agricoles
Autres biens	Automobiles	Garantie TGN	Extension de garantie CatNat

Légende : régime assurantiel « de marché» / fonds publics / système mixte (PPP)

B.1. Textes de lois - Qualité des données

Figure 6.2.3: Extraits des textes réglementaires de niveau 2 : Règlement délégué adopté par la Commission Européenne le 17 janvier 2015

Article 19 – Définition des critères de qualité

« Les données utilisées dans le calcul des provisions techniques ne sont considérées comme :

1. Exhaustives (. . .) que lorsque les données incluent suffisamment d'informations historiques pour qu'il soit possible d'apprécier les caractéristiques des risques sous-jacents et de dégager des tendances d'évolution des risques et que des données sont disponibles pour chacun des groupes de risques homogènes utilisés dans le calcul des provisions techniques, aucune donnée pertinente n'étant exclue de ce calcul sans justification.

2. Exactes (. . .) que lorsqu'elles sont exemptes d'erreurs importantes, qu'elles sont cohérentes, même si provenant de périodes de temps différentes, et qu'elles sont enregistrées en temps utile et de manière cohérente dans la durée.

3. Appropriées (...) que lorsqu'elles sont adaptées aux fins pour lesquelles elles doivent être utilisées ; que leur volume et nature sont propres à garantir que les estimations formulées sur leur fondement pour le calcul des provisions techniques ne sont pas entachées d'une erreur d'estimation importante ; qu'elles sont cohérentes avec les hypothèses sous-tendant les techniques actuarielles et statistiques qui leur sont appliquées pour le calcul des provisions techniques ; reflètent adéquatement les risques auxquels l'entreprise d'assurance ou de réassurance est exposée au regard de ses engagements d'assurance ou de réassurance ; qu'elles ont été collectées, traitées et appliquées de manière transparente et structurée sur la base d'une procédure documentée correcte ».

Article 19 – Données externes

« L'utilisation de données provenant d'une source externe est possible à la condition d'être en mesure de démontrer que l'utilisation de ces données est plus adaptée que l'utilisation de données provenant exclusivement d'une source interne, de connaître l'origine de ces données ainsi que les hypothèses ou méthodes utilisées pour les traiter, d'identifier toutes tendances d'évolution des données externes ainsi que toutes variations, dans le temps ou entre données, des hypothèses ou méthodes utilisées pour traiter ces données et que ces hypothèses et méthodes reflètent les caractéristiques de son portefeuille d'engagements d'assurance ou de réassurance ».

Article 20 - Limites des données

« Lorsque les données ne satisfont pas aux dispositions de l'article 19, les entreprises d'assurance et de réassurance documentent de manière appropriée les limites de ces données, y compris en indiquant si et comment il y sera remédié et en précisant quelles fonctions de leur système de gouvernance seront responsables de ce processus. Les données sont enregistrées et stockées de manière appropriée avant de faire l'objet d'ajustements destinés à remédier à leurs limites »

Article 219 - « Critères relatifs aux données

1. Les données utilisées pour le calcul des paramètres propres à l'entreprise ne sont considérées comme **exhaustives, exactes et appropriées** que si elles satisfont aux critères suivants :
(a) les données répondent aux conditions énoncées à l'article 19, (...);
(b) les données peuvent être **intégrées** aux méthodes standardisées ;
(c) les données n'empêchent pas l'entreprise d'assurance ou de réassurance de se conformer

aux exigences de l'article 101, paragraphe 3, de la directive 2009/13/CE ;
(d) les données respectent toute autre exigence supplémentaire en matière de données nécessaire à l'utilisation de chaque méthode standard ;
(e) les données et le processus de leur élaboration sont **pleinement documentés**, y compris en ce qui concerne :

i) **la collecte des données et l'analyse** de leur qualité, la documentation requise comprenant un répertoire des données qui précise leur source, leurs caractéristiques et leur usage, et les caractéristiques de la collecte, du traitement et de l'application des données ;

ii) **le choix des hypothèses** utilisées pour élaborer et ajuster les données, y compris les ajustements en ce qui concerne les créances de réassurance et les sinistres catastrophiques ainsi que la répartition des dépenses, la documentation requise comprenant un répertoire de toutes les hypothèses sur lesquelles se fondent le calcul des provisions techniques et une justification du choix des hypothèses ;

iii) **le choix des méthodes actuarielles et statistiques** aux fins de l'élaboration et de l'ajustement des données, et leur application ;

iv) **la validation des données**.

2. **Lorsque des données externes sont utilisées**, elles satisfont aux critères supplémentaires suivants :
(a) **le processus de collecte des données est transparent et vérifiable par audit**, et il est connu de l'entreprise d'assurance ou de réassurance qui fonde le calcul des paramètres propres à l'entreprise sur ces données ;
(b) lorsque les données proviennent de différentes sources, les hypothèses sous-tendant la collecte, le traitement et l'application des données assurent que celles-ci sont comparables ;
(c) les données proviennent d'entreprises d'assurance ou de réassurance ayant une activité et un profil de risque similaires à ceux de l'entreprise d'assurance ou de réassurance dont les paramètres propres à l'entreprise sont calculés sur la base de ces données ;
(d) les entreprises utilisant les données externes sont en mesure de s'assurer que des éléments statistiques suffisants montrent que les distributions de probabilité qui sous-tendent leurs propres données et celles des données externes ont un degré élevé de similitude, en particulier en ce qui concerne le niveau de volatilité qu'elles impliquent ;
(e) les données externes ne comprennent que des données provenant d'entreprises présentant un profil de risque similaire, semblable également à celui de l'entreprise utilisant les données, en ce sens notamment que les données externes comprennent des données d'entreprises dont la nature de l'activité et le profil de risque, en ce qui concerne les données externes, sont similaires, et pour lesquelles il existe des éléments statistiques suffisants montrant que les distributions de probabilité sous-jacentes aux données externes

présenteront un degré élevé d'homogénéité.

