Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Omar Jerrari

Titre : Modélisation dynamique du coût des inondations historiques en France

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

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Secrétariat

Bibliothèque :

Entreprise :

Nom : *Axa Global P&C*

Signature :

Directeur de mémoire en entreprise :

Nom : Théo Sermet

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

"Essentially, all models are wrong, but some are useful."

George E.P. Box ; Norman R. Draper (1987).
Empirical Model-Building and Response Surfaces , p 424*, Wiley.*

Note de confidentialité

Les informations, liées au groupe AXA et à ses différentes entités, présentes dans ce mémoire sont de nature hautement confidentielle.

Dès lors, leur usage et leur reproduction, sous quelque forme que ce soit, est formellement interdite sans l'accord préalable d'AXA Global P&C.

Aussi, la plupart des données numériques contenues dans le présent mémoire ont été volontairement altérées et n'ont qu'une valeur indicative. Néanmoins, ces modifications ont été réalisées de manière à ce que les valeurs restent comparables entre elles.

Remerciements

Ce mémoire n'aurait pu être réalisé sans l'aide de plusieurs personnes que je tiens à remercier.

J'ai eu la chance de réaliser ce travail au sein de l'équipe d'actuariat réassurance d'AXA Global P&C. Mes premiers remerciements vont à Théo Sermet, mon maître de stage, et à Pierre Thérond, mon tuteur pédagogique, pour leur encadrement et leurs précieux conseils.

Je souhaite aussi remercier tous les membres de l'équipe pour leurs suggestions et leurs patientes explications ainsi que leur accueil chaleureux et leur sympathie tout au long de cette étude.

Note de synthèse

Contexte et problématique

Les catastrophes naturelles représentent un risque majeur pour les compagnies d'assurance car elles peuvent remettre en question leur solvabilité par l'importance des pertes économiques qu'elles génèrent. Afin de mieux connaître leur exposition et de choisir de manière optimale les fonds propres à détenir et les structures de réassurance à utiliser, les assureurs éprouvent le besoin de développer leur propre appréciation des risques naturels à travers des modèles spécifiques qui modélisent l'évènement physique lui-même avant de quantifier les pertes qui en découleraient. Les sorties de ces modèles permettent une quantification du risque en simulant des évènements fictifs donnant lieu, après application à l'exposition courante (c'est-à-dire l'ensemble des sites assurés par la compagnie), à une estimation de la distribution stochastique annuelle des pertes.

La construction de tels modèles requiert des expertises physiques pointues et leur validation par le régulateur peut alors être compliquée au vu de leur opacité sous-jacente. De plus, l'existence du régime CAT NAT en France et de la protection financière de l'État à travers la Caisse Centrale de Réassurance contre certains risques naturels ne favorisent pas le développement de tels modèles sur le marché contrairement à d'autres pays.

C'est dans ce contexte que l'on s'intéresse dans ce mémoire à construire un modèle intermédiaire entre une vision purement historique et une vision purement stochastique du risque d'inondation en France. En effet, on se donne pour ambition de modéliser de manière dynamique ce risque à partir d'estimations du coût des évènements historiques subis par AXA. Le caractère dynamique du modèle signifie que celui-ci permettra de produire des estimations qui varieront au cours du temps et qui dépendront de la distribution géographique des sites assurés par AXA et de l'aménagement urbain au moyen de protections physiques contre les inondations mises en œuvre par les villes concernées.

Enfin, le développement d'une vision fine et actualisée du coût de chaque évènement historique, dont chacun peut se rappeler de l'ampleur, permettra de créer une métrique intuitive dans le cadre de la gestion des risques et de la politique de souscription.

Démarche de résolution

Cadre de l'étude

La méthode utilisée jusqu'à présent consistait à appliquer l'inflation et la croissance du portefeuille sur les coûts historiques. Notre modèle apportera une vision plus fine et tiendra compte de l'évolution géographique du portefeuille d'AXA.

Pour construire un tel modèle, nous analysons 26 évènements d'inondation ayant eu lieu entre 1999 et 2012 (au sens d'une catastrophe naturelle) et leur impact sur le portefeuille d'une entité du groupe AXA.

En termes de données, nous disposons de :

- la base des sinistres,
- la base des arrêtés CAT NAT, qui recense l'ensemble des communes où l'état de catastrophe naturelle a été décrété suite à une inondation et la date de l'arrêté correspondant,
- le portefeuille de l'année 2016 sur lequel on fera nos estimations,

- les portefeuilles historiques de 2003 à 2012 qui nous permettront de calibrer nos estimations des évènements de 1999 à 2012,
- les hauteurs d'eau du réseau fluvial national à différentes stations de jaugeage (relevés en temps réel gérés par un institut de recherche publique),
- les précipitations historiques,
- un modèle numérique de terrain d'une résolution horizontale de 75 m et verticale de 1 m ainsi qu'une cartographie du réseau fluvial français.

Articulation du mémoire

Après avoir présenté les données à disposition et fixé les périmètres et hypothèses de l'étude, nous abordons quelques généralités sur les risques assurantiels liés aux catastrophes naturelles et les spécificités liées à leur modélisation en nous concentrant sur l'inondation. Dans la suite, nous explicitons l'algorithme derrière le modèle physique de propagation de l'eau utilisé afin de construire les empreintes historiques permettant de reproduire le plus fidèlement possible les évènements passés. Nous présentons par la suite les différents outils d'apprentissage (*GLMs*, *random forest* et *gradient boosting*) et les mesures d'erreur utilisées avant de les calibrer sur la sinistralité historique et ainsi produire les estimations du coût de chaque évènement sur le portefeuille 2016. Dans un dernier temps, nous construisons un modèle fréquence-coût à partir des pertes estimées et, s'agissant d'un risque à occurrence peu fréquente et dont l'historique n'est donc pas suffisamment représentatif des pertes futures, nous le crédibilisons à l'aide de scénarios généraux calibrés par des experts afin de produire le modèle final pour le risque d'inondation. Ce modèle est enfin utilisé à des fins d'optimisation de réassurance au vu du capital économique requis par la réglementation et de l'appétit au risque propre à l'assureur.

Choix de la modélisation effectuée

Les différentes étapes de la modélisation sont visibles sous forme de schéma sur la figure 7.

L'étape clé de notre étude consiste à reproduire les évènements passés. En effet, il n'existe pas de données fiables relevant les hauteurs d'eau dans les zones sinistrées pendant la durée des inondations. Nous construirons alors ce que l'on appelle des empreintes en appliquant un algorithme de propagation de l'eau à l'aide des hauteurs d'eau historiques le long du réseau fluvial et de la cartographie du territoire. Ces empreintes permettent de décrire l'ensemble des zones inondées et la hauteur d'eau associée à chacune d'elles comme on peut le voir sur la figure 8 avec une empreinte modélisée de la crue de la Seine de 1910. Les empreintes obtenues sont enfin fiabilisées à l'aide des informations à notre disposition sur les arrêtés CAT NAT et la sinistralité historique. Nous construisons également à partir du modèle numérique de terrain des cartes de ruissellement qui nous permettront de prendre en compte l'épuisement de la capacité du sol à absorber l'eau.

Forts de ces variables physiques, nous utilisons les données restantes pour calibrer les différents modèles d'apprentissage envisagés. Nous sélectionnons le modèle ayant la meilleure statistique de déviance et un sur-apprentissage limité que l'on teste en réalisant des validations croisées et qui permet, enfin, de prédire des pertes agrégées par évènement très proches des pertes observées.

Nous estimons ensuite les pertes sur le portefeuille actuel avec le modèle sélectionné et calibrons avec celles-ci un modèle fréquence-coût. Les lois de fréquence envisagées sont les lois de comptage classiques de Poisson et binomiale négative. Nous avons retenu pour les lois de sévérité des distributions à support positif et qui sont utilisées pour modéliser des valeurs extrêmes et sont donc particulièrement adaptées à notre situation où l'on est face à des pertes rares et

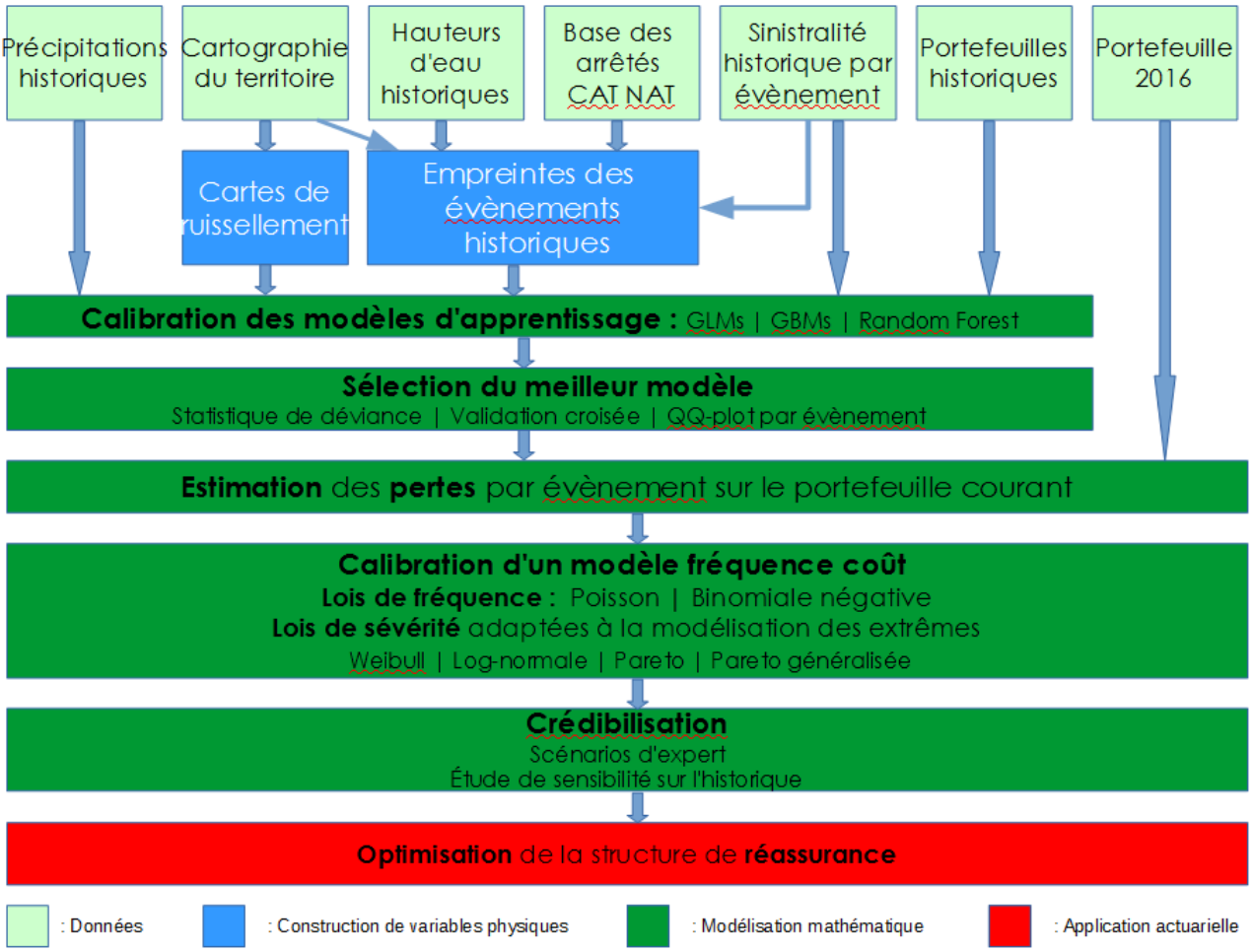

FIGURE 1 – Schéma synthétisant les différentes étapes de modélisation effectuées.

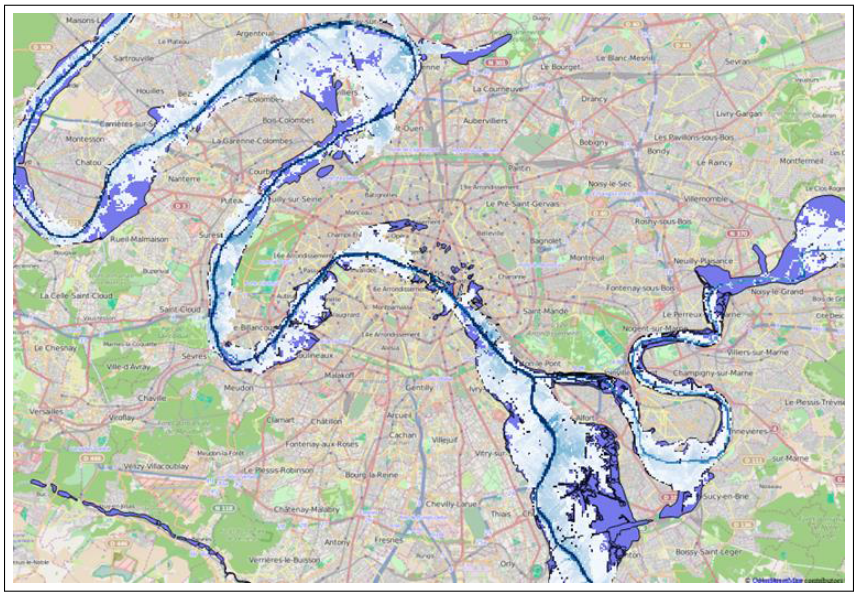

FIGURE 2 – Empreinte modélisée de la crue de la seine de 1910

très sévères. Ainsi, nous sélectionnerons une loi parmi les lois de Weibull, de Pareto, Pareto généralisée et log-normale à l'aide de critères d'ajustement tels que la distance de Kolmogorov-Smirnov.

Les points utilisés pour calibrer le modèle sont issus d'observations sur une période de 14 ans, ce qui est relativement court lorsque l'on cherche à modéliser des évènements peu fréquents tels que des inondations. Nous allons alors crédibiliser le modèle obtenu avec des scénarios calibrés par des experts sur les conséquences financières d'une crue de la Seine équivalente à celle de 1910. Une étude de sensibilité sur l'historique nous permettra de calibrer cette crédibilisation et de construire le modèle final.

Enfin, nous utilisons le modèle produit afin de sélectionner la structure de réassurance optimale en termes de fonds propres requis par la réglementation Solvabilité 2 en introduisant des notions telles que la création de valeur et l'appétit au risque.

Résultats

Différentes mesures d'erreurs et critères de sélection de modèle ont mené à estimer le coût d'un sinistre par une forêt aléatoire et la probabilité de sinistre par un modèle *GBM*. L'importance des variables explicatives dans ces deux modèles ont montré une certaine robustesse des empreintes développées et la spécificité de destruction propre à chaque évènement, confirmant nos motivations dans la réalisation de cette étude. L'étage du site assuré apparaissant comme variable assurantielle la plus importante dans l'estimation de la probabilité de sinistre est conforme à notre intuition vis-à-vis du risque modélisé. Enfin, la variable hauteur d'eau s'est avérée peu pertinente et a été retiré du modèle, remettant en question la qualité du module de propagation. Cette statistique est représentée graphiquement dans les figures 9 et 10.

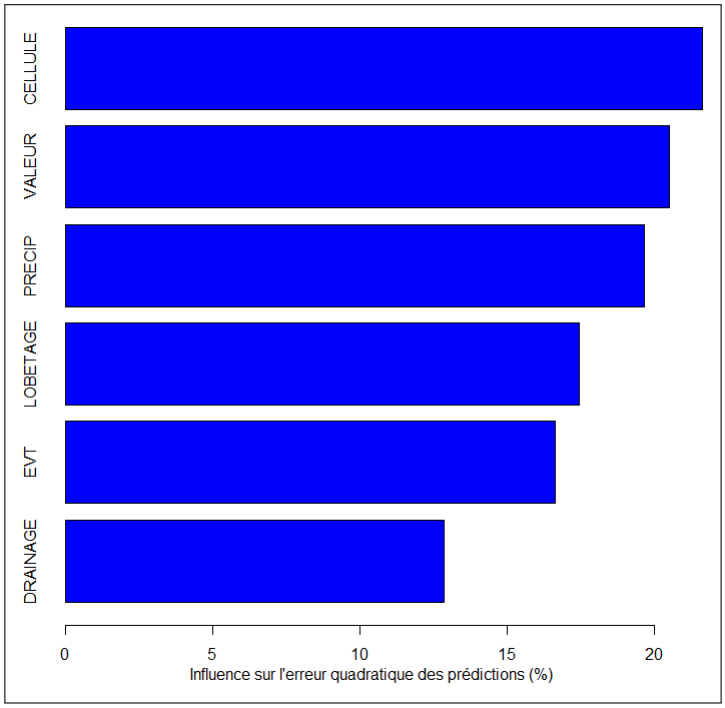

FIGURE 3 – Importance de chaque variable explicative dans le modèle d'estimation du coût de sinistre (CELLULE : distance à la cellule inondée modélisée la plus proche, EVT : nom de l'évènement, PRECIP : quantités de précipitations pendant la durée de l'évènement, DRAINAGE : variable de ruissellement, LOBETAGE : branche d'activité couplée à la variable d'étage, VALEUR : valeur assurée du bien).

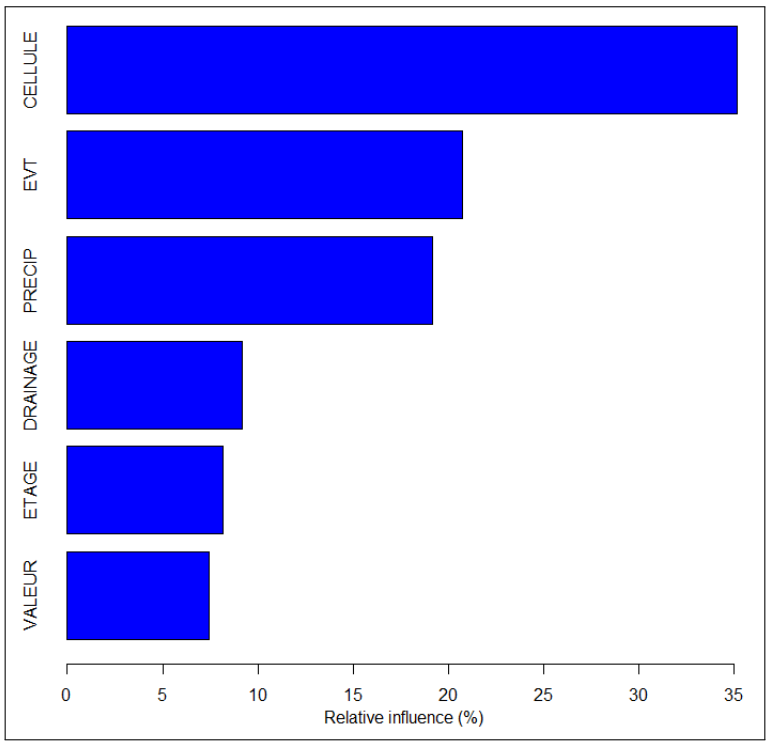

FIGURE 4 – Importance de chaque variable explicative dans le modèle d'estimation de la probabilité de sinistre (CELLULE : distance à la cellule inondée modélisée la plus proche, EVT : nom de l'évènement, PRECIP : quantités de précipitations pendant la durée de l'évènement, DRAINAGE : variable de ruissellement, LOBETAGE : branche d'activité couplée à la variable d'étage, VALEUR : valeur assurée du bien).

Le QQ-plot du modèle sur les données utilisées visible sur la figure 11 permet de juger de sa qualité de prédiction. On y voit que les valeurs prédites suivent la même tendance que les valeurs observées avec des écarts contenus sauf pour un point extrême surestimé par le modèle de 35%.

Enfin, la figure 12 permet de comparer les estimations faites sur le portefeuilles 2016 avec le modèle développé et la méthode d'inflation par l'indice des prix de construction et l'évolution globale de la somme assurée du portefeuille. Les évolutions restent contenues et s'étalent de -22% à $+14\%$ et l'évolution moyenne est de -4%. Il n'est pas surprenant d'avoir une évolution globale à la baisse car l'assureur adapte régulièrement sa politique de souscription à la suite d'évènements d'inondation.

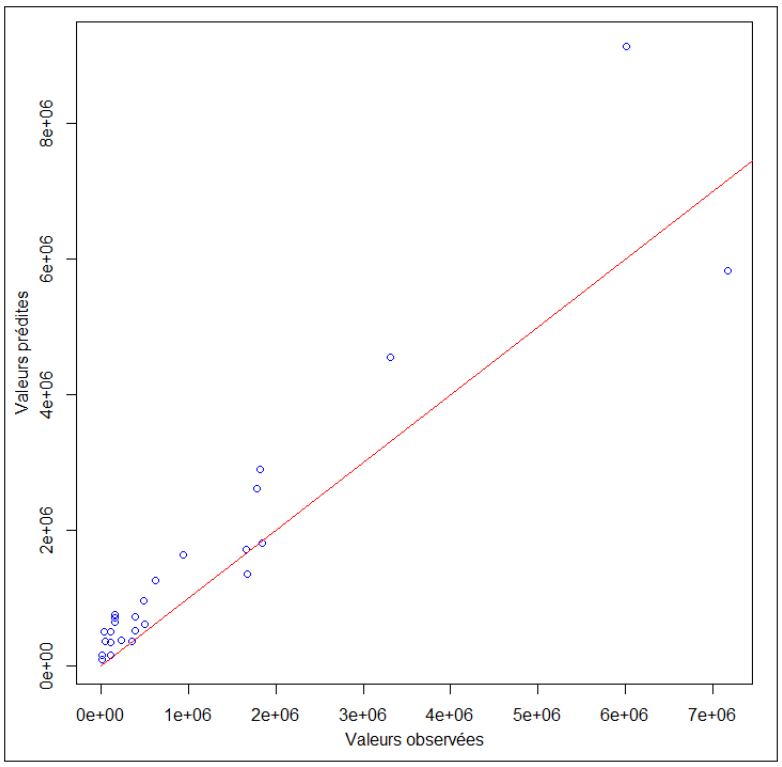

FIGURE 5 – QQ-plot par évènement du modèle.

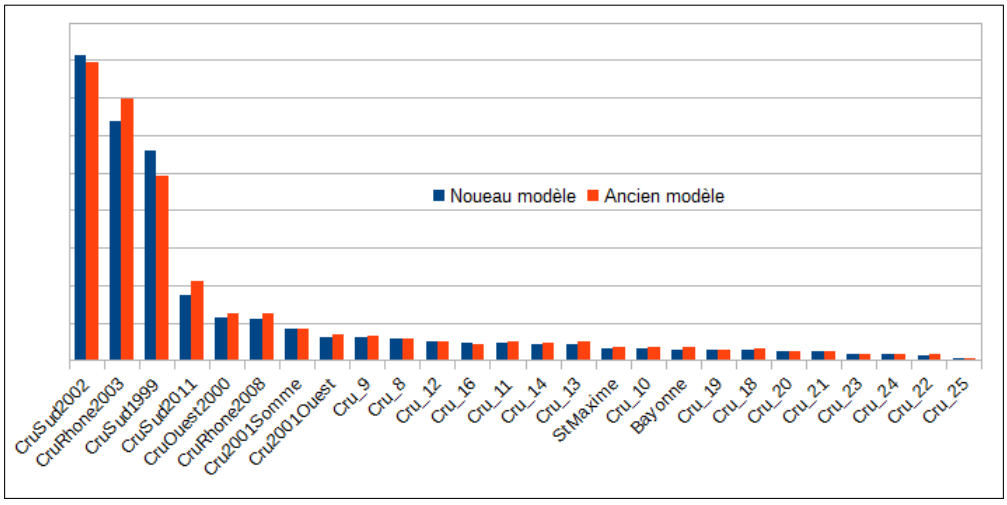

FIGURE 6 – Coût estimé pour chaque évènement, comparaison avec la méthode actuelle.

Conclusion

Cette étude a permis de développer une vision fine et actualisée de l'impact financier des évènements historiques et qui peut être généralisée à d'autres portefeuilles et prendre en compte les évolutions du territoire en terme d'aménagement urbain et de protection contre les inondations.

Elle a également permis de construire des empreintes historiques pouvant être utilisées en souscription et en gestion des risques de manière intuitive en faisant référence à un évènement passé et vécu. La robustesse de ces empreintes, montrée par les modèles d'apprentissage, renforce la confiance que l'on peut y accorder.

Ce travail a également permis de constater l'apport potentiel du *machine learning* en actuariat par rapport aux méthodes plus classiques telles que les *GLMs*. Nous avons également montré que son aspect "boîte noire" peut être nuancé grâce à des statistiques telles que l'importance des variables explicatives.

En outre, nous avons montré comment le modèle peut être utilisé afin de piloter le risque d'inondation porté par l'assureur à travers la construction d'un modèle fréquence coût et l'optimisation de la structure de réassurance en maximisant la création de valeur espérée au vu de l'impact sur le capital économique tout en répondant aux exigences d'appétit au risque de la compagnie. Le modèle développé a alors permis une estimation plus fine du capital économique et de l'efficience de la réassurance.

Enfin, nous retenons deux pistes d'amélioration importantes du modèle. Le module de propagation de l'eau peut être amélioré afin de rendre pertinente la variable hauteur d'eau et améliorer potentiellement la capacité prédictive du modèle. La deuxième amélioration possible est dans la prise en compte des données manquantes qui peut être revue en essayant différentes techniques disponibles dans la littérature.

Executive summary

Context and objective

Natural catastrophes represent a major risk for insurance companies because they can question their solvency due to the huge amount of economic losses that can be caused. In order to better understand their exposure and to optimally choose the economic capital that must be held and the reinsurance structures to use, insurers need to develop their own assessment of natural risks through specific models that model the physical event before quantifying the resulting losses. The outputs of these models allow a quantification of the risk by simulating fictitious events giving rise to an estimate of the annual losses distribution which depend on the current exposure (that is to say all the sites insured by the company).

The construction of such models requires sharp physical exertise and need to be approved by the regulator. Moreover, the existence of the NAT CAT regime in France and the financial protection of the State through the Caisse Centrale de Réassurance against certain natural risks do not favor the development of such models on the market unlike other countries.

This is why we are interested in this memoir to build an intermediate model between a purely historical vision and a purely stochastic vision of flood risk in France. Indeed, our ambition is develop a dynamic model of this risk on the basis of estimates of the cost of historical events suffered by AXA. The dynamic nature of the model means that it will produce estimates that will vary over time and which will depend on the geographic distribution of AXA insured sites and urban development through physical protection against flooding by involved cities.

Finally, the development of a finer and up-to-date vision of the cost of each historical event, which everyone can recall the magnitude, will provide an intuitive metric within the framework of risk management and underwriting policy.

Solving approach

Study framework

The method cuurently in place was to apply inflation and portfolio growth to historical costs. Our model will provide a finer view and take into account the geographic evolution of AXA's portfolio.

To build such a model, we analyze 26 flood events that occurred between 1999 and 2012 (in the sense of a natural disaster) and their impact on the portfolio of an entity in the AXA group.

To do so, we have the following databases at our disposal :

- claims databse,
- NAT CAT decrees databse, which lists all the municipalities where the state of natural disaster was decreed following a flood event and the date of the corresponding decree,
- 2016 portfolio that will be used to perform final estimations,
- the historical portfolios from 2003 to 2012 that will allow us to calibrate our estimates of events from 1999 to 2012,
- the water levels of the national river system at different gauging stations (real-time readings managed by a public research institute),

- historical rainfall,
- a digital terrain model with a horizontal resolution of 75 m and vertical resolution of 1 m as well as a map of the French river network.

Paper outline

After presenting the available data and defining the scope and hypotheses of the study, we will discuss some general aspects of insurance risks related to natural disasters, with a focus on flood peril, and the specificities associated with their modeling. Further, we will explain the algorithm behind the physical model of water propagation used in order to construct the historical fingerprints making it possible to reproduce as closely as possible the historical events. We then present the different fitting tools (GLMs, random forests and gradient boosting) and the error measures used before calibrating them on the historical claims and thus producing the estimated cost of each event on the 2016 portfolio. Finally, we construct a frequency-cost model based on the estimated losses and, as it is a low frequency risk whose history is not sufficiently representative of future losses, we will use general scenarios calibrated by experts to produce the final model. This model will be finally used for optimal reinsurance purposes under economic capital and risk appetite constraints.

Modeling assumptions

The different steps of the modeling are visible in the form of a diagram in the figure 7.

The key step in our study is to replicate past events. Indeed, there are no reliable data on the water levels in the areas affected during the flood events. We will then construct events fingerprints by applying a water propagation algorithm using historical water levels along the river system and using land mapping. These fingerprints make it possible to describe all the flooded zones and the height of water associated with each of them, we can see in figure 8 a modeled fingerprint of the flood of the Seine river in 1910. Fingerprints obtained are finally reliabilized using the available information on NAT CAT decrees and the historical losses. We also construct model runoff maps that will allow us to take into account the depletion of the soil's ability to absorb water.

Using these physical variables, we use the remaining data to calibrate the various models. We select the model with the best deviance statistics and a limited over-fitting which is tested by performing cross-validation tests and which aggregated losses per event predictions are very close to the observed losses.

We then estimate the losses on the current portfolio with the selected model and calibrate a frequency-cost model. The frequency laws taken into consideration are the traditional counting laws : Poisson and negative binomial. Regarding severity, we have chosen distributions with positive support and which are used to model extreme values and are therefore particularly adapted to our situation where one is faced with rare and very severe losses. Thus, we will select a law from Weibull, Pareto, Pareto generalized and log-normal distributions using adjustment criteria such as Kolmogorov-Smirnov distance.

The data points used to calibrate the model are derived from observations over a period of 14 years, which is relatively short when attempting to model rare events such as floods. We will then credibilize the model obtained with scenarios calibrated by experts on the potential financial consequences of a flood of the Seine river close to the 1910 one. A study of sensitivity on the history will allow us to calibrate this credibility and build the model final.

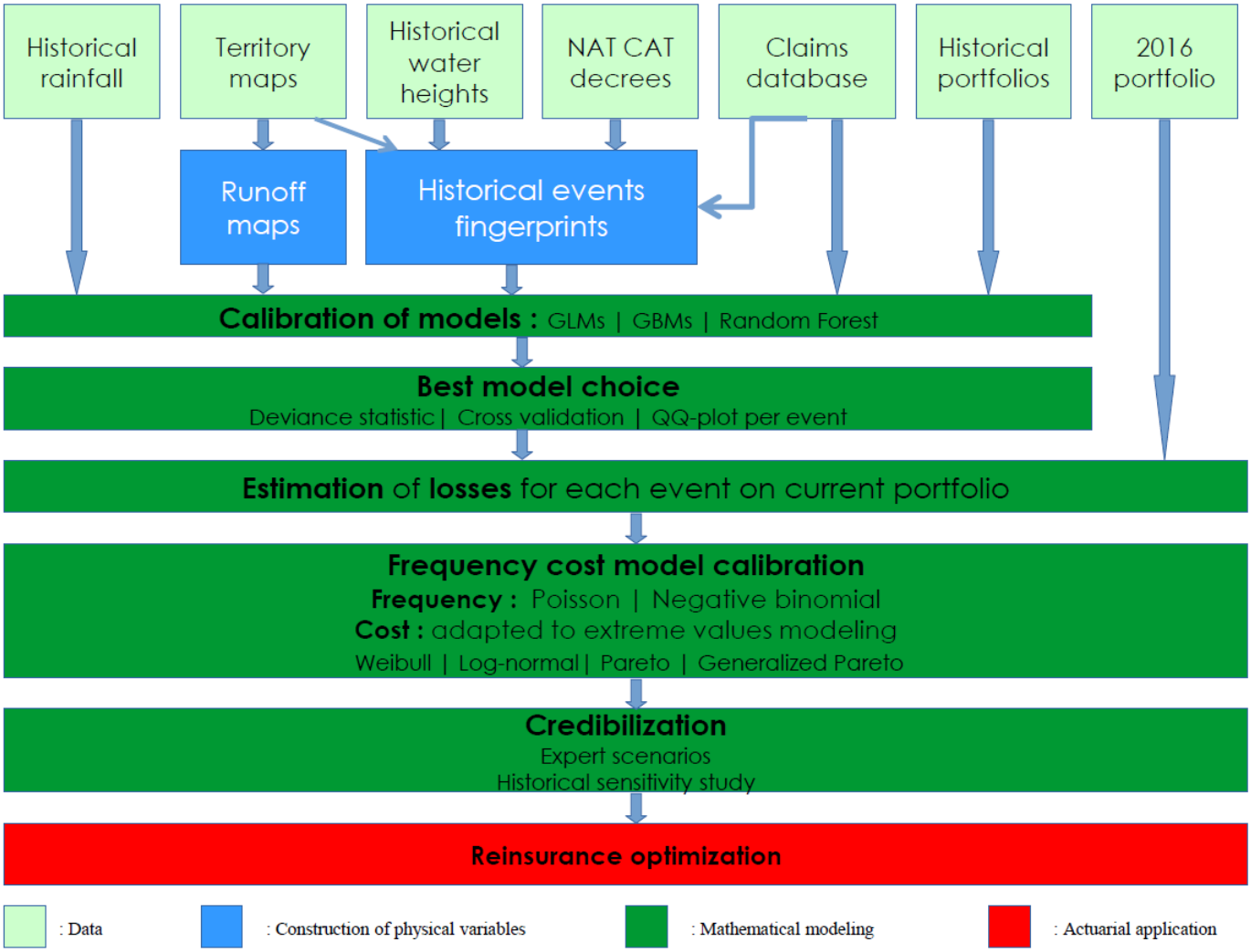

FIGURE 7 – Diagram synthesizing the different steps of modeling carried out.

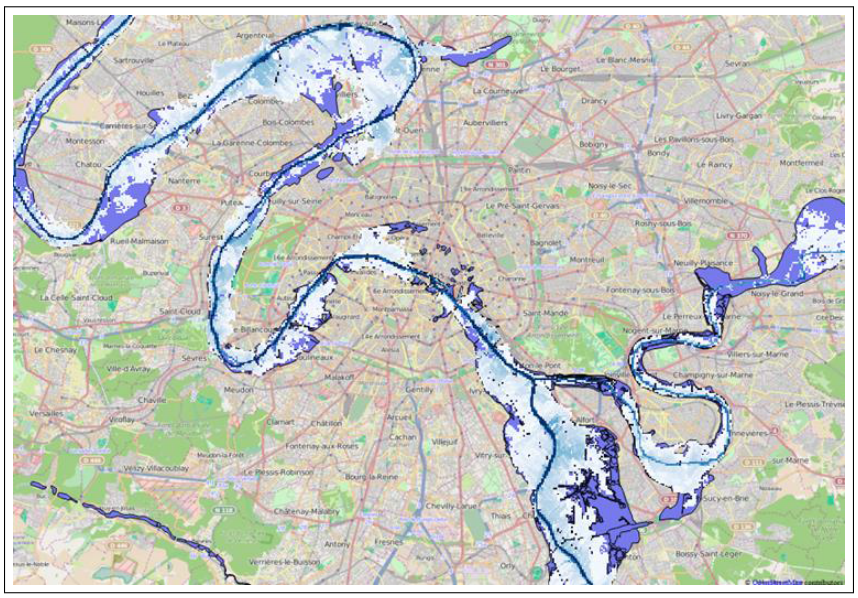

FIGURE 8 – Empreinte modélisée de la crue de la seine de 1910

Finally, we use the developed model to select the optimal reinsurance structure under constraints such as the economic capital required by the Solvency II regulation. To do so, we will introduce concepts such as value creation and risk appetite.

Results

Various error measurements and model selection criteria have led to estimate the cost of a disaster by a random forest model and the probability of a claim to occur by a GBM model. The importance of the explanatory variables in these two models showed a certain robustness of the built fingerprints and the specific kind of destruction of each studied event. These results confirm our original motivation in this study. The floor level of the insured site appears as the most important policy variable in the estimation of the probability of loss, this is in line with our intuition with respect to flood risk. Finally, the water height variable proved to be irrelevant and was removed from the model, questionning the quality of the propagation module. This statistic is represented graphically in the figures 9 and 10.

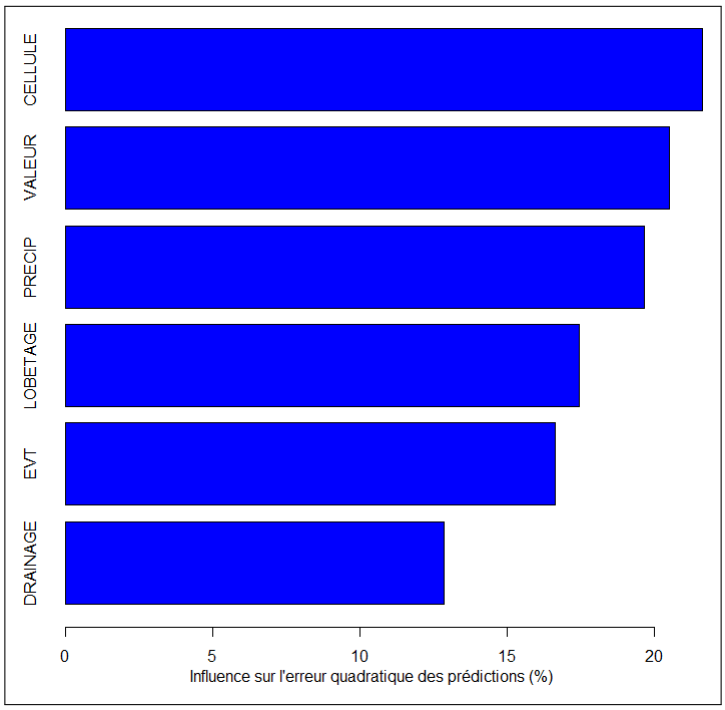

FIGURE 9 – Importance of each explanatory variable in cost estimation model (CELLULE : distance to the nearest modeled flooded cell, EVT : event namet, PRECIP : rainfall quantities during the event, DRAINAGE : runoff variable, LOBETAGE : line of business coupled with floor level, VALEUR : insured value).

The figure 11 shows the QQ-plot of the final model, it can be used to judge the prediction quality. It shows that the predicted values follow the same trend as the observed ones with limited deviations except for one extreme point overestimated by the model by 35%

Finally, figure 12 compares the 2016 estimates with the developed model and the current method which consists in inflating losses with construction price indices and the evolution of the total sum insured. Differences remain limited and range from -22% to $+14\%$ and the average gap is -4%. It is not surprising to see a global downward trend as the insurer regularly adapts its underwriting policy following flood events.

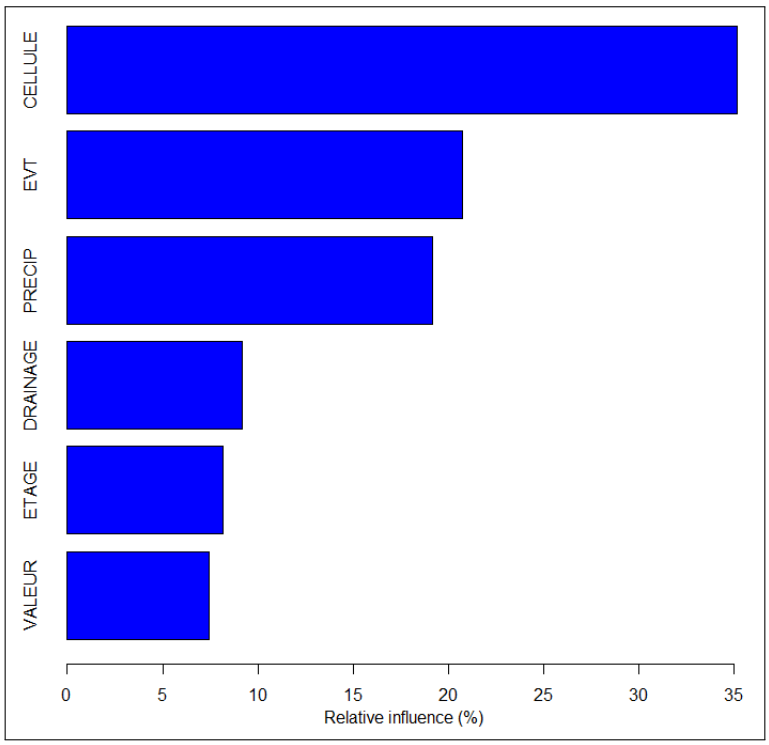

FIGURE 10 – Importance of each explanatory variable in claim occurrence probability estimation model (CELLULE : distance to the nearest modeled flooded cell, EVT : event namet, PRECIP : rainfall quantities during the event, DRAINAGE : runoff variable, LOBETAGE : line of business coupled with floor level, VALEUR : insured value).

