Mémoire présenté devant l'ENSAE ParisTech
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 21/02/2019

Par : **Mariame El ALJ**

Titre : **Sophistication de la tarification du risque catastrophe naturelle**

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

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

Membres présents du jury de la filière

Entreprise : AXA GIE
Nom :EL HASSANI Hamza
Signature :

Membres présents du jury de l'Institut des Actuaires

Directeur du mémoire en entreprise :

Nom :
Signature :

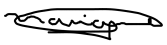
Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Secrétariat :

Signature du candidat

Bibliothèque :



Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Remerciements

Ce mémoire représente le fruit d'un travail réalisé au sein de AXA GIO. Il n'aurait pas abouti sans l'aide et l'implication des membres de l'équipe Actuarial & CAT modeling et des équipes pédagogiques de l'ENSAE, et je tiens à ce titre à tous les remercier.
En premier lieu, je tiens à adresser mes sincères remerciements et toute ma gratitude à Hamza El Hassani, mon encadrant pour ce mémoire. Rien n'aurait été possible sans son support, son écoute, sa disponibilité, ses suggestions, ses conseils et ses innombrables idées.
Je voudrais également remercier Nicolas Baradel, qui a encadré ce mémoire pour le compte du mastère actuariat à l'ENSAE pour ses conseils, ses suggestions et ses retours.

Enfin, merci à mes amis, ma famille et en particulier mes parents pour leur soutien inconditionnel.

Résumé

Catastrophes naturelles, Réassurance non-vie, Souscription, Taux de destruction, Courbes de rabais, Modélisation, Tarification dynamique

L'accélération du réchauffement climatique à travers le monde exacerbe les phénomènes météorologiques extrêmes et multiplie le risque de catastrophes climatiques.
La prise en main de ces risques constitue alors un véritable défi pour les compagnies d'assurance et de réassurance qui assistent à une augmentation du coût des dommages liés à ces catastrophes susceptible de menacer leur solvabilité.

Aujourd'hui, avec le développement des modèles stochastiques de catastrophes naturelles, les assureurs peuvent désormais quantifier ce risque et définir leur besoin en réassurance de façon plus précise.
Par ailleurs, une connaissance accrue et une maîtrise de ces phénomènes permet de développer un avantage compétitif sur le marché par le biais d'une meilleure segmentation du risque et d'actions de souscriptions plus réfléchies.

Ce mémoire a pour vocation d'apporter un support technique à l'assureur *AXA Mexico*, filiale du groupe *AXA*, à travers une évaluation complète des différentes méthodologies d'intégration de la modélisation *CAT* dans les processus de tarification et de souscription. Cette démarche permet d'envisager à terme l'élaboration d'un outil de souscription ayant pour but premier une sophistication de la tarification (y compris celle des structures complexes) et dans un second temps un meilleur contrôle de l'exposition en identifiant des zones géographiques ou segments où la diversification du risque serait mise à profit pour optimiser le coût de réassurance.

La sophistication de la tarification a été considérée suivant plusieurs approches : la première consistant à calibrer des modèles statistiques usuels type *GLMs* ou des modèles plus complexes tel que les *GBMs* avec une décomposition fréquence-sévérité classique sur les pertes simulées site à site d'un modèle *CAT*. Cette méthode permet d'obtenir des résultats convenables et reproduit relativement bien les simulations du modèle original. Cependant, cette approche a montré dans un second temps ses limites dans le cadre de la tarification de structures complexes, mettant en lumière la difficulté de la prise en compte des conditions financières lorsqu'un unique paramètre de la distribution de perte (moyenne) est modélisé.

Une discussion est alors menée afin de mettre en place une méthodologie permettant de se défaire du cadre classique de la calibration en modélisant cette fois la totalité de la distribution de pertes individuelles pour chacune des polices soit avec des algorithmes de calibration et simulation statistiques soit par une approche purement informatique.

Pour des raisons de confidentialité, l'ensemble des données figurant dans ce mémoire ont été modifiées. Il est donc conseillé d'apprécier la pertinence de la démarche, plus que les résultats chiffrés.

Abstract

Natural disasters, Non-life reinsurance, Underwriting, Destruction rate, Exposure curves, Modeling, Dynamic pricing

Global warming acceleration exacerbates extreme weather events and increases the risk of climate disasters. Taking control of these risks is a real challenge for insurance and reinsurance companies that have witnessed an increase in the cost of damage related to these catastrophes, which could threaten their solvency.

Today, natural disaster stochastic models development enables insurers to quantify this risk and define their need for reinsurance.
In addition, an increased knowledge and control of these phenomena makes it possible to develop a competitive advantage on the market through better risk segmentation and more thoughtful underwriting actions.

This paper fulfills the needs of the insurer AXA Mexico in the achievement of its performance objectives through a comprehensive evaluation of different methodologies for incorporating catastrophe modeling into ratemaking and underwriting.
The key role of cat models in the value creation chain is emphased and ultimately allows the development of an underwriting tool whose primary objective is the pricing sophistication (including complex structures) and then to better control exposure by identifying geographical areas or segments where risk diversification would be used to optimize reinsurance cost.

The pricing sophistication has been considered through several approaches : the first being to calibrate standard statistical models such as *GLMs* or more complex models such as *GBMs* with a classical frequency-severity decomposition on *CAT* model simulated site-to-site losses. This method provides suitable results and reproduces relatively well the simulations of the original model. However, this approach has subsequently shown its limitations in the pricing of complex structures, highlighting the difficulty of taking into account financial conditions when a single parameter of the loss distribution is modelled (mean).

A discussion is then conducted in order to set up a methodology to overcome the traditional calibration framework by modelling this time the total distribution of individual losses for each of the policies either with calibration and simulation statistical algorithms or by a purely computerized approach.

For confidentiality reasons, all sensible information in this paper have been modified. It is therefore advisable to assess the relevance of the approach, more than quantified results.

Note de synthèse

Contexte et problématique

L'ère moderne de l'évaluation des risques de catastrophes naturelles à l'aide de modèles *CAT* se développe depuis les années 1980. Cependant ce n'est qu'en 1992, suite à la survenance de l'ouragan *Andrew* dont la violence a surpris le monde entier qu'une réelle prise de conscience se produit dans le monde des assurances.
Depuis, les outils de simulation de catastrophes naturelles sont en développement constant permettant alors une meilleure connaissance des risques extrêmes potentiels au sein d'un portefeuille et par ce biais une meilleure structuration des traités de réassurance associés.

Ces modèles stochastiques reposent sur la construction d'un catalogue d'évènements *CAT* fictifs mais probables et de leurs pertes associées.
Dès lors, il est intéressant pour l'assureur d'exploiter ces résultats sur l'ensemble de la chaîne de valeur, notamment en souscription et en tarification.
Ce mémoire met donc en exergue différentes approches d'intégration de la modélisation *CAT* dans ces processus.

Approches proposées

Les approches de tarification et de souscription présentées dans ce mémoire s'appliquent de manière générale à tous les périls ayant fait l'objet d'une modélisation *CAT*. Néanmoins, le périmètre de l'étude se limitera à la modélisation du risque ouragan pour le compte d'*AXA Mexico*, filiale du groupe *AXA* opérant au Mexique. Le modèle sous-jacent utilisé tout au long de ce mémoire est un modèle développé en interne au sein des équipes d'*AXA GIO*.
Ce modèle est à la base des différentes méthodes de tarification proposées. Nous présentons notamment, une approche fréquence sévérité classique, une approche par calibration conjointe d'une loi statistique et d'une simulation *Monte-Carlo* et enfin une méthode reposant sur une restructuration de l'architecture informatique du modèle.

Approche Fréquence Sévérité

La première approche considérée repose sur un modèle fréquence-sévérité mettant en œuvre des modèles paramétriques et non paramétriques sur les pertes simulées par le modèle *CAT*. Dans un premier temps, une attention particulière est donc accordée à la composition du portefeuille à fournir en entrée.
En effet, une hétérogénéité de ce portefeuille entraînerait une erreur d'appréciation des relativités des variables explicatives potentielles. Afin d'éliminer tout biais inhérent aux données d'exposition propres au portefeuille mexicain, un portefeuille fictif constitué de contrats dont les caractéristiques sont uniformes et couvrant l'ensemble des zones géographiques est alors construit.
La deuxième étape consiste à calibrer la fréquence et la sévérité au moyen de modèles paramétriques tels que les GLMs et d'autres méthodes d'apprentissage du type machine learning (GBMs, Forêts aléatoires) afin de comparer leurs performances.
La dernière étape consiste à appliquer des courbes de rabais pour tenir compte des conditions financières. Cette approximation est alors remise en cause, lorsque nous considérons un nombre limité de courbes de rabais ou lorsque les conditions financières à appliquer sont relativement complexes (multi-layering, réassurance facultatives, franchises min/max).

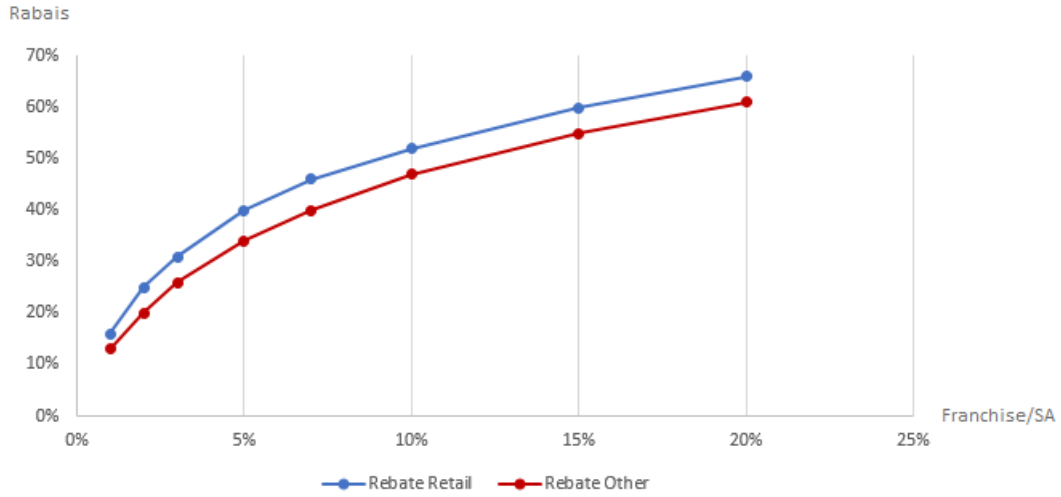


FIGURE 1 – Courbes de rabais utilisées pour le modèle Mexique Ouragan

Pour permettre une meilleure prise en compte de ces effets non-linéaires sur la prime pure et réduire les approximations effectuées, nous envisageons donc dans la suite de ne plus modéliser un unique paramètre de la distribution de pertes individuelles (actuellement la moyenne), mais un vecteur de paramètres permettant alors de caractériser la totalité de la distribution de pertes.

Approche calibration de loi et Monte-Carlo

Cette approche peut être décrite en trois grandes étapes. La première consiste à calibrer les lois statistiques ajustant le mieux la distribution de la sévérité observée par événements et par sites et celle de la fréquence observée par années et par sites.
A cette étape, nous explorons plusieurs options quant à l'ajustement de la sévérité :

- Le taux de destruction sans changement d'échelle avec des lois telles que la *mbbefd* ou la *bêta*.
- Le taux de destruction avec changement d'échelle permettant l'usage des lois de valeurs extrêmes.

L'étape suivante vise à prédire les paramètres de ces lois à l'aide de modèles de statistiques.
Deux types d'erreurs sont introduites pour quantifier le degré de confiance que nous avons pour chaque étape :

- Erreur d'ajustement : Supposer que les observations décrivant chaque site sont tirées de distributions classiques dépendant d'un nombre restreint de paramètres constitue une hypothèse forte.
- Erreur de prédiction : En supposant que la loi sous-jacente aux pertes est connue, quelle serait la qualité de prédiction lors de l'estimation de ses paramètres.

Enfin la dernière étape consiste à simuler la distribution annuelle de perte calibrée pour chaque site en utilisant du Monte-Carlo.
Afin d'évaluer la qualité de cette solution, cette dernière a été appliquée sur l'ensemble du portefeuille pour calculer l'erreur relative faite sur les quantiles de pertes à différentes périodes de retour.

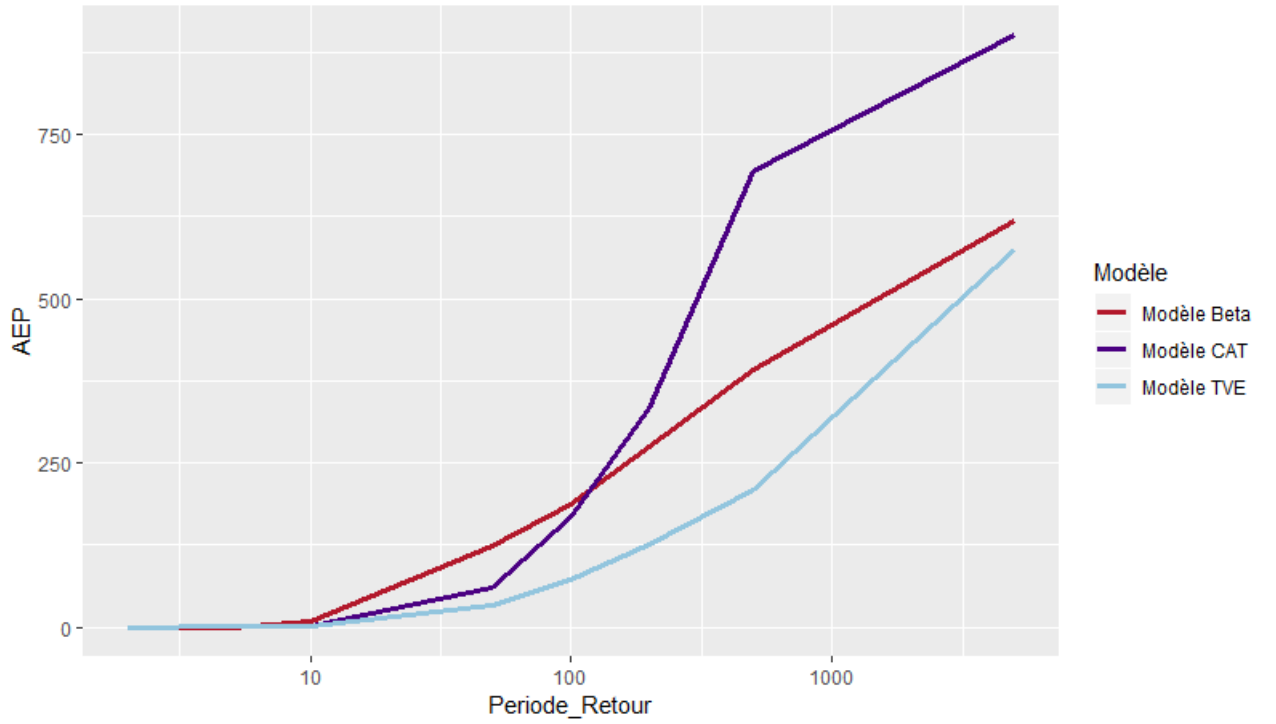


FIGURE 2 – Exemple des courbes AEP d'un site risqué obtenues pour différents modèles

Approche restructuration de l'architecture du modèle

L'idée de cette méthode consiste à répliquer le processus de calcul de pertes du modèle *CAT* mais avec un changement de paradigme.
L'architecture des modèles *CAT* et les ressources nécessaires à leur déploiement sont optimales lorsqu'il s'agit de tarifer un portefeuille et non un nombre réduit de polices.
Pour répondre à cette problématique, cette solution opérationnelle d'aide à la souscription vise à restructurer le catalogue stochastique d'événements en utilisant un système de gestion de base de données.
Une fois que les informations concernant la police sont rentrées, les coordonnées géographiques seront utilisées pour extraire toutes les valeurs de vitesse de vent correspondant à tous les événements l'ayant touché. Les caractéristiques physiques du site seront croisées aux courbes de vulnérabilité pour obtenir un taux de destruction. Enfin les conditions financières rentrées seront appliquées à la distribution de pertes brutes pour obtenir une distribution des pertes nettes.

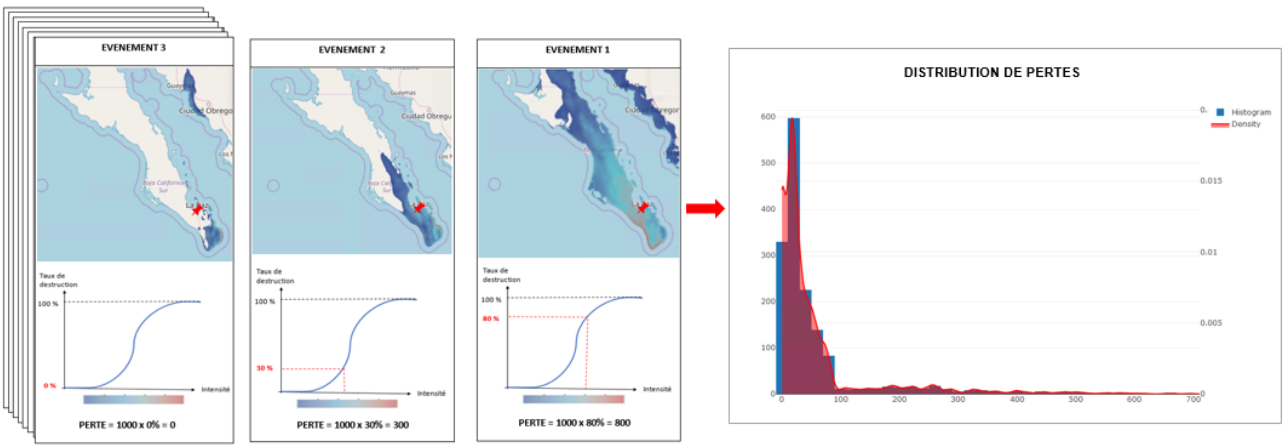


FIGURE 3 – Le process de l'outil

Contrairement aux approches présentées, cette dernière méthode ne repose que sur le modèle *CAT*, en d'autres termes il n'y a aucun autre modèle statistique derrière.
De ce fait, la distribution de perte que génère cette solution de façon instantanée est identique à celle que générerait le modèle.

Résultats

Le fondement statistique des approches fréquence sévérité et celle de la calibration de loi reposent sur des hypothèses fortes rarement vérifiées d'un point de vue théorique. De plus la difficulté technique de leur mise en oeuvre conduit plutôt à sélectionner l'approche du rétro-engineering informatique.
Cette approche basée totalement sur le modèle *CAT*, présente des atouts considérables pour l'assureur et constitue un outil de souscription à part entière. Parmi les fonctionnalités intégrées :

- Tarification des structures de couverture en réassurance complexes notamment les facultatives sans passer par les courbes de rabais.
- Calcul du MPL *Maximum Possible Loss* ou en français la perte maximale possible ou les pertes à des périodes de retour différentes.
- Calcul de l'impact d'un changement de conditions d'assurance ou de réassurance. Par exemple : une baisse/augmentation de participation, l'introduction d'une franchise plus haute ou encore la cession de facultatives.
- Prise en compte de l'accumulation géographique, en calculant l'impact d'un nouveau risque ou l'annulation d'un risque existant sur le bi-centenaire du portefeuille ou tout autre période de retour.

A noter aussi que tous ces résultats peuvent être obtenus à n'importe quel niveau de granularité : site, police ou grand compte.

Executive Summary

Context and objective

The modern era of Natural Hazard Risk Assessment using Catastrophe Models (CAT) has been evolving since the 1980s. However, it was not until 1992, following the onset of the hurricane Andrew whose violence has surprised the world that a real awareness is happening in the insurance world. Since then, disaster simulation tools are in constant development and reinsurance treaties are increasingly structured.
These stochastic models produce loss distributions at a fine grained level. Therefore, it is interesting for the insurer to exploit these results across the entire value chain, from underwriting to cost of capital optimisation.
This paper highlights different approaches for integrating catastophes modeling into pricing and underwriting processes.

Proposed approaches

This paper does discuss and present ratemaking and underwriting approaches from a general level and irrespective of peril. Nevertheless, the scope of the study will be limited to hurricane risk modelling on behalf of AXA Mexico, an AXA Group subsidiary operating in Mexico. The underlying model used throughout this thesis is a model developed internally within AXA GIO. This model is at the origin of the different pricing methods proposed. We compare in particular, the frequency severity approach, the law calibration approach and Monte-Carlo to the computer retro-engineering approach.

Frequency severity approach

This approach is based on a severity frequency model implementing parametric and non-parametric models on losses simulated by the *CAT* model. Particular attention is therefore given to the composition of the portfolio to be provided in input. Indeed, a heterogeneity of this portfolio would lead to an error of appreciation of the potential explanatory variables relativities. In order to eliminate any bias inherent in the Mexican portfolio's exposure data, a notional portfolio of hypothetical properties with a uniform set of base features covering all geographic areas is constructed. The second step is to calibrate frequency and severity using parametric models such as GLMs and other machine learning methods (GBM, Random Forest) to compare their performance.
The final step involves appling discount curves to reflect financial conditions. This approximation can be quite rough, when considering a limited number of discount curves or when the financial conditions to be applied are relatively complex (multi-layering, reinsurance facultatives, min / max deductibles).

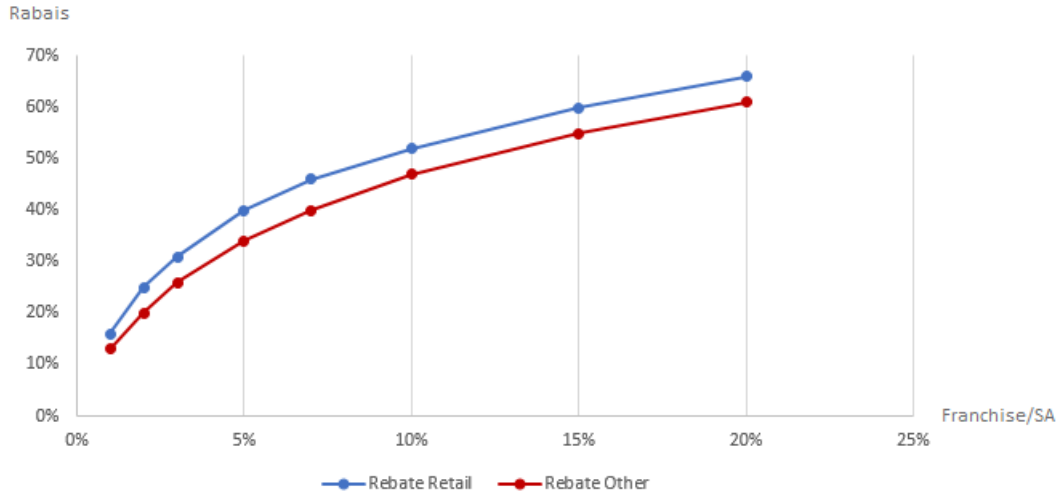


FIGURE 4 – Rebate curve used for the model Mexico hurricane

Therefore, we propose in the following to no longer model a single parameter of the loss distribution (currently the mean), but a vector of parameters characterizing the entire distribution. This should suffice for a better modelling of rebate rates nonlinear effect.

Distribution fitting and Monte-Carlo

This approach can be described in three main steps. The first is to calibrate the statistical laws that best adjust observed severity and frequency, for each site. The next step is to predict the parameters of these laws using statistical models.
Two types of errors are introduced to quantify the degree of confidence we have for each step :

- Adjustment error : Assuming that the observations describing each site are derived from simple distributions depending on a limited number of parameters, is a strong assumption.
- Prediction error : Assuming that the underlying loss distribution is known, what would be the prediction quality when estimating its parameters.

Finally the last step is to simulate the annual loss distribution calibrated for each site using Monte Carlo.
In order to assess the quality of this solution, this latter was applied across the entire portfolio to compute the relative error made on quantiles of losses at different return periods.

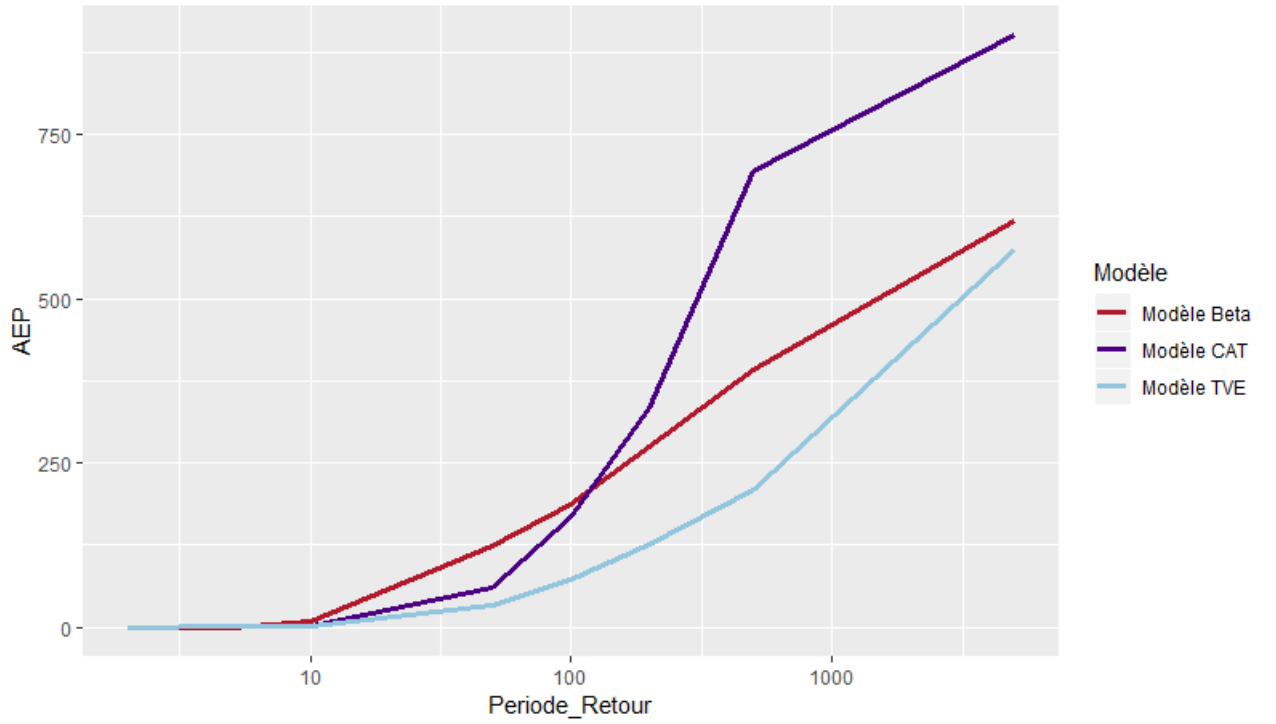


FIGURE 5 – AEP curves of a risky property site location using different methods

Computer reverse engineering approach

The idea behind this method is to replicate the loss calculation process of the *CAT* model but with a paradigm shift. The architecture of the *CAT* models and the resources needed to deploy them are optimal when it comes to pricing a portfolio and not just one or few policies. To address this issue, an elegant and ergonomic IT solution was to restructure the stochastic catalog of events using a database management system.

Once the incoming risk information is filled, the geographic coordinates will be used to extract wind speed values of all events hitting this location. Physical characteristics of the site will be crossed to vulnerability curves to obtain a destruction rate. Finally, the financial conditions entered will be applied to gross loss distribution to obtain the net version.

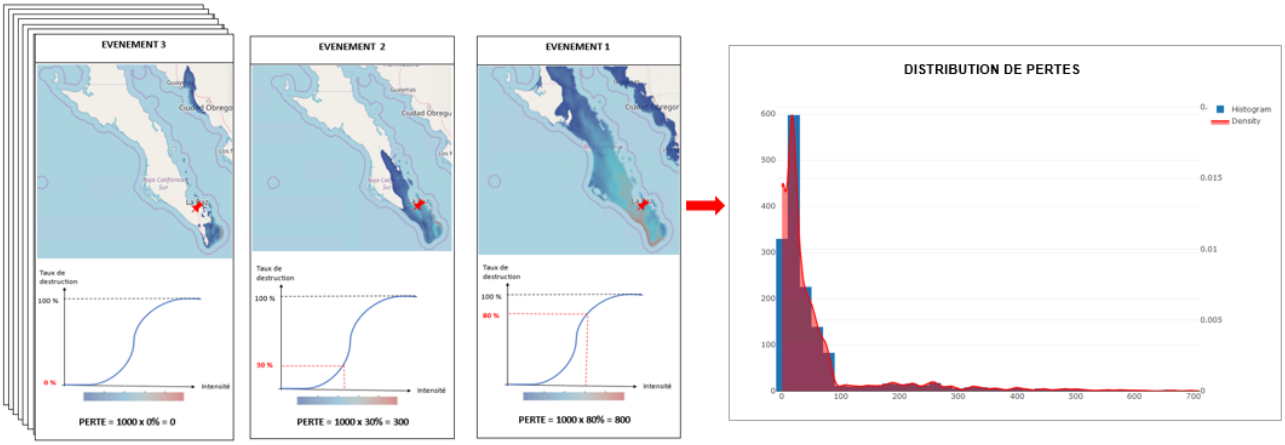


FIGURE 6 – Le process de l'outil

Unlike previous approaches, this last method relies only on the CAT model, in other words there is no other statistical model behind it.

As a result, the loss distribution generated instantaneously by this solution is identical to that generated by the model.

Results

The statistical basis of the frequency severity and the law calibration approaches are based on strong hypotheses rarely verified from the theoretical point of view. Moreover, the technical difficulty of their implementation leads rather to select the approach of the computer reverse engineering.
This approach based entirely on the CAT model, offers considerable advantages for the insurer and is a fully-fledged underwriting tool. Among built-in functionalities, we have :

- Pricing of complex reinsurance coverage structures including facultatives without rebate curves
- Calculates MPL Maximum possible loss or quantiles at different return periods.
- Measures the impact in insurance or reinsurance conditions changes (reduction/increase in policy shares, introduction of higher deductibles, assignment of facultatives)
- Takes into account geographical accumulation, by calculating the impact of an incoming risk or the cancellation of an existing one on portfolio bicentenary or at any other return period.

It should also be noted that all these results can be obtained at any level of granularity : site, police or client account.

Table des matières

Introduction **14**

1 Catastrophes naturelles en assurance **16**
1.1 Définition et chiffres clés . . . 16
1.2 Principes de la modélisation des catastrophes naturelles . . . 18
1.2.1 Module Exposition . . . 19
1.2.2 Module Alea . . . 20
1.2.3 Module Vulnérabilité . . . 21
1.2.4 Module Financier . . . 24
1.3 Résultats principaux des modèles CAT . . . 25

2 Rappels en réassurance **28**
2.1 Principes de la réassurance . . . 28
2.2 La réassurance proportionnelle et non proportionnelle . . . 29
2.2.1 Proportionnelle . . . 29
2.2.2 Non proportionnelle . . . 31
2.3 Tarification d'un traité . . . 33

3 Première approche de la tarification CAT **34**
3.1 Tarification en assurance non-vie versus tarification CAT . . . 34
3.2 Génération d'un portefeuille fictif . . . 34
3.3 Modèles linéaires généralisés GLM . . . 36
3.3.1 Théorie . . . 36
3.3.2 Application . . . 37
3.3.2.1 Modèle de sévérité . . . 38
3.3.2.2 Modèle de fréquence . . . 43
3.4 Modèles non paramétriques . . . 44
3.4.1 Théorie . . . 44
3.4.1.1 Random Forest . . . 44
3.4.1.2 Gradient Boosting . . . 46
3.4.2 Application . . . 47
3.4.2.1 Modèle de sévérité . . . 47
3.4.2.2 Modèle de fréquence . . . 49
3.5 Analyse des résultats . . . 51
3.5.1 Modèles de sévérité . . . 51
3.5.2 Modèles de fréquence . . . 53
3.6 Conditions financières : Courbes de rabais . . . 54
3.6.1 Définition des courbes de rabais . . . 54
3.6.2 La construction des courbes de rabais . . . 54
3.7 Comparaison des tarifs modélisées et observées . . . 58

4 Nouvelle approche statistique de tarification **59**
4.1 Calibration sur les données de pertes brutes . . . 60
4.1.1 Présentation des données et du contexte . . . 60
4.1.2 Choix de la loi statistique et calibration des paramètres . . . 61
4.1.2.1 Loi de fréquence . . . 63
4.1.2.2 Loi de sévérité . . . 65
4.2 Estimation des paramètres du modèle prédictif . . . 73
4.2.1 Sévérité en terme de taux de destruction . . . 73

4.2.2 Sévérité avec changement d'échelle 74
4.3 Comparaison des pertes modélisées et observées 76
4.4 Comparaison des tarifs modélisées et observées 77
4.5 Limites de cette approche 78

5 Prix technique et post-ajustement **79**
5.1 Chargement technique 79
5.2 Ajustement à la charge réelle 79

6 Restructuration de l'architecture informatique du modèle CAT **81**
6.1 Méthode développée 81
6.2 Implémentation de la méthode 84
6.2.1 Inputs 84
6.2.2 Outputs 84
6.2.2.1 Tarification 84
6.2.2.2 Impact marginal sur le PML 86
6.2.2.3 Coût de la réassurance 88
6.3 Changement de paradigme 90

7 Conclusion **93**

Bibliographie **94**

Introduction

L'avancée la plus significative dans la quantification des risques de Catastrophes Naturelles est le recours à la modélisation *CAT*. Ces modèles mathématiques fondés à partir d'hypothèses issues de la dynamique des aléas naturels et de lois statistiques, permettent de quantifier les dommages matériels et les pertes financières dus à un péril spécifique.
Depuis la fin des années 80, l'industrie de l'assurance utilise des modèles commerciaux développés par trois sociétés qui se partagent le marché mondial : RMS, AIR et EQECAT, constituant ainsi un oligopole dans ce secteur.
En 2012, le marché de la modélisation du risque de catastrophes naturelles accueille l'initiative *Oasis* qui vise à rassembler une large communauté composée d'assureurs, réassureurs et courtiers autour d'une même organisation afin de mieux appréhender les phénomènes climatiques et apporter de la transparence à la modélisation.

Étant donné son empreinte internationale dans des pays à forte sinistralité *CAT*, le groupe *AXA* devient un membre influant de cette communauté. Il dispose à présent de nombreux modèles pour différents périls et ce pour les différents pays de ses entités.
Actuellement, les modèles produits sont de plus en plus sophistiqués puisqu'ils disposent d'une résolution géographique de plus en plus fine et de catalogues d'événements stochastiques reposant sur des considérations scientifiques et physiques avancées.
Cette sophistication nous pousse à la réflexion quant à l'usage de ces modèles. Hormis les rôles qui leur sont connus à l'échelle macro pour calculer des niveaux réglementaires de fonds propres exigés ou encore l'optimisation du transfert de risques par le biais de la réassurance, les modèles *CAT* peuvent être utilisé à un niveau micro, notamment en tarification.
Ces derniers produisent une distribution de pertes par site et par événements dont peut être dériver plusieurs informations concernant les distributions de la fréquence et de la sévérité et des quantiles de pertes à différentes périodes de retour. Ces résultats permetteraient alors à l'assureur d'optimiser l'ensemble de sa chaine de valeur, de la souscription au coût du capital.