Article 244 « Contenu minimum de la documentation

La documentation du modèle interne contient l'ensemble des informations suivantes :
(a) une liste de tous les documents constitutifs de cette documentation ;
(b) la politique de modification du modèle interne visée à l'article 115 de la directive 2009/138/CE ;
(c) une description des politiques, contrôles et procédures régissant la gestion du modèle interne, y compris les responsabilités attribuées à cet égard aux membres du personnel de l'entreprise d'assurance ou de réassurance ;
(d) une description de la technologie informatique utilisée dans le modèle interne, y compris de tout plan d'urgence lié à cette technologie informatique ;
(e) toutes les hypothèses pertinentes sur lesquelles le modèle interne est fondé et leur justification, conformément à l'article 230, paragraphe 2 ;
(f) l'explication, visée à l'article 230, paragraphe 2, point c), de la méthode selon laquelle ces hypothèses sont choisies, qui contient les informations suivantes :
    i) les données d'entrée sur lesquelles ce choix se fonde ;
    ii) les objectifs de ce choix et les critères utilisés pour établir son caractère approprié ;
    iii) toute limite que présente ce choix ;
(g) un répertoire des données utilisées dans le modèle interne, indiquant leur source, leurs caractéristiques et l'utilisation qui en est faite ;
(h) la spécification de la manière dont les données ont été collectées, traitées et appliquées, visée à l'article 231, paragraphe 3, point e) ;
(i) lorsque, dans le modèle interne, des données ne sont pas utilisées de manière cohérente dans la durée, une description des incohérences, assortie d'une justification ;
(j) la spécification des indicateurs qualitatifs et quantitatifs de la couverture des risques, visée à l'article 233 ;
(k) une description des techniques d'atténuation du risque prises en considération dans le modèle interne, visées à l'article 235, et une explication de la manière dont le modèle interne tient compte des risques découlant de l'utilisation de techniques d'atténuation du risque ;
(l) une description des futures décisions de gestion prises en considération dans le modèle interne, visées à l'article 236, et une description des écarts significatifs visés à l'article 236, paragraphe 2 ;
(m) les spécifications afférentes à l'attribution des profits et des pertes, visées à l'article 240, paragraphe 1 ;
(n) les spécifications afférentes au processus de validation du modèle, visées à l'article 241, paragraphe 3 ;

(o) les résultats de la validation, en termes de respect de l'article 101 de la directive 2009/138/CE ;

(p) pour les modèles et données externes :

i) le rôle joué par ces modèles et données externes dans le modèle interne ;

ii) les raisons pour lesquelles des modèles externes sont préférés à des modèles développés en interne, et des données externes à des données internes ;

iii) les alternatives à l'utilisation de modèles et données externes envisagées par l'entreprise d'assurance ou de réassurance et une explication de la décision de privilégier un modèle externe particulier ou un ensemble particulier de données externes ».

B.2. Test MCAR de Little

B.2.1. Théorie Test MCAR de Little - 1988

Le test de Little créé en 1988 est une extension multivariée du t-test. Tout comme le t-test, le test de Little évalue les différentes moyenne entre les sous-groupes de cas qui partagent le même schéma de données manquantes. La statistique de test est la suivante:

$$d^2 = \sum_{j=1}^{J} n_j \left(\hat{\mu_j} - \hat{\mu_j}^{(ML)}\right)^T \hat{\sum_j}^{-1} (\hat{\mu_j} - \hat{\mu_j}^{(ML)})$$

où :

- n_j est le nombre de données manquantes dans le pattern j
- $\hat{\mu_j}$ contient pour chaque pattern j la valeur moyenne
- $\hat{\mu_j}^{(ML)}$ contient l'estimation de la moyenne par maximum de vraisemblance
- $\hat{\sum_j}$ correspond à l'estimation du maximum de vraisemblance de la matrice de covariance

La statistique d^2 représente la somme pondérée de scores J au carré. Plus précisément, la différence $\hat{\mu_j} - \hat{\mu_j}^{(ML)}$ correspond à un score d'écart. Lorsque les données sont MCAR, les moyennes des sous-groupes doivent être comprises dans l'erreur d'échantillonnage des moyennes ML de sorte que les petites déviations soient compatibles avec un mécanisme MCAR, c'est-à-dire que le score doit être petit.

Lorsque l'hypothèse nulle est vraie, c'est-à-dire lorsque les données sont MCAR, la statistique d^2 est approximativement distribuée comme une statistique du χ^2 avec $\sum k_j - k$ degré de liberté. Où k_j est le nombre de variables complètes pour le modèle j et k le nombre total de variables.