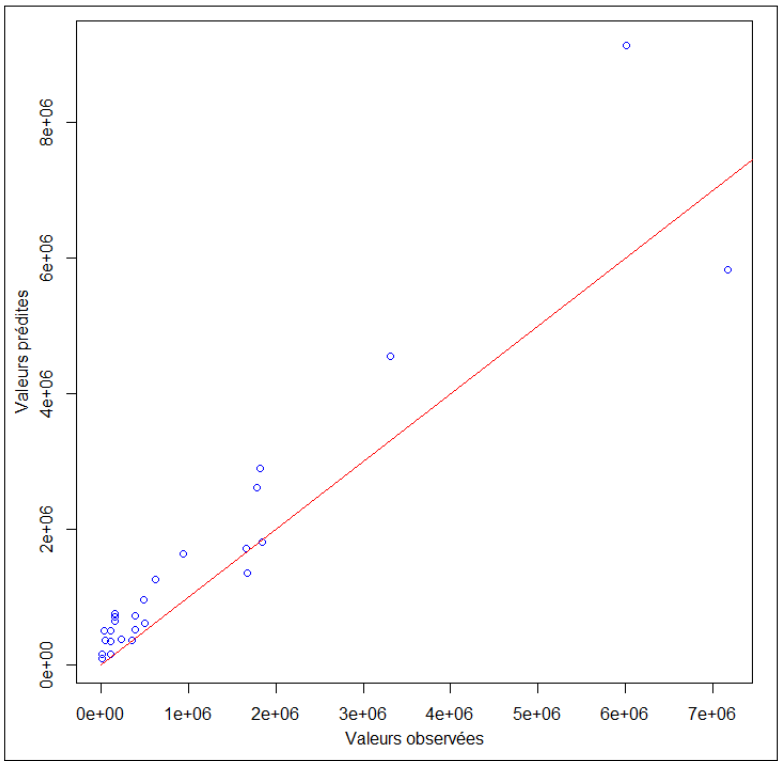

FIGURE 11 – QQ-plot per event of the final model.

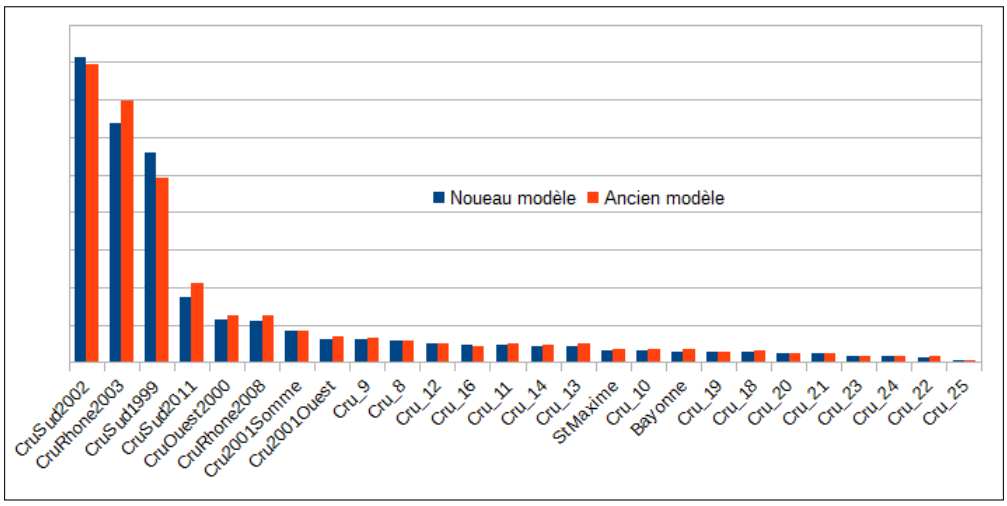

FIGURE 12 – Estimated cost for each event, compared to current method.

Conclusion

This study allowed to develop a finer and updated vision of the financial impact of historical events, the model can be generalized to other portfolios and take into account the evolutions of the territory in terms of urban planning and protection against floods.

It has also made it possible to build historical fingerprints that can be used in underwriting and risk management in a very intuitive manner by referring to a past and lived event. The robustness of these fingerprints, as demonstrated by the fitting models, reinforces the confidence that can be placed in them.

This work also revealed the potential contribution of machine learning in actuarial science compared to more classical methods such as GLMs. We have also shown that its " black box " aspect can be subdued by statistics such as the importance of explanatory variables.

In addition, we have shown how the model can be used to control the flood risk carried by an insurer through the construction of a frequency-cost model and the optimization of the reinsurance structure by maximizing the expected value creation with respect to the impact on economic capital while meeting the company's risk appetite requirements. The developed model allowed a more refined estimate of the economic capital and the efficiency of the reinsurance.

Finally, the model can be improved in two major ways. The water propagation module can be improved in order to make the water height variable relevant and potentially improve the predictive capacity of the model. The second possible improvement is to work on how the missing data can be taken into account, the methodology we use can be reviewed by trying different techniques available in the literature.

Table des matières

Introduction . . . 21
1 Descriptif des données à disposition . . . 22
1.1 Base sinistres . . . 22
1.2 Base des arrêtés CAT NAT . . . 24
1.3 Portefeuille d'AXA France IARD 2016 . . . 24
1.4 Portefeuilles de 2003 à 2012 . . . 24
1.5 Hauteurs d'eau aux stations de jaugeage . . . 25
1.6 Précipitations historiques . . . 26
1.7 Cartographie du territoire . . . 26
1.8 Périmètre de l'étude et hypothèses . . . 27
2 Catastrophes naturelles et leur modélisation . . . 28
2.1 Spécificités du risque de catastrophes naturelles . . . 28
2.1.1 Définition . . . 28
2.1.2 Régime d'assurance des catastrophes naturelles en France et assurabilité du risque d'inondation . . . 29
2.1.3 Spécificités de la modélisation . . . 31
2.2 Construction d'un modèle CAT . . . 32
2.3 Description du risque d'inondation . . . 32
3 Construction des empreintes historiques . . . 34
3.1 Module physique de propagation . . . 34
3.2 Affinement des empreintes obtenues . . . 36
3.3 Prise en compte du ruissellement . . . 38
4 Modèles Linéaires Généralisés . . . 39
4.1 Définition . . . 39
4.2 Famille exponentielle revisitée . . . 39
4.3 Modèles linéaire généralisés . . . 39
4.4 Estimation des coefficients de régression . . . 40
4.5 Choix de la fonction de lien . . . 41
4.6 En pratique . . . 41
4.6.1 Régression Gamma . . . 42
4.6.2 Régression logistique . . . 42
4.7 Régression Bêta . . . 43
4.8 Conclusion . . . 44
5 Mesures d'erreurs . . . 46
5.1 Sur-apprentissage (*overfitting*) . . . 46
5.2 Mesures d'erreur . . . 46
5.2.1 Statistique de déviance . . . 46
5.2.1.1 Déviance Gaussienne . . . 47
5.2.1.2 Déviance de Gamma . . . 47
5.2.1.3 Déviance de Bernoulli . . . 48
5.2.1.4 Déviance de Bêta . . . 48

5.2.2 Ratio de l'erreur de validation . . . 48
5.2.3 Mesure pseudo-R^2 . . . 49
5.2.4 Erreur de Laplace . . . 49
5.2.5 Courbe ROC pour la classification . . . 49
6 Apprentissage automatique (*machine learning*) . . . 51
6.1 Généralités . . . 51
6.2 Arbres de régression . . . 52
6.3 Apprentissage ensembliste . . . 55
6.4 Conclusion . . . 55
7 Forêts aléatoires (*random forest*) . . . 56
7.1 Définition . . . 56
7.2 Erreur de validation et convergence d'une forêt . . . 56
7.3 Importance des variables explicatives . . . 59
7.4 *Random forest* en pratique . . . 59
8 Gradient Boosting Machine (GBM) . . . 60
8.1 Définition . . . 60
8.2 Réduction de l'overfitting . . . 61
8.2.1 Learning rate et nombre d'itérations . . . 61
8.2.2 Sous échantillonnage . . . 62
8.3 Choix de la fonction coût . . . 63
8.4 Influence relative des variables explicatives . . . 63
8.5 En pratique . . . 64
9 Calibration de la vulnérabilité . . . 66
9.1 Variables explicatives utilisées . . . 66
9.2 Modèles linéaires généralisés . . . 66
9.2.1 Régression Gamma . . . 67
9.2.2 Régression bêta . . . 70
9.3 Gradient Boosting Machine . . . 73
9.4 Forêts aléatoires . . . 77
9.5 Conclusion . . . 78
10 Calibration de la probabilité de sinistre . . . 80
10.1 Régression logistique . . . 80
10.2 Forêts aléatoires . . . 80
10.3 Gradient Boosting Machine . . . 84
10.4 Conclusion . . . 87
11 Estimation des pertes nettes . . . 88
11.1 Estimation du coût des évènements sur le portefeuille 2016 . . . 88
11.1.1 Prise en compte des conditions financières . . . 88
11.1.2 Prise en compte des données manquantes . . . 89
11.2 Résultats numériques . . . 92
11.3 Calibration du modèle fréquence sévérité . . . 92
11.3.1 Lois de fréquence . . . 93
11.3.2 Lois de sévérité . . . 93
11.3.3 Critères d'ajustement . . . 93
11.4 Crédibilisation par des scénarios . . . 95
12 Impact sur la réassurance . . . 98
12.1 Introduction à la réassurance . . . 98
12.2 Capital économique requis par Solvabilité 2 . . . 100
12.3 Réassurance optimale pour le risque d'inondation . . . 100
12.3.1 Tarification d'un traité XS . . . 100

12.3.2 Création de valeur espérée . . . 101
12.3.3 Appétit au risque . . . 102
12.3.4 Conclusion . . . 103
12.3.5 Résolution numérique . . . 103
Conclusion . . . 106
Bibliographie . . . 106
A Annexe :Démonstrations et outils mathématiques . . . 109
A.1 *GLMs* : Démonstration de l'équation (4.3) . . . 109
A.2 Estimateur du maximum de vraisemblance . . . 109
A.2.1 Fonction de vraisemblance . . . 110
A.2.2 Matrice d'information de Fischer . . . 110
A.2.3 Convergence en loi . . . 110
A.3 Fonction $\Gamma(\cdot)$ . . . 111
A.4 Probabilité d'appartenir à OOB data dans la construction d'un arbre de forêt 111

Introduction

Les catastrophes naturelles représentent un risque majeur pour les compagnies d'assurance car elles peuvent remettre en question leur solvabilité par l'importance des pertes économiques qu'elle génèrent. Afin de mieux connaître leur exposition et de choisir de manière optimale les fonds propres à détenir et les structures de réassurance à utiliser, les assureurs éprouvent le besoin de développer leur propre appréciation des risques naturels à travers des modèles spécifiques qui modélisent l'évènement physique lui même avant de quantifier les pertes qui en découleraient. Les sorties de ces modèles permettent une quantification du risque en simulant des évènements fictifs donnant lieu, après application à l'exposition courante, à une estimation de la distribution stochastique annuelle des pertes.

La construction de tels modèles requièrent des expertises physiques pointues et leur validation par le régulateur peut alors être compliquée au vu de l'opacité sous-jacente. De plus, l'existence du régime CAT NAT en France et de la protection financière de l'état à travers la Caisse Centrale de Réassurance contre les risques naturels ne favorisent pas le développement de tels modèles sur le marché contrairement à d'autres pays.

C'est dans ce contexte que l'on s'intéresse dans ce mémoire à construire un modèle intermédiaire entre une vision purement historique et une vision purement stochastique du risque d'inondation en France. En effet, on se donne pour ambition de modéliser de manière dynamique ce risque à partir d'estimations du coût des évènements historiques subis par AXA. Le caractère dynamique du modèle signifie que celui-ci permettra de produire des estimations qui varieront au cours du temps et qui dépendront de la distribution géographique des sites assurés par AXA et de l'aménagement urbain et des moyens de protection physique contre les inondations mis en œuvre par les villes concernées.

Pour ce faire, le mémoire s'articule en quatre blocs retraçant les étapes majeures de la réalisation de cette étude.

Le premier bloc présente les données à disposition et fixe les périmètres et hypothèses de l'étude. Il permet également d'aborder quelques généralités sur les risques assurantiels liés aux catastrophes naturelles et les spécificités liées à leur modélisation en se concentrant sur l'inondation.

Un deuxième bloc permet de traiter le sujet le plus délicat de l'étude qui consiste à repérer les zones historiquement inondées par les évènements passés et les hauteurs d'eau associées, nous expliciterons alors l'algorithme derrière le modèle physique de propagation de l'eau utilisé afin de construire ces empreintes historiques.

Le troisième bloc présente théoriquement les différents outils d'apprentissage (*GLMs*, *random forest* et *gradient boosting*) et les mesures d'erreur utilisées dans la construction du modèle prédictif des coûts de chaque évènement sur le portefeuille 2016.

Un dernier bloc propose d'utiliser les estimations produites afin de construire le modèle du risque d'inondation porté par AXA France à travers un modèle fréquence-coût crédibilisé par un modèle de marché au vu de la faible longueur de l'historique. Ce modèle est enfin utilisé à des fins d'optimisation de réassurance au vu du capital économique requis par la réglementation et l'appétit au risque propre à l'assureur.

1 Descriptif des données à disposition

Nous proposons dans cette partie de définir le périmètre de notre étude en décrivant l'ensemble des données auxquelles nous avons accès et qui seront utilisées dans les différentes approches qui seront appliquées par la suite.

1.1 Base sinistres

Nous disposons d'une base de données recensant les sinistres de dommages directs (i.e. hors perte d'exploitation) déclarés à AXA France IARD provenant de 27 évènements d'inondation ayant eu lieu en France entre 1999 et 2012.

Ces sinistres peuvent provenir de six branches d'activité différentes :

- AGR : Polices permettant de se couvrir contre les risques liés aux exploitations agricoles
- COL : Polices souscrites par les collectivités permettant d'assurer des biens publics
- IMM : Polices permettant d'assurer des immeubles
- MIE : Polices qui permettent de couvrir les risques courus par des locaux à usage industriel
- MRH : Polices Multi-Risques Habitation, elles permettent d'assurer les appartements et maisons à usage résidentiel
- MRP : Polices Multi-Risques Professionnels, elles permettent d'assurer les risques courus par des locaux à usage commercial

Notons que toutes ces branches couvrent le péril inondation en France [1] et sont donc toutes concernées par notre modélisation.

Nous disposons pour chaque sinistre des informations suivantes :

- Charge du sinistre avant application des éventuelles franchises
- Valeur du bien sinistré
- Branche du contrat sinistré
- Coordonnées du lieu du sinistre
- Évènement ayant causé le sinistre

Malheureusement, cette base n'est pas complète et on y retrouve plusieurs sinistres avec des informations manquantes et qui sont essentielles à la suite de l'étude comme la branche du contrat sinistré, la valeur assurée du contrat ou l'information géographique du sinistre. Ces sinistres représentent 54% de l'ensemble de la base et comptent pour 59% de la charge totale, ils ne seront pas utilisés par la suite mais seront pris en compte à travers une renormalisation des résultats finaux.

Afin que les comparaisons entre les différents sinistres soient cohérentes, nous devons réévaluer les charges et les valeurs assurées en tenant compte de l'inflation et les ramener à une même année de référence, nous avons choisi l'indice de construction donné par la Fédération Française du Bâtiment (FFB [2]) qui rend compte des variations de coût des différents éléments qui entrent dans la composition de la construction d'un immeuble de type courant à Paris. Il est pertinent d'utiliser cet indice car la majeure partie de la charge d'un sinistre inondation provient des travaux de rénovation et de remise en état du bien détruit et cet indice était justement conçu au départ pour l'indexation des polices d'assurance. Notons néanmoins qu'il aurait été plus précis d'appliquer un

1. Voir partie 2.1.2
2. http://www.ffbatiment.fr/federation-francaise-du-batiment/le-batiment-et-vous/en_chiffres/indices-index/Chiffres_Index_FFB_Construction.html

indice plus adapté à la partie de la charge sinistre liée au remplacement du mobilier détruit mais la base de données à notre disposition ne nous permet pas de séparer la charge selon le type de frais.

Nous pouvons voir sur les figures 1.1 et 1.2 respectivement la répartition du nombre et de la charge des sinistres (de la base complète) par branche d'activité. On remarque que la branche MRH représente une grande partie des sinistres mais que la charge par sinistre est plus importante sur les autres branches.

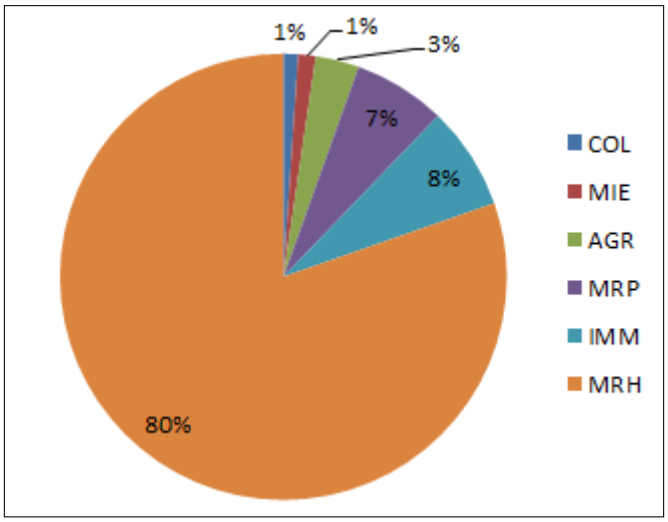

FIGURE 1.1 – Répartition du nombre de sinistres inondation par branche

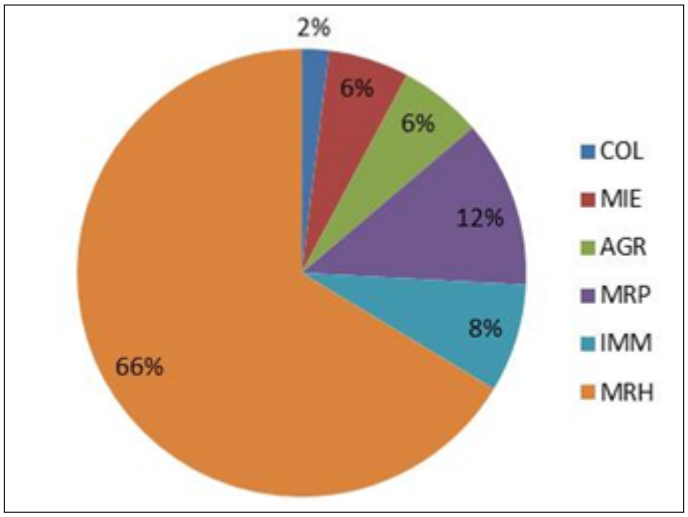

FIGURE 1.2 – Répartition de la charge des sinistres inondation par branche

L'histogramme 1.3 représente les coûts en euros courants des évènements d'inondation survenus en France entre 1999 et 2012, on remarque que trois évènements principaux se démarquent des autres, il s'agit de la crue dans l'Aude de novembre 1999 ,celle du Sud-Est de septembre 2002 et celle du Rhône de décembre 2003.

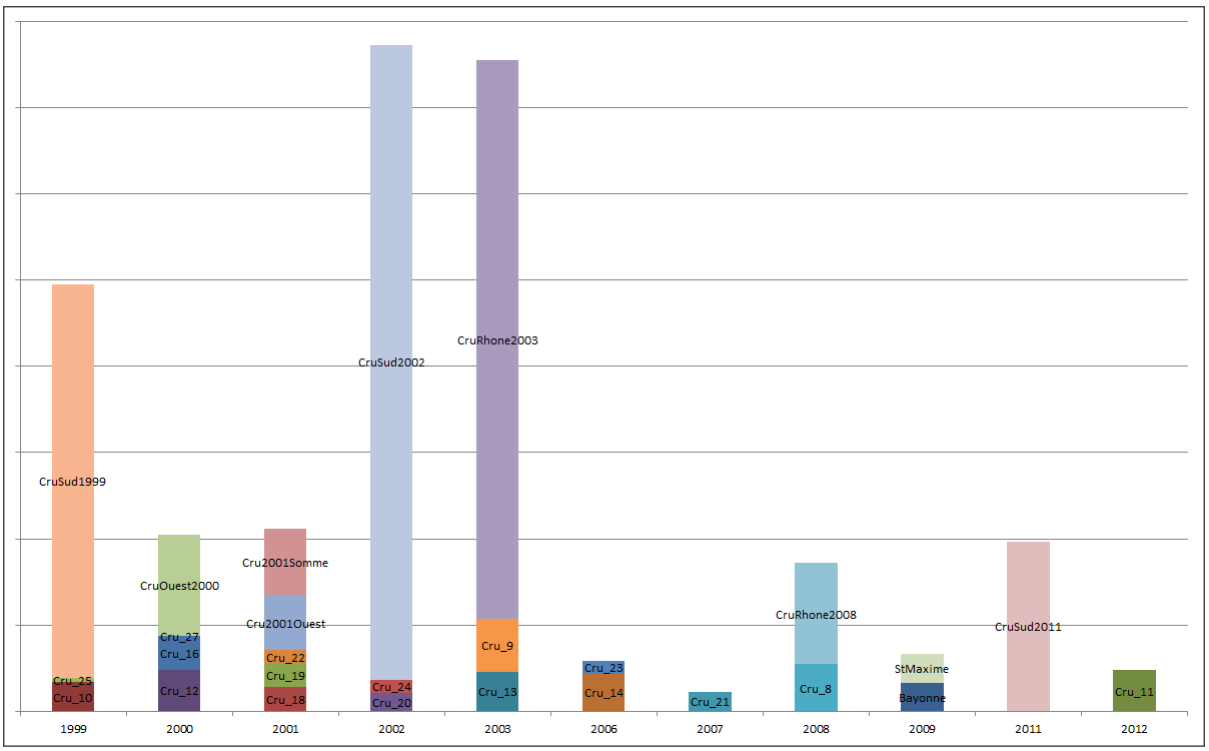

FIGURE 1.3 – Évènements d'inondation ayant eu lieu en France entre 1999 et 2012 et leurs coûts en euros courants.

1.2 Base des arrêtés CAT NAT [3]

Nous aimerions aussi prendre en compte les zones géographiques qui ont été inondées par les évènements mais qui n'ont pas été sinistrées, soit parce que il n'y avait aucun site assuré par AXA au moment de l'évènement ou parce que l'inondation n'était pas assez intense pour générer un sinistre. Nous disposons pour cela de la base Gaspar [4] qui est entretenue par des services départementaux et qui recense, entre autres, l'ensemble des communes où l'état de catastrophe naturelle a été reconnu suite à une inondation et la date de l'arrêté correspondant.

1.3 Portefeuille d'AXA France IARD 2016

Afin de pouvoir évaluer le coût des évènements présentés ci-dessus, nous avons besoin d'informations sur l'exposition actuelle d'AXA France IARD. Nous avons à notre disposition le portefeuille de l'année 2016 avec pour chaque police les informations suivantes :

- Valeur du bien assuré
- Branche de la police
- Coordonnées géographiques du bien assuré
- Diverses informations sur la nature du bien (étage, matériaux, etc)

1.4 Portefeuilles de 2003 à 2012

Afin de pouvoir se replacer dans le contexte des évènements, nous avons besoin de connaître l'exposition historique d'AXA France lors de leurs survenances. Nous disposons pour cela des

3. Le régime CATNAT français est présenté avec plus de détails dans la partie 2.1.2
4. http://macommune.prim.net/gaspar/

portefeuilles d'AXA France IARD de 2003 à 2012 de la branche MRH avec pour chaque police les informations suivantes :

- Valeur du bien assuré
- Géolocalisation partielle du site assuré
- Diverses informations sur la nature du bien (étage, matériaux, etc)

Nous remarquons immédiatement une baisse dans la qualité des données par rapport au portefeuille 2016 :

- Tous les sites assurés ne sont pas géocodés et leur proportion se détériore au fur et à mesure que l'on recule dans le temps. Nous disposons néanmoins du nombre de sites assurés total pour chaque année depuis 1999. Nous ferons alors l'hypothèse que les sites non géocodés sont géographiquement répartis de la même manière que les sites géocodés afin d'approximer le portefeuille dans sa totalité.
- Nous n'avons pas les portefeuilles des autres branches d'activité, ils seront approximés à partir du portefeuille MRH en faisant l'hypothèse que les proportions des différentes branches sont les mêmes que dans le portefeuille 2016
- Enfin, nous n'avons pas de portefeuille pour exploiter les évènements survenus entre 1999 et 2002 mais on connaît néanmoins l'évolution globale de la taille des portefeuilles entre 1999 et 2003 en terme de somme totale assurée, on approximera donc ces portefeuilles en utilisant le portefeuille 2003 corrigé de l'évolution de l'exposition globale. En utilisant cette méthode, on fait l'hypothèse que la répartition géographique des sites assurés est constante entre 1999 et 2003.

1.5 Hauteurs d'eau aux stations de jaugeage

Les évènements historiques étudiés ont eu lieu suite à un débordement de rivière. Afin de pouvoir les reproduire physiquement, nous comptons utiliser les hauteurs d'eau relevées à l'aide de stations de jaugeage situés le long du réseau fluvial Français pendant la période des évènements. Celles-ci sont fournies par la banque de données HYDRO [5] qui est gérée par le Ministère de l'Ecologie, du Développement Durable et de l'Energie, des identifiants sont nécessaires pour pouvoir y accéder.

Après avoir traité et éliminé les données inexploitables dans le cadre de notre étude (historiques trop courts ne couvrant pas l'ensemble des évènements), nous gardons un total de 1459 stations de jaugeage avec les hauteurs d'eau historiques associées. La figure 1.4 illustre la répartition géographique de ces stations.

5. http://www.hydro.eaufrance.fr/

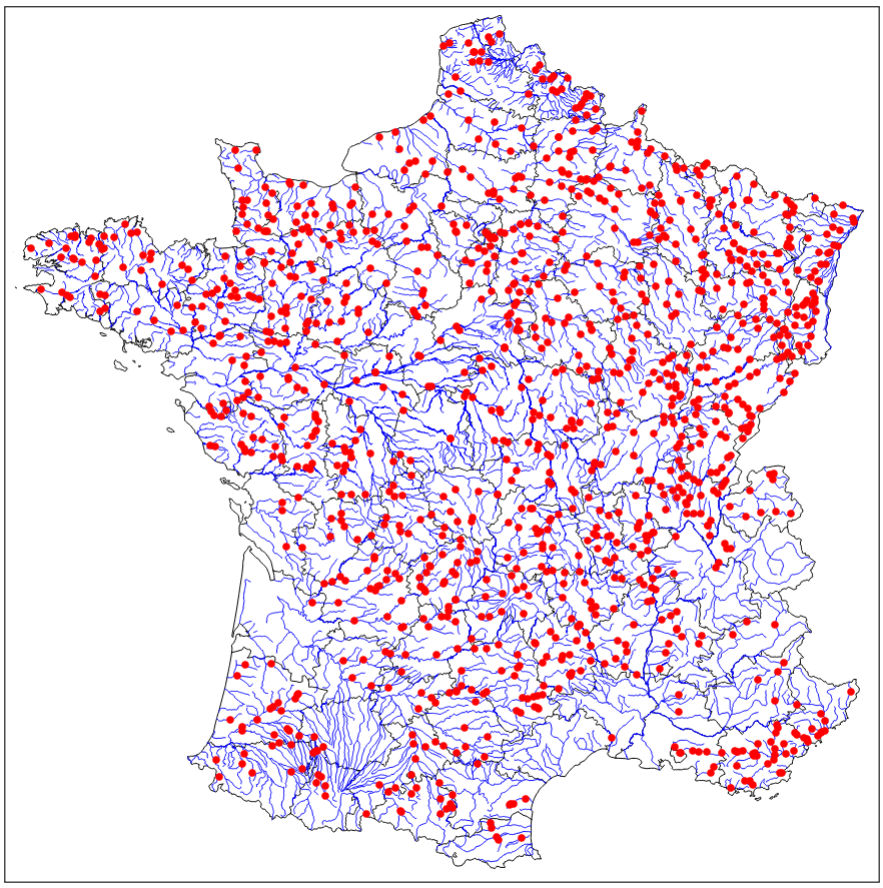

FIGURE 1.4 – Répartition géographique des stations de jaugeage utilisées

1.6 Précipitations historiques

Les débordements de rivière ayant causé les évènements étudiés sont dûs en premier lieu à des accumulations de précipitations intenses, il se peut alors, en plus des zones inondées directement par débordement de rivière, que certaines zones aient été inondées par ruissellement [6]. Nous nous intéressons alors aux précipitations historiques ayant eu lieu pendant la durée de chaque évènement dans la zone géographique concernée. Nous y avons accès à travers la banque de données Era Interim [7](publique dans le cadre de recherches académiques uniquement) qui contient toutes les quantités précipitations journalières ayant eu lieu en Europe depuis 1979 sur un maillage géographique d'une résolution de 80 kms.

1.7 Cartographie du territoire

A chaque fois que des études géographiques seront nécessaires, la France sera représentée par une grille spatiale (que l'on appelle *raster*) quadrillant le territoire avec une résolution de $75m \times 75m$. On parlera ainsi de cellule pour décrire une case de cette matrice.

Nous disposons également d'un Modèle Numérique de Terrain (MNT) fourni par l'Institut Géographique National (IGN) qui contient l'altitude de chacune des cellules avec une résolution d'$1m$.

6. Le ruissellement a lieu lorsque les sols atteignent leurs capacités d'absorption d'eau
7. http://apps.ecmwf.int/datasets/

Enfin le réseau fluvial qui sera utilisé provient de la base Carthage 4 fournie par l'IGN et qui décrit un réseau fluvial d'une longueur de 125 000 kms, des résolutions plus fines existent permettant de prendre en compte des fleuves plus petits mais celle-ci est suffisante dans notre cas où l'on analyse des évènements extrêmes seulement.

1.8 Périmètre de l'étude et hypothèses

Cette partie nous a permis de fixer le périmètre de notre étude à travers les données auxquelles nous avons pu accéder et aussi d'émettre des premières hypothèses afin de combler à la mauvaise qualité de certaines d'entre elles.

Il en ressort notamment que l'on étudiera les coûts des dommages directement (les pertes d'exploitations ne sont pas étudiées) causés par 27 évènements historiques d'inondations ayant eu lieu en France entre 1999 et 2012 sur six branches d'activités différentes. On cherchera à reproduire ces évènements en modélisant les débordements de rivières qui ont pu avoir lieu et les accumulations des précipitations à certains endroits ayant épuisé la capacité d'absorption d'eau des sols.

Par ailleurs, nous ferons l'hypothèse que la répartition géographique des sites assurés était constante entre les années 1999 et 2002 et que ceux-ci étaient uniformément répartis à l'échelle de chaque commune entre les années 1999 et 2012.

2 Catastrophes naturelles et leur modélisation

Nous nous intéressons dans cette partie à décrire les spécificités des risques de catastrophes d'origine naturelle par rapport aux autres risques courus par une compagnie d'assurance et à expliquer les différentes alternatives de modélisation qui nous sont offertes.

2.1 Spécificités du risque de catastrophes naturelles

2.1.1 Définition

Nous distinguons en assurance non vie trois types de risques :

- Les **risques attritionnels** représentent des sinistres fréquents mais dont les coûts sont relativement faibles dans la mesure où ils sont pris en charge facilement par l'assureur (ex : bris de glace, vol automobile, etc).
- Les **risques atypiques** représentent a contrario des sinistres peu fréquents mais dont l'impact est très élevé (ex : un incendie détruisant un site industriel, accident ferroviaire, etc). Ils nécessitent une modélisation différente qui se concentre sur la queue de distribution de la sinistralité globale ; en pratique on les distingue des risques atypiques à partir d'un seuil de coût calculé en se basant sur la sinistralité globale historique.
- Les **risques de catastrophes** qui représentent les sinistres causés par des évènements rares de forte ampleur impactant une large zone géographique et générant des dégâts sur plusieurs sites assurés. Ils peuvent être d'origine naturelle (ex : gel, sécheresse, tempête, tremblement de terre, etc) ou humaine (ex : accident nucléaire, terrorisme, etc).

(a) Sinistre attritionnel : bris de glace auto

(b) Sinistre atypique : incendie d'un site industriel

FIGURE 2.1 – Exemples de sinistres attritionnels et atypiques

(a) Sinistre de catastrophe naturelle : séisme

(b) Sinistre de catastrophe humaine : attentat terroriste

FIGURE 2.2 – Exemples de sinistres catastrophes

Par le fait qu'elles touchent de larges zones géographiques, les catastrophes naturelles peuvent impacter la quasi-totalité du portefeuille d'un assureur et remettre en question sa solvabilité car le bénéfice dû à la mutualisation des risques s'en retrouve très limité. Le recours aux réassureurs, qui profitent d'une diversification des risques à une échelle géographique plus large, permet de palier à cette carence de mutualisation. Ainsi, chaque assureur éprouve le besoin de développer sa propre vision du risque de catastrophes naturelles auquel il est soumis afin de définir de manière optimale ses besoins en réassurance et la capital économique à immobiliser en face de ce risque.

2.1.2 Régime d'assurance des catastrophes naturelles en France et assurabilité du risque d'inondation

Pour pallier à la carence de couverture des risques naturels, la France dispose depuis 1982 d'un régime particulier d'assurance des catastrophes naturelles nommé CATNAT. Il concerne les catastrophes naturelles telles qu'elles sont définies dans l'article L125-1 du code des assurances :

> *"Sont considérés comme les effets des catastrophes naturelles, au sens du présent chapitre, les dommages matériels directs non assurables ayant eu pour cause déterminante l'intensité anormale d'un agent naturel, lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises."*

Ainsi, le régime CATNAT ne s'applique pas à certaines catastrophes d'origine naturelle qui sont normalement assurables au titre de garanties contractuelles comme les orages de grêle, l'effet du gel sur les plantations ou les incendies et tempêtes [8]

Nous nous interrogeons maintenant sur les conditions à réunir afin qu'un risque soit assurable. Les aspects essentiels de l'assurance tels que les présente Willi Gruss dans [1] permettent de déterminer le degré d'assurabilité d'un risque :

- **La communauté d'intérêts** : les personnes menacées doivent être nombreuses et former une communauté de risques.
- **Le besoin** : la survenance de l'évènement redouté doit précariser la situation économique de l'assuré.
- **L'estimation** : la charge des sinistres escomptés doit pouvoir être chiffrée.
- **Le caractère aléatoire** : le moment où survient l'évènement assuré ne doit pas être prévisible ; l'évènement doit se produire indépendamment de la volonté du preneur d'assurance.
- **La rentabilité** : les personnes assurées constituent une communauté dont l'objet est de couvrir les besoins financiers futurs selon un plan bien concret.
- **La menace similaire** : tous les individus formant la communauté d'assurance doivent être exposés aux mêmes dangers ; la survenance de l'évènement redouté doit affecter tous les patrimoines de façon similaire.

La condition de communauté d'intérêts permet à elle seule de montrer que le risque d'inondation n'est pas assurable. En effet, les bâtiments fortement menacés par les inondations représentent généralement moins de 1% du parc immobilier d'un pays [9]. La communauté de risques est donc trop restreinte pour trouver une solution satisfaisante à la fois pour le preneur d'assurance et l'assureur. Prenons l'exemple d'un particulier propriétaire de son habitation principale qui est

8. L'article L 122-7 du code des assurances 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, ouragan ou cyclones, sur les biens faisant l'objet de tels contrats. Les tempêtes, ouragans ou cyclones sont donc exclus du régime CATNAT.

9. Swiss Re, 1999 - Voir [2]

située dans une zone inondable de période de retour 5 ans. En supposant un taux de destruction du bien de 5% [10] en cas d'inondation, la prime pure annuelle s'élèverait à 1% de la valeur assurée du bien, soit quelques milliers d'euros pour une résidence principale pour se couvrir contre le risque d'inondation uniquement. En outre, le chargement technique aurait un impact important sur la prime finale proposée en raison d'un bénéfice de mutualisation très faible sur ce risque là. Il est donc économiquement difficile pour la population de se couvrir contre le risque d'inondation de cette manière.

Par ailleurs, l'article L125-1 stipule également que tous *" les contrats d'assurance garantissant les dommages d'incendie ou tous autres dommages à des biens situés en France, ainsi que les dommages aux corps de véhicules terrestres à moteur, ouvrent droit à la garantie de l'assuré contre les effets des catastrophes naturelles "*. Les inondations d'intensité anormale étant considérées comme des catastrophes naturelles, elles sont donc garanties par l'ensemble des contrats [11] étudiés dans le cadre de ce mémoire. Ces garanties sont financées à travers une contribution uniforme de l'ensemble des assurés de chaque branche d'activité indépendamment de leur exposition réelle au risque de catastrophes naturelles. Cette surprime est imposée par le régime CATNAT et s'élève à :

- 12% de la prime de base pour les biens autres que véhicules à moteur
- 6% des primes de base liées au vol et aux incendies (ou, à défaut 0.50% de la prime de dommages) pour les véhicules terrestres à moteur

Notons aussi que 12% de cette surprime permet de financer le fonds de prévention des risques naturels majeurs (dit Fonds Barnier) qui permet de financer toute initiative visant à renforcer les moyens de protection contre les risques naturels.

Maintenant que la condition de la communauté d'intérêts est vérifiée, il reste à remplir la condition de rentabilité. En effet, tous les preneurs d'assurance doivent pouvoir être correctement indemnisés en cas de sinistre. Or, on a vu que la solvabilité des assureurs locaux peut rapidement être mise en péril en cas de survenance d'évènements extrêmes en raison du faible bénéfice de mutualisation. Le régime CATNAT permet là aussi de combler cette carence en proposant une solution de réassurance avec une garantie illimitée à travers la Caisse Centrale de Réassurance (CCR) détenue en totalité par l'Etat. Ainsi, chaque assureur assurant les effets d'une catastrophe naturelle a la possibilité de se réassurer auprès de la CCR à travers le programme de réassurance suivant :

- Un traité en quote part à 50%, signifiant que l'assureur et la CCR se partagent de manière égale les pertes générées par les effets d'une catastrophe naturelle.
- Un traité en excédent de perte annuelle, signifiant que la CCR prend en charge entièrement la partie de la perte annuelle dépassant une certaine franchise, ce qui plafonne ainsi le reste à charge annuel de l'assureur.
- Enfin, l'Etat prend en charge les pertes qui ne peuvent pas être absorbées par les réserves de la CCR.

Rappelons enfin que l'ensemble de ces garanties sont appliquées uniquement si un évènement est reconnu comme effet d'une catastrophe naturelle au sens de l'article L125-1 du code des assurances à travers un arrêté interministériel.

Ainsi, le régime CATNAT français permet d'offrir des garanties contre les risques non assurables en proposant un mécanisme de solidarité entre assurés d'une part et entre l'Etat et les

10. Le taux de destruction est défini par le montant des dégâts divisé par la valeur assuré du bien - Taux de 5% généralement estimé par les modèles d'inondation développés par AXA Global P&C

11. Voir partie 1.1

assureurs locaux d'autre part.

La condition d'estimation peut également constituer un frein à l'assurabilité de ces risques dans la mesure où l'on ne dispose pas de suffisamment d'observations qui auraient permis une quantification fiable du risque couvert, c'est tout l'objet du développement récent de techniques spécifiques de modélisation des catastrophes naturelles témoignant d'un réel besoin de quantification des ces risques de la part des assureurs. On peut néanmoins considérer que le régime CATNAT permet à l'assureur d'être indifférent à cette condition à travers le mécanisme de réassurance offert. Il est enfin immédiat que les trois conditions d'assurabilité restantes (besoin, caractère aléatoire et menace similaire) sont réunies.

2.1.3 Spécificités de la modélisation

Notons que contrairement aux autres risques, l'objectif de la modélisation d'un risque de catastrophes naturelles n'est pas de simuler les pertes à l'échelle d'un sinistre mais de développer une vision des pertes agrégées à l'échelle de l'évènement. La méthode de modélisation classique qui consiste à calibrer des distributions de pertes à partir des pertes historiques en vue de prédire les pertes futures ne s'applique pas aux risques de catastrophes naturelles pour plusieurs raisons :

- Le caractère peu fréquent des évènements en question limite la quantité de données à notre disposition et ne nous permettrait donc pas d'être confiant dans la distribution obtenue après calibration.
- Cette méthode repose très fortement sur l'hypothèse que le passé est représentatif du futur. Cette hypothèse n'est pas vérifiée dans le cas des catastrophes naturelles car les pertes occasionnées dépendent de la répartition géographique du portefeuille de l'assureur qui peut évoluer de manière significative d'une année à l'autre. Par ailleurs, l'aléa physique sous-jacent dépend de plusieurs paramètres météorologiques amenés à varier considérablement au cours du temps et dont l'évolution est impactée par le contexte actuel du changement climatique.
- L'activité moderne d'assurance existant depuis peu longtemps, la longueur de l'historique des observations à disposition des assureurs va rarement au delà d'une vingtaine d'années. La modélisation des évènements extrêmes est alors plus compliquée et il devient impossible d'estimer de manière confiante la perte survenant en moyenne une fois tous les 200 ans requise pour définir le capital à immobiliser en face du risque de catastrophes naturelles dans le cadre de Solvabilité 2.

Face à ces obstacles, nous avons besoin de modéliser directement l'évènement physique et les zones géographiques qu'il est susceptible de toucher avant de chercher à estimer les pertes assurantielles qui en découleraient. Devant la difficulté technique et le besoin conséquent en ressources (hydrologues, sismologues, météorologues, etc), la quasi totalité des assureurs achètent des modèles auprès des trois spécialistes dans ce type de modélisation et qui servent de référence dans le marché de la réassurance pour la tarification des programmes de réassurance couvrant les catastrophes naturelles par exemple, il s'agit d'*AIR*, *EQECat* et *RMS*. Néanmoins, le coût des licences annuelles de ces logiciels de modélisation et leur aspect "boîte noire" poussent certains (ré)assureurs à construire leurs propres modèles en interne. Ils développent ainsi une vision plus précise et une meilleure compréhension du risque modélisé, peuvent orienter le modèle à des usages autres que la tarification des programmes de réassurance (améliorer la politique de souscription par exemple) et produisent des estimations plus représentatives de leur exposition et politique de souscription car le modèle est calibré exclusivement sur leurs propres données contrairement aux logiciels qui utilisent plusieurs bases de données de source différente. La construction de ces modèles sera détaillée dans la partie 2.2.