L'objectif de ce mémoire est de présenter différentes manières d'exploiter les résultats des modèles *CAT*, notamment en tarification. De plus, ce mémoire cherche à promouvoir le développement des modèles en interne en mettant en lumière l'intérêt qu'ils représenteraient en terme de création de valeur.
L'étude sera composée de six grandes parties :

— Les deux premières parties définissent les mécanismes et concepts généraux autour desquels s'articule le mémoire. La première partie présente les fondamentaux de la modélisation *CAT* ; les différents modules qui la composent, le processus de calcul de pertes ainsi que les résultats. La seconde, quant à elle, comprend des rappels en réassurance sur les différentes formes de couverture et leur méthodologie de tarification associée.

— La troisième partie propose une première méthode de tarification par une approche fréquence sévérité sur les résultats obtenus en sortie de modèle. Dans un premier temps, sera décrite la construction d'un portefeuille fictif ayant servi à l'obtention de cette base de modélisation. Dans un second temps, sera modélisée la fréquence et la sévérité moyenne calculées sur l'ensemble des événements composant le catalogue stochastique en utilisant des modèles paramétriques et non paramétriques.
Cette méthode repose sur la notion des courbes de rabais pour le passage de la prime brute à une prime nette de conditions financières. Nous montrons alors que cette approximation peut être remise en cause.

— La quatrième présente l'implémentation d'une nouvelle approche de tarification permettant de s'affranchir des courbes de rabais présentées à la section précédente. Cette dernière consiste à ne plus modéliser uniquement la moyenne des pertes mais tous les paramètres caractérisant leur distribution.

— La cinquième partie met en lumière l'étape de la mise à l'ultime de la tarification *CAT* garantissant ainsi la profitabilité de l'assureur.

— La sixième et dernière partie illustre l'intérêt de la modélisation en interne en proposant une approche innovante de tarification qui de par ses fonctionnalités permettra la conception d'un outil de souscription. Cet outil exploite le module d'aléa et de vulnérabilité pour reproduire le processus de calcul de pertes du modèle *CAT*. Grâce à sa nouvelle architecture informatique, il permet entre autre l'analyse prospective d'un nouveau risque et ce en évaluant l'impact marginal potentiel qu'aurait ce risque à différentes périodes de retour.

1 Catastrophes naturelles en assurance

1.1 Définition et chiffres clés

Une catastrophe naturelle est caractérisée par l'intensité anormale d'un agent naturel (inondation, coulée de boue, tremblement de terre, avalanche, sécheresse...) lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises.[1]
Chaque année des phénomènes météorologiques d'ampleur importante touchent différentes parties du monde : tempêtes de vent, ouragans, tremblements de terre, ...
En effet, en 2017 ces phénomènes ont amplifié le montant des pertes économiques. Le réassureur Munich Re a évalué le coût des catastrophes naturelles dans le monde à 330 milliards USD soit un montant deux fois plus important qu'en 2016.
Les dommages de 2017 sont principalement dus à trois événements : les ouragans *Harvey*, *Irma* et *Maria*. Ces ouragans survenus au mois de septembre dans l'Atlantique nord comptent pour 65% de l'ensemble des pertes, soit 215 milliards USD.

La graphique ci-dessous met en évidence l'écart entre les pertes assurées et les pertes économiques au fil du temps. Cet écart appelé *Insuffisance de protection* ou *Protection gap* en anglais, représente les pertes financières engendrées par des catastrophes naturelles et qui ne sont pas couvertes par l'assurance. En analysant la moyenne mobile des pertes économiques, nous constatons une tendance à la hausse de cette dernière. Par conséquent, le besoin d'une bonne maîtrise du risque *CAT* constitue un véritable atout auprès des compagnies d'assurance et de réassurance.

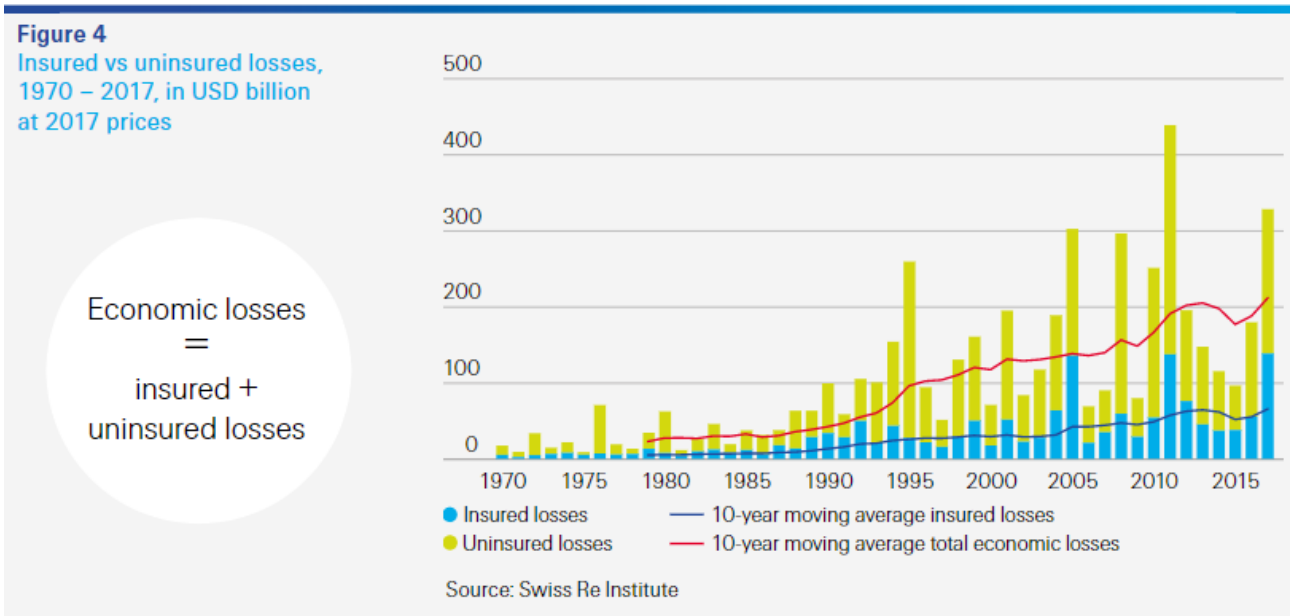


FIGURE 7 – Coût des catastrophes naturelles 1970 - 2017 en milliards USD au prix de 2017

En France, un arrêté interministériel constate l'état de catastrophe naturelle. Il permet l'indemnisation des dommages directement causés aux biens assurés, en vertu de la loi du 13 juillet 1982, relative à l'indemnisation des victimes de catastrophes naturelles.
Ce régime spécifique d'indemnisation, sous forme de partenariat public-privé, garantie à chaque citoyen français une indemnisation correcte en cas de sinistre causé par un phénomène naturel. En plus d'être obligatoire, la garantie pour les risques de catastrophes naturelles est automatique puisqu'elle est incluse dans tous les contrats d'assurance de dommage.

Un taux de surprime indifférencié, quelque soit le type de risque et l'exposition aux périls naturels, fixé par l'Etat et exprimé en pourcentage de la prime dommage, est payé par l'assuré. Il s'élève à 12 % de la prime relative aux garanties dommages pour les biens autres que véhicules à moteur et 6 % des primes vol et incendie pour les véhicules terrestres à moteur.[2]
Dans d'autres pays et contrairement au régime français où la tarification est non segmentée, s'exprime alors le besoin d'établir des stratégies de tarification et de souscription plus discriminantes.

1.2 Principes de la modélisation des catastrophes naturelles

De par sa nature, le risque de catastrophes naturelles est sujet au problème de la rareté des données historiques. Leur faible fréquence rend les approches actuarielles classiques hors d'usage et peu fiables pour évaluer les pertes futures potentielles. [6]

Les modèles CAT sont des modèles stochastiques permettant de réaliser des simulations de tous les événements probables sur une période longue et de mesurer les pertes potentielles engendrées. Ils reposent sur quatre modules différents :

— *Module d'exposition* : Ce premier module consiste à fournir des informations concernant la localisation géographique des sites assurés, leurs caractéristiques (ligne de métier,le nombre d'étage, l'année de construction, l'occupation du risque, le type de structure,...) et leur valeur assurée.

— *Module d'aléa* : Ce module permet de générer un catalogue complet d'événements probables selon des hypothèses géophysiques et météorologiques. Il résulte de ce module pour chaque localisation géographique et pour chaque événement une intensité exprimée selon la nature du péril en termes de : vitesse du vent, accélération de la vitesse du sol, hauteur d'eau, ...

— *Module de vulnérabilité* : Ce module permet d'assigner à chaque grandeur physique de l'aléa, pour chaque événement et type de bien un taux de destruction. Pour chaque combinaison d'événements simulés et sites une perte économique est alors calculée.

— *Module financier* : Ce dernier module calcule les pertes nettes de conditions d'assurance spécifiques à chaque contrat souscrit (franchises, limites,...) tenant également compte de la part de coassurance et conditions de réassurance.

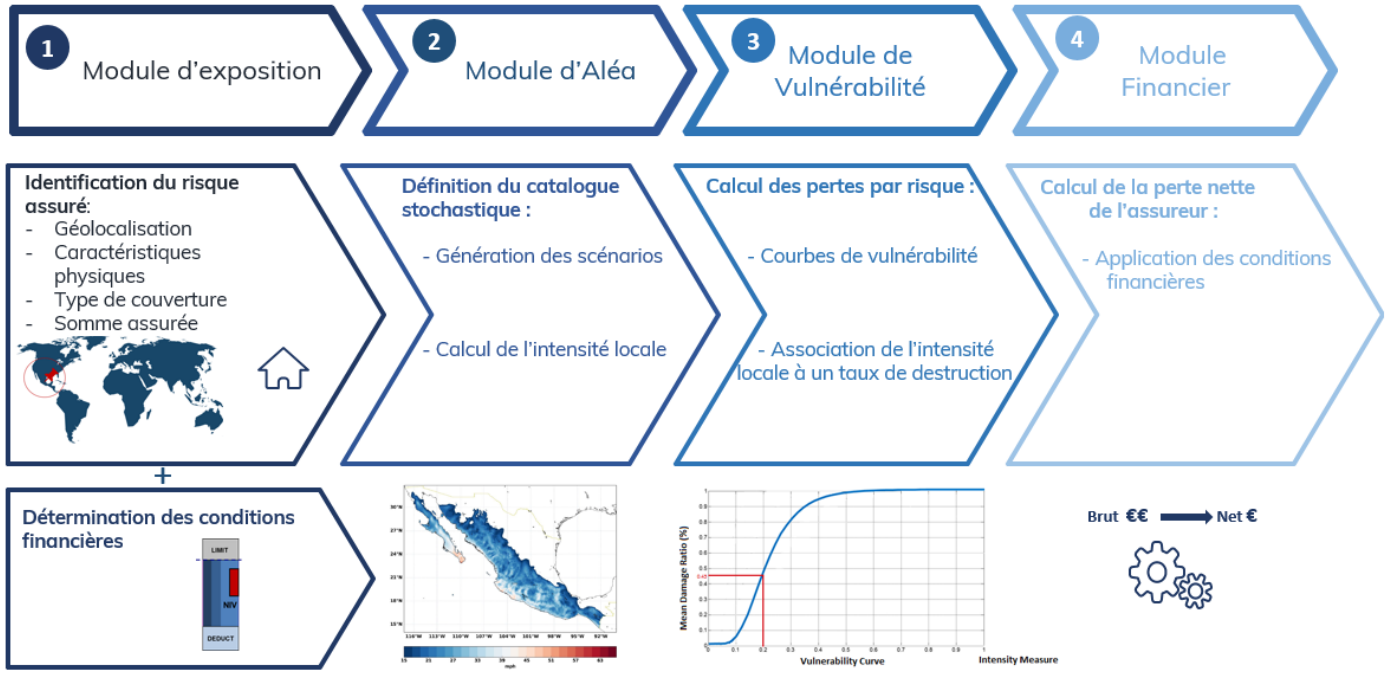


FIGURE 8 – Description d'un modèle CAT

1.2.1 Module Exposition

Ce premier module est le module via lequel nous importons des informations concernant les caractéristiques des sites présents dans le portefeuille et donc du risque auquel nous sommes exposés.

La vulnérabilité de chaque site peut dépendre de plusieurs de ses caractéristiques parmi elles nous retrouvons :

- *La Localisation* : Un événement est plus susceptible de toucher un site appartenant à son empreinte géographique et moins ceux qui n'y sont pas.
- *La branche d'activité* : La sensibilité des biens assurés aux différentes intensité du paramètre d'aléa est différente selon les branches d'activité. Les bâtiment industriels ou commerciales obéissent généralement à des normes plus prudentes visant à limiter voire éviter les pertes liées à des risques extrêmes.
- *Les matériaux de construction* : Le taux de destruction d'un bien construit en bois est plus vulnérable qu'un bien construit en acier ou en béton.
- *L'année de construction* : La vétusté des constructions assurées et donc leur vulnérabilité dépend bien évidemment de leur ancienneté d'une part. D'une autre part, suivant l'évolution du code de construction, de nouvelles normes et dispositifs sont établies permettant de renforcer la résistance du bâtiment.
- *Le nombre d'étages* : Il est intuitif de concevoir par exemple pour les tremblements de terre que plus les biens proches du sol subissent un taux de destruction plus important.
- *La somme assurée* : Il s'agit du montant garantie dans le contrat d'assurance qui est toujours divisé en trois garanties. La garantie bâtiment qui couvre les dégâts touchant le bâtiment (Building), la garantie contenu qui couvre les dégâts touchant les biens assurés (Contenu) et la garantie interruption d'activité qui couvre les dégâts causant une interruption d'activité (Business Interruption).

L'incertitude est au cœur de la modélisation du risque et exige une appréciation de la qualité et l'exhaustivité des données concernant l'exposition. En effet, la sophistication des modèles de catastrophes les rend avide de données et plus sensibles à leur qualité.
Les résultats peuvent changer de façon significative en fonction du niveau de résolution géographique et des détails concernant les caractéristiques physiques fournis au module d'exposition. Plusieurs études sur le marché montrent que l'amélioration de la qualité des données entraîne une réévaluation drastique du niveau des pertes estimées.
Les pertes modélisées sont donc aussi robustes que les données saisies qui leur sont associées.

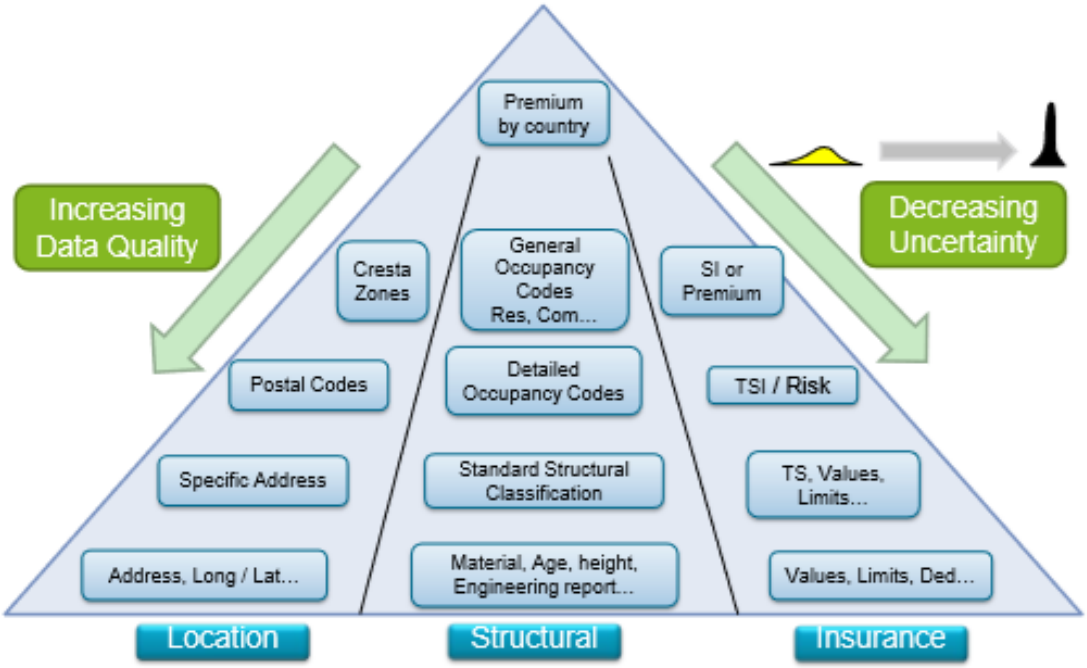


FIGURE 9 – Pyramide de la qualité de l'exposition

1.2.2 Module Alea

L'exposition à des phénomènes naturels à caractère aléatoire est prise en compte au travers de la distribution géographique, la fréquence de survenance du phénomène et sa sévérité. Afin d'avoir une vision réaliste du risque que représente un portefeuille, il est indispensable de disposer d'un catalogue avec tous les événements probables.
Plus l'historique des données est important en termes d'ancienneté et de complétude plus ces derniers refléteront une image conforme à la réalité du risque sous-jacent. Hors les données historiques catastrophes naturelles à notre disposition remontent approximativement à 30 ans. Durant cette période, il est tout à fait possible qu'aucun événement extrême ne soit survenu ou qu'un site ne soit jamais touché alors que d'autres sites voisins le soient.
L'objectif de ce module est donc de simuler un catalogue d'événements fictifs probables liés au phénomène *CAT* en question en se basant sur les données historiques et des hypothèses physiques issues de la dynamique des aléas naturels.

Malgré le fait que ces événements soient fictifs et n'ont jamais eu lieu auparavant, d'un point de vue scientifique il n'y a aucune raison pour que ces derniers ne surviennent pas dans le futur. Néanmoins, ces événements doivent être cohérents par rapport aux événements historiques à moins qu'ils y aient des raisons scientifiques qui justifient cette potentielle différence.

Selon le péril étudié, le module d'aléa peut comprendre des dizaines ou centaines de milliers d'événements qui constituent un modèle avec des plages de temps s'étalant sur plusieurs milliers d'années voire des dizaines de milliers d'années.

Pour modéliser les pertes causées par un événement potentiel, il faut connaître l'intensité de cet événement. Cette mesure d'intensité doit décrire de façon précise les caractéristiques d'un événement à un lieu donné et doit être corrélée au maximum possible avec les dommages causés. Par exemple pour les tremblements de terre sont utilisés l'accélération du sol PGA (*Peak Ground Acceleration*), la rigidité du sol vs30 (*velocity average over the top 30m of the soil*) ou encore la magnitude de l'échelle de Richter.

Pour chaque événement simulé, le module d'aléa produit alors une *empreinte géographique* qui associe à chaque localisation une intensité.

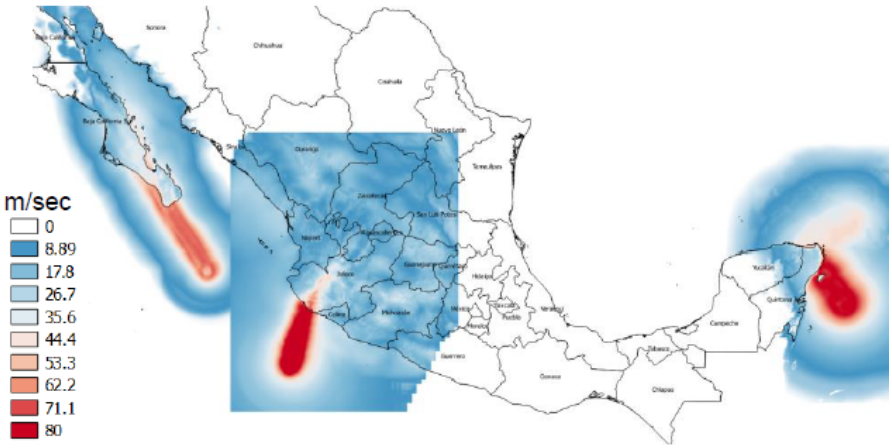


FIGURE 10 – Empreintes géographiques historiques d'ouragans mexicains : Odile 2014, Patricia 2015, Emily 2005

1.2.3 Module Vulnérabilité

A chaque occurrence d'une catastrophe naturelle, les dommages causés peuvent varier énormément d'un site à l'autre même dans le cas où ces derniers se voient attribuer la même intensité d'événement. En effet, la vulnérabilité d'un site dépendra aussi des matériaux avec lesquels il est construit, son âge et sa hauteur. Lorsqu'il s'agit de la garantie couvrant le *contenu* du site, le montant de perte sera différent selon que le site touché soit un magasin, une plate-forme pétrolière ou un jardin publique.

Le taux de destruction, défini comme étant le ratio des pertes divisés par la somme assurée ou en anglais la *Total Insured Value*, dépend donc de l'intensité de l'événement mais aussi de plusieurs caractéristiques intrinsèques du site assuré.

$$Taux\ de\ destruction = \frac{Pertes}{TIV}$$

TIV : Somme assurée (Total Insured Value).

Le module de vulnérabilité a donc pour objectif de définir les taux de destruction moyens en fonction des caractéristiques des biens assurés et de l'intensité de l'événement modélisé. En d'autres termes, à l'issue de ce module sont produites plusieurs *courbes de vulnérabilité* qui dépendent des caractéristiques du site, et qui associent un taux de destruction à chaque intensité de l'aléa modélisé (mouvement du sol, vitesse de vent,...).

Lors de la construction d'un modèle, les différents sites assurés sont regroupés selon des classes de risque pour lesquels une courbe de vulnérabilité unique peut être utilisée.
Par exemple, les sites de types bâtiments résidentiels seront groupés en une classe représentée par une unique courbe de vulnérabilité.

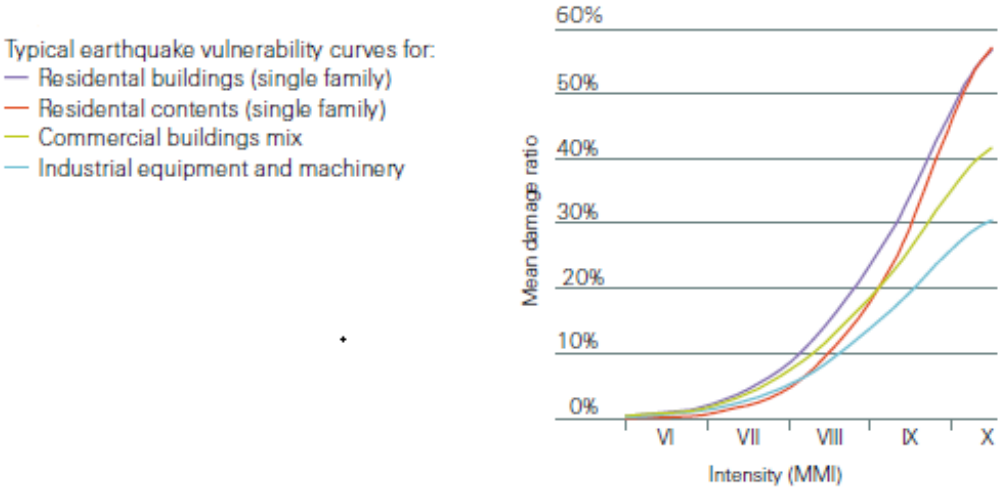


FIGURE 11 – Exemples de courbes de vulnérabilité

La construction des courbes de vulnérabilité se basent sur deux sources différentes. Idéalement, elles reposent sur les **historiques de sinistres**, plus ces derniers sont complets et récents meilleure sera notre estimation des taux de destructions. Cependant, pour des sinistres extrêmes dont la survenance est rare, les courbes de vulnérabilité reposent sur une **expertise d'ingénieurs** analysant la fragilité d'un bâtiment suite à différents niveaux de l'aléa physique considéré (mouvement du sol, vitesse de vent,...).[8]

Deux structures exposées au même niveau d'aléa et avec des caractéristiques similaires peuvent en réalité subir des pertes différentes. Il s'agit d'une incertitude dite de *second degré*, c'est l'incertitude quant à la perte structurelle associée à un risque physique, une situation géographique lorsqu'un événement survient. Elle comprend différentes sources :

- *Incertitude du modèle* : Il existe une incertitude par rapport à la valeur de l'aléa physique que l'on associe à une situation géographique lors de la construction du modèle catastrophe. En fonction des hypothèses faites et des bases de données utilisées il existe plusieurs modèles alternatifs permettant de calculer des intensités locales.
- *Traduction du niveau d'aléa en taux de destruction* : Les courbes de vulnérabilité se basent sur des historiques de sinistres ainsi que des expertises d'ingénieurs pour définir la fragilité d'un bâtiment face à une intensité d'aléa physique. Néanmoins, il existera toujours des caractéristiques intrinsèques au site que l'on aura du mal à capter et qui impacteront sa sensibilité.
- *Précision des informations de l'exposition* : Les informations concernant l'exposition peuvent être mal renseignées. Imputer des valeurs par défaut au informations manquantes est inéluctablement une source d'incertitude supplémentaire.

Les modèles CAT tiennent compte de cette variabilité. La prise en compte de cette dernière dépend du péril étudié et fait que chaque valeur d'intensité d'aléa n'est pas associée à un taux de destruction moyen unique mais une distribution de taux de destruction.

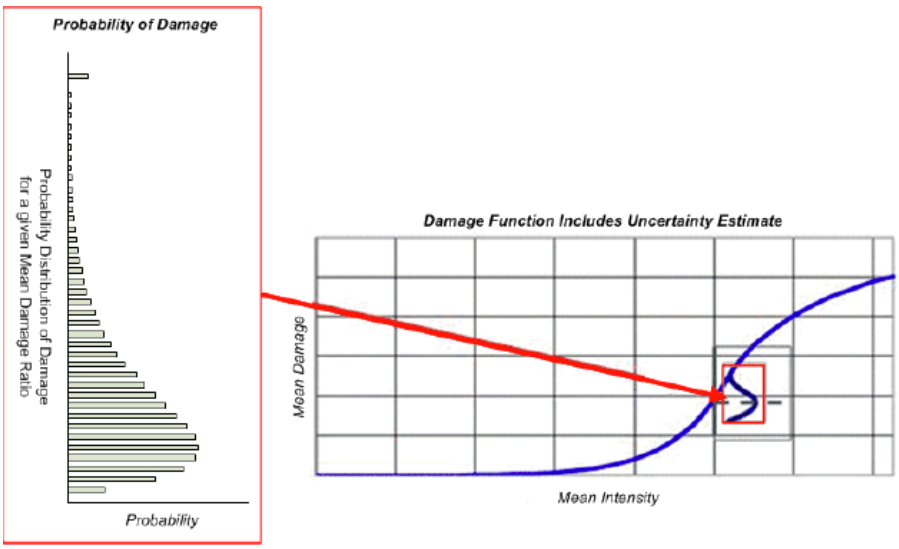


FIGURE 12 – Distribution autour du taux de destruction moyen

Pour la plupart des périls, tenir compte de cette variabilité consiste à appliquer un coefficient de variation COV appelé aussi *event volatility factor*. Ce coefficient suit une loi log-normale centrée sur 1 et permet de modéliser l'écart au taux de destruction moyen. Il est calibré sur les données sinistres et varie selon la branche d'activité (lob) par exemple. Nous pouvons le définir de la façon suivante :

$$COV_{lob} = \frac{\sigma_{lob}}{\mu_{lob}}$$

avec :

— σ_{lob} : Ecart-type des pertes associé à une branche d'activité

— μ_{lob} : Moyenne des pertes associée à une branche d'activité

1.2.4 Module Financier

Le module financier vient compléter la vision de la *perte économique* en appliquant les conditions d'assurance à la perte brute résultant des modules décrits précédemment. Ces conditions financières permettent aux assureurs ou réassureurs de limiter leur pertes éventuelles.
En effet, ces dernières permettent d'une part de réduire l'exposition de l'assureur/ réassureur. D'une autre part, elles permettent d'alléger les tâches administratives et ce en se débarrassant de la gestion des petits sinistres qui sont assez fréquents.
Les conditions d'assurance varient en fonction des marchés, des périls et des polices assurés. Elles comportent souvent des *franchises* et des *plafonds*.

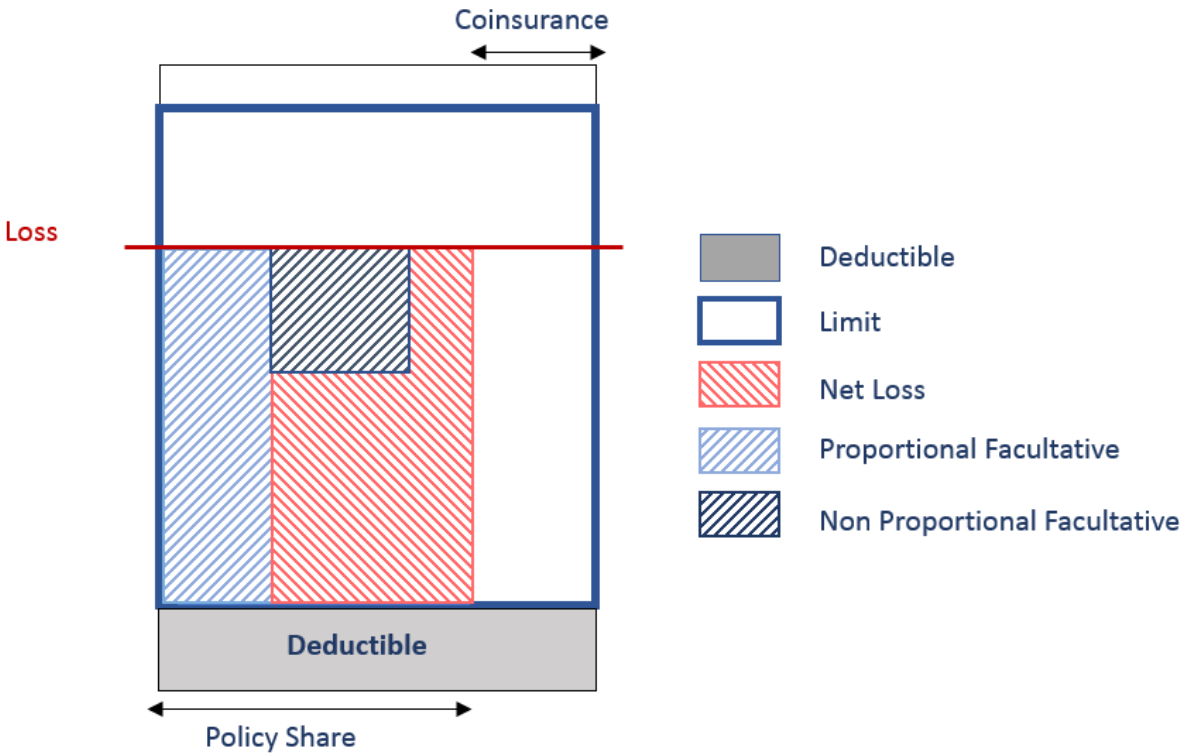


FIGURE 13 – Application des conditions financières

Ces derniers peuvent être appliqués à différents niveaux. En effet, ils peuvent être appliqués au niveau granulaire le plus fin : le site, comme ils peuvent être appliqués au niveau de la police lorsque le site est couvert avec d'autres au sein d'une même police (*multi sites*). Ils peuvent également varier selon le type de garantie : Building, Content et Business interruption.

Afin de limiter son exposition sur les risques individuels, il est courant qu'un assureur veuille partager le risque avec un ou plusieurs assureurs *(principe de coassurance)* ou de souscrire à des *facultatives*.
Par conséquent, les architectures de protection de l'assureur peuvent être complexes incluant ainsi des conditions d'applications et d'agrégations spécifiques.
A l'issue de ce module, il est alors possible d'extraire :

— GUL *(Ground Up Loss)* : Les pertes avant application des conditions d'assurance
— GROSS *(Gross Loss)* : Les pertes après application des franchises et limites
— NET *(Net Pre Cat Loss)* : Les résultats après application de la réassurance par risque

1.3 Résultats principaux des modèles CAT

Il résulte du modèle CAT plusieurs sorties permettant de quantifier le risque d'un portefeuille d'assurance face à l'aléa naturel sous-jacent. Parmi ces sorties nous retrouvons les *courbes EP* et la *Event Loss Table*.

Les courbes EP

Les courbes EP (*Exceedance Probability*) expriment, comme leur nom l'indique, la probabilité que les pertes excèdent un seuil donné.
Ces courbes sont régulièrement utilisées en réassurance et surtout en modélisation CAT. En effet, elles permettent d'identifier la probabilité qu'un portefeuille subisse une perte durant sa période de couverture ainsi que sa probabilité de perte totale. De plus, elles constituent des benchmarks pertinants lorsqu'il s'agit de comparer différents risques ou différents horizons temporels.
Les courbes EP peuvent représenter deux grandeurs différentes :

Notations :
Y_i *la variable aléatoire représentant le montant de perte causé par l'événement* i
N *la variable aléatoire représentant le nombre d'événements*

— Les courbes OEP(*Occurrence Exceedence Probability*), associent une période de retour (en année) au coût maximal d'un événement.

$$OEP(x) = P(\max_{i=1,\dots,N}(Y_i) > x)$$

Ensuite la période de retour en années est simplement donnée par :

$$RP(x) = \frac{1}{OEP(x)}$$

L'OEP ne tient compte que de l'événement le plus destructeur de chaque année, cette grandeur ne reflète donc qu'une partie des pertes annuelles d'un portefeuille.

— Les courbes AEP (*Aggregate Exceedence Probability*) associent une période de retour (en année) au coût total des événements.

$$AEP(x) = P(\sum_{i}^{N} Y_i > x)$$

Le choix de la métrique d'intérêt se fait en fonction de la nature des risques couverts. Si une garantie ne couvre que des événements extrêmes, l'assureur s'intéressera à l'OEP.
En revanche, si la garantie couvre également la sinistralité attritionnelle, il sera plus pertinent d'analyser l'AEP.

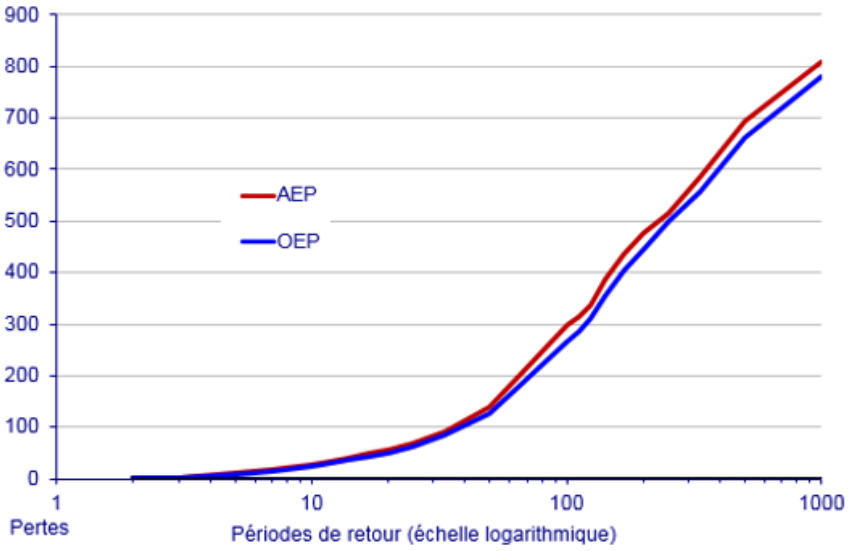


FIGURE 14 – Exemple d'OEP et AEP

Dans cet exemple, un événement générant une perte de 350 millions $ se produit en moyenne une fois tous les 200 ans (quantile à 99,5%). En d'autres termes, chaque année ce portefeuille a une chance sur 200 de subir une perte de 350 millions $ suite à un unique événement.
Nous pouvons également remarquer que la différence entre l'AEP et l'OEP sur la figure 14 n'est pas importante. C'est souvent le cas lorsqu'il s'agit de catastrophes naturelles puisqu'une année n'est que très rarement constituée de plus d'un événement extrême.