Tout comme l'approche du t-test, le test de Little a un certain nombre de limites. Tout d'abord, il n'identifie pas les variables spécifiques qui violent l'hypothèse MCAR mais seulement si une variable est MCAR. Ensuite, ce test suppose que les modèles de données manquantes partagent une matrice de covariance communes. Ce qui pose un problème puisque les mécanismes MAR et MNAR peuvent produire des modèles de données manquantes avec des variances et des covariances différentes, phénomènes non détectés par le test. Enfin, des études (Thoemmes & Enders, 2007) suggèrent que le test de Little

n'est pas robuste lorsque le nombre de variables violant l'hypothèse MCAR est faible. Le test produit alors des erreurs de type II.

B.2.2. Illustration du Test MCAR de Little

Pour illustrer le test MCAR de Little, reprenons l'exemple proposé par Craig K. Enders [Enders2010] :

Nous travaillons avec la table suivante :

IQ	Psychological well-being	Job Performance
78	13	-
84	9	-
84	10	-
85	10	-
87	-	-
91	3	-
92	12	-
94	3	-
94	13	-
96	-	-
99	6	7
105	12	10
105	14	11
106	10	15
108	-	10
112	10	10
113	14	12
115	14	14
118	12	16
134	11	12

Cette table contient 4 patterns différents de données manquantes :

- Le cas où toutes les données sont présente (9 lignes concernées)
- Le cas où seulement le QI est présent (2 lignes concernées)
- Le cas où le QI et le score de bien-être sont présent (8 cas)
- et enfin le cas où le QI et le score de performance au travail sont présents (1 ligne concernée)

En appliquant les différentes équations vues précédemment on trouve que :

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{JP} \\ \hat{\mu}_{WB} \end{bmatrix} = \begin{bmatrix} 100 \\ 10.23 \\ 10.27 \end{bmatrix}$$

et

$$\hat{\sum} = \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP} & \hat{\sigma}_{IQ,WB} \\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP} & \hat{\sigma}_{JP,WB} \\ \hat{\sigma}_{WB,IQ} & \hat{\sigma}_{WB,JP} & \hat{\sigma}^2_{WB} \end{bmatrix} = \begin{bmatrix} 189.60 & 22.31 & 12.21 \\ 22.31 & 8.68 & 5.61 \\ 12.21 & 6.50 & 11.04 \end{bmatrix}$$

Considérons le cas où seulement les données représentant le QI sont présentes. Ce pattern a une moyenne de 91.50 donc la statistique de test est :

$$d^2 = \frac{2 * (91.50 - 100)^2}{189.60} = 0.762$$

Regardons maintenant le cas des 8 lignes où sont seulement présent le QI et le score du bien-être. Les moyennes de ces lignes sont respectivement 87.75 et 9.13 pour le QI et le bien-être. La statistique de test devient alors :

$$d^2_j = 8 * \left(\begin{bmatrix} 87.75 \\ 9.13 \end{bmatrix} - \begin{bmatrix} 100 \\ 10.27 \end{bmatrix} \right)^T \begin{bmatrix} 189.60 & 12.21 \\ 12.21 & 11.04 \end{bmatrix}^{-1} \left(\begin{bmatrix} 87.75 \\ 9.13 \end{bmatrix} - \begin{bmatrix} 100 \\ 10.27 \end{bmatrix} \right) = 6.432$$

Il reste encore la dernière statistique de test à calculer. Une fois cela fait, il suffit d'additionner les d^2_j et on trouve : $d^2 = 14.63$. En se référençant à la statistique de test du χ^2 à 5 degré de liberté, on obtient du p_value de 0.01 donc les données sont MCAR.

B.3. Imputation des données manquantes par équations chaînées

Parmi l'ensemble des méthodes d'imputation, la méthode d'imputation multiples par équations chaînées, aussi appelée FCS (Fully Conditional Specification) a été choisie pour reconstruire les données présentes dans notre jeu de données. Cette méthode, reposant sur les travaux de Royston et d'Oudshoorn, permet de générer plusieurs jeux de données où les valeurs manquantes sont complétées par plusieurs valeurs plausibles.

L'intérêt principal de cette méthode est qu'elle reflète correctement l'incertitude des valeurs manquantes tout en préservant les aspects importants des distributions ainsi que des relations entre variables.