2.2 Construction d'un modèle CAT

La construction d'un modèle de catastrophes naturelles s'articule en trois modules :

- **Le module aléa :** Ce module part de l'historique de données physiques (différentes selon le péril modélisé) et des caractéristiques de la zone géographique modélisée (altitude, dynamique météorologique, capacité d'absorption du sol, etc) et permet de générer un catalogue d'un certain nombre d'années (généralement 10000) d'évènements avec la zone géographique impactée ainsi que l'intensité associée. L'intensité est représentée par une variable physique permettant de décrire l'évènement (vitesse de vent pour le risque de tempête, hauteur d'eau pour les inondations, taille des grêlons pour le risque de grêle, etc).
- **Le module vulnérabilité :** À chaque évènement du catalogue obtenu, ce module évalue les dégâts causés à chaque site assuré impacté par l'évènement. Cette évaluation tient compte, en plus de l'intensité de l'évènement, des caractéristiques de vulnérabilité des bâtiments assurés (étage, matériaux de construction, construction para-sismique, etc) et de la valeur des sites impactés. On complète ainsi le catalogue des évènements par les pertes brutes qui leurs sont associées.
- **Le module financier :** Ce module permet enfin de transformer les pertes brutes en pertes nettes à charge de l'assureur après application des conditions financières des polices touchées (éventuels plafonds et franchises, contrats de co-assurance et programmes de réassurance déjà en place). Nous obtenons ainsi une distribution empirique des pertes nettes annuelles causées par le risque modélisé.

L'articulation de ces modules peut être visualisée sur la figure 2.3.

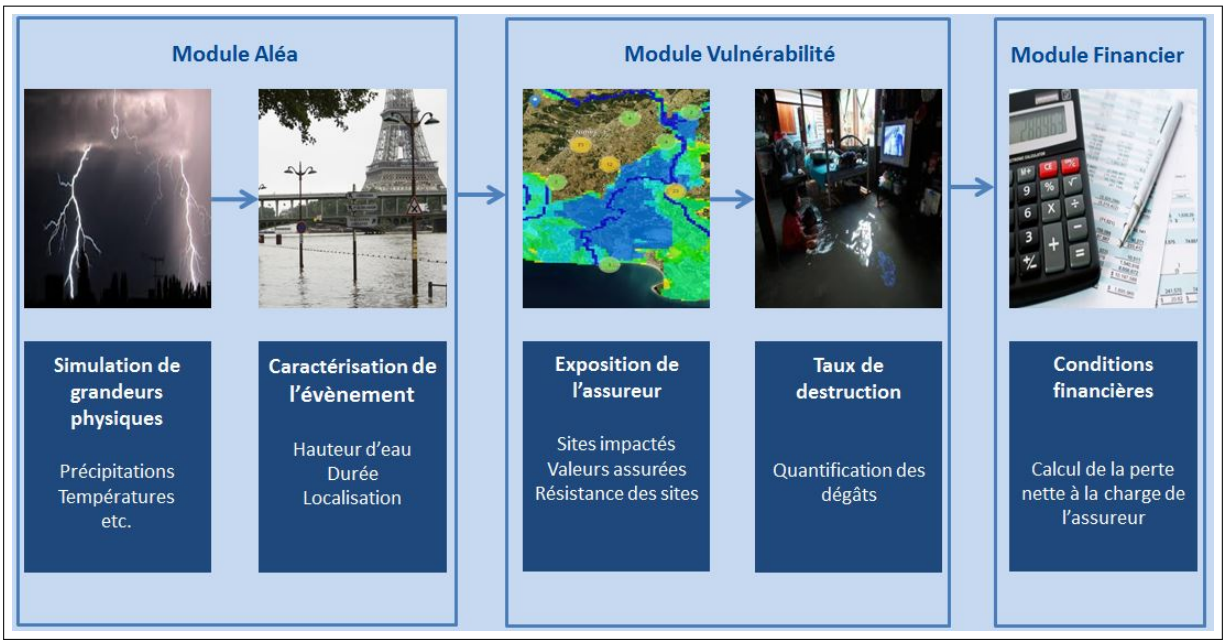

FIGURE 2.3 – Décomposition d'un modèle CAT en 3 modules

2.3 Description du risque d'inondation

Il n'existe pas de définition exacte et précise du risque d'inondation car ces évènements diffèrent selon les pays. On peut néanmoins définir [12] ,à titre indicatif, le risque d'inondation comme

12. SwissRe, 1999

le risque de submersion limitée dans le temps, partielle ou totale de zones normalement sèches par l'eau et des matériaux en suspension ou du matériau de charriage. Plusieurs phénomènes peuvent être à l'origine d'une inondation :

- **Tsunami :** d'immenses vagues sous-marines, se propageant à grande vitesse (de l'ordre de la centaine du km/h) dans les eaux profondes, qui finissent par émerger et s'écraser sur le littoral.
- **Raz-de-marée :** des masses d'eau énormes poussées par des tempêtes en période de marée haute près du littoral et pouvant inonder de grandes étendues de terres.
- **Inondation soudaine :** des précipitations brèves et violentes sur une zone restreinte font monter le niveau des petits cours d'eau jusqu'à inonder une partie des terres ; ce phénomène peut être accentué en milieu urbain où la surcharge des canalisations provoque des refoulements d'égout.
- **Débordement de rivière :** des précipitations abondantes pendant plusieurs jours font déborder les cours d'eau pouvant inonder des milliers de km^2 pendant plusieurs semaines.
- **Embâcle :** la fonte des neiges et des blocs de glace du printemps s'accumule contre un pont ou tout autre obstacle pouvant obstruer une partie de la rivière qui peut déborder localement.
- **Rupture de barrage :** peut être causée essentiellement par des précipitations abondantes et des glissements de terrain.
- **Lahar :** de violentes précipitations combinées à une éruption volcanique projetant de grandes quantités de cendrés se transforment en coulée boueuse dévalant les pentes du volcan.
- **Coulée boueuse :** un talus de terre meuble et fortement imbibée d'eau qui cède et se précipite dans la vallée à travers les rigoles et les lits des ruisseaux.

On se limite dans ce mémoire aux évènements provoqués par des inondations soudaines et des débordements de rivière.

3 Construction des empreintes historiques

La première étape de l'étude consiste à générer l'empreinte historique de chacun des évènements étudiés. Il s'agit concrètement d'associer à chaque évènement une carte répertoriant l'ensemble des zones géographiques inondées avec la hauteur d'eau correspondante. Comme il n'existe pas d'outils de mesure des hauteurs d'eau couvrant l'ensemble du territoire, il est impossible de savoir de manière certaine quelles sont les zones qui ont été inondées et avec quelle hauteur d'eau. La difficulté réside donc dans le fait de reproduire le plus fidèlement possible l'évènement en question. Nous partons alors de la base de données des hauteurs d'eau aux différentes stations de jaugeage, des précipitations historiques et du modèle physique de propagation de l'eau développé en interne par AXA Global P&C.

3.1 Module physique de propagation [13]

Le développement de ce module fait appel à des connaissances poussées en hydrologie et dépasse donc la finalité de ce mémoire, nous proposons néanmoins d'en décrire les grandes lignes par souci d'exhaustivité.

Le but du module physique de propagation est de générer une première empreinte avant de l'affiner à partir des données à notre disposition. Pour chaque évènement il s'agit de simuler les débits d'eau ayant eu lieu sur le réseau fluvial français avant de simuler son éventuelle propagation à l'intérieur des terres. Pour ce faire, une première étape est de modéliser les hauteurs d'eau à des intervalles de 20 kms le long des rivières. Nous disposons déjà de 1459 stations de jaugeage avec l'historique des hauteurs d'eau associé et plus de 10000 stations virtuelles ont donc été ajoutées afin de couvrir l'ensemble du réseau avec la résolution souhaitée.

Il s'agit maintenant de reproduire les hauteurs d'eau aux stations virtuelles, on utilise pour cela l'historique [14] des précipitations et des températures journalières observées entre 1958 et 2014. Des équations hydrologiques [15] tenant compte de l'altitude du terrain, des précipitations, de l'évapotranspiration et de l'impact de la température sur la fonte des neiges permettent de convertir ces données en hauteur d'eau aux différentes stations (réelles et virtuelles) à disposition. Les stations réelles permettent par ailleurs de mesurer l'erreur du modèle en comparant les hauteurs observées et les hauteurs simulées, on peut voir à titre illustratif cette comparaison sur la figure 3.1.

Enfin, en croisant temporellement les hauteurs d'eau avec les dates des évènements, on fait déborder l'eau sur les cellules concernées dès lors que la hauteur d'eau à une station dépasse le niveau du lit de la rivière. Il existe plusieurs algorithmes de propagation et de diffusion de l'eau pouvant prendre en compte différents paramètres d'influence comme la rugosité et d'autres variables géologiques. Pour des raisons de simplicité et de temps de calcul, nous avons fait le choix de prendre en considération uniquement la différence d'altitude entre cellules voisines et de faire déborder les cellules de manière récursive :

- Une cellule ne peut être inondée que par une cellule voisine dont la hauteur (altitude + niveau d'eau) dépasse l'altitude de la cellule à inonder.

13. Module développé par l'équipe Actuariat Réassurance d'AXA Global P&C en 2015
14. Base de données produite par Météo France à travers la base SAFRAN
15. Le lecteur intéressé pourra se diriger vers la documentation des modèles GR4J développé par l'Institut de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture (IRSTEA) pour comprendre le fonctionnement du module en détail http://webgr.irstea.fr/modeles/journalier-gr4j-2/fonctionnement_gr4j/

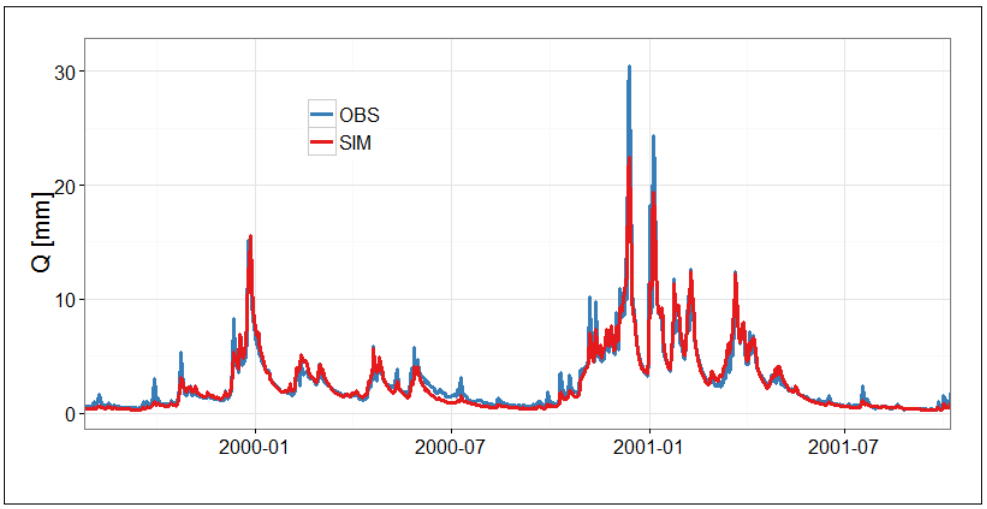

FIGURE 3.1 – Comparaison entre les hauteurs d'eau simulées et observées au niveau d'une station entre septembre 1999 et septembre 2001

- Le niveau d'eau de la nouvelle cellule inondée correspond au niveau d'eau maximal provenant d'une seule cellule voisine (une cellule ne peut pas être inondée par plusieurs cellules).
- Le niveau d'eau de la nouvelle cellule est égal à la hauteur d'eau de la cellule inondante dépassant son altitude.

Cet algorithme peut être visualisé sur la figure 3.2.

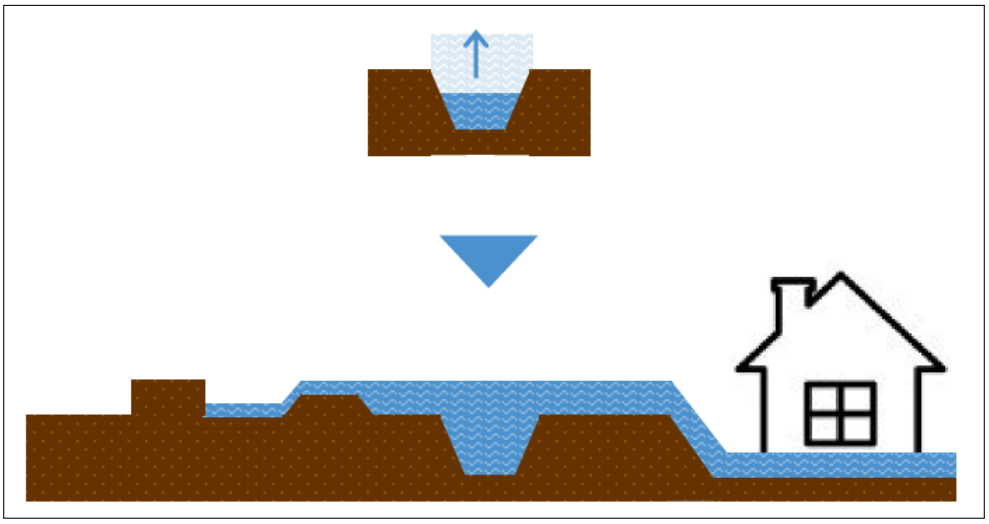

FIGURE 3.2 – Algorithme de diffusion de l'eau

Ce module permet ainsi d'obtenir en sortie une carte (au format *raster*) par évènement décrivant l'ensemble des cellules inondées avec la hauteur d'eau associée.

Limites du modèle

On remarque immédiatement que ce modèle inonde les cellules plus que la réalité car la notion de volume n'est pas prise en compte. En effet, un même volume d'eau peut ici se propager sur toutes les cellules voisines situées à la même altitude et en gardant la même quantité d'eau alors

qu'en réalité cette quantité diminuerait au fur et à mesure de la propagation jusqu'à disparaître.

En outre, même si elle est prise en compte à travers le Modèle Numérique de Terrain, l'existence éventuelle des moyens de défense et de protection contre les crues (digues notamment) constitue un facteur limitant car il existe peu de base de données structurées représentant cette information rendant une modélisation fidèle très difficile.

3.2 Affinement des empreintes obtenues

Compte tenu des contraintes que l'on vient de présenter, il est essentiel de retraiter les empreintes obtenues en supprimant les cellules qui auraient été inondées à tort. Ce retraitement est fait en exploitant les informations à notre disposition sur les évènements : la base GASPAR et la base sinistres.

Compte tenu de la taille importante du portefeuille d'AXA en France et de son homogénéité, nous pouvons raisonnablement affirmer qu'une cellule se situant dans une commune où aucun sinistre n'a été déclaré pendant la durée de l'évènement n'a fort probablement pas été inondée, surtout si aucun arrêté CAT NAT n'y a été déclaré.

Ainsi, nous gardons parmi les cellules inondées par notre modèle que :

- Les cellules se situant dans une commune où au moins un sinistre a été déclaré pendant l'évènement.
- Les cellules se situant dans une commune où un arrêté CAT NAT a été déclaré et qui se situe dans un département avec au moins un sinistre AXA. Cette deuxième condition sur le département permettent d'éliminer les communes touchées par des intempéries ayant eu lieu simultanément mais sans lien réel avec l'évènement étudié.

Nous pouvons comparer sur les figures 3.3 et 3.4 l'empreinte de la crue du Rhône de 2003 avant et après retraitement. On s'apperçoit effectivement que de nombreuses zones très éloignées du Rhône se retrouvent inondées par le modèle et sont retirées de la carte après le retraitement.

Notons que nous ne chercherons pas à corriger la surestimation des hauteurs d'eau en considérant que les rapports d'ordre entre cellules sont conservées par notre algorithme et que cette erreur sera donc annulée lors de la calibration des méthodes d'estimation du coût des sinistres.

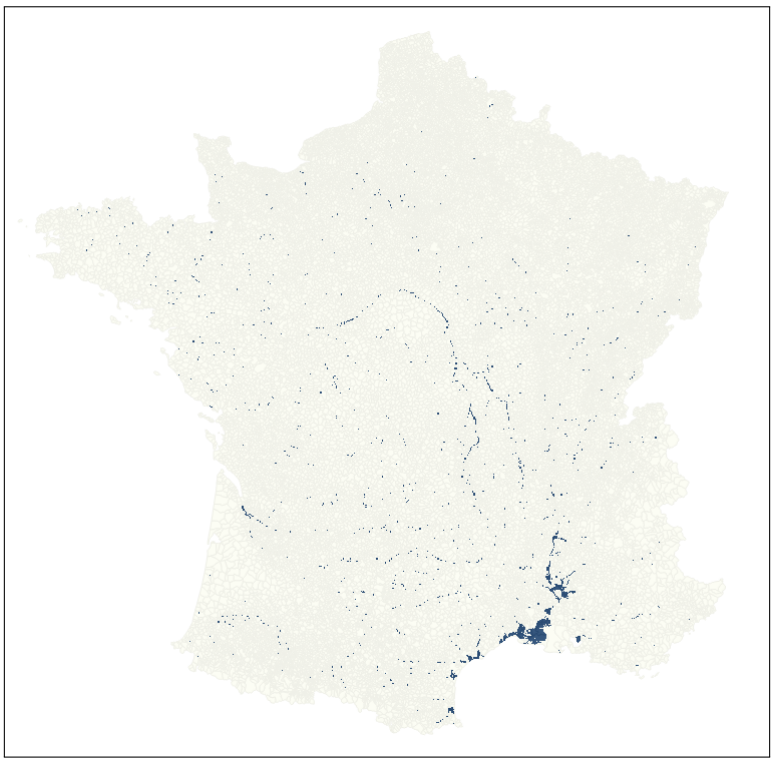

FIGURE 3.3 – Empreinte de la crue du Rhône 2003 sortie par le modèle, **avant** retraitement. Les zones inondées sont en bleu.

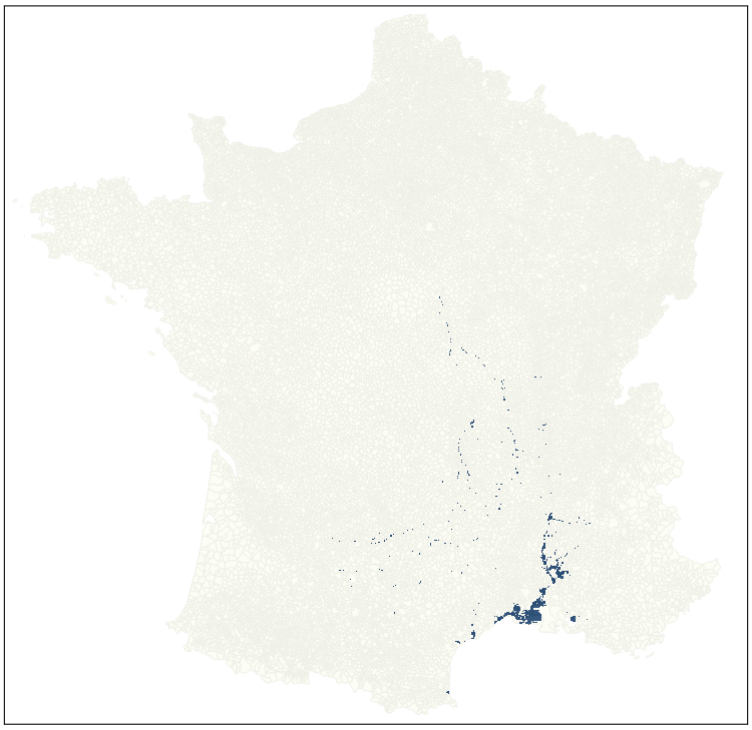

FIGURE 3.4 – Empreinte de la crue du Rhône 2003 sortie par le modèle, **après** retraitement. Les zones inondées sont en bleu.

3.3 Prise en compte du ruissellement

Enfin, nous souhaitons enrichir ces cartes pour tenir compte du ruissellement, c'est-à-dire des zones où de fortes précipitations ont eu lieu jusqu'à épuiser les capacités d'absorption des sols conduisant à une accumulation de l'eau à certains endroits.

Pour prendre en compte ce phénomène [16], on utilise le Modèle Numérique de Terrain pour attribuer à chaque cellule un score reflétant à quel point l'eau est susceptible de la traverser pour être évacuée. Après avoir défini un seuil sur ces scores, on obtient un réseau de drainage [17] à l'échelle de toute la France. Nous calculons, dans une deuxième étape, la hauteur relative de chaque cellule au point du réseau de drainage auquel elle conduit. C'est cette variable, associée aux quantités de précipitations ayant eu lieu pendant l'évènement, qui seront utilisées lors de la calibration du modèle pour tenir compte du ruissellement. Nous pouvons voir sur la figure 3.5 un exemple de carte de drainage.

3	2	3	6	7	13	15
6	0	0	0	3	11	17
14	7	1	0	4	9	16
19	13	4	2	0	5	4
17	15	13	4	1	0	0
16	16	15	12	7	5	3
18	17	19	17	17	11	7

FIGURE 3.5 – Exemple de carte de drainage. Le réseau de drainage est constitué par les cases légèrement bleues avec des zéros rouges. Les autres cases contiennent la hauteur au dessus du point du réseau auquel elles conduisent.

16. Le lecteur intéressé pourra se documenter sur les algorithmes TauDEM (*Terrain Analysis Using Digital Elevation Models*) et HAND (*Height Above the Nearest Drainage*) qui ont été utilisés pour générer ces cartes.
17. Réseau emprunté par l'eau pour être évacué pendant les précipitations

4 Modèles Linéaires Généralisés

Nous présentons dans cette partie un premier outil mathématique de régression et de classification qui sera utilisé pour l'estimation des pertes : les modèles linéaires généralisés (*GLMs*).

4.1 Définition

Contrairement au modèle linéaire classique où l'on suppose que la variable réponse suit une loi normale, que son espérance dépend linéairement des variables explicatives et que sa variance est constante (homoscédasticité), les *GLMs* permettent de gérer d'autres distributions, un lien entre espérance et variables explicatives non forcément linéaire et une variance de la variable réponse non nécessairement constante, elle dépend le plus souvent de l'espérance (hétéroscédasticité).

On désignera par Y la variable réponse et par $X = (X^1, \dots, X^p)$ le vecteur des variables explicatives. On suppose qu'il existe une certaine fonction F telle que $E(Y|X) = F(X)$ et on cherche à l'approcher par une fonction $\hat{F}$.

L'hypothèse du modèle linéaire classique selon laquelle les variables aléatoires $Y|X$ sont générées de manière indépendante est maintenue.

4.2 Famille exponentielle revisitée

Soit Y une variable aléatoire dont la loi de probabilité dépend d'un paramètre $\gamma \in \mathbb{R}^d$. On dit que la loi de Y appartient à la **famille exponentielle** si sa fonction de densité (par rapport à la mesure de comptage sur $\mathbb{N}$ et la mesure de Lebesgue sur $\mathbb{R}$) peut s'écrire sous la forme :

$$f(y|\gamma) = exp\left(\sum_{j=1}^{d} a_j(y)\,\alpha_j(\gamma) \;+\; b(y) \;+ \beta(\gamma)\right)$$

Dans les *GLMs*, on considère des lois de la famille exponentielle mais dont la fonction de densité peut s'écrire sous la forme :

$$f(y|\theta, \phi) = exp\left(\frac{y\,\theta \;-\; b(\theta)}{\phi} \;+\; c(y,\, \phi)\right) \tag{4.1}$$

où b et c sont des fonctions différentiables et θ et ϕ des paramètres réels. θ est appelé le paramètre naturel et ϕ le paramètre de dispersion. On suppose que ϕ est connu, notons que c'est cette hypothèse qui nous permet d'affirmer que la loi de Y appartient à la famille exponentielle.

4.3 Modèles linéaire généralisés

On suppose que les variables réponses $Y_1, \dots, Y_n$ sont indépendantes et que chaque Y_i suit une loi de densité :

$$f(y_i|\theta_i, \phi) = exp\left(\frac{y_i\,\theta_i \;-\; b(\theta_i)}{\phi} \;+\; c(y_i,\, \phi)\right)$$

Notons encore que ϕ ne dépend pas de i, il est le même pour l'ensemble des données Y_i.

On pose $E(Y_i) = \mu_i$ et on suppose qu'il existe une fonction g, appelée fonction de lien, telle que :

$$g(\mu_i) = \beta x_i^T \tag{4.2}$$

où $x_i = (x_i^1, \cdots, x_i^p)$ est le vecteur de variables explicatives lié à la variable réponse Y_i et $\beta = (\beta_1, \ldots, \beta_p)$ un vecteur à estimer.

4.4 Estimation des coefficients de régression

Lorsqu'une variable aléatoire Y a une densité de la forme (4.1), on montre (voir annexe A.1) que :

$$E(Y) = b'(\theta) \text{ et } Var(Y) = b''(\theta)\,\phi \tag{4.3}$$

Il en ressort dans notre cas que μ_i dépend uniquement de θ_i et il s'en suit que la variance est le produit du paramètre de dispersion et d'une fonction de l'espérance.

Il s'agit à présent d'estimer le vecteur β et ϕ avec les observations $((y_1, x_1), \ldots, (y_n, x_n))$ par l'estimateur du maximum de vraisemblance (voir annexe A.2). On cherche donc à maximiser la log-vraisemblance donnée par :

$$l(\beta, \phi) = \sum_{i=1}^{n} \frac{y_i\,\theta_i - b(\theta_i)}{\phi} + \sum_{i=1}^{n} c(y_i, \phi)$$

(rappelons que par (4.2) et comme μ_i dépend uniquement de θ_i, on a que θ_i est une fonction de β).

On s'occupe dans un premier temps uniquement du paramètre β. Il vient, en appliquant la règle de dérivation par chaîne :

$$\frac{\partial}{\partial \beta_j}(\cdot) = \frac{\partial \theta_i}{\partial \beta_j}\,\frac{\partial}{\partial \theta_i}(\cdot) \text{ pour chaque i}$$

que :

$$\nabla l(\beta) \equiv s(\beta) = \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi}\,\nabla\theta_i(\beta) \tag{4.4}$$

Or, $g(\mu_i) = \beta x_i^T$ et comme $\mu_i = b'(\theta_i)$, on a :

$$g(b'(\theta_i)) = \beta x_i^T \tag{4.5}$$

en dérivant (4.5) par rapport à β, on obtient :

$$g'(b'(\theta_i))\; b''(\theta_i)\; \nabla\theta_i(\beta) = x_i$$

En remplaçant dans (4.4), on obtient finalement :

$$s(\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{g'(\mu_i)\,V_i} x_i^T$$

où $V_i = Var(Y_i) = \phi\, b''(\theta_i)$.

On dérive ensuite la fonction l par rapport à ϕ et on obtient le vecteur score $\begin{pmatrix} \frac{\partial l}{\partial \phi} \\ s(\beta) \end{pmatrix}$. L'estimation $(\hat{\phi}, \hat{\beta})$ est donnée en résolvant :

$$\begin{pmatrix} \frac{\partial l}{\partial \phi} \\ s(\beta) \end{pmatrix} = 0 \tag{4.6}$$

On verra par la suite que, pour la plupart des lois, $\phi = 1$ et qu'il n'a souvent pas besoin d'être estimé.

L'équation (4.6) n'a souvent pas de solution explicite. Nous sommes alors contraints de la résoudre par des méthodes numériques itératives telles que la méthode de scoring de Fisher ou l'algorithme de Newton-Raphson fermé, l'explication et le choix de la méthode à utiliser en pratique sont détaillés dans [8] mais on retiendra que le scoring de Fisher est généralement plus adapté aux *GLMs* et permet un calcul immédiat de l'intervalle de confiance des estimations finales.

On retiendra également que ces méthodes font intervenir l'inversion de la matrice $X^T X$ où $X = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^p \\ 1 & x_2^1 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & \cdots & x_n^p \end{pmatrix}$, le caractère inversible de $X^T X$ est garanti par l'hypothèse de d'indépendance des variables explicatives.

4.5 Choix de la fonction de lien

Revenons sur le choix de la fonction de lien g, c'est la fonction qui permet de lier l'espérance de Y aux variables explicatives et elle doit être strictement monotone, deux fois différentiable et définie sur le support de la variable réponse, on choisira par exemple la fonction logarithmique si l'on on veut un lien multiplicatif. Il existe néanmoins une fonction que l'on note g_* et que l'on appelle la fonction de lien canonique, elle dépend de la distribution supposée de Y et elle est de telle sorte que la résolution numérique de (4.6) soit plus précise (voir [3] pour plus de détails). On retiendra que celle-ci doit vérifier $g_*(\mu_i) = \theta_i$ et comme $\mu_i = b'(\theta_i)$, on a alors :

$$g_*(\cdot) = \left(b^{-1}\right)'(\cdot)$$

4.6 En pratique

Récapitulons les différentes étapes d'une régression par GLM :

- On choisit une loi de probabilité pour décrire Y dont la densité s'écrit de la forme (4.1)
- On choisit une fonction de lien g bijective adaptée pour décrire le lien entre l'espérance de Y et les variables explicatives, la fonction de lien canonique g_* étant un choix optimal
- On estime le vecteur $\hat{\beta}$ par maximum de vraisemblance
- L'approximation de $E(Y|X)$ est alors donnée par $g^{-1}\left(\hat{\beta}^T X\right)$

Enfin, comme $\hat{\beta}$ est estimé par maximum de vraisemblance, on a que $\sqrt{n}\left(\hat{\beta} - \beta\right)$ est approximativement de loi $\mathcal{N}_p\left(0, \mathcal{I}_1^{-1}(\beta)\right)$ où $\mathcal{I}_1(\beta)$ est la matrice d'information de Fisher pour un échantillon de taille 1 et $\mathcal{N}_p$ la loi normale dans $\mathbb{R}^p$.

Ce qui permet de construire des tests d'hypothèses asymptotiques sur la non-nullité des coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_p$ et un intervalle de confiance asymptotique pour β et donc pour l'estimation finale.

Notons que l'hypothèse d'appartenance à la famille exponentielle revisitée (équation (4.1)) de la distribution supposée de la variable réponse n'est pas une condition nécessaire à la construction d'un *GLM* comme nous le verrons dans le cas de la régression Bêta dans la partie 4.7. Cette hypothèse permet simplement de construire un cadre de construction de modèles commun à un

grand nombre de distributions de probabilité.

Nous donnons par la suite quelques exemples de GLMs qui seront utilisés dans ce mémoire pour modéliser la probabilité qu'un site assuré soit sinistré et le coût du sinistre causé par l'inondation associée.

4.6.1 Régression Gamma

On suppose ici que Y suit une loi Gamma de paramètres α et λ dont la fonction de densité définie sur $\mathbb{R}_+^*$ s'écrit :

$$f(y|\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, exp(-\lambda\, y)\, y^{\alpha-1}$$

(la fonction Γ est définie en annexe)

En notant $\mu = E(Y) = \frac{\alpha}{\lambda}$, la densité se réécrit :

$$\begin{aligned} f(y|\alpha,\mu) &= \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^\alpha exp\left(-\frac{\alpha}{\mu}\, y\right) y^{\alpha-1} \\ &= exp\left(\frac{-\frac{1}{\mu}\, y - \left(-ln\left(\frac{1}{\mu}\right)\right)}{\frac{1}{\alpha}} + c(y,\phi)\right) \end{aligned}$$

qui appartient à la famille exponentielle (forme de l'équation (4.1)) avec $\theta = -\frac{1}{\mu}$, $b(\theta) = -ln(-\theta)$, $\phi = \frac{1}{\alpha}$ et

$$c(y,\phi) = \left(\frac{1}{\phi} - 1\right) ln(y) - ln\left(\Gamma\left(\frac{1}{\phi}\right)\right) + \frac{1}{\phi}\, ln\left(\frac{1}{\phi}\right)$$

La fonction de lien canonique est la fonction inverse : $g_*(\mu) = \frac{1}{\mu}$, la fonction ln est généralement aussi utilisée.

La régression Gamma est souvent utilisée pour modéliser des variables positives comme les coûts de sinistres. Notons enfin que la distribution Gamma inclut la distribution exponentielle (en fixant $\alpha = 1$).

4.6.2 Régression logistique

On suppose ici que Y suit une loi de Bernoulli de paramètre p dont la fonction de densité définie sur $\{0,1\}$ est donnée par :

$$\begin{aligned} f(y|p) &= p^y\, (1-p)^{1-y} \\ &= exp(y\, ln(p) + (1-y)\, ln(1-p)) \\ &= exp\left(y\, ln\left(\frac{p}{1-p}\right) - ln\left(\frac{1}{1-p}\right)\right) \end{aligned}$$

et qui appartient donc à la famille (4.1) avec $\theta = ln\left(\frac{p}{1-p}\right)$, $\phi = 1$, $b(\theta) = ln(1+exp(\theta))$, $c(y,\phi) = 0$ et $g_*(\mu) = ln\left(\frac{\mu}{1-\mu}\right)$.

La régression logistique est généralement utilisée à des fins de classification où les Y_i représentent une classe d'appartenance (typiquement 0 ou 1) et l'on cherche à prédire les p_i, probabilité que Y_i appartiennent à la classe 1.

4.7 Régression Bêta

Tandis que les régressions Gaussienne et Gamma permettent de construire des modèles dont la variable réponse est supposée suivre une distribution de support infini, comment pouvons nous modéliser correctement et par des régressions des variables réponses continues de support fini (que l'on ramènera à l'intervalle $]0,1[$, ex : taux, proportions, etc) ? [18]

Il est commun dans ces cas d'avoir recours à une régression Gaussienne après avoir opéré une transformation logit ($\tilde{y} = ln(y/(1-y))$), cette méthode présente deux inconvénients majeurs, le premier réside dans le fait que les paramètres sont estimés par rapport à la moyenne de $\tilde{y}$ et non de y (la fonction logit n'tant pas linéaire), l'autre limitation vient de la non-adaptation du modèle aux données hétéroscédastiques.

Ferrari et Cibrari-Neto proposent en 2004 dans [7] d'utiliser une régression Bêta pour modéliser ce type de données à l'aide d'une paramétrisation différente de la densité de la loi Bêta. La loi Bêta n'appartenant pas à la famille exponentielle revisitée (mais appartient à la famille exponentielle), la calibration des paramètres ne suit pas la méthodologie commune présentée plus haut.

On suppose ici que Y suit une loi Bêta de paramètre p et q dont la fonction de densité définie sur $]0,1[$ est donnée par :

$$f(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)}\, y^{p-1}\,(1-y)^{q-1}$$

Que nous réécrivons, en posant $\mu = \frac{p}{p+q}$ et $\phi = p+q$:

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\,\phi)\,\Gamma((1-\mu)\,\phi)}\, y^{\mu\phi\,-1}\,(1-y)^{(1-\mu)\phi\,-1},\ 0<y<1$$

.

avec $0<\mu<1$ et $\phi>0$. On obtient alors $E(Y) = \mu$ et $Var(Y) = \frac{\mu\,(1-\mu)}{1+\phi}$, notons que ϕ représente ici le paramètre de précision (à μ fixé, plus ϕ est grand plus la variance de Y est petite) et non de dispersion (qui serait ϕ^{-1}).

En gardant toujours les mêmes notations que précédemment, le modèle de régression est posé par :

$$g(\mu_i) = \beta x_i^T \equiv \eta_i$$

et les paramètres β sont estimés par maximum de vraisemblance.

La log-vraisemblance du vecteur de données $(y_1,\cdots,y_n)$ est donnée par $l(\beta,\phi) = \sum_{i=1}^{n}\ l_i(\mu_i,\phi)$ où :

$$\begin{aligned} l_i(\mu_i,\phi) &= ln\,(\Gamma(\phi)) - ln\,(\Gamma(\mu_i\phi)) - ln\,(\Gamma((1-\mu_i)\phi)) + (\mu_i\phi-1)\,ln(y_i) && (4.7)\\ &\quad + ((1-\mu_i)\phi-1)\ ln(1-y_i) && (4.8)\end{aligned}$$

De même que dans la partie 4.4, on cherche à calculer $s(\beta) = \nabla l(\beta)$.
En observant que :

18. Nous verrons dans la partie 9.2.2 pourquoi nous avons besoin de modéliser de telles variables dans notre étude.

$$\frac{\partial l(\beta,\phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i,\phi)}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

On obtient après développement :

$$s(\beta) = \phi \sum_{i=1}^{n} \frac{(y_i^* - \mu_i^*)}{g'(\mu_i)} x_i^T$$

où :

- $y_i^* = ln\left(\frac{y_i}{1-y_i}\right)$
- $\mu_i^* = \psi(\mu_i\,\phi) - \psi((1-\mu_i)\,\phi)$ où ψ est la fonction digamma définie par

$$\psi(z) = \frac{d\,ln(\,\Gamma(z)\,)}{dz}$$

On obtient ensuite les paramètres estimés par maximum de vraisemblance $(\hat{\phi},\,\hat{\beta})$ en annulant le vecteur score $\begin{pmatrix} \frac{\partial l}{\partial \phi} \\ s(\beta) \end{pmatrix}$ par la méthode du scoring de Fisher.

Remarques pratiques

Étant en dehors du cadre des modèles linéaires généralisés, il n'existe pas de fonction de lien canonique simplifiant les calculs lors de l'estimation des paramètres et plusieurs choix sont possibles. Parmi eux, citons :

- la fonction logit : $g(\mu) = ln\left(\frac{\mu}{1-\mu}\right)$
- la fonction probit : $g(\mu) = \Phi^{-1}(\mu)$ où Φ est la fonction de répartition de la loi normale centrée réduite
- la fonction complémentaire log-log : $g(\mu) = ln\,(-ln(1-\mu))$
- la fonction log-log : $g(\mu) = -ln\,(-ln(\mu))$

En présence de valeurs nulles ou égales à 1 dans le jeu de données, Smithson et Verkuilen proposent dans [9] d'appliquer à ces valeurs la transformation [19] :

$$y' = \frac{y\,(n-1) + 1/2}{n}$$

où n est la taille du jeu de données. Ce qui revient à ajouter $\frac{1}{2n}$ aux valeurs nulles et à retrancher $\frac{1}{2n}$ aux valeurs égales à 1.

4.8 Conclusion

Même si les *GLMs* permettent de contourner les hypothèses très contraignantes du modèle linéaire classique et de modéliser de manière plus précises certains aléas, ils imposent néanmoins de faire des hypothèses sur la distribution de la variable réponse.

19. Le choix de cette transformation est empirique et s'inspire de la théorie de détection du signal, le biais qui en découle est négligeable tant que le nombre de telles données reste peu significatif. Le lecteur intéressé trouvera plus de détails dans la page personnelle de M. Smithson www.michaelsmithson.online/stats/betareg/Readme.pdf et dans MacMillan and Creelman (2005) :"Detection Theory : A User's Guide", pp 8-9 .

Nous présentons par la suite des techniques de régression et de classification par *machine learning* qui sont plus flexibles en termes d'hypothèses et qui sont de plus en plus accessibles avec le développement des outils informatiques.

5 Mesures d'erreurs

On présente dans cette partie la méthode qui sera utilisée pour mesurer la précision des différentes estimations qui seront faites dans ce mémoire. On note toujours Y la variable à estimer, $X = (X^1, \ldots, X^p)$ le vecteur des variables explicatives et $\hat{F}$ la fonction permettant d'estimer Y à partir de X.

5.1 Sur-apprentissage (*overfitting*)

L'overfitting correspond au cas où la fonction $\hat{F}$ devient trop spécialiste du jeu de données qui a servi à sa calibration et qui aura du mal à faire des prédictions sur des données différentes.

Ainsi, on a besoin de calculer l'erreur de prédiction de $\hat{F}$ sur un jeu de données qui n'a pas servi à sa construction afin de mesurer sa capacité à faire des prédictions sur de nouvelles données. On présente ici la méthode du K-fold cross-validation qui sera utilisée pour calibrer les différents modèles présentés dans ce mémoire :

- on divise le jeu de données initial en K sous-jeux différents
- on réalise K calibrations, la k ème calibration utilise $K-1$ sous-jeux (appelés jeux d'apprentissage) pour calculer $\hat{F}$ et le dernier sous-jeu (jeu de validation) est utilisé pour mesurer l'erreur E^k
- l'erreur de validation (ou de généralisation) est donnée par :

$$E_{validation} = \sum_{i=1}^{K} E^k$$

On s'intéresse maintenant aux différentes mesures qui peuvent être utilisées pour calculer les E^k.

5.2 Mesures d'erreur

Plusieurs mesures d'erreur en modélisation sont proposées dans la littérature statistique, on retient, dans ce mémoire, celle de la déviance. Mesurer l'erreur d'une estimation implique d'analyser les résidus, c'est-à-dire l'écart entre les valeurs estimées $\hat{y}_i$ et les valeurs observées y_i. La déviance offre plus de souplesse dans la manière de prendre en compte les écarts résiduels que la traditionnelle mesure des erreurs quadratiques (voir [10]).

5.2.1 Statistique de déviance

La déviance est l'écart entre la log-vraisemblance obtenue en les points estimés et celle obtenue avec un modèle saturé (c'est à dire un modèle parfait qui estimerait correctement tous les points), elle correspond aussi à la somme des carrés des résidus de déviance :

$$D = 2\,\phi \sum_{i=1}^{n} |l(y_i) - l(\hat{y}_i)|$$

où ϕ est le paramètre de dispersion présenté dans la partie 4.2 et

$$l(\lambda) = ln\left(f_Y(y_i;\lambda)\right) \text{ et } \hat{y}_i = \hat{F}(x_i)$$

où f_Y désigne la densité de la distribution choisie pour décrire Y, $l(\lambda)$ désigne ainsi la log-vraisemblance de la variable Y en y_i en supposant une moyenne λ.

