Tarification à l'adresse en assurance habitation individuelle

Pierre Chatelain

UNIVERSITÉ DE LYON

Université Claude Bernard Lyon 1

THÈSE de DOCTORAT DE L'UNIVERSITÉ CLAUDE BERNARD LYON 1

École Doctorale n° 486
Sciences économiques et de gestion

Discipline : Actuariat

Soutenue publiquement le 12/04/2023, par :

Pierre Chatelain

Tarification à l'adresse
en assurance habitation individuelle

Devant le jury composé de :

Katrien Antonio Professeure d'Amsterdam, Pays-Bas et à l'université Leuven, Belgique	**Rapporteur**
Arthur Charpentier Professeur à l'université du Québec à Montréal, Canada et l'université Rennes 1	**Rapporteur**
Stéphane Loisel Professeur des universités à l'université Lyon 1	**Directeur de thèse**
Maria Mercèdes Claramunt Bielsa Professeure à l'université de Barcelone, Espagne	**Examinatrice**
Xavier Milhaud Maître de conférences Aix Marseille Université	**Co-Directeur de thèse**
Denys Pommeret Professeur des universités à l'université Aix Marseille Université et à l'université Lyon 1	**Président**
RACHDI Nabil Docteur en mathématiques, Addactis France	**Invité**

Remerciement

J'exprime toute ma reconnaissance aux membres du comité Pr Antonio, Pr Charpentier, Pr Mercè Claramunt Bielsa et Pr Pommeret. Je suis honoré que vous ayez accepté de lire et d'évaluer ce travail. Je remercie d'autant plus les rapporteurs Pr Antonio et Pr Charpentier qui ont accepté d'étudier ce travail malgré leur emploi du temps chargé.

Mes premiers remerciements s'adressent tout d'abord à mes directeurs de thèse, Pr Stéphane Loisel et Pr Xavier Milhaud ainsi que Nabil Rachdi. C'est à travers leurs encadrements et leurs encouragements que j'ai pu produire le document tel qu'il est aujourd'hui.

Il n'est pas possible de commencer les remerciements sans remercier Addactis France d'avoir accepté de financer cette thèse CIFRE avec l'aide de l'ANRT. Il me faut donc remercier les nombreux collaborateurs d'Addactis qui m'ont accompagnés durant ces plus de 3 ans de thèse. Merci notamment à Guillaume Rosolek pour sa confiance et permis de travailler sur le projet SHOP. Par ordre alphabétique, je souhaite citer les personnes de l'équipe P & A qui ont très largement contribué à l'enrichissement technique de cette thèse. Merci Arshak pour geopanda. Merci à Bilal Sadou et sa sœur Lou Absidal. Merci à Bruno, mon marathonien préféré. Merci à Cédric - le meilleur des alternants. Merci à Franck - pour ses bières sur le billard. Merci à FX de m'avoir fait connaitre Addactis. Merci à Linda pour les flocons d'avoines. Merci à Marie - promotrice immobilière prometteuse. Merci à Médérick pour sa bonne humeur et ses problèmes mathématiques. Merci à Montassar pour son management même dans les moments difficiles. Merci à Quang - le maitre Yoda du *data.table*. Merci à Silvia - pour ton goût du vin et du fromage italien. Merci à Victoria - le taux maximal non-vie est maintenant à 0.47%. Merci à Yasser - pour ton talent et ta bonne humeur. Merci aussi à Margot, Marketa, Mathilde, Simon, Yu, Junior pour votre travail.

Cependant, Addactis, c'est aussi des rencontres avec des personnes toutes plus intéressantes que les autres. Vous êtes bien trop nombreux pour tous vous citer mais merci notamment à Romain Gauchon de nous avoir guidé avec Tachfine sur ces premiers instants de thèse. Merci à Pascal et Brahim de leur réactivité et leur bonne humeur, sans vous pas de thèse. Merci à Eve, et son chanteur préféré David Castello-Lopes. Merci à Michaël - l'ange qui dépeint de si belles blagues.

Mes remerciements vont aussi à tous les doctorants et post-docs dont les années Covids ne nous ont pas permis de partager autant de moments ensembles que souhaitée en particulier Amal, Behzad, Etienne, Haïfa, Karim, Léonie, Marwa, Mathias, Natalia, Nesrine, Pierre, Pierrick, Rayane, Sarah B., Sarah M. et Sarra G. Il n'est pas possible de terminer ces remerciements sans citer le maître incontesté de l'IFRS 17, des CSMs ou des RAs, déléguée des doctorants Tachfine El Alami, mon jumeau de thèse. Je pense que tu sais que je ne peux trouver des mots suffisamment forts pour te remercier sincèrement.

Un merci infini à toute ma famille, mes grands-parents, mes parents, mes frères, mes cousins, mes oncles et tantes. Merci aussi à mes amis d'ici ou d'ailleurs, d'école ou d'enfance, de jeu ou de bar. Une pensée pure, profonde, pétillante et particulière pour, toi, Julie, de m'avoir soutenu, supporté durant ces quelques années d'excursion intellectuelle.

D'ailleurs en ce temps léthargique,
Sans gaîté comme sans remords,
Le seul rire encore logique,
C'est celui des têtes de morts.

此外，在个慵的代，
没有笑和悔恨，
唯一得通的笑声，
是死人的那个。

Verlaine, Claire Lenoir de Villiers de l'Isle Adam.

Résumé

Cette thèse étudie les problématiques d'assurance habitation individuelle (MRH) basée sur la géolocalisation, appelée tarification à l'adresse. Des solutions sont proposées pour la prise en compte de la qualité de données dans les modèles et une méthodologie adaptée à la tarification à l'adresse est développée pour les différents risques liés aux habitations. Cette thèse est composée d'une introduction générale et de sept chapitres. L'ensemble des chapitres résultent des travaux réalisés avec plusieurs assureurs lors de la mise en place de la tarification à l'adresse. Ces recherches ont permis le développement du produit nommé *Smart Home Pricing* **SHoP**.

Le chapitre 1 détaille le principe de la tarification en assurance non-vie et sa déclinaison opérationnelle et mathématique. Le chapitre 2 présente les variables produits par un fournisseur de données, indispensables pour mettre en place la tarification à l'adresse. Succinctement, le chapitre 2 revient sur les processus de géolocalisation et ses limites ainsi que sur les types de données disponibles à l'aide de quelques exemples avec leurs atouts et leurs obstacles.

À partir d'un portefeuille d'assureur sur la France métropolitaine, le chapitre 3 propose une méthodologie adaptée à la tarification à l'adresse et au processus d'obtention des données et de leur qualité. Sur les sinistres de hautes fréquences et de montants modérés, deux résultats intéressants se profilent. Tout d'abord, les données à l'adresse permettent d'améliorer significativement les modèles traditionnels. Pour des performances équivalentes, elles permettent de remplacer les questionnaires de souscriptions sauf l'âge de l'assuré et la formule d'assurance choisie.