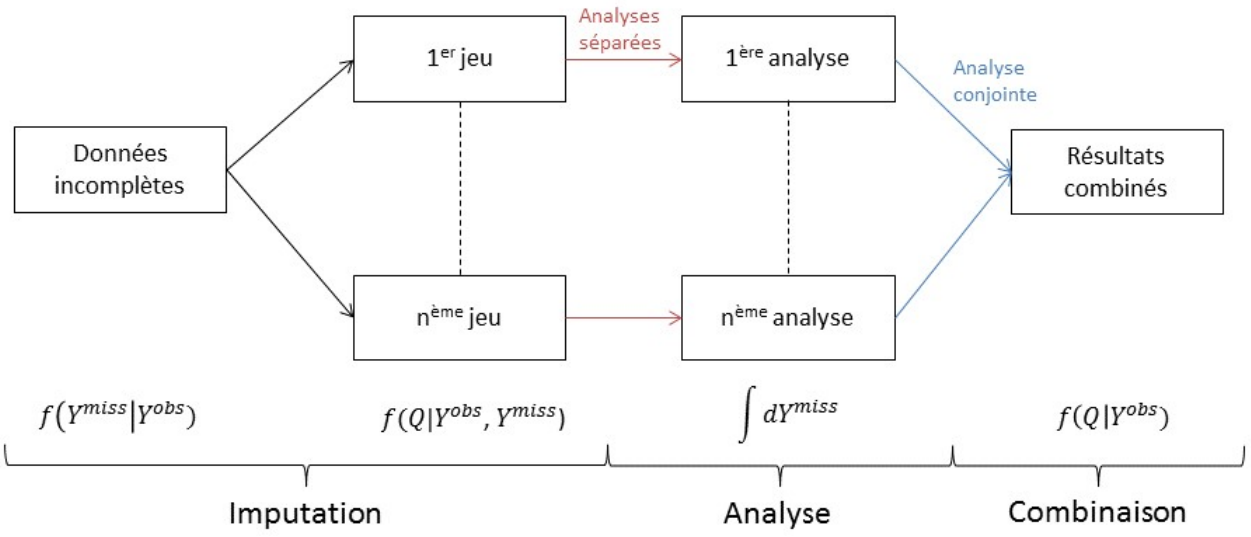


Figure 6.2.4: Schéma descriptif de la méthodologie de l'imputation multiple

L'imputation par équations chaînées se déroule en quatre étapes :

1. **Vérification des hypothèses**
   Test de l'hypothèse MCAR [1] de Little et détermination du type de données manquantes

2. **L'imputation**
   L'imputation multiple repose sur une approche bayésienne du modèle d'inférence:

[1] voir détail du test en annexe

on cherche à estimer l'espérance de la loi a posteriori d'une quantité d'intérêt Q (par exemple une moyenne, une proportion, un coefficient de corrélation), ainsi que sa variance de façon à construire un intervalle de crédibilité pour Q.

Cette étape consiste a créer m>1 jeux de données complets où chaque donnée manquante est remplacée par une valeur simulée. Plusieurs méthodes peuvent être utilisées pour le remplacement des données manquantes en fonction de leur type.

L'intérêt de ce type d'imputation est que l'on spécifie pour chaque variable la méthode appropriée. Le choix dépend de l'échelle de la variable à imputée. Dans notre cas, trois méthodes ont été utilisées :

- Méthode PMM : predictive mean matching
- Méthode CART : classification and regression tree
- Méthode RF : random forest

3. L'analyse
4. La combinaison des résultats

C.1. Modification des poids

C.1.1. Démonstration de la modification des poids

Démontrons que la règle de modification des poids est égale à

$$\Delta w_{ij} = \lambda \delta_i x_{ij}$$

avec : $\begin{cases} \forall i \in S : \delta_i = a_i(1-a_i)(d_i - a_i) \\ \forall i \in C : \delta_i = a_i(1-a_i)\sum_{k \in Succ(i)} \delta_k w_{ik} \end{cases}$

Preuve :
On sait que :

$$\Delta w_{ij} = -\lambda \frac{\partial E}{\partial w_{ij}}$$

avec $\partial E = \partial E_s(\overrightarrow{w}) = \frac{1}{2}\sum_{i \in S}(a_i - d_i)^2$

Les poids n'ayant d'impact sur la sortie du réseau qu'au travers de l'entrée totale de la cellule, on peut appliquer la règle du chaînage des dérivées partielles de la façon suivante :

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_i}\frac{\partial y_i}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} * x_{ij}$$

car : $\frac{\partial y_i}{\partial w_{ij}} = \frac{\partial \sum_{j \in Pred(i)} w_{ij} x_{ij}}{\partial w_{ij}}$

Il reste maintenant à calculer l'équation $\frac{\partial E}{\partial y_i}$. LE résultat de cette équation dépend de l'emplacement de la cellule i (couche cachée ou couche de sortie). Il est donc nécessaire de distinguer les deux cas.

Cas d'une cellule de la couche de sortie

comme les y_i n'impactent la sortie du réseau que par l'intermédiaire de la sortie obtenue a_i, il est possible d'appliquer la règle du chaînage des dérivées partielles de la façon suivantes :

$$\frac{\partial E}{\partial y_i} = \underbrace{\frac{\partial E}{\partial a_i}}_{(1)} * \underbrace{\frac{\partial a_i}{\partial y_i}}_{(2)}$$

$$(1) : \frac{\partial E}{\partial a_i} = \frac{\partial \frac{1}{2}\sum_{i \in S}(a_i - d_i)^2}{\partial a_i}$$

Tous les termes ont une dérivée nulle à l'exception du ième terme donc :

$$(1) : \frac{\partial E}{\partial a_i} = \frac{\partial \frac{1}{2}(a_i - d_i)^2}{\partial a_i} = -(a_i - d_i)$$

$$(1) : \frac{\partial a_i}{\partial y_i} = \frac{\partial \phi(y_i)}{\partial y_i} = \phi(y_i)(1 - \phi(y_i)) = a_i(1 - a_i)$$

d'où :

$$\frac{\partial E}{\partial y_i} = -(a_i - d_i)a_i(1 - a_i)$$

donc :

$$\frac{\partial E}{\partial w_{ij}} = (d_i - a_i)a_i(1 - a_i)x_{ij}$$

La modification des poids étant :

$$\Delta w_{ij} = -\lambda \frac{\partial E}{\partial w_{ij}}$$

On a donc dans le cas d'une cellule de sortie :

$$\Delta w_{ij} = \lambda \delta_i x_{ij} \text{ avec } \delta_i = a_i(1 - a_i)(d_i - a_i)$$