On développe ici l'expression de la déviance pour quelques lois de probabilités dont la densité s'écrit sous la forme (4.1) :

5.2.1.1 Déviance Gaussienne

On suppose que Y_i suit une loi normale de moyenne μ_i et de variance σ^2. La fonction de densité est alors donnée par :

$$\begin{aligned} f_{Y_i}(y_i) &= \frac{1}{\sigma\sqrt{2\pi}}\, exp\left(-\frac{(y_i-\mu_i)^2}{2\,\sigma^2}\right) \\ &= exp\left(\frac{y_i\,\mu_i - \frac{1}{2}\,\mu_i^2}{\sigma^2} \;+\; c(y_i,\sigma^2)\right) \end{aligned}$$

On identifie alors le paramètre $\phi = \sigma^2$ et on a :

$$l_{Gaussienne}(y_i) = \frac{y_i^2 - \frac{1}{2}\,y_i^2}{\sigma^2} + c(y_i,\sigma^2)$$

et

$$l_{Gaussienne}(\hat{y}_i) = \frac{y_i\,\hat{y}_i - \frac{1}{2}\,\hat{y}_i^2}{\sigma^2} + c(y_i,\sigma^2)$$

D'où

$$\begin{aligned} D_{Gaussienne}(y_i,\hat{y}_i) &= 2\,\phi\sum_{i=1}^{n} 2|l_{Gaussienne}(y_i) - l_{Gaussienne}(\hat{y}_i)| \\ &= 2\,\sigma^2\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{2\sigma^2} \\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \end{aligned}$$

On remarque que l'on tombe sur l'erreur quadratique qui est une mesure très utilisée en statistique car les données souvent supposées Gaussiennes.

5.2.1.2 Déviance de Gamma

On suppose que Y_i suit une loi Gamma de moyenne μ_i et de variance $\frac{\mu_i^2}{\alpha}$. On rappelle la fonction de densité déjà développée dans la partie 4.6.1 :

$$f_{Y_i}(y_i) = exp\left(\frac{-\frac{1}{\mu}\,y - \left(-ln\left(\frac{1}{\mu}\right)\right)}{\frac{1}{\alpha}} + c(y,\phi)\right)$$

.

D'où :

$$l_{Gamma}(y_i) = \frac{-\frac{y_i}{y_i} - ln(y_i)}{\phi} + c(y_i,\alpha)$$

et

$$l_{Gamma}(\hat{y}_i) = \frac{-\frac{y_i}{\hat{y}_i} - ln(\hat{y}_i)}{\phi} + c(y_i,\alpha)$$

D'où

$$\begin{aligned} D_{Gamma}(y_i, \hat{y}_i) &= 2\,\phi \sum_{i=1}^{n} 2|l_{Gamma}(y_i) - l_{Gamma}(\hat{y}_i)| \\ &= 2\,\phi \sum_{i=1}^{n} \left| \frac{\frac{y_i - \hat{y}_i}{\hat{y}_i} - ln\left(\frac{y_i}{\hat{y}_i}\right)}{\phi} \right| \\ &= 2 \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} - ln\left(\frac{y_i}{\hat{y}_i}\right) \right| \end{aligned}$$

5.2.1.3 Déviance de Bernoulli

On suppose que Y_i suit une loi de Bernoulli, on a vu dans la partie 4.6.2 que $\phi = 1$, on obtient facilement :

$$\begin{aligned} D_{Bernoulli}(y_i, \hat{y}_i) &= 2\,\phi \sum_{i=1}^{n} |l_{Bernoulli}(y_i) - l_{Bernoulli}(\hat{y}_i)| \\ &= 2 \sum_{i=1}^{n} \left| y_i ln\left(\frac{y_i}{\hat{y}_i}\right) - (1 - y_i)\, ln\left(\frac{1 - y_i}{1 - \hat{y}_i}\right) \right| \end{aligned}$$

5.2.1.4 Déviance de Bêta

À partir de l'expression de la log-vraisemblance de la loi Bêta explicitée dans l'équation 4.8, on obtient (ϕ représente ici le paramètre de précision et non de dispersion) :

$$\begin{aligned} D_{Beta}(y_i, \hat{y}_i) &= 2\,\phi^{-1} \sum_{i=1}^{n} |l_{Beta}(y_i, \phi) - l_{Beta}(\hat{y}_i, \phi)| \\ &= 2\,\phi^{-1} \sum_{i=1}^{n} \left| ln\left(\frac{\Gamma(\hat{y}_i\,\phi)\;\Gamma((1 - \hat{y}_i)\,\phi)}{\Gamma(y_i\,\phi)\;\Gamma((1 - y_i)\,\phi)}\right) + \phi\;(y_i - \hat{y}_i)\; ln\left(\frac{y_i}{1 - y_i}\right) \right| \end{aligned}$$

La valeur de ϕ utilisée est celle obtenue par l'estimation des paramètres du *GLM*.

5.2.2 Ratio de l'erreur de validation

Dans ce mémoire, les E^k seront calculés par la statistique de déviance. Seulement, l'erreur de validation ne permet pas à elle seule de juger la qualité d'un modèle, on a besoin de la comparer à la déviance propre du jeu de données initial. On calcule alors le rapport de la déviance du modèle et de la déviance originale :

$$R_{validation} = \frac{E_{validation}}{D(y_i, \bar{y}_n)}$$

où $\bar{y}_n$ désigne la moyenne des $y_1, \ldots, y_n$.

Le mieux, bien sûr, est d'avoir un rapport proche de 0. En pratique, on est satisfait lorsque le rapport est inférieur à 30%.

5.2.3 Mesure pseudo-R²

Dans le même esprit que le ratio de validation, le pseudo R^2 mesure un ratio de déviance mais qui ne porte pas sur des jeux de validation, il porte sur les données ayant servi à l'apprentissage. Il permet de se faire une idée sur la capacité des variables explicatives à expliquer les variations de la variable réponse, on parle même de déviance expliquée (ou de variance expliquée lorsqu'il s'agit de la déviance Gaussienne). Il est défini par :

$$R^2 = 1 - \frac{D(y_i, \hat{y}_i)}{D(y_i, \bar{y}_n)}$$

Là aussi, le mieux est d'avoir un R^2 proche de 1 et on est généralement satisfait à partir de 70%. Cette mesure est également appelée le pseudo R^2 de McFadden. Enfin, on parlera également de déviance expliquée sur le jeu de validation en faisant référence au complémentaire de l'erreur de validation.

5.2.4 Erreur de Laplace

En plus de la déviance Gaussienne et de Gamma, on peut également utiliser l'erreur absolue (aussi appelée erreur de Laplace) comme mesure dans une régression, elle est définie par :

$$E_{Laplace}(y_i, \hat{y}_i) = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

5.2.5 Courbe ROC pour la classification

En classification, en plus de la déviance de Bernoulli, il existe une autre mesure souvent utilisée en machine learning et qui est celle retenue dans ce mémoire, il s'agit de l'aire sous la courbe *ROC* (*Receiver Operating Characteristic*). Avant de définir la courbe *ROC*, nous définissons dans un premier temps les ratios de vrais et faux positifs de seuil s :

À partir des probabilités $\hat{p}_i$ prédites par le modèle, on définit un seuil s qui classe y_i comme positif lorsque $\hat{p}_i \geq s$ et négatif sinon.

On calcule le Ratio de Vrais Positifs (RVP ou sensibilité) de seuil s donné par :

$$\begin{aligned} RVP_s &= \frac{\text{nombre de bonnes estimations positives}}{\text{nombre de donnees positives}} \\ &= \frac{\sum_{i=1}^{N} 1_{\{\hat{p}_i \geq s\}} \times 1_{\{y_i = 1\}}}{\sum_{i=1}^{N} 1_{\{y_i = 1\}}} \end{aligned}$$

et le Ratio de Faux Positifs (RFP ou complémentaire de la spécificité) de seuil s donnés par :

$$\begin{aligned} RFP_s &= \frac{\text{nombre de fausses estimations positives}}{\text{nombre de donnees negatives}} \\ &= \frac{\sum_{i=1}^{N} 1_{\{\hat{p}_i \geq s\}} \times 1_{\{y_i = 0\}}}{\sum_{i=1}^{N} 1_{\{y_i = 0\}}} \end{aligned}$$

La courbe ROC est alors obtenue en traçant les points (RFP_s, RVP_s) pour différents seuils s et l'aire est calculée par des méthodes numériques classiques de calcul d'intégrales (formule d'Euler ou somme de Riemann par exemple).

Interprétation

Cette mesure présente deux avantages majeurs par rapport à la mesure classique du taux de prédiction :

- Le premier réside dans l'interprétation de cette mesure : A. Hanley et J. McNeil montrent dans [25] l'équivalence avec la statistique de Wilcoxon qui estime la probabilité que le modèle classera dans le bon ordre une observation positive et une observation négative choisies de manière aléatoire. Formellement, on a :

$$AUC = P_{I,J}\left(\hat{p}_I > \hat{p}_J\right)$$

où, I est la variable aléatoire représentant un indice parmi les indices des observations positives,
J est la variable aléatoire représentant un indice parmi les indices des observations négatives.

Ainsi, cette mesure juge de la capacité du modèle à distinguer deux sites assurés, l'un étant plus vulnérable que l'autre. Cette mesure paraît alors plus pertinente dans notre étude. En effet, nous observons dans nos données des réalisations d'une variable aléatoire suivant une loi de Bernouilli et non le paramètre de celle-ci (représentant la probabilité d'être sinistré). De plus, le fait qu'un site ait été sinistré ne signifie pas que la probabilité qu'il le soit est supérieur à 1/2. Ainsi, il n'est pas pertinent de chercher à comparer les classes observées avec les probabilités prédites ni avec les classes prédites.

- Par ailleurs, cette mesure est insensible à la disproportion entre les classes positives et négatives dans le jeu de données (comme c'est le cas dans le notre). Exemple : dupliquer les observations négatives dans la base de test doublera le nombre de fausses estimations positives laissant ainsi le RFP constant. Cette propriété est rarement vérifiée par les mesures classiques telles que le taux de prédiction qui est très sensible à la distribution des classes dans le jeu de données servant à la mesure [20].
  Enfin, notons qu'un classificateur aléatoire aura pour courbe ROC la droite d'équation $y = x$ et donc une aire sous la courbe égale à 1/2. Ainsi, il n'existe pas de modèle raisonnable avec une AUC inférieure à 1/2.

20. T. Fawcett présente dans [26] plus en détails les propriétés de cette mesure et son usage pratique en *machine learning*

6 Apprentissage automatique (*machine learning*)

L'être humain semble capable d'apprendre et d'intégrer de nouveaux concepts sans aménagement spécifique. Le *machine learning* consiste à automatiser cette capacité. En d'autres termes, cela consiste à développer un algorithme qui prend en entrée un flux de données et qui soit capable d'en extraire différentes informations pour apprendre de nouveaux concepts en vue de faire des prédictions, et ce quelle que soit la nature du concept à apprendre, sans programmation spécifique et en un temps raisonnable (i.e. de complexité polynomiale).

En actuariat, le *machine learning* est utilisé dans un but statistique (certains parlent d'ailleurs de *statistical learning*), concrètement il s'agit de s'en servir pour des travaux de régression et de classification. De même que pour les *GLMs*, il s'agit dans notre cas d'utiliser les données à notre disposition sur les évènements historiques d'inondation, la sinistralité associée et l'exposition actuelle d'AXA afin de prédire le coût de ces évènements s'ils venaient à se reproduire.

Le cadre de ce mémoire ne permet pas de faire une description détaillée des différentes techniques et applications du *machine learning* qui constitue un domaine vaste de l'informatique. Nous présentons succinctement quelques généralités nécessaires à l'introduction des méthodes qui ont été choisies pour la modélisation et qui seront présentées dans les sections 7 et 8.

6.1 Généralités

Auteur d'un article pionnier en *machine learning*, L.G. Vailiant caractérise en 1984 dans [13] de manière formelle les concepts apprennables (*learnable*) par un algorithme (dans le sens *machine learning* défini en introduction). Le cadre formel qu'il pose concerne les problèmes de classification mais on fait l'hypothèse raisonnable qu'un algorithme a les mêmes propriétés qu'il soit utilisé pour une classification ou pour une régression.

Nous distinguons deux types de prédicteurs : les prédicteurs faibles (*weak learner*) et les prédicteurs forts (*strong learner*) : nous disposons de vecteurs de variables explicatives $x_1, \dots, x_n$ générés selon une distribution D et d'une fonction concept c qui associe une étiquette $y = c(x)$ (typiquement 0 ou 1 en classification), le but de l'algorithme est, à partir d'un certain jeu de données $\{(x_1, y_1), \dots, (x_n, y_n)\}$, "d'apprendre" la fonction c et de l'approximer par une fonction h qu'on appelle hypothèse et qui servira à faire des prédictions sur des données qui n'ont pas servi à l'apprentissage. L'algorithme doit fonctionner quelque soit la distribution D. Un *weak learner* doit produire une fonction h avec une erreur légèrement inférieure à celle d'un prédicteur aléatoire (qui se trompe en moyenne une fois sur deux), un *strong learner* doit avoir une erreur qui soit faible, nous en donnons ici une définition formelle :

- L'erreur de l'hypothèse $h(x)$ produit par l'algorithme est donnée par :
$$err_{c,D} = P_{x\sim D}\left(\, h(x) \neq c(x) \,\right)$$
- Un algorithme $A(\epsilon, \delta)$ produisant une fonction hypothèse h est appelé ***strong learner*** pour une classe de fonctions concept H si pour toute distribution D, on a :
$$\forall\, \epsilon > 0,\, \delta < \frac{1}{2},\, c \in H,\; P_D(\, err_{c,D} \leq \epsilon\,) > 1 - \delta$$
De plus la complexité de A doit être polynomiale en $\frac{1}{\epsilon}$, $\frac{1}{\delta}$ et $|x|$ où $|x|$ désigne la longueur du vecteur x.
- Un algorithme $A(\eta, \delta)$ produisant une fonction hypothèse h est appelé ***weak learner*** pour une classe de fonctions concept H si pour toute distribution D, on a :
$$\forall\, 0 < \eta < \frac{1}{2},\, \delta < \frac{1}{2},\, c \in H,\; P_D(\, err_{c,D} \leq \frac{1}{2} - \eta\,) > 1 - \delta$$

De plus la complexité de A doit être polynomiale en $\frac{1}{\delta}$ et $|x|$.

On remarque bien sûr que la complexité augmente quand on cherche à réduire la probabilité d'erreur et que les *strong learners* sont destinés à avoir des complexités supérieures aux *weak learners* (la complexité d'un *strong learner* est en plus croissante avec $\frac{1}{\epsilon}$).

En 1988 et 1989, M.J. Kearns et L.G. Vaillant se posent la question de l'équivalence entre les *weak learners* et les *strong learners* dans [15] et [16], le terme de *boosting* est alors utilisé pour décrire le passage d'un *weak learner* à un *strong learner*. R.E. Shapire y répond dans [17] où il propose un premier algorithme de *boosting* et démontre ainsi l'équivalence entre les deux types de prédicteurs, par équivalence on signifie qu'un problème pouvant être résolu par un *weak learner* peut l'être par un *strong learner* et vice-versa. En 1995, Y. Freund et R.E. Shapire mettent au point AdaBoost (*Adaptive boosting*) dans [18] qui est le premier algorithme de *boosting* performant et qui permettra à ses auteurs de remporter le prestigieux prix Gödel en 2003.

6.2 Arbres de régression

Nous présentons les arbres de régression et de classification qui sont considérés comme des *weak learners* (on nuancera ce propos un peu plus bas) et qui sont souvent utilisés en *machine learning*. Les arbres de régression et de classification sont apparus en 1984 et font suite aux travaux de L. Breiman dans [19].

Concentrons-nous maintenant sur la construction de l'arbre, il s'agit de séparer les données selon les valeurs de leurs variables explicatives. A chaque nœud de l'arbre, on cherche la variable explicative et le seuil optimaux qui permettront de séparer les données en deux groupes différents, le choix optimal est celui qui maximise le gain de variance :

- pour un arbre de régression, minimiser l'erreur des moindres carrés intra groupe des nœuds fils obtenus :

$$RSS = \sum_{y \in \text{gauche}} (y - \bar{y}_G)^2 \; + \sum_{y \in \text{droit}} (y - \bar{y}_D)^2 \tag{6.1}$$

où $\bar{y}_G$ désigne la moyenne des observations du noeud gauche
$\bar{y}_D$ désigne la moyenne des observations du noeud droit

- pour un arbre de classification à K classes, minimiser le critère de Gini défini par :

$$Gini = n_G \sum_{k=1,\dots,K} p_{k,G}\,(1 - p_{k,G}) \; + n_D \sum_{k=1,\dots,K} p_{k,D}\,(1 - p_{k,D}) \tag{6.2}$$

où $p_{k,G}$ désigne la proportion de classe k dans le noeud gauche
$p_{k,D}$ désigne la proportion de classe k dans le noeud droit
n_G désigne le nombre d'observations dans le noeud gauche
n_D désigne le nombre d'observations dans le noeud droit

Chaque feuille de l'arbre ainsi obtenu est caractérisée par un sous ensemble des valeurs pouvant être prises par les variables explicatives. La grandeur modélisée (*coût du sinistre par exemple*) (ou la classe pour un arbre de classification) associée à chaque feuille est estimée par la moyenne sur les observations du nœud (pour un arbre de classification, c'est soit la classe majoritaire à l'intérieur de la feuille soit les probabilités empiriques d'appartenance à chacune des classes).
Afin de limiter le sur-apprentissage (*overfitting*), nous pouvons contrôler le nombre de données

minimum à avoir à l'intérieur de chaque nœud et la profondeur maximum de l'arbre (qui détermine le nombre maximum de feuilles puisque c'est un arbre binaire) ; le pire cas d'overfitting correspond à un arbre avec autant de feuilles que de données initiales (et forcément une seule observation par feuille). Afin de ne pas perdre en temps de calcul, on fixe aussi un minimum de gain de variance (appelé paramètre de complexité dans le package rpart sous R) en deçà duquel on arrête le développement de l'arbre, cela permet de ne pas faire de calculs pour des gains de variances négligeables.

Le nombre de paramètres donne une certaine flexibilité aux arbres en terme de précision et de temps de calcul et c'est dans cette mesure que l'on ne peut pas toujours affirmer sans nuance qu'il s'agit d'un *weak learner*, il faut s'assurer que les paramètres choisis impliquent une complexité raisonnable, il est néanmoins communément admis qu'un arbre de profondeur 1 est un *weak learner*.

Exemple :

Considérons l'exemple du coût des sinistres MRH liés à un évènement d'inondation, pour des raisons de simplicité, nous nous limiterons à 3 variables explicatives : la valeur assurée du bien, la hauteur d'eau et la distance à la cellule inondée la plus proche.

1. On commence par chercher la variable explicative et le seuil qui permettent de faire une première séparation optimale, on trace (voir figure 6.1) pour chaque variable explicative la courbe *RSS* en faisant varier le seuil de décision.

   La variable explicative qui maximise le gain de variance ici est la distance à la cellule inondée la plus proche, le seuil associé est de 3.71 *kms*, ce qui nous permet d'initialiser l'arbre de régression (voir figure 6.2).

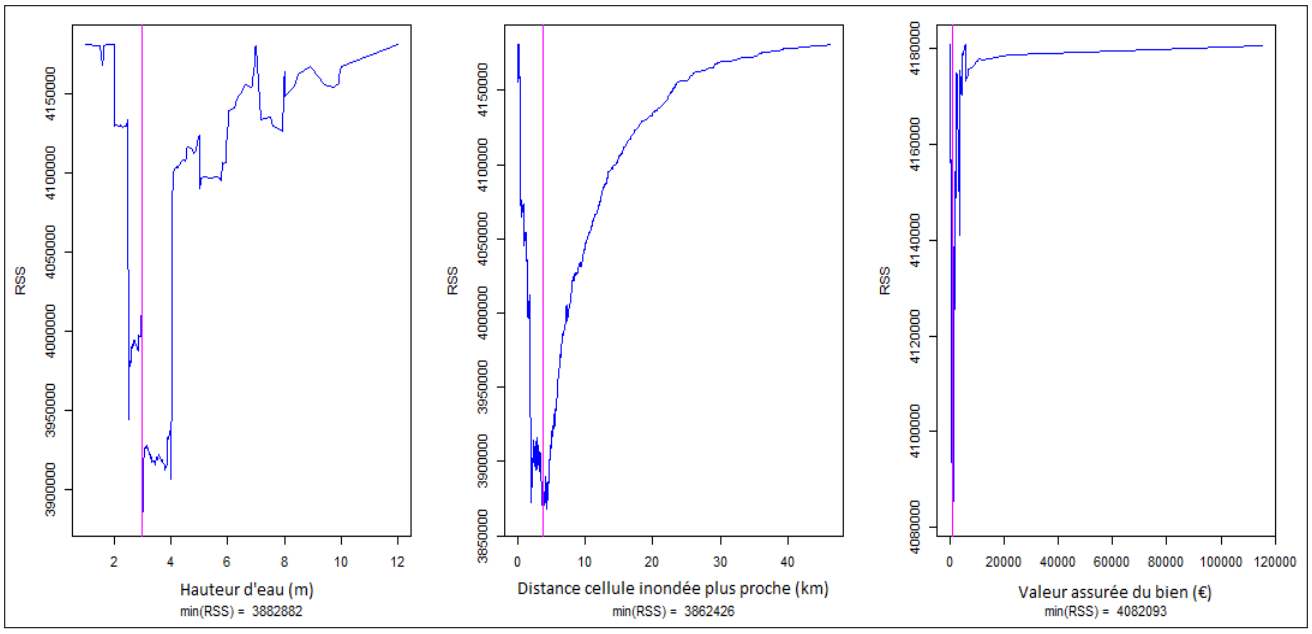

FIGURE 6.1 – Exemple de construction : séparation optimale pour la racine de l'arbre

2. On cherche ensuite le choix optimal pour le nœud gauche (les données considérées maintenant sont uniquement celles qui appartiennent au nœud gauche, c'est-à-dire dont la variable distance

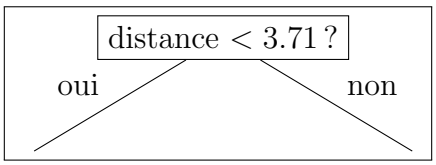

FIGURE 6.2 – Exemple de construction : initialisation de l'arbre

est inférieure à 3.71 kms) (voir figure 6.3)

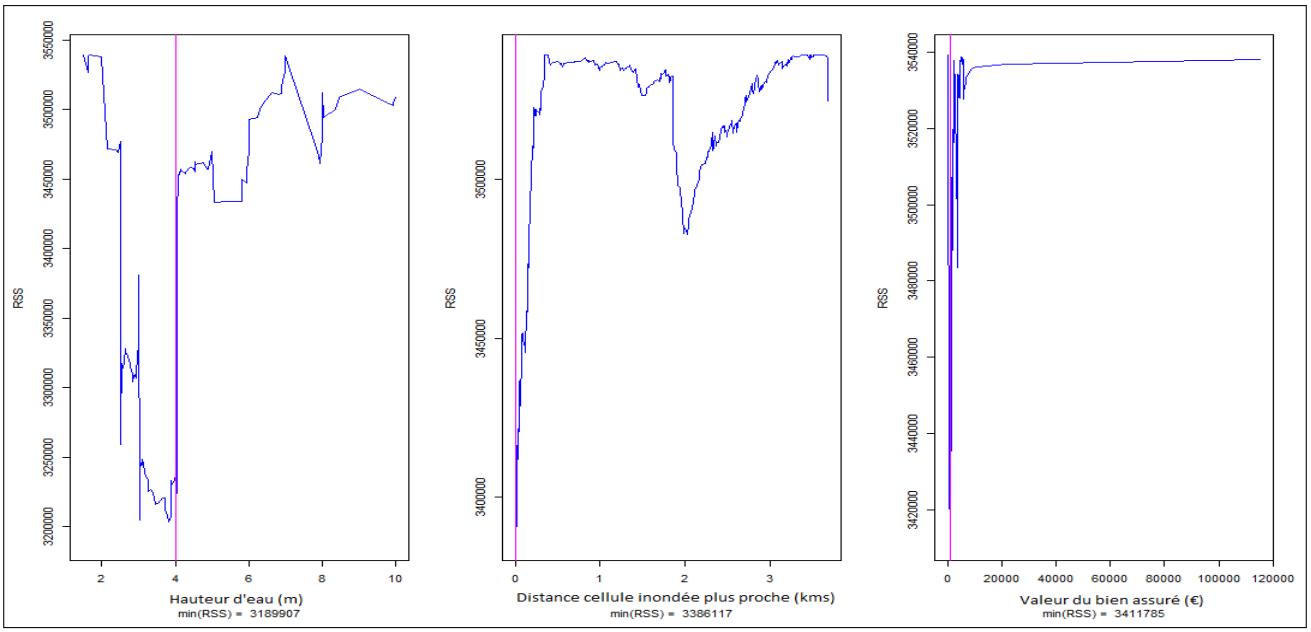

FIGURE 6.3 – Exemple de construction : séparation optimale pour la première branche gauche de l'arbre

3. On continue ainsi jusqu'à obtenir la complexité de l'arbre souhaitée, par exemple pour une complexité de profondeur 2, on obtient l'arbre de régression visible sur la figure 6.4. Le nombre de sinistres dans chaque feuille est calculé en faisant une moyenne sur l'ensemble des données vérifiant les critères de la feuille.

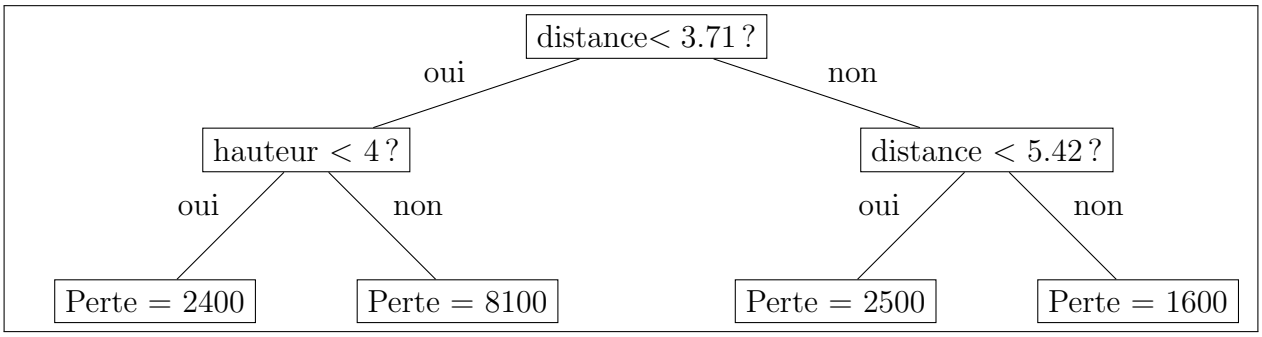

FIGURE 6.4 – Exemple de construction : arbre de régression avec une complexité de profondeur 2

6.3 Apprentissage ensembliste

On revient ici sur le *boosting* introduit plus haut, on parle aussi d'apprentissage ensembliste (*ensemble learning*). Concrètement, on génère plusieurs *weak learner*s (typiquement les arbres présentés plus haut), chacun utilisant uniquement une partie des données. Les différents *weak learner*s produits sont combinés pour réduire la variance et le biais des prédicteurs (voir section 4), le but étant que le prédicteur obtenu en sortie soit un *strong learner*.

Pour que le *boosting* ait des chances de réussir, il faut que les prédicteurs de base soient complémentaires, c'est-à-dire qu'ils apportent des informations complémentaires qui, en les combinant, donnent lieu à une information plus précise. C'est pour cela que chaque prédicteur utilise un jeu de données différent des autres. Il faut aussi que le temps de calcul de chaque prédicteur soit très court afin que la combinaison des prédicteurs puisse se faire en un temps raisonnable.

On distingue deux manières de faire, produire plusieurs arbres de manière indépendante avant de les combiner (on appellera cette méthode la méthode parallèle) ou construire les arbres de manière séquentielle en tenant compte à chaque construction de l'apprentissage réalisé par les arbres précédents (méthode séquentielle). Dans ce mémoire, nous utiliserons chacune des deux méthodes avec les forêts aléatoires pour la méthode parallèle et le gradient *boosting* qui est une amélioration de l'algorithme Adaboost pour la méthode séquentielle. Ces méthodes sont décrites dans les sections 7 et 8 et permettront de se faire une idée plus concrète de l'apprentissage ensembliste.

6.4 Conclusion

L'avantage de ces méthodes par rapport aux méthodes classiques est qu'il n'est pas besoin d'émettre d'hypothèse sur la distribution des variables réponses ou sur le lien avec les variables explicatives ; la seule hypothèse émise est que les variables réponses sont indépendantes et générées de manière identique à partir du vecteur des variables explicatives. Ainsi, un arbre de prédiction permet de capturer des interactions non linéaires entre les différentes variables et ne garde que les variables explicatives les plus importantes. Ces techniques requièrent néanmoins un temps de calcul plus conséquent que les techniques usuelles mais les dernières avancées en informatique permettent de pallier cet obstacle.

Poser un cadre mathématique précis autour du *machine learning* est difficile et dépasse le cadre de ce mémoire, ainsi on ne montre pas qu'un arbre de régression est un *weak learner*, mais partant du principe que l'on peut améliorer sa performance et au vu de sa faible complexité, on peut dire que ce n'est pas un *strong learner* ; on ne montre pas non plus que l'on peut toujours transformer des *weak learners* en *strong learners*.

7 Forêts aléatoires (*random forest*)

Nous présentons ici la première méthode d'apprentissage ensembliste utilisée dans ce mémoire : les forêts aléatoires.

7.1 Définition

Les forêts aléatoires sont constituées d'une collection d'arbres de régression ou de classification, elles sont un cas particulier du bagging[20] et ont été introduites par L. Breiman en 2001 dans [21]. C'est une méthode de *boosting* parallèle utilisant comme weak learner des arbres de prédiction où chaque arbre est construit sur la base d'un échantillon aléatoire de l'ensemble des données et sur une sélection aléatoire des variables explicatives. Les prédictions finales sont obtenues en faisant une moyenne sur l'ensemble des arbres de la forêt (ou en prenant le vote majoritaire pour la classification). La construction d'une forêt est décrite dans l'algorithme 1.

Algorithme 1 : Algorithme de construction d'une forêt aléatoire

Données : *donnees* : données initiales
nbArbres : nombres d'arbres à développer dans la forêt
nbFeuilles : nombre de feuilles maximum par arbre
minNbObs : nombre minimum de données dans chaque feuille de l'arbre
M : nombre de variables explicatives
m : nombre de variables explicatives à sélectionner à chaque nœud

```
début
    arbres ← liste de longueur nbArbres
    pour n allant de 1 à nbArbres faire
        donnee_tmp ← échantillon de données de même taille que donnees
        arbres[n] ← arbre de prédiction basé sur les données donnees_tmp
        tant que arbres[n] n'a pas atteint la complexité nbFeuilles et minNbObs faire
            développer un nouveau nœud de la manière suivante :
                échantillonner m variables explicatives parmi les M variables
                faire la séparation optimale sur les m variables sélectionnées
            ajouter le noeud à arbres[n]
        fin
    fin
    combiner les éléments de arbres en faisant une moyenne pour la régression et un vote
    pour la classification
fin
```

Ainsi, la fonction hypothèse d'une forêt aléatoire est $h(x) = \frac{1}{n} \sum_{k=1}^{n} h_k(x)$ où les $h_k(x)$ sont les fonctions hypothèses des arbres de prédiction qui constituent la forêt.

7.2 Erreur de validation et convergence d'une forêt

En *machine learning*, l'erreur de validation est une fonction permettant de mesurer à quel point la machine est capable de faire des prédictions sur des données qui n'ont pas servi à l'apprentissage ou à la calibration.

Un seul arbre de prédiction peut être vu comme une observation particulière qui utilise une seule vision de l'ensemble des données et en utilisant toutes les variables explicatives disponibles. L'idée derrière les forêts aléatoires est de combiner plusieurs de ces points de vues en utilisant

différents échantillons de données et différents jeux de variables explicatives, chaque arbre étant une observation d'un tirage aléatoire "de points de vues". On obtient ensuite l'information complète en faisant la combinaison de tous ces points de vues et c'est ainsi que l'on espère améliorer la précision de notre estimations. On s'intéresse ici à la convergence des forêts aléatoires qui est due à la loi des grands nombres :

Étant donnés k arbres de prédictions $h_1(x), h_2(x), \ldots, h_K(x)$ qui constituent une forêt aléatoire, y une variable réponse et x le vecteur des variables explicatives,

- Pour la classification, on définit la fonction marge par :

$$mg(x,y) = \frac{1}{K}\sum_k I\left(h_k(x) = Y\right) - max_{j\neq Y}\left(\frac{1}{K}\sum_k I(h_k(x) = j)\right)$$

où $I(.)$ est la fonction indicatrice. La fonction marge mesure à quel point le vote moyen pour la bonne classe excède le vote moyen pour n'importe quelle autre classe, plus la marge est grande plus on a de confiance dans la classification.

Ainsi l'erreur de validation de l'ensemble de la forêt est donnée par :

$$EV = P_{X,Y}\left(mg(X,Y) < 0\right)$$

l'indice X, Y indique que la probabilité porte sur l'espace probabilisé de l'ensemble des données x, y.

On introduit maintenant la variable aléatoire Θ qui détermine l'échantillonnage du jeu de données et le choix des variables explicatives pour chaque arbre. Ainsi, on note dorénavant $h_k(X) = h(X, \Theta_k)$ car l'arbre k est entièrement déterminé par le tirage de Θ.

L'erreur de validation de la forêt converge lorsque le nombre d'arbres tend vers l'infini, par la loi forte des grands nombres, vers :

$$P_{X,Y}\left(P_\Theta\left(\, h(X,\Theta) = Y\,\right) - max_{j\neq Y}\left(P_\Theta(\, h(X,\Theta) = j)\,\right) < 0\right)$$

- Pour la régression, l'erreur dépend de la mesure d'erreur choisie que l'on notera L (le choix de la mesure d'erreur est un problème différent, celui-ci est développé dans la section 8), elle peut être l'erreur des moindres carrés, la déviance de Poisson, etc, et elle dépend de la nature des données à modéliser.
Avec les mêmes notations que pour la classification, l'erreur de validation de la forêt est donnée par :

$$E_{X,Y}\left(L\left(Y, \frac{1}{K}\sum_k h(X,\Theta_k)\right)\right)$$

qui converge quand le nombre d'arbres tend vers l'infini, par la loi forte des grands nombres vers

$$E_{X,Y}\left(\, L\left(Y, E_\Theta\left(h(X,\Theta)\right)\right)\,\right)$$

On en conclut que l'erreur d'une forêt aléatoire converge vers l'erreur espérée d'un arbre de prédiction. On a aussi montré au passage que le nombre d'arbres ne génère pas d'overfitting

puisque l'erreur converge vers une quantité finie : intuitivement, chaque arbre est tiré de manière indépendante selon le même aléa (c'est d'ailleurs l'hypothèse de la loi des grands nombres) et n'utilise pas l'information construite par les autres arbres, aucun overfitting ne peut ainsi être généré. Nous pouvons visualiser cette propriété sur la figure 7.1.

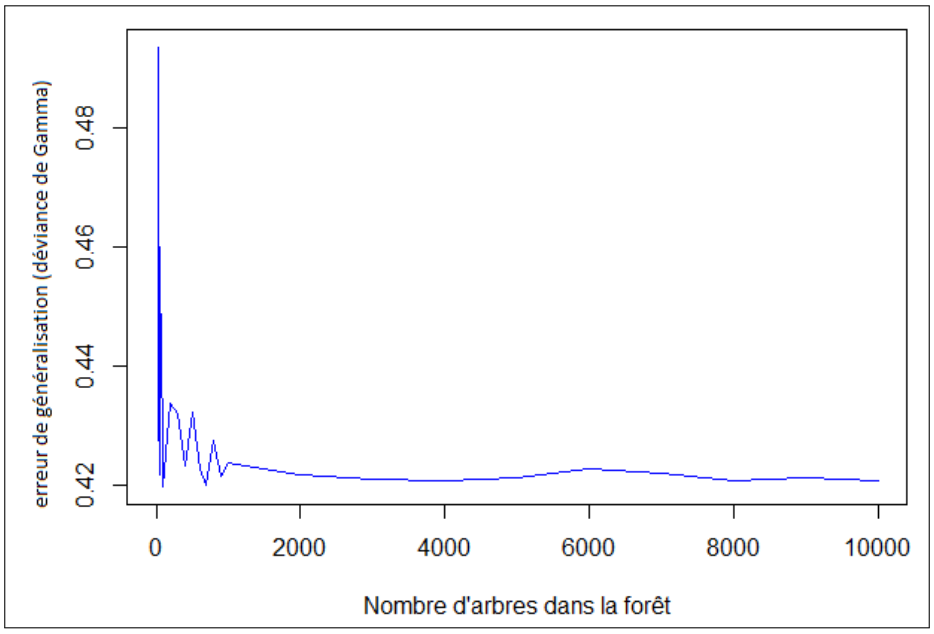

FIGURE 7.1 – Convergence des forêts aléatoires, exemple avec le coût des sinistres en MRH

Jeu de validation

L'autre grand avantage des forêts aléatoires est qu'il n'est pas nécessaire de consacrer une partie des données à l'estimation de l'erreur de validation : en effet, comme chaque arbre est construit sur un échantillon du jeu de données, chaque donnée a une probabilité de 37% (démonstration en annexe) de ne pas appartenir à l'échantillon et de ne pas servir à la construction de l'arbre. L'ensemble des données n'ayant pas servi à la construction d'un arbre est appelé *out-of-bag data ou OOB data*.

Chaque donnée est ensuite prédite uniquement avec les arbres dont elle n'a pas servi à la construction (en moyenne 37% des arbres) : pour une donnée (x, y), la prédiction est donnée par

$$\tilde{h}(x) = \text{moyenne}\,\{h_k(x)\ /k\ tel\ que\ x \in OOB_k\}$$

On calcule l'erreur sur ces prédictions et on la considère comme étant l'erreur de validation (on l'appelle aussi erreur *out-of-bag*).

Cette méthode présente l'avantage d'économiser du temps de calcul par rapport à la cross-validation qui nécessite de calculer une forêt par fold pour estimer l'erreur en plus du calcul de la forêt finale avec l'ensemble des données alors qu'un seul calcul de forêt suffit à tout faire dans notre cas. Aussi la cross-validation peut mener à des interprétations trompeuses car le faible nombre de folds (généralement 5 ou 10 car limité par le temps de calcul) peut mener à des découpages particuliers, on peut se retrouver par exemple avec un sous ensemble ne contenant que les plus grandes valeurs dont l'écart est significatif avec les autres données et se retrouver avec une grande erreur, la méthode présentée ici permet de travailler avec beaucoup plus de sous ensembles et de gommer ces éventuels "découpages malheureux".

7.3 Importance des variables explicatives

Pour gagner en lisibilité et proposer une mesure qualitative sur l'interaction des variables explicatives, Breiman propose dans son article [21] de mesurer l'importance des variables en étudiant la sensibilité, sur chaque arbre, de l'erreur OOB par rapport aux variables explicatives.

Pour calculer l'importance de la variable j, pour chaque arbre construit,

1. sélectionner le jeu de données OOB correspondant à l'arbre,
2. permuter les valeurs de la j-ème variable du jeu sélectionné,
3. calculer les prédictions de l'arbre sur le nouveau jeu de données,
4. calculer l'erreur de prédiction obtenue (erreur quadratique pour les régressions et taux de prédiction pour les classifications),
5. calculer le taux d'augmentation de l'erreur par rapport au vrai jeu de données OOB.

La moyenne des taux d'augmentation d'erreur de chaque variable explicative sur l'ensemble des arbres sert de mesure de son importance, plus l'augmentation d'erreur est grande, plus la variable est importante.

7.4 *Random forest* en pratique

Nous présentons ici comment utiliser les forêts aléatoires dans la pratique. Dans le cadre de ce travail, le package randomForest [32] sous R a été utilisé.

Les 4 points suivants sont à prendre en considération :

- le nombre d'arbres à développer dans la forêt, on a vu que le nombre d'arbres ne générait pas d'overfitting, la seule contrainte est alors le temps de calcul, l'idée est de commencer avec un certain nombre d'arbres et d'en ajouter progressivement jusqu'à ce que l'erreur *out of bag* se stabilise
- le nombre de variables explicatives m_{try} à garder à chaque noeud parmi les p variables disponibles, les valeurs par défaut du package sont $p/3$ pour la régression et $\sqrt{p}$ pour la classification, il faut explorer différentes valeurs et choisir celle qui minimise l'erreur *out-of-bag*, notons que le bagging consiste à prendre $m_{try} = p$
- l'importance des variables, il est important de lancer une première calibration uniquement dans le but d'éliminer les variables les moins importantes. Cette étape permettra de se débarrasser des variables bruitées et des variables inutiles dans le but d'éviter de se retrouver à certains arbres uniquement avec des variables ne pouvant aucunement améliorer les estimations finales d'un point de vue de réduction de l'erreur de validation.
- il faut également veiller à ce qu'il n'y ait pas de variables redondantes (i.e. représentant la même information) afin qu'aucune information ne soit sur-représentée lors de la sélection aléatoire des variables

Notons enfin que l'on peut gagner en temps d'exécution du fait de l'indépendance des arbres en construisant parallèlement les arbres de manière séparée avant de les recombiner pour construire la forêt.