Les chapitres 4, 5 et 6 s'intéressent à la problématique de la crédibilité des données afin d'intégrer les indices de qualité disponibles et aussi afin d'évaluer l'impact des erreurs de géolocalisation. Une structure sur la qualité de donnée est développée pour intégrer des indices de qualités individualisés dans les modèles linéaires et les modèles linéaires généralisés. Sous certaines hypothèses, plusieurs remarques sont faites. L'évolution des coefficients s'explique en partie par l'évolution de la performance du géocodage. La perte en qualité des données réduit la performance des modèles sous certaines hypothèses. La structure de qualité définie une relation d'ordre. Finalement, la qualité des données impacte et complexifie l'anti-sélection dans un monde concurrentiel.

Le chapitre 7 s'intéresse aux conséquences de l'ajout de nouvelles données sur la prise en compte du risque de subsidence en assurance habitation en France. Après un état de l'art des méthodologies pour modéliser les risques climatiques, un article questionne l'assurabilité des sécheresses et des évolutions récentes de la législation à partir de données marchés d'un assureur.

Abstract

This thesis deals with the development of an individual home insurance product based on geolocation, called address-based pricing or *Smart Home Pricing* abbreviated **SHoP**. This thesis is composed of five chapters and a general introduction allowing to put in perspective the work in front of the existing product. All the chapters are the result of work with several insurers during the implementation of pricing using the address.

The chapter 1 reviews the principle of pricing in non-life insurance and its operational and mathematical application. The chapter 2 presents the essential elements to set up the pricing by address provided by a data provider **Nam.R**. In a broad and non-exhaustive way, the chapter 2 returns on the processes of geolocalisation and its limits as well as on the types of data available with their advantages and disadvantages using some examples.

The chapter 3 proposes a methodology adapted to the data, to the process of obtaining the data and their quality based on a portfolio of insurers in metropolitan France. For high frequency and moderate amount claims, two interesting results are presented : the address data allow to improve greatly the traditional models and it is possible to obtain equivalent results using only the address data, the age of the person and the coverage chosen.

Chapters 4, 5 and 6 follows directly from chapter 3 as it addresses the issue of data credibility for integrating the quality indices available in SHoP but also assesses the impact of geolocation errors. In developing a data quality framework for incorporating individualized quality indexes into linear models and generalized linear models, four complements hypothesize the following points. Lower data quality reduces model performance. The evolution of the coefficients can be explained as a function of the evolution of the geocoding performance. An order relationship can be established on the quality of the variables. Data quality impacts and complicates anti-selection in a competitive world.

Finally, the chapter 7 looks at the consequences of the addition of new data on the consideration of drought/subsidence risk in household insurance in France. After a state of the art of the methodologies put in place to model climatic risks, an article based on market data from an insurer questions the insurability of droughts and recent changes in the legislation.

Table des matières

0 Introduction Générale **1**
- 0.1 Contexte du sujet . . . 1
- 0.2 Contributions de cette thèse : . . . 2

1 Pratique du marché actuel et des outils actuariels **9**
- 1.1 Le marché Multi-Risques Habitations (MRH), ses contraintes opérationnelles et son fonctionnement . . . 9
- 1.2 Les outils statistiques et méthodologies utilisés . . . 34

2 Géolocalisation et les données à l'adresse **49**
- 2.1 La géolocalisation à partir d'une adresse . . . 49
- 2.2 Les informations externes . . . 59

3 Produit MRH à partir de la tarification à l'adresse **69**
- 3.1 Le contexte du projet . . . 69
- 3.2 La problématique de la géolocalisation et le choix de la base de modélisation . . . 74
- 3.3 La sélection des variables et des modèles . . . 79
- 3.4 Les résultats sur l'ajout de données externes . . . 92

4 Modèles linéaires et la qualité de données **99**
- 4.1 Introduction . . . 100
- 4.2 Integration of data quality . . . 101
- 4.3 Estimation process with quality indexes . . . 102
- 4.4 Theoretical results . . . 103
- 4.5 Simulations with data quality inputs . . . 108
- 4.6 Real-life applications . . . 113
- 4.7 Proofs . . . 118
- 4.8 Multivariate case . . . 122

5 Modèles linéaires généralisés et la qualité de données **125**
- 5.1 Introduction . . . 125
- 5.2 Data problems and imputation . . . 126
- 5.3 Frameworks studied . . . 129
- 5.4 Adapting the prediction to data quality . . . 134
- 5.5 Simulation study - M1 estimator . . . 136
- 5.6 Discussion . . . 139
- 5.7 Theoretical framework . . . 143
- 5.8 Various operational remarks . . . 149
- 5.9 Proofs . . . 151

6 Compléments sur la qualité de données et le lien avec la tarification à l'adresse **163**
- 6.1 Des compléments sur la qualité de données . . . 163
- 6.2 Le lien avec la théorie des valeurs manquantes . . . 172
- 6.3 La qualité des données et la tarification à l'adresse . . . 177

7 Données à l'adresse : application à la sécheresse **187**
7.1 La modélisation des risques climatiques . . . 187
7.2 L'assurabilité des risques de subsidences . . . 193
7.3 Data and subsidence . . . 194
7.4 Modeling the CatNat frequency, claim cost and legal declaration . . . 199
7.5 Reserving, prevention and insurability statement . . . 210
7.6 Conclusion . . . 214

8 Conclusion et perpectives **219**

Appendices **221**

A Quelques commentaires sur l'histoire de l'assurance habitation **223**

B Comparaison du modèle de coût sur une base par sinistres individuels ou par sinistres moyens **225**

C Détection de changement de distribution : graves **227**
C.1 La théorie des valeurs extrêmes . . . 227
C.2 QQ plot . . . 228
C.3 Mean excess plot . . . 229
C.4 Estimateur d'Hill . . . 230
C.5 Gertensgarbe ou test Séquentiel de Mann-Kendall itératif . . . 233
C.6 Structure attritionnelle apprenante des graves . . . 238

D Différentes approches pour les Ginis **239**
D.1 La courbe de Lorenz et de concordance . . . 239
D.2 L'AUC et les Ginis . . . 240

E Exemple de calculette tarifaire **241**

Chapitre 0

Introduction Générale

0.1 Contexte du sujet

En France, la loi Quillot de 1972 (article 1733 du Code civil) et la loi Méhaignerie du 6 juillet 1989 ont rendu obligatoire la souscription d'une assurance habitation locative avec les quatre garanties suivantes : Incendies, Vols, Dégâts Des Eaux et Bris de Glace. De part ce caractère obligatoire, l'assurance habitation individuelle est un des piliers de l'assurance non-vie. Selon la FFA (ou FA [1]) en 2020, 117 milliards d'euros de cotisations sont alloués à la protection des biens particuliers soit 19.5 % du marché non-vie. Pour les particuliers en moyenne, 30.1 % des cotisations d'assurances (hors vie et prévoyance) sont allouées à l'assurance habitation [2]. En ce moment, le marché habitation se trouve dans une phase de croissance annuelle lente entre 2 à 3 %. Selon la FA, avec 45 millions de contrats, l'assurance habitations individuelle (MRH) voit sa fréquence et ses coûts moyens augmenter significativement surtout en TGN (Tempête Grêle Neige) avec plus de 200 % d'augmentation entre 2021 et 2022 et pour les garanties moins climatiques comme l'incendie, le vol et les dommages électriques. Seuls les dégâts des eaux et la responsabilité civile tendent à diminuer en fréquence. Récemment, l'augmentation des coûts moyens est largement portée par l'inflation comme montre l'accroissement de l'indice FFB [3] (entre 8 et 10 % en 2022).