Cas d'une cellule de la couche cachée

Les cellules internes sont directement influencées par y_i. L'équation $\frac{\partial E}{\partial y_i}$ devient donc dans le cas d'une cellule interne :

$$\frac{\partial E}{\partial y_i} = \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} * \frac{\partial y_k}{\partial y_i} = \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} * \frac{\partial y_k}{\partial a_i} * \frac{\partial a_i}{\partial y_i}$$

On a vu précédemment dans le cas d'une cellule de la couche cachée que :

$$\frac{\partial a_i}{\partial y_i} = a_i(1 - a_i)$$

De plus :

$$\frac{\partial y_k}{\partial a_i} = \frac{\partial y_k}{\partial \phi(y_i)} = w_{ik}$$

Donc :

$$\frac{\partial E}{\partial y_i} = a_i(1 - a_i) \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} w_{ki}$$

d'où :

$$\frac{\partial E}{\partial w_{ij}} = a_i(1 - a_i) \sum_{k \in Succ(i)} \frac{\partial E}{\partial y_k} w_{ki} x_{ij}$$

La modification des poids étant :

$$\Delta w_{ij} = -\lambda \frac{\partial E}{\partial w_{ij}}$$

On a :

$$\Delta w_{ij} = \lambda \delta_i x_{ij} \text{ avec } \delta_i = a_i(1 - a_i) \sum_{k \in Succ(i)} w_{ki}$$

En conclusion, on a :

$$\Delta w_{ij} = \lambda \delta_i x_{ij}$$

avec : $\begin{cases} \forall i \in S : \delta_i = a_i(1 - a_i)(d_i - a_i) \\ \forall i \in C : \delta_i = a_i(1 - a_i) \sum_{k \in Succ(i)} \delta_k w_{ik} \end{cases}$

C.1.2. Algorithmes de modification des poids

Algorithm 2: Algorithme Rprop

Data: $\forall i,j : \Delta_{ij}(t) = \Delta_0$
$\forall i,j : \frac{\partial E}{\partial w_{ij}}(t-1) = 0$

while *la convergence n'a pas eu lieu* **do**
  Calcul du gradient $\frac{\partial E}{\partial w}(t)$
  for *chaque poids et biais* **do**
    if $(\frac{\partial E}{\partial w_{ij}}(t-1) * \frac{\partial E}{\partial w_{ij}}(t) > 0)$ **then**
      $\Delta_{i,j}(t) = minimum(\Delta_{i,j}(t-1) * \eta^+, \Delta_{max})$
      $\Delta w_{i,j}(t) = -sign(\frac{\partial E}{\partial w_{ij}}(t)) * \Delta_{i,j}(t)$
      $w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t)$
      $\frac{\partial E}{\partial w_{ij}}(t-1) = \frac{\partial E}{\partial w_{ij}}(t)$
    else if $(\frac{\partial E}{\partial w_{ij}}(t-1) * \frac{\partial E}{\partial w_{ij}}(t) < 0)$ **then**
      $\Delta_{i,j}(t) = maximum(\Delta_{i,j}(t-1) * \eta^-, \Delta_{min})$
      $\frac{\partial E}{\partial w_{ij}}(t-1) = 0$
    else
      $\Delta w_{i,j}(t) = -sign(\frac{\partial E}{\partial w_{ij}}(t)) * \Delta_{i,j}(t)$
      $w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t)$
      $\frac{\partial E}{\partial w_{ij}}(t-1) = \frac{\partial E}{\partial w_{ij}}(t)$

Algorithm 3: Algorithme Rprop+

Data: $\forall i,j : \Delta_{ij}(t) = \Delta_0$
$\forall i,j : \frac{\partial E}{\partial w_{ij}}(t-1) = 0$

while *la convergence n'a pas eu lieu* **do**
  Calcul du gradient $\frac{\partial E}{\partial w}(t)$
  for *chaque poids et biais* **do**
    if $(\frac{\partial E}{\partial w_{ij}}(t-1) * \frac{\partial E}{\partial w_{ij}}(t) > 0)$ **then**
      $\Delta_{i,j}(t) = minimum(\Delta_{i,j}(t-1) * \eta^+, \Delta_{max})$
      $\Delta w_{i,j}(t) = -sign(\frac{\partial E}{\partial w_{ij}}(t)) * \Delta_{i,j}(t)$
      $w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t)$
    else if $(\frac{\partial E}{\partial w_{ij}}(t-1) * \frac{\partial E}{\partial w_{ij}}(t) < 0)$ **then**
      $\Delta_{i,j}(t) = maximum(\Delta_{i,j}(t-1) * \eta^-, \Delta_{min})$
      $w_{i,j}(t+1) = w_{i,j}(t) - \Delta w_{i,j}(t-1)$
      $\frac{\partial E}{\partial w_{ij}}(t-1) := 0$
    else
      $\Delta w_{i,j}(t) = -sign(\frac{\partial E}{\partial w_{ij}}(t)) * \Delta_{i,j}(t)$
      $w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t)$