Year Loss Table
La table YLT est une table qui se présente sous la forme suivante :

YearID	EventID	LocationID	Loss
1	Hurricane1	1	21552
1	Hurricane2	2	31524
1	Hurricane3	2	572
2	Hurricane4	3	0
2	Hurricane5	4	1435
2	Hurricane6	5	0
...	...	...	...
10000	Hurricane4225	5000	254

TABLE 1 – Year Loss Table

— YearID : L'identifiant de l'année simulée
— EventID : L'identifiant de l'événement
— LocationID : L'identifiant du site touché par l'événement
— Loss : La perte du site suite à l'occurrence de l'événement

La table YLT permet d'obtenir une distribution de pertes annuelles. La construction de cette dernière se fait en deux étapes. Dans un premier temps, nous effectue du Monte-Carlo pour simuler un certain nombre d'années (10000 années dans cet exemple).
Ensuite, pour chaque année simulée, nous tirons aléatoirement des événements. Les empreintes associées à ces événements sont croisées avec les sites constituants le portefeuille et leurs pertes

respectives.

Annual Aggregate Loss

Les résultats des modèles CAT permettent également de calculer le coût des périls CAT appelé AAL pour *Annual Aggregate Loss en anglais*.
L'AAL peut être utilisé à des fins de tarification puisqu'il correspond à la moyenne des pertes annuelles agrégées sur l'ensemble des événements du catalogue stochastique. Cette métrique peut être calculée à tous les niveaux de granularité de la façon suivante :

$$AAL_i = \frac{\sum_{j=1}^{n_{sim}} \sum_{k=1}^{n_j} DR_{ikj} TIV_i}{n_{sim}}$$

n_j : Le nombre d'événements survenu l'année j
n_{sim} : Le nombre d'années simulées sur le catalogue stochastique
DR_{ikj} : Taux de destruction de i subi suite à l'événement k de l'année j
TIV_i : Valeur assurée de i

Enfin, la validation de l'utilisation d'un modèle CAT repose en grande partie sur la notion de Back-testing. Le test rétroactif de validité ou back-testing en anglais consiste à tester la pertinence d'une modélisation en s'appuyant sur un large ensemble de données historiques réelles.
Concrètement, il s'agit de resimuler des événements historiques et les comparer aux pertes réelles mises en «*as-if*» c'est à dire corrigées de l'inflation et des évolutions de l'exposition du portefeuille.

2 Rappels en réassurance

La réassurance permet aux assureurs de céder la part de risque qui excède leur capacité de souscription, ou des risques qu'ils ne veulent pas supporter seuls.
Ce transfert de risque peut prendre différentes formes et contenir des clauses multiples. Il est donc nécessaire voire indispensable de connaître tous ces concepts afin d'évaluer un risque à sa juste valeur.
Cette section sera dédiée aux rappels des principes et définitions de base en réassurance que l'on retrouvera à maintes reprises tout au long de ce mémoire. Nous allons dans un premier temps définir les mécanismes de la réassurance et ses fonctions. Ensuite, nous allons présenter les traités de réassurance classiques pour finalement analyser leurs avantages et inconvénients respectifs.

2.1 Principes de la réassurance

Il est d'usage de distinguer trois modes de réassurance traditionnelle que l'on peut discerner selon les dispositions du traité de réassurance : nature du risque couvert, spécificité des obligations contractuelles et part de polices concernées.

— La réassurance obligatoire : On dit de la réassurance qu'elle est obligatoire lorsque la cédante s'engage à céder tous ses risques et la compagnie de réassurance s'engage à les accepter. Ce risque porte donc sur l'ensemble du portefeuille de la cédante.

— La réassurance facultative : On dit de la réassurance qu'elle est facultative lorsque la cédante a le choix de céder un risque au réassureur qui est libre après analyse de l'accepter ou de le refuser. Ce type de contrat est généralement mis en place lorsqu'il s'agit de grands risques industriels ou des risques atypiques de part leur exposition ou leur rareté.

— La réassurance Facultative Obligatoire : Ce contrat donne à la cédante la liberté d'inclure ou non au traité de réassurance les risques de son choix. Le réassureur quant à lui s'engage à accepter toutes les affaires que la cédante lui soumettra (sous réserve que l'engagement ne dépasse pas un montant maximum fixé initialement). Autrement dit, dans ce genre de contrat, l'assureur est en facultative et le réassureur est en obligatoire. De ce fait, le poids de ce type de traité reste traditionnellement limité en réassurance.

— La réassurance Obligatoire Facultative : Ce contrat engage la cédante à présenter tous ses risques et donne au réassureur la liberté de les accepter ou pas. Le réassureur est dans ce cas en facultative et l'assureur est en obligatoire. Cette forme de réassurance n'existe pas en pratique.

Ces différents modes de réassurance présentent des avantages et des inconvénients du côté du réassureur et de la cédante que l'on peut résumer à travers le tableau suivant :

Types	Vision	Avantages	Inconvénients
Facultatives	*Réassureur*	— Vision parfaite du risque — Liberté de souscription	— Anti-sélection possible (L'assureur ne montre que ce qu'il veut et garde les bons risques) — Beaucoup de facultatives pour mutualiser un seul risque — Négociation longue et coûteuse, ainsi qu'une gestion exigeante
Obligatoires	*Cédante*	— Sécurité de transférer les risques — Gestion allégée par rapport aux facultatives	— Existence d'exclusions — Obligation systématique de céder

TABLE 2 – Avantages et inconvénients des différents modes de réassurance

2.2 La réassurance proportionnelle et non proportionnelle

On distingue deux formes de réassurance obligatoire : la réassurance proportionnelle et la réassurance non-proportionnelle.

2.2.1 Proportionnelle

Un traité de réassurance de type proportionnel se caractérise par l'égalité entre la proportion de primes reçues par le réassureur et la proportion du coût des sinistres transférés au réassureur. Ce principe de proportionnalité comprend une notion de partage des coûts. En effet, le réassureur verse à la cédante une commission de réassurance à titre de participation aux frais d'acquisition et de gestion des affaires qui lui sont cédées.
Deux formes de traité existent en proportionnel : le traité en quote-part, en anglais Quota share (QS) et le traité en excédent de plein ou Surplus (SP) en anglais.

Le traité en Quote Part

Dans la quote-part, le réassureur prend en charge une proportion constante sur tous les risques du portefeuille. Dans l'exemple ci-dessous, ce taux est fixé à 30%.

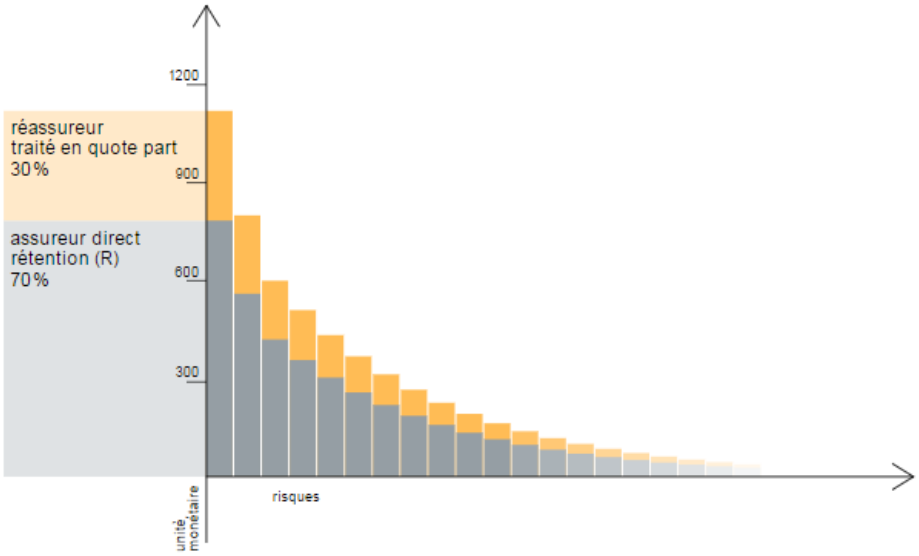


FIGURE 15 – Exemple d'un traité en Quote-Part

Le traité en excédent de pleins

Dans l'excédent de plein, le réassureur prend en charge uniquement la portion des risques dépassant un niveau de capital constant appelé plein de rétention. Dans l'exemple ci-dessous, ce montant est fixé à 300. Contrairement à la quote part, le taux de cession n'est pas fixe. Ce dernier varie selon la somme assurée de la police et peut être calculé comme suit :

$$\tau_i = \frac{min(max(S_i - K, 0), C)}{S_i}$$

avec :

- S_i : la somme assurée de la police i
- K : le plein de rétention de la cédante
- C : la capacité qui représente le montant à hauteur duquel le réassureur s'engage à payer son client.

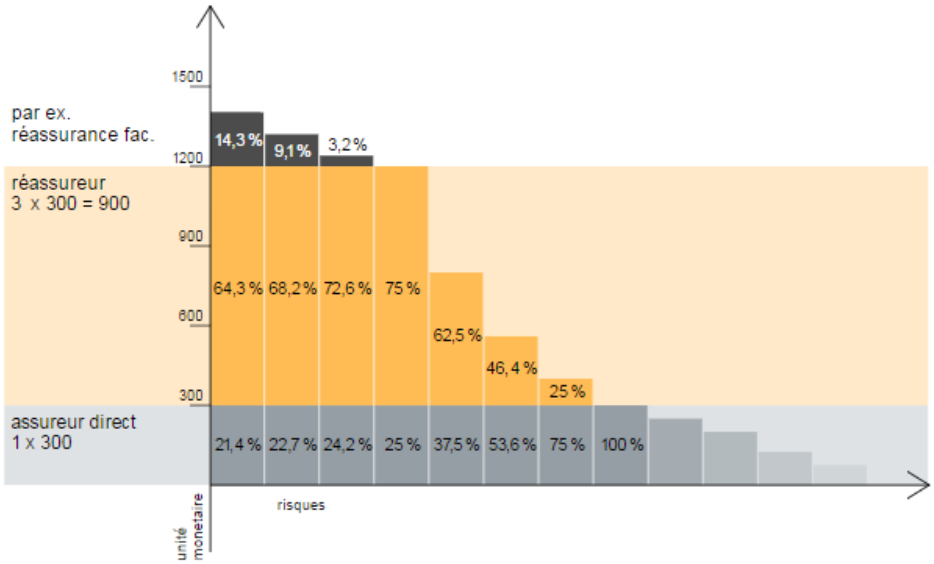


FIGURE 16 – Exemple d'un traité en excédent de plein

2.2.2 Non proportionnelle

Les traités non-proportionnels sont définis par une priorité et une portée. La priorité étant le seuil à partir duquel le réassureur intervient dans la prise en charge du sinistre, et la portée étant une limite à l'engagement du réassureur.
La réassurance non proportionnelle distingue deux types de contrats :

— L'excédent de sinistres noté XS

— L'excédent de perte annuelle

L'excédent de sinistre XS

L'excédent de sinistre XS ou *excess-of-loss* en anglais sont des traités où le réassureur prend à sa charge le montant de sinistre qui excède la priorité dans la limite de la portée. On les note :

Portee XS Priorite

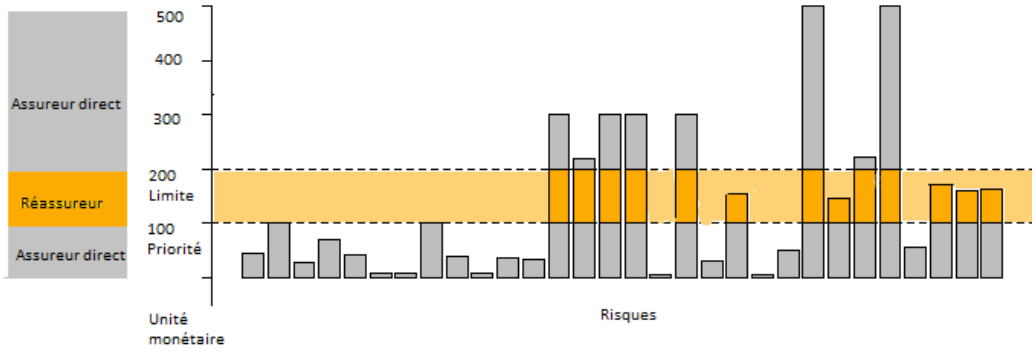


FIGURE 17 – Exemple d'un traité en excédent de sinistre

Par conséquent, si S est le montant du sinistre, le montant à la charge du réassureur sera :

$$min(max(S - priorite, 0), portee)$$

Il est aussi à noter que ces traités peuvent être définis **par risque**, c'est à dire que le réassureur intervient à chaque fois qu'un sinistre supérieur à une priorité survient pour une police donnée. Comme ils peuvent être définis **par événement**, dans ce cas le réassureur intervient à chaque survenance d'événement.

L'excédent de perte annuelle

L'excédent de perte annuelle ou *stop loss* en anglais est utilisé lorsque la sinistralité se mesure à l'année. La particularité du *stop loss* par rapport à l'*excess-of-loss* est le fait que la priorité et la portée sont définies en fonction d'un pourcentage de la sinistralité rapportée à l'assiette de prime de la cédante. Ce traité permet par conséquent l'équilibre du bilan technique annuel de cette dernière. Un traité en excédent de perte est noté suivant le même modèle que le XS :

Portee SL Priorite

avec :

— Priorité : Seuil d'intervention exprimé en ratio de perte

— Portée : Capacité d'intervention exprimée en ratio de perte

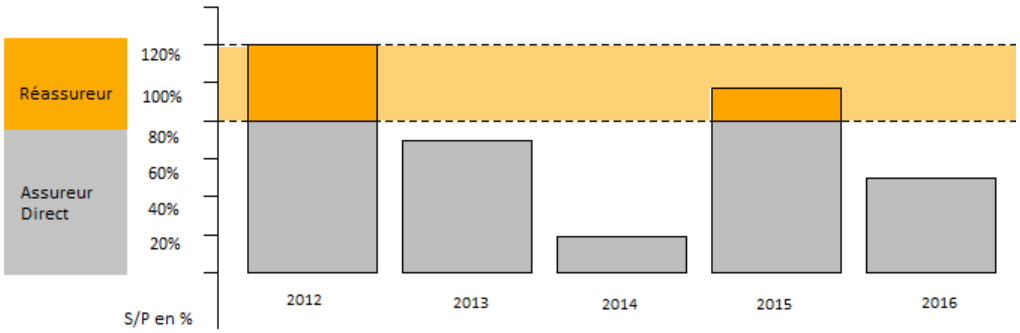


FIGURE 18 – Exemple d'un traité en excédent de perte annuelle 120% SL 80%

La notion de reconstitution

En pratique, les termes de l'excédent de sinistre, la prime de réassurance versée par la cédante au réassureur ne constitue qu'une prime de base dédiée à couvrir la première portée. Pour éviter que la cédante soit à découvert dans le cas où cette première portée soit totalement absorbée, les réassureurs ont mis en place **la reconstitution de garantie**. Cela signifie qu'après la survenance de chaque sinistre, l'assureur est obligé de payer une prime de reconstitution permettant le renouvellement de la portion de portée absorbée par le sinistre précédent et ce dans la limite de la garantie annuelle.
Lorsque le réassureur accorde N reconstitutions de garantie, l'assureur peut bénéficier de N+1 fois de la portée de son contrat XS. A priori lorsque rien n'est précisé, cela implique que les reconstitutions sont illimitées et gratuites.
Par exemple : Si S représente le montant total, π la prime pure de base, et P la portée. En considérant que le réassureur s'engage à renouveler la portée initiale de l'XS r fois, avec un prime de reconstitution à ri % de la prime initiale.
Alors la prime additionnelle π_i associée à la ième reconstitution s'écrit :

$$\pi_i = r_i \frac{\pi}{p}(S - (i-1)p)\mathbb{1}_{(i-1)p \leq S \leq ip} + p\mathbb{1}_{S \geq ip}$$

Dans la suite du mémoire, nous considérerons que les primes de reconstitutions sont gratuites et illimitées.

2.3 Tarification d'un traité

Un traité de réassurance couvre un ensemble de risques d'un portefeuille. En fonction des risques couverts, il existe une multitude de méthode de tarification : la tarification sur expérience, la tarification probabiliste ainsi que la tarification sur exposition.

La méthode de tarification présentée dans cette section est basée sur le modèle *CAT* et concernera les traités en excédent de sinistres par événement qui est la couverture privilégié pour les risques CAT NAT. Cette méthode ne tient pas compte des primes de reconstitutions qui sont considérées comme étant gratuites et illimitée.

Comme mentionné à la section dédiée aux rappels en réassurance, le montant à la charge du réassureur dans un traité en excédent de sinistre C XS D prend la forme suivante à chaque survenance d'événement :

$$min(max(X - D, 0), C)$$

La prime pure du traité peut donc s'écrire de la sorte :

$$PP = \frac{1}{n_{evt}} \sum_{i=1}^{n_{evt}} \min(max(X_i - D, 0), C)$$

La cédante fait payer à l'assuré une prime pure *CAT*, qui contient l'ensemble de la distribution des pertes possible de la police. De ce fait, la prime pure versée au réassureur PP est déjà contenue dans la prime versé par l'assuré.

Une seconde composante vient s'ajouter à la prime pure. Cette dernière est liée à l'accumulation de risques dans une même zone ; un facteur non pris en compte dans la prime CAT. Cette deuxième partie correspond au coût de réassurance et celui du capital. En effet, l'accumulation de risques entraîne a priori une volatilité importante qui pousse la cédante à se couvrir via la réassurance. Il s'agit donc de la prime commerciale du traité que l'on formule ainsi :

$$P_C = \frac{PP + C\sigma}{1 - g}$$

avec :

PP	:	La prime pure du traité
σ	:	La volatilité des récupérations de réassurance
C	:	Coefficient de chargement de sécurité
g	:	Chargement rémunérant la gestion du traité, des sinistres associés, etc

Le coût de réassurance peut être obtenu à partir de la différence entre la prime commerciale et la prime pure :

$$CoR = P_C - PP = \frac{C\sigma + gPP}{1 - g}$$

La couverture en réassurance permet la diminution du montant requis en capitaux propres de la cédante. En revanche, le coût de la réassurance augmente.
La « *création de valeur* » est la différence entre l'économie de capital obtenue (positive) et le coût de la réassurance (négatif).

$$\text{Création de valeur} = \Delta SCR(\phi - \alpha) - CoR$$

avec :

ϕ	:	Le taux de rendement des actifs
α	:	Le taux de rendement des actionnaires

Toutes choses égales par ailleurs, la création de valeur consiste en une optimisation entre le coût de la réassurance et le coût du capital.

3 Première approche de la tarification CAT

3.1 Tarification en assurance non-vie versus tarification CAT

Pour les garanties d'assurance en automobile par exemple, un historique de données considérable est exploité afin d'estimer les primes selon différents critères de risque.
Pour ce type de garantie, l'importance des données sinistres et l'indépendance entre les accidents survenus nous permet d'utiliser les méthodes actuarielles classiques pour évaluer le risque.
Les données historiques de catastrophes naturelles quant à elles sont peu nombreuses et l'hypothèse d'indépendance n'est pas plausible. L'exploitation classique de ces données ne nous permet donc pas d'estimer la probabilité de survenance d'un péril ni l'intensité des dommages causés.
Rappelons aussi que ces données historiques remontent approximativement à 30 ans, alors que certains périls comme les tremblement de terre peuvent avoir des périodes de retour de dizaines de milliers d'années. L'historique des données n'est donc pas assez important en termes d'ancienneté et de complétude pour refléter une image conforme à la réalité du risque sous-jacent.
Les modèles *CAT* permettent l'utilisation des informations physiques de l'aléa naturel (via leur catalogues stochastiques et historiques). Ils habilitent donc les assureurs à allouer proprement le risque d'un péril naturel entre les différentes parties prenantes de la même façon que pour des risques attritionnels comme en automobile.

La tarification de ce risque repose donc sur les résultats des modèles *CAT*. En revanche, afin d'éviter de relancer la totalité de la chaîne de modélisation présentée à la section précédente pour tarifer chaque nouveau site souscrit, nous souhaitons calibrer un modèle statistique sur les pertes résultant de la modélisation *CAT*.

Nous partons donc du principe que la construction d'un modèle prédictif robuste qui retranscrit au mieux les résultats du modèle *CAT* nous permettrait non seulement d'obtenir un prix de façon immédiate et à moindre coût mais aussi d'identifier l'impact des différentes variables sur la vulnérabilité.

Cette section sera donc dédiée à la présentation des résultats des différentes méthodes d'apprentissage utilisées pour le risque Ouragan au Mexique. Dans un premier temps, nous expliquerons l'importance de l'exposition fournie en input et la manière dont elle a été construite. Ensuite, nous modéliserons la fréquence ainsi que la sévérité en utilisant des GLM et d'autres algorithmes machine learning. Enfin, nous introduirons le concept des courbes de rabais permettant le passage de la prime brute à la prime nette et leurs constructions. Nous clôturerons cette section sur la problématique qui découle de l'utilisation de ses courbes et en proposant de nouvelles méthodes pour y répondre.

3.2 Génération d'un portefeuille fictif

Précédemment, lors de notre présentation des modèles *CAT*, nous avions défini le premier module - le module d'exposition - comme étant le module qui consistait à fournir l'exposition pour laquelle les pertes sont calculées.
Pour une rétro-conception efficace des résultats du modèle *CAT*, le portefeuille fourni en input dans ce premier module doit vérifier deux conditions :

— Une couverture sur l'ensemble du pays avec une résolution géographique relativement

proche de celles des empreintes d'événements.

— Une répartition géographique uniforme des différents profils de risque définis lors du module de vulnérabilité.

L'hétérogénéité du portefeuille fourni en input induit une erreur d'appréciation des relativités des variables explicatives potentielles du modèle. Afin de l'illustrer, imaginons qu'on veuille tarifer un nouveau site situé dans une zone excentrée et plus risquée par rapport aux zones où on souscrit d'habitude. Ayant négliger cette nouvelle zone et tous les événements qui la touchent, nous aurons sous estimé le risque qu'elle représente. Nous introduisons alors un biais dont l'importance dépend du degré de dis similarité de cette nouvelle zone par rapport au reste du portefeuille, tant dans l'estimation de l'impact de valeur de l'aléa naturel que dans la segmentation des zones de risques.
Par conséquent, l'idéal serait donc de créer un portefeuille fictif qui attribuera à chaque point du Mexique autant de sites que de profils de risques.
Cependant, le run du portefeuille permettant d'obtenir les sorties CAT est coûteux en termes de temps et puissance de calcul d'où le besoin de restreindre le nombre de ligne du portefeuille. L'objectif est donc de créer un portefeuille fictif représentatif du risque par zone assurée à moindre coût. Le nombre de ligne du portefeuille dépend de plusieurs facteurs dont il faudra tenir compte pour réduire son volume sans perte d'informations :

— Les hypothèses du modèle : certaines hypothèses sont très importantes et jouent un rôle primordial lors de la définition des courbes de vulnérabilité, c'est notamment le cas pour les hypothèses caractérisant le site : le type de garantie, les matériaux de construction, l'année de construction, ...

— Granularité géographique : Les modèles *CAT* considèrent des granularité géographiques très fines qui peuvent aller jusqu'à une résolution de 25 mètres pour les modèles inondations. Le fait d'opter pour une granularité plus grossière réduirait la taille du portefeuille. Par conséquent, il est envisageable de substituer une répartition uniforme des profils de risque au niveau site (un point) par une répartition à une granularité moins fine.

Il y a donc arbitrage entre le nombre de points à considérer qui doit prendre en compte toutes les valeurs de l'aléa, le nombre d'hypothèses du modèle déterminants le taux de destruction et la qualité de représentation du risque.

Nous cherchons à calibrer un GLM pour le risque ouragan au Mexique. Initialement, la résolution géographique adoptée au sein de l'équipe de modélisations pour les ouragans est de 1km.
La première étape consiste à choisir subtilement un nombre de points ainsi que leur espacement de façon à ce que les valeurs de l'aléa sous-jacent au péril étudié soient assez représentatifs du risque de la zone. A l'issu de cette étape, le nombre de points considérés passe de 1.6 millions de points répartis sur l'ensemble du pays à une centaine de milliers.
La seconde étape consiste à réduire le nombre d'hypothèses du modèle *CAT*. Comme mentionné lors de la présentation des modèles CAT, différentes courbes de vulnérabilité sont construites selon les différents profils de risques considérés. Notre objectif à ce stade sera alors de réduire le nombre de profils de risque en utilisant des algorithmes de *clustering*.

Les profils de risques sont définis comme étant le produit cartésien fini de toutes les hypothèses du modèle CAT. En d'autres termes, il s'agit de tous les k-uplets dont les composantes appartiennent aux k variables du site prise en compte dans le modèle.
A l'étape finale, pour chaque point sera tiré aléatoirement un profil de chaque cluster.

FIGURE 19 – Grille originale

FIGURE 20 – Grille après réduction

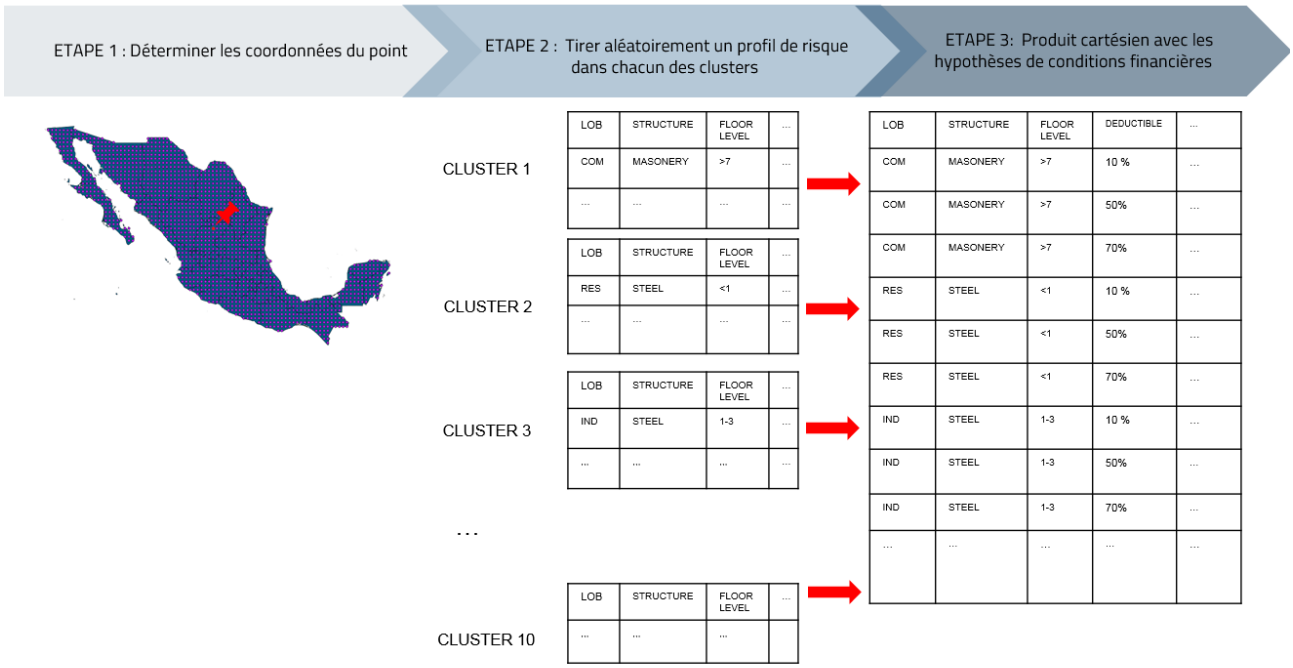


FIGURE 21 – Processus génération du portefeuille

3.3 Modèles linéaires généralisés GLM

3.3.1 Théorie

Les modèles linéaires généralisés sont définis par deux hypothèses : une hypothèse concernant la distribution de la variable réponse et une hypothèse concernant sa structure . La première hypothèse consiste en l'appartenance de la distribution de la variable réponse à la famille de distribution exponentielle.

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

avec :
θ : Le paramètre canonique ou le paramètre de position qui indique la tendance centrale de la distribution
ϕ : Le paramètre de dispersion ou paramètre d'échelle
w : Le poids attribué aux observations
L'hypothèse structurelle concerne l'inversibilité de la fonction de lien g et détermine la manière avec laquelle la variable réponse est liée à une combinaison linéaire des variables explicatives.

$$g(\mu) = x'\beta$$

$$g(\mathbb{E}(Y|x_1, ..., xp)) = \sum_{k=1}^{p} \beta_k x_k$$

Pour une sous famille donnée de la famille exponentielle, nous pouvons choisir plusieurs fonctions de lien dont la fonction de lien canonique. Cette dernière est la transformation inversible g tel que le paramètre canonique est égal à la combinaison des variables explicatives $x'\beta$:

$$\theta = x'\beta \iff g(b'(\theta)) = \theta$$

La fonction de lien $g(\mu) = log(\mu)$ permet par exemple de modéliser le logarithme de l'espérance et conduit au modèle multiplicatif suivant :

$$\mathbb{E}(Y|x) = exp(\beta x) = exp(\sum_{k=1}^{p} \beta_k x_k)$$

Qualité d'ajustement

Afin d'évaluer la qualité d'ajustement du modèle sur la base des différences entre observations et estimations. Différents critères sont proposés :

Déviance

Le modèle estimé est comparé au modèle saturé, c'est-à-dire le modèle possédant autant de paramètres que d'observations et estimant donc exactement les données. Cette comparaison est basée sur l'expression de la déviance notée D des log-vraisemblances $\mathcal{L}$ et $\mathcal{L}_{sat}$:

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

Cette dernière représente le logarithme du carré du rapport des vraisemblances. Ce rapport "généralise" l'usage des sommes de carrés propres au cas gaussien et donc à l'estimation par moindres carrés.
D suit une loi du χ^2 à n-p-1 degrés de liberté. Cela nous permet de construire un test de rejet ou d'acceptation du modèle selon que la déviance est jugée significativement ou non importante.

Test de Pearson

Un test du χ^2 est aussi utilisé pour comparer les valeurs observées y_i à les valeurs prédites par le modèle. La statistique du test est définie par :

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{\hat{Var}(\hat{\mu_i})}$$

elle suit asymptotiquement la même loi que la déviance. Normalement si l'approximation de la loi asymptotique est bonne ces deux approches conduisent à des résultats peu différents.

3.3.2 Application

Après avoir lancer le modèle Mexique Ouragan sur le portefeuille simulé précédemment, nous récupérons les pertes relatives aux sites de ce portefeuille. Ces résultats constituent les données sur lesquelles nous nous basons pour construire nos modèles.
Nous adopteront alors une approche fréquence sévérité modélisant ainsi la fréquence et la sévérité annuelle calculées sur une base de 10000 années simulées. La majorité des variables à notre disposition proviennent du portefeuille et représentent les caractéristiques physiques des sites assurés.

— **Type de structure** : Structure en béton armé, béton et acier

— **Ligne de métier** : Collective, résidentiel, commercial, industriel

— **Le nombre d'étages**

— **Coordonnés géographiques** : longitude et latitude du site

— **TIV** : Somme assurée du site

Afin d'enrichir notre base et de prendre en compte l'exposition géographique au risque ouragan nous avons crée de nouvelles variables :

— **Distance à la côte** : Grâce aux coordonnées géographiques des sites, nous avions calculé la distance à la côte.

— **CRESTA** : En utilisant les coordonnées géographiques, nous avions associé à chaque site la cresta (pour *Catastrophe Risk Evaluation and Standardizing Target Accumulations*) à laquelle il appartient.

De plus, dans le but de prendre en compte la vitesse de vent, nous utilisons une carte de risque conçues à partir de trajectoires historiques d'ouragans.
La carte mise à notre disposition se base sur les enregistrement des fréquences de cyclone à différentes intensités sur un maillage de 10km*10km avec prise en compte de l'aléa spatial et statistique.
En utilisant la segmentation faite sur les vitesses de vent, nous pouvons définir différents niveaux de risques :

— **Zone de Risques** : Variable catégorielle allant de 0 (pour les zones sûres) à 5 (pour les zones les plus risquées).

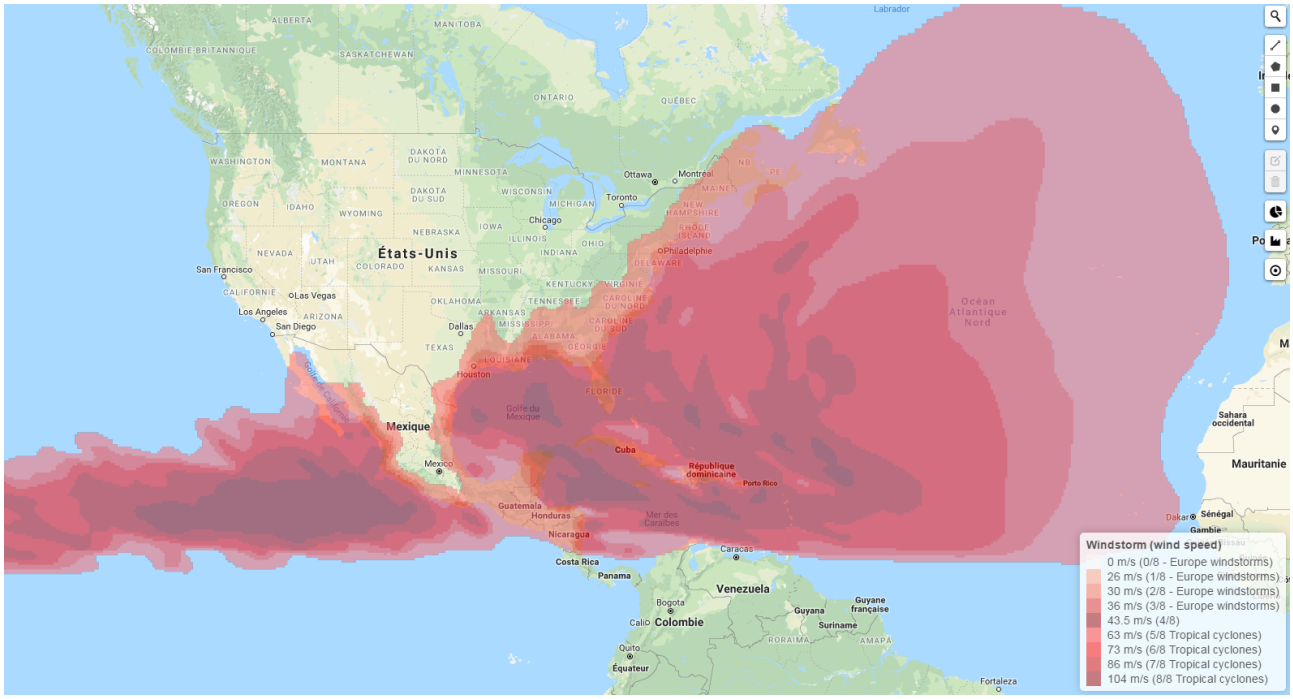


FIGURE 22 – Carte de risque d'ouragans tropicaux

3.3.2.1 Modèle de sévérité

Pour modéliser la sévérité, plutôt que de considérer le coût annuel en montant absolu, nous allons nous intéresser au taux de destruction défini comme étant le ratio des coûts annuels relatifs à la somme assuré.