Avec un ratio sinistres sur primes aux alentours de 54.6 % en 2021 et un ratio combiné de l'ordre de 98 %, la concurrence entre les assureurs s'est largement accentuée ces dernières années. Le marché français MRH (Multi-risques habitation) est un marché plutôt stable avec un taux de résiliation entre 11.9 et 13.4 % et un taux d'affaires nouvelles entre 12.9 % et 15.7 % sur les dix dernières années. Pour accroître leurs portefeuilles, de plus de 3 %, les acteurs se doivent soit de réduire leur propre taux de résiliation ou accroître leur nombre d'affaires nouvelles au détriment de leurs concurrents. Un peu plus de 80 % du marché est détenu par 10 acteurs et aucun d'entre eux n'a d'évolutions dépassant largement le nombre de nouveaux contrats. Cependant, la volatilité de certains risques comme les risques climatiques ou catastrophes naturelles peuvent être contrôlés lors de la souscription. Plus que l'accroissement du chiffre d'affaires, c'est la sélection et le refus de certains assurés à travers la connaissance des risques qui deviennent primordiaux.

Dans ce climat concurrentiel, la connaissance du risque se confronte à de nombreux aspects marketing et client. Avant l'avènement d'internet, les données provenaient exclusivement des informations de souscriptions. Depuis, les assureurs veulent se différentier par l'utilisation de données nouvelles et disruptives. Néanmoins, leur caractère parcellaire, leur coût et leur disponibilité sont des freins à leur intégration opérationnelle. De plus en plus de données sont disponibles grâce aux efforts constants et fructueux de la législation française. C'est pourquoi la tarification à partir de données externes commence à émerger, mais de nombreuses problématiques sont à résoudre.

L'obtention de données externes est limitée et nécessite des investissements non négligeables. La motivation est avant tout financière, mais ne doit pas désorganiser lourdement les activités de services assurantiels. À partir du moment où l'assureur utilise des données externes, les erreurs dans la donnée sont de sa responsabilité alors qu'avant elles étaient de l'assuré ! Pour l'instant, leur utilisation est cachée dans des indicateurs de risques ou des variables géographiques - comme les zoniers.

Pour clôturer ces quelques lignes de contexte, cette thèse considère l'intégration des données externes

1. France Assureur.
2. L'assurance automobile avec l'assurance habitation représente 85.3 % du montant des primes individuelles.
3. Fédération Française du Bâtiment

sous l'angle de la tarification actuarielle au sens large. D'autres services sont aussi fortement impactés par l'ajout de données externes et les utilisent plus aisément. En effet, les services Solvabilité II ou de Réassurances peuvent se permettre d'incorporer des données externes sans que cela impacte les processus B2C [4].

0.2 Contributions de cette thèse :

Avant de s'intéresser aux aspects précis de l'assurance habitation, cette thèse résume dans un premier temps les pratiques marchés en tarification non-vie en ces débuts d'années 2020. Les procédures opérationnelles de tarifications sont souvent obscures pour les personnes assurées. Un premier apport de cette thèse est d'expliquer concrètement les tenants et aboutissants de la tarification d'un contrat d'assurance. La plupart des livres ou des articles sur la tarification discutent peu des choix opérationnels des différentes méthodes (Werner et Modlin 2010, [21] ou Esbjörn et Björn, 2010 [13]). Pour être adaptée à toutes les législations et spécificités des pays, la tarification y est simplifiée, connaissant les caractéristiques $\mathbf{X}$ du contrat c, à l'évaluation de la sinistralité S espérée sur un an : $\mathbb{E}(S(c)|\mathbf{X}(c))$.

Dans une première partie, les données et leur provenance ainsi que l'historique de sinistralités utilisé servant à la modélisation sont dépeints. Dans la seconde partie, nous discutons comment, pour des raisons actuarielles et législatives, une prime d'assurance est calculée. Une prime d'assurance est assujettie à des contraintes comportementales. Le niveau de segmentation d'une prime doit être suffisamment élevé pour éviter l'anti-sélection et gérer le niveau d'aléa moral. La littérature scientifique questionne la différentiation entre mutualisation et hyper-segmentation. La compréhension de ces notions sont très importantes surtout lors d'ajouts de données individualisantes. Dans certains cas, lorsque le risque est trop individualisé, la question de l'assurabilité se pose.

Ensuite, l'arsenal mathématique dont dispose un actuaire est décrit. L'article de Wuthrich et Buser 2021 [22] présente très bien l'ensemble des méthodes statistiques disponibles en les appliquant sur les données en télématique. Les modèles linéaires généralisés (GLM) sont les plus utilisés grâce à leurs simplicités, leurs versatilités et leurs interprétabilités. [5] L'inclusion de zoniers en lien avec les données externes fera l'objet d'explications détaillées. Récemment, des méthodes dites Machine Learning ont vu leur popularité s'accroitre même s'il existe des débats sur leur utilisation. De mon point de vue, leur rôle dans un processus tarifaire est d'accompagner le modélisateur pour accélérer ou améliorer la modélisation à l'aide de GLM. Pour finir, cette thèse explique les problématiques de sélections de variables ou de bases de modélisations.

0.2.1 Création d'un produit d'assurance se basant sur les données externes

Depuis longtemps, les données externes ont été incorporées : notamment dans les zoniers. Récemment de nombreux travaux sur les données télématiques ont été développés en assurance automobile. L'idée est d'ajouter un boitier à une voiture qui envoie des informations sur la conduite. Ces données externes à la souscription permettent d'adapter la tarification, *Pay-as-you-drive* ou *Pay-how-you-drive* (Tselentis et al., 2016 [18]). Les aspects opérationnels, d'élasticité aux prix, de prise en compte de données externes ont fait l'objet d'études approfondies par Henckaerts et Antonio 2022 [7]. Les données à l'adresse sont l'équivalent logique des données télématiques mais appliquées à l'assurance habitation individuelle. Des tentatives d'utilisations de ces données dans certaines assurances ont été faites, mais aucune n'a encore abouti. De plus, aucun article ne fait état de l'usage des données aux bâtiments à partir de la géolocalisation de celui-ci.

La principale contribution actuarielle de cette thèse est de démontrer l'intérêt des données externes aux bâtiments sur la connaissance des risques habitations. Cette contribution se divise entre une étude attentive des données et une méthodologie pour prendre en compte la qualité des données et du géocodage. Deux questions se posent : Quelles informations utilisables pour la tarification d'une prime, apportent les données à l'adresse ? Est-ce que les données à l'adresse peuvent remplacer les questionnaires de souscriptions ?

Afin de permettre au lecteur de comprendre la difficulté de l'exercice, le chapitre 2 parcourt le processus de géolocalisation d'un bâtiment à partir d'une adresse. La géolocalisation est un processus difficile à industrialiser et surtout le taux de réussite est différent spatialement. La bonne connaissance

4. *Business to Client*, les échanges l'assureur envers les assurés.