8 Gradient Boosting Machine (GBM)

Nous présentons dans cette section une deuxième méthode d'apprentissage ensembliste introduite par J.H. Friedman dans [22] et [23] : le gradient boosting qui est une méthode de boosting séquentielle.

8.1 Définition

Contrairement aux forêts aléatoires qui font une moyenne de l'ensemble des prédicteurs tirés de manière aléatoire et indépendante, les GBMs proposent une stratégie constructive, les arbres s'ajoutent à l'ensemble de manière progressive et la construction d'un nouvel arbre se fait en tenant compte de l'apprentissage qui a été réalisé par les arbres précédents.

L'idée est toujours de chercher une fonction hypothèse $\hat{h}$ qui minimise l'erreur de généralisation :

$$\hat{h}(x) = \arg\min_{h(x)} E_{Y,X}(\, L(Y,\, h(X))\,)$$

où L est une fonction coût à choisir. Le gradient boosting propose d'utiliser la méthode de descente du gradient qui permet de minimiser des fonctions différentiables sur un espace hilbertien (espace vectoriel de dimension finie ou infinie muni d'un produit scalaire) en plusieurs itérations.

Concrètement, il s'agit de minimiser la fonction vectorielle de dimension n :

$$J(h) = \sum_{i=1}^{n} L(y_i, h(x_i))$$

Le gradient négatif de $J(h)$ indique la meilleure direction locale qui minimise $J(h)$, on met ensuite à jour la fonction $\hat{h}$:

$$\hat{h} \leftarrow \hat{h} - \rho \nabla J(h)$$

où ρ est le pas à effectuer dans la direction indiquée par le gradient. Notons que nous sommes contraints de choisir une fonction coût qui soit différentiable. Telle qu'elle est, cette méthode ne permet pas d'atteindre notre but. On approxime h aux points observés et on a à chaque pas autant de degrés de liberté que d'observations disponibles, ce qui génèrera un overfitting important. C'est là que les prédicteurs faibles interviennent, Friedman propose d'utiliser une certaine classe de fonctions qui limitent l'overfitting pour approximer le gradient, typiquement des arbres de prédiction. La construction de cette méthode est décrite dans l'algorithme 2.

Par construction, il est clair que cette méthode permet d'obtenir une précision meilleure que celle d'un arbre puisqu'on utilise à chaque étape l'information pouvant être donnée par un arbre pour améliorer la précision de manière séquentielle. La partie numérique permettra de confirmer ces résultats.

Toutefois, à la différence des forêts aléatoires, les arbres ne sont ici pas indépendants et chaque étape exploite l'information construite par les étapes précédentes. Le gradient boosting peut générer ainsi de l'overfitting en fonction des paramètres choisis. On présente dans la sous section qui suit différentes manières d'éviter et de réduire l'overfitting.

Algorithme 2 : Algorithme de la méthode du Gradient Boosting

Données : $\{(x_1, y_1), \ldots, (x_n, y_n)\}$: données initiales
$nbIter$: nombres d'itérations à faire
L : fonction coût choisie (différentiable)

1 **début**

2
$$h(x) \leftarrow \underset{\gamma \in \mathbb{R}}{argmin} \sum_{i=1}^{n} L(y_i, \gamma)$$

3

4 **pour** $m = 1$ *à* $nbIter$ **faire**

5
$$\text{pour } i \in \{1, \ldots, n\},\ r_{i,m} \leftarrow -\left.\frac{\partial L(y_i, z)}{\partial z}\right|_{z=h(x_i)}$$

6

7
$$h_m(x) \leftarrow \text{arbre de regression avec le jeu de donnes}\{(x_i, r_{i,m})\}$$

8

9
$$\gamma_m \leftarrow \underset{\gamma \in \mathbb{R}}{argmin} \sum_{i=1}^{n} L(y_i,\ h(x_i) + \gamma\, h_m(x_i))$$

10

11
$$h(x) \leftarrow h(x) + \gamma_m\, h_m(x)$$

12

13 **fin**

14 **retourner** $h(x)$

15 **fin**

8.2 Réduction de l'overfitting

8.2.1 Learning rate et nombre d'itérations

Comme nous l'avons déjà évoqué pour les forêts aléatoires, un arbre de prédiction représente un certain point de vue et qui peut contenir beaucoup de bruit, Friedman introduit ainsi un coefficient de contrôle (*shrinkage ou learning rate*) qu'on note ν et qui permet de réduire le pas effectué à chaque descente de gradient afin que les éventuels "mauvais arbres" n'aient pas d'impact important sur l'optimisation. La ligne 11 de l'algorithme 2 est ainsi modifiée par

$$h(x) \leftarrow h(x) + \nu\, \gamma_m\, h_m(x)$$

Clairement, le nombre d'itérations M a un impact direct sur l'overfitting car chaque étape augmente la quantité d'information apprise jusqu'à ce qu'à détériorer la capacité de généralisation. Il y a une dépendance mutuelle entre le learning rate et le nombre d'itérations, plus le learning rate est petit plus le nombre d'itérations doit être grand pour avoir un même résultat. En pratique, on fixe d'abord ν et on choisir le nombre d'itérations optimal par cross-validation.

On peut voir sur la figure 8.1 l'effet du learning rate et du nombre d'itérations dans l'exemple de la modélisation du coût des sinistres, on fixe le nombre d'itérations à 500 et on observe l'évolution

de l'erreur de généralisation en fonction du learning rate. On distingue deux parties dans le graphe, une première où le learning rate est trop bas pour un nombre d'arbres de 500 générant du sous-apprentissage et une deuxième partie où le nombre d'arbres est trop important générant du sur-apprentissage.

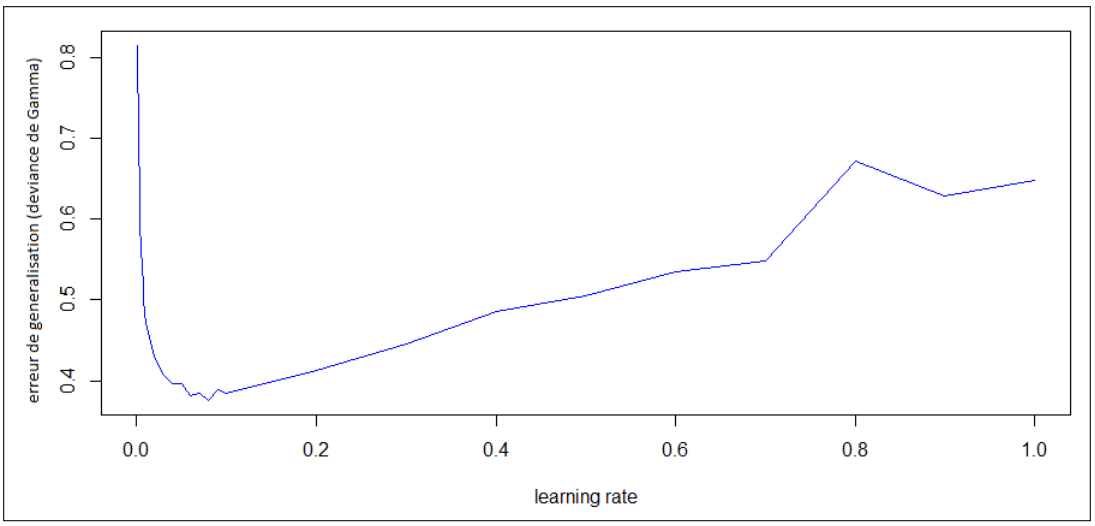

FIGURE 8.1 – Effet du learning rate sur l'erreur de validation, nombre d'arbres fixé à 500

8.2.2 Sous échantillonnage

Friedman propose également d'ajouter une composante stochastique à l'algorithme. Dans le même esprit que les forêts aléatoires, il propose à chaque étape de construire l'arbre permettant d'approximer le gradient en prenant uniquement un sous échantillon (sans remise) de l'ensemble des données. Il reste alors à choisir la taille du sous échantillon. C'est ce qu'on appelle le gradient boosting stochastique.

Cela permet également d'avoir comme pour les forêts aléatoires un OOB dataset à chaque étape et qui fournit un jeu n'ayant pas servi à la construction de l'étape et qui permet de mesurer l'évolution de l'erreur par rapport à l'étape précédente. Toutefois, on remarque en pratique que cet indicateur n'est pas très fiable du fait que les données OOB ont servi à la construction des étapes précédentes.

On peut voir sur la figure 8.2 l'effet du sous échantillonnage sur l'erreur de généralisation dans l'exemple du coût des sinistres, le nombre d'arbres et le learning rate étant fixes.

S'ajoutent à ces paramètres les paramètres des arbres de prédiction vus précédemment (profondeur de l'arbre et nombre minimum d'observations par feuille). Classiquement, les valeurs optimales à choisir pour tous ces paramètres sont celles qui minimisent l'erreur de généralisation (calculée par cross validation).

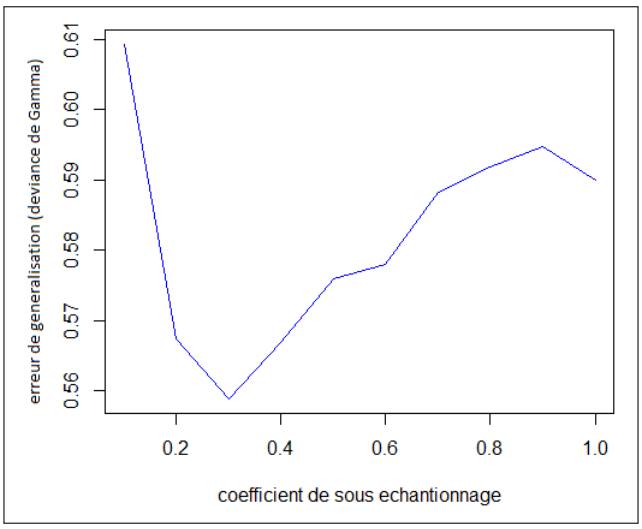

FIGURE 8.2 – Effet du sous échantillonnage sur l'erreur de validation

8.3 Choix de la fonction coût

On s'intéresse dans cette partie au choix de la fonction coût L. Mise à part la contrainte qu'elle doit être différentiable, il n'y a aucune autre règle quant au choix de cette fonction. Le choix peut donc être arbitraire mais toutes les fonctions coûts n'aboutissent pas aux mêmes résultats, c'est le plus souvent un travail d'exploration qui permet de choisir la fonction la plus appropriée en fonction de la nature des données à estimer :

- Pour des **données de classification**, on utilisera par exemple la déviance de Bernoulli.
- Pour des **données de comptage**, on utilisera par exemple déviance de Poisson ou la déviance binomiale négatives.
- Pour des **données continues**, on utilisera l'erreur des moindres carrés (déviance gaussienne) ou la déviance de Laplace. Il faut se demander dans ces cas là à quel point on veut prendre en compte les grandes valeurs dans notre estimation, l'erreur des moindres carrés est par exemple très impactée par les grands écarts à la moyenne, la déviance de Laplace permet de minimiser leur impact. Pour réduire l'impact des points extrêmes, il existe également des fonctions hybrides comme la fonction Huber (à paramètre δ fixé) qui combine la déviance gaussienne et celle de Laplace :

$$L_{Huber,\delta}(y,f) = \begin{cases} \frac{1}{2}(y-f)^2 & \text{si } |y-f| \leq \delta \\ \delta\left(|y-f| - \frac{\delta}{2}\right) & \text{si } |y-f| > \delta \end{cases}$$

ou encore la fonction quantile (à paramètre α fixé) définie par :

$$L_{quantile,\alpha}(y,f) = \begin{cases} (1-\alpha)|y-f| & \text{si } |y-f| \leq 0 \\ \alpha|y-f| & \text{si } |y-f| > 0 \end{cases}$$

Pour des données continues positives comme les coûts de sinistre, on préfèrera la déviance de Gamma puisque la distribution sous-jacente est généralement bien adaptée pour représenter les coûts de sinistre.

Pour une étude détaillée des différentes fonctions coût, on pourra se référer à [24].

8.4 Influence relative des variables explicatives

L'une des mesures les plus utiles après la calibration du modèle est l'influence relative de chacune des variables explicatives. Outre la description qualitative des données offerte par cette

mesure, celle-ci permet également de détecter d'éventuelles variables bruitées ou inutiles et de les éliminer.

Toujours dans l'article [22], Friedman propose de mesurer l'importance d'une variable par son impact sur la calibration de la fonction finale $\hat{h}$. Elle est alors estimée par le nombre de fois que la variable est sélectionnée pour séparer les données, pondéré par le taux de réduction de l'erreur (voir équations 6.1 et 6.2) découlant de la séparation des données, et moyenné sur l'ensemble des arbres. Formellement, l'influence relative de la variable j est donnée par :

$$\hat{I}_j = \frac{1}{M} \sum_{m=1}^{M} \hat{I}_j(T_m)$$

où, M désigne le nombre d'arbres total
T_m désigne le m-ème arbre de la forêt

avec,

$$\hat{I}_j(T_m) = \sum_{t=1}^{J} i_t \, 1_{\{v_t = j\}}$$

où, J désigne le nombre de noeuds non terminaux de l'arbre T_m
v_t désigne le numéro de la variable explicative utilisée pour la séparation du t-ème noeud
i_t désigne le taux de réduction de l'erreur suite à la séparation du t-ème noeud

On normalise enfin chacune de ces estimations de sorte à ce que la somme sur toutes les variables explicatives soit égale à 1.

8.5 En pratique

Dans le cadre de ce mémoire, nous avons utilisé le package gbm [35] et le package dismo [34] qui permet de mieux exploiter le parckage gbm.

En plus de la fonction de coût à choisir, il y a 4 paramètres à fixer principalement :

- le learning rate ν et le nombre d'itérations M : on a dit que ces deux paramètres sont mutuellement dépendants et qu'il fallait en fixer un pour avoir la valeur optimale de l'autre. L'idéal serait d'avoir un ν aussi petit que possible mais nous sommes limités par le temps de calcul, plus ν est petit plus M doit être grand et le nombre d'arbres à construire est directement lié au temps de calcul et le temps de stockage en mémoire. En pratique, il est raisonnable de prendre $0.001 \leq \nu \leq 0.01$ et de limiter le nombre d'itérations à 10000. Le lien entre le learning rate et le nombre d'itérations optimal n'est pas proportionnel, i.e. si $\nu = 0.01$ donne $M = 300$ on n'a pas forcément $M = 3000$ pour $\nu = 0.001$. On peut voir sur la figure 8.3 l'évolution du nombre optimal d'arbres en fonction du learning rate qui varie entre 0.001 et 0.1 pour la régression sur le coût des sinistres MRH. On obtient par ailleurs une erreur de généralisation de 38% (arrondi à 10^{-1}) quelque soit le learning rate choisi tant que la régression est faite avec le nombre optimal d'arbre correspondant, toutefois on a une erreur de 63% si on choisit $\nu = 1$, ce qui montre que le learning rate permet bien d'améliorer le modèle.
- le coefficient de sous échantillonnage f : il correspond au coefficient de réduction de la taille des données dans le sous échantillonnage réalisé à chaque étape, en général on prend $0.5 \leq f \leq 0.8$

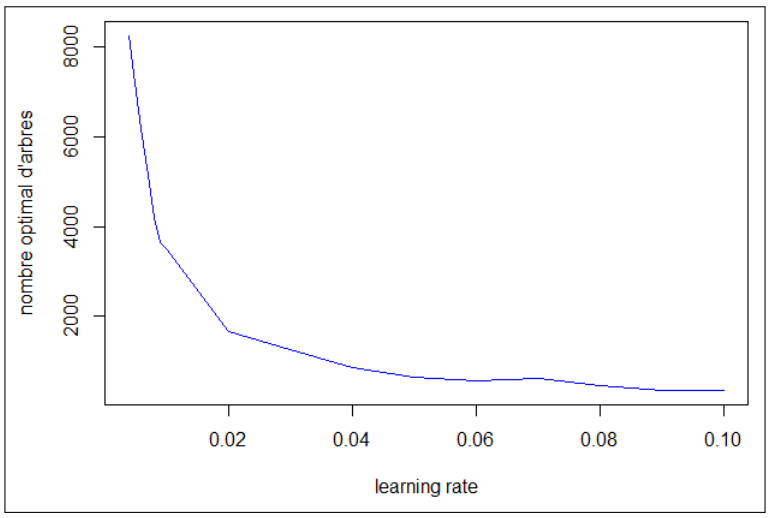

FIGURE 8.3 – Dépendance entre learning rate et nombre d'itérations, exemple avec le coût des sinistres

- complexité des arbres : la valeur par défaut de la profondeur des arbres dans le package dismo est 5 et semble convenir. Les résultats de la régression ne sont pas très sensibles à ce paramètre car le nombre d'arbres est adapté en conséquence.

Enfin, à la différence des forêts aléatoires, l'importance des variables explicatives n'a pas d'impact sur la qualité des estimations car on utilise à chaque étape l'ensemble des variables explicatives et on a vu que les arbres de prédiction sélectionnaient automatiquement les variables les plus importantes. Ce propos doit néanmoins être nuancé lorsque l'on fixe un coefficient de sous échantillonnage inférieur à 1 auquel cas des variables bruitées peuvent être sélectionnées pour séparer certains sous-ensembles du jeu de données initial. Il convient alors de lancer une première calibration avec un coefficient de sous échantillonnage égal à 1 afin de sélectionner uniquement les données d'influence relative significative. De plus, même si elles sont sans impact sur le résultat final, supprimer les variables d'importance nulle permet d'optimiser le temps de calcul lié à la calibration du modèle.

9 Calibration de la vulnérabilité

Nous proposons dans cette partie de calibrer les différents algorithmes d'apprentissage afin de modéliser le coût engendré par un site sinistré. La méthode de la validation croisée aura un rôle essentiel dans le choix du modèle final et de ses paramètres, les différents jeux d'apprentissage et de validation seront tirés de manière aléatoire tout en conservant les mêmes proportions de chaque évènement dans chacun des jeux.

9.1 Variables explicatives utilisées

Nous présentons dans un premier temps les différentes variables explicatives à notre disposition issues des données présentées dans la partie 1 et des analyses hydrologiques présentées dans la partie 3 :

- **Valeur estimée du bien sinistré** : estimation faite par l'assureur du montant de sinistre maximal brut de conditions financières pouvant survenir au vu des garanties offertes par la police d'assurance.
- **Branche d'activité sous laquelle le bien est assuré**.
- **Étage du bien** : cette variable va nous servir d'indicateur de la hauteur du bien par rapport au sol, elle est essentiellement utile pour la branche MRH car les biens des autres branches sont souvent situés au niveau du sol. Elle peut prendre trois valeurs :
  - Maison : bien située au niveau du sol et ne faisant pas partie d'un immeuble.
  - Rez-de-chaussée : Appartement situé au rez-de-chaussée d'un immeuble.
  - Étage : Appartement situé en étage d'un immeuble.
- **Évènement associé au sinistre** : notre motivation derrière l'intégration de cette variable est de nous permettre de capturer les éventuelles spécificités de chaque évènement qui n'ont pas pu être expliquées par les autres variables, il peut s'agir par exemple des systèmes de défense mis en place contre les crues ou encore du mode de construction dans la zone inondée qui peut aggraver ou atténuer les dégâts causés.
- **Distance à la cellule débordée la plus proche** : l'idée derrière cette variable est de contourner les limites du modèle de propagation présenté dans la partie 3.1 en passant d'une variable binaire (une cellule est débordée ou ne l'est pas) à une variable continue permettant de distinguer les cellules qui n'ont vraisemblablement pas été inondées (une cellule se trouvant par exemple à plusieurs kilomètres de la cellule débordée la plus proche) et les cellules pour lesquelles il serait imprudent d'affirmer avec certitude qu'elles n'ont pas été inondées (une cellule se trouvant à quelques mètres d'une cellule inondée). Cette variable permet aussi de donner du sens aux sinistres se trouvant dans des cellules non inondées et non concernées par le ruissellement. Notons que cette variable prend la valeur 0 lorsque le sinistre se trouve dans une cellule inondée. De plus, elle permet également d'amortir l'imprécision liée à notre modèle numérique de terrain dont la résolution horizontale de $75m$ n'est pas suffisamment fine pour un modèle d'inondation.
- **La hauteur d'eau simulée de la cellule débordée la plus proche**.
- **Hauteur relative au réseau de drainage le plus proche** (voir partie 3.3).
- **La quantité maximale de précipitations journalières ayant eu lieu sur la zone du sinistre pendant la durée de l'évènement**.

9.2 Modèles linéaires généralisés

Nous allons tenter, dans cette partie, de construire un *GLM* permettant d'estimer le coût de chaque sinistre causé par un des évènements étudiés. Nous allons partir de la base des sinistres et

calibrer des modèles de régression Gamma et Bêta.

Nous vérifions dans un premier temps l'indépendance des variables explicatives en calculant leur matrice de corrélation. Les corrélations en valeur absolue vont de 0.01% à 12% avec une moyenne de 2%. Ces corrélations relativement basses nous permettent d'être confiants quant à l'indépendance des variables utilisées.

9.2.1 Régression Gamma

Commençons par lancer le modèle avec l'ensemble des variables explicatives à notre disposition, nous avons retenu la fonction de lien canonique ln. Voici un extrait des résultats de calibration obtenus sous R :

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.017e+01  1.796e-01  56.618  < 2e-16 ***
BC_VALUE                   2.614e-10  1.111e-10   2.353 0.018651 *
nearestCell               -9.291e-03  9.092e-04 -10.219  < 2e-16 ***
precip_max                 1.092e-02  1.489e-03   7.332 2.37e-13 ***
carthage                  -1.985e-02  2.203e-03  -9.011  < 2e-16 ***
waterHeight               -6.498e-03  1.154e-02  -0.563 0.573499
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nous remarquons qu'à l'exception de la hauteur d'eau (nommée *waterHeight* dans les résultats), toutes les variables explicatives passent le test de Student sur la non-nullité des coefficients avec une confiance confortable.

Le rapport de déviance expliquée sur ce premier modèle est de 16.6%.

La non-pertinence de la hauteur d'eau dans le modèle n'était pas inattendue au vu des limites du module de propagation de l'eau mentionnées dans la partie 3.1. En effet, il est plus prudent d'apprécier le caractère binaire de cette variable (inondée/non inondée) que la valeur de la hauteur d'eau en elle-même. Nous allons tout de même essayer d'en extraire de l'information en définissant trois intervalles de hauteurs d'eau qui permettront d'exprimer la sévérité du débordement de chaque cellule en espérant que celle-ci soit moins bruitée que la hauteur d'eau en elle-même.

Pour définir les bornes de ces intervalles, nous traçons la moyenne cumulée des taux de destruction[21] des sites assurés au fur et à mesure que la hauteur d'eau décroît. L'observation de sauts dans la moyenne cumulée, indiquant une variation significative de la sévérité aux hauteurs d'eau correspondantes, permettra de définir les bornes des intervalles.

La figure 9.1 laisse apparaître des sauts (entraînant une baisse de la moyenne) au niveau des hauteurs 5.2 m et 6.2 m.

Nous lançons une nouvelle calibration de notre *GLM* en exprimant la variable hauteur d'eau à l'aide de trois modalités : faible, modéréé, élevée. Voici un extrait des résultats :

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
```

21. Le taux de destruction est défini comme la charge du sinistre divisée par la valeur assurée du bien sinistré

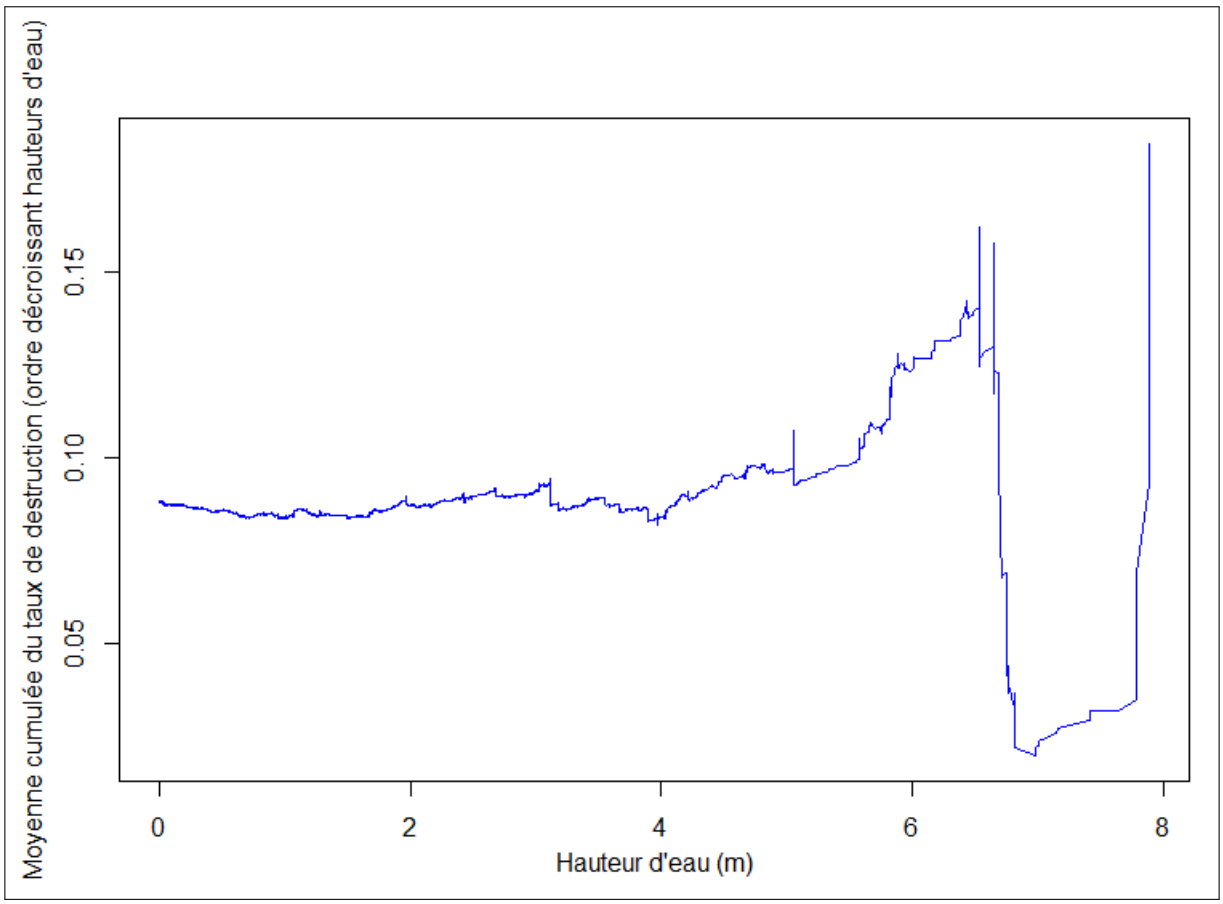

FIGURE 9.1 – Moyenne cumulée du taux de destruction dans l'ordre décroissant des hauteurs d'eau.

```
(Intercept)          9.896e+00  2.100e-01  47.120  < 2e-16 ***
BC_VALUE             2.623e-10  1.108e-10   2.367 0.017963 *
nearestCell         -9.210e-03  9.062e-04 -10.163  < 2e-16 ***
precip_max           1.084e-02  1.484e-03   7.301 2.98e-13 ***
carthage            -1.964e-02  2.200e-03  -8.925  < 2e-16 ***
waterHeightSevere    4.872e-01  1.797e-01   2.712 0.006700 **
waterHeightWeak      2.720e-01  1.124e-01   2.419 0.015568 *
```

La variable passe maintenant le test de non nullité des coefficients avec une confiance confortable. Néanmoins, on observe un coefficient positif pour la modalité hauteur d'eau faible, ce qui signifie que l'on estime des sinistres plus sévèrement pour des hauteurs d'eau faibles que des hauteurs d'eau modérées (celle-ci étant la modalité de référence). Cela contredit le sens physique que l'on attribuait à cette variable. De plus, en modifiant légèrement la borne de 5.2 m à 5.0 m, la variable ne passe plus le test de non nullité avec les résultats suivants :

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.014e+01  1.988e-01  51.011  < 2e-16 ***
BC_VALUE             2.622e-10  1.111e-10   2.360 0.018281 *
nearestCell         -9.213e-03  9.083e-04 -10.143  < 2e-16 ***
precip_max           1.088e-02  1.488e-03   7.310 2.79e-13 ***
carthage            -1.995e-02  2.206e-03  -9.042  < 2e-16 ***
waterHeightSevere    2.387e-01  1.649e-01   1.448 0.147722
```

```
waterHeightWeak            1.903e-02  8.678e-02    0.219 0.826464
```

Cette variable nous semble alors beaucoup trop bruitée comme nous pouvons le voir sur la figure 9.2 qui trace le taux de destruction en fonction de la hauteur d'eau. Nous décidons alors de l'éliminer. Nous vérifions que les tests de non nullité sont toujours validés avec le nouveau modèle retenu :

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.016e+01  1.796e-01  56.600  < 2e-16 ***
BC_VALUE                   2.627e-10  1.111e-10   2.363 0.018124 *
nearestCell               -9.262e-03  9.081e-04 -10.200  < 2e-16 ***
precip_max                 1.096e-02  1.488e-03   7.365 1.85e-13 ***
carthage                  -1.989e-02  2.203e-03  -9.029  < 2e-16 ***
```

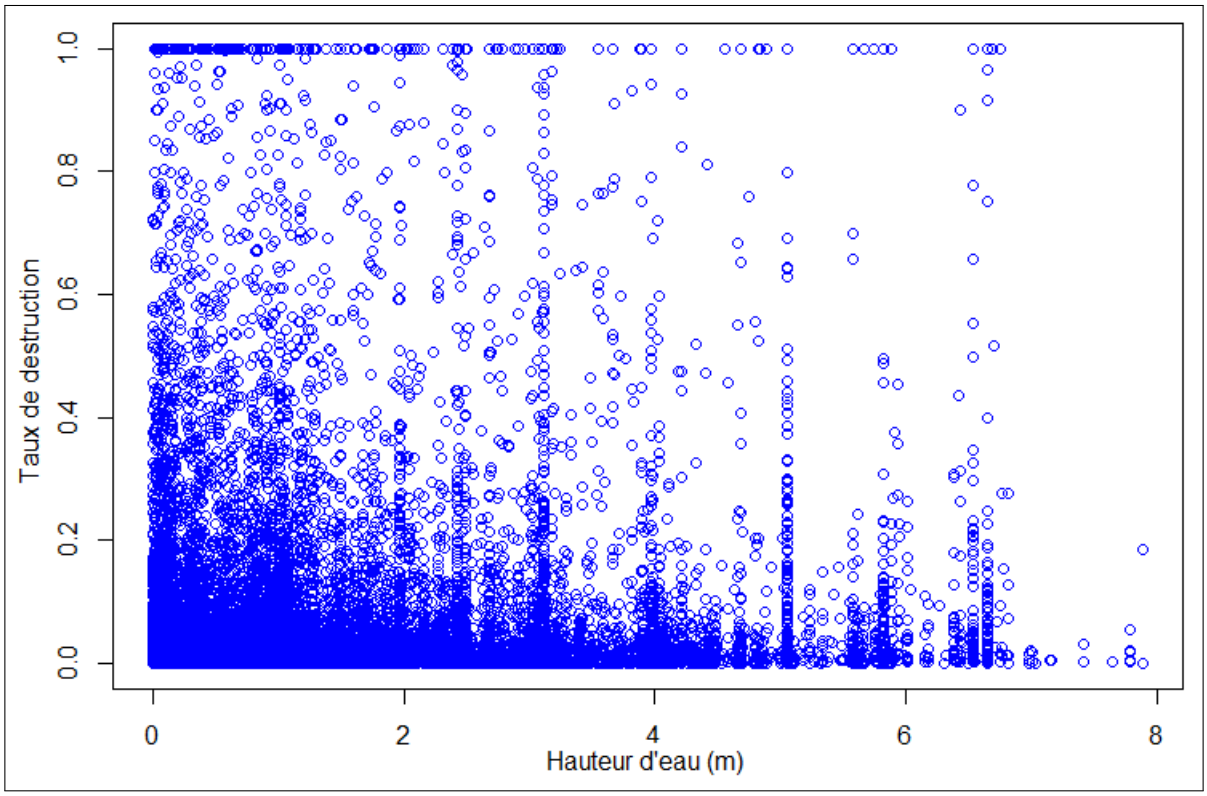

FIGURE 9.2 – Nuage de points des taux de destruction en fonction des hauteurs d'eau.

Nous gardons un rapport de déviance expliquée de 16.6%, nous ne sommes donc pas pénalisés par le retrait de la variable hauteur d'eau.

Enfin, nous vérifions que notre modèle ne fait pas de sur-apprentissage en calculant l'erreur de validation par la méthode de validation croisée avec $K = 10$. Les jeux d'apprentissage et de validation ont été tirés de manière aléatoire et uniforme. On obtient :

- Déviance expliquée moyenne sur les jeux d'apprentissage : 16.6%
- Déviance expliquée moyenne sur les jeux de validation (complémentaire de l'erreur de validation) : 15.4%

Comme les deux valeurs sont proches, nous pouvons affirmer que le sur-apprentissage est limité dans notre modèle.

Nous pouvons voir sur la figure 9.3 le QQ-plot du modèle après agrégation des résultats par évènement ainsi que l'intervalle de confiance à 95% des valeurs prédites. On remarque que l'intervalle de confiance est très étroit, l'estimation de la somme des sinistres étant moins volatile que l'estimation des sinistres pris un à un.

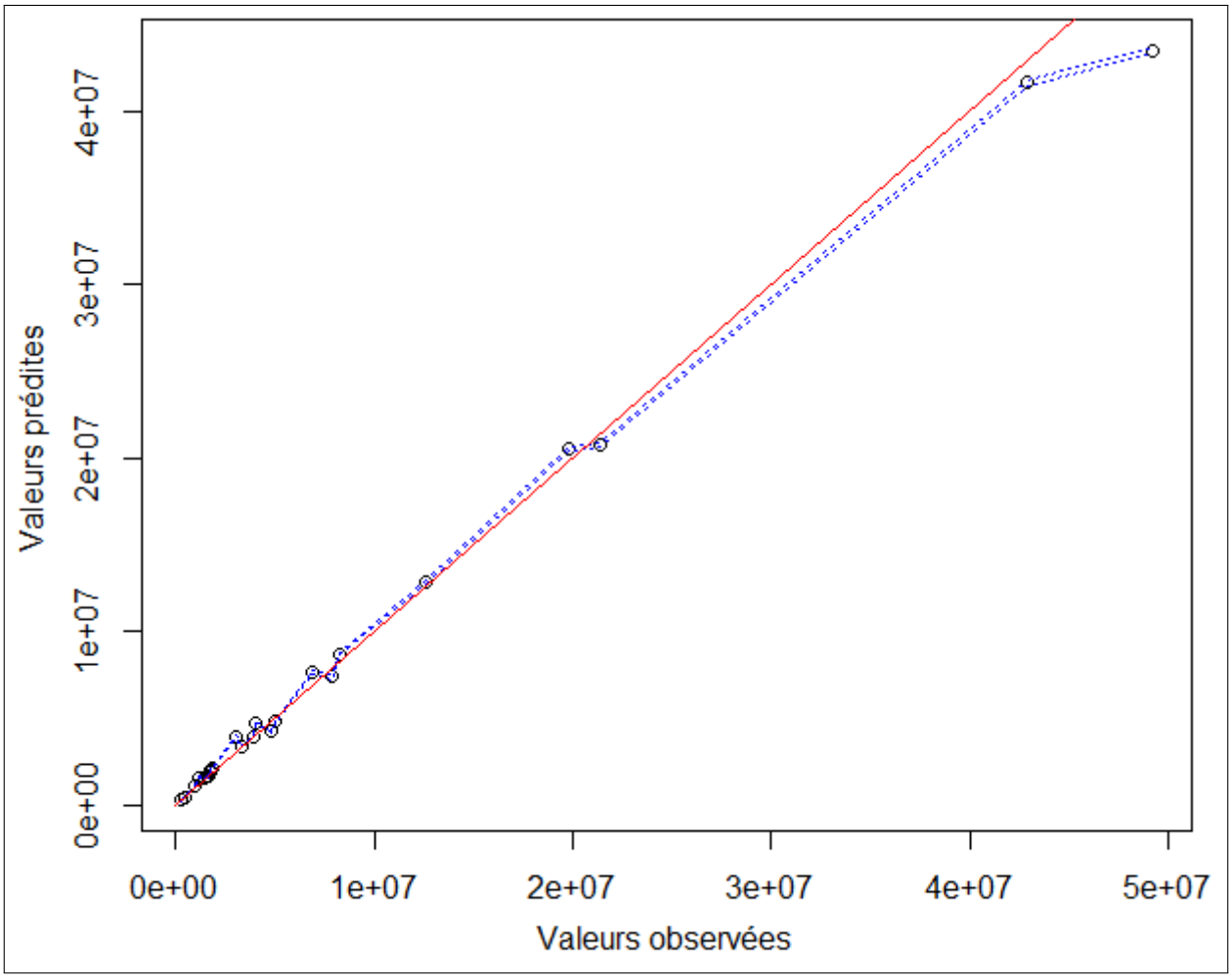

FIGURE 9.3 – QQ-plot du modèle *GLM* Gamma. Données agrégées par évènement. En bleu : intervalle de confiance à 95% des valeurs prédites.

9.2.2 Régression bêta

Dans le modèle de la régression Gamma, nous avons cherché à estimer la charge du sinistre étant donné un ensemble de variables explicatives dont la valeur assurée qui s'est avérée, sans surprise, pertinente. En effet, toutes choses étant égales par ailleurs, le montant d'un sinistre est une fonction croissante de la valeur du bien sinistré. La première limitation du modèle Gamma est qu'il accorde un poids constant à la valeur assurée menant à supposer, lorsqu'on utilise la fonction de lien logarithmique, une relation proportionnelle entre le montant du sinistre et la valeur du bien et donc un taux de destruction constant.

Or, il est généralement admis et empiriquement observé que la taille du risque (approchée par la valeur assurée) a un impact à la baisse sur le taux de destruction dû à ce que l'on appellera "l'effet château/maison", illustré sur la figure 9.4 pour le risque d'incendie. Il est effectivement raisonnable de penser que plus le bien touché est grand, moins il sera touché par une inondation : un château sera touché uniquement au premier niveau qui représente une toute petite partie du

château alors qu'une maison constitué d'un seul niveau sera entièrement touchée, les risques de grandes tailles ont aussi tendance à être mieux protégés.

FIGURE 9.4 – Illustration de "l'effet château/maison" pour le risque d'incendie.

Face à cette limitation, nous allons modéliser le taux de destruction du site sinistré au lieu de la charge sinistre. Le taux de destruction ayant pour support l'intervalle $[0, 1]$, la régression Bêta est particulièrement adaptée. De plus, comme on peut le voir sur la figure 9.5, la loi Bêta offre l'avantage d'avoir une courbe de densité pouvant prendre plusieurs formes différentes, être convexe ou concave, symétrique ou non. Ainsi, la régression Bêta est adaptée à plusieurs types de données quel que soit le comportement de la variable-réponse.

Nous lançons alors l'estimation des paramètres avec les variables explicatives retenues dans le modèle Gamma. Plusieurs estimations ont été faites avec les différentes fonctions de lien présentées dans la partie 4.7, nous avons retenu la fonction de lien complémentaire log-log qui offrait le meilleur rapport de déviance expliquée. Voici un extrait des résultats :