Afin d'appréhender les relations de nature non linéaire, capturer les effets de groupes, mais aussi détecter les valeurs aberrantes ou extrêmes, il est très utile de représenter graphiquement la variable dépendante en fonction des variables explicatives.

FIGURE 23 – Taux de destruction moyen par CRESTA

La figure 23 représente la distribution géographique du taux de destruction (pour mille). Comme escompté, les crestas sur les côtes atlantiques et pacifiques sont les plus exposées.
En effet, un ouragan qui se développe en mer jusqu'à atteindre une puissance incroyable, s'éteint quand il atteint les côtes et traversent les terres : il perd ses principales sources d'énergie et s'effondre.

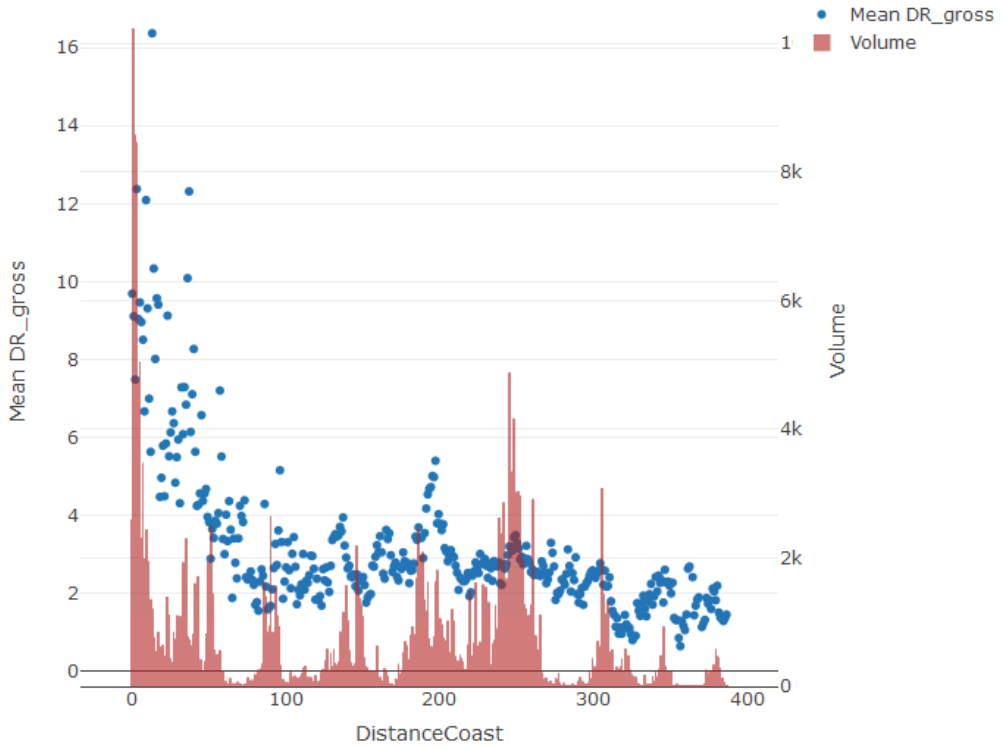


FIGURE 24 – Taux de destruction moyen en fonction de la distance à la côte

La figure 24 représente le taux de destruction moyen et le volume des valeurs prises par la variable représentant la distance à la côte. Cette figure appuie le constat précédent, les côtes subissent les taux de dommage les plus importants. Il est aussi indispensable de noter l'effet non linéaire de cette variable sur le taux de destruction. Ceci soulève un point très important qui se pose souvent lors de l'intégration de variables continues. Plusieurs méthodes sont spécifiquement prévues à cet effet :

— Transformation monotones des variables explicatives : exponentiel, logarithme, exposants,...

— Introduction de régression polynomiale

— Introduction de régression splines

— Transformation en variables catégorielles

De manière générale, l'idéal est de relâcher en amont l'hypothèse de linéarité du prédicteur que l'on impose dans le GLM, en utilisant des fonctions splines et des modèles GAM puis d'identifier en fonction de l'allure de la courbe la façon la plus convenable pour introduire la variable dans le modèle.

Les fonctions Splines Le principe général des fonctions splines en analyse statistique des données repose sur trois étapes :

- Les valeurs de la variable continue X sont découpées en (k+1) intervalles. On choisit pour cela k valeurs numériques définissant les bornes des intervalles, appelées *noeuds* .
- Au sein de chaque intervalle, la relation entre la variable dépendante Y et la variable explicative X est modélisée par un polynôme de degré d.
- Les coefficients des polynômes sont choisis de sorte que la courbe totale soit régulière autrement dit de sorte à ce qu'elle soit lisse et sans rupture. Mathématiquement parlant, cela veut dire que la courbe doit être continue, et dérivable (d-1) fois.

Une fonction spline S(x) de degré d avec les nœuds s_1, ..., s_k est alors définie par :

$$S(x) = \alpha_0 + \sum_{i=1}^{d} \alpha_i x^i + \sum_{j=1}^{k} \beta_j P_d(x, s_j)$$

avec $P_d(x, s_j) = (x - s)_+^d$

Dans le cadre de notre étude, nous avons deux variables continues qui sont la distance à la côte et le nombre d'étages. Regardons à présent l'influence de ces deux variables sur la sévérité. Nous nous limitons au cas d'une seule variable explicative et nous ajustons le taux de destruction grâce à des splines cubiques.

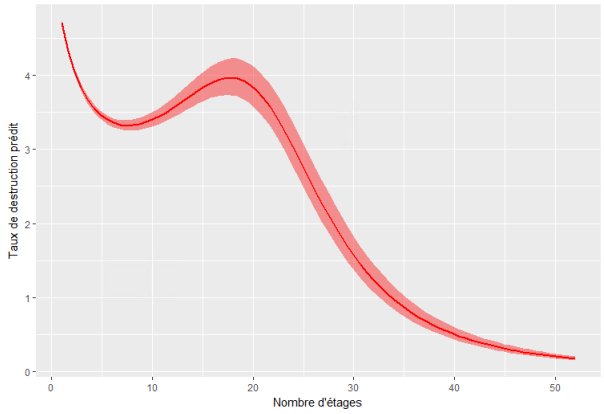


FIGURE 25 – Effet de la variable du nombre d'étages sur la sévérité

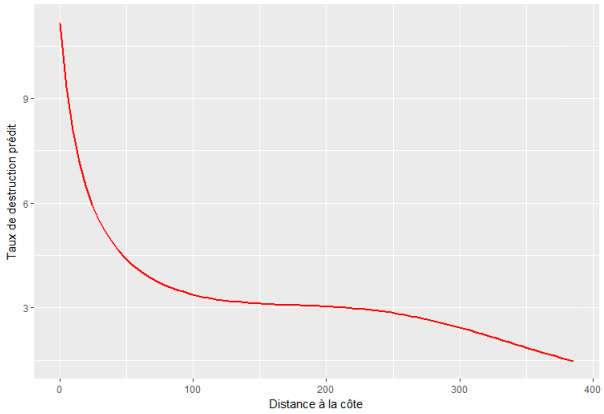


FIGURE 26 – Effet de la variable de distance à la côte sur la sévérité

Les figures montrent une influence clairement non linéaire des deux variables sur la sinistralité. Nous remarquons que l'effet marginal de la distance à la côte est en adéquation avec ce qui a été observé dans la figure 24. Nous constatons également que les sites ayant plusieurs étages sont moins vulnérables. Cependant, cette figure est à interpréter avec prudence du fait de la rareté des sites ayant plus de 5 étages et donc de la dispersion des taux de destruction qui leur sont associés.
Pour des raisons de confidentialité, nous ne présenterons ni les coefficients estimés ni les variables retenues du modèle. Néanmoins, nous pouvons montrer les relativités de quelques variables.

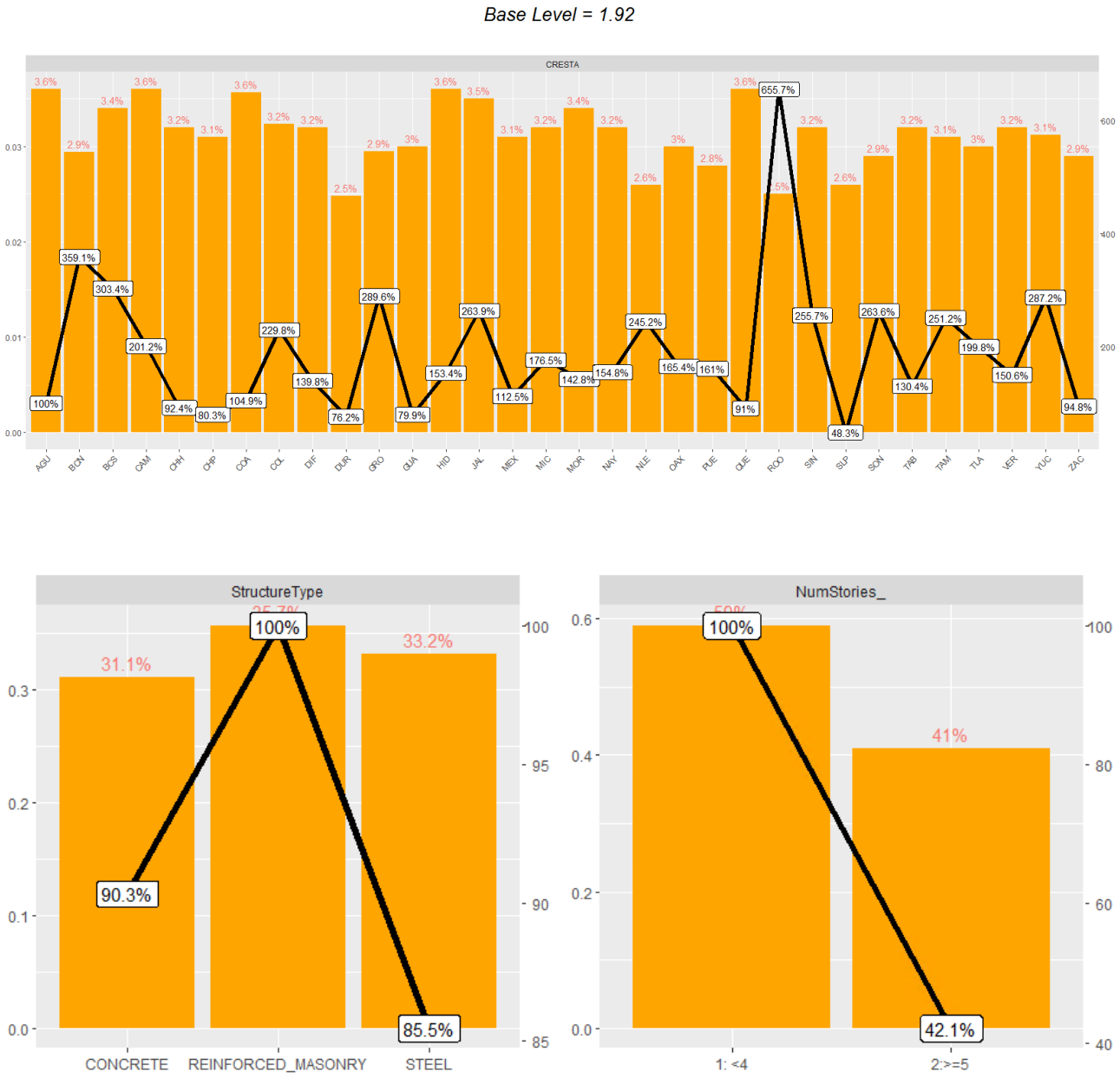


FIGURE 27 – Relativités du GLM de sévérité

Le niveau de base de 1,92 correspond au taux de destruction pour mille pour les modalités de référence dans le modèle.
Toutes choses étant égales par ailleurs, nous constatons que la région d'Aguascalientes est 3 fois moins risquée que la basse californie et 6 fois moins que le Quintana Roo. De même pour la structure, toutes choses égales par ailleurs un bâtiment en structure armé est plus vulnérable qu'un ouvrage en béton de 10% et de 15% par rapport à ceux en acier. Enfin pour le nombre d'étages, le résultat confirme les constats de l'étape d'investigation. Un site ayant moins de 5 étages est plus de deux fois plus risqué qu'un site plus haut *ceteris paribus*.

3.3.2.2 Modèle de fréquence

La modélisation de la fréquence se fait usuellement avec les lois de Poisson, binomiale négative et binomiale.

— Poisson : V(N) = E(N) , on parle d'équi-dispersion.

— Binomiale négative : V(N) > E(N) , on parle de sur-dispersion.

— Binomiale : V(N) < E(N), on parle de sous-dispersion.

Ces lois sont donc utilisées sur des données de comptage autrement dit des valeurs entières non négatives. Dans notre cas, la variable réponse représente le ratio du nombre d'événements touchant le site sur le nombre d'années simulées considérées par le modèle CAT.

Cette situation pourrait être abordée par l'analyse de deux modèles. Dans le premier, nous supposons des erreurs binomiales pour calibrer la fréquence, alors que dans le deuxième, nous utiliserons des erreurs de poisson pour approximer le nombre total d'événement sur la totalité des années simulées.
Néanmoins, le modèle binomial et de Poisson se ressemblent fortement. Il est connu que si $X \sim B(n,p)$, alors la valeur ajustée de X est $E(X) = np$ et la variance est $V(X) = np(1-p)$. Comme la taille de la base de données est grande (n grand) et les probabilités de survenance sont faibles (p petit), la distribution de Poisson de paramètre $\lambda = np$ est une bonne approximation. Le modèle binomial et de Poisson fournissent alors des résultats sensiblement identiques. De ce fait, nous commençons par construire un modèle de poisson classique, et on teste la sur-dispersion pour voir si les propriétés de la loi de poisson (soit l'égalité entre la moyenne et la variance) sont vérifiées.

Test de sur-dispersion
Le test peut se présenter de la façon suivante : on suppose que :

$$V(Y/X) = E(Y/X) + \tau E(Y/X)^2$$

On cherche à tester :

$$H_0 : \quad \tau = 0 \quad \textit{contre} \quad H_1 : \quad \tau > 0$$

Une des statistiques de test serait :

$$T = \frac{\sum_{i=1}^{n}[(Y_i - \hat{\mu}_i)^2 - Y_i]}{\sqrt{2\sum_{i=1}^{n}\hat{\mu_i}^2}}$$

qui suit sous H_0 une loi normale centrée réduite.

Sans grande surprise, dans notre cas nous rejetons l'hypothèse nulle et nous optons pour une négative binomiale pour tenir compte de la sur-dispersion.

Comme pour le modèle de sévérité, ne nous détaillerons pas les résultats du modèle de fréquence. Néanmoins nous présenterons les coefficients estimés pour l'une des variables les plus influentes du modèle : AXAZONE.

Le niveau de base 371.76 correspond au nombre d'événements survenus sur l'ensemble des années constituant le catalogue stochastique pour les modalités de référence dans le modèle.
Toutes choses étant égales par ailleurs, nous constatons que la zone la plus risquée , en l'occurrence la zone 3 est 4 fois plus exposée que la zone 0 plutôt centrale.

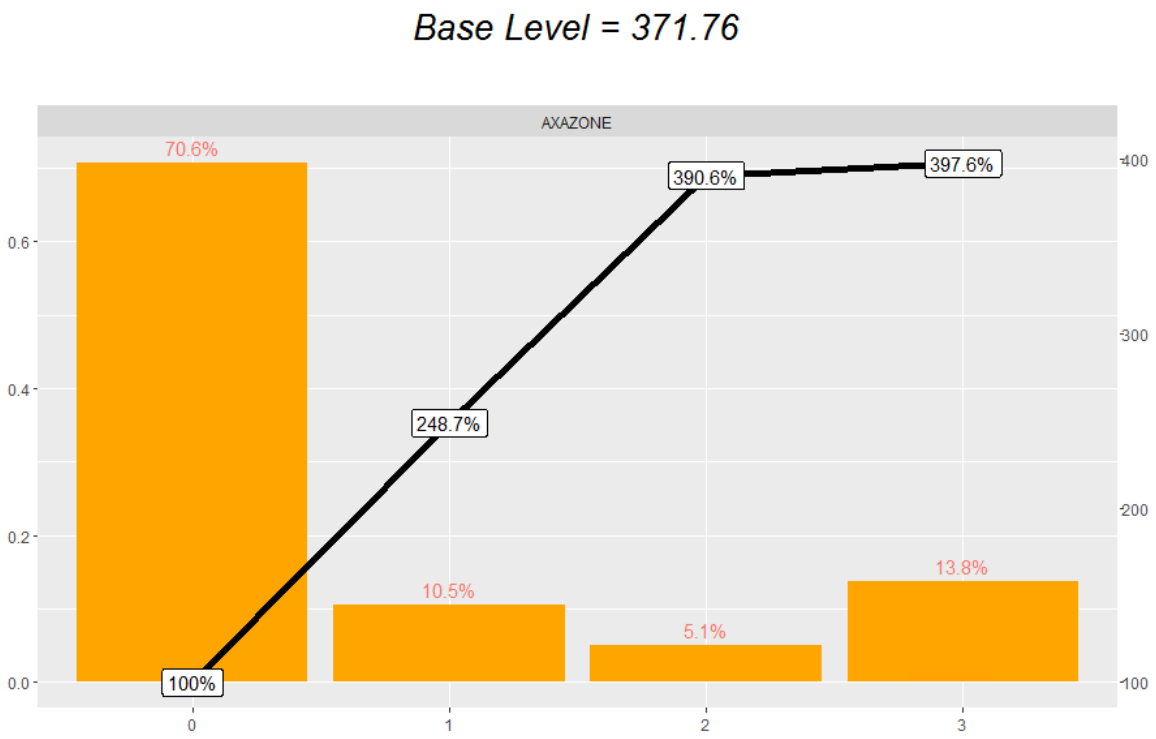


FIGURE 28 – Relativité du GLM de fréquence

3.4 Modèles non paramétriques

De nos jours tous les secteurs d'activité sont concernés par la data science y compris le secteur de l'assurance / réassurance. Le big data aide à la réalisation de leurs objectifs principaux en améliorant la conformité, la compréhension du risque et l'expérience client.
Certains modèles sont complexes à opérer et se déclinent en plusieurs sous-objectifs. Lorsque l'objectif est prédictif, c'est à dire lorsque l'on veut atteindre un niveau de précision optimale, les algorithmes deviennent plus complexes et plus dur à interpréter.
La réglementation en assurance est extrêmement exigeante lorsqu'il s'agit de justifier et expliquer des modèles tarifaires c'est pourquoi les modèles linéaires généralisés sont encore largement utilisés. En effet, ils permettent une compréhension du modèle, donc de l'impact des variables, attributs ou facteurs.
Néanmoins, ces modèles imposent des hypothèses fortes concernant la distribution des données et engendre une perte de précision quand les variables explicatives ne sont pas au bon ordre ou lorsque des interactions sont omises .
Les actuaires doivent donc décider du degré d'intelligibilité à sacrifier pour améliorer les performances du modèle en terme de précision.
Nous allons tester deux modèles non paramétriques dont le but commun est d'améliorer la performances du classifieur faible qu'est l'arbre de décision mais dont les approches sont différentes : les forêts aléatoires (*Random Forest*) et les Gradient Boosting Machine. Les performances de ces modèles seront ensuite comparées à celles des GLM calibrés précédemment.

3.4.1 Théorie

3.4.1.1 Random Forest

L'algorithme des forêts aléatoires fait partie des algorithmes de machine learning les plus populaires. Il repose sur une méthode d'agrégation appelée **B**ootstrap **Agg**regat**ing** ou bagging que l'on applique aux arbres de décisions CART.

Le but est de construire plusieurs arbres de décisions avec des variables initiales différentes et provenant d'échantillons différents. Introduire de l'aléa dans la construction de chaque arbre nous permet ainsi de décorréler la collection d'arbres provenant du Bagging.
Le Bagging est une application de la procédure Bootstrap aux arbres de décisions CART connu comme étant un algorithme machine learning à forte variance. La collection de prédicteurs obtenus à partir de B échantillons bootstrap est alors agrégée en faisant simplement une moyenne ou un vote majoritaire. Dans un contexte de régression, les forêts aléatoires procèdent comme suit :

1 : for b in 1, . . . , B do :
2 : Construire un échantillon bootstrap ξ_b
3 : Construire un arbre de régression à partir de ξ_b : $\hat{\Phi}_b$ en répétant de manière récursive les étapes suivantes :
Pour chaque nœud terminal, jusqu'à atteindre une règle d'arrêt. Il en existe plusieurs, celle proposée par le package R *RandomForest* : on ne découpe pas un nœud qui contient moins d'observations qu'un nombre fixé au préalable n_{min} .

- Séléctionner de façon aléatoire m variables des p variables.
- Sélectionner la coupure minimisant la fonction de coût de CART sur ces m variables
- Découper le nœud en deux nœuds fils

4 : end for
5 : Construction de l'estimateur $\hat{\Phi}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{\Phi}_b(x)$

Dans un contexte de classification, l'étape finale d'agrégation dans l'algorithme consiste à faire voter les arbres à la majorité. On sait que l'erreur moyenne quadratique peut être décomposée de la sorte :

$$MSE = Variance + Biais^2$$

Les tirages bootstrap sont effectués de la même manière et indépendamment les uns des autres. Les arbres sont donc identiquement distribués ce qui signifie que le biais obtenu à l'issu de l'agrégation des arbres est similaire à celui obtenu par un seul arbre de décision. Ce biais peut être minimisé en construisant des arbres assez profonds.
Le bagging permet de minimiser l'erreur moyenne quadratique en minimisant la variance. En notant $\sigma^2(x)$ la variance des estimateurs qu'on agrège et $\rho(x)$ la corrélation entre deux estimateurs que l'on agrège (calculés sur deux échantillons bootstrap).

$$Var(\frac{1}{B}\sum_{b=1}^{B}\hat{\Phi}_b(x)) = \frac{1}{B^2}\sum_{b=1}^{B}\sum_{j=1}^{B} cov(\hat{\Phi}_b(x), \hat{\Phi}_j(x))$$

$$= \frac{1}{B^2}(\sum_{b=1}^{B} Var(\hat{\Phi}_b(x)) + \sum_{1\leq j\neq b\leq B}^{B} cov(\hat{\Phi}_b(x), \hat{\Phi}_j(x)))$$

$$= \frac{B(B-1)\sigma^2(x)\rho(x) + \sigma^2(x)B}{B^2}$$

$$= \frac{(B-1)\sigma^2(x)\rho(x)}{B} + \frac{\sigma^2(x)}{B}$$

$$= \sigma^2(x)\rho(x) - \frac{\sigma^2(x)\rho(x)}{B} + \frac{\sigma^2(x)}{B}$$

$$= \sigma^2(x)\rho(x) + \frac{\sigma^2(x)(1-\rho(x))}{B}$$

Le premier terme de l'expression ci-dessus montre que cette variance baisse en réduisant la corrélation entre les arbres ρ qui est à son tour une fonction décroissante du nombre de variables sélectionnées m. Le deuxième terme de cette équation montre que pour un ρ constant et $\rho \leq 1$ la variance baisse en fonction de B le nombre d'arbres. Le choix de B n'est donc pas crucial pour la performance de l'estimateur, il est recommandé de prendre le plus grand B possible (en fonction du temps de calcul).

3.4.1.2 Gradient Boosting

Le gradient boosting est considéré comme étant l'une des méthode les plus puissantes pour construire des modèles prédictifs.
L'idée du boosting vient de la volonté d'améliorer le pouvoir prédictif d'un classifieur faible ou "weak learner". On appelle weak learner tous les algorithmes dont la performance est au moins aussi bien que celle du hasard.
Le principe général du boosting consiste à construire une famille d'estimateurs qui sont ensuite agrégés par une moyenne pondérée des estimations en régression ou un vote à la majorité en classification.
Cette méthode est généralement plus performante que celle des forêts aléatoires présentées précédemment. Contrairement au bagging dont les classifieurs faibles sont identiquement distribués, le boosting construit les classifieurs de manière récursive. Chaque estimateur fourni est une version adaptative du précédent en donnant plus de poids aux observations mal ajustées ou mal prédites.
Pour faire l'analogie avec les forêts aléatoire, cette section propose une présentation de l'algorithme boosting en régression de Drucker (1997) qui étudient les performances empiriques en relation avec CART.

Soit x_0 à prévoir et z = $(x_1, y_1), ..., (x_n, y_n)$ un échantillon
Initialiser p par la distribution uniforme $p = p_i = 1/n; i = 1, ..., n$
for m = 1 à M do :

1 : Tirer avec remise dans z un échantillon z_m^* suivant p

2 : Estimer $\hat{f_m}$ sur l'échantillon z_m^*

3 : Calculer à partir de l'échantillon initial z :

$$l_m(i) = l(y_i, \hat{f_m}(x_i)) \quad i = 1, ..., n; (l : fonction\, perte)$$

$$\hat{\epsilon_m} = \sum_{i=1}^{n} p_i l_m(i)$$

$$w_i = g(l_m(i))p_i \qquad (g : fonction\ \ continue\, non\, decroissante)$$

4 : Calculer les nouvelles probabilités :$pi = \frac{w_i}{\sum_{i=1}^{n} w_i}$
end for
Calculer $\hat{f}(x_0)$ moyenne ou médiane des prévisions $\hat{f_m}(x_0)$ pondérées par des coefficients $log(\frac{1}{\beta_m})$.
Si on note $L_m = sup_{i=1,..,n} l_m(i)$ le maximum de l'erreur observée par le modèle $\hat{f_m}$ sur l'échantillon initial. La fonction g est définie par :

$$g(l_m(i)) = \beta_m^{1-l_m(i)/L_m}\ avec\ \beta_m = \frac{\hat{\epsilon_m}}{L_m - \hat{\epsilon_m}}$$

3.4.2 Application

Afin de nous assurer de la fiabilité de nos modèles et éviter le sur-apprentissage, nous utilisons la k-fold cross validation afin de sélectionner les paramètres de nos modèles.

On rappelle que cette technique de validation croisée consiste à découper le jeu de données en k parties (folds en anglais) à peu près égales. Tour à tour, chacune des k parties est utilisée comme jeu de test. Le reste (autrement dit, l'union des k-1 autres parties) est utilisé pour l'entraînement.

Pour calibrer une forêt aléatoire sur nos sorties du modèle CAT tout en évitant le sur-apprentissage, il est nécessaire d'ajuster le paramètre clé de cet algorithme. Ce paramètre n'est autre que le nombre de variables sélectionnées aléatoirement comme candidates à chaque split, noté m précédemment.
Pour le gradient boosting, nous ajusterons le paramètre de rétrécissement évitant les corrections excessives *eta* et la profondeur des arbres *max.depth*.

3.4.2.1 Modèle de sévérité

Les figures 29 et 30 représentent les résultats de la calibration des deux modèles à l'issue d'une validation croisée en 5 parties.
En temps normal, arrivée à un certain seuil de complexité du modèle, la racine de l'erreur quadratique moyenne devrait augmenter : c'est le phénomène de sur-apprentissage.
Ceci dit, vu que nous travaillons sur des résultats de modèle, intégrant très peu de bruit,

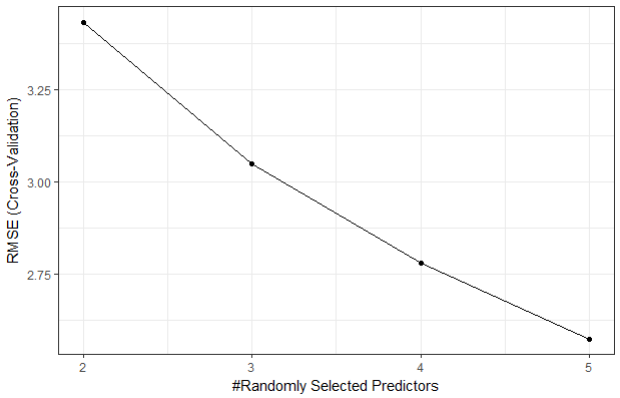


FIGURE 29 – Calibration du random forest

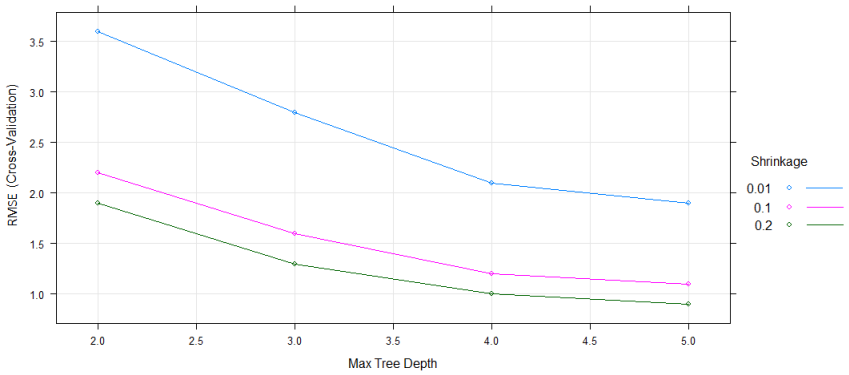


FIGURE 30 – Calibration du gradient boosting

l'allure des courbes parait tout à fait normale.

Une fois nos deux modèles calibrés, nous souhaitons connaître les variables les plus discriminante de chaque modèle.
Contrairement aux GLM, en ce qui concerne les modèles non paramétriques, il ne s'agit pas d'éliminer des variables ou de retenir seulement les variables les plus discriminantes pour obtenir une bonne prédiction. Cette tâche n'est pas nécessaire lorsque l'on utilise des arbres de classification ou régression. En effet par construction, cet algorithme hiérarchise toutes les variables explicatives selon l'importance de leur effet sur la variable réponse. De tels indices d'importance permettent par conséquent d'établir un classement des variables, de la plus importante à la moins importante.

Les figures 35 et 36 mettent en évidence ce classement. Nous remarquons que les variables les plus discriminantes sont similaires pour les deux modèles. A la tête du classement, nous retrouvons toujours la distance à la côte, la ligne de business résidentielle et quelques CRESTA tel que celui du Quintana Roo.
On peut aussi voir que la contribution des variables dans le modèle forêt aléatoire est plus homogène que celle observée pour le gradient boosting.
Ce constat n'est pas surprenant compte tenu du fonctionnement et du paramétrage de chacun des deux algorithmes. Vu que pour le random forest un tirage aléatoire est effectué parmi un nombre restreint de variable à chaque nœud, plusieurs d'entre elles sont mises à contribution.
A contrario, pour le gradient boosting les arbres ne sont pas très profonds et il n'existe aucune contrainte quant à la sélection de variable lors de la construction des arbres.

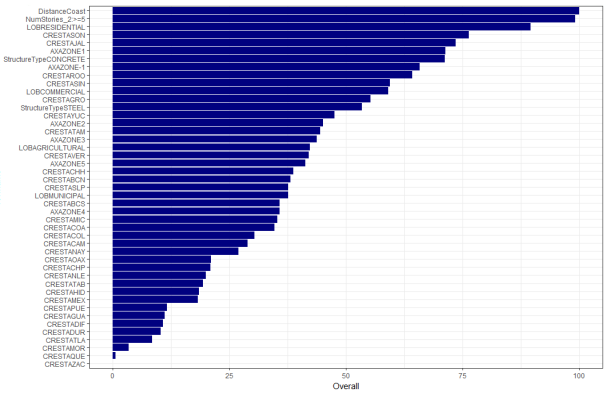


FIGURE 31 – Importance des variables selon Random Forest

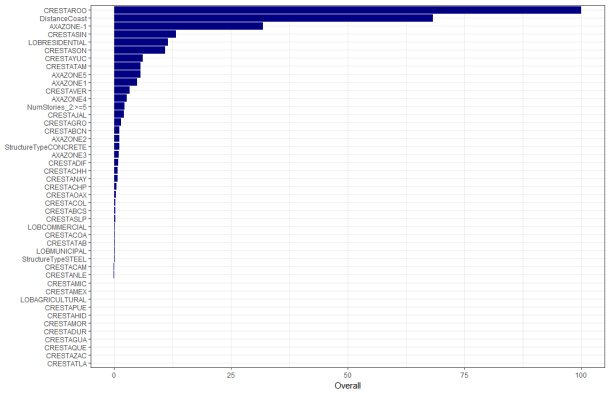


FIGURE 32 – Importance des variables selon gbm

3.4.2.2 Modèle de fréquence

Comme pour les modèles de sévérité, nous utilisons une validation croisée pour calibrer l'algorithme des forêts aléatoires et celui des arbres boostés.

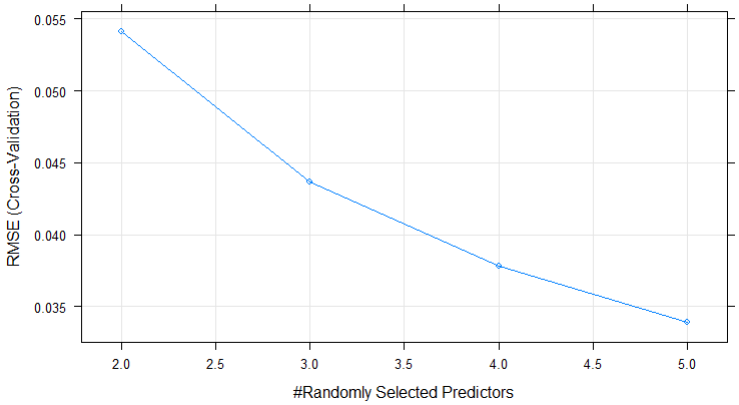


FIGURE 33 – Calibration du random forest

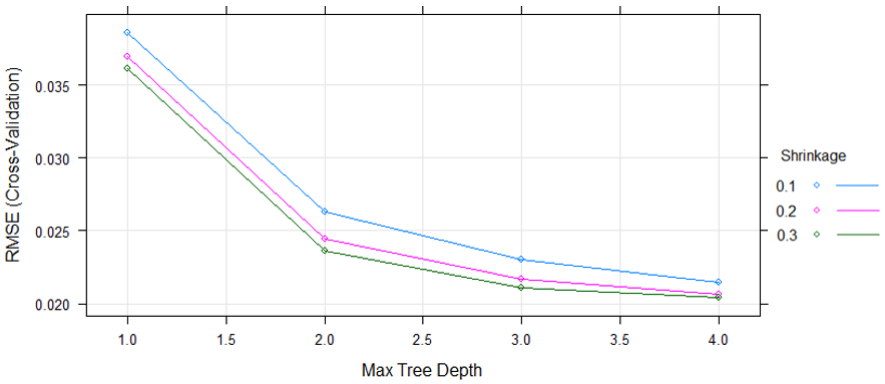


FIGURE 34 – Calibration du gradient boosting

Concernant l'importance des variables, nous observons une fois de plus que le random forest fait contribuer plus de variable. De plus, nous remarquons que la séparation de la prime pure en une composante de fréquence et de sévérité n'est pas que mathématique. Les facteurs explicatifs ne sont pas les même. En effet, nous remarquons que les variables les plus importantes pour

la fréquence sont des variables liées à la situation géographique tandis que la sévérité est aussi expliquée par les caractéristiques physiques du site.