5. D'après Guven et Werner, 2007 [20], les GLMs ont vu leur popularité croitre dans les années 1990. D'ailleurs, cet article explique simplement les avantages opérationnels des GLMs.

du processus de géolocalisation permet d'anticiper les impacts sur les modélisations. Dans un second temps, il est nécessaire d'expliquer succinctement les types de données externes à disposition à travers d'exemples. Ces derniers permettent à tout à chacun de comprendre les particularités, les erreurs et le type d'informations accessibles pour l'évaluation des risques.

Le chapitre 3 est un résumé des premiers travaux de modélisations sur les sinistres attritionnels appliqués à un portefeuille d'assurance habitation individuelle. La méthodologie développée a permis de comprendre les données, les corriger et de valider leur intérêt actuariel. Les résultats finaux montrent que les sinistres attritionnels en habitation individuelle sont significativement mieux appréhendés à l'aide des données externes à l'adresse qu'avec des données open data. Même si la performance des données externes permet de limiter, à performance égale, la souscription aux questions sur l'âge et sur les couvertures voulues, les modèles pour une souscription plus rapide peuvent avoir des conséquences négatives à cause de l'anti-sélection induite.

0.2.2 Intégration de la qualité des données dans la modélisation

Dès le lancement du projet, le fournisseur de données a procuré des indices de qualité associés à la donnée envoyée. Les premières interrogations qui sont apparues étaient : Que mesure cet indice de qualité ? Comment peut-on intégrer de tels indices dans la modélisation ? L'intégration de la qualité des données dans la modélisation est un sujet vaste. Tout d'abord, la notion de qualité est multivariée (Rogova et Bossse 2010 [14], Todoran et al. 2014 [17]) avec plusieurs dimensions : imprécision, complétude, temporalité et crédibilité. Sur la complétude ou l'imprécision des données, la littérature est riche de solutions et de modèles. Pour les problèmes de complétions, les travaux (par exemple : Van Buuren, 2018 [19], Little et Rubin 2019 [10]) se basent souvent sur les définitions suivantes MCAR (*Missing Completely At Random*), MAR (*Missing At Random*), or MNAR (*Missing Not At Random*). Pour la prise en compte d'imprécisions, des *Errors-In-Variable* modèles Van Huffel et Lemmerling [115] ou des modèles bien spécifiques, Trabelsi et al. 2016 [113], Tami et al. 2018 [16] essayent de les prendre en compte.

Cependant, pour la qualité de données basées sur la crédibilité des données, de rares méthodes essayent de la prendre au compte comme RANSAC (RANdom SAmple Consensus, Fischler et Bolles 1981[4]) et ses extensions, mais celles-ci n'adaptent pas les résultats face à la qualité individuelle. Une contribution statistique de cette thèse est de mettre en avant une structure sur la qualité de données pour comprendre les impacts de la crédibilité de la donnée sur les modèles linéaires ou les modèles linéaires généralisés. Notons $\mathcal{M}_{n\times(p+1)}(I)$ l'ensemble des matrices de taille $n\times p+1$ dont tous les éléments appartiennent à I. La structure s'inspire des recherches de Julie Josse sur les valeurs manquantes :

$$\mathbf{X} = \mathbf{X}^{real} \circ \mathbf{\Omega} + \mathbf{Z} \circ (J_{n,p+1} - \mathbf{\Omega})$$

où :

- $\mathbf{X} = (X_{ij}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ sont les données disponibles mais de qualité imparfaite ;
- $\mathbf{X}^{real} = (X_{ij}^{real}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ sont les données idéales, de qualité parfaite, mais inconnues ;
- $\mathbf{Z} = (Z_{ij}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ ce sont les erreurs que l'on suppose être distribuées comme $\mathbf{X}^{real}$;
- $\mathbf{\Omega} = (1, \Omega_j)_{j=1,\dots,p} = (\omega_{ij}) \in \mathcal{M}_{n\times(p+1)}(0,1)$ une matrice binaire (composée de 0 et de 1) qui vaut 1 en i et j quand l'élément $\mathbf{X}_{ij}^{real}$ est observé.
- $\circ$ au produit d'Hadamard (terme à terme) ;
- $J_{n,p+1}$ est la matrice identité $n \times (p+1)$ pour la multiplication d'Hadamard (composée que de 1).

Cependant, $\mathbf{\Omega}$ est inconnue en pratique sinon ce serait trop simple ! De ce fait, l'hypothèse est que $\mathbf{\Omega}$ est une variable aléatoire. Les indices de qualité de données observées $\mathbf{Q} = (1, Q_j)_{j=1,\dots,p} = (Q_{ij}) \in \mathcal{M}_{n\times(p+1)}([0,1])$ sont supposés être l'espérance de $\mathbf{\Omega}$ c'est-à-dire :

$$\mathbb{E}(\omega_{ij}) = \mathbb{P}(\omega_{ij} = 1) = Q_{ij}.$$

Il est nécessaire de mettre en avant différents cas de figures. Des hypothèses entre les corrélations entre les variables sont faites et justifiées par les différents cas observés lors du projet de la tarification à l'adresse. Les principaux théorèmes sont prouvés sous des hypothèses simples. À l'aide de cette structure, l'article **Efficient algorithm with quality index : application to linear regression** co-écrit avec Xavier Milhaud - chapitre 4 - propose un algorithme pour obtenir les modèles qui ne sont pas biaisés par la qualité de la donnée. Il permet aussi de prendre en compte la qualité de donnée de chaque individu dans les prédictions du modèle. Deux méthodes statistiques sont étudiées : les modèles linéaires dans le précédent article et les GLMs dans l'article **Integrating data quality into GLM for insurance pricing**

- chapitre 5. Notons $\hat{\beta}$ les coefficients obtenus . L'exposant M_2 réfère au modèle appris sur les données $\mathbf{X}$ et son absence sur les données $\mathbf{X}^{real}$. L'article 4 se place dans le cas où $\mathbf{\Omega}$, $\mathbf{Z}$ et $\mathbf{X}^{real}$ sont indépendants (cas nommé (C1) ensuite), sous les hypothèses que les marginales $\mathbf{X}^{real}$ sont *i.i.d* notée (X-A1), que les erreurs $\mathbf{Z}$ sont indépendantes notée (Z-A1) ou gardent la même structure de corrélation que $\mathbf{X}^{real}$ notée (Z-A2). Notons $\bar{Q}_j$ pour tout $j = 1, ..., p$, la moyenne des indices de qualité. Le théorème suivant donne une simple relation proportionnelle des coefficients.

Théorème 0.2.1

Sous (X-A1) et (Z-A1) ou (Z-A2), pour tout $j = 1, ..., p$, quand $n \to +\infty$:

$$\hat{\beta}_j^{M_2} / \bar{Q}_j \to \beta_j, \quad \text{presque sûrement.}$$

L'intercepte n'est pas modifié, *c-â-d* $\hat{\beta}_0^{M_2} \to \beta_0$.