```
Coefficients (mean model with cloglog link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.927e+00  8.652e-02 -22.266  < 2e-16 ***
BC_VALUE                -9.286e-11  5.421e-11  -1.713  0.08671 .
nearestCell             -4.575e-03  4.357e-04 -10.500  < 2e-16 ***
precip_max               4.482e-03  7.126e-04   6.289 3.19e-10 ***
carthage                -1.122e-02  1.058e-03 -10.604  < 2e-16 ***
```

La pertinence des variables explicatives est conservée (avec une confiance moindre pour la valeur assurée) mais nous augmentons comme nous l'espérions le rapport de déviance expliquée qui est maintenant de 73%.

Malheureusement, l'agrégation des résultats après multiplication par les valeurs assurées permettant d'obtenir les pertes par évènement donnent des estimations très éloignées des valeurs observées comme on peut le voir sur le QQ-plot de la figure 9.6. Ceci est dû à la volatilité de la variable valeur assurée qui vient s'ajouter à la volatilité des prédictions des taux de destruction.

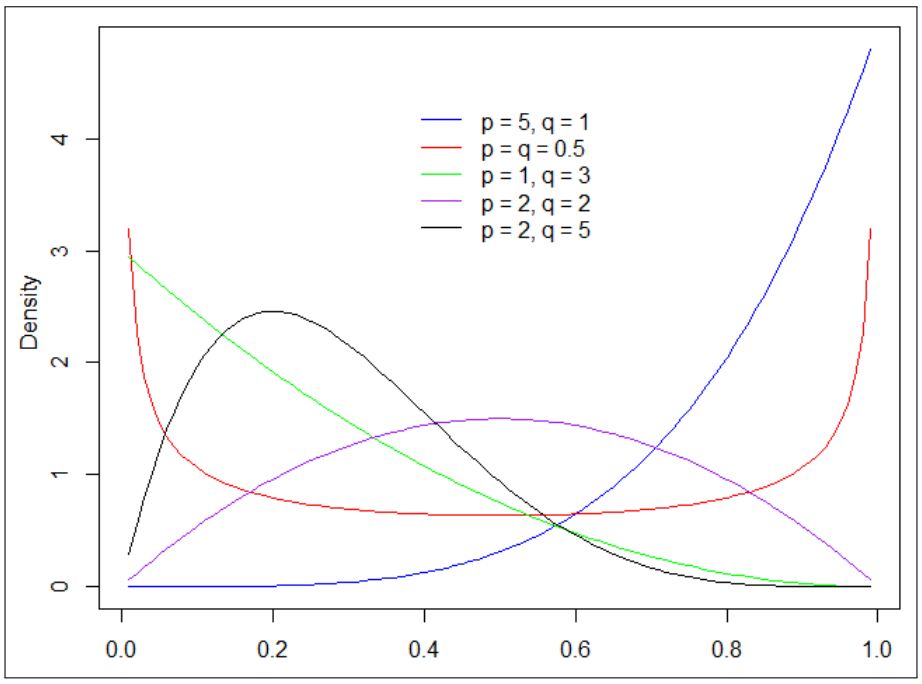

FIGURE 9.5 – Différentes densités de la loi Bêta

Ainsi, les petits écarts entre valeurs observées et prédites sur les taux de destruction sont amplifiés après multiplication par les valeurs assurées et la volatilité de cette variable empêche la réduction de ces écarts lors de l'agrégation des prédictions.

Ainsi, même si le modèle Bêta estime mieux les taux de destructions que le modèle Gamma estime les charges sinistres, le modèle Gamma semble le plus apte à prédire les charges agrégées par évènement.

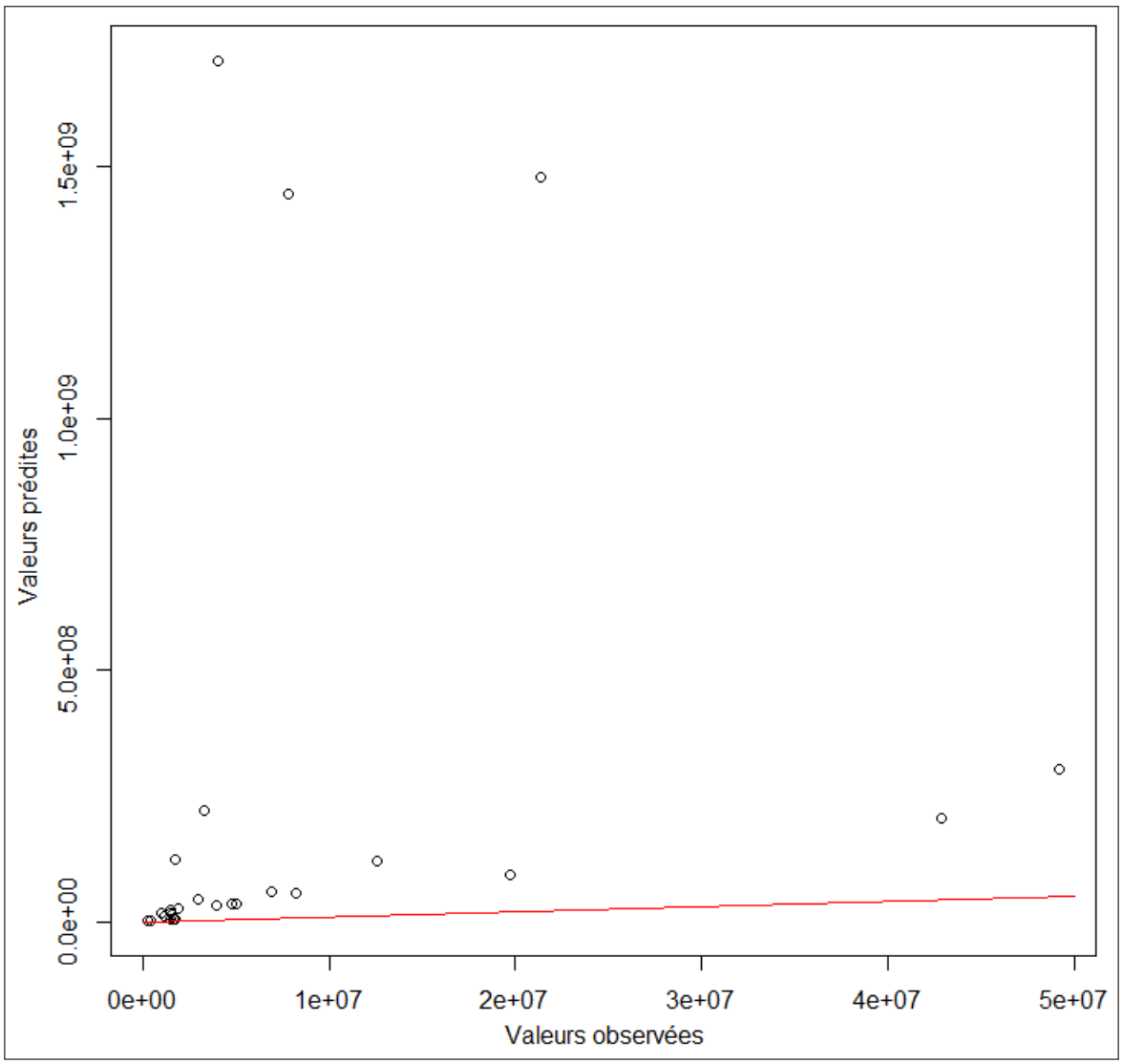

FIGURE 9.6 – QQ-plot du modèle *GLM* Bêta. Données agrégées par évènement.

9.3 Gradient Boosting Machine

Nous cherchons maintenant à calibrer le modèle à l'aide d'un *GBM* en testant les fonctions de coût de Laplace et Gaussienne (voir partie 8.3) et en fixant les paramètres de nombre d'arbres, *learning rate* et coefficient de sous-échantillonnage de sorte à minimiser l'erreur de validation.

Nous allons calibrer les modèles avec trois différentes fonctions de coût : la déviance de Laplace, de Gauss et Gamma. Au vu des remarques faites sur la variable de la hauteur d'eau lors de la calibration des *GLMs* dans la section précédente, nous construirons pour chaque fonction de coût un modèle avec toutes les variables et un modèle sans la variable hauteur d'eau. Les résultats d'ajustement décrits par la part de déviance expliquée sur les jeux de validation et d'apprentissage sont présentés dans le tableau 9.7.

Nous remarquons trois points essentiels à partir de ces résultats :

- Sans grande surprise, la variable hauteur d'eau n'a jamais d'impact significativement positif sur les résultats et sa suppression permet de diminuer l'erreur d'apprentissage du modèle basé sur la déviance de Gauss. Au vu de son caractère très bruité et donc d'aucune contribution possible à la capacité de prédiction du modèle, nous abandonnons définitivement cette variable.

	Laplace	Laplace*	Gauss	Gauss*	Gamma	Gamma*
Erreur de validation	72.5%	72.5%	93%	92.8%	66.7%	66.8%
Erreur d'apprentissage	70.1%	70.2%	71.4%	69.6%	65.9%	66%

FIGURE 9.7 – Erreurs d'apprentissage et de validation des modèles *GBMs*. * désigne le modèle calibré sans la variable hauteur d'eau.

- Tandis que les modèles de Laplace et Gamma ont des erreurs d'apprentissage et de validation très proches, l'écart important entre ces deux erreurs pour le modèle de Gauss laissent apparaître un sur-apprentissage important. Cela est dû à la grande sensibilité de la fonction de coût gaussienne aux valeurs extrêmes du jeu de données rendant le modèle très lié au jeu d'apprentissage. Nous éliminons donc la fonction de coût Gaussienne.
- La fonction de coût Gamma est celle qui permet de produire le modèle avec la plus petite erreur et le plus faible écart entre erreurs d'apprentissage et de validation et donc le moins de sur-apprentissage. La loi Gamma étant particulièrement adaptée pour la modélisation des coûts de sinistres, choisir sa déviance comme fonction de coût pénalise moins les écarts à la moyenne estimée car leur distribution sous-jacente se rapproche vraisemblablement d'une loi Gamma.

Les figures 9.8 et 9.9 représentent l'influence relative (en pourcentage) de chaque variable explicative dans le cas des deux modèles Gamma. On remarque d'abord que la variable hauteur d'eau est effectivement la moins influente même si son bruit permet d'expliquer une partie de la variabilité totale. À l'inverse, la valeur assurée et la modalité de l'évènement ont une grande influence sur les prédictions. Ce qui nous réconforte dans la théorie de l'effet "Château/Maison" mentionnée dans la section précédente mais également dans nos motivations de modélisation quant à la spécificité de destruction de chaque évènement que l'on cherche à capter. Il est aussi intéressant de noter que l'influence de la variable hauteur d'eau est, après sa suppression, essentiellement transférée aux variables de distance à la cellule la plus proche et de hauteur au dessus du réseau de drainage, qui sont également des variables de nature hydrologique.

Enfin, la figure 9.10 montre le QQ-plot des trois modèles (sans la variable hauteur d'eau). Il en ressort le sur-apprentissage flagrant du modèle de Gauss et l'erreur de prédiction du modèle de Laplace une fois les pertes agrégées par évènement. Ainsi, c'est la loi la plus adaptée à la modélisation des coûts de sinistres qui définit la fonction de coût du modèle *GBM* le plus robuste. Nous retiendrons parmi ces trois modèles, en effet, le modèle Gamma par la suite.

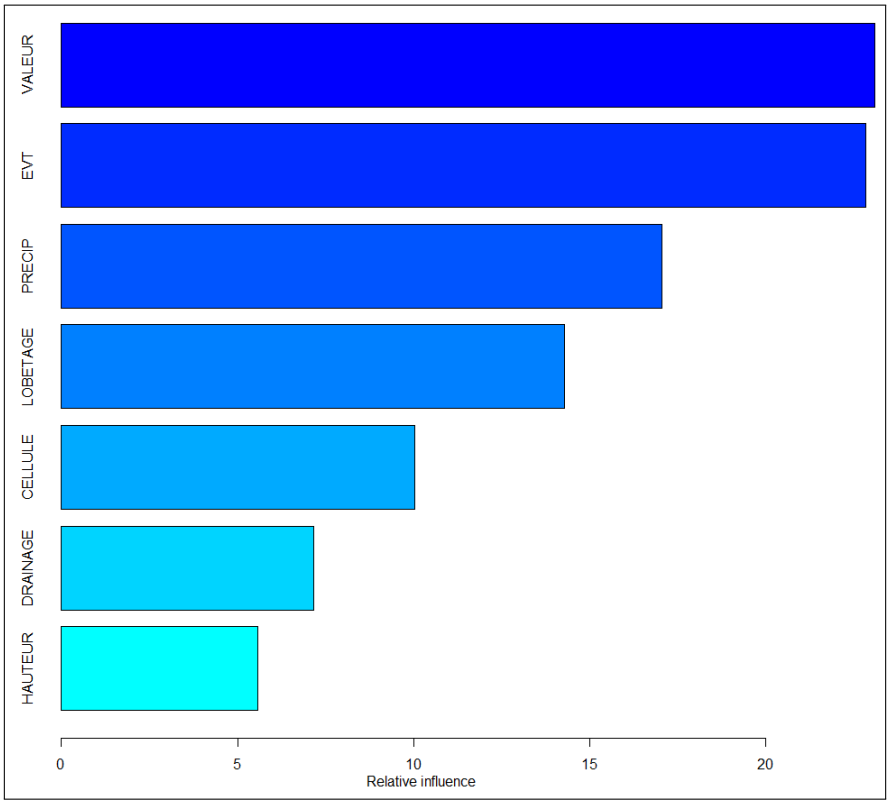

FIGURE 9.8 – Influence relative de toutes les variables explicatives avec la fonction de coût Gamma.

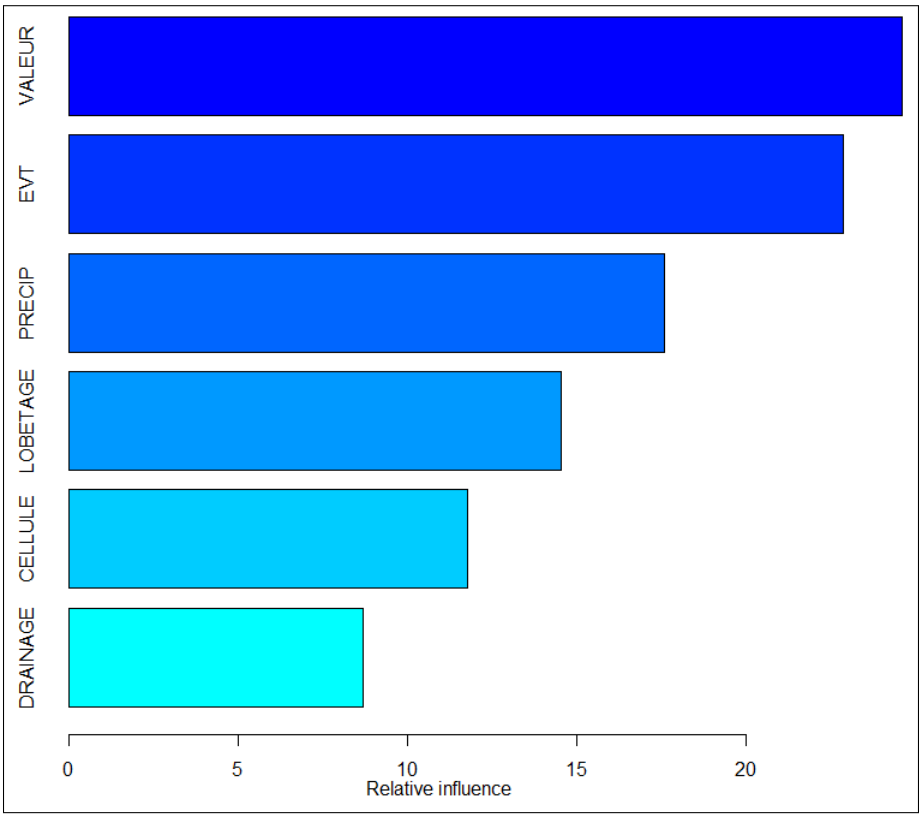

FIGURE 9.9 – Influence relative des variables explicatives sauf la hauteur d'eau avec la fonction de coût Gamma.

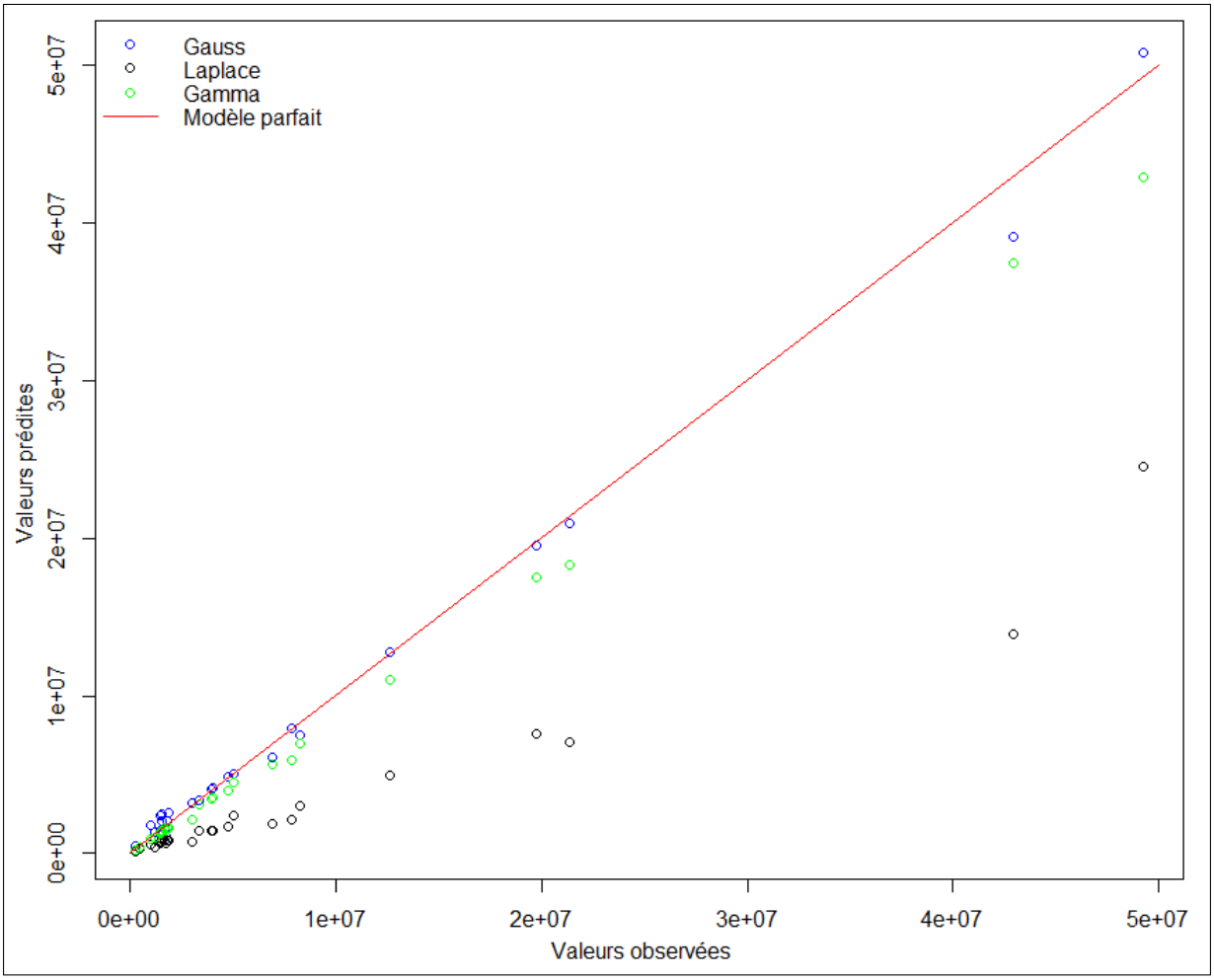

FIGURE 9.10 – QQ-plot des trois modèles *GBMs* sans la variable hauteur d'eau.

9.4 Forêts aléatoires

Enfin, nous essayons de prédire le coût des sinistres en construisant une forêt aléatoire. L'erreur *OOB* converge à partir de 300 arbres et nous construisons la forêt finale avec 500 arbres.

Une première calibration incluant la variable hauteur d'eau confirme encore une fois son caractère bruité que nous éliminons donc.

Nous obtenons un rapport de déviance Gamma expliquée de 22% sur le jeu d'apprentissage et de 19% sur le jeu de validation (les données *OOB*), ce qui permet d'affirmer que le sur-apprentissage est très limité dans ce modèle.

Nous pouvons voir sur la figure 9.11 le QQ-plot du modèle construit. Malgré des rapports de déviance moins bons que le *GBM* Gamma, nous remarquons une meilleure qualité d'ajustement lorsque l'on agrège les pertes par évènement. De plus, le graphe nous confirme l'absence de sur-apprentissage dans le modèle puisque les prédictions des pertes par évènement faites sur les données out-of-bag se superposent de manière quasi-parfaite avec les prédictions faites sur les données d'apprentissage.

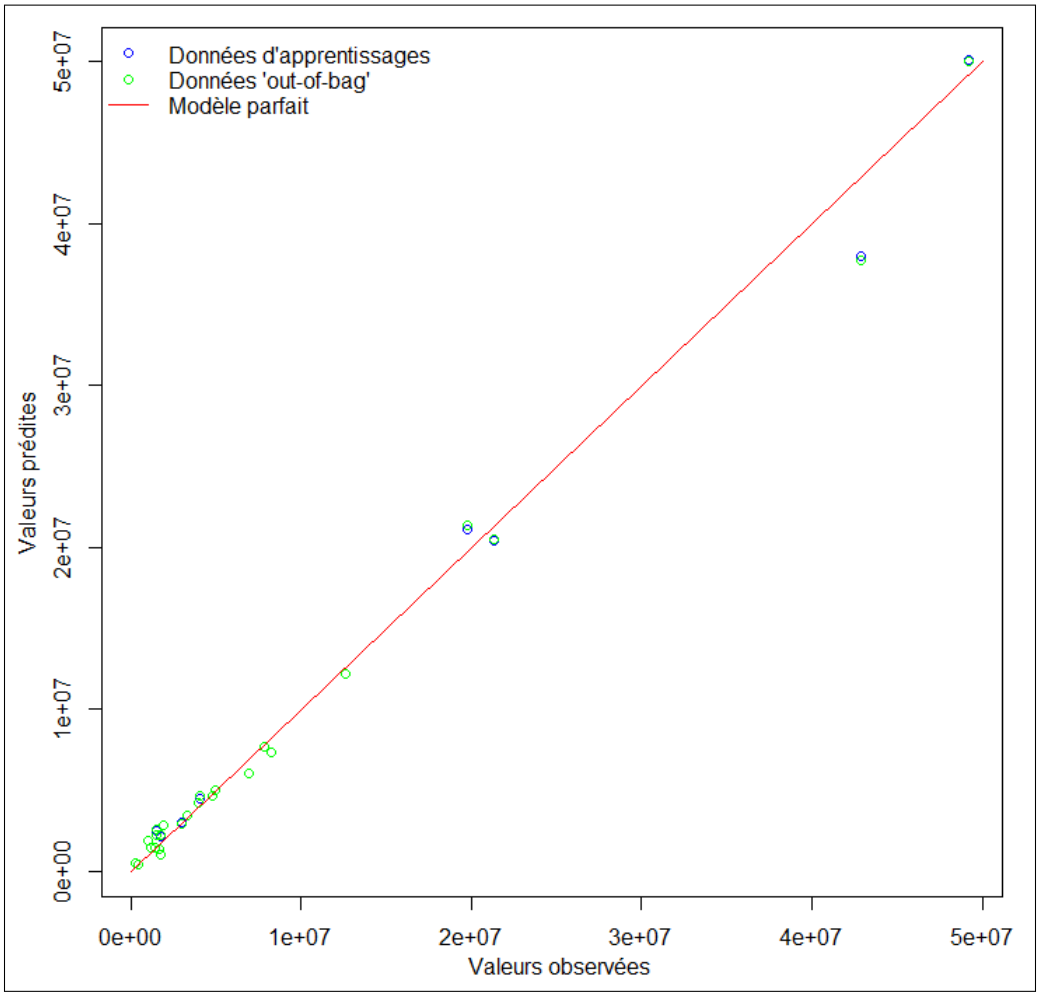

FIGURE 9.11 – QQ-plot du modèle de forêt aléatoire. En bleu, prédictions sur le jeu d'apprentissage. En vert, prédictions sur le jeu de validation

Enfin, nous pouvons voir sur la figure 9.12 l'importance mesurée de chacune des variables

explicatives utilisées. Nous remarquons que la distance à la cellule inondée la plus proche est celle qui a le plus d'impact sur la réduction d'erreur quadratique, ce qui renforce notre confiance dans les travaux menés dans la reproduction physique la plus fidèle possible des évènements historiques. Globalement, toutes les variables ont un certain impact et leur importance est assez homogène.

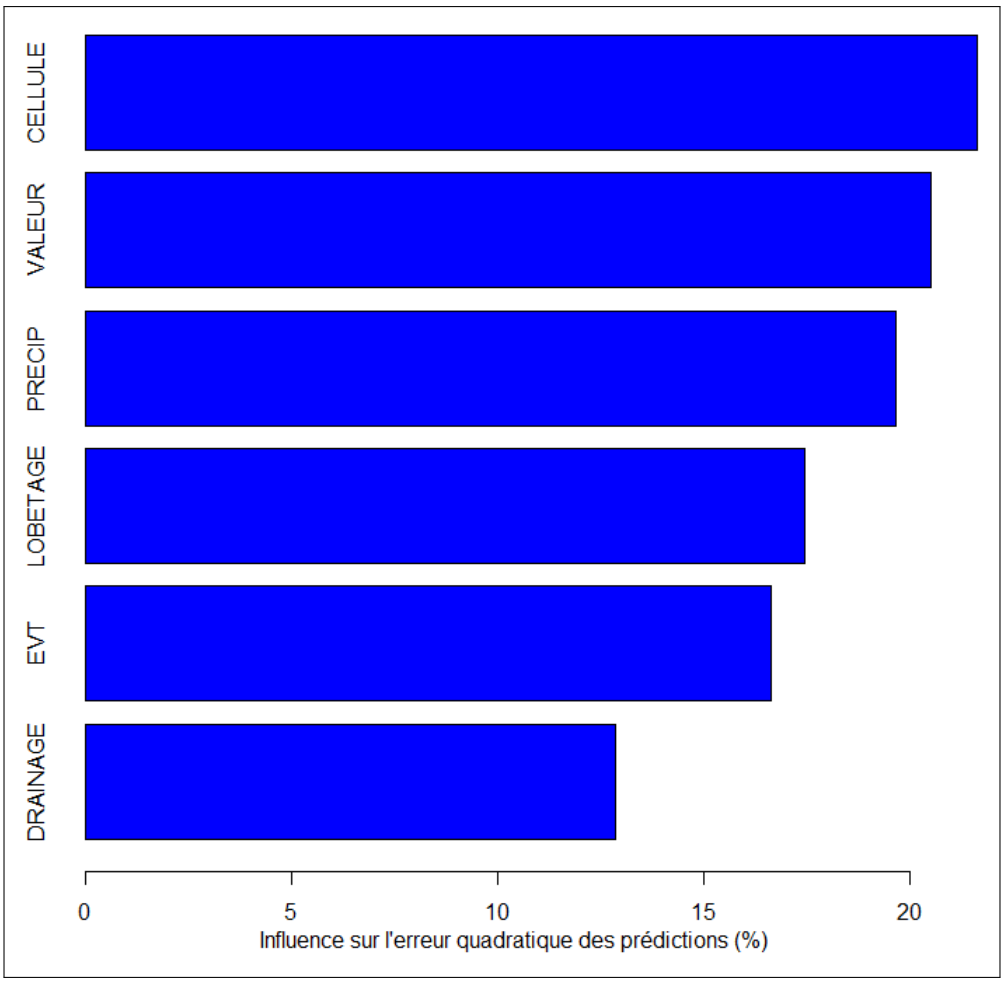

FIGURE 9.12 – Importance de chaque variable explicative dans le modèle de forêt aléatoire.

9.5 Conclusion

Nous avons montré dans cette partie de manière numérique les limites des *GLMs* face aux algorithmes d'apprentissage automatique tels que les *GBMs* et les forêts aléatoires. En effet, ceux-ci avaient de meilleurs rapports de déviance expliquée et présentaient des capacités de prédictions agrégées par évènement plus performantes.

Le choix entre le modèle *GBM* Gamma et la forêt aléatoire est moins évident. D'une part, le modèle *GBM* présente un bien meilleur rapport de déviance expliquée (34% vs 22%) mais la forêt aléatoire présente une plus grande robustesse dans la prédiction des pertes agrégées par évènement. Il n'est pas très surprenant que le *GBM* ait un meilleur rapport de déviance expliquée puisque c'est ce rapport même que l'algorithme cherche à maximiser au fur et à mesure de la construction des arbres lorsque la fonction de coût spécifiée est la déviance Gamma, tandis que ce rapport n'est aucunement pris en considération dans la construction de la forêt aléatoire. Ce rapport ne peut donc pas être retenu comme critère absolu puisqu'il ne représente pas l'objectif de notre démarche

qui est de prédire les pertes agrégées par évènement.

Ainsi, le modèle *GBM* aurait été le meilleur si l'on pouvait affirmer que le coût des sinistres était, conditionnellement aux variables explicatives, distribué selon une loi Gamma. Nous garderons alors la forêt aléatoire comme modèle d'estimation du coût des sinistres.

Enfin, notons que c'est le modèle qui requiert le moins d'hypothèses qui s'est avéré être le plus performant. Nous prouvons ainsi dans cette partie tout l'intérêt de la flexibilité et du peu d'hypothèses à formuler du *machine learning* .

10 Calibration de la probabilité de sinistre

De même que pour la vulnérabilité, nous allons ici calibrer les différents modèles d'apprentissage afin de modéliser le taux de sinistre, c'est-à-dire, la probabilité qu'un site assuré soit sinistré étant donné un évènement. En effet, le fait qu'un site assuré se trouve dans l'empreinte de l'évènement ne suffit pas à conclure que celui-ci sera sinistré à cause de l'incertitude sur l'empreinte et les moyens de protection du site contre les infiltrations d'eau. Nous considérons néanmoins que les sites assurés dont la cellule inondée la plus proche est à plus de 5 kms ne sont pas sinistrés, cette condition est vérifiée par la base sinistres que nous utilisons.

Les variables explicatives utilisées seront les mêmes que pour la vulnérabilité.

Enfin, il est à noter que le jeu de données est fortement déséquilibré puisqu'à peine 0.65% des sites assurés sont sinistrés. Ainsi, on cherchera à garder la même proportion des observations sinistrées dans chacun des jeux d'apprentissage et de validation, qui seront tirés de manière aléatoire tout en respectant cette condition.

10.1 Régression logistique

Nous commençons par la construction d'un modèle de régression logistique avec la fonction de lien *logit*.

Cette première méthode s'est avérée sans succès puisqu'elle mène à des probabilités de sinistre prédites toutes égales à 0. Cela est dû à la forte présence de zéros dans notre jeu de données limitant de manière importante le poids des sites sinistrés dans la calibration des paramètres. En effet, B. Owen montre dans son article sur les régressions logistique à déséquilibre infini [12] que lorsque la proportion des zéros tend vers l'infini, le coefficient nul diverge vers $-\infty$ et les autres coefficients convergent, menant après application de l'inverse de la fonction *logit* à des probabilités prédites nulles.

10.2 Forêts aléatoires

Nous calibrons maintenant un modèle de forêt aléatoires en fixant un taux de tolérance très bas. Le taux de tolérance fixe le seuil en deçà duquel le gain de variance lors d'une séparation de nœud n'est plus significative pour continuer à développer l'arbre. Fixer un seuil très bas permettrait de capter l'impact de la classe minoritaire sur les estimations finales.

On peut voir sur la figure 10.1 l'importance de chacune des variables explicatives utilisées. On remarque que les variables les plus discriminantes sont de nature physique. La forte importance de la distance à la cellule inondée la plus proche et de l'évènement concerné nous conforte encore une fois dans notre construction d'empreintes et dans notre motivation à vouloir capter les spécificités de destruction de chaque évènement historique. La variable la plus importante liée au site assuré est l'étage, ce qui correspond à la nature du péril modélisé. Enfin, nous constatons encore une fois la pertinence quasi nulle de la variable hauteur d'eau.

La courbe *ROC* du modèle est visible sur la figure 10.2 , l'aire sous la courbe est de 90.8% sur le jeu d'apprentissage et de 87.4% sur le jeu de validation. Les rapports de déviances de Bernoulli expliquées sont en revanche moins bonnes puisque nous obtenons à peine 6.3% sur le jeu d'apprentissage et de 4.2% sur le jeu de validation. Ce qui signifie que même si le modèle arrive à discriminer les sites sinistrés et le non sinistrés correctement la plupart du temps, il ne le fait avec

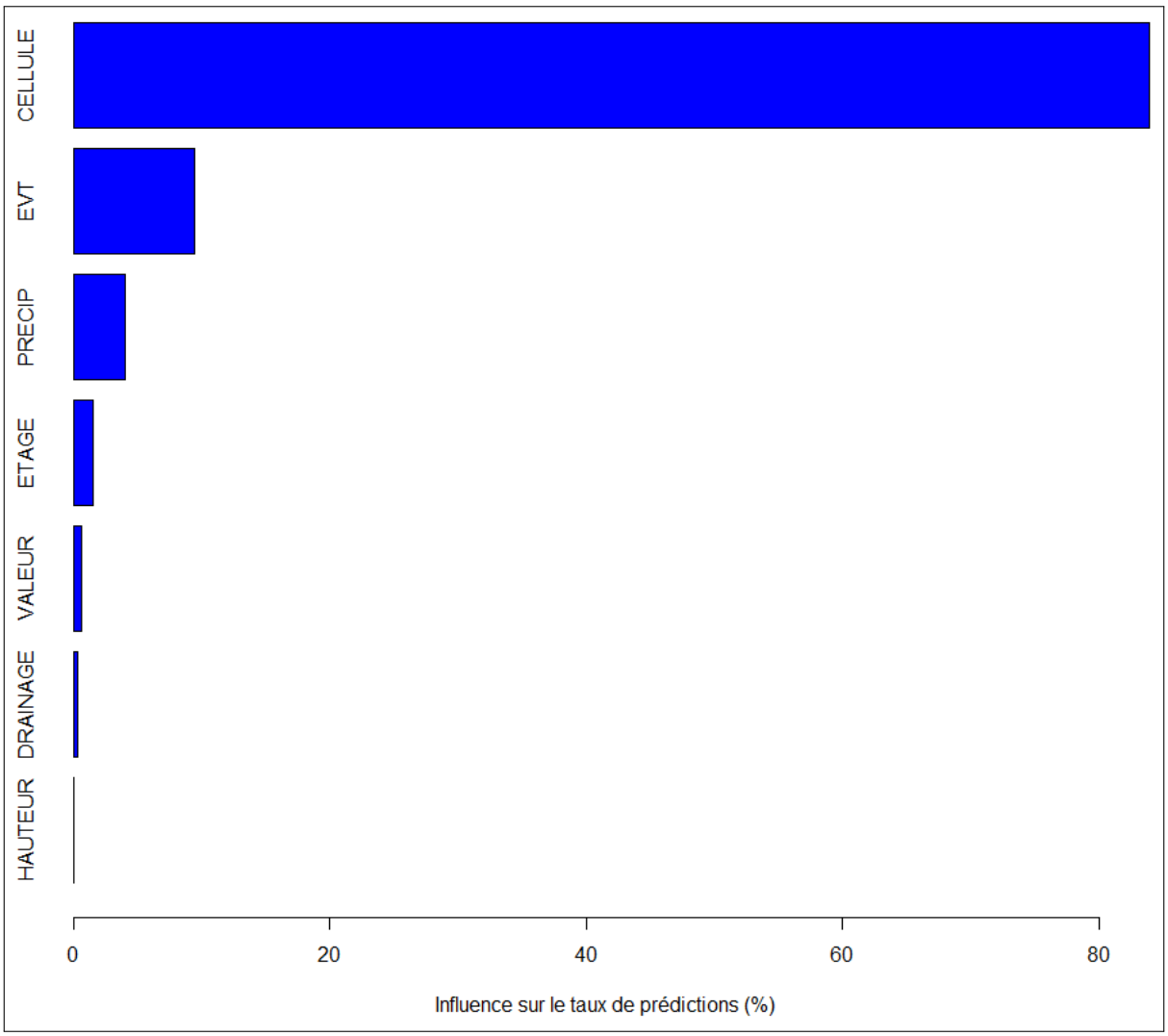

FIGURE 10.1 – Importance de chaque variable explicative dans le modèle de forêts aléatoires.

un écart suffisamment important pour expliquer significativement la variabilité sous-jacente.

Enfin, nous pouvons voir sur la figure 10.3 le QQ-plot par évènement de ce modèle combiné avec le modèle de la forêt aléatoire retenu pour la modélisation de la vulnérabilité dans la partie 9. On peut voir que les valeurs prédites sont éloignées des valeurs observées sans tendance particulière et que le modèle manque complètement l'estimation des évènements les plus coûteux.

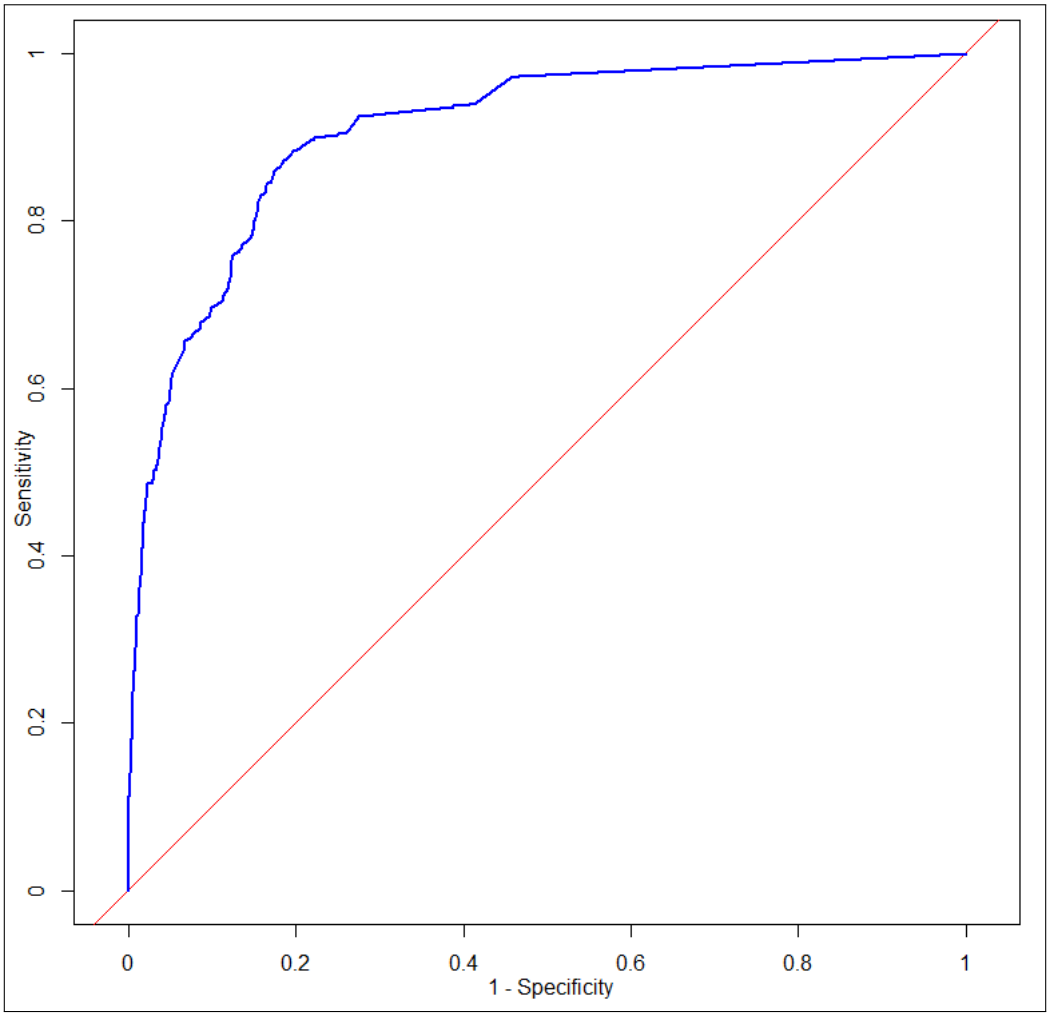

FIGURE 10.2 – Courbe *ROC* obtenue à partir des prédictions de la forêt aléatoire.

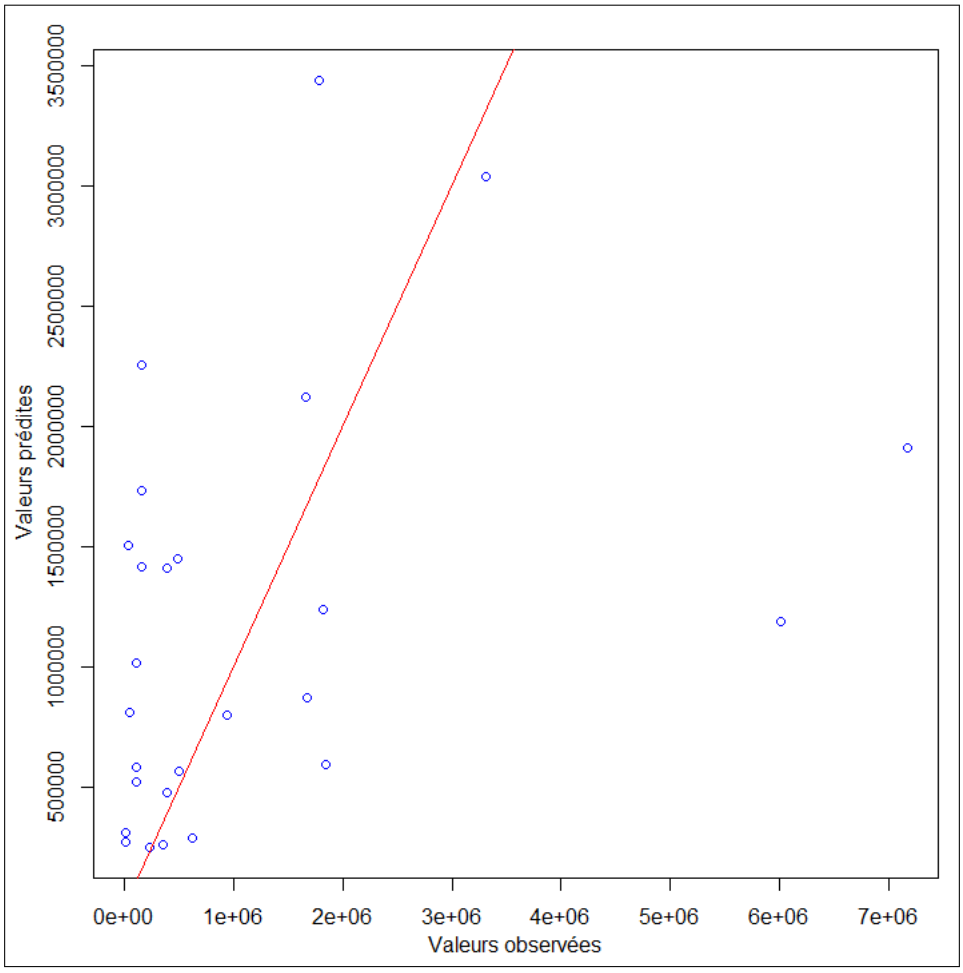

FIGURE 10.3 – QQ-plot par évènement avec les valeurs prédites par la forêt aléatoire combinés aux prédictions de sévérité obtenue dans la partie 9.

10.3 Gradient Boosting Machine

Nous allons maintenant calibrer un modèle *GBM* pour l'estimation de la probabilité de sinistre. Les paramètres retenus sont ceux qui permettent de minimiser la déviance de Bernoulli sur les jeux de validation et nous fixons encore une fois un seuil de tolérance à l'amélioration du modèle très bas afin de capter le poids très faible des sites sinistrés dans le jeu de données.

Nous pouvons voir sur la figure 10.4 l'influence relative de chacune des variables explicatives dans le modèle. L'importance des variables physiques et particulièrement celles de l'évènement et de la distance à la cellule la plus proche est encore confirmée.

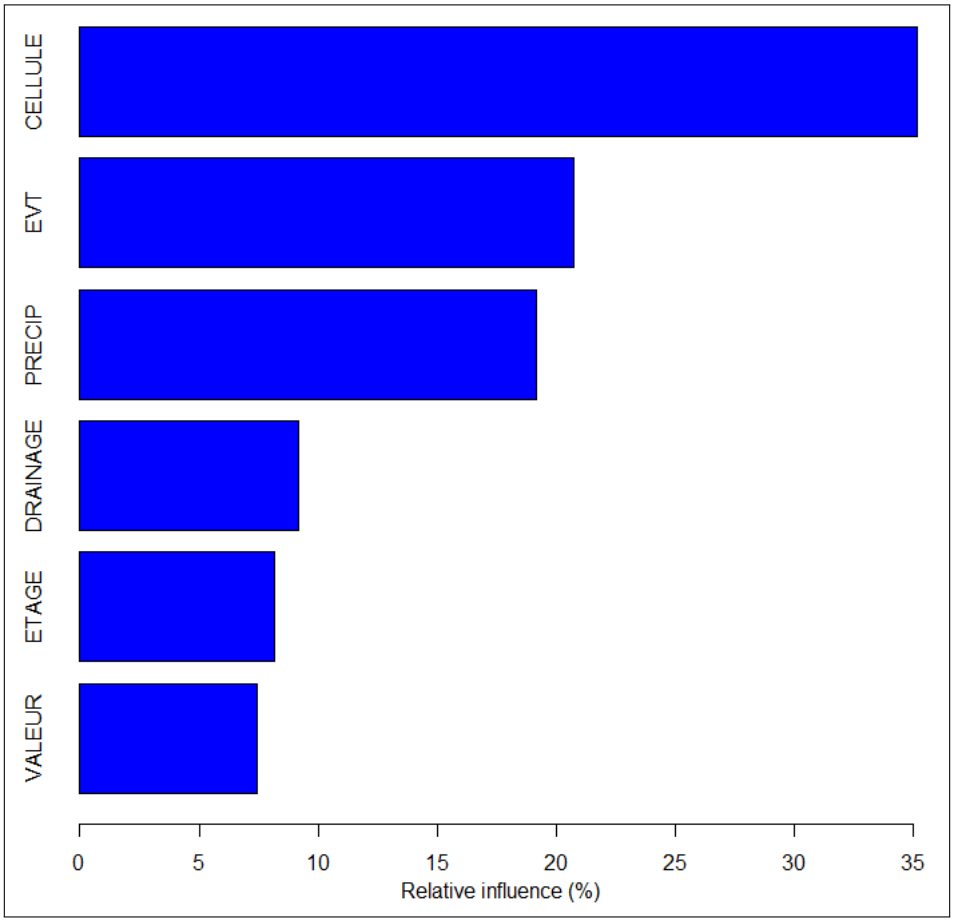

FIGURE 10.4 – Importance de chaque variable explicative dans le modèle *GBM*.

La courbe *ROC* du modèle est visible sur la figure 10.5 , l'aire sous la courbe est de 95.7% sur le jeu d'apprentissage et de 94.3% sur le jeu de validation. Les rapports de déviances de Bernoulli expliquées sont de 31.7% sur le jeu d'apprentissage et de 31.4% sur le jeu de validation. Ces résultats sont plus en ligne avec ce que l'on observait lors de la calibration de la vulnérabilité dans la partie 9. De plus, ils montrent que le sur-apprentissage est quasi-négligeable dans le modèle.

Enfin, nous pouvons voir sur la figure 10.6 le QQ-plot par évènement de ce modèle combiné avec le modèle de la forêt aléatoire retenu pour la modélisation de la vulnérabilité dans la partie 9. Les valeurs prédites suivent la même tendance que les valeurs observées avec des écarts contenus sauf pour un point extrême surestimé par le modèle de 35%.

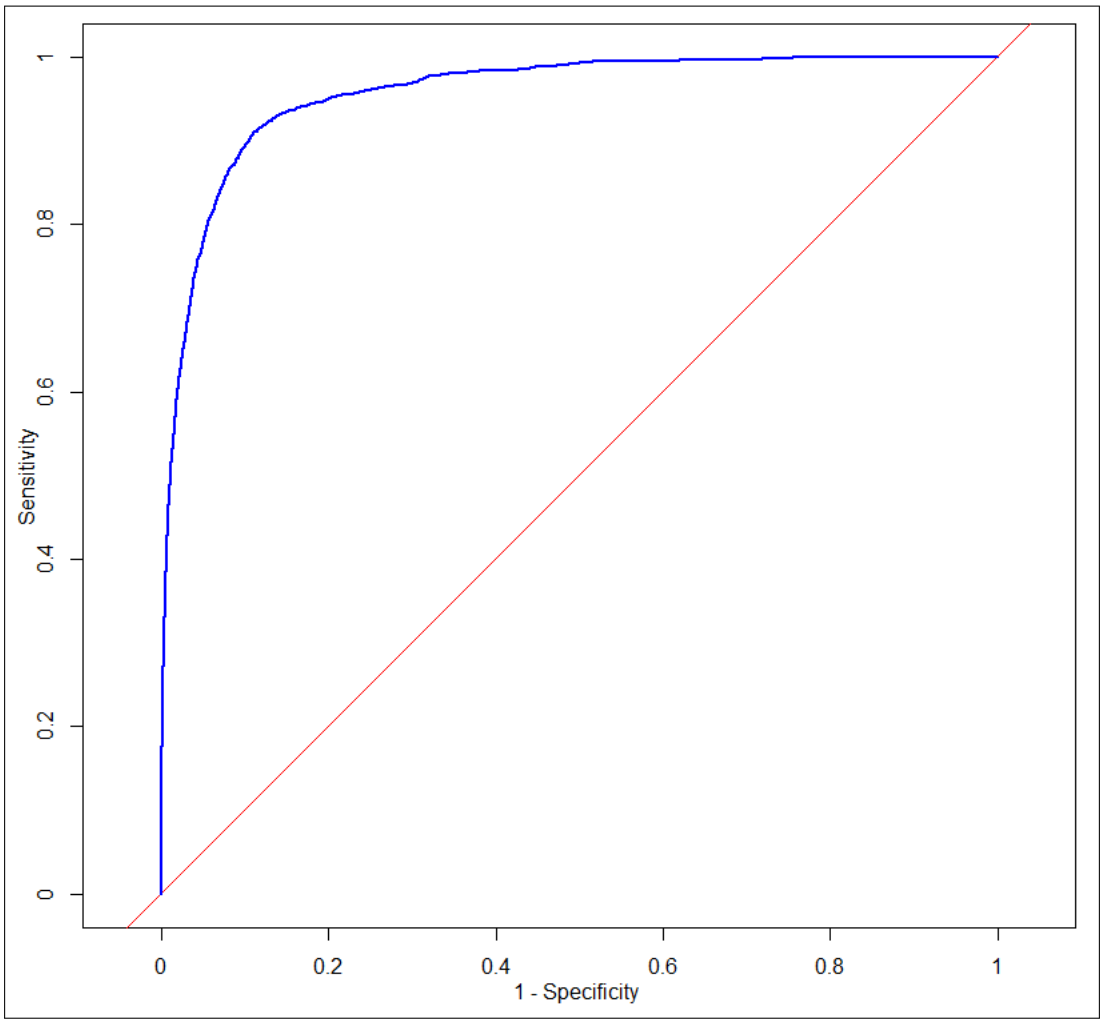

FIGURE 10.5 – Courbe *ROC* obtenue à partir des prédictions du modèle *GBM*.

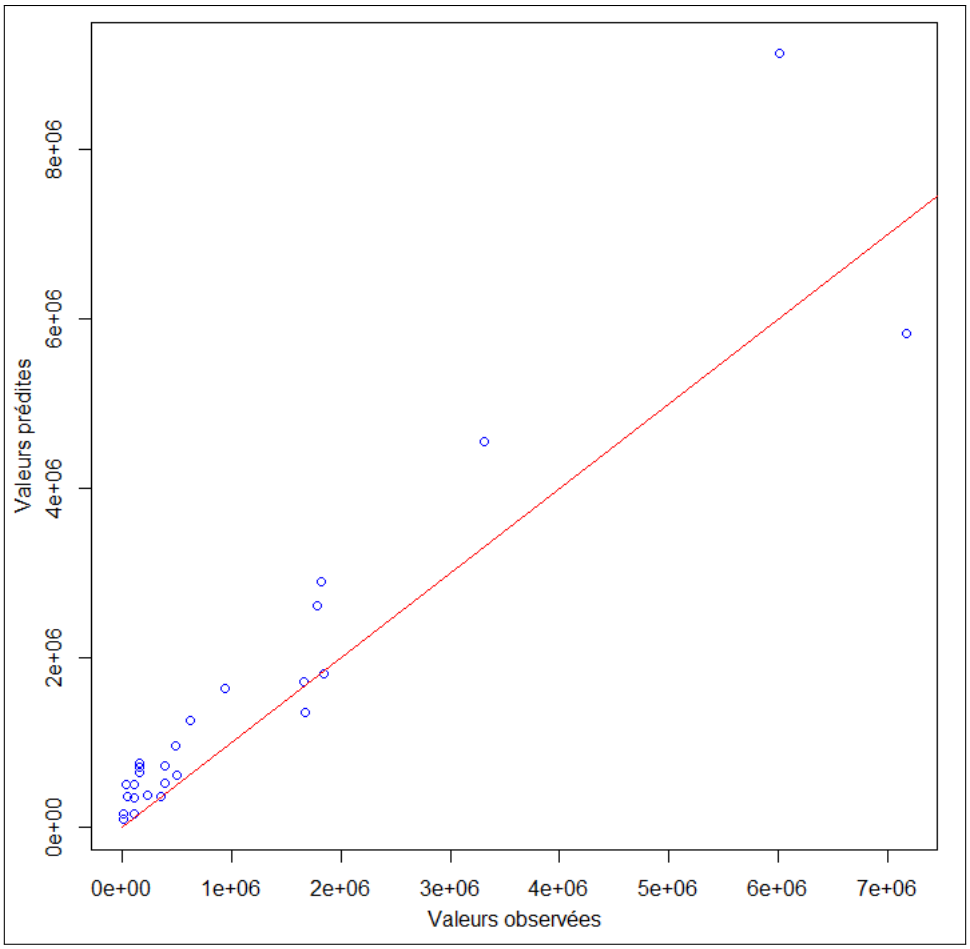

FIGURE 10.6 – QQ-plot par évènement avec les valeurs prédites par le modèle *GBM* combinés aux prédictions de sévérité obtenues dans la partie 9.

10.4 Conclusion

Au vu des différents indicateurs sur les prédictions de la forêt aléatoire et du modèle *GBM*, le dernier semble le plus approprié pour modéliser la probabilité de sinistre. Les rapports de déviance expliquée sont significatifs et la comparaison, après inclusion du modèle de vulnérabilité retenu dans la partie 9, entre les pertes par évènement prédites et les pertes observées montre une capacité de prédiction satisfaisante.

Cette différence avec la forêt aléatoire, peut s'expliquer par l'utilisation de la déviance de Bernoulli comme fonction de coût pour construire les différents arbres au lieu du critère de Gini. Les résultats numériques présentés dans cette partie laissent alors apparaître une meilleure robustesse des *GBMs* face aux jeux de données fortement déséquilibrés que les forêts aléatoires.

11 Estimation des pertes nettes [22]

11.1 Estimation du coût des évènements sur le portefeuille 2016

Nous avons construit deux modèles qui, en les combinant, nous permettent d'estimer la charge espérée d'un évènement E pour un site assuré de caractéristiques X_i. Le modèle de régression estime la charge sinistre sachant que le site est effectivement sinistré, nous noterons le résultat $R(X_i,\, E)$. Le modèle de classification permet d'estimer la probabilité que le site soit sinistré, nous noterons le résultat $C(X_i,\, E)$.

Pour un évènement E, la charge totale espérée brute de franchises sur le portefeuille 2016, notée S_E^{brut} est donnée par :

$$\begin{aligned} S_E^{brut} &= \sum_{i=1}^{N} E(S_i^E|X_i) \\ &= \sum_{i=1}^{N} E(S_i^E|X_i,\, S_i^E > 0) \;\times\; P(S_i^E > 0|X_i) \\ &= \sum_{i=1}^{N} R(X_i,\, E) \times C(X_i,\, E) \end{aligned}$$

où, N désigne le nombre de sites assurés dans le portefeuille 2016
S_i^E la variable aléatoire représentant la charge brute du i-ème site assurée causée par l'évènement E

11.1.1 Prise en compte des conditions financières

Nous devons à présent prendre en compte la franchise de chaque police afin de calculer la charge totale espérée par évènement nette de conditions financières.

Pour y parvenir, nous avons besoin de calculer l'écart type des prédictions $R(X_i,\, E)$. Comme le modèle retenu pour la vulnérabilité est une forêt aléatoire, il est aisé de calculer l'écart type des prédictions en calculant celui-ci sur chaque arbre (sur le nœud final permettant de faire la prédiction) avant de les agréger grâce à l'indépendance des arbres composant la forêt [23].

Nous devons maintenant associer une distribution aux charges prédites afin de pouvoir y appliquer les conditions financières. Nous présentons deux méthodes pour y parvenir :

- Pour chaque prédiction, nous pouvons choisir une distribution parmi une famille de lois de probabilités à support positif et adaptées à la modélisation d'une charge sinistre (loi log-normale, gamma, exponentielle, Pareto, etc). Ce choix pourra être fait à partir du vecteur des observations présentes dans tous les nœuds finaux dont est issue la prédiction et en se basant sur des tests statistiques comme celui de Kolmogorov Smirnov.

22. Note au lecteur, nous utiliserons dans cette partie les termes polices et sites sans distinction par abus de langage.

23. Notons que cette méthode n'aurait pas pu être appliquée avec un modèle *GBM* car les arbres ne sont pas indépendants. Il aurait fallu faire l'apprentissage directement sur les pertes nettes en incluant la franchise dans les variables explicatives, donnée qui n'était pas toujours disponible dans notre base.

- La deuxième méthode repose sur la théorie de l'information. Elle consiste à choisir la distribution qui maximise l'entropie conditionnellement à l'information disponible. L'entropie étant une mesure d'incertitude, l'idée suggérée par cette méthode est de choisir la distribution qui minimise la quantité d'information a priori contenue dans la loi choisie [24]. Par ailleurs, nous pouvons montrer que la distribution qui maximise l'entropie parmi toutes les lois à support strictement positif, d'espérance et d'écart type connus est la loi log-normale [25].

Pour limiter la complexité du modèle et la volatilité qui en découlerait, nous choisissons la deuxième méthode. Ainsi, pour un évènement E, la charge totale nette espérée sur le portefeuille 2016, notée S_E est donnée par :

$$\begin{aligned} S_E &= \sum_{i=1}^{N} E((S_i^E - d(X_i))_+ | X_i) \\ &= \sum_{i=1}^{N} E((S_i^E - d(X_i))_+ | X_i,\, S_i^E > 0) \;\times\; P(S_i^E > 0 | X_i) \\ &= \sum_{i=1}^{N} f(R(X_i,\, E),\; \sigma(R(X_i,\, E)),\; d(X_i)) \;\times C(X_i,\, E) \end{aligned} \tag{11.1}$$

où, $d(X_i)$ désigne la franchise du i-ème site assuré
$f(a,\, b,\, c) = E((Y - c)_+)$ où Y suit une loi log-normale d'espérance a et d'écart type c, simulée par la méthode de Monte Carlo.

Dans la suite on notera :

$$R'(X_i,\, E) \;=\; f(R(X_i,\, E),\, \sigma(R(X_i,\, E)),\, d(X_i))$$

11.1.2 Prise en compte des données manquantes

Comme indiqué dans la partie 1 lors de la description des données à notre disposition, plusieurs données ne sont pas exploitables et le modèle actuel n'en tient pas encore compte. Ainsi, la formule d'estimation de la charge espérée par évènement 11.1 se limite au cadre des données exploitables et ne capture pas la totalité de la charge historiquement observée.

Pour rappel, voici les données qui ont été exclues dans la phase de calibration des modèles d'apprentissage :

- Modèle de régression :
  - Sinistres sans information sur la branche d'activité
  - Sinistres non géolocalisés
- Modèle de classification :
  - Portefeuilles historiques des branches autres que la branche MRH (sauf pour 2016)
  - Les polices résidentielles non exploitables des portefeuilles historiques car les coordonnées géographiques sont manquantes
  - Les sinistres utilisées dans le modèle de régression des polices non exploitables dans les portefeuilles historiques

24. Ces notions sont présentées plus en détail dans l'article de K. Conrad (2013) [28]
25. Lorsque l'écart type n'est pas connu, la distribution d'entropie maximale est la loi exponentielle. Celle-ci aurait alors été utilisée si le modèle retenu était un *GBM*.

Nous devons alors associer une distribution aux informations manquantes afin de pouvoir estimer l'espérance des pertes totales pour chaque évènement. Ne disposant d'aucune information discriminante, nous ferons l'hypothèse que les informations manquantes sont indépendantes de leurs valeurs (non observables). Ce qui revient à approximer la distribution des informations manquantes par celle des valeurs observées, en d'autres termes, nous faisons l'hypothèse que les informations manquantes suivent la même distribution que celles des observations exploitables.

Selon s'il s'agit du modèle de régression ou de classification, nous appliquerons un facteur multiplicatif aux estimations de sorte à prendre en compte respectivement la charge sinistre non exploitable ou le nombre de polices et de sinistres non exploitables.

Formellement, cela revient à définir :

$$R''(X_i,\, E) \;=\; R'(X_i,\, E) \times \frac{C_{expl.}(E,\, B(X_i)) \,+\, C_{excl.}(E,\, B(X_i)) \,+\, p(E,\, B(X_i))\, C(E,\, inconnue)}{C_{expl.}(E,\, B(X_i))}$$