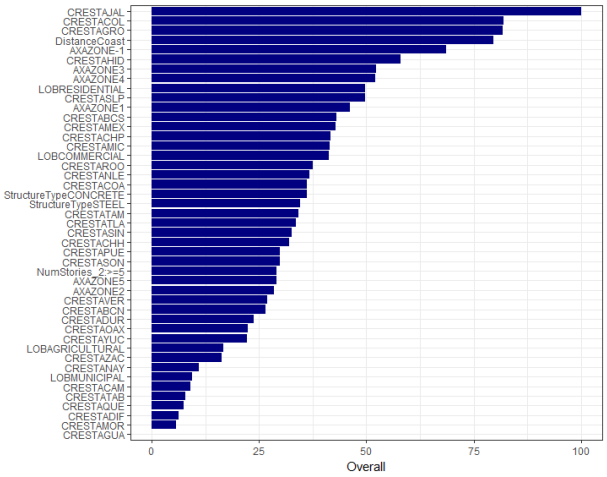


FIGURE 35 – Importance des variables selon Random Forest

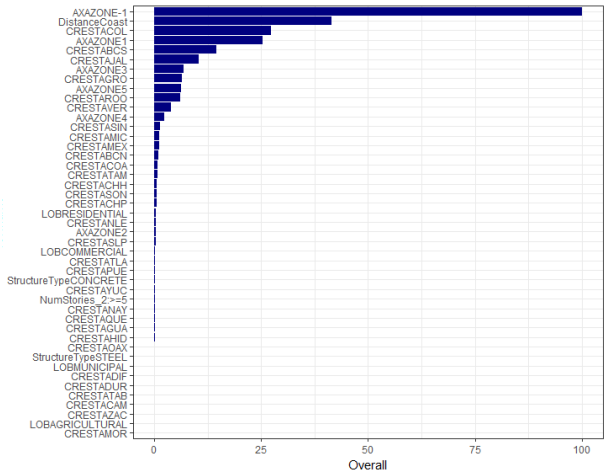


FIGURE 36 – Importance des variables selon gbm

3.5 Analyse des résultats

3.5.1 Modèles de sévérité

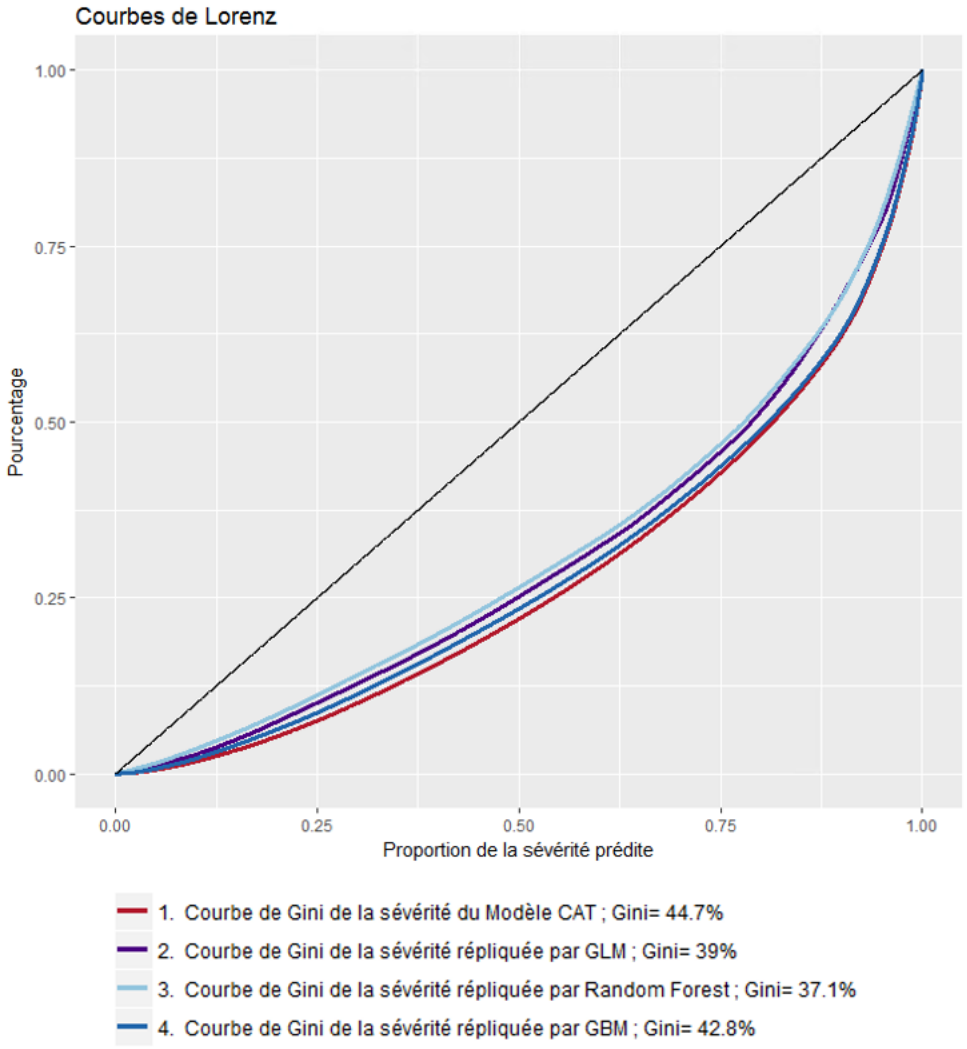


FIGURE 37 – Comparaison des modèles de sévérité en termes d'indice de Gini

La figure 37 représente les courbes de Lorenz des différents modèles de sévérité que nous avons expérimentés. La courbe relative au modèle CAT représente la tarification "parfaite" que l'on souhaite retranscrire et qui assigne à chaque site une prime égale au risque qu'il représente. De plus, la précision de nos modèles est représentée par l'aire qui se trouve entre leur courbes respectives et celle-ci. De ce fait, plus la courbe d'un modèle tend vers la courbe du modèle *CAT* meilleure sera sa performance. D'après la figure, le modèle de sévérité le plus discriminant est celui du GBM suivi du GLM.
Afin d'évaluer la précision des modèles, nous utilisons la base de test pour mesurer les écarts entre les valeurs prédites et les valeurs observées. Concrètement, nous souhaitons évaluer la racine l'erreur quadratique moyenne sur l'échantillon test pour chaque méthode.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\phi}_{Method}(x_i))^2}{n}}$$

avec :

$\hat{\phi}_{Method}(x_i)$: fonction de prédiction associée à la méthode

$i \in [1, ..., n]$ et n la taille de l'échantillon test

Le tableau suivant met en évidence les valeurs prises par cette métrique pour les différentes méthodes :

Méthode	RMSE
GLM Gamma	3.2752
Random Forest	2.4951
GBM	1.1262

En termes de précision, les modèles de sévérité non paramétriques sont de loin les plus performants. Dans notre cas, le pouvoir prédictif de la gamma est entaché par l'instabilité des coefficients estimés en présence de queues épaisses.

3.5.2 Modèles de fréquence

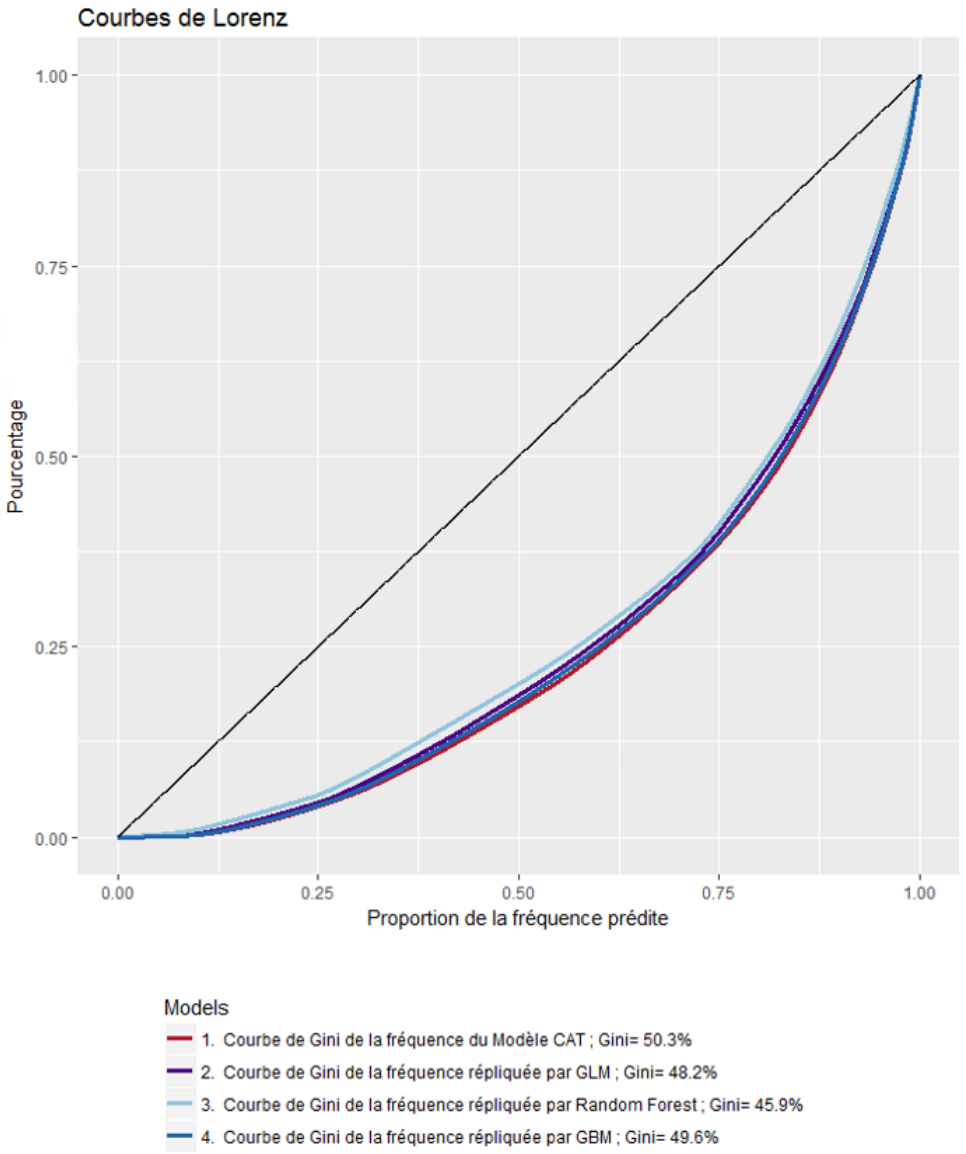


FIGURE 38 – Comparaison des modèles de fréquence en termes d'indice de Gini

De tous les modèles de fréquence le GBM est encore une fois l'algorithme le plus performant. Ce résultat s'explique par le fait que le boosting est supérieur au bagging sur des données non bruitées telles que les nôtres. Cependant de manière générale, les performances des différents modèles se valent.

Méthode	RMSE
GLM Poisson	0.0291
Random Forest	0.0336
GBM	0.0201

3.6 Conditions financières : Courbes de rabais

La tarification actuelle repose sur une estimation d'une prime pure brute que l'on transforme en prime nette en appliquant des courbes de rabais.
Ces courbes ont pour avantage d'être lisible et rendent le passage du prix brut au prix net plus intelligible. Cependant, dans certains cas cette approximation peut s'avérer grossière compte tenu de la volatilité du taux de rabais suivant l'allure de la distribution de pertes sous-jacente.

3.6.1 Définition des courbes de rabais

Les courbes de rabais sont dérivées des courbes d'exposition. Si la courbe d'exposition G est le rapport entre la prime de l'assureur et la prime de base que l'on peut formuler comme suit : G : [0, 1] → [0, 1]

$$G(d) = \frac{E(min(x,d))}{E(min(x,1))}$$

avec :

x : Taux de destruction défini comme $\frac{\text{Montant de sinistre}}{\text{Somme assurée}}$

d : Priorité (ou franchise) d'un traité XS en % définie comme $\frac{\text{Franchise}}{\text{Somme assurée}}$

La courbe de rabais représente la variation de la prime de l'assureur suite à l'introduction d'une franchise et peut être notée de la sorte : H : [0, 1] → [0, 1]

$$H(d) = \frac{E(min(x,1)) - E(min(x,d))}{E(min(x,1))}$$

3.6.2 La construction des courbes de rabais

La première étape consiste à construire la fonction de répartition empirique du taux de destruction.

Taux de destruction x%	Nombre de sinistres avec $\frac{X}{SA_i} \leq x\%$	Sinistres avec $\frac{X}{SA_i} \leq x\% : A_x$	Premiers% des sinistres vérifiant $\frac{X}{SA_i} > x\% : B_x$	Fonction de répartition
0%	0	0	0	0%
...	...	...	...	...
α	$\sum_i \mathbb{1}_{\frac{X_i}{SA_i} \leq \alpha\%}$	$\sum_i X_i \mathbb{1}_{\frac{X_i}{SA_i} \leq \alpha\%}$	$\alpha\% SA_i(\mathbb{1}_{\sum_i \frac{X_i}{SA_i} > \alpha\%})$	$\frac{A_x + B_x}{\sum_i X_i}$
...	...	...	...	...
100%	N	$\sum_i X_i$	0	100%

TABLE 3 – Construction d'une courbe d'exposition empirique

Par conséquent nous pouvons formaliser cette fonction de répartition empirique :

$$F(x\%) = \frac{\sum_i X_i \mathbb{1}_{\frac{X_i}{SA_i} \leq \alpha\%} + x\% * SA_i * (\mathbb{1}_{\sum_i \frac{X_i}{SA_i} > x\%})}{\sum_i X_i}$$

Suite aux travaux de *S.Bernegger* [3] sur la modélisation des courbes d'exposition, nous utilisons le taux de destruction comme une variable aléatoire issue d'une distribution MBBEFD *(Maxwell-Boltzmann Bore-Einstein FermiDirac)* dont la fonction de répartition est de la forme :

$$F(x) = \begin{cases} 1 & \text{si } x = 1 \\ 1 - \frac{(a+1)b^x}{a+b^x} & \text{si } 0 \leq x < 1 \end{cases}$$

avec $a \in \mathbb{R}$ et $b \in]0, +\infty[$
Le but est donc d'ajuster une courbe MBBEFD sur notre courbe empirique, c'est-à-dire qu'il nous faut estimer les coefficients a et b qui reflète au mieux nos données empiriques.

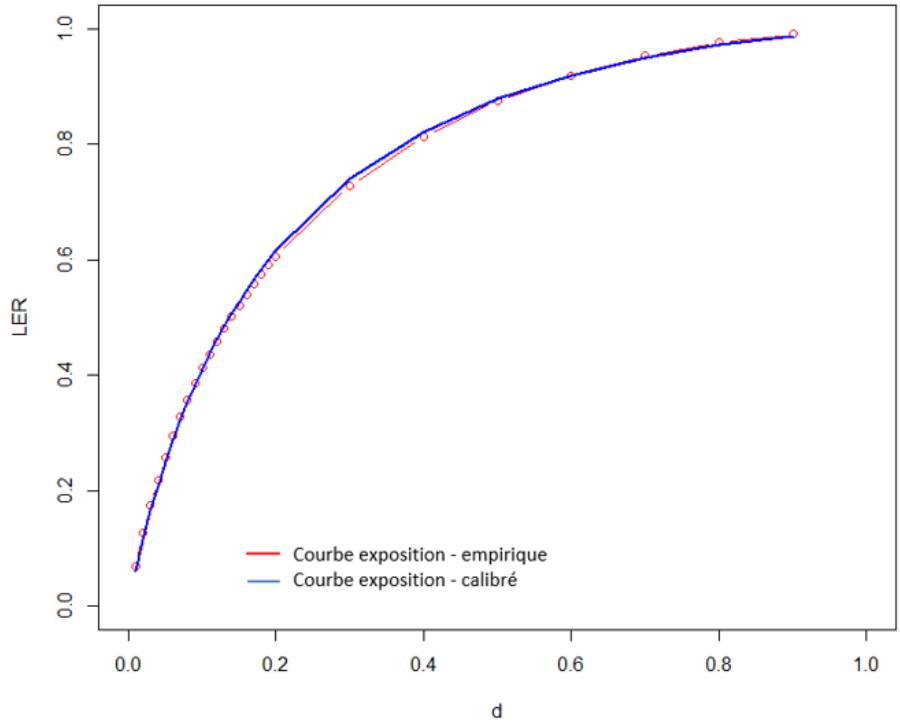


FIGURE 39 – Ajustement par Maximum Vraisemblance d'une MBBEFD

Il y a plusieurs remarque à faire concernant les courbes d'expositions. Le premier point à soulever réside dans le fait que ces courbes fonctionnent pour des traités à portée infinie. En effet, les bornes supérieures ne sont pas modélisées lors de la construction de la fonction de répartition.
Afin de tarifer des tranches d'XS limitées, il suffira de faire la soustraction de deux tranches illimitées.

$$PrimePure_i = E(X_i)(G(l) - G(d))$$

Il est à noter également que les études faites autour de ces courbes ont décelé un lien entre ces dernières et la notion de capital assuré. De ce fait, nous trouvons différentes courbes pour différentes tranches de capitaux assurés.
L'idéal serait de pouvoir trouver une courbe d'exposition spécifique à chaque capital assuré.

Pour la tarification actuelle pour le péril ouragan au Mexique, nous ne distinguons que deux courbes de rabais : une courbe pour la branche de détail (ou particuliers) et une autre pour les autres branches. Les sommes assurées étant moins élevées pour cette branche, elles sont donc plus sensibles à l'introduction d'une franchise.

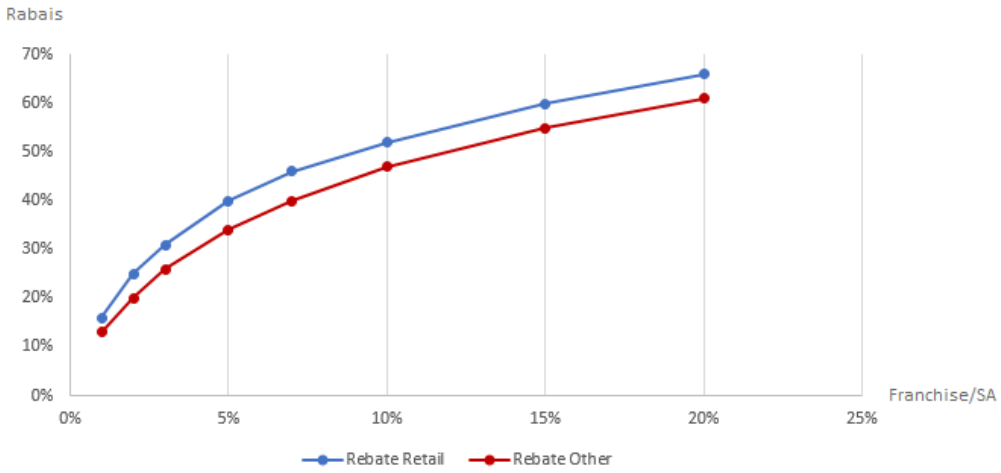


FIGURE 40 – Courbes de rabais Mexico Hurricane

Pour des sites relativement homogènes, le taux de rabais estimé à travers ces deux courbes peut s'avérer être une bonne approximation. Néanmoins pour que les courbes de rabais puissent fournir une meilleure approximation, nous devrions en calibrer d'avantage. Comme mentionné précédemment, l'idéal serait d'envisager des courbes bi-dimensionnelle, prenant en compte non seulement le niveau de franchise mais aussi le montant du capital assurée.
Pour illustrer l'approximation grossière apportée par les courbes de rabais dans certains cas, imaginons que l'on ait un CAT Modèle fictif avec 10 événements, chacun survenant à une année différente et que l'on veuille tarifer un site avec une valeur assurée d'1 million de dollars et une franchise de 10000 dollars, soit 1% de sa valeur assurée.
Selon les courbes de rabais calibrées de la figure 40, 1% de franchise correspond à 15 % de rabais.

EvenementID	Pertes Brutes	Pertes Nettes
1	5000	0
2	0	0
3	0	0
4	32000	22000
5	0	0
6	40000	30000
7	2000	0
8	0	0
9	0	0
10	1000	0

$$\text{La prime brute (AAL)} = \frac{80000}{10} = 8000\$$$

$$\text{La prime nette observée} = \frac{52000}{10} = 5200\$$$

$$\text{La prime nette estimée} = 8000(100\% - 15\%) = 6800\$$$

En faisant l'exercice avec différent niveau de franchise :

Franchise	Courbe de rabais	Rabais observé
1%	15%	18%
2%	25%	35%
5%	45%	100%

Les taux de rabais réels ne sont pas du tout dans le même ordre de grandeur que ceux provenant de la courbe de rabais.

3.7 Comparaison des tarifs modélisées et observées

Les opérations d'assurance ou de réassurance sont caractérisés par une inversion du cycle de production économique. Le prix de revient n'est connu qu'à posteriori : le paiement de l'indemnité par l'assureur est conditionnel à la survenance du risque. L'un des enjeux de l'assureur consiste à construire un tarif lui permettant au moins d'honorer ses engagements et payer l'ensemble des sinistres survenus.
La spécificité de notre tarification réside dans le fait qu'elle se base sur des données stochastiques issues d'un modèle CAT et non de données historiques. Le mode d'évaluation du tarif élaboré dans cette section reposera donc sur une comparaison des primes modélisées aux pertes annuelles agrégées AAL (en anglais *Annual aggregate loss*).

Le tarif est jugé cohérent si la somme des primes récoltées sur tout le portefeuille est au moins égale à la somme des moyenne des pertes annuelles agrégées des sites qui le constituent. Le tableau ci-dessous reprend l'indice des primes estimées sur le portefeuille (en base 100) en fonction des algorithmes utilisés et avec les résultats du modèle de *CAT* comme référence.

Primes (M $)	Pertes espérées	Prime GLM	Prime Random Forest	Prime Gradient Boosting
Brut	100	99.25	99.40	99.87
Net	100	128.01	127.28	128.32

Toutes les primes modélisées sont du même ordre de grandeur en brut. Malgré le fait que les modèles non paramétriques soient plus conservateurs que les modèles GLM, les primes modélisés ne couvrent pas totalement les pertes observées.
En Net, les primes obtenues après application des courbes de rabais sont presque deux fois plus importantes que celles sorties par le modèle CAT.
En effet et comme nous l'avions expliqué précédemment, les courbes de rabais dépendent du comportement et de la forme de la distribution de pertes sous-jacentes. Deux sites ayant la même prime brute et le même niveau de franchises peuvent en réalité avoir un taux de rabais très différent (du double au quadruple en moyenne).

Les approches proposées dans les sections suivantes proposeront des techniques permettant d'estimer la distribution de pertes au niveau le plus granulaire, celui du site. L'obtention de cette distribution nous permettra non seulement de tarifer de façon précise toutes les structures de réassurance relativement complexes mais aussi d'avoir une vision parfaite et dynamique de l'exposition du portefeuille.

4 Nouvelle approche statistique de tarification

La tarification actuelle repose sur une estimation d'une prime pure brute puis par l'application de courbes de rabais pour le calcul d'une prime nette.
La section précédente présente une multitude de méthodes élaborées pour l'estimation de cette prime. Elle traite également du biais que peut entraîner l'usage d'une (ou un nombre limité) de courbes de rabais par modèle.
Nous avons conclu que l'approximation apportée par ces courbes peut être assez grossière compte tenu de la volatilité du taux de rabais suivant la forme de la distribution de pertes sous-jacente.
Nous proposons donc dans les sections qui suivent de ne plus modéliser un unique paramètre de la distribution de pertes (actuellement la moyenne), mais un vecteur de paramètres permettant de caractériser la totalité de la distribution de pertes.
Une bonne estimation de cette distribution nous permettra non seulement de modéliser de manière plus fine les effets non-linéaires à l'origine du taux de rabais mais aussi d'avoir une vision "parfaite" du risque et de son accumulation.

Pour ce faire, nous proposons une méthode qui peut être décrite en deux grandes étapes :

- **La première étape consiste à calibrer les loi(s) statistiques** ajustant le mieux la distribution des pertes observées par événements et par sites.
- **La deuxième étape consiste à prédire les paramètres de cette (ou de ces) loi(s)** à l'aide de modèles de prédiction classiques que sont les GLM.

Il découlera de chacune des étapes précédentes un type d'erreur :

- **Une erreur d'ajustement** : La distribution de perte au niveau site est souvent irrégulière. Supposer que ces pertes suivent une loi particulière entraîne une faible qualité d'ajustement.
- **Une erreur de prédiction** : En supposant que la loi statistique régissant les pertes du site ainsi que ses paramètres soient connus, quelle est l'erreur dans l'estimation de ces paramètres lorsque nous les prédisons avec un modèle type GLM.

Nous proposerons dans un premier temps d'estimer cette distribution en calibrant une loi de sévérité sur la distribution de pertes conditionnelles et une loi de fréquence sur le nombre de sinistres annuels.
Nous discuterons ensuite des différentes difficultés rencontrées notamment sur la complexité d'ajuster une loi statistique paramétrique pour des distributions très fortement asymétriques.
Cette discussion s'appuiera sur différentes métriques et tests permettant d'apprécier la qualité d'ajustement des lois choisies (AIC, statistique KS, Anderson etc..) et leur capacité à répliquer les pertes extrêmes individuelles.
Dans la seconde partie, nous nous intéresserons à la deuxième étape consistant à prédire à partir de variables explicatives présentes dans le portefeuille (ou intégrées dans celui-ci) les paramètres clés caractérisant les lois statistiques calibrées à la première étape.
Nous présenterons alors les différents « inputs » requis pour permettre cette calibration et les différentes questions qui lui sont reliées.
Enfin, nous illustrerons cette approche par un exemple en considérant un site en particulier puis en l'appliquant sur l'ensemble du portefeuille. Par une méthode de Monte-Carlo, nous réaliserons alors des simulations afin de vérifier que l'identification des paramètres sur les pertes simulées redonne bien des valeurs proches des valeurs réelles . Cette étape nous permettra de juger la robustesse de nos modèles et la crédibilité de cette approche.

4.1 Calibration sur les données de pertes brutes

4.1.1 Présentation des données et du contexte

Le modèle CAT considéré simule plus de 60.000 événements sur 10.000 années. Par agrégation des pertes des événements de la même année, nous pouvons récupérer pour chacun des sites du portefeuille la perte totale pour chaque année simulée du catalogue stochastique. Ainsi, à chaque site du portefeuille est associé un vecteur de pertes annuelles de taille 10.000.
Dans la suite et de façon similaire à la méthode de tarification actuelle, nous normalisons les pertes par la somme assurée pour travailler avec des taux de destruction.
Tandis que dans la méthode actuelle de tarification, nous agrégeons cette distribution en calculant la moyenne des pertes, ici nous conserverons l'ensemble de la distribution.

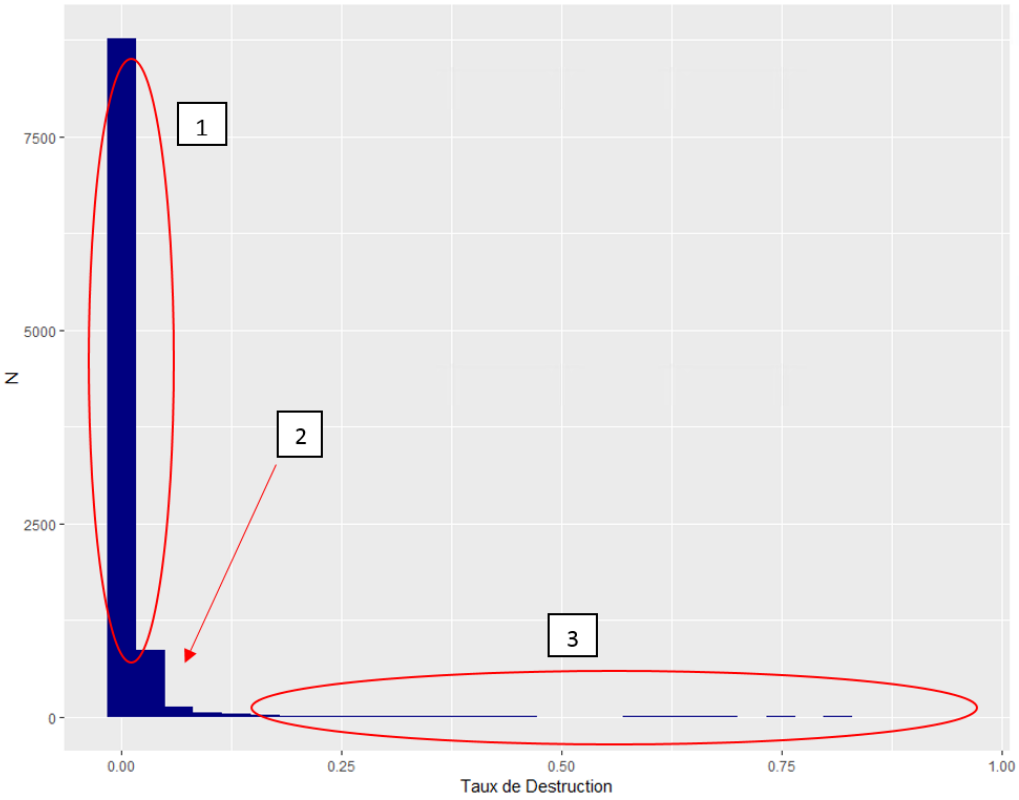


FIGURE 41 – Distribution annuelle de pertes en taux de destruction pour un site donné

La figure 41 représente cette distribution pour un site donné. Elle se compose généralement de trois parties :

— Une masse de dirac en 0 reflétant la probabilité de non-génération de sinistre qui est plus ou moins élevée en fonction de la nature du site, de sa localisation géographique et des caractéristiques internes de celui-ci.

— Une partie avec des taux de destruction dits attritionnels le plus souvent proches de 0 et issus d'années où peu d'événements forts ont été simulés.

— Une partie avec de forts taux de destruction générés par quelques événements du catalogue stochastique. Ces taux reflètent une destruction quasi-totale voire totale du site concerné. La probabilité de survenance de tels événements est faible, Cependant, ce sont ces valeurs-là qui nous intéressent pour répliquer correctement le comportement de la distribution empirique puisque ce sont celles-ci qui vont contribuer grandement au calcul du taux de rabais.

Afin de pouvoir appliquer correctement les conditions financières pour un site donné, il est nécessaire de de disposer la distribution des pertes par événement et non pas annuelles de ce site (ces deux distributions peuvent être égales pour un site qui sera touché par exactement un événement par année simulée).
Pour cela, nous faisons le choix de travailler dans la suite sur la distribution conditionnelle de sévérité de ce site et la distribution de fréquence annuelle.
Par conséquent, chaque site ne sera donc plus représenté par son vecteur de pertes annuelles (pour rappel de taille 10.000) mais par deux distributions :

- **La distribution de pertes par événement de taille variable égale aux nombres événements occasionnant des pertes sur l'ensemble du catalogue stochastique**
- **La distribution du nombre de sinistres annuel (de taille 10.000) avec une masse en Dirac en 0.**

4.1.2 Choix de la loi statistique et calibration des paramètres

Pour juger de l'adéquation de la distribution des valeurs d'un caractère à une loi de probabilité $\mathcal{L}$, nous distinguons deux approches complémentaires :

Analyses Graphiques

- *Méthode de l'histogramme* : Nous traçons l'histogramme des fréquences correspondantes aux données, pour ensuite superposer les valeurs de la "densité" associée à la loi $\mathcal{L}$ en estimant éventuellement les paramètres inconnus de celle-ci.
- *Méthode de la fonction de répartition* : Nous traçons le graphe de la fonction de répartition empirique définie par :

  $$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \leq x}$$

  Pour ensuite superposer le graphe de la fonction de répartition associée à la loi $\mathcal{L}$ en estimant éventuellement les paramètres inconnus de celle-ci.
- *Méthode de l'approximation de la densité* : Nous utilisons une estimation de la densité inconnue. Puis nous superposons les valeurs de la densité associée à la loi $\mathcal{L}$ en estimant éventuellement les paramètres inconnus de celle-ci.
- *Méthode du QQ plot (quantile-quantile plot)* : Cette méthode consiste à comparer les quantiles empiriques et les quantiles théoriques. Soit F(x) la fonction de répartition de X et x_p le quantile d'ordre p tel que $x_p = inf\{x \in \mathbb{R}; F(x) \geq p\}$.
  En notant $x_{(1)}, x_{(2)}, ...x_{(n)}$ les statistiques d'ordres de X, nous pouvons écrire la fonction de répartition empirique de la sorte :

  $$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \leq x} = \begin{cases} 0 & \text{si } x < x_i \\ \frac{k}{n} & \text{si } x_{(k)} \leq x \leq x_{(k+1)}, \quad k \in \{1, ..., n-1\} \\ 1 & \text{si } x \geq x_{(n)} \end{cases}$$

  Soit $(p_1, p_2, ..., p_n)$ une suite strictement croissante de n réels vérifiant $p_k \in]\frac{k-1}{n}, \frac{k}{n}[$, $k \in \{1, ..., n\}$, de sorte que $x_{(k)} = inf\{x \in \mathbb{R}; F_n(x) \geq p_k\}$.
  Le QQ-plot désigne alors le nuage de points $\mathcal{N}$ dans le repère orthonormé (O, I, J) défini par :

  $$\mathcal{N} = \{(x_{p_1}, x_{(1)}), (x_{p_2}, x_{(2)}), ..., (x_{p_n}, x_{(n)})\}$$

Si X suit la loi $\mathcal{L}$, les données font que F_n est une bonne estimation de F et, a fortiori, $x_{(k)}$ devient une bonne estimation de x_{p_k} pour tout $k \in \{1, ..., n\}$. Les points du nuage doivent donc être proche de la droite diagonale d'équation : y = x.