Plus les hypothèses se complexifient, plus les effets de la qualité des données sont difficiles à comprendre. Sous l'hypothèse que les marginales $\mathbf{X}^{real}$ sont corrélées deux à deux (X-A2), le théorème suivant peut être montré pour les modèles linéaires.

Théorème 0.2.2

Pour tout $j \neq k$, si $|\rho| = |\rho_{jk}^{real}| \neq 1$, les convergences presque sûres sont vraies sous (X-A2) quand $n \to +\infty$:

$$\text{Sous (Z-A1)} : \frac{1}{1-\rho^2}\left(\frac{\hat{\beta}_j^{M_2}}{\bar{Q}_j}(1-\rho^2\,\bar{Q}_j\bar{Q}_k) + \sqrt{\frac{Var(X_k)}{Var(X_j)}}\frac{\hat{\beta}_k^{M_2}}{\bar{Q}_k}\rho\,(\bar{Q}_j\bar{Q}_k - 1)\right) \to \beta_j,$$

$$\begin{aligned}\text{Sous (Z-A2)} : \frac{1}{1-\rho^2}\Big(&\frac{\hat{\beta}_j^{M_2}}{\bar{Q}_j}(1-\rho^2\,(1+2\bar{Q}_j\bar{Q}_k - \bar{Q}_j - \bar{Q}_k)) \\ &+ \sqrt{\frac{Var(X_k)}{Var(X_j)}}\frac{\hat{\beta}_k^{M_2}}{\bar{Q}_k}\rho\,(2\bar{Q}_j\bar{Q}_k - \bar{Q}_j - \bar{Q}_k)\Big) \to \beta_j.\end{aligned}$$

Dès que l'on s'éloigne des modèles linéaires, les relations entre les coefficients ne sont plus explicites. Il est nécessaire de recalculer la log-vraisemblance. Pour les modèles multiplicatifs (log-gamma et log-Poisson), les relations entre log-vraisemblance $log(\mathcal{L})$ s'écrivent bien. Par exemple, pour le modèle log-Poisson, $log(\mathcal{L}(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q}))$ peut s'estimer par :

$$\begin{aligned}\frac{1}{\bar{Q}_p}\Big[&log(\mathcal{L}^{M_2}(\hat{\beta}; \mathbf{Y}|\mathbf{X}) \\ &- (1-\bar{Q}_p) \times log(\mathcal{L}(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}^{real}_{(*p)})) \\ &- (1-\bar{Q}_p) \times \sum_{i=1}^{n} v_i e^{\hat{\beta}^{*p}\mathbf{X}^{real}_{i,(*p)}}(1 - M_{X_p}(\hat{\beta}_p))\Big],\end{aligned}$$

quand uniquement la dernière variable X_p a une qualité imparfaite et est indépendante des autres variables. $\mathbf{X}^{real}_{(*p)}$ représente les autres variables avec v_i l'exposition en offset de la ligne i et $M_{X_p}(.)$ la fonction génératrice des moments de X_p. Les cas plus complexes avec des corrélations entre les $\mathbf{\Omega}$, ou les corrélations entre les erreurs permettent d'étendre la structure de la qualité aux cas pratiques de la tarification à l'adresse. D'ailleurs, celle-ci sera le cas d'étude et de l'application de la structure de qualité de données.

Deux contributions actuarielles sont permises à l'aide de cette structure mathématique dans le chapitre 6.

La première contribution est de comprendre l'évolution des coefficients d'une régression linéaire lorsqu'il y a une amélioration du taux de géocodage. Cette contribution a permis de comprendre les évolutions observées dans le cadre de tarification à l'adresse. En fonction des hypothèses et des variables, les évolutions des coefficients peuvent être comprises par le type d'évolution du géocodage.

La seconde contribution est de comprendre l'anti-sélection qu'induit l'utilisation de données de qualité moindre dans un monde concurrentiel. L'ajout de variables externes de qualités imparfaites peut dans certains cas réduire la marge des compagnies l'utilisant. Cette contribution montre que même dans des cas simples sous des hypothèses simples, l'anti-sélection n'est pas toujours en faveur du modèle le plus segmentant. Dans un mode idéaliste, les assurés comparent et choisissent deux assureurs en fonction de la prime. Dans ce cas, l'assureur utilisant un modèle plus segmentant avec une variable ayant un problème de qualité peut avoir un profit négatif à cause de l'anti-sélection induit par la qualité de données. Même s'il est toujours avantageux d'utiliser de tels modèles dans la plupart des cas, l'utilisation de nombreuses variables avec une qualité de données imparfaite obligent à la plus grande prudence.

0.2.3 Étude de l'assurabilité des risques climatiques liée aux arrêtés CatNat avec un focus sécheresse.

L'évolution de l'assurance habitation a des impacts sociétaux importants en particulier à cause du risque climatique. Quel impact a l'augmentation des primes vis-à-vis des risques climatiques ? Nyce et al., 2015 [12] ont fait un état de l'art des différents impacts des primes d'assurances sur le prix des maisons. Tout d'abord, le prix des maisons est inélastique sur un court terme aux risques climatiques (Nyce, 1999). Des modèles microéconomiques récents comme anciens (Bin et al. en 2008 [1], Macdonald et al. en 1987 [11]) montrent que le changement de tarification impacte les prix immobiliers. Ces modèles prennent en compte la corrélation spatiale pour montrer l'influence des tarifs d'assurance sur le prix des maisons, les études sont essentiellement américaines souvent proche de Miami.

Chaque type d'assurance a des impacts sociétaux différents. Il semble que l'assurance tempête est plus porteuse d'informations que l'inondation. Gron et al. 1994 [5] et Doherty et Posey 1997 [3] ont observé que les arrivées soudaines de catastrophes naturelles impliquent une augmentation significative des primes d'assurances. Mais il n'y a pas d'impact direct puisque les primes se stabilisent sur le long terme. Harrison, Smersh, and Schwartz (2001) [6] en Floride expliquent que la différence entre les zones de risques existe, mais qu'elle ne correspond pas à la différence des primes d'assurances. De nombreuses études mettent en valeur la corrélation négative entre primes d'assurances et prix des maisons. Bin et al. 2008 [1] ont étudié la corrélation entre l'accès à la côte, l'assurance et le prix de la maison. En conclusion, l'impact des zones inondables est significatif. L'augmentation des prix pour s'approcher d'un degré d'une plage est de 995 dollars, mais le risque d'inondation réduit le prix de 11%. Zietz et al., 2008 [23] utilisent la base de données de Floride pour montrer la corrélation négative entre les risques et le prix des maisons. Quid de la France ? À ma connaissance, aucune étude n'existe pour l'instant. En effet, la garantie Catastrophe Naturelle (la plus grosse partie des risques dits climatiques) est tarifée comme 12% de la somme des primes de dommages pour tous les contrats.