où, $B(X_i)$ est la branche d'activité du site assuré numéro i,

$C_{expl.}(E,\, B(X_i))$ est la charge totale historique exploitable causée par E pour la branche $B(X_i)$,

$C_{excl.}(E,\, B(X_i))$ est la charge totale historique exclue (pas d'information géographique) causée par E pour la branche $B(X_i)$,

$C(E,\, inconnue)$ est la charge totale historique de branche d'activité inconnue causée par E,

$p(E, B) \;=\; \dfrac{C_{expl.}(E,\, B)}{\sum_{branche\ b} C_{expl.}(E,\, b)}$ est la proportion de la charge exploitable causée par E de la branche B parmi toutes les branches.

et, en vérifiant que nous n'aurons pas des valeurs de C' strictement supérieures à 1 :

$$C'(X_i,\, E) \;=\; C(X_i,\, E) \times \frac{S_{expl.}(E,\, MRH)}{S_{retrouv.}(E,\, MRH)} \times \frac{P_{expl.}(A(E),\, MRH)}{P_{tot.}(A(E),\, MRH)} \times \frac{S_{expl.}(E, B(X_i))}{S_{expl.}(E, MRH)} \times \frac{P_{tot.}(2016,\, MRH)}{P_{tot.}(2016,\, B(X_i))}$$

où, $A(E)$ est l'année d'occurrence de l'évènement E,

$S_{expl.}(E, b)$ est le nombre total de sinistres exploitables (branches d'activité connues et géolocalisés) causés par E pour la branche b,

$S_{retrouv.}(E, MRH)$ est le nombre total de sites MRH sinistrés par E et qui ont pu être retrouvés dans le portefeuille exploitable,

$P_{expl.}(a, MRH)$ est le nombre de polices MRH exploitables dans le portefeuille historique de l'année a,

$P_{tot.}(a, b)$ est le nombre de sites de la branche b assurées pendant l'année a.

Notons que nous ne cherchons pas à capturer la partie des sinistres non exploitables dans la renormalisation des estimations de taux de sinistres car cette part est déjà prise en compte dans la renormalisation de l'estimation des charges sinistre. Le but de l'étude étant d'estimer la charge finale par évènement et non le nombre de sinistres, il est plus pertinent de chercher à récupérer la charge sinistre non exploitable plutôt que leur nombre, une telle approche supposerait que le coût moyen par sinistre est le même dans chacune des deux parties de la base des sinistres.

Enfin, nous aurions pu prendre en compte les données manquantes directement lors de la calibration des modèles en attribuant un poids à chaque observation. Ceux-ci auraient permis de pondérer les différents calculs de mesures d'erreur et de log-vraisemblance et ainsi les prédictions finales (avec des moyennes et votes pondérés dans chaque nœud terminal des *GBMs* et forêts aléatoires). Il y aurait alors eu deux manières de s'y prendre :

- Les poids des observations sont déterminés par les facteurs calculés plus haut. Cette méthode permet de capturer la charge totale mais présente néanmoins l'inconvénient majeur de confondre une observation de poids m et m occurrences d'une même observation et donc de négliger la variabilité de la variable réponse conditionnellement aux variables explicatives de l'observation. Cela revient à surestimer le poids accordé à ces observations dans les prédictions finales par rapport aux observations moins censurées (et donc de poids inférieurs) et qui montreront plus de variabilité [26].
- En intégrant les observations exclues de la manière suivante :
  - Déterminer les observations exploitables similaires (i.e. de même évènement et, si disponible, même branche d'activité) à l'observation exclue.
  - Dupliquer l'observation exclue autant de fois qu'il y a de vecteurs de variables explicatives différents dans les observations exploitables similaires.
  - Attribuer à chaque duplication un vecteur de variables explicatives possibles.
  - Le poids de chacune des duplications est déterminé par la probabilité empirique (taux d'occurrence) du vecteur de variables explicatives associé parmi toutes les valeurs possibles.

  Cette méthode repose toujours sur l'hypothèse que les variables explicatives des observations exclues suivent la même distribution que celle des observations exploitables et présente l'avantage par rapport à notre méthode d'inclure toutes les données et ainsi de supprimer l'hétérogénéité de la proportion des données manquantes entre les différents évènements. Elle introduit néanmoins plus

26. J.M. Brick et G. Kalton expliquent ces propos plus en détail et les différentes manières de gérer les données manquantes dans [29].

de complexité et augmente le risque de présenter des données bruitées à la calibration du modèle [27] impactant l'évaluation de la relation entre les variables explicatives et la variable réponse.

Nous adaptons donc la formule 11.1 d'estimation de la charge espérée par évènement :

$$S_E = \sum_{i=1}^{N} (R''(X_i, E) \times C'(X_i, E)) \tag{11.2}$$

11.2 Résultats numériques

Nous pouvons comparer sur la figure 11.1 les estimations de chaque évènement obtenus avec le modèle développé et le modèle actuellement utilisé qui consiste à réactualiser les pertes historiques en appliquant le taux de l'inflation et l'évolution globale de la somme totale assurée sur tout le portefeuille. Les évolutions restent contenues et s'étalent de -22% à $+14\%$ et l'évolution moyenne est de -4%. Il n'est pas surprenant d'avoir une évolution globale à la baisse car l'assureur adapte régulièrement sa politique de souscription à la suite d'évènements d'inondation.

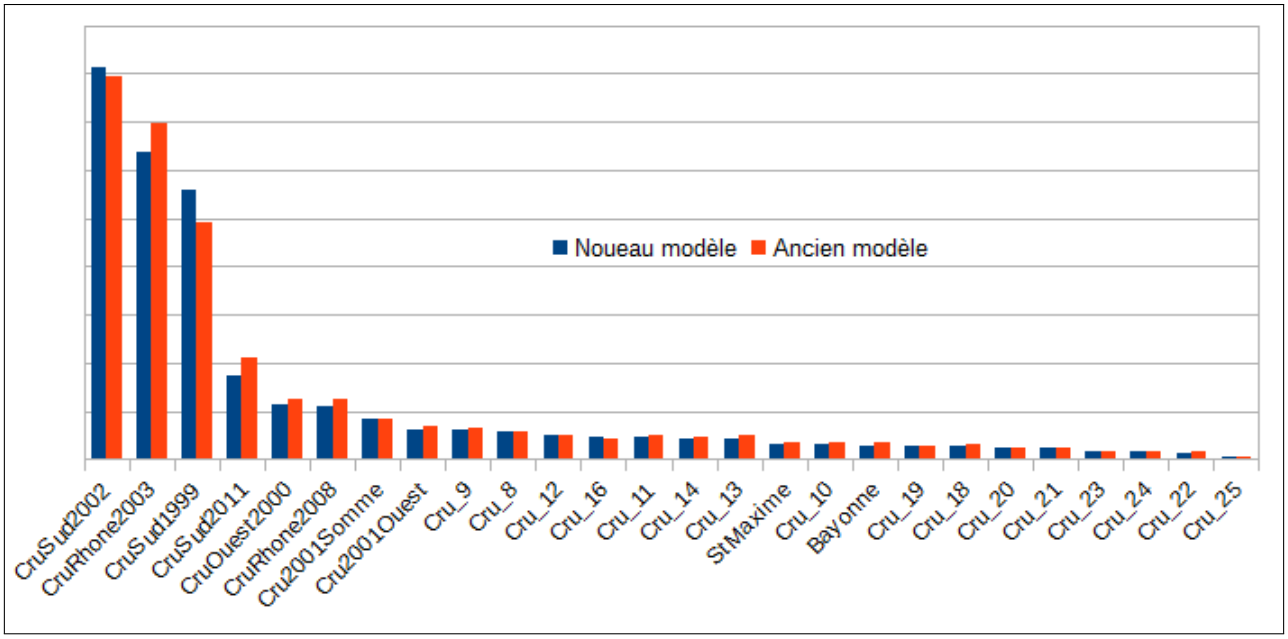

FIGURE 11.1 – Coût estimé pour chaque évènement, comparaison avec la méthode actuelle.

11.3 Calibration du modèle fréquence sévérité

Nous allons maintenant utiliser les estimations de coût de chaque évènement $(S_1, \ldots, S_{27})$ afin de modéliser la charge annuelle liée au risque d'inondations extrêmes porté par notre portefeuille. Cette charge sera représentée par une variable aléatoire que l'on notera X.

La fréquence des évènements étant indépendante de leurs coûts, nous choisissons de modéliser Y par un modèle fréquence coût :

$$X = \sum_{i=1}^{N} X_i$$

27. Cette approche est intéressante et mérite d'en étudier l'impact. Malheureusement, les délais de cette étude n'ont pas permis de l'implémenter.

où, N est la variable aléatoire représentant le nombe d'évènements survenant dans l'année,
X_i est la variable aléatoire représentant le coût du i-ème évènement.

Il s'agit maintenant de modéliser N et X_i par des lois de probabilités que l'on ajustera à partir de l'historique.

11.3.1 Lois de fréquence

Les lois prises en considération pour la modélisation de la fréquence N sont classiquement les lois **de Poisson** et **binomiale négative**. La différence majeure entre ces deux lois réside dans la dispersion de l'aléa que l'on cherche à modéliser. En effet, la loi de Poisson a la propriété d'avoir une variance égale à l'espérance, on parle alors de distribution équi-dispersée, alors que la loi binomiale négative permet de modéliser des aléas montrant plus de variabilité puisque qu'elle présente une variance supérieure à l'espérance [28].

11.3.2 Lois de sévérité

Comme présenté en introduction, nous ne cherchons pas à modéliser le risque d'inondation dans sa totalité mais uniquement les évènements générant des pertes supérieures à 1 million d'euros. Ainsi, nous modélisons la sévérité X_i avec des lois appartenant à la famille des distributions des valeurs extrêmes généralisées [29]. Nous distinguons dans cette famille trois sous-familles, selon la vitesse de décroissance vers 0 des densités des distributions dont on cherche à modéliser les extrêmes [30] :

- la famille **de Fréchet** dite à queue épaisse (loi non bornées et décroissance exponentielle),
- la famille **de Gumbel** dite à queue fine (loi non bornée et décroissance polynomiale),
- la famille **de Weibull** dite à queue finie (loi bornée à droite).

Ainsi, nous choisissons une loi de chacune des trois familles pour modéliser la sévérité : la loi log normale (famille de Gumbel), la loi de Weibull tronquée à droite [31] (famille de Weibull) et la loi de Pareto (famille de Fréchet). Enfin, nous examinerons également la loi de Pareto Généralisée qui est adaptée à la modélisation des dépassement de seuils extrêmes.

11.3.3 Critères d'ajustement

Nous estimons les paramètres de chacune des lois prises en considération par la méthode de maximum de vraisemblance et des moments.

Parmi les différents indicateurs numériques et tests statistiques permettant de juger l'adéquation d'une distribution avec les observations, nous retenons les suivants :

- La distance de **Kolmogorov Smirnov** mesure la distance entre les fonctions de répartition modélisée et empiriques en calculant :

$$D_n = sup\,|F(x) - U_n(x)|$$

28. On peut d'ailleurs montrer qu'une loi binomiale négative correspond à une loi de Poisson dont le paramètre suit une loi Gamma. Ainsi, une loi binomiale négative correspond à une loi de Poisson après introduction d'une variabilité sur son paramètre.
29. Cette famille de distributions, issue de la théorie des valeurs extrêmes, permet de modéliser les maximas d'un vaste ensemble de distributions.
30. On parle aussi d'épaisseur de la queue de distribution.
31. La valeur maximum possible pourra alors être donnée par l'exposition totale au risque d'inondation

où n est le nombre d'observations, F la fonction de répartition modélisée et U_n la fonction de répartition empirique.

On peut montrer que $\sqrt{n}\,D_n$ suit asymptotiquement une loi de Kolmogorov sous l'hypothèse :

$$H_0 : F = U$$

où U est la distribution théorique de la variable aléatoire que l'on cherche à modéliser.
Ce qui nous permet de construire un test statistique sur l'adéquation de la distribution modélisée lorsque n est suffisamment grand.

- La distance de **Cramér-von Mises** est également une distance entre les fonctions de répartition théoriques et empiriques obtenue en calculant :

$$T_n = n \int_{-\infty}^{+\infty} [U_n(x) - F(x)]^2 \, dF(x)$$

On peut montrer, pour les valeurs observées rangées dans l'ordre croissant $x_1, \ldots, x_n$, que :

$$T_n = \frac{1}{12\,n} + \sum_{i=1}^{n} \left[\frac{2\,i-1}{2n} - F(x_i) \right]^2$$

.
On montre que T_n suit une loi de Cramér-von Mises à n degrés de libertés [32].

Ce qui nous permet de construire un test statistique sur l'hypothèse H_0. Le test de Cramér-von Mises peut alors être une alternative au test de Kolmogorov Smirnov lorsque nous ne disposons pas de suffisamment de données. Néanmoins, il ne peut pas être utilisé pour des distributions discrètes.

- Le critère **AIC** [33] qui est une mesure utilisée en théorie de l'information pour estimer la quantité d'information perdue lorsqu'un modèle est utilisé pour décrire une réalité. En statistiques, il tient compte de l'adéquation du modèle par rapport aux observations et de sa complexité. Il s'agit néanmoins d'un critère relatif permettant de comparer deux modèles et non d'un critère absolu permettant de réaliser un test statistique. Il est calculé par :

$$AIC = 2k - l$$

où k représente le nombre de paramètres utilisés dans le modèle et l la log-vraisemblance du modèle.

Comme nous disposons de peu de données pour utiliser le test de Kolmogorov Smirnov pour les lois discrètes, le choix entre la loi binomiale négative et de Poisson se fera selon la dispersion des observations et le critère AIC.

Résultats

Pour la fréquence, les observations montrent une variance supérieure de 40% à l'espérance montrant alors une sur-dispersion et le critère AIC est minimisé par loi binomiale négative qui est donc retenue.

32. Le lecteur intéressé par la formulation exacte et l'algorithme d'implémentation de cette loi pourra consulter : Csörgo, S. and Faraway, J.J. (1996) The exact and asymptotic distributions of Cramér-von Mises statistics. Journal of the Royal Statistical Society, Series B 58, 221–234.

33. Akaike Information Criterion

Pour la sévérité, les distances de Kolomogrov Smirnov et Cramér-von Mises sont respectivement minimisées par la loi de Pareto généralisée et la loi log normale dont les paramètres ont été estimés par maximum de vraisemblance. Par ailleurs, le critère AIC est minimisé par la loi log normale qui sera alors retenue.

Néanmoins les p-valeurs des deux tests utilisés sont très faibles (inférieures à 1%), ce qui permettrait de rejeter avec une confiance confortable l'adéquation des lois choisies. Ce résultat n'est pas surprenant compte tenu de la faible occurrence des évènements modélisés et de la longueur relativement courte de notre historique. C'est pourquoi nous chercherons dans la partie suivante à crédibiliser notre modèle par des scénarios de pertes de marché qui proviennent de la simulation des conséquences financières d'une éventuelle crue de la Seine.

11.4 Crédibilisation par des scénarios

Outre la courte période de temps observée, nous ne disposons que de 3 évènements pouvant être jugés comme extrêmes comme nous pouvons le voir sur la figure 11.1. Ainsi, notre historique n'est pas suffisamment exhaustif pour être représentatif des évènements majeurs tels que la crue de la Seine de 1910 par exemple. Nous proposons alors de le crédibiliser par une courbe de marché, c'est-à-dire se servir du modèle construit pour modéliser les évènements à forte probabilité où l'historique est suffisant et modéliser les évènements extrêmes à faible probabilité à partir d'études de marché moins spécifiques au portefeuille et l'historique d'AXA. Le scénario retenu est celui d'une crue de la Seine, celui-ci est particulièrement intéressant au vu des accumulations des risques présentes dans la région parisienne.

Nous introduisons pour cela la notion de période de retour d'un évènement de coût x qui correspond à l'inverse de la probabilité d'avoir un évènement durant l'année d'un coût supérieur à x. Formellement,

$$\begin{aligned} T(x) &= \frac{1}{1 - P(\underset{1\leq i\leq N}{max}(X_i))} \\ &= \frac{1}{1 - M_N(F_{X_i}(x))} \end{aligned}$$

où, M_N est la fonction génératrice des moments de la variable aléatoire N,
F_{X_i} est la fonction de répartition de la sévérité.

On définit également, par abus de langage, la courbe OEP[34], qui associe à une période de retour t la perte x dont la probabilité de dépassement est un $\frac{1}{t}$. Formellement,

$$OEP : t \mapsto T^{-1}(t)$$

On dit ainsi, qu'une perte x a pour période de retour t lorsque $OEP(t) = x$ et on peut montrer que le temps d'attente entre deux évènements de coûts supérieurs à x est en moyenne égal à t, d'où l'appellation de période de retour.

34. *Occurrence Exceedance Probability* représente en théorie la probabilité de survenance d'un évènement dépassant un certain coût.

Notons respectivement OEP_1 et OEP_2 les courbes OEP du modèle historique et du modèle de marché. La courbe OEP issue d'un mélange des deux modèles est obtenue par :

$$OEP(t) = \begin{cases} OEP_1(t), & \text{si } t \leq T_1 \\ (1-\alpha(t))\, OEP_1(t) + \alpha(t)\, OEP_2(t), & \text{si } T_1 < t \leq T_2 \\ OEP_2(t), & \text{si } t > T_2 \end{cases}$$

$$\text{où,} \quad \alpha(t) = \frac{t - T_1}{T_2 - T_1} \text{ est la fonction de poids des deux courbes,}$$
T_1 et T_2 sont les seuils de mélange des deux courbes.

Ainsi, nous accordons une crédibilité totale à notre historique sur les faibles périodes de retour inférieures à T_1, celle-ci diminue ensuite linéairement jusqu'en T_2 où la crédibilité devient nulle.

Afin de fixer les seuils T_1 et T_2, nous allons étudier la sensibilité de la courbe OEP de notre historique par rapport à chacun de ses points. Nous allons construire 26 modèles fréquence coût en retirant à chaque tour un des 26 évènements étudiés et tracer les courbes OEP qui en découlent. Nous pourrons ainsi évaluer la robustesse de notre modèle en fonction des périodes de retour. Les seuils T_1 et T_2 seront alors déterminés par les périodes de retour où l'on observe une déviation par rapport à la courbe OEP de référence (avec tous les évènements) de respectivement 5% et 10%. Nous pouvons voir sur la figure 11.2 les résultats de cette étude, les seuils retenus sont alors de 9 et 23 ans.

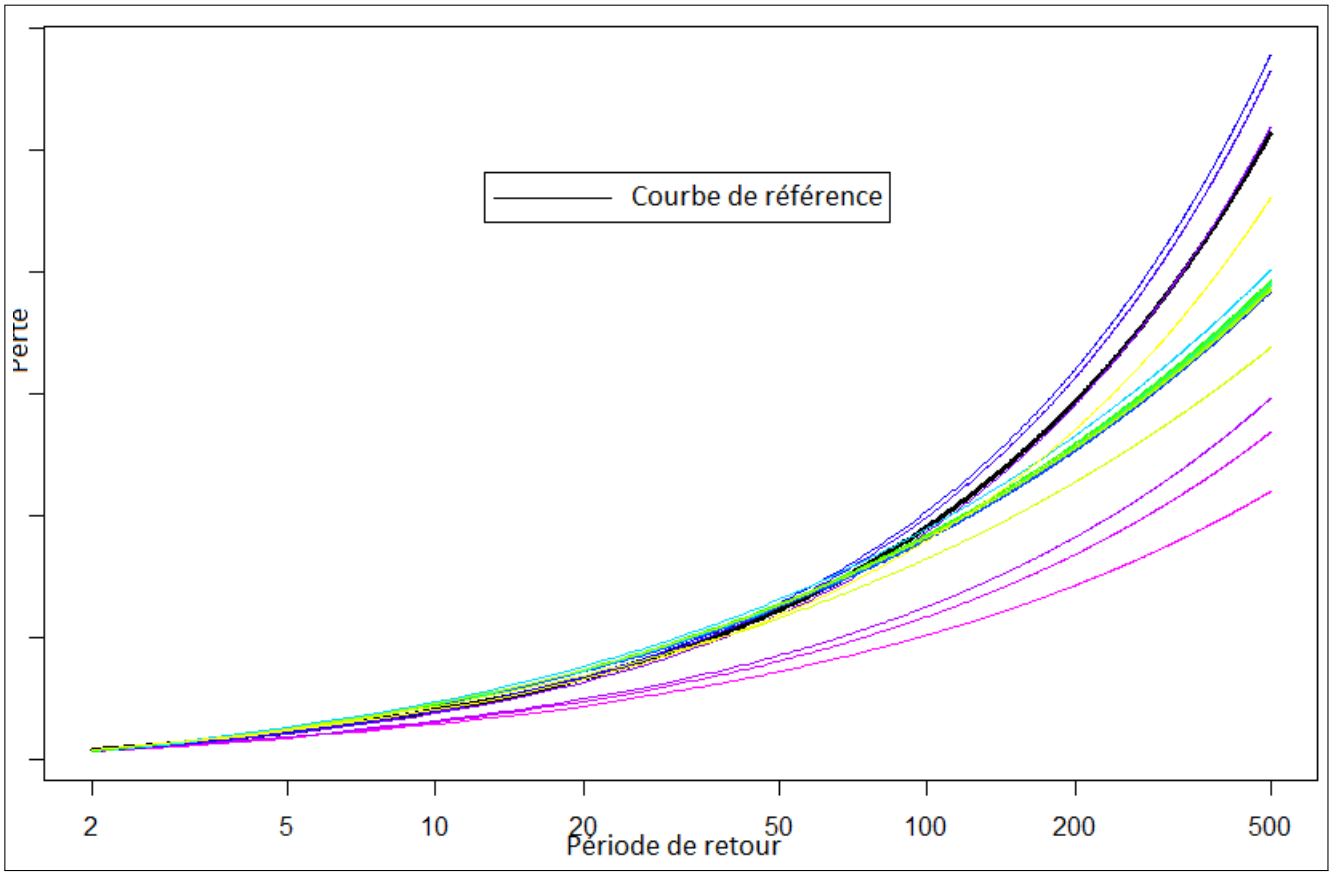

FIGURE 11.2 – Sensibilité de la courbe OEP du modèle. Courbe de référence en noir.

La fonction de répartition finale crédibilisée est obtenue en inversant la fonction génératrice des moments par des méthodes numériques ou analytique lorsque c'est possible comme pour les lois binomiale négative et de Poisson.

12 Impact sur la réassurance

Nous proposons dans cette partie d'exploiter les résultats du modèle développé à des fins d'optimisation du résultat technique et du capital économique de l'assureur. Nous cherchons pour cela la structure optimale de réassurance répondant à nos besoins.

12.1 Introduction à la réassurance [35]

Un contrat de réassurance permet à l'assureur de céder une partie de son risque afin de réduire la volatilité de son résultat technique, limiter son besoin en capital économique ou encore à réduire simplement l'incertitude éventuelle sur sa souscription. Nous décrivons brièvement les structures de réassurance généralement proposées :

- **La réassurance proportionnelle** consiste à partager les coûts de chaque sinistre entre la cédante et le réassureur selon un taux de cession fixé au contrat. Ce type de réassurance est rarement intéressant car l'assureur cède l'ensemble de la distribution de son risque et lui permet uniquement de diminuer la charge des sinistres. De plus, elle coute cher car elle implique des frais de gestion (sinistre par sinistre) importants.
  Elle est souvent utilisée lorsque l'assureur n'a pas suffisamment d'expertise dans les risques souscrits (lors du lancement d'un nouveau produit d'assurance par exemple).
- **La réassurance en excédent de sinistre** fait intervenir le réassureur uniquement lorsque le montant de sinistre dépasse un certain seuil appelé la priorité et son engagement est limité à ce que l'on appelle la portée. Ce type de réassurance permet de céder la partie supérieure à un certain seuil du risque porté par l'assureur. Cette structure est bien plus avantageuse pour la cédante car elle lui permet de céder la partie la plus volatile de son risque et qui mobilise le plus de capital économique.

Au vu de la flexibilité offerte et de la confiance que nous avons dans la qualité de notre portefeuille, nous privilégierons la structure non proportionnelle comme choix de réassurance pour le risque d'inondations extrêmes de notre portefeuille.

Les traités de réassurance en excédent de sinistre (ou *Excess of loss* ou simplement XS) sont notés *portée* XS *priorité*. La figure 12.1 montre par exemple le fonctionnement d'un traité 50m XS 10m, c'est à dire un traité de priorité 10 millions d'euros et de portée 50 millions pour 5 montants de sinistre différents.

Nous distinguons les traités XS par risque et par évènement :

- Un traité XS **par risque** s'applique à chaque police d'assurance présente dans le portefeuille. Si aucune police ne subit de sinistre supérieur à la priorité du traité, celui-ci n'est pas enclenché.
- Un traité XS **par évènement** s'applique à toutes les polices du portefeuille touchées par un seul et même évènement tel qu'il est défini dans les termes du contrat [36]. La priorité et la portée s'appliquent à la somme des sinistres causés par l'évènement.

La différence entre ces deux types de traités est visible à travers un exemple de 4 sinistres provoqués par un même évènement sur la figure 12.2 .

Ainsi, on préfèrera un traité par risque pour se couvrir contre des risques qui génèrent peu de sinistres mais de montants importants et on choisira un traité par évènement pour se couvrir

35. Partiellement extrait du mémoire d'actuaire de R. Chiche, Construction d'un modèle catastrophe stochastique d'inondation, 2013.

36. Il regroupe généralement tous les sinistres résultant d'une même cause, dans la même zone géographique et pendant une période limitée (de 504 heures généralement pour les inondations).

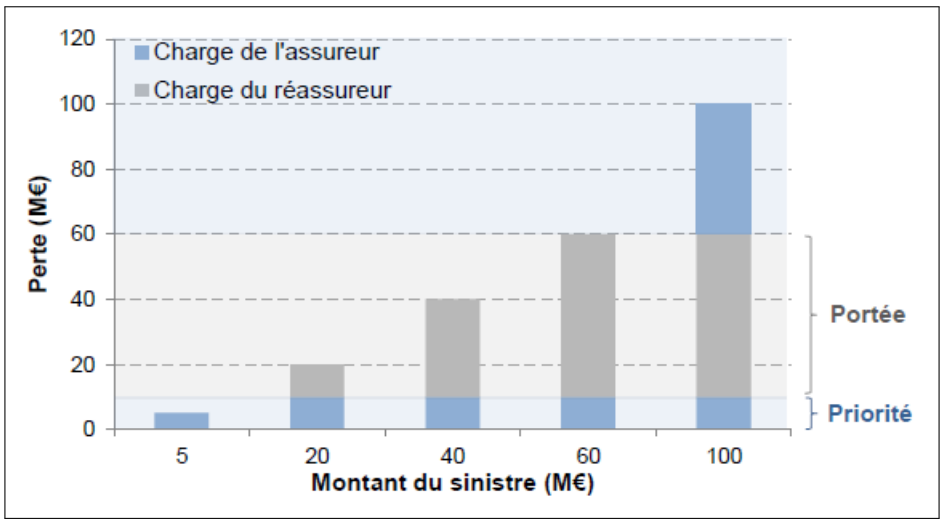

FIGURE 12.1 – Exemple d'application d'un traité 50m XS 10m

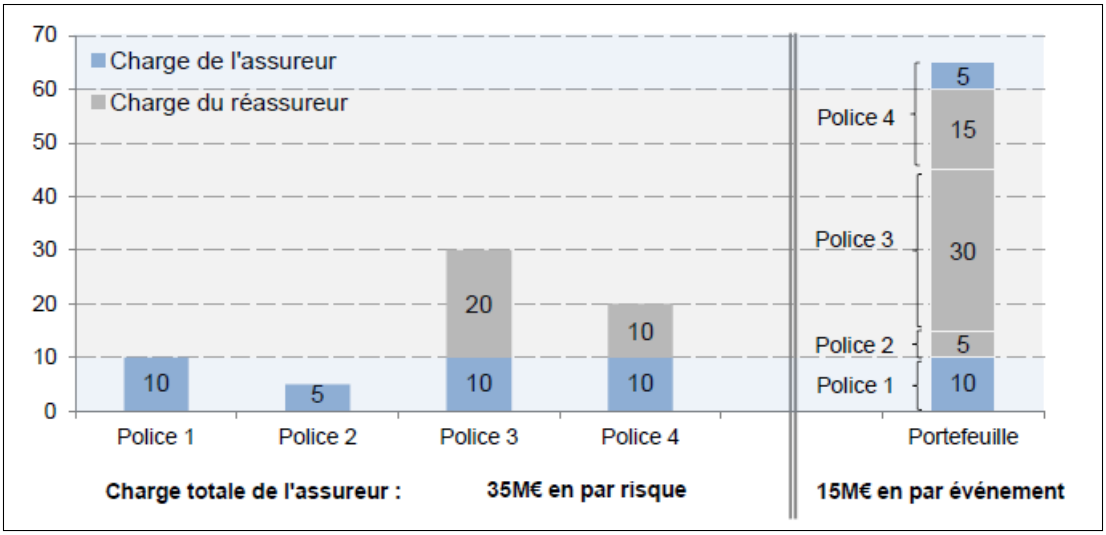

FIGURE 12.2 – Différence entre un 50m XS 10m par risque et par évènement

contre les risques qui génèrent beaucoup de sinistres de faibles montants. Le choix pour le traité par évènement pour se couvrir contre le risque d'inondation semble alors immédiat. En effet, un évènement d'inondation provoque des pertes colossales mais le coût moyen par sinistre reste relativement faible.

Enfin, la portée représente aussi le montant maximal pouvant être versé à l'assureur pendant l'année de couverture. Ainsi, après l'occurrence d'un premier évènement consommant une partie de la portée offerte par le traité, la couverture de l'assureur pour le reste de l'année est alors diminuée du montant versé par le réassureur. Pour éviter d'être en sous-couverture pendant une partie de l'année, on introduit la notion de reconstitutions. Elles permettent à l'assureur de reconstituer la partie de la portée qui a été consommée après un sinistre afin de renouveler sa couverture pour le reste de l'année. Celles-ci peuvent se faire gratuitement ou contre un nouveau paiement de la prime (au *pro rata* de la couverture consommée) du traité[37]. Enfin, le contrat précise le nombre

37. Le prix des reconstitutions peut aussi représenter un pourcentage de la prime du traité. Nous étudions uniquement le cas des reconstitutions gratuites et payantes avec un prix égal à la prime du traité pour simplifier notre propos.

de reconstitutions possibles pendant l'année.

Pour récapituler, les paramètres à fixer dans le choix d'un traité XS sont alors :

- La priorité,
- la portée,
- le nombre de reconstitutions,
- si les reconstitutions sont payantes ou non.

12.2 Capital économique requis par Solvabilité 2

La réglementation Solvabilité 2, entrée en vigueur au 1er Janvier 2016, requiert d'une compagnie d'assurance de détenir suffisamment de fonds propres permettant d'absorber la pire perte annuelle survenant avec une probabilité de 0.5%. Chaque assureur doit alors calculer le capital économique (appelé *Solvency Capital Requirement* dans la réglementation et que l'on notera par la suite SCR) à apporter en plus du *Best Estimate* pour faire face à ses obligations. Son calcul se fait de manière indépendante pour chacun des risques (regroupés par module) portés par la compagnie d'assurance avant d'être agrégés pour permettre de dégager un bénéfice de mutualisation des différents risques.

Le risque d'inondation étudié intervient dans le module CAT regroupant tous les risques catastrophes présentés dans la partie 2.1.1. Le calcul du SCR à proprement parler sort du cadre de ce mémoire et ne sera pas détaillé, on retiendra simplement qu'il correspond, pour le risque inondation pris de manière isolée, à la différence entre le pire résultat technique (noté R^T) avec une probabilité de 0.5% et le résultat technique moyen :

$$SCR_{inondation} = E(R^T) - VaR_{0.5\%}(R^T)$$

Enfin, on fera l'hypothèse simplificatrice que le risque inondation se diversifie avec tous les autres risques portés par AXA France à 80% [38].

12.3 Réassurance optimale pour le risque d'inondation

Nous cherchons enfin dans cette partie à optimiser la structure de réassurance contre le risque d'inondation d'un point de vue d'économie de capital et d'aversion au risque.