Le critère d'information d'Akaike AIC

Le critère d'information d'Akaike permet de pénaliser les modèles en fonction du nombre de paramètres à estimer. Ce critère satisfait le principe de parcimonie : les hypothèses suffisantes les plus simples doivent être préférées. Ce dernier concorde parfaitement avec la logique de l'approche que l'on présente. En effet, minimiser le nombre de paramètres nous permettra de réduire le nombre de modèles statistiques à calibrer à la deuxième étape et donc de minimiser l'erreur de prédiction. En revanche il est important de garder en tête que ce test ne dit rien de la qualité absolue du modèle. Il ne rendrait ainsi pas compte du fait que tous les modèles candidats ne produisent pas de bons ajustements.

$$AIC = 2k - ln(L)$$

avec k le nombre de paramètres à estimer du modèle et L le maximum de la fonction de vraisemblance du modèle

Tests statistiques d'adéquation à une loi

— *Test du Chi-deux* :

Ce test consiste à comparer l'histogramme des fréquences et la distribution de la loi de probabilité servant de modèle théorique. Pour ce faire, nous découpons au préalable l'intervalle d'observation en k classes, pour construire un indice mesurant l'écart constaté entre les effectifs observés et les effectifs théoriques espérés sous H_0.

$$\chi^2_n = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}$$

avec :

— n_i : effectif observé dans la classe i

— n : effectif total observé

— np_{i0} : effectif théorique attendu sous H_0 dans la classe i

Soit $K \to \chi^2_{(k-r-1)}$sous H_0 où r le nombre de paramètres qui ont été estimés pour que la loi théorique soit entièrement déterminée. Alors la p-valeur associée au test du Chi-deux d'adéquation à une loi est : $p - valeur = \mathbb{P}(K \geq \chi^2_{obs})$.
En pratique, ce test est surtout utilisé pour les variables discrètes.

— *Test de Kolmogorov-Smirnov* :

Soit $X_1, ..., X_n$ un échantillon de loi F_X et soit F_0 une loi continue. Nous souhaitons tester :

$$\begin{cases} H_0 : & F_X = F_0 \\ H_1 : & F_X \neq F_0 \end{cases}$$

Le test de Kolmogorov-Smirnov consiste à mesurer la plus grande distance entre la distribution empirique F_n et la distribution théorique F_0. Plus cette distance est importante plus le rejet de H_0 est significatif (idée similaire à l'analyse graphique de la méthode de

la fonction de répartition).
La statistique de Kolmogorov-Smirnov est alors définie par :

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

Le test au seuil α associé à cette statistique est défini par la région critique suivante :

$$\{D_n \geq c_{n,\alpha}\}$$

où $c_{n,\alpha}$ est le quantile (1-α) de la table Kolmogorov-Smirnov.
En pratique, ce test est utilisé pour les variables à densité.

— *Test d'Anderson Darling* :
Le test d'Anderson Darling se base également sur un calcul de distance entre la fonction de répartition empirique et la fonction de répartition théorique. La statistique d'Anderson -Darling est définie comme suit :

$$A_n^2 = n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)$$

Le test au seuil α associé à cette statistique est défini par la région critique suivante :

$$\{A_n^2 \geq c_\alpha\}$$

où c_α est le quantile (1-α) de la table d'Anderson Darling.

Cependant comme le font remarquer *Marie Laure Delignette Muller, Christophe Dutang, auteurs du package R fitdistrplus* [5], il n'est pas raisonnable de rejeter une distribution juste parce qu'un test d'ajustement la rejette, surtout en présence d'un nombre important d'observations. En réalité, les tests d'ajustements ont tendance a rejeté automatiquement les distributions quand l'échantillon est suffisamment grand.

Le fait de supposer que les observations décrivant chaque site sont tirées de distributions classique dépendant d'un nombre restreint de paramètres constitue une hypothèse forte. Ces modèles ne peuvent donc pas décrire de façon parfaite les distributions de pertes dont nous disposons.
Néanmoins, notre objectif consiste à trouver des distributions paramétriques choisies dans des familles prédéfinies de distributions qui décrivent au mieux les distributions empiriques de fréquence et de sévérité conditionnelle.
Par conséquent, le choix de la distribution la plus adaptée reposera d'avantage sur l'examen visuel des ajustements. Nous pourrons aussi nous appuyer sur les statistiques de tests présentées pour comparer les différents modèles. Ces dernières mesurent la distance entre la distribution théorique et la distribution empirique. De ce fait, le choix du meilleur modèle revient à choisir le modèle avec la statistique de test la plus faible.
La statistique de Kolmogorov-Smirnov est plutôt sensible lorsqu'il existe un décalage à tendance centrale entre les distributions contrairement à la statistique d'Anderson Darling qui accorde plus d'importance aux queues de distribution. Dans notre cas, nous préférerons donc cette dernière en cas de conclusions divergentes entre les deux tests.

4.1.2.1 Loi de fréquence

L'échantillon du nombre de sinistres annuels par site est composé de 10000 observations. Il présente un nombre de sinistres annuels généralement faible avec une masse importante en 0.

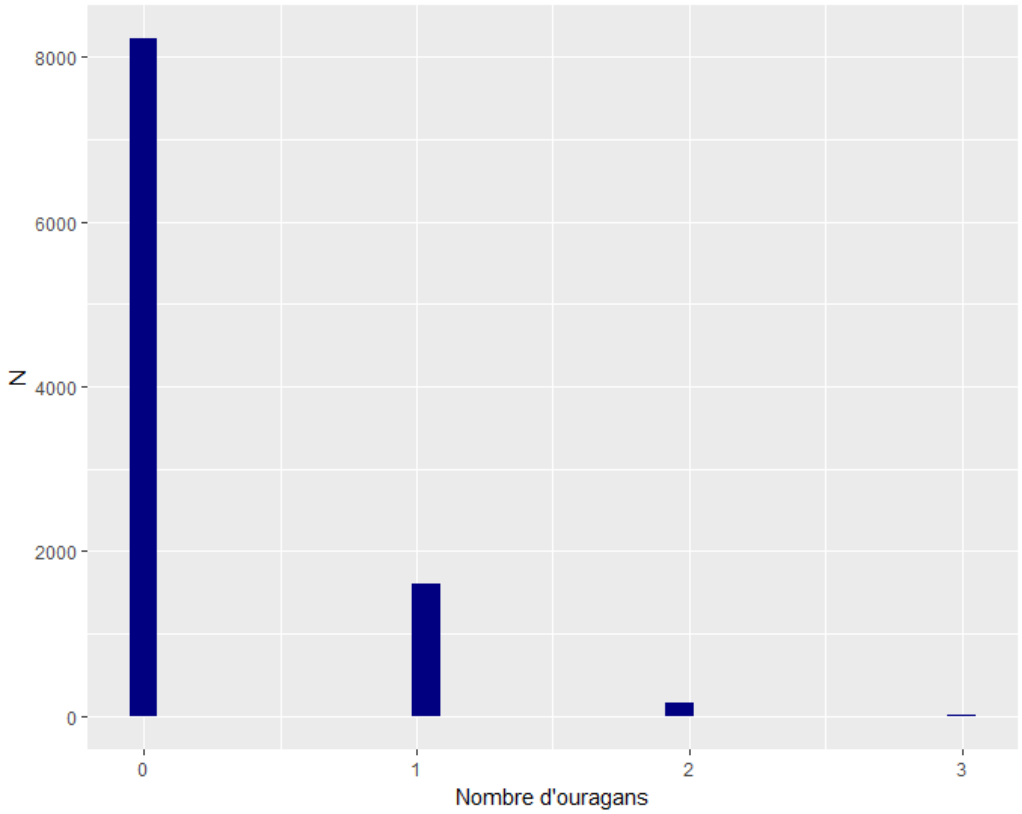


FIGURE 42 – Distribution annuelle de pertes en taux de destruction pour un site donné

L'analyse de ce résultat permet de mettre en lumière la nécessité d'ajuster le nombre de sinistre à des lois non seulement modélisant les événements rares, comme la loi de Poisson, mais aussi représentant la sur-dispersion des observations liée à la masse en 0. Pour cette dernière caractéristique, nous présentons et comparons à la loi de Poisson, la Binomiale Négative.

	Poisson	Négative binomiale
Densité $f(x)$	$\frac{\lambda^k e^{-\lambda}}{k!}$	$\frac{\Gamma(k+v)p^v(1-p)^k}{\Gamma(k+1)\Gamma(v)}$ avec : $p = \frac{v}{\lambda+v}$
Espérance $E(x)$	λ	$\frac{(1-p)v}{p}$
Variance $V(x)$	λ	$\lambda + \frac{\lambda^2}{p}$

TABLE 4 – Lois de fréquence

Après avoir lancer un programme calibrant la fréquence sur les lois de poisson et binomiale négative pour tout les sites du portefeuille, nous faisons l'extraction des statistiques du chi-

deux et l'aic pour comparer l'ajustement des deux lois.
Les tests de calibration montrent qu'une loi de Poisson permet d'ajuster de façon correcte la grande majorité des distributions du nombre de sinistre annuel. En effet, nous remarquons qu'en terme d'AIC les performances des lois de poisson et binomiale négative se valent et qu'en terme de statistique du chi-deux la négative binomiale améliore légèrement l'ajustement. Dans la suite, nous décidons de sélectionner la loi de poisson pour sa simplicité et le fait qu'elle ne soit caractérisée par un unique paramètre λ. Ce choix nous permettra de nous limiter à un seul GLM pour la fréquence et de réduire ainsi les sources d'erreur de prédiction.

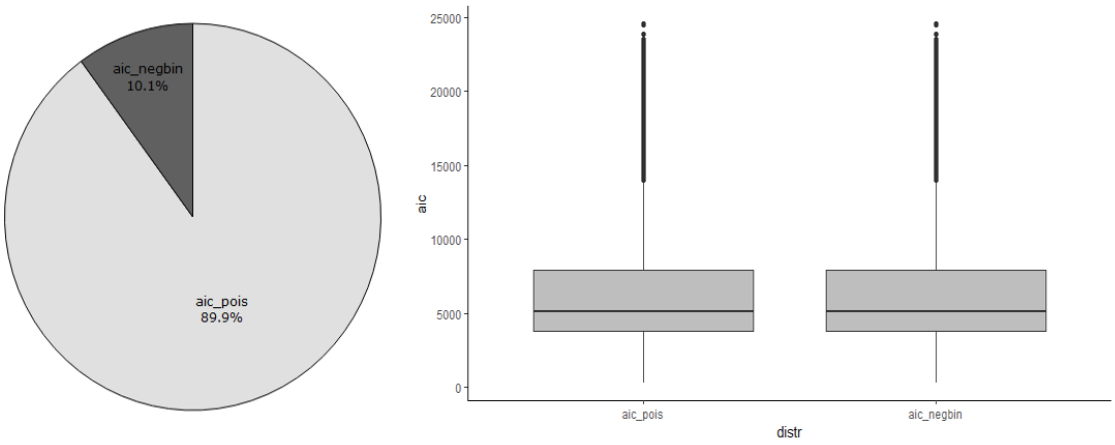


FIGURE 43 – Ajustement sur la fréquence de sites selon le critère de l'AIC

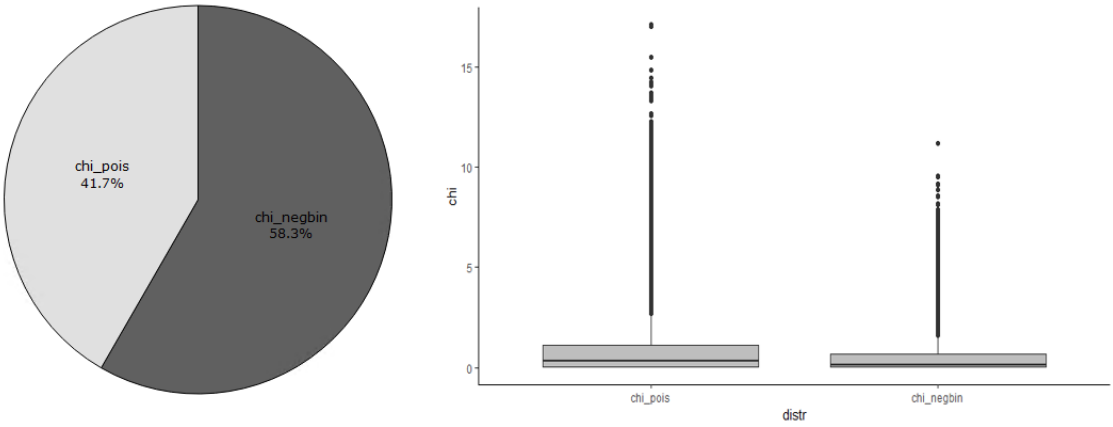


FIGURE 44 – Ajustement sur les fréquences des sites selon le critère de la statistique du chi-deux

4.1.2.2 Loi de sévérité

L'évaluation du risque diffère selon si cette évaluation est faite au niveau individuel - c'est à dire au niveau site - ou au niveau portefeuille en agrégeant plusieurs risques individuels.
Lorsque l'on évalue les pertes d'un portefeuille, il y a une sorte de nivellement entre les différents sites assurés à une zone donnée qui produit une distribution de pertes assez régulière. Ce n'est pas le cas au niveau site, où la distribution des pertes est assez prononcée.
Un portefeuille ne subira jamais de taux de destruction à 100% suite à un événement CAT ce qui est parfaitement envisageable au niveau site. Par conséquent, une seule loi permet rarement d'ajuster correctement la sinistralité, aussi bien sur la partie centrale de la distribution qu'en queue de distribution.
Les premières tentatives de calibration de loi de sévérité sur les distributions hautement

volatiles se révèlent décevantes car l'asymétrie de ces distributions nous empêche de calibrer correctement une loi sur les extrêmes.

La première idée consiste à faire un changement d'échelle et à travailler avec les logarithmes des taux de destruction. Cela permet de « contracter » la distribution des taux de destruction et d'améliorer de façon très significative l'ajustement sur les quantiles extrêmes.
Nous ne souhaitons modéliser que des taux de destruction appartenant à l'intervalle]0,1] (0 étant ouvert car nous ne travaillons que sur la distribution de sévérité conditionnelle) et le passage en log impose d'utiliser des lois à support sur $]-\infty, 0[$ (ou par symétrie sur $]0, +\infty[$) .
Ce changement d'échelle nous permet d'élargir la sélection des distributions candidates pour calibrer la sévérité.

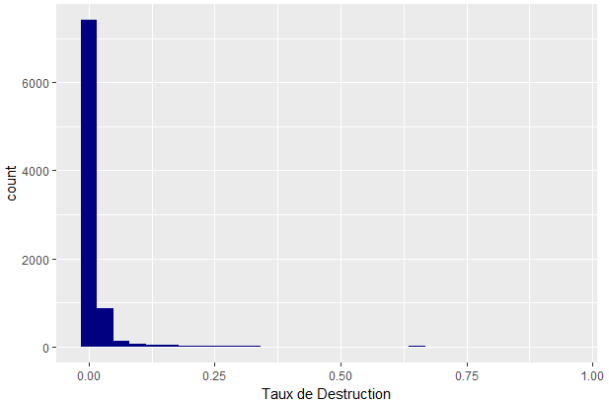


FIGURE 45 – Taux de destruction d'un site donné

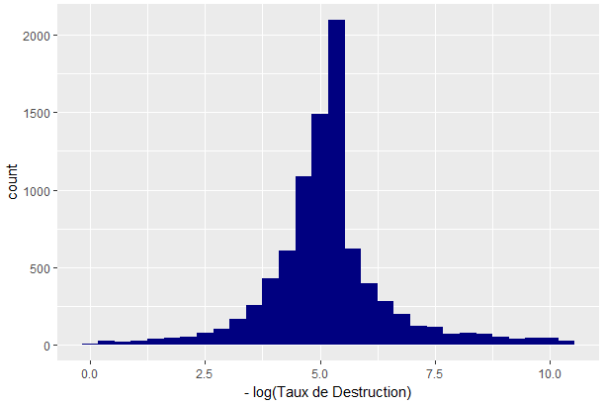


FIGURE 46 – Taux de destruction transformé d'un site donné

La figure 33 montre alors que la forme de la distribution est beaucoup plus classique et devrait permettre un meilleur ajustement sur les quantiles extrêmes.
Dans de premiers tests, nous ajustons la sévérité sur plusieurs distributions à support sur $\mathbb{R}+*$ tel que : Weibull, Gamma, log-normale, Pareto.
Ce pendant, nous ne présenterons que la loi de weibull, lognormal et gamma :

	Weibull	Lognormal	Gamma
Densité $f(x)$	$\begin{cases} \beta\alpha x^{\alpha-1}e^{-\beta x^{\alpha}} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$	$\begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} exp(-\frac{(ln(x)-\mu)^2}{2\sigma^2}) & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$	$\begin{cases} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1}e^{-\beta x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$
Espérance $E(x)$	$\frac{\Gamma(1+\frac{1}{\alpha})}{\beta^{\frac{1}{\alpha}}}$	$e^{\mu+\frac{\sigma^2}{2}}$	$\frac{\alpha}{\beta}$

TABLE 5 – Lois de sévérité à support sur $\mathbb{R}+$

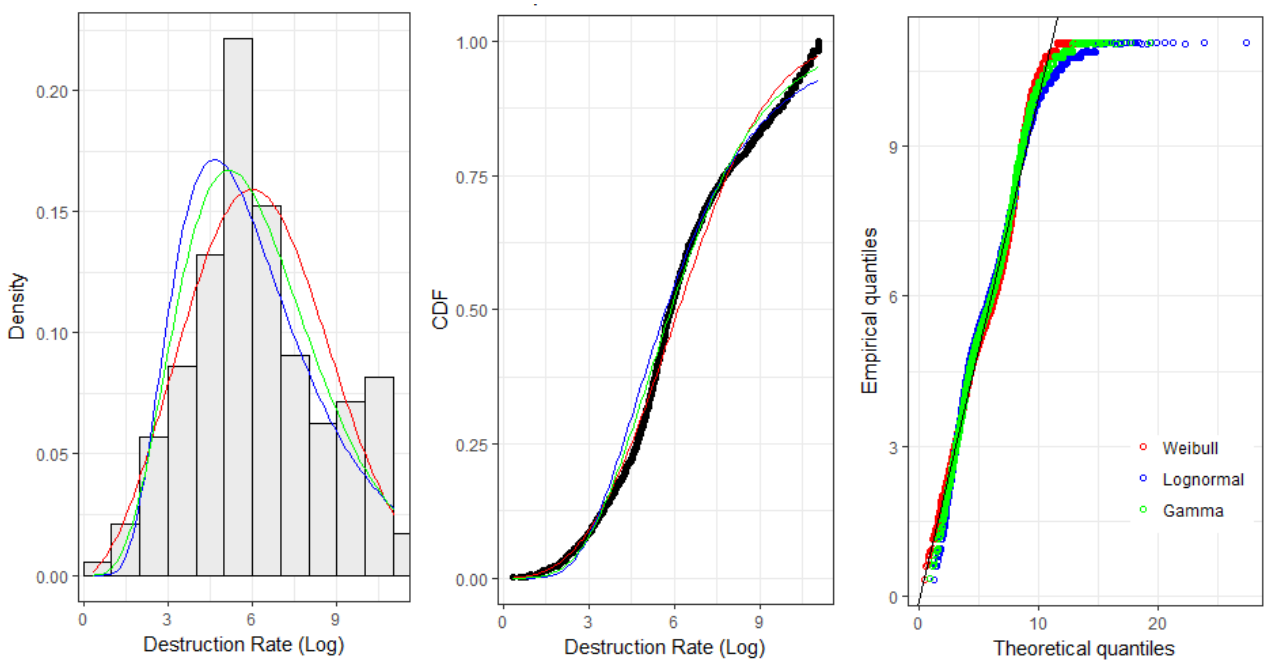


FIGURE 47 – Ajustement du taux de destruction transformé pour un site donné

D'autres tests consistent à n'effectuer aucun changement d'échelle et d'utiliser des lois à support sur [0,1] ou sur]0,1[tels que la bêta ou encore la loi mbbefd dont nous rappelons les caractéristiques dans le tableau suivant :

	Bêta 1-inflatée	Mbbefd
Fonction de répartition $F(x)$	Caractérisée par : $\frac{\beta(x;a,b)}{\beta(a,b)}$ avec : $\beta(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}$	Selon la première paramétrisation : $\begin{cases} (1-\frac{(a+1)b^x}{(a+b^x)})\mathbb{1}_{[0,1[}(x) + \mathbb{1}_{[1,+\infty[}(x) & \text{si } a(1-b)>0 \\ (1-b^x)\mathbb{1}_{[0,1[}(x) + \mathbb{1}_{[1,+\infty[}(x) & \text{si } a=+\infty, b<1 \\ \mathbb{1}_{[1,+\infty[}(x) & \text{si a=0 ou b=1} \end{cases}$ et $P(X=1) = \frac{(a+1)b}{a+b} = p_{a,b}$
Espérance $E(x)$	$\frac{a}{a+b}$	$\frac{ln(\frac{a+b}{a+1})}{ln(b)}(a+1)$

TABLE 6 – Lois de sévérité à support sur $(0,1)$

La calibration avec ces lois risque de fournir un ajustement assez médiocre. La loi mbbefd, par exemple considère un taux de destruction pouvant être de 1 avec une probabilité non nulle. Par conséquent, comme l'illustre la figure 48, dans le cas où un site ne subit aucune destruction totale sur l'ensemble des événements constituant le catalogue stochastique, cela ne marche pas très bien.

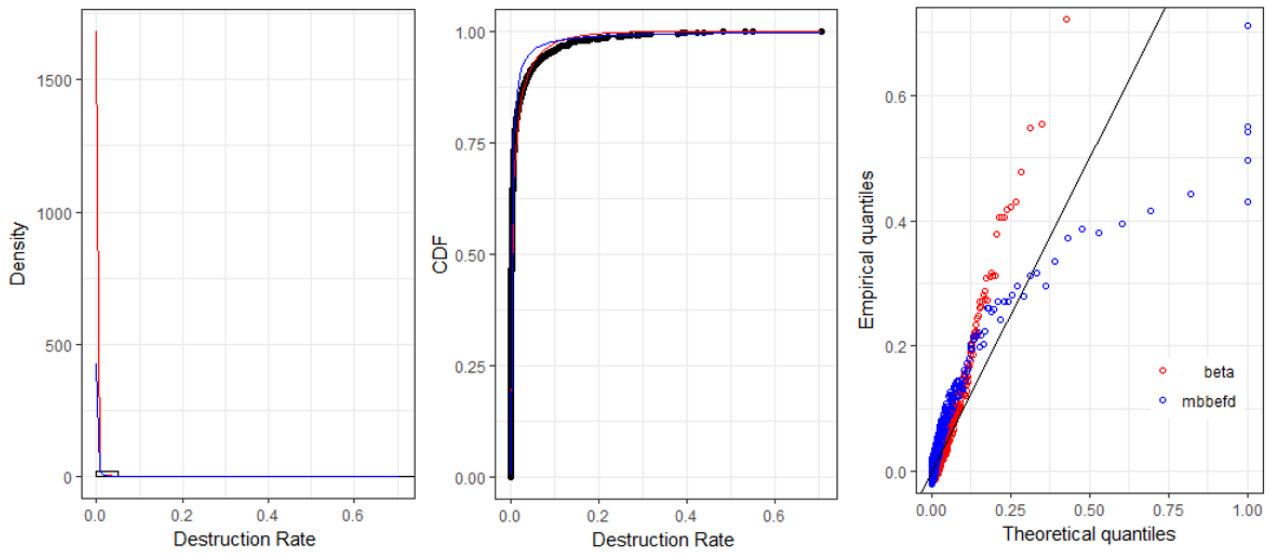


FIGURE 48 – Ajustement du taux de destruction pour un site donné

Nous souhaitons quand même comparer les deux approches : la sévérité en taux de destruction brut sans changement d'échelle et avec changement d'échelle.
Dans un premier temps, nous focalisons l'analyse sur l'ajustement avec changement d'échelle. Nous avons calibré les distributions de pertes de chaque site du portefeuille sur les lois de gamma, lognormal et weibull. Pour chaque site nous comparons les statistiques de Kolmogorov-Smirnov, d'Anderson-darling et AIC relatives aux différentes lois. Nous retenons alors les lois qui minimisent chacune des statistiques.
Les résultats peuvent être résumés dans le tableau de contisuivant :

	Gamma	Log-normal	Weibull
Kolmogorov-Smirnov	84499	79183	112812
Anderson-darling	32689	162338	81467
AIC	32618	148167	95709

TABLE 7 – Analyse des résultats de la calibration sur des lois à support sur $]0, +\infty[$

Nous remarquons que globalement la loi log-normale fournit un bon ajustement pour la majorité des sites, suivi de la weibull et la loi de gamma dont les queues de distributions ne sont visiblement pas assez lourdes pour calibrer les valeurs extrêmes.
Néanmoins, ce résultat nous pousse alors à nous interroger sur les spécificités des sites qui pourrait expliquer l'adéquation de leur sévérité avec une distribution plutôt qu'une autre.
Les figures 49 et 50 révèlent les lois sur représentées dans chacune des crestas mexicaines selon les critères d'Anderson-Darling et de l'AIC. La première remarque que nous pouvons faire concerne la loi de gamma qui ne concerne que très peu de crestas non risquées et qui ne fera désormais plus des distributions candidates.
Nous constatons également une cohérence entre les deux statistiques puisque ces dernières s'accordent sur les lois régissant les pertes de la majorité des crestas.
De plus, en analysant la figure 49, nous remarquons que malgré le fait que les caractéristiques de l'exposition impactant la vulnérabilité diffèrent d'une cresta à une autre, les zones les plus exposées sont mieux ajustées par des weibull et les moins risquées par des log-normales.
Ce constat est en accord avec la théorie, vu que la weibull a une queue plus lourde que celle de la loi Log-normale.

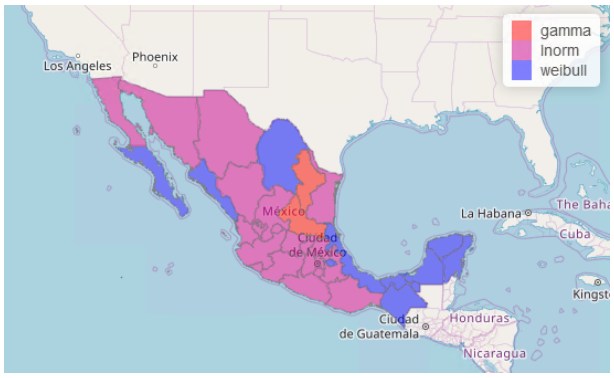


FIGURE 49 – Distribution géographique des lois selon la statistique d'Anderson Darling

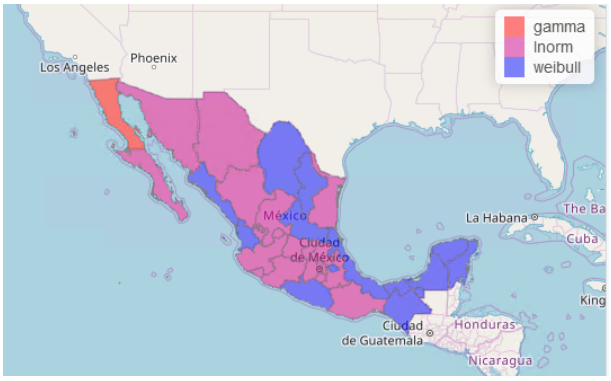


FIGURE 50 – Distribution géographique des lois selon l'AIC

Nous serions tenté alors de construire un modèle de classification dont la variable à prédire serait binaire représentant la weibull et la log-normale.
Nous aurions pu également améliorer l'ajustement en considérant une « loi mélange ». La première étape aurait consisté à analyser la distribution et définir un point qui séparerait le « cœur » de la distribution de sa queue.
Dans une deuxième étape, nous aurions estimé une loi sur chacune des deux parties pour aboutir à une densité ayant la forme suivante :

$$f(y) = pf_1(y) + (1-p)f_2(y), \forall y \in \mathbb{R}_+$$

avec p la part des observations de l'une des parties. [4]
Cependant cette dernière méthode engendre une autre problématique, celle de la détermination du point d'entrée dans la région extrême pour chaque site. Ce seuil doit être suffisamment grand pour pouvoir appliquer les résultats de convergence en loi mais pas trop grand afin d'avoir suffisamment de données pour obtenir des estimateurs de bonne qualité. Cette méthode est intéressante et pourrait être le sujet d'un autre mémoire mais ne sera pas traité.
La méthode consistant à prédire le type de lois à ajuster, elle n'est intéressante que si les qualités d'ajustements sont significativement différentes avec les deux distributions.
Au vu des résultats fournis, la quasi totalité des sites peuvent être calibrés sur une loi log-normale ou une weibull. Dans notre cas, vu que les lois candidates ont le même nombre de paramètres alors utiliser l'AIC revient un peu à utiliser les rapports de vraisemblance. Il est alors égal à :

$$exp(\frac{AIC_{log-normale} - AIC_{weibull}}{2})$$

Si par exemple sa valeur est égal à 0.8 alors la probabilité pour que la weibull soit celle qui minimise la perte d'information est 0.8 fois la probabilité pour que ce soit la log-normale. Le tableau ci dessous rapporte les quantiles de cette quantité sur tout le portefeuille :

0%	25%	50%	75%	100%
0	6.68×10^{-38}	1.36×10^{-7}	$2,82 \times 10^{6}$	Inf

TABLE 8 – Quantiles des rapports de vraisemblances du portefeuille

Nous remarquons alors que sur au moins un quart du portefeuille la weibull fournit de meilleurs ajustements que la log-normale. Un modèle permettant de prédire la loi générant les pertes est alors construit.

La variable à prédire est une variable à deux modalité : weibull ou log-normale. Il s'agit de la loi minimisant la statistique d'Anderson-Darling.
Au vu de sa performance sur les modèles précédents, nous faisons le choix d'utiliser l'algorithme du gradient boosting se basant sur des arbres de décision. La métrique d'évaluation pour ce modèle de classification binaire est la ROC.

La ROC Curve (*Receiver Operating Characteristic*) permet de visualiser la performance d'un classificateur binaire.
La ROC représente la courbe décrivant l'évolution de la sensibilité en fonction du complémentaire de la spécificité, pour toutes les valeurs-seuil envisageables du test.

$$\text{Sensibilité} = \frac{TP}{P}$$

$$\text{Spécificité} = \frac{TN}{N}$$

avec :

$$\begin{array}{lcl} TP & : & \text{Vrai positif} \\ TN & : & \text{Vrai négatif} \\ FP & : & \text{Faux positif} \\ FN & : & \text{Faux négatif} \\ P & = & TP + FN \text{ le nombre de positif} \\ N & = & TN + FP \text{ le nombre de négatif} \end{array}$$

Plus cette courbe se rapproche du coin supérieur gauche, plus le classifieur optimise le taux de vrais positifs et minimise le taux de faux positifs). La courbe d'un classifieur aléatoire fera plus ou moins la même proportion d'erreurs que de classifications correctes quelque soit le seuil et se rapprochera fortement de la diagonale. La courbe ROC peut être synthétiser par l'aire sous cette courbe l'AUC (Area Under Curve) qui doit être supérieure 0,5 (AUC du classifieur aléatoire).

La figure ci-dessous représente la courbe ROC du modèle. Avec un AUC de 97.2%, celui-ci semble très performant.

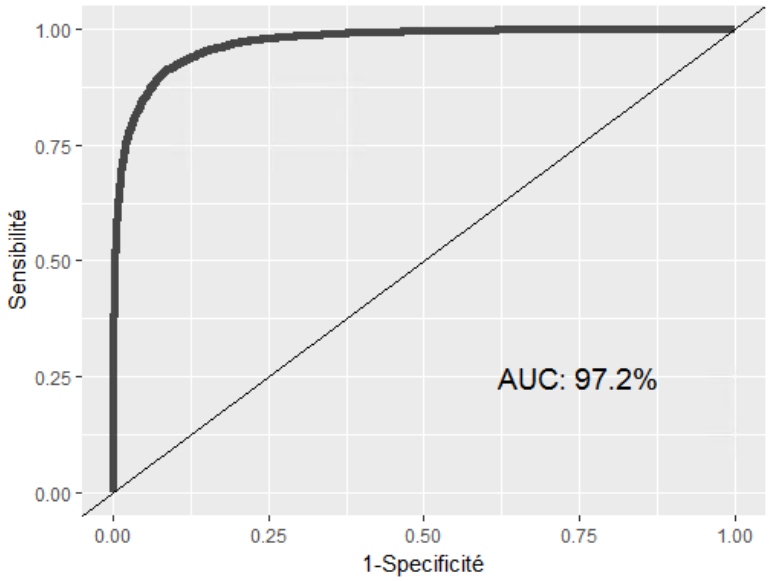


FIGURE 51 – Courbe ROC du modèle de classification

Les variables introduites dans le modèle sont toujours des variables relatives aux caractéristiques physiques des sites traduisant leur vulnérabilité ainsi que des variables relatives à la situation géographique et à l'aléa.
De nouvelles variables issus du module d'aléa ont été injectées au modèle. Parmi elles se trouve la variable *Max*. Cette variable est issue d'une carte de risque que nous avons créée en fonction de la valeur maximum de vitesse de vent sur tout le catalogue stochastique du modèle.

FIGURE 52 – Carte de risque créée à partir du catalogue stochastique d'événements du modèle

La création de cette variable très granulaire permet la détection des zones risquées à un niveau très fin. En affichant l'importance des variables, nous constatons que cette nouvelle variable est celle qui pèse le plus dans la prédiction.

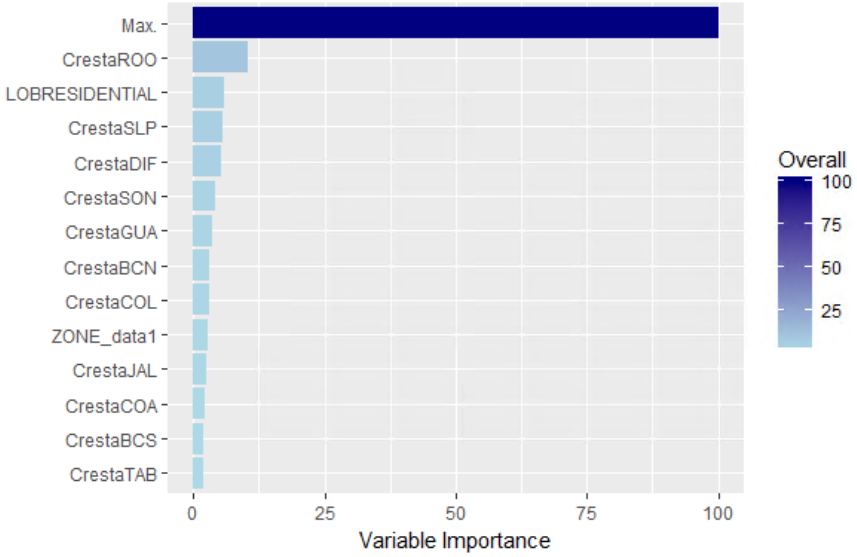


FIGURE 53 – Importance des variables

L'analyse de la sévérité sans changement d'échelle s'avère plus compliquée. Après avoir calibré les distributions de pertes de chaque site du portefeuille sur les lois de bêta et mbbefd, nous comparons les statistiques de Kolmogorov-Smirnov, d'Anderson-darling et AIC relatives aux différentes lois et retenons alors les lois qui minimisent chacune des statistiques.