Algorithm 4: Algorithme Rprop+

Data: $\forall i,j : \Delta_{ij}(t) = \Delta_0$
$\forall i,j : \frac{\partial E}{\partial w_{ij}}(t-1) = 0$

while *la convergence n'a pas eu lieu* **do**
  Calcul du gradient $\frac{\partial E}{\partial w}(t)$
  for *chaque poids et biais* **do**
    if $(\frac{\partial E}{\partial w_{ij}}(t-1) * \frac{\partial E}{\partial w_{ij}}(t) > 0)$ **then**
      $\Delta_{i,j}(t) = minimum(\Delta_{i,j}(t-1) * \eta^+, \Delta_{max})$
    else if $(\frac{\partial E}{\partial w_{ij}}(t-1) * \frac{\partial E}{\partial w_{ij}}(t) < 0)$ **then**
      $\Delta_{i,j}(t) = maximum(\Delta_{i,j}(t-1) * \eta^-, \Delta_{min})$
    $w_{i,j}(t+1) = w_{i,j}(t) - sign(\frac{\partial E}{\partial w_{ij}}(t)) * \Delta_{i,j}(t)$

C.1.3. Exemple concret et détaillé d'un PMC

Voici un exemple issu de l'ouvrage de Larene Fausett [FAU94] dans lequel l'échantillon d'apprentissage vaut : $(X_1 = 0, X_2 = 1; Y = 1)$ et $\epsilon = 0,25$. Le réseau comporte une couche d'entrée composée de deux neurones, une couche cachée composée également de deux neurones et une couche de sortie comprenant un seul neurone.

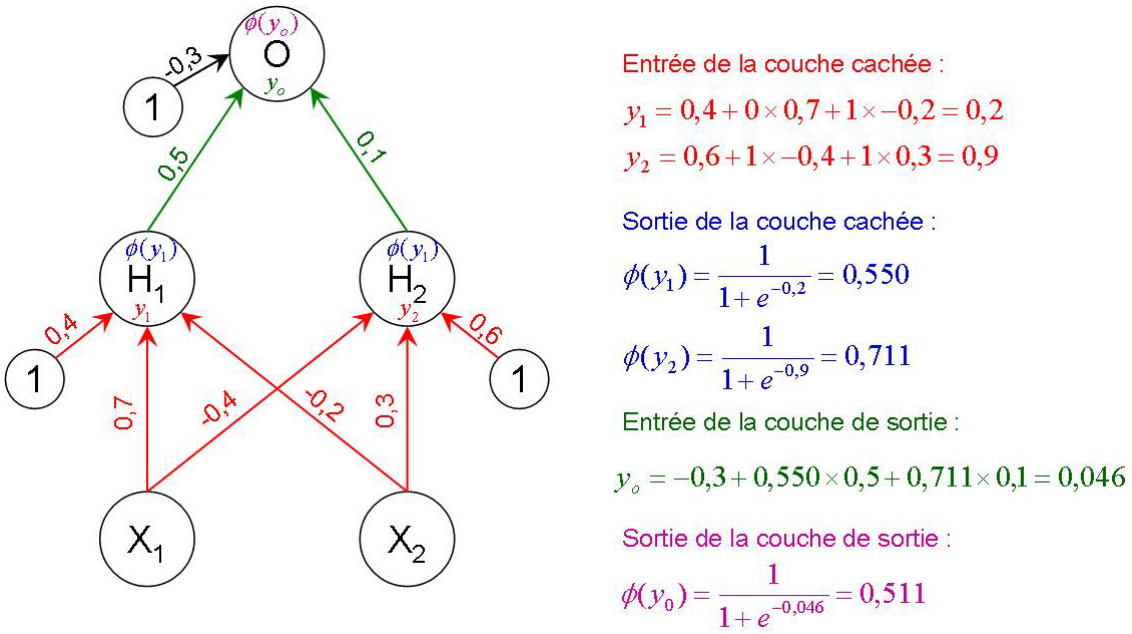
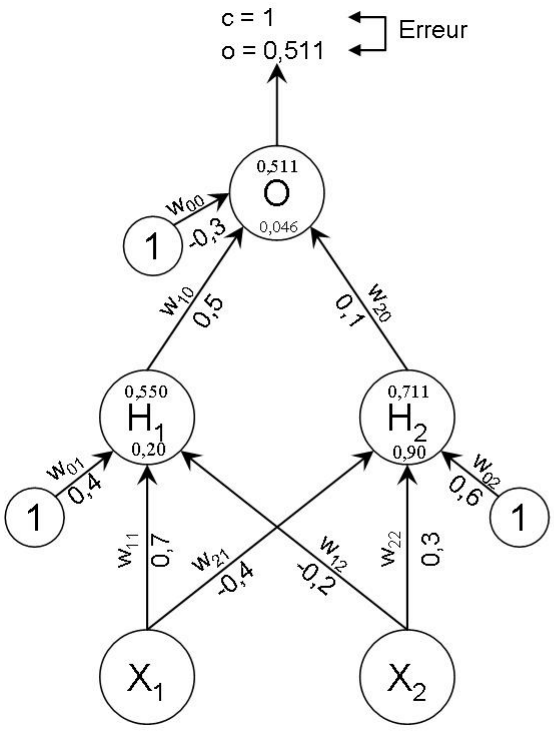