12.3.1 Tarification d'un traité XS

Une première étape dans la démarche d'optimisation consiste à choisir un modèle de tarification d'un traité de réassurance. Nous proposons pour notre application le modèle suivant :

$$P = \frac{E(S) + \alpha\,\sigma(S)}{1-\beta}$$

38. Là encore, le calcul de l'allocation du bénéfice de diversification est bien plus complexe et sort du cadre de ce mémoire. En pratique, le facteur de diversification d'un risque n'est pas constant mais dépend du SCR isolé de chacun des risques portés par l'assureur.

où, S est la variable aléatoire représentant la somme des récupérations versées par le réassureur au titre du traité selon les sinistres survenus dans l'année,
$\sigma(S)$ représente l'écart type des récupérations versées par le traité,
α est le taux de chargement de sécurité,
β représente le taux de chargement sur la prime totale pour frais de gestion.

De plus, la somme des récupérations versées par un traité XS par évènement s'expriment par :

$$S = min\,(\,(n+1)\,L,\,Y\,)\,-\,\frac{min(n\,L,\,Y)}{L}\;\times\;\tau\,P$$

où, n représente le nombre de reconstitutions possibles dans le traité,
τ prend la valeur 1 ou 0, indique si les reconstitutions sont payantes ou non,
L est la portée du traité (aussi appelée limite),
Y est une variable alétoire décrivant les récupérations d'un traité avec les mêmes portée et priorité mais avec des reconstitutions gratuites et illimitées.

Enfin, la variable aléatoire Y est définie par :

$$Y = \sum_{i=1}^{N} min(L,\,(X_i\,-\,R)_+)$$

où, N est la variable aléatoire représentant le nombe d'évènements survenant dans l'année,
X_i est la variable aléatoire représentant le coût du i-ème évènement,
R est la prioritée du traité (aussi appelée rétention).

On remarque donc que la somme des récupérations S dépend de la prime payée P lorsque les reconstitutions sont payantes, il est alors pertinent de noter celle-ci S^P. Ainsi le prix P d'un traité XS est la solution de l'équation :

$$P\,-\,\frac{E(S^P)+\alpha\,\sigma(S^P)}{1-\beta}\;=\;0$$

En pratique, on calculera ce prix à l'aide de simulations de Monte Carlo et l'équation ci-dessus se simplifie en équation quadratique en P.

12.3.2 Création de valeur espérée

Nous introduisons dans cette partie la notion de création de valeur induite par la réassurance. Il s'agit de calculer l'écart espéré entre les résultats techniques bruts et nets de réassurance après déduction du coût de la réassurance et des économies liées à la réduction du capital.

Notons respectivement R^T et R'^T les résultats techniques nets d'impôts bruts et nets de réassurance, en gardant les mêmes notations que dans la partie précédente de tarification et en considérant un traité XS par évènement de priorité R, de limite C, de prix P, avec n reconstitutions de taux de prime τ (valant 0 ou 1), on a :

$$\begin{aligned} E(R'^T - R^T) &= (1-\gamma) \times E\left((P_a + S - P - X) - (P_a - X)\right) \\ &= (1-\gamma) \times \left(E(S) - P\right) \end{aligned}$$

où, P_a représente les primes acquises par l'assureur au titre du risque inondation,

$X = \sum_{i=1}^{N} X_i$ est la variable aléatoire représentant la perte totale annuelle brute subie par l'assureur au titre du risque inondation,

γ est le taux d'impôts sur le résultat [39].

De même, nous calculons l'économie du capital immobilisé au titre du risque d'inondation isolé donné par :

$$\begin{aligned} \Delta SCR &= (1-\rho) \times (SCR'_{inondation} - SCR_{inondation}) \\ &= (1-\rho) \times \left(P - VaR_{0.5\%}(S - X) + VaR_{0.5\%}(-X)\right) \\ &= (1-\rho) \times \left(P + VaR_{99.5\%}(X - S) - VaR_{99.5\%}(X)\right) \end{aligned}$$

où ρ est le facteur de diversification du $SCR_{inondation}$ dans le SCR total de l'assureur.

Cette baisse de capital entraîne deux économies :

- la diminution du coût du capital demandé par les actionnaires que l'on modélisera par un taux fixe ψ
- la baisse des bénéfices financiers, imposés également au taux γ, issus du placement du capital. On modélisera aussi ce rendement financier par un taux fixe φ.

Ainsi, la création de valeur moyenne induite par le traité étudié est donnée par :

$$CV(R, C, n, \tau) = (1-\gamma) \times (E(S) - P) - \Delta SCR \times (\psi - \varphi(1-\gamma))$$

12.3.3 Appétit au risque

La directive Solvabilité 2 requiert, à travers le pilier qualitatif, des compagnies d'assurance à développer un système de gouvernance et de gestion des risques. L'ORSA (*Own Risk and Sovency Assessment)*, défini dans l'article 45 de la directive, impose entre autres, à la compagnie d'assurance de mettre en place un processus d'appétit au risque qui consiste à définir le niveau de risque agrégé que la compagnie accepte de prendre en vue de la poursuite de son activité et d'atteinte de ses objectifs stratégiques. Celui-ci s'exprime sous la forme d'une mesure de risque qui tient compte des différentes parties prenantes dans l'activité de l'assureur (actionnaires, détenteurs de la dette, etc) et des régulateurs et agences de notations [40].

Plusieurs mesures de risque peuvent être utilisées afin de fournir un indicateur quantitatif de l'appétit au risque. Nous exprimerons dans cette étude l'appétit au risque comme la déviation

39. Nous appliquons ce taux même si le résultat est négatif. En effet, un résultat négatif sur cette branche dégagerait une économie d'impôt sur le résultat total de l'assureur (et nous supposons que celui-ci est toujours positif).

40. M. Juillard et F. Planchet, Winter. Le pilier 2, la gestion des risques et l'ORSA, Juillet 2011.

maximale d_{max} du résultat technique permise par la compagnie à l'horizon d'une année avec une certaine probabilité p. Formellement cela revient à respecter la condition suivante :

$$\frac{E(R^T_{total}) - VaR_p(R^T_{total})}{E(R^T_{total})} \leq d_{max}$$

Dans la suite nous prendrons les valeurs $d_{max} = 5\%$ et $p = 5\%$. En prenant le même facteur de diversification que pour le capital économique à 80% [41] pour le risque d'inondation, la condition d'appétit au risque pour le risque d'inondation s'exprime par :

$$\frac{E(R^T) - VaR_{5\%}(R^T)}{E(R^T)} \leq 25\%$$

La réassurance peut alors être vue comme un outil de pilotage permettant de céder suffisamment de risques afin de respecter à tout moment les limites de l'appétit au risque de la compagnie.

12.3.4 Conclusion

Notre optimisation de réassurance consistera à la recherche du traité maximisant la création de valeur espérée tout en cédant suffisamment de risques nous permettant de respecter les limites de l'appétit au risque. Formellement, cela revient à résoudre le problème d'optimisation sous contraintes suivant :

$$\begin{aligned} \underset{R,C,n,\tau}{\text{argmax}} \quad & CV(R, C, n, \tau) \\ \text{s. c.} \quad & \frac{E(R^T) - VaR_{5\%}(R^T)}{E(R^T)} \leq 25\% \end{aligned}$$

Et nous prendrons les valeurs $\alpha = 10\%$, $\beta = 15\%$, $\gamma = 34\%$, $\rho = 80\%$, $\psi = 6\%$ et $\varphi = 3\%$.

12.3.5 Résolution numérique

Nous faisons les calculs de création de valeur et d'appétit au risque pour différentes priorités, portées et nombre de reconstitutions payantes ou gratuites. Nous procédons par dichotomie en plusieurs étapes en commençant par des intervalles et des pas larges et affinons ceux-ci petit à petit jusqu'à obtenir une précision satisfaisante.

Les résultats obtenus montrent que le traité XL optimal est de priorité $6m$ et de limite $140m$ avec une reconstitution payante. La création de valeur est alors de $-6.6m$ et la déviation du résultat technique avec une probabilité de 5% est à 24%.

Nous avons donc une création de valeur négative, ce qui signifie que la réassurance est beaucoup trop chère au vu de l'économie de fonds propres réalisée. Néanmoins, elle est nécessaire pour être conforme à l'aversion au risque de l'assureur et passer d'une déviation du résultat technique avec une probabilité de 5% de 150% sans réassurance à 24%.

Nous pouvons voir sur le graphique 12.3 l'évolution de la création de valeur en fonction de priorité et de la limite pour l'ensemble des traités XL avec une reconstitution payante.

Les simulations réalisées ont aussi montré qu'aucun traité avec des reconstitutions gratuites n'a permis d'atteindre une création de valeur aussi grande. En effet, le traité optimal avec des

41. Nous faisons l'hypothèse simplificatrice que le facteur de diversification de la déviation du résultat technique à 5% est le même que celui de la déviation à 0.5%.

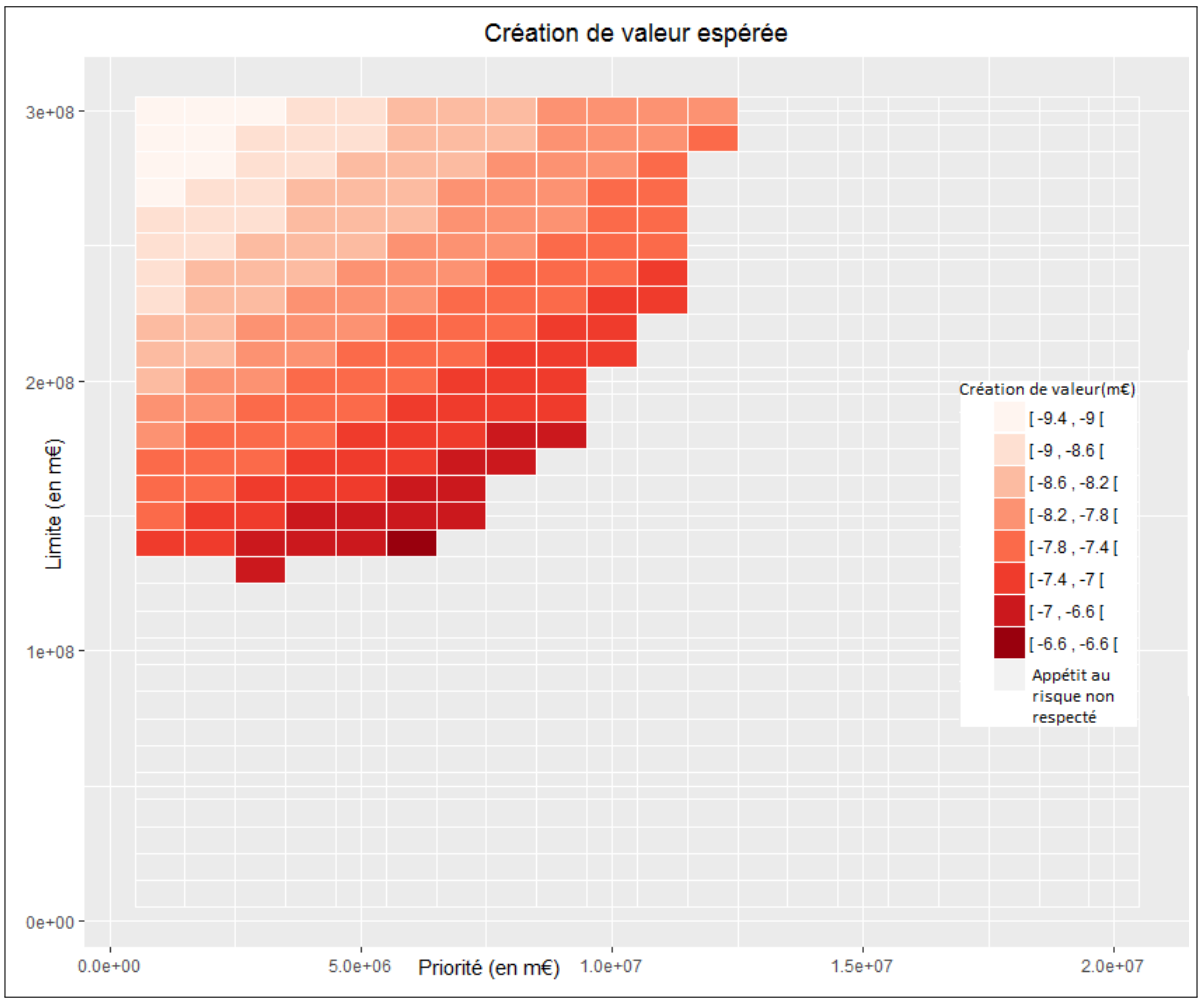

FIGURE 12.3 – Évolution de la création de valeur pour les traités avec une reconstitution payante.

reconstitutions gratuites a une rétention de $14m$, une limite de $140m$ et zéro reconstitution. La création de valeur est de $-6.75m$ et la déviation du résultat technique avec une probabilité de 5% est de 24.8%. On a ainsi plus de volatilité sur le résultat et une création de valeur moindre. Cela peut s'expliquer par la faible fréquence des inondations rendant la couverture contre le risque de fréquence à travers des reconstitutions gratuites peu efficient. Cela a néanmoins un impact minime sur la volatilité du résultat et le capital économique rendant les traités à reconstitutions payantes plus optimaux pour ce risque.

L'évolution de la création de valeur en fonction de la priorité et de la limite pour l'ensemble des traités XL sans reconstitution est également visible sur la figure 12.4.

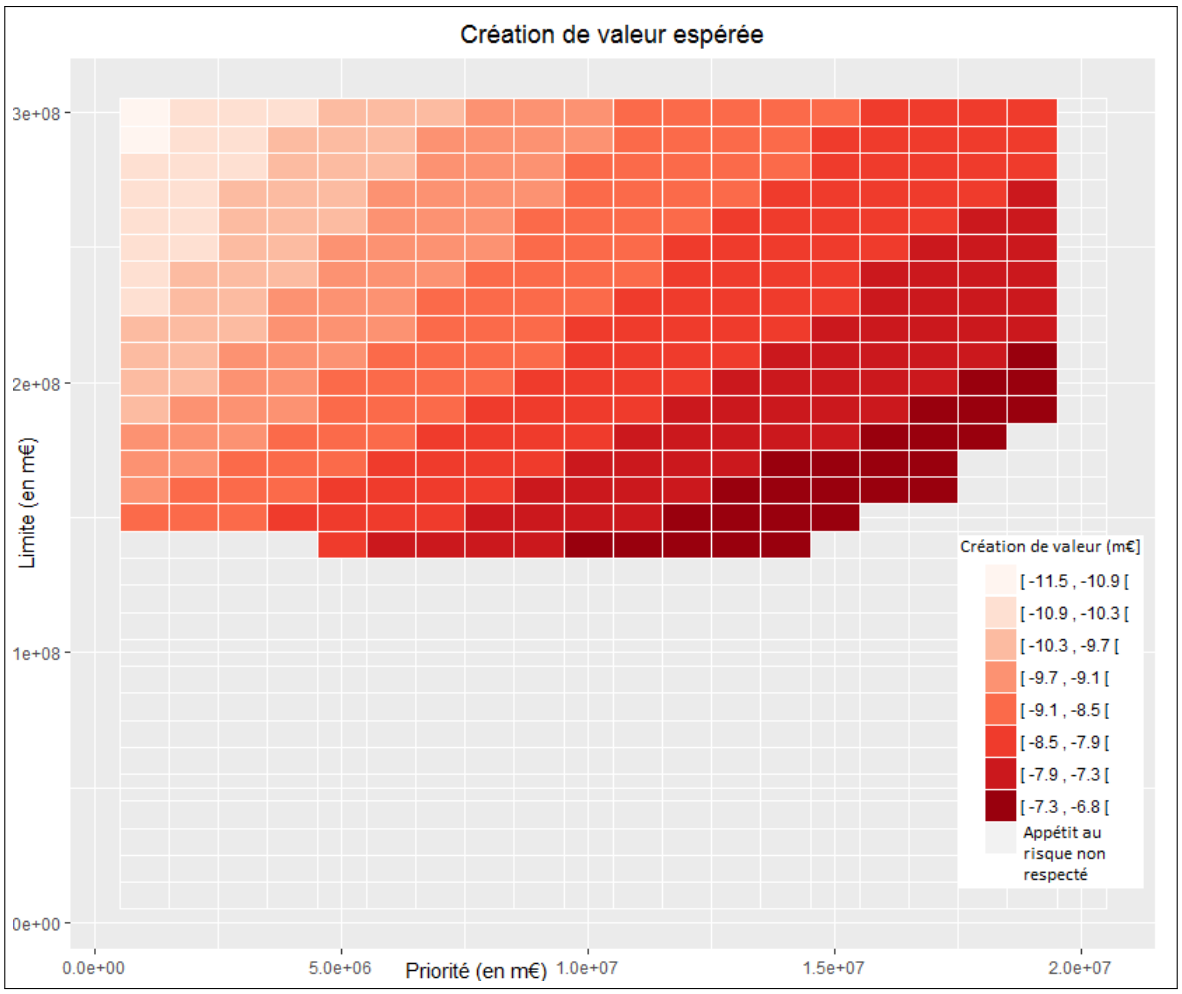

FIGURE 12.4 – Évolution de la création de valeur pour les traités sans reconstitution.

Conclusion

Le but de cette étude était de développer une vision fine et actualisée de l'impact des évènements historiques sur l'exposition courante d'AXA. Nous avons alors utilisé un modèle physique de propagation de l'eau pour construire des premières empreintes historiques que nous avons ensuite nettoyer à l'aide de la base des arrêtés CAT NAT et de la sinistralité historique. Nous avons ensuite utilisé ces empreintes et les différentes données à notre disposition pour calibrer différents modèles d'apprentissage et produire une estimation fine du coût de chaque évènement historique.

Il était alors intéressant de constater l'apport potentiel du *machine learning* en actuariat par rapport aux méthodes plus classiques telles que les *GLMs*. Néanmoins, l'aspect très opaque de ces techniques récentes complique la communication des résultats à la direction d'une compagnie d'assurance et ralentit leur usage en pratique. De plus, l'inexistence presque totale de propriétés statistiques vérifiées par ces prédictions ne permet pas de quantifier leur pertinence contrairement aux méthodes linéaires. Notons toutefois l'existence de statistiques de nature différentes telles que l'importance des variables explicatives qui permet de juger de la pertinence des choix de modélisation effectués. Nous avons ainsi, grâce à cette statistique, montré dans notre étude la spécificité de destruction de chaque évènement et la robustesse de nos empreintes historiques. L'étage du site assuré apparaît également comme la caractéristique assurantielle la plus importante, ce qui est conforme à notre intuition vis-à-vis du risque modélisé et nous rassure dans la pertinence du modèle. La puissance de calcul requise pour calibrer de tels modèles peut également constituer un obstacle à leur intégration dans les modèles mais le développement récent et la démocratisation des outils informatiques devrait limiter ce frein.

Par ailleurs, les empreintes développées fournissent un outil de souscription et de gestion des risques intuitif [42] dans le sens où l'on décrit des évènements réels dont chacun se rappelle de l'ampleur. Le modèle développé peut également être qualifié de dynamique dans le sens où il adapte les estimations des pertes des évènements historiques en fonction de la répartition géographique des sites assurés et de l'aménagement actuel du territoire, prenant ainsi en compte les éventuelles mises en place de moyens de protection contre les débordements de rivières par exemple. Le modèle peut par ailleurs être appliqué à de nouveaux portefeuilles qui n'existaient pas au moment des évènements.

Enfin, nous avons vu comment le modèle peut être utilisé afin de piloter le risque d'inondation porté par l'assureur à travers la construction d'un modèle fréquence coût et l'optimisation de la structure de réassurance en maximisant la création de valeur espérée au vu de l'impact sur le capital économique tout en répondant aux exigences d'appétit au risque de la compagnie. Le modèle développé permet alors une estimation plus fine du capital économique et de l'efficience de la réassurance.

Les limites et pistes d'amélioration du modèle restent multiples. L'impertinence de la variable de la hauteur d'eau, soupçonnée lors de la construction des empreintes, et confirmée par la statistique d'importance des variables explicatives lors de la calibration des modèles d'apprentissage ; justifie la nécessité d'améliorer le module de propagation de l'eau, ce qui pourrait permettre de capturer davantage de variabilité et gagner en capacité de prédiction. Citons également la prise en compte des données manquantes qui peut être améliorée en testant les différentes méthodes proposées dans la littérature.

42. Une application concrète serait par exemple la gestion des accumulations dans les zones historiquement sinistrées.

Bibliographie

[1] W. Gruss. L'industrie de l'assurance, 1985.

[2] P. Hausmann. Les inondations, un risque assurable ? 1999. http://www.swissre.com/publications/Les_inondations_un_risque_assurable.html.

[3] P.M.E Altham. Introduction to generalized linear modelling, August 2011. Statistical Laboratory, University of Cambridge http://www.statslab.cam.ac.uk/~pat/All.pdf .

[4] A. Charpentier. Actuariat iard - act2040, Hiver 2013. http://freakonometrics.free.fr/ .

[5] O. Gaudoin. Principes et méthodes statistiques, février 2013. Notes de cours, Grenoble INP - Ensimag.

[6] O. Gaudoin. Satistique inférentielle avancée, février 2014. Notes de cours, Grenoble INP - Ensimag.

[7] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7) :799–815, 2004.

[8] Ioannis Kosmidis and David Firth. Bias reduction in exponential family nonlinear models. *Biometrika*, page asp055, 2009.

[9] Michael Smithson and Jay Verkuilen. A better lemon squeezer ? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1) :54, 2006.

[10] M Mittlbock and Harald Heinzl. Pseudo r-squared measures for generalized linear models. In *Proceedings of the 1st European Workshop on the Assessment of Diagnostic Performance, Milan, Italy*, pages 71–80, 2004.

[11] F Planchet and G Serdeczny. Modèles fréquence-coût : quelles perspectives d'évolution, Mars 2014. Version 0.7 : http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/0/39B54166464089AFC12572B0003D88C2/$FILE/IARD_IA_20140321.pdf .

[12] Art B Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8(Apr) :761–773, 2007.

[13] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11) :1134–1142, 1984.

[14] J Kun. Weak learning, boosting, and the adaboost algorithm, May 2015. http://jeremykun.com/2015/05/18/boosting-census/ .

[15] MJ Kearns. Thoughts on hypothesis boosting, 1988. *ML class project*, 319 :320.

[16] Michael J Kearns and Leslie G Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, 41(1) :67–95, 1989.

[17] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2) :197–227, 1990.

[18] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[19] L Breiman, JH Friedman, R Olshen, and CJ Stone. Classification and regression trees. 1984.

[20] Leo Breiman. Bagging predictors. *Machine learning*, 24(2) :123–140, 1996.

[21] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

[22] Jerome H Friedman. Greedy function approximation : A gradient boosting machine. *mh (xa m)*, 1000 :0, february 1999.

[23] Jerome H Friedman. Stochastic gradient boosting. *mh (x ; am)*, 1000 :0, March 1999.

[24] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7, 2013.

[25] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1) :29–36, 1982.

[26] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8) :861–874, 2006.

[27] Caisse Centrale de Réassurance. En france, l'indemnisation des catastrophes naturelles. principes et fonctionnement, Mars 2015. https://www.ccr.fr/blobs/com.cardiweb.cardiboxv6.cm.business.Article/3028161110579609690/documentJoint/1/indemnisation%20cat-nat.pdf .

[28] Keith Conrad. Probability distributions and maximum entropy. *retrieved November*, 14 :2013, 2013.

[29] J Michael Brick and Graham Kalton. Handling missing data in survey research. *Statistical methods in medical research*, 5(3) :215–238, 1996.

[30] Donald A Darling. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4) :823–838, 1957.

[31] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[32] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3) :18–22, 2002.

[33] Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart : Recursive Partitioning and Regression Trees*, 2015. R package version 4.1-9.

[34] Robert J. Hijmans, Steven Phillips, John Leathwick, and Jane Elith. *dismo : Species Distribution Modeling*, 2015. R package version 1.0-12.

[35] Greg Ridgeway with contributions from others. *gbm : Generalized Boosted Regression Models*, 2015. R package version 2.1.1.

[36] Greg Ridgeway. Generalized boosted models : A guide to the gbm package. *Update*, 1(1) :2007, 2007.

[37] G. Grothendieck. *sqldf : Perform SQL Selects on R Data Frames*, 2014. R package version 0.4-10.

[38] M Dowle, T Short, S Lianoglou, A Srinivasan with contributions from R Saporta, and E Antonyan. *data.table : Extension of data.frame*, 2014. R package version 1.9.4.

[39] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc : an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12 :77, 2011.

[40] Julian Faraway, George Marsaglia, John Marsaglia, and Adrian Baddeley. *goftest : Classical Goodness-of-Fit Tests for Univariate Distributions*, 2017. R package version 1.0-4.

A Annexe :Démonstrations et outils mathématiques

A.1 *GLMs* : Démonstration de l'équation (4.3)

Soit Y une variable aléatoire dont la loi de probabilité a une fonction de densité de la forme :

$$f(y|\theta,\phi) = exp\left(\frac{y\,\theta\ -\ b(\theta)}{\phi}\ +\ c(y,\,\phi)\right)$$

où b et c sont des fonctions et θ et ϕ des paramètres réels.

- **Preuve de $E(Y) = b'(\theta)$:**
  On a alors :

$$\begin{aligned}\frac{\partial f}{\partial \theta} &= \frac{\partial\, ln(f)}{\partial\theta}\, f \\ &= \frac{y\,\theta\ -\ b(\theta)}{\phi}\, f \end{aligned} \tag{A.1}$$

  Il vient, comme $\int_y f(y|\theta,\phi)\,dy = 1$ et sachant que l'on peut interchanger $\frac{\partial}{\partial\theta}$ et $\int_y$ dans le cas de la fonction exponentielle :

$$\begin{aligned}\phi\,\frac{\partial}{\partial\,\theta}\int_y f(y|\theta,\phi)\,dy = \ 0 \ &= \int_y y\, f(y|\theta,\phi)\,dy - \ b'(\theta)\ \int_y f(y|\theta,\phi) \\ &= E(Y) - b'(\theta)\end{aligned} \tag{A.2}$$

  Ce qui conclut.

- **Preuve de $Var(Y) = b''(\theta)\,\phi$:**
  On a par l'équation (A.1) :

$$\begin{aligned}\phi\,\frac{\partial^2 f}{\partial\,\theta^2} &= (y - E(Y))\,\frac{\partial f}{\partial\theta} - b''(\theta)\, f \\ &= \frac{(y - E(Y))^2}{\phi}\, f - b''(\theta)\, f\end{aligned}$$

  Et de même que dans l'équation (A.2) :

$$\begin{aligned}\phi\,\frac{\partial^2}{\partial\,\theta^2}\int_y f(y|\theta,\phi)\,dy = \ 0 \ &= E\left(\frac{(Y - E(Y))^2}{\phi}\right) - b''(\theta) \\ &= \frac{Var(Y)}{\phi} - b''(\theta)\end{aligned}$$

  Ce qui conclut.

A.2 Estimateur du maximum de vraisemblance

Soient $Y_1,\ldots,Y_n$ des variables aléatoires indépendantes et de même loi dépendant d'un paramètre $\theta \in \mathbb{R}^d$ de fonctions de densités respectives $f(\cdot,\theta)$. Soient $y_1,\ldots,y_n$ des observations de ces variables aléatoires.

A.2.1 Fonction de vraisemblance

La fonction de vraisemblance pour l'échantillon $(y_1, \ldots, y_n)$ est définie par :

$$\mathcal{L}(\theta) \equiv \mathcal{L}(\theta, y_1, \ldots, y_n) = \Pi_{i=1}^{n}\ f(y_i, \theta)$$

Et on définit la fonction de log-vraisemblance :

$$l(\theta) \equiv ln\, \mathcal{L}(\theta) = \sum_{i=1}^{n}\ ln(f(y_i, \theta))$$

L'estimation de θ par maximum de vraisemblance est la valeur $\hat{\theta}$ qui maximise la log-vraisemblance :

$$\hat{\theta} = \underset{\theta}{argmax}\ l(\theta)$$

.

$\hat{\theta}$ est alors solution de l'équation :

$$\nabla_{\hat{\theta}}\ l(\theta) = 0$$

.

A.2.2 Matrice d'information de Fischer

La matrice d'information de Fisher $\mathcal{I}_n(\theta)$ est la matrice de variance covariance du gradient, de terme général :

$$\mathcal{I}_{|,\|}(\theta) = Cov\left[\frac{\partial l(\theta, Y_1, \ldots, Y_n)}{\partial \theta_j}\ ;\ \frac{\partial l(\theta, Y_1, \ldots, Y_n)}{\partial \theta_k}\right]$$

On a par ailleurs l'égalité :

$$\mathcal{I}_n(\theta)\ =\ n\,\mathcal{I}_1(\theta)$$

A.2.3 Convergence en loi

On dit qu'une suite de variable aléatoire $(X_n)_{n\geq 0}$ converge en loi vers la loi de probabilité de fonction de répartition F si et seulement si

$$\underset{n\to+\infty}{lim}\ F_{X_n}(x) = F_X(x)$$

en tout point x où F est continue. Cela signifie que, quand n est suffisamment grand, X_n suit approximativement la loi de probabilité de fonction de répartition F.

Lorsqu'on estime θ par maximum de vraisemblance, on a le résultat de convergence suivant :

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{\mathcal{L}} \mathcal{N}_d\left(0, \mathcal{I}_1^{-1}(\theta)\right)$$

où $\mathcal{I}_1(\theta)$ est la matrice d'information de Fisher pour un échantillon de taille 1 et $\mathcal{N}_d$ la loi normale dans $\mathbb{R}^d$.

A.3 Fonction $\Gamma(\cdot)$

La fonction Γ est définie pour $a > 0$ par

$$\Gamma(a) = \int_0^{+\infty} e^{-x}\, x^{a-1}\, dx$$

.

Et on a :
$\forall n \in \mathbb{N}^*,\ \Gamma(n) = (n-1)!,\ \Gamma(1) = 1,\ \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},$

$$\forall a \in]1, +\infty[, \Gamma(a) = (a-1)\,\Gamma(a-1)$$

A.4 Probabilité d'appartenir à OOB data dans la construction d'un arbre de forêt

Soit un jeu de données $X_n = (x_1, \ldots, x_n)$, et soit $\tilde{X}_n$ un échantillon avec remise de X_n de taille n, on a alors :

$$\begin{aligned}
P(x_1 \notin \tilde{X}_n) &= \left(1 - \frac{1}{n}\right)^n \\
&= exp\left(n\, ln\left(1 - \frac{1}{n}\right)\right) \\
&\underset{n\to+\infty}{=} exp\left(n\left(-\frac{1}{n} + o\left(\frac{1}{n}\right)\right)\right) \\
&\underset{n\to+\infty}{=} exp\left(-1 + o(1)\right) \\
&\underset{n\to+\infty}{\to} 0.368
\end{aligned}$$

Ainsi, la probabilité pour une donnée d'appartenir à OOB tend vers 37%, en réalité la convergence est très rapide (voir tableau 1) , ce qui nous permet d'affirmer que la probabilité est toujours de 37% (on utilise rarement des jeux de données taille inférieure à 50).

n	$P(x_1 \notin \tilde{X}_n)$
5	0.328
10	0.349
20	0.358
50	0.364
100	0.366

TABLE 1 – Convergence de la probabilité de non appartenance à OOB data