En tenant compte de tous les critères, les résultats d'ajustements sur tous les sites du portefeuille penchent légèrement en faveur de la loi de bêta, cette dernière s'ajuste à plus de site que la mbbefd.
L'étude de la distribution des critères d'aic et de la statistique d'Anderson- Darling montre une certaine équivalence des deux lois avec moins de déviance extrême pour la bêta.

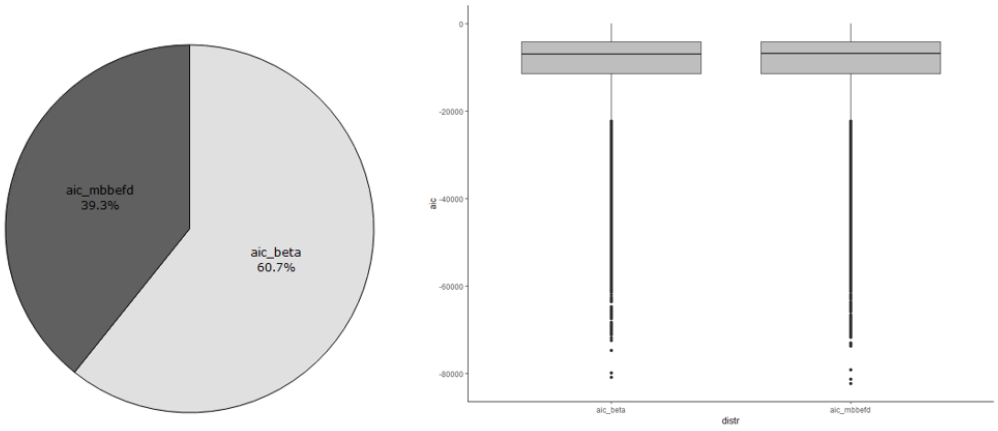


FIGURE 54 – Ajustement de la sévérité des sites selon le critère de l'AIC

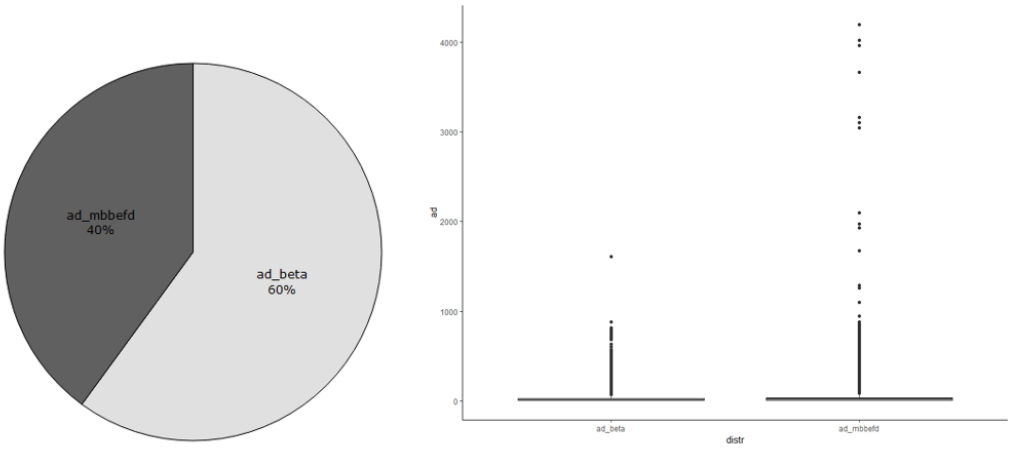


FIGURE 55 – Ajustement de la sévérité des sites selon la statistique d'Anderson-Darling

Contrairement aux ajustements sur les logarithmes de taux de destruction, la répartition géographique des ajustements de la loi mbbefd et de la bêta ne permettent pas de faire des conclusions concernant la relation de la loi générant la sévérité de chaque site et leur vulnérabilité.
A l'issu de cette approche et au vu des résultats de l'analyse comparative de l'ajustement des deux lois à support sur [0,1], la loi de bêta est sélectionnée.

L'objectif de cette section était de choisir parmi différentes lois de sévérité et de fréquence celles qui s'ajustent au mieux aux distributions générées par le modèle CAT. Nous avions alors décider de sélectionner la loi de poisson pour la fréquence. Quant aux lois de sévérité, nous en avons retenues trois :

— **Une loi de bêta lorsque la sévérité ne subit aucune transformation**

— **Des lois de weibull et log-normale lorsque la sévérité subit une transformation en passant par un méta-modèle permettant de déterminer laquelle des deux est la plus adéquate.**

Dans la suite et en fonction des conclusions faites à cette étape, nous étudions la prédiction des paramètres pour chaque approche. Nous établirons ensuite une analyse comparative de chacune tant en termes d'ajustement sur les distributions "réelle" que sur le plan opérationnel.

4.2 Estimation des paramètres du modèle prédictif

4.2.1 Sévérité en terme de taux de destruction

Les deux premiers moments de la loi de bêta s'écrivent de la forme suivante :

$$\begin{cases} E(dr) &= \frac{a}{a+b} \\ V(dr) &= \frac{ab}{(a+b)^2(a+b+1)} \end{cases}$$

$$\begin{cases} a &= \frac{E(dr)(1-E(dr)-V(dr))}{V(dr)} \\ b &= \frac{(1-E(dr)-V(dr))(1-E(dr))}{V(dr)} \end{cases}$$

Étant face à un système à deux équations et deux inconnus, il est possible de retrouver les paramètres a et b à partir de la moyenne et de la variance de la sévérité conditionnelle.
La connaissance des moments d'ordre 1 et 2 nous permettent donc de reproduire la sévérité caractérisant le site.
L'espérance conditionnelle de la sévérité est une composante de la prime pure, plusieurs modèles statistiques ont été donc construit autour de cette métrique à la section 3.
La construction d'un modèle prédictif de la variance ou l'écart-type conditionnel(le) permet donc de retrouver les paramètres de la loi de bêta.
Les modèles de fréquence également calibrés lors de la construction de la prime CAT à la section 3, nous permettent de retrouver le lambda définissant la distribution du nombre d'ouragans par année simulée.
Le choix de la loi de bêta présente donc un avantage considérable. Cette loi nous permet de générer la loi régissant la distribution de pertes tout en maintenant la lisibilité des modèles GLM constituant la prime pure.
Étant donné que l'écart-type conditionnel de la sévérité n'est pas une des composantes de la prime pure et qu'il permet uniquement de générer la distribution de perte, cette métrique sera modélisée au moyen de plusieurs algorithmes machine learning. Nous retiendrons l'algorithme qui apportera le maximum de précision.
Les détails concernant le calibrage de ces modèles non-paramétriques ne présentent aucun intérêt pour le lecteur et ne seront donc pas fournis.
La figure 56 représente le nuage de points des écarts-type prédits en fonction des observés sur les données test. Comme nous pouvons le constater, la majorité des points se trouvent dans la zone bleue qui représente une marge d'erreur de 20%. Les résultats du modèle semblent donc raisonnables.

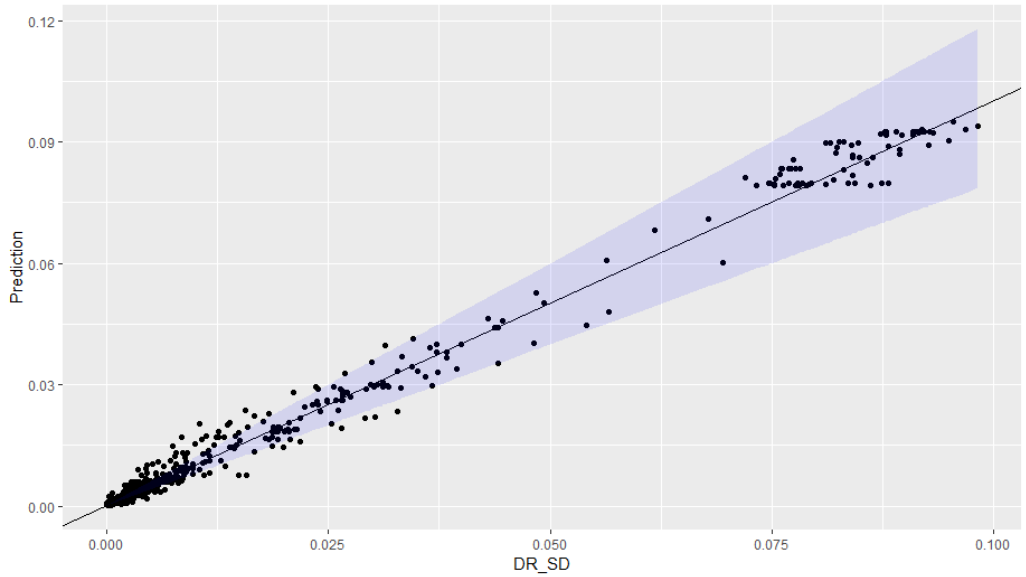


FIGURE 56 – L'écart type conditionnel prédit en fonction de l'observé

4.2.2 Sévérité avec changement d'échelle

Le changement d'échelle que subit la sévérité dans cette approche ne nous permet plus d'utiliser le modèle de sévérité construit à la section 3. Pour l'illustrer, prenons par exemple les deux premiers moments de la loi log-normale :

$$\begin{cases} E(\rho(dr)) &= exp(\mu + \frac{\sigma^2}{2}) \\ V(\rho(dr)) &= exp(\sigma^2 - 1)exp(2\mu + \sigma^2) \end{cases}$$

avec $\rho(x) = -log(x)$
Les paramètres des lois sont exprimés en fonction des deux premiers moments de $\rho(dr)$, c'est le cas aussi pour la loi de weibull. L'utilisation de ces lois implique par conséquent la construction de quatre modèles permettant de prédire chaque paramètre (deux modèles pour chaque loi).
L'examen visuel de la distribution empirique de certains paramètres et le calcul des coefficients d'asymétrie et d'aplatissement ne nous permet pas de sélectionner des distributions candidates. En effet la plupart des paramètres ont une distribution qu'il est difficile d'approcher avec un modèle paramétrique (encore moins la famille exponentielle pour un GLM).
Ils seront donc plutôt modélisés au moyen de modèles non-paramétriques. Le pouvoir prédictif de chacun de ces modèles est représenté par un nuage de point représentant le paramètre prédit en fonction de l'observé sur la base test.
Nous constatons alors que les modèles concernant la moyenne du logarithme pour la loi log-normale et le paramètre d'échelle pour la weibull fournissent de meilleurs résultats que les autres modèles. Ceci est dû en grande partie au fait que la distribution de ces paramètres est moins asymétriques que celles des autres.

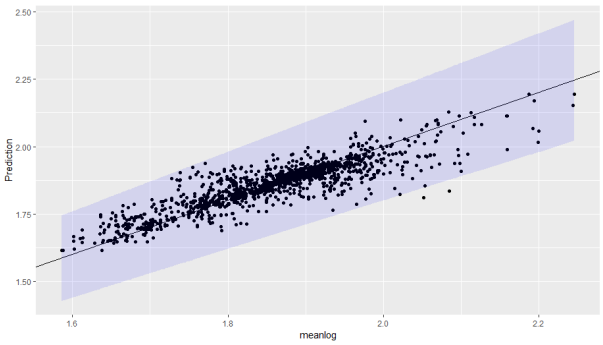


FIGURE 57 – Prédiction de la moyenne du logarithme en fonction de l'observé

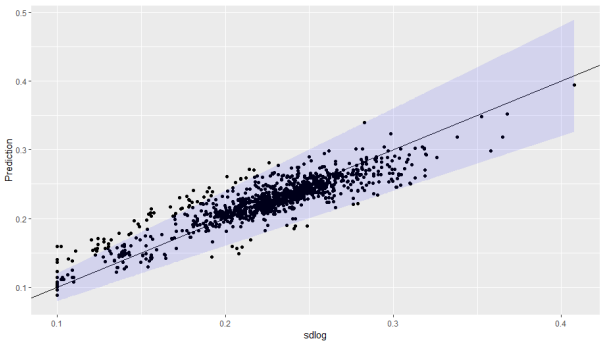


FIGURE 58 – Prédiction de l'écart-type du logarithme en fonction de l'observé

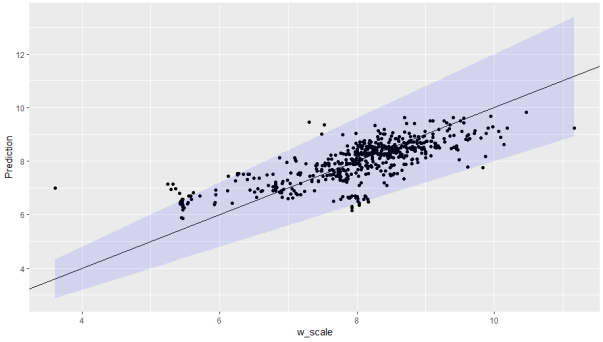


FIGURE 59 – Prédiction du paramètre d'échelle en fonction de l'observé

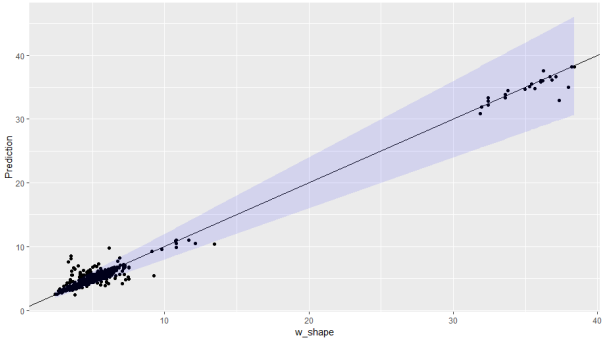


FIGURE 60 – Prédiction du paramètre de forme en fonction de l'observé

A l'issue de l'étape consistant à calibrer des lois sur la sévérité observée, il s'est avéré que la transformation du taux de destruction fournissait un meilleur ajustement.
De plus, l'utilisation des modèles non paramétriques pour modéliser les paramètres permet de réduire l'erreur de prédiction. A priori, cette approche devrait fournir des distributions de pertes raisonnables par rapport à celles générées par le modèle *CAT*.
Malheureusement, cette amélioration de fit devrait se faire au détriment de l'intelligibilité et la lisibilité des modèles de tarification.

4.3 Comparaison des pertes modélisées et observées

Pour simuler la distribution annuelle de perte calibrée pour un site, connaissant sa loi de sévérité et sa loi de fréquence, nous utilisons la méthode Monte-Carlo dont les étapes sont les suivantes :

1. Simulation de N années
2. Pour chaque année simulée :
   (a) Tirage aléatoire dans la loi de poisson de paramètre λ afin de déterminer le nombre NB de sinistres annuels.
   (b) Tirage aléatoire de NB valeurs dans la loi de sévérité selon l'approche étudiée.
3. Consolidation de la courbe par agrégation des pertes par année simulée. La simulation étant très rapide, nous pouvons simuler plus d'années que d'années disponible dans la catalogue stochastique (pour rappel 10.000) pour assurer une convergence.

Plusieurs tests ont été effectués sur des sites avec des profils de risque très variés. Les figures 61 et 62 représentent les AEP de deux sites présentant deux profils de risque extrêmement différents : un site non risqué et un site très risqué. Nous constatons que l'ajustement des taux de destruction transformés au moyen de distributions à queues épaisses fournit des résultats médiocres quelque soit le profil de risque du site.
En revanche, la méthode consistant à calibrer les taux de destruction sans changement d'échelle est beaucoup plus simple et fournit de meilleurs résultats .
Les performances de cette méthode d'ajustement dépend évidemment du profil de risque mais reste raisonnable dans tous les cas.

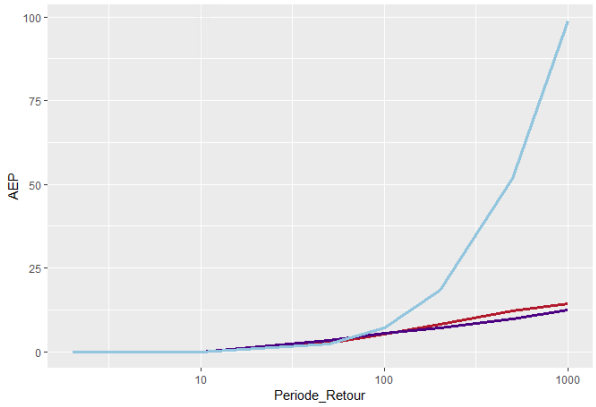


FIGURE 61 – Exemple d'AEP d'un site non risqué

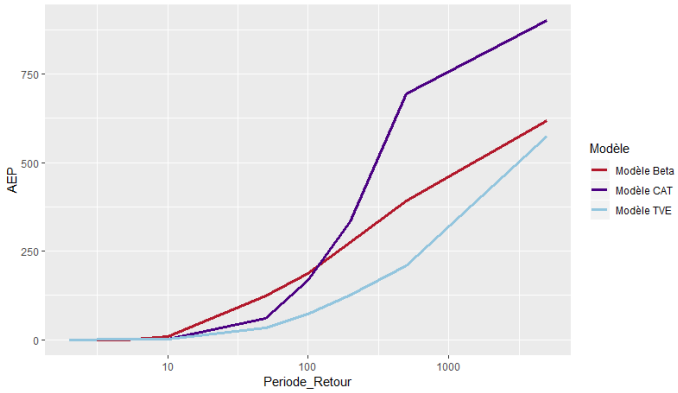


FIGURE 62 – Exemple d'AEP d'un site risqué

Pour évaluer l'adéquation des distributions de pertes simulées ajustées au moyen de la bêta et les distributions de pertes observées sur l'ensemble du portefeuille, nous mesurons le pourcentage d'erreur faite sur l'AEP à différentes périodes de retour.
La distribution de la valeur absolue des erreurs relatives sont résumées dans le tableau suivant :

	Min	25%	50%	75%
50	0.01%	8.2%	19.5%	31.8%
100	0.003%	7.4%	12.5%	25.7%
200	0.04%	9.3%	14.5%	27.1%
500	0.05%	9.9%	17.1%	30.3%

TABLE 9 – Distribution des erreurs en fonction des périodes de retour

Le tableau 9 représente les quantiles de pertes relatives absolues. Nous remarquons que sur les trois quarts du portefeuille nous ne faisons pas plus de 32% d'erreur et ce même sur une période de retour de 500 ans.

4.4 Comparaison des tarifs modélisées et observées

Pour être en accord avec nos analyses de primes CAT faites à la section 3, nous comparons les pertes annuelles moyennes AAL calculées sur les distributions simulées et celles des distributions issues du modèle CAT. Le tableau suivant résume les résultats obtenus (en base 100) :

Primes (M $)	Pertes espérées	Prime modélisées
Brut	100	100,42
Net	100	112.92

Les primes modélisées et les pertes espérées sont du même ordre de grandeur en brut et en net. Les primes nettes issues des distributions de pertes simulées fournissent une meilleure approximation que celles obtenues par application des courbes de rabais à la section précédente. Ce résultat est tout à fait normal compte tenu du fait que nous avons accès à la distribution de pertes par événement, les primes nettes peuvent être calculées facilement et de façon précise.
Pour nous convaincre de la précision apportée par la méthode, nous avons comparer pour chaque niveau de franchise les taux de rabais réels observés sur la prime d'un site tiré aléatoirement avec les taux de rabais induits par cette nouvelle méthode et ceux provenant la courbe calibrée pour le mexique présentés à la section 3.

Franchise	Taux de rabais observés	Taux de rabais actuel	Taux de rabais modélisés
1%	39%	13%	36%
2%	49%	20%	50%
3%	55%	26%	59%
5%	61%	34%	70%
10%	71%	47%	85%
15%	77%	55%	89%
20%	82%	61%	90%

Encore une fois, nous constatons que les taux de rabais observés ne sont pas du même ordre de grandeur que ceux provenant de la courbe de rabais.
Nous remarquons également que la nouvelle approche tend à surestimer le taux de rabais, cela s'explique par le fait que la loi de bêta retenue pour la modélisation de la sévérité ne réplique pas les valeurs extrêmes observées qui contribuent de façon significative au calcul du taux de rabais.
De façon générale, nous pourrons remarquer que l'on aura tendance à sous estimer le prix net

de franchises pour les sites assez risqués. En réalité, les pertes extrêmes sont peu sensibles à l'introduction d'une franchise ce qui explique l'importance des primes nettes.
De ce fait, une sous estimation de ces valeurs extrêmes entraîne automatiquement une sous-estimation de la prime nette.

4.5 Limites de cette approche

Bien que les lois paramétriques simples (deux paramètres au plus) soient très utilisées pour modéliser la sévérité, il arrive souvent que les données soient complexes pour ces distributions. En effet, une distribution de perte à un niveau aussi granulaire que celui du site nécessite l'introduction de nouvelles lois plus flexibles et capables de s'adapter aux particularités (asymétrie positive et queue épaisse...etc.) de ces pertes.
Le recours à des familles de distributions, dites généralisées ou à des lois mélanges est en mesure d'apporter plus d'ajustement entre données et modèles.
Cependant l'automatisation du processus de calibrage des pertes sur des lois mélanges est souvent compliqué et relève du cas par cas.
Les familles de distributions généralisées, elles, recouvrent plusieurs lois de probabilité connues et sont donc en mesure de s'ajuster à différents échantillons. En revanche, ces dernières contiennent des lois ayant au moins quatre paramètres. La modélisation de la moyenne et l'écart-type de taux de destruction ne suffit plus. Afin d'assurer l'identification de chaque paramètre, ce type de loi nécessitera la modélisation d'au moins quatre métriques, typiquement : un paramètre de localisation (médiane ou moyenne), un paramètre d'échelle (volatilité ou écart type), un paramètre d'asymétrie et un paramètre d'aplatissement.
Faute de temps, nous n'avions pas eu l'occasion de tester ces lois et aller au bout de cet exercice.

La prédiction de l'écart type à la seconde étape est également une source d'erreur dont la magnitude dépend des modèles statistiques utilisés.
De plus pour la sévérité, l'incertitude des paramètres impacte différemment le tarif des tranches suivant leur hauteur. Ceci tient au fait qu'une des composantes de la prime pure à la tranche est l'aire au-dessus de la distribution cumulée de sévérité calculée entre la franchise et la limite de la tranche. Par conséquent, plus la tranche est élevée, plus l'incertitude autour du paramètre estimé engendre de l'incertitude de la charge individuelle à la tranche.

5 Prix technique et post-ajustement

Les différentes méthodes de tarification présentées ci-dessus permettent une sophistication de la segmentation du risque d'Ouragan au Mexique sur tout le portefeuille considéré. En réalité, pour assurer sa rentabilité et minimiser un risque de déviation incontrôlée de son résultat, il est courant pour un assureur de charger sa prime pure d'un montant supplémentaire constituant ainsi un *"matelas de sécurité"*. Ci-dessous, nous présentons la mise à l'ultime du tarif proposé par le biais de deux ajustements : un chargement technique de la prime pure permettant de couvrir les déviations de la sinistralité individuelle et d'un ajustement plus global du tarif permettant d'équilibrer les primes modélisées à la charge réelle du portefeuille.

5.1 Chargement technique

Afin de garantir sa rentabilité, l'assureur introduit un chargement technique qui vient s'ajouter à la prime pure et qui prend en compte la volatilité potentielle de la sinistralité individuelle de la police.

$$\text{Prix technique} = \frac{PP + \alpha\sigma}{1 - g}$$

avec :

PP	:	La prime pure de la police
σ	:	La déviation de la distribution de perte de la police
α	:	Coefficient de chargement de sécurité
g	:	Chargement lié aux frais de gestion

La prime pure de la police ainsi que la déviation de sa distribution de perte peuvent être calculées en utilisant les résultats la modélisation *CAT*. La définition des chargements de sécurité et de frais de gestion, quant à elle, dépend de plusieurs éléments intrinsèques à la politique de l'assureur et sort du cadre de ce mémoire.

5.2 Ajustement à la charge réelle

Cette dernière étape permet de garantir l'égalité entre la somme des primes modélisées et la charge réelle de l'assureur.

$$\boxed{\text{Charge réelle} = \text{Le coût de réassurance } + \text{Le coût de la rétention}}$$

Dans la suite ne sera considéré que le cas d'une réassurance non proportionnelle , les ajustements dans le cas d'une structure proportionnel étant évidents.
Imaginons que le traité *CAT* couvrant l'assureur Mexicain se compose des quatre tranches suivantes :

	Limite	Priorité	AAD	Reconstitutions
XL1	12M	3M	3M	2 à 100%
XL2	35M	15M		1 à 100%
XL3	200M	50M		1 à 100%
XL3	500M	250M		1 à 100%

TABLE 10 – Structure du traité de réassurance CAT

Le coût de la rétention - Les pertes nettes de réassurance

Les pertes nettes de réassurance représentent les coûts des sinistres restant à la charge de l'assureur après application de la réassurance.
Ce montant peut être estimé à partir des bases de sinistres historiques actualisées "*as if*". Il s'agit de la moyenne des coûts historiques après application de la tranche de rétention. Cette méthode n'est alors consistante que si l'historique à disposition est assez fourni.
Le montant peut être également déterminé à partir de l'utilisation de modèles CAT. Dans ce cas, il correspond à la moyenne des pertes annuelles relatives à la tranche non couverte par le traité sur toutes les années simulées du modèle.
Ce montant peut également être calculé à partir d'une moyenne pondérée combinant les coûts historiques et les coûts simulées et dont les poids sont établis par avis d'experts.

Le coût de réassurance

Dans la suite, nous considérons que la structure non proportionnelle de réassurance de l'assureur mexicain couvre plusieurs périls.
L'estimation des coûts de réassurance à allouer au risque étudié, en l'occurrence le risque d'Ouragan, nécessite encore une fois l'usage de la modélisation *CAT* pour définir une clé de répartition correcte par péril.

Peril	AAL	Clé de répartition
Inondations	10M	9%
Tremblement de terre	55M	50%
Ouragan	45M	41%

TABLE 11 – Structure du traité de réassurance CAT

Le coût du risque d'Ouragan représente donc 41% du coût total du traité. L'enveloppe de primes allouée au risque Ouragan est alors égale à la somme des coûts de réassurance qui lui sont associés et du coût de rétention.

La définition de l'assiette de prime représente une étape importante pour assurer la solvabilité de l'assureur. La définition des primes pures, elle, permet de discriminer entre les sites et d'allouer proprement sa capacité de souscription. A l'issue de ces deux étapes, est alors construit un tarif *CAT* dont la formule est la suivante :

$$\text{Prime CAT} = \frac{\text{Charges réelles}}{\text{Somme des primes modélisées}}\text{Prix technique}$$

Cette prime *CAT* individuelle reflète le risque intrinsèque à la police mais ne tient pas compte du principe de l'accumulation du risque par zone qui impacte fortement le coût de réassurance.

Dans la section suivante, nous présenterons un outil de souscription répliquant le processus de calcul de pertes des modèles CAT. Cette solution permet de reproduire les distributions de pertes en utilisant les modules d'aléa et de vulnérabilité du modèle interne et répond à toutes les réflexions relatives à l'accumulation du capital et l'optimisation prospective du portefeuille.

6 Restructuration de l'architecture informatique du modèle CAT

Comme mentionné précédemment pour obtenir les primes nettes exactes et outrepasser les courbes de rabais, nous avons besoin de la distribution des pertes au niveau site. L'idée de la méthode précédente était de prédire la distribution des pertes d'un site en essayant de prédire les paramètres de la loi sous-jacente.
Contrairement aux approches présentées précédemment, cette technique de tarification repose uniquement sur le modèle catastrophe naturelle, autrement dit, elle ne repose sur aucun autre modèle statistique.
Cette dernière met en valeur l'intérêt du développement de modèles *CAT* en interne pour l'optimisation prospective du portefeuille et la mise en place d'une tarification dynamique.

6.1 Méthode développée

L'outil développé permet la prise en main du modèle *CAT* et l'exploitation des résultats du module d'aléa et de vulnérabilité pour pouvoir générer des distributions de pertes au niveau site. Cette distribution nous permettra alors de tarifer un site en prenant en compte toutes les structures de réassurance même les plus complexes et d'intégrer la notion d'accumulation du risque.
Dans un souci pédagogique, les différentes procédures et calculs sont illustrés à l'aide d'un schéma, à travers un exemple fictif simplifié.

Exemple : Imaginons que l'on veuille tarifer un site à usage commercial construit dans les années 90 en béton se situant à la ville de *La Paz* en Basse Californie du sud dont la valeur assurée est estimée à 1000$. Nous supposons aussi que ce site est soumis à des conditions d'assurance soit une franchise de 100 $ et une limite contractuelle de 500 $ par événement.

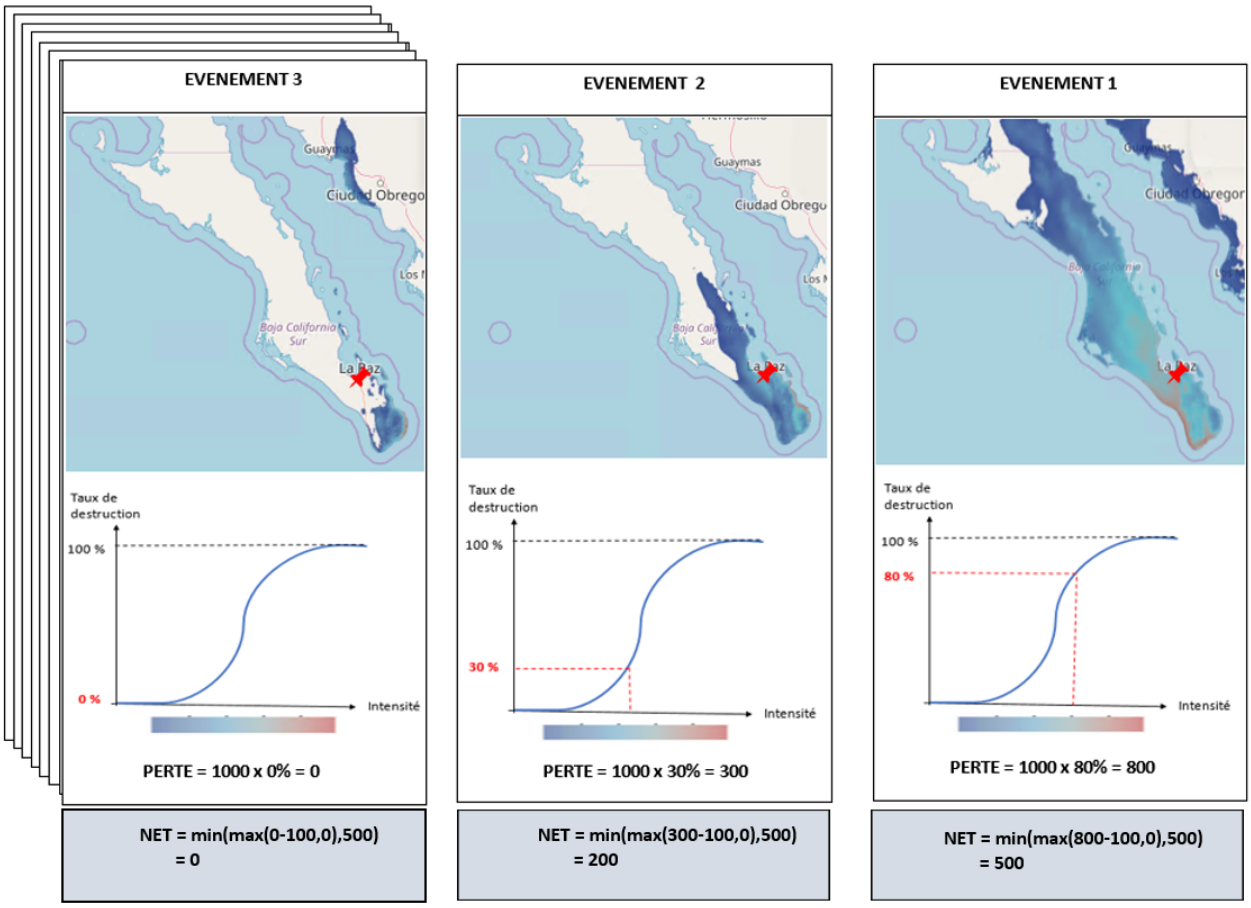


FIGURE 63 – Le process de l'outil

— La première étape consiste à requêter dans tous les fichiers d'événements la cellule correspondant au coordonnées du site entré en input pour extraire la valeur d'aléa correspondante.

— L'étape suivante a pour but dans un premier temps d'utiliser les informations impactant la vulnérabilité du site entrées en input (telles que la ligne de business, la structure du bâtiment, le type de couverture,...) afin de retrouver la courbes de vulnérabilité correspondant au profil de risque du site. Ensuite, dans un second temps, nous associons à chaque valeur d'aléa obtenue lors de la première étape un taux de destruction et une *chance of Loss*.
En faisant le produit de ces grandeurs, nous obtenons une prime pure en terme de taux de destruction.
Il est à noter que pour une meilleure lisibilité de la figure 63 les courbes de vulnérabilités représentent déjà le produit *Sévérité et Chance of Loss*

— La troisième étape permet la prise en compte de l'incertitude secondaire en choquant les taux de destruction avec un coefficient de variation COV. Pour rappel, ce facteur suit une loi log-normale centrée en 1 et permet de modéliser la volatilité autour d'un taux de destruction moyen. Variant selon la branche d'activité (lob), il est défini par la formule suivante

$$COV_{lob} = \frac{\sigma_{lob}}{\mu_{lob}}$$

avec :

— σ_{lob} : Ecart-type des pertes associé à une branche d'activité

— μ_{lob} : Moyenne des pertes associée à une branche d'activité

On peut fractionner cette étape en trois sous-étapes :

1. Générer pour chaque branche d'activité des observations selon une log-normale centrée en 1 et un écart type égal à COV.
2. Tirer aléatoirement dans cette loi des valeurs de COV
3. Multiplier ces COV tirés un bon nombre de fois par le taux de destruction moyen qui lui est déterministe
4. Calculer la moyenne de la distribution des taux de destruction obtenue à l'issu de l'étape précédente

— La quatrième étape consiste à multiplier la distribution des pertes en termes de taux de destruction, par la somme assurée pour avoir la distribution des pertes du site en montant absolu.

— La cinquième étape consiste à appliquer les conditions financières d'assurance.

— La dernière étape tient compte de la structure de réassurance à tous les niveaux , du niveau site au niveau grands comptes.