Figure 6.2.5: Propagation avant



Erreur :

$$c - o = 1 - 0{,}511 = 0{,}488$$

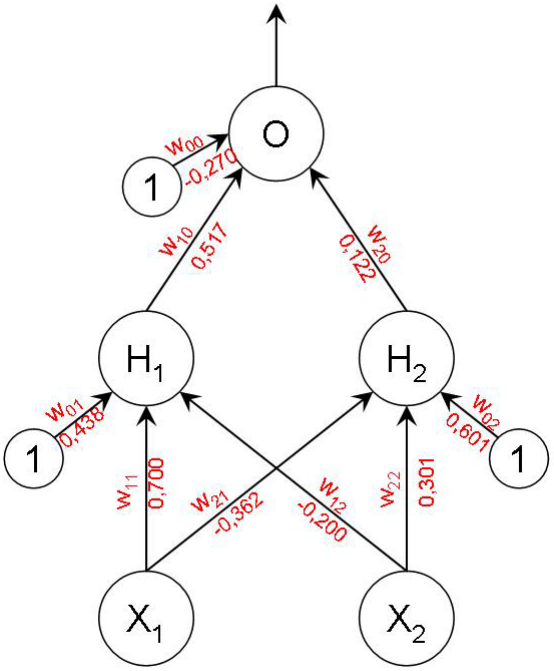
Calcul des δ :

$$\delta_k = o(1-o)(c-o)$$
$$\delta_k = 0{,}511(1-0{,}511)(1-0{,}511)$$
$$\delta_k = 0{,}122$$

$$\delta_{j_1} = o_1(1-o_1)\delta_k w_1$$
$$\delta_{j_1} = 0{,}550 \times (1-0{,}550) \times 0{,}122 \times 0{,}5$$
$$\delta_{j_1} = 0{,}015$$

$$\delta_{j_2} = o_2(1-o_2)\delta_k w_2$$
$$\delta_{j_2} = 0{,}711 \times (1-0{,}711) \times 0{,}122 \times 0{,}1$$
$$\delta_{j_2} = 0{,}0025$$

Figure 6.2.6: Calcul des delta



Calcul des Δw_{ij} :

$$\Delta w_{00} = \varepsilon\delta_k = 0{,}25 \times 0{,}122 = 0{,}0305$$
$$\Delta w_{10} = \varepsilon\delta_k o_1 = 0{,}25 \times 0{,}122 \times 0{,}550 = 0{,}0168$$
$$\Delta w_{20} = \varepsilon\delta_k o_2 = 0{,}25 \times 0{,}122 \times 0{,}711 = 0{,}0217$$

$$\Delta w_{01} = \varepsilon\delta_{j_1} = 0{,}25 \times 0{,}015 = 0{,}038$$
$$\Delta w_{11} = \varepsilon\delta_{j_1} x_1 = 0{,}25 \times 0{,}015 \times 0 = 0$$
$$\Delta w_{21} = \varepsilon\delta_{j_1} x_2 = 0{,}25 \times 0{,}015 \times 1 = 0{,}038$$

$$\Delta w_{02} = \varepsilon\delta_{j_2} = 0{,}25 \times 0{,}0025 = 0{,}0006$$
$$\Delta w_{12} = \varepsilon\delta_{j_2} x_1 = 0{,}25 \times 0{,}0025 \times 0 = 0$$
$$\Delta w_{22} = \varepsilon\delta_{j_2} x_2 = 0{,}25 \times 0{,}0025 \times 1 = 0{,}0006$$

Figure 6.2.7: Modification des poids

Bibliographie

References

[ANA05] A.D. ANASTASIADIS. “Neural Networks training and applications using biological Data”. In: (2005), p. 50.

[AOU10] JM. AOUIZERATE. “Mémoire d’actuariat : alternative neuronale en tarification santé”. In: (2010).

[BBH07] P. BORNE, M. BENREJEB, and J. HEGGEGE. “Les réseaux de neurones”. In: (2007), pp. 137–142.

[BG11] STEF VAN BUUREN and K. GROOTHUIS-OUDHOORN. “Mice : Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* (2011).

[CHA08] A. CHARPENTIER. “Insurability of Climate Risks”. In: (2008).

[CL18] P. CHABER and M. LAWRYNCZUK. “Pruning of recurrent neural models: an optimal brain damage approach”. In: (2018).

[CLR96] G. CHRYSSOLOURIS, M. LEE, and A. RAMSEY. *Confidence interval prediction for neural networks models",IEEE Trans. Neural Networks*. Vol. 7. 1996, pp. 229–232.

[COU15] F. COUIBAULT. *Les grands principes de l’assurance*. l’argus de l’assurance, 2015.

[CV17] G. CIABURRO and B. VENKATESWAREN. “Neural Networks with R”. In: (2017).

[CYB89] G. CYBENKO. “Approximation by superpositions of a sigmoidal function”. In: 2 (1989), pp. 303–314.

[DS96] J.E DENNIS and R.B SCHNABEL. “Numerical Methods for Unconstrained Optimization and nonlinear equations”. In: (1996).

[EL10] C.K. ENDERS and T.D. LITTLE. “Applied Missing Data Analysis”. In: (2010).