Cependant, les primes d'assurances ne sont qu'une partie de la procédure de souscription. Avec l'augmentation de la fréquence des sécheresses en France, le risque de refus d'assurances est de plus en plus élevé. Le risque de subsidence suite à une sécheresse est un risque complexe à modéliser comme concluent Charpentier et al., 2021 [2]. Les sécheresses, leur intensité comme leur fréquence font l'objet de nombreuses études ([8], [9], [15]). En France, c'est le risque climatique le plus suivi depuis 2018 suite à de nombreuses années de sécheresses fréquentes et couteuses (le sinistre est de l'ordre en moyenne de 25 000 €). Selon France assureur, l'année 2018 est associée à 1,482 milliard d'euros de dommages. La sécheresse en cours (2022) est estimée entre 1,6 et 2,4 milliards d'euros pouvant dépasser le record de 2003 à 2,3 milliards d'euros. De plus, des évolutions importantes sur le régime CatNat ont été mises en place et le risque de subsidences a une place bien spécifique.

L'article **Subsidence and household insurances in France : geolocated data and insurability** du chapitre 7 contribue à mettre en avant une méthode de modélisation permettant d'utiliser les données à l'adresse et d'ensuite questionner l'assurabilité de ce risque. Les données à l'adresse permettent de capturer l'ensemble des informations géographiques de la sinistralité subsidence. Les résultats obtenus permettent de montrer des gains dans le cadre de la connaissance du risque pour du provisionnement.

L'envers du décor est que les modèles permettent de sélectionner précisément les habitations très fortement exposées avec une sinistralité espérée allant au-delà de 100 euros. Avec l'augmentation de la prise en charge des plus sinistrés du fait de la législation, nous mettons en garde sur l'assurabilité de ces habitations. Nous proposons de mutualiser par palier le régime CatNat pour que toutes les habitations restent assurables. Nous subdivisons la France en trois zones en fonction de la fréquence de retours, les habitations pourraient payer 12 % des primes dommages en zone 0, 14 % en zone 1 et 18 % en zone 2 pour le régime CatNat. Il serait nécessaire de refaire l'exercice pour les inondations ou les séismes.

L'idée est de maintenir la mutualisation spatiale sans que les assureurs soient tentés financièrement de refuser certains contrats.

Références

[1] Bin, O., Kruse, J. B., and Landry, C. E. (2008). Flood hazards, insurance rates, and amenities : Evidence from the coastal housing market. *Journal of Risk and Insurance*, 75(1) :63–82.

[2] Charpentier, A., James, M. R., and Ali, H. (2021). Predicting drought and subsidence risks in france. *Natural Hazards and Earth System Sciences Discussions*, 2021 :1–27.

[3] Doherty, N. and Posey, L. (1997). Availability crises in insurance markets : optimal contracts with asymmetric information and capacity constraints. *Journal of Risk and Uncertainty*, 15(1) :55–80.

[4] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6) :381–395.

[5] Gron, A. (1994). Capacity constraints and cycles in property-casualty insurance markets. *The RAND Journal of Economics*, pages 110–127.

[6] Harrison, D., T. Smersh, G., and Schwartz, A. (2001). Environmental determinants of housing prices : the impact of flood zone status. *Journal of Real Estate Research*, 21(1-2) :3–20.

[7] Henckaerts, R. and Antonio, K. (2022). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. *Insurance : Mathematics and Economics*, 105 :79–95.

[8] Ionita, M. and Nagavciuc, V. (2021). Changes in drought features at the european level over the last 120 years. *Natural Hazards and Earth System Sciences*, 21(5) :1685–1701.

[9] Jonathan, S., Gustavo, N., Jürgen, V., and Barbosa, P. (2016). Meteorological droughts in europe : Events and impacts – past trends and future projections. Technical Report EUR 27748 EN, Publications Office of the European Union, Luxembourg.

[10] Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

[11] MacDonald, D. N., Murdoch, J. C., and White, H. L. (1987). Uncertain hazards, insurance, and consumer choice : evidence from housing markets. *Land Economics*, 63(4) :361–371.

[12] Nyce, C., Dumm, R. E., Sirmans, G. S., and Smersh, G. (2015). The capitalization of insurance premiums in house prices. *Journal of Risk and Insurance*, 82(4) :891–919.

[13] Ohlsson, E. and Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*, volume 2. Springer.

[14] Rogova, G. L. and Bosse, E. (2010). Information quality in information fusion. In *2010 13th International Conference on Information Fusion*, pages 1–8. IEEE.

[15] Spinoni, J., Naumann, G., Carrao, H., Barbosa, P., and Vogt, J. (2014). World drought frequency, duration, and severity for 1951–2010. *International Journal of Climatology*, 34(8) :2792–2804.

[16] Tami, M., Clausel, M., Devijver, E., Dulac, A., Gaussier, E., Janaqi, S., and Chebre, M. (2018). Uncertain trees : Dealing with uncertain inputs in regression trees. *arXiv preprint arXiv :1810.11698*.

[17] Todoran, I.-G., Lecornu, L., Khenchaf, A., and Le Caillec, J.-M. (2014). Toward the quality evaluation of complex information systems. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*, volume 9091, page 90910N. International Society for Optics and Photonics.

[18] Tselentis, D. I., Yannis, G., and Vlahogianni, E. I. (2016). Innovative insurance schemes : Pay as/how you drive. *Transportation Research Procedia*, 14 :362–371. Transport Research Arena TRA2016.

[19] Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

[20] Werner, G. and Guven, S. (2007). Glm basic modeling : avoiding common pitfalls. In *Casualty Actuarial Society Forum, Winter*, pages 257–272. Citeseer.

[21] Werner, G. and Modlin, C. (2010). Basic ratemaking. In *Casualty Actuarial Society*, volume 4, pages 1–320.

[22] Wuthrich, M. V. and Buser, C. (2021). Data analytics for non-life insurance pricing. *Insurance : Mathematics and Economics*, pages 16–68.

[23] Zietz, J., Sirmans, G. S., and Smersh, G. (2008). The impact of inflation on home prices and the valuation of housing characteristics across the price distribution. *Journal of Housing Research*, 17(2) :119–137.

Chapitre 1

Pratique du marché actuel et des outils actuariels

1.1 Le marché Multi-Risques Habitations (MRH), ses contraintes opérationnelles et son fonctionnement

Si nous voulons modifier la manière de tarifer un produit d'assurance, il est nécessaire de comprendre les liens entre la tarification et les éléments de la chaîne opérationnelle assurantielle. Ces derniers se sont construits autour du fonctionnement des contrats d'assurances. Cette partie objective les contraintes opérationnelles. Les notions indispensables à la tarification seront mentionnées et explicitées dans le cadre de l'assurance habitation.

Attention :
Ce chapitre entreprend une description généraliste et représentative du marché français. Pour autant, chaque entreprise a un mode de fonctionnement différent. En effet, les types de souscriptions influencent grandement les méthodes pouvant être implémentées. Si nous nous aliènons du francocentrisme de cette thèse, les législations changent ; les assurances habitations ne sont pas toujours obligatoires, ne regroupent pas toujours plusieurs types d'assurances en même temps, sans même parler de RGPD (Règlement Général sur la Protection des Données). En fonction des pays, les processus et les données se transposent différemment, bouleversant les points d'attentions et les actions déployées. Les montants donnés sont en €.