A noter qu'il s'agit là d'une simplification parce qu'en réalité l'outil permet d'obtenir une distribution de pertes pour chaque site et pour chaque garantie. Par conséquent, l'ensemble des étapes définies précédemment sont appliquées pour chaque site et pour chaque garantie.[9]

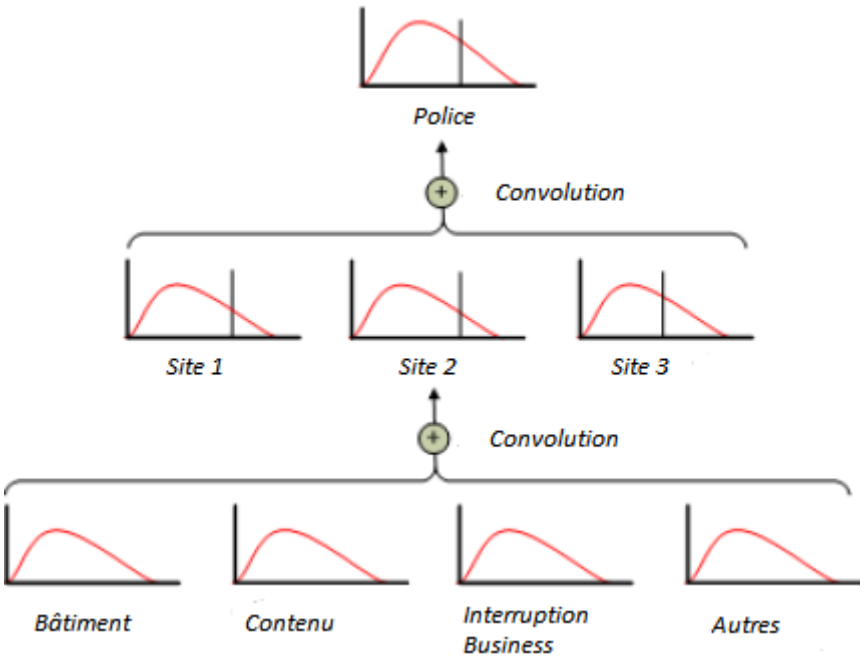


FIGURE 64 – Gestion d'une police multi-sites à plusieurs garanties

Une fois la distribution des pertes du site par événement à disposition, nous sommes aptes à appliquer toutes les conditions financières aussi bien au niveau du site qu'au niveau police.
Le programme développé permet le calcul de toutes les métriques qui nous intéressent dont la prime pure, l'écart type conditionnel et les OEP et AEP ainsi que l'impact de la police tarifée sur ces derniers.
Cette nouvelle approche est simple et fournit des résultats dont les débouchés potentiels sont extrêmement intéressants. Elle permet aux souscripteurs d'évaluer le coût réel d'une police pour l'assureur peu importe la complexité de son programme de réassurance.

Nous pouvons également l'utiliser pour déterminer la contribution marginale des sites au PML (perte bi-centenaire) établissant ainsi un classement des polices les plus risquées et/ou plus rentables.

L'atout majeure de cette technique réside dans sa flexibilité et le fait qu'elle soit applicable à n'importe quel péril et portefeuilles du moment que l'on dispose d'un modèle CAT permettant de récupérer les distributions de pertes brutes par événement et par sites.

6.2 Implémentation de la méthode

6.2.1 Inputs

Lorsque l'on veut tarifer une police, il suffit alors de renseigner pour chacun de ses sites les informations suivantes :

1. Les caractéristiques de l'exposition : Coordonnées géographiques, indicateurs de vulnérabilité (ligne de business, les matériaux de construction, l'année de construction, le nombre d'étage,...), les sommes assurées par garantie et les conditions d'assurance du site.
2. Les conditions d'assurance de la police.
3. Les facultatives des polices où chaque ligne correspondant à une tranche pour les non-proportionnelles avec le point d'attachement de la tranche, sa limite et son taux de cession.
4. Les facultatives des sites contenant les même informations que celui de la police
5. Conditions du traité CAT

6.2.2 Outputs

L'obtention des distributions de pertes au niveau site permet de tarifer de façon précise n'importe quelles structures y compris celles dont les conditions financières (ou agrégation de conditions financières) sont complexes.
Les perspectives ne se limitent pas à la tarification, elle permet également le calcul d'une multitude de mesures de risques tels que la perte bi-centenaire du site ou de la police appelée aussi *MPL* (*Maximum probable loss*) ou une mesure plus stable telle que la *TVaR*(*Tail value at risk*)[7].
Cette méthode octroie la possibilité d'évaluer l'optimalité des structures de couverture en réassurance notamment en analysant l'impact d'une augmentation/baisse de la rétention, l'introduction d'un niveau de franchise supérieur ou encore la cession d'une facultative.
Enfin, l'outil habilite l'assureur à contrôler l'accumulation de son exposition en analysant l'impact de la souscription d'une nouvelle affaire ou la résiliation d'une police sur le bi-centenaire de son portefeuille et sur le coût de réassurance.

6.2.2.1 Tarification

L'outil permet de retrouver les pertes à tout niveau de granularité par événement. Ceci nous permet de tarifer ou du moins fournir une bonne approximation du prix de n'importe quel structure d'assurance ou de réassurance.

— Calcul du prix net : L'outil permet à son utilisateur de connaître exactement le montant qui reste à sa charge pour n'importe quel risque et à n'importe quel niveau de granularité. L'application des conditions d'assurance et de réassurance à des niveaux de granularité différents est assez complexe et peut être résumée en 6 grandes étapes :

1. Application des conditions d'assurance (Franchise /limite par type de garantie) au niveau site
2. Application des conditions d'assurance (Franchise /limite par type de garantie) au niveau police
3. Back allocation ou répartition des pertes nettes de conditions d'assurance sur tous les sites.
4. Application des facultatives au niveau site.
5. Application des facultatives au niveau police.
6. Back allocation ou répartition des pertes nettes d'assurance et de réassurance sur les sites de la police.

— Tarification d'une facultative : La tarification d'une tranche de facultative non proportionnelle se fait selon la formule suivante :

$$Prix_{tranche} = \frac{min(max(Pertes - \text{Point attachement}_{tranche}, 0), \text{Capacité}_{tranche})\text{Taux cession}}{N_{sim}}$$

avec N_{sim} le nombre d'années simulées dans le modèle.
La détermination du prix d'une tranche de facultative permet d' évaluer sa rentabilité en comparant le prix modélisé au prix proposé par le marché. Cela permet aussi de définir le ROL de chaque tranche et de déterminer leur amortissement.
Concrètement l'outil pourrait être un support pour l'optimisation des conditions d'assurance et de réassurance.

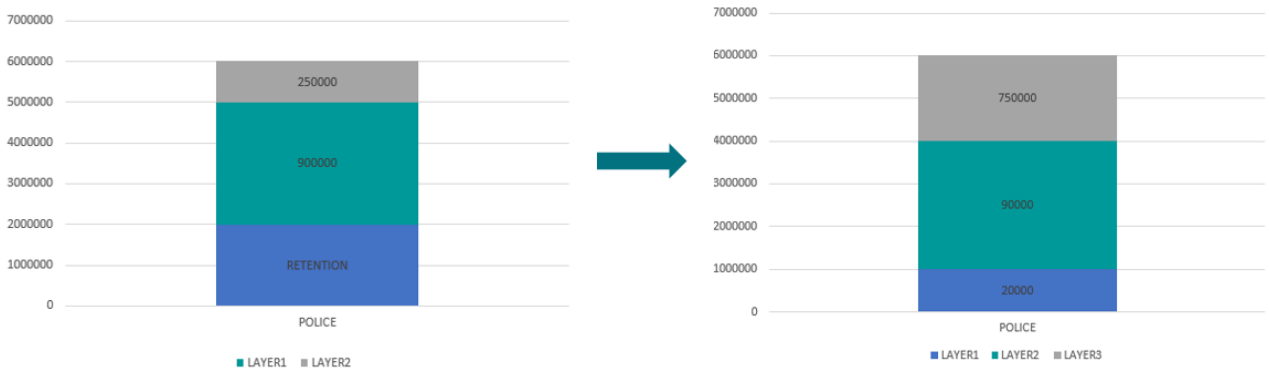


FIGURE 65 – Impact d'un changement structurel des conditions d'assurance ou de réassurance (baisse/hausse de la participation, introduction d'une franchise plus haute,...).

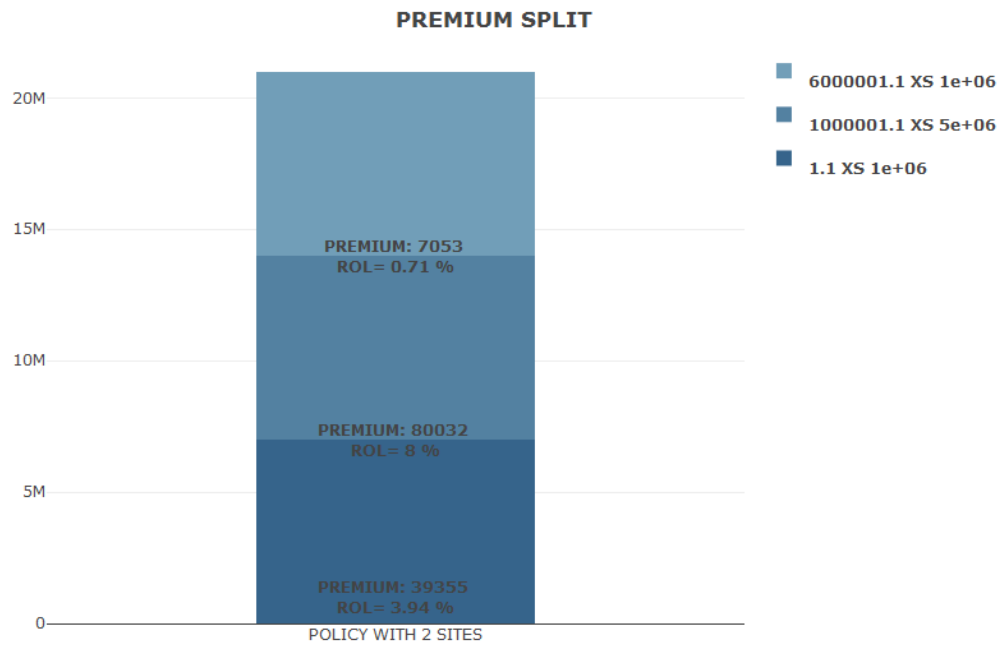


FIGURE 66 – Tarification des tranches d'une facultative et calcul du ROL

6.2.2.2 Impact marginal sur le PML

Hormis le scoring de risque, les assureurs doivent contrôler l'accumulation de leur exposition dans les zones à risque afin de réduire le volume des pertes lors de la survenance d'événements extrêmes.
En effet, dans un contexte tel que celui des catastrophes naturelles, un seul événement peut toucher plusieurs sites assurés.
Assurer 1000 sites dans une zone non risquée n'est pas problématique. En revanche, 1000 sites sur la côte de Quintana Roo expose l'assureur à un niveau de pertes inacceptable.
La gestion du risque liée à un portefeuille dans sa globalité permet de garantir la solvabilité de la compagnie. De plus, elle permet de réduire le coût de réassurance et de maîtriser le nombre de sinistres à gérer suite à l'occurrence d'un événement.
La sélection du risque par le passé était basée sur une séquence récursive de règles de décision en fonction des caractéristiques du site ou établie à dire d'experts.
L'outil fournit une évaluation du risque qui englobe le portefeuille de l'assureur dans son intégralité. En effet, il permet de mesurer l'impact de l'intégration d'une nouvelle police ou de la résiliation d'une existante sur les courbes EP.

La différence de pertes à chaque période de retour de la courbe EP relative au portefeuille original et celle du portefeuille suite à l'intégration de la police représente la perte marginale induite par l'introduction de ce nouveau risque.
L'impact marginal sur le *PML* prend en compte la diversification géographique de l'exposition. Cette mesure de risque varie en fonction de l'accumulation de l'exposition dans chaque zone et son niveau de risque.

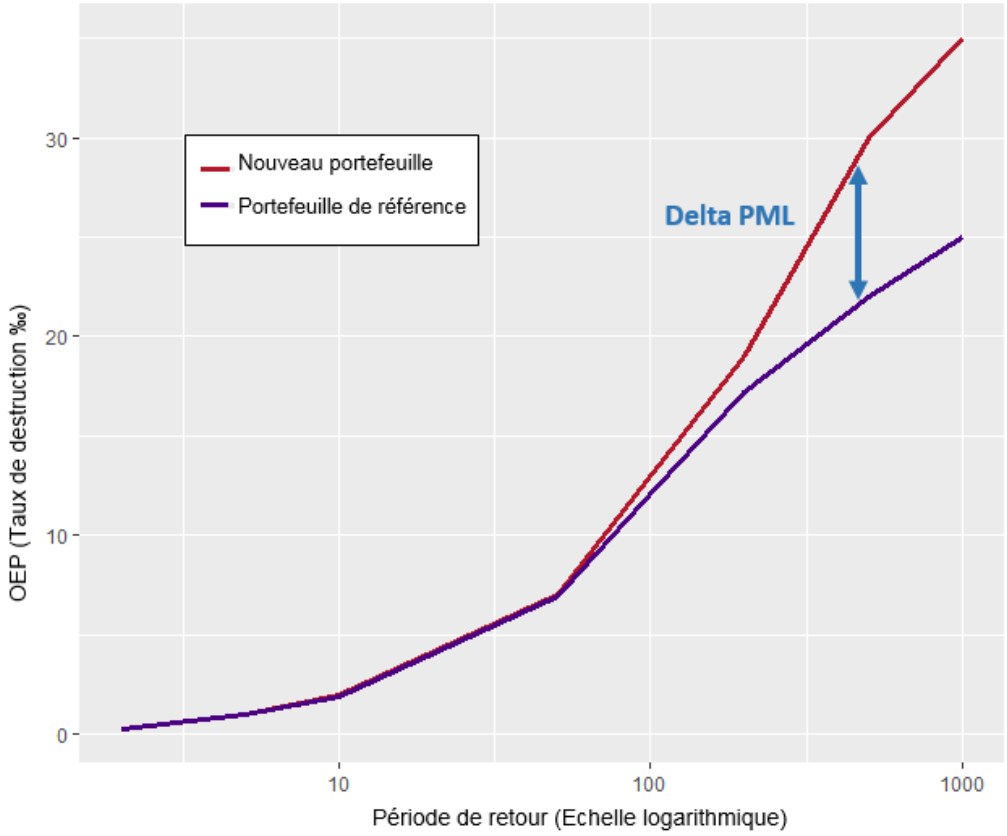


FIGURE 67 – Impact Marginal d'une police sur le PML

L'objectif est donc de pouvoir mesurer cet impact marginal sur le *PML* du portefeuille. Le *PML* (ou l'*OEP*) n'est pas une mesure de risque additive contrairement à l'*AAL*. En terme d'*AAL*, l'impact marginal d'une nouvelle police sur le portefeuille de référence est égal à sa prime pure (son *AAL*). En terme de *PML*, la méthodologie est différente. Le *PML* d'un portefeuille est inférieur à la somme des *PML* des risques qui le composent.
En temps normal, il aurait fallu redéployer tout le modèle *CAT* pour le calcul de la métrique, cela aurait pris quelques heures.
L'outil fournit cette information en quelques secondes en sommant les YLT du portefeuille de référence et celui de la nouvelle police par événement qu'il utilise pour mettre à jour les courbes EP.
Reprenons l'exemple du *CAT* Modèle simplifié avec 10 événements chacun survenant à des années différentes.

EvenementID	Portefeuille de référence	Nouveau compte	Nouveau Portefeuille
1	50	0	50
2	7	2	9
3	13	0	13
4	32	0	32
5	46	8	54
6	25	2	27
7	27	0	27
8	31	3	34
9	6	6	12
10	10	0	10

TABLE 12 – Exemple calcul de l'impact marginal

L'impact marginal du nouveau compte sur le *PML* avec une période de retour de 10 ans est de 4M$.

La figure 68 synthétise le mécanisme de calcul de l'impact marginal d'un nouveau risque sur le *PML*.

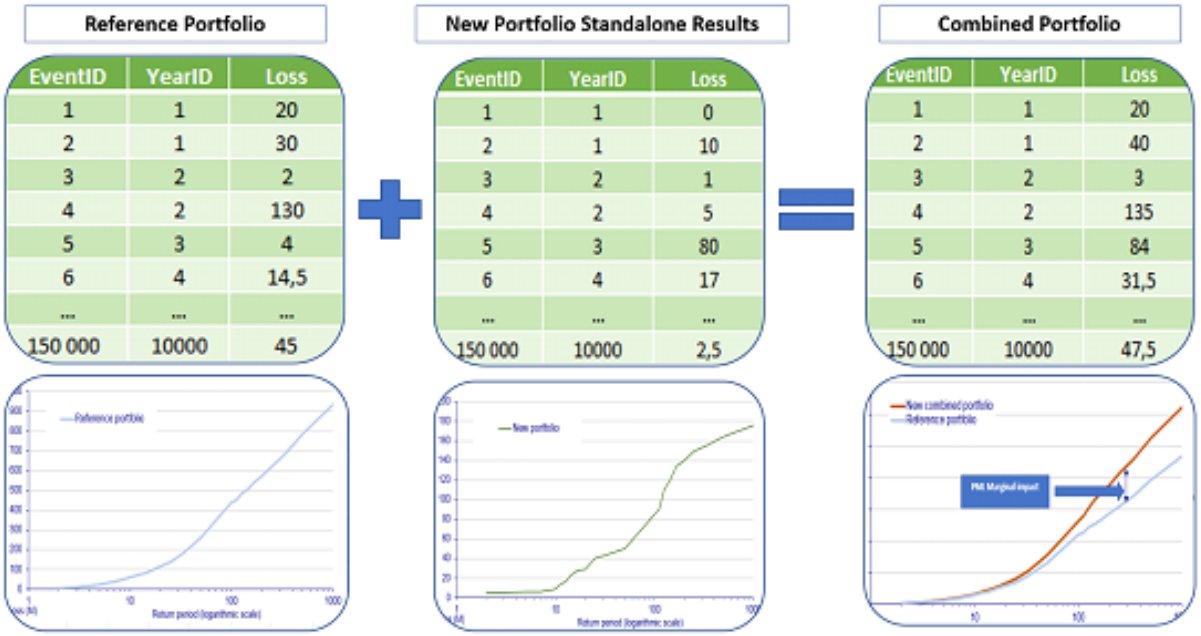


FIGURE 68 – Exemple du mécanisme de calcul

Pour chaque événement du catalogue stochastique, les pertes du nouveau portefeuille sont ajoutées aux pertes du portefeuille de référence afin que le nouvel *YLT* comprenne les pertes combinées des deux ensembles d'exposition lorsqu'un événement les touche simultanément et les pertes pour un seul ensemble d'exposition lorsque l'événement a un impact uniquement sur le portefeuille de référence ou le nouveau portefeuille.

6.2.2.3 Coût de la réassurance

Dans le paragraphe précédant, nous avions vu qu'un nouveau risque pouvait impacter ou pas le *PML* du portefeuille en fonction de sa diversification géographique.
Dans le cas où un nouveau compte aurait un impact(surtout sur les longues périodes de retour), il est intéressant pour l'utilisateur de déterminer le coût de réassurance qu'il engendrera et s'assurer que cela soit en phase avec l'appétence au risque de l'assureur.

Le coût de réassurance varie selon plusieurs facteurs(les conditions du marché, l'appétence au risque du réassureur, ...), cependant l'outil pourrait fournir un bon proxy.
L'impact d'un nouveau risque sur le coût de réassurance est calculé suivant trois formules différentes :

— Impact sur l'*AAL* : Le proxy le plus simple consiste à réévaluer l'*AAL* du nouveau portefeuille et considérer que l'évolution de l'AAL impacte directement le prix de la réassurance.

$$\Delta\text{Coût} = \text{Coût de réassurance } * \Delta AAL(\%)$$

— Impact sur la capacité de la tranche la plus haute : Nous pouvons aussi considérer que l'impact marginal d'un nouveau risque sur le PML entraîne uniquement une plus grande

capacité de la tranche la plus haute (*Top layer en anglais*). Connaissant le *rate of line* ROL, le :

$$\Delta\text{Coût} = ROL * \Delta\text{Capacité}_{\text{Top Layer}}$$

— Réévaluer le coût technique de la réassurance du nouveau portefeuille : La manière la plus juste mais aussi la plus coûteuse est de retarifer le traité CAT en utilisant les résultats des modèles CAT sur le nouveau portefeuille.

6.3 Changement de paradigme

La technique actuelle reprend le processus de calcul de pertes des modèles *CAT* sous un angle différent.

— Architecture Informatique : Le calcul de pertes des modèles *CAT* s'appuie sur une plateforme cloud faisant coopérer des centaines de machines virtuelles. Le catalogue d'événement est alors distribué entre les différentes machines. L'outil lui tourne en local.

— Objectif : Les *CAT* modèles sont conçus pour tarifer un portefeuille en entier. La complexité des prétraitements requis et les ressources mobilisées lors de leurs déploiements sont conséquentes. Par conséquent la tarification d'une ou quelques polices par le biais d'un modèle *CAT* n'est pas une solution optimale.

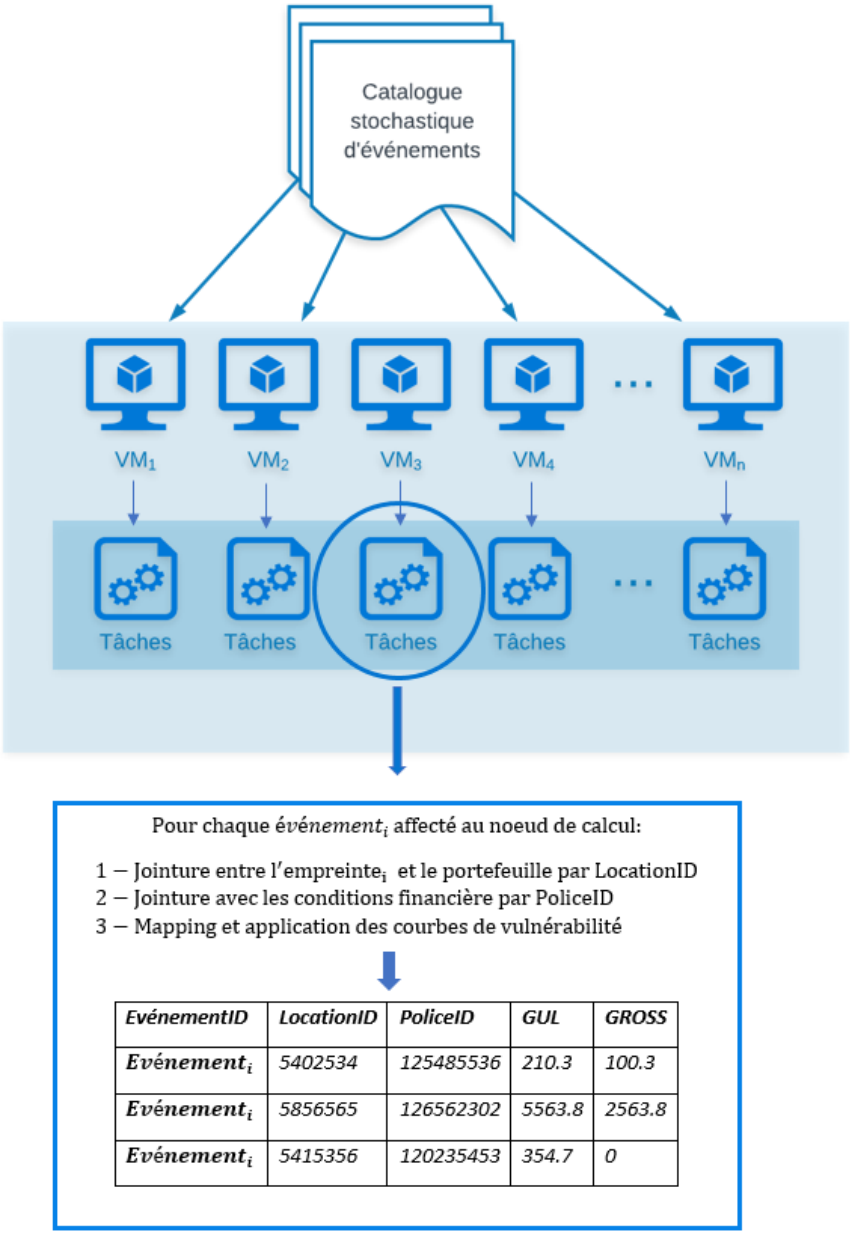


FIGURE 69 – Architecture lancement du modèle CAT

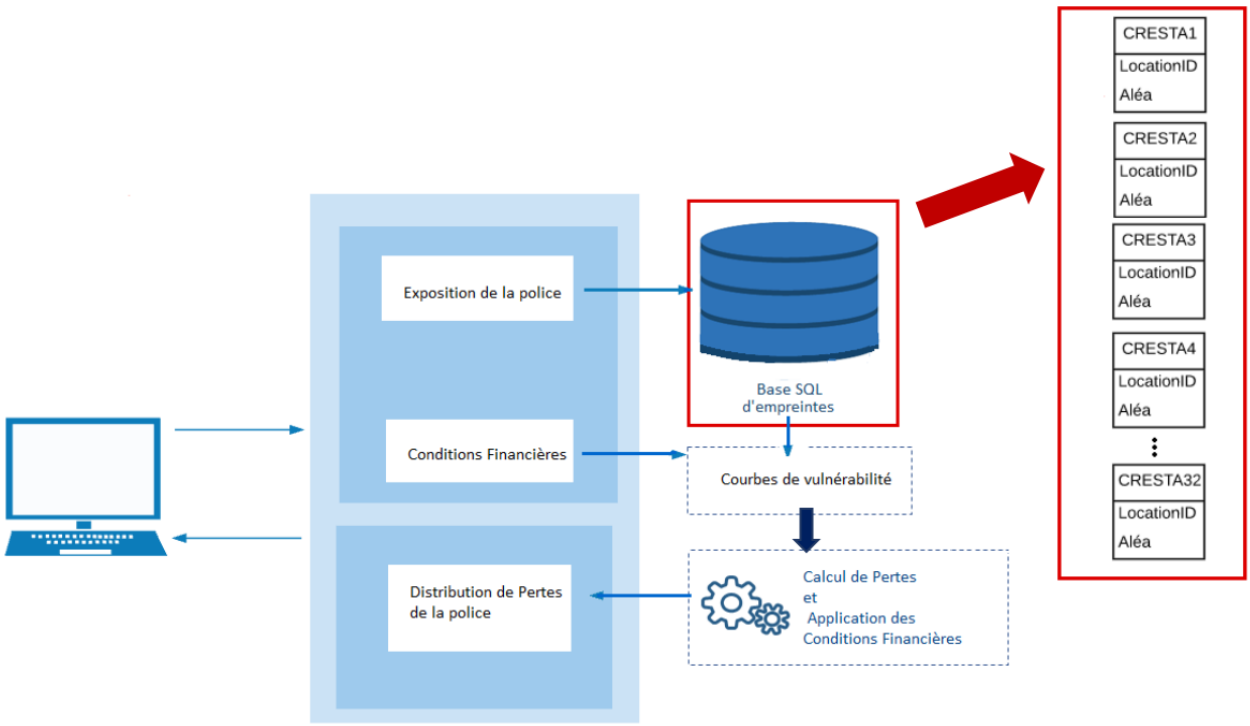


FIGURE 70 – Architecture de l'outil

L'empreinte géographique d'un événement est une matrice de pixels (également appelés cellules) nous fournissant une valeur d'intensité d'aléa pour chaque cellule.
Les modèles *CAT* comprennent plusieurs événements stochastiques qui peuvent atteindre pour certains périls une centaine de milliers. Les empreintes relatives à ces événements sont stockées sur des disques durs Azur (une solution de cloud computing proposée par *Windows*) et occupent des dizaines de Gigabits.

Comme mentionné précédemment, le but est de pouvoir requêter le numéro de cellule associé à la localisation d'un site sur l'ensemble des empreintes de façon quasi instantanée.
La solution la plus naive consiste à concaténer toutes les empreintes puis d'extraire les événement touchant la cellule concernée et la valeur d'aléa correspondante.
Dans un premier temps, afin d'alléger cette table, nous avons décidé de découper au préalable toutes les empreintes de notre catalogue pour se débarrasser de toutes les cellules ne faisant pas partie du territoire mexicain et celle dans les océans.

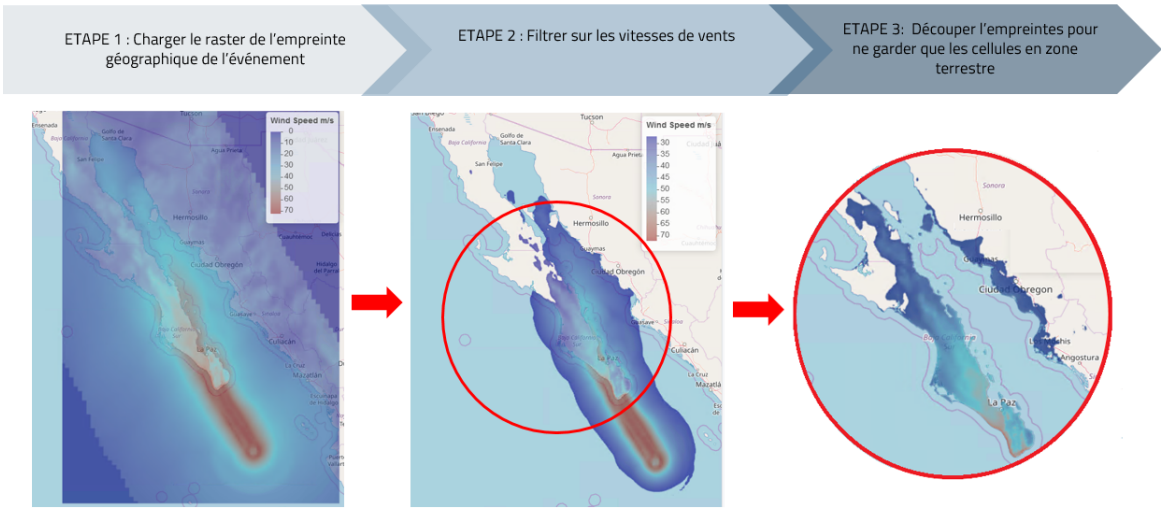


FIGURE 71 – Traitements des empreintes géographiques

Malgré un gain très important, la table comporte toujours plus de 750 millions de lignes ce qui équivaut à 30 Gigabits. Une telle base met 20 minutes à s'ouvrir sur une machine virtuelle possédant 250 Gigabits de mémoire vive. Afin de récupérer que les événements touchant une cellule en particulier, nous stockons la table suivant une architecture de base de donnée composée de 32 tables correspondants aux 32 crestas mexicaines.
Chaque table contient donc l'ensemble des lignes correspondant aux cellules composant une cresta en particulier. Cette structuration de la base de donnée suivant une logique géographique nous permet d'optimiser le temps de traitement de la requête qui ne sollicite désormais qu'une table par localisation. Cette solution s'adapte aux spécificités de ce péril et n'est donc pas adaptée à d'autre périls. Les tremblement de terres à titre d'exemple, sont conçus sur une résolution moins fine mais possèdent un catalogue stochastique d'événements beaucoup plus dense. D'autres solutions seront envisagées pour l'application de cette approche sur des modèles très granulaire tels que les modèles inondations.

7 Conclusion

Les études effectuées tout au long de ce mémoire avaient pour objectif de comparer plusieurs méthodes permettant d'intégrer les modèles *CAT* à la tarification et à la souscription.
Nous avions porté notre choix sur le risque Ouragan appliqué au portefeuille mexicain.
Ceci dit, l'extension à d'autres périls et d'autres portefeuilles est possible à condition de disposer d'un modèle *CAT* permettant de récupérer les distributions de pertes brutes par événements et par sites.

La première approche suggérée consistait à construire des modèles de prime pure classiques en utilisant les résultats du modèle de catastrophes comme base de modélisation.
Cette approche permet l'élaboration d'un tarif clair, lisible et multiplicatif en utilisant des *GLM*. Pour reconstituer un tarif brut, il suffit alors de multiplier le taux de fréquence de base par les relativités du *GLM* de fréquence et faire de même pour le taux de sévérité de base et enfin multiplier les deux estimations pour obtenir le tarif brut.
L'introduction d'algorithmes de type machine learning aboutit à de meilleurs résultats. Cependant la complexité et la subtilité de ces algorithmes et de leurs paramétrages font que les *GLM* sont encore largement utilisés dans le monde de l'assurance.
La prime brute est ensuite convertie en nette au moyen de courbes de rabais. Cette méthode présente de nombreux avantages comme la lisibilité et la facilité de l'interprétation du taux de rabais, mais peut manquer de précision dans certains cas.

Les approches proposées par la suite visent à s'affranchir de ces courbes. Contrairement aux méthodes classiques de tarification dont le but est de modéliser une moyenne de perte, les travaux réalisés ont pour objectif de restituer la totalité de la distribution de pertes du site.

La première solution consistait à calibrer cette distribution au moyen de lois statistiques dont les paramètres seront à prédire dans une deuxième étape.
Cette technique repose donc sur des hypothèses fortes qui sont en réalité difficilement vérifiables. De plus, l'exercice est intéressant du point de vue théorique mais sa réalisation effective représente une contrainte opérationnelle de taille.

La seconde méthode est une solution informatique intelligente qui présente le modèle CAT sous un nouveau paradigme. Cette dernière approche offre à travers sa vision réaliste et complète du risque, toutes les dimensions nécessaires à l'optimisation du portefeuille de l'assureur. Elle présente néanmoins une grande contrainte en termes de lisibilité et partage des résultats. La construction du tarif et la prise en compte des conditions financières devient beaucoup plus obscure et moins lisible que l'approche actuelle.

Dans ce contexte et pour répondre aux contraintes opérationnelles, nous pouvons réfléchir à une solution permettant de maintenir l'approche classique pour une tarification de *"masse"* que l'on pourrait combiner à un second outil qui affinerait les hypothèses et l'évaluation des risques.

Références

[1] Catastrophes naturelles. http ://www.statistiques.developpement-durable.gouv.fr, Mis en ligne le 13 octobre 2016.

[2] Assurance des catastrophes naturelles. https ://www.ffa-assurance.fr/, Mis en ligne le 16 octobre 2018.

[3] Stefan Bernegger. The swiss re exposure curves and the mbbefd distribution class 1. *ASTIN Bulletin : The Journal of the IAA*, 27(1) :99–111, 1997.

[4] A. Charpentier. *Computational Actuarial Science with R*. Chapman & Hall/CRC The R Series. CRC Press, 2015.

[5] Marie Laure Delignette-Muller, Christophe Dutang, et al. fitdistrplus : An r package for fitting distributions. *Journal of Statistical Software*, 64(4) :1–34, 2015.

[6] Patricia Grossi, Howard Kunreuther, and Don Windeler. An introduction to catastrophe models and insurance. In *Catastrophe modeling : A new approach to managing risk*, pages 23–42. Springer, 2005.

[7] Canadian Institute of Actuaries. Incorporation of flood and other catastrophe model results into pricing and underwriting. www.casact.org/cms/pdf/IncorporationofFlood.pdf, July 2018.

[8] Swiss RE. Catastrophes naturelles et réassurance. www.swissre.com/sigma/, Mis en ligne le 16 septembre 2015.

[9] AIR Worldwide. Financial module. Internal documentation, January 2014.