[EQE13] EQECAT. “La modélisation des catastrophes naturelles”. In: (2013). DOI: 18mars2013.

[FAU94] V. FAUSSET. “Fundamentals of neural networks: architectures, algorithmes and applications”. In: *Prentice Hall* (1994).

[FFA18] Fédération Française de l’Assurance FFA. “La garantie tempête grêle neige en 2016”. In: (2018).

[GAG16] C. GAGNE. “Apprentissage et reconnaissance”. In: *cours Université LAVAL* (2016).

[GAR91] G.D. GARSON. “Interpreting neural network connection weights”. In: 6 (1991), pp. 46–51.

[GF10] Frauke Günther and Stefan Fritsch. “NeuralNet : Training of Neural Network”. In: (2010).

[GIO+08] R. GIORGI et al. “The performance of multiple Imputation for Missing Covariate Data within the Contest of Repression Relative survival Analysis”. In: (2008).

[GOH95] A.T.C GOH. “Back-propagation neural networks for modeling complex systems”. In: 9 (1995), pp. 143–151.

[HSW94] K. HORNIK, M. STINCHOMBE, and H. WHITE. “Neuronal Computation”. In: 6 (1994), pp. 1262–1278.

[JAM16] S. JAMAL. “Mémoire d’actuariat- construction du taux de rachat structurel en Epargne : approximation non linéaire et agrégation de modèles”. In: (2016).

[KAL02] R. KALLEL. “Thèse : Evaluation du Boostrap pour le choix d’un modèle Neuronal”. In: (2002).

[KM95] M. KARPINSKI and A. MACINTYRE. “Polynomial bounds for VC dimension of sigmoidal neural networks”. In: (1995), pp. 200–208.

[LR02] RJA. LITLLE and D.B. RUBIN. “Statistical Analysis with Missing Data”. In: *2nd Edition, John Wiley and sons* (2002).

[MAR70] K.V. MARDIA. “Measures of multivariate skewness and kurtosis”. In: 51 (1970), pp. 519–530.

[Min13] Protection des risques naturels Ministère de l’Ecologie du Développement Durable et de l’Energie. “Les tempêtes”. In: (2013).

[MP69] M. MINSKY and S. PAPERT. *Perceptrons, Introduction*. Cambridge MA : MIT Press, 1969, pp. 1–20.

[NOC92] J. NOCEDAL. “Theory of algorithms for unconstrained optimization”. In: 1 (1992), pp. 199–242.

[OD04] J.D OLDEN and R.G DEALTH. “An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data”. In: (2004), pp. 389–397.

[OJ02] J.D. OLDEN and D.A. JACKSON. “Illuminating the "black-box" : a randomization approach for understanding variable contributions in artificial neural network”. In: (2002), pp. 135–154.

[PAR04] M. PARIZAU. “Le perceptron multicouche et son algorithme de rétropropagation des erreurs”. In: (2004).

[RAK17] R. RAKOTAMALALA. “Classifcation automatique sous R – CAH et K-Means”. In: (2017).

[RB93] RIEDMILLIER and BRAUN. “A direct adaptive method for faster backpropagation learning : the RPROP algorithm”. In: (1993), pp. 586–591.

[ROS58] F. ROSENBLATT. “The Perceptron : a probabilistic model for information storage and organization in the brain”. In: 65 (1958), pp. 386–408.

[ROY82] P. ROYSTON. “An extension of Shapiro and Willk’s Test for Normality to Large Samples”. In: 31 (1982), pp. 115–124.

[RUB87] D.B. RUBIN. “Multiple imputation for nonresponse in survey”. In: *John Wiley and Sons* (1987).

[SAP06] G. SAPORTA. *Probabilités analyse de données et statistiques*. Editions TECHNIP, 2006.

[SCA17] P. SCALART. “Introduction au Data Mining - Réseaux de neurones”. In: *Cours : Modélisation Paramétrique, Filtrage optimal et Adaptatif* (2017).

[SCH97] J.L. SCHAFER. “Analysis of incomplete multivariate data by simulation”. In: *New York : Chapman and Hall* (1997).

[SG02] J.L SCHAFER and J.W GRAHAM. “Missing Data: Our view of the state of the art. Psychological Methods”. In: (2002).

[SO98] J.L. SCHAFER and M.K. OLSEN. “Multiple Imputation for Multvariate Missing Data Problems : a Data Analysts Perspective”. In: (1998).

[Sof] NCSS Statistical Software. *Negative Binomial Regressive*. Chap. 326, pp. 326–336.

[SR07] D.M. STASINOPOULOS and R.A. RIGBY. “Generalized Additive Models for Location Scale and Shape (GAMLSS) in R”. In: (2007).

[Tan17] Tutoriels Tanagra. “Analyse de tweets sous R”. In: (2017).

[TAS04] P. TASSI. *Méthodes statistiques*. Editions Economica, 2004.

[VEA+98] R.D. DE VEAUX et al. “Prediction intervals for neural networks via nonlinear regression”. In: 40 (1998), pp. 273–282.

[WAL16] T. WALTERS. “Impute Housing Data”. In: *website* (2016).

[WOL71] P. WOLFE. “Convergence conditions for ascent methods. II: Some corrections”. In: 13 (1971), pp. 185–188.