1.1.1 Le processus de tarification et les données

Le processus de tarification s'intègre pleinement dans la chaîne de valeur assurantielle d'un contrat d'assurance. Cette section expose la provenance des données et les contraintes associées. Le graphique 1.1 résume les interactions entre les concepts actuariels et les bases de données utiles. Après avoir défini le contrat d'assurance dans la sous-section 1.1.1.a, le processus de souscriptions - sous-section 1.1.1.b sera détaillé. Les contraintes de la souscription sont détaillées dans la sous-section 1.1.1.c. Suite à l'acceptation du contrat, la vie de celui-ci est liée aux indemnisations de sinistres détaillées dans la sous-section 1.1.1.d. Les échanges d'informations entre B&C sont sauvegardés dans la base de contrats - sous-section 1.1.1.e et dans la base de sinistralités - sous-section 1.1.1.f. La jointure des deux bases aboutit à la création des bases d'études pour les travaux actuariels dont ceux de tarification. Cependant, les informations disponibles à date sont encore incertaines. Dès lors, il est nécessaire d'évaluer au mieux le coût final des sinistres en exploitant les éléments de provisionnements que la sous-section 1.1.1.g commente. Finalement, la sous-section 1.1.1.h met en avant les exigences que doit vérifier la tarification en lien avec les données utilisées.

1.1.1.a Un contrat assurantiel : un échange de risque payant et contractuel

L'objet social des assureurs est de pratiquer des opérations d'assurances *c-à-d* des services payants d'échanges de risques. Un prospect/client (**le souscripteur** potentiel) consulte un assureur pour qu'il

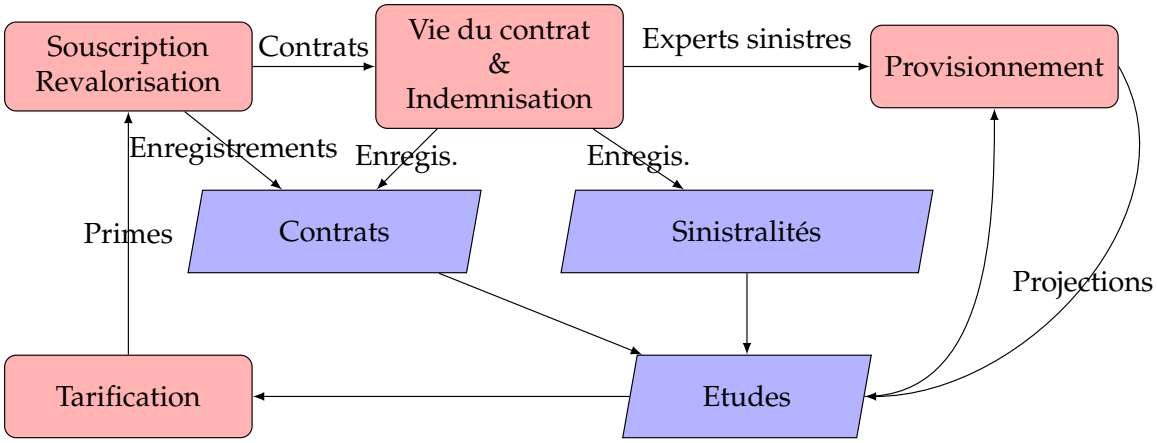


FIGURE 1.1 – Tarification et chaîne de valeur assurantielle. Les squircles rouges correspondent aux notions actuarielles et les trapèzes bleus aux bases de données.

endosse à sa place une perte probable (**risque**) sur une durée déterminée (**son exposition** au risque), contre un paiement monétaire (la **prime commerciale** ou **cotisation**). En France, cet échange se contractualise et s'organise à l'aide du droit de l'assurance. Les parties créent un **contrat** d'assurances. Seules les règles sur les contrats d'assurances les plus impactantes seront mentionnées.

Cette thèse se cantonne à l'assurance **IARD** (Incendie, Accident et Risque Divers) et plus particulièrement aux risques **MRH** (Multi-Risques Habitations). À noter que les directions d'actuariats gèrent conjointement l'ensemble des produits comme les produits habitations individuelles (**MRH**) , les produits habitations professionnelles (**MRP**) , les produits automobiles ou ceux des deux-roues. Par conséquent, les méthodes de tarifications entre les produits sont analogues. Un contrat **IARD** s'articule plus autour de la notion de **bien assuré** que celle de bénéficiaire [1]. En France, deux points communs entre les contrats **MRH** et **Autos** existent : ils possèdent une partie **obligatoire** et sont **multigaranties** [2]. Ces spécificités influencent grandement les caractéristiques et les comportements liés au contrat.

L'assurance habitation se scinde en deux groupes prédominants ; les contrats portant sur les **Logements individuels**, les **Logements collectifs** et un autre groupe plus annexe ; ceux sur les **Logements exceptionnels**. En substance, ils se résument à une distinction entre les maisons, les appartements et les autres types d'habitations. La barrière entre les deux premiers critères n'est pas clairement définie et elle dépend des assureurs. Pour les habitations exceptionnelles par exemple les châteaux, les manoirs, les maisons de retraites, les troglodytes, les assureurs composent avec ce type d'habitations *in concreto*. Ce type de produit n'est pas traité dans cette thèse.

Les obligations et les protections proposées divergent en raison de **la qualité du souscripteur**. Par conséquent, un produit MRH se subdivise généralement en trois classes : les locataires (**LOC**), les propriétaires (**PO**) et les propriétaires non-occupants (**PNO**). Ces obligations se retrouvent dans la législation française qui dresse les contours d'un contrat d'assurance.

Le Code des Assurances se subdivise en trois parties : la partie législative, la partie réglementaire et la partie des arrêtés complétant celle réglementaire. En particulier, la partie législative (Article L100-1 à L561-1) définit le contrat d'assurance dans les articles L100-1 à L195-1 et les assurances obligatoires dans la partie suivante (L200-1 à L271-1). Les arrêtés qui nous préoccupent sont les arrêtés portant principalement sur les catastrophes naturelles. À travers cet attirail législatif en assurance de dommages, plusieurs remarques transparaissent :

1. La majorité des articles législatifs régit les règles de souscriptions, de résiliations ou d'indemnisations. En second lieu, elles définissent les règles de provisionnement ;
2. Les règles de tarification des primes d'assurances sont quasiment inexistantes. En effet, les assureurs sont libres de choisir le prix des services (Article L410-2) selon les principes du jeu de la concurrence. Quelques limitations globales existent d'après l'article L111-7 et L111-8, sur les utilisations de critères de sexes (sauf si justifiable actuariellement), de grossesse, de maternité ou le don d'organes ou de gamètes. Les principes de calcul des primes associées aux catastrophes

1. Sauf pour la Responsabilité Civile en MRH et en Auto.
2. Les contrats d'assurances français MRH ou automobiles assurent le bien sur plusieurs risques en même temps d'où la notion de multirisques/multigaranties.

naturelles et aux attentats sont fixés par la législation. En assurance automobile, le principe des bonus-malus ainsi que les critères à considérer sont fixés par la législation et impactent directement la tarification. Un dernier article contraint les évolutions de primes en permettant à l'assuré de solliciter une réduction du montant de la prime si son risque diminue. Ce sont les seules règles de tarification IARD qui existent ;

3. Il n'existe pas une agence gouvernementale contrôlant principalement les aspects de la tarification. Le Bureau Central de Tarification (**BCT**) s'occupe uniquement des personnes qui n'ont pas pu souscrire un contrat obligatoire sur la demande du prospect. Les attributions du BCT ne concernent pas les montants des primes proposées ou les critères de tarifications. Seule l'Autorité de Contrôle Prudentielle et de Résolution (ACPR) a un mandat pour contrôler la tarification à travers les fonctions clefs actuarielles.

Sommairement, rien n'empêche légalement un actuaire de proposer un tarif extrême avec des critères extrêmes. Inversement, un tarif unique pourrait être proposé sans nécessiter une direction de tarification. Toutefois, la législation contraint d'autres aspects d'un contrat d'assurance pendant sa durée de vie (provisionnement, souscription, indemnisations. . .). Dans ces conditions, la tarification est contrainte indirectement. D'autres contraintes comportementales restreignent aussi la flexibilité des tarifs.

En assurance, une portion de la prime est associée aux indemnisations futures, encore inconnues. Le **cycle inversé de production** [3] caractérise les contrats d'assurances. Les parties du contrat sont obligées d'échanger des informations pour le bon déroulement du contrat. Ces échanges d'informations sont traduits informatiquement en données et sont les briques élémentaires de la tarification actuarielle.

1.1.1.b La souscription d'un contrat d'assurance

Les premiers échanges interviennent dès la souscription. La souscription débute lorsqu'un **potentiel souscripteur/prospect** entreprend des démarches pour souscrire auprès d'un assureur. La souscription se clôture par l'acceptation ou le refus de contractualisation par l'un des parties.

Comment se déroule la souscription d'un contrat habitation individuelle ?

Différents **canaux de distributions** co-existent pour qu'un client puisse souscrire un contrat d'assurances, physiquement chez un agent général, un courtier d'assurances, dans une banque ou virtuellement par internet. Dans le dernier cas, plusieurs plateformes de souscription existent : le site d'un assureur, ceux des comparateurs d'assurances ou des courtiers. Chacun des canaux opère différemment, néanmoins le processus de souscription reste sensiblement le même [4]. Financièrement, les frais et les chargements sont différents.

Id client	Canaux de distribution	Produit	Nom	Prénom	Age	Adresse	Garanties	Réponses
153467	26	MRH 2019	Dupont	Jade	43	...	...	...
154309	26	MRH 2019	Martin	Léo	67	...	...	...

Informations obligatoires

Réponses aux questions (du questionnaires, du commercial,...)

FIGURE 1.2 – Transformation des informations envoyées par l'assuré en données tabulaires.

Etape 1 : Le client répond aux questions soit par l'intermédiaire d'un questionnaire, soit aux questions du commercial. L'enregistrement de ces informations correspond à la création d'une nouvelle ligne (Figure 1.2) dans la base des devis. En **MRH**, les informations sont de plusieurs types : les informations de gestions (identifiants de l'assuré, le(s) canal de distributions, le produit proposé sont nécessaires pour les créations des bases), l'identité de l'assuré (Nom, Prénom, Age), les informations pour identifier son bien (en MRH, l'adresse du bâtiment et la plaque d'immatriculation en assurance automobile). Obligatoires, ces informations assurent l'unicité du bien assuré. D'autres questions pour un objectif tarifaire concernent les éléments assurés : le type de logement, la qualité de l'occupant, les options souscrites, les caractéristiques du bâtiment et la sinistralité passée.

3. C'est-à-dire un produit de l'assurance est vendu avant que l'entreprise ne connaisse le coût définitif, contrairement à un constructeur de téléphones.
4. Certains processus utilisent plusieurs canaux en même temps.

Etape 2 : À partir de ces informations, l'assureur communique le montant de la prime au commercial (s'il existe) et la propose au potentiel souscripteur. L'outil de calcul sous-jacent est appelé **calculette tarifaire**. Dans certains cas, le commercial peut aussi proposer une ristourne. Après quoi, le souscripteur peut accepter ou non le contrat. La base finalement constituée ressemble à la Figure 1.3 qui s'appelle la base des devis. Informatiquement, les données sont sous forme tabulaire. Cette base permet notamment de calculer le **taux de transformation**[5]. Ce taux est utilisé pour positionner un tarif par rapport à la concurrence ou contrôler les refus de souscriptions d'autres assureurs.

Id client	Canaux de distribution	Produit	Nom	Prénom	Age	Adresse	Garanties	Réponses	Prime proposée	Prime acceptée	Transformation
153467	26	MRH 2019 maison	Dupont	Jade	43	...	...	...	110, 00	110, 00	Oui
154309	26	MRH 2019 maison	Martin	Léo	67	...	...	...	80,00	Na	Non

FIGURE 1.3 – Version simplifiée d'une base des devis. Les devis transformés sont nommés **affaires nouvelles** (**A.N.**).

1.1.1.c Les contraintes opérationnelles de souscriptions

De nombreuses contraintes apparaissent dès la souscription. Les primes d'assurances habitations se paient soit annuellement, soit semestriellement, soit trimestriellement, soit mensuellement. Cette prime doit aussi être constante sur l'année. Une fois que la prime est acceptée, celle-ci ne peut être ajustée qu'annuellement.

Contrainte 1: Constante annuellement

Une prime d'assurance doit être constante dans le temps lors de la période de couverture.

Il est important de rappeler que, même si l'assurance habitation est obligatoire, le souscripteur n'a aucune obligation de souscrire le contrat proposé auprès d'un assureur en particulier. Pour l'assureur, les primes sont les éléments principaux déterminant la rentabilité. Il est impératif que le produit permette de dégager une marge. Pour être précis, pour certains **produits dits "d'appels"**, il est possible que la marge soit légèrement négative pour attirer de nouveaux clients sur d'autres produits. Ainsi, l'estimation de la rentabilité d'un contrat en amont est nécessaire pour la tarification et pour la contraindre.

Contrainte 2: Borné inférieurement

Le montant d'une prime d'assurance est bornée inférieurement en fonction du coût estimé du contrat.

Une contrainte opérationnelle non négligeable concerne la durée du processus de souscription. Le temps disponible pour la souscription d'une assurance est limité, d'autant plus sur les canaux internet. Lors d'une souscription dans une agence, le commercial compose avec un temps limité pour vendre un produit d'assurance, surtout quand d'autres produits sont à proposer. En conséquence, le nombre de questions est limité. De même, la complexité et la pertinence des questions sont adaptées. Ainsi, le souscripteur doit proposer des réponses sans ambiguïté ou sans chercher dans son acte de vente par exemple[6]. Un exemple récurrent concerne la date de construction d'un bâtiment ; les locataires connaissent rarement cette information. Le non-respect de cette contrainte provoque des abandons lors de la souscription plus fréquents ou des réponses équivoques ou erronées. De surcroît, les questions sont toujours fermées ou à choix multiples. D'autre part, le calcul de la prime doit se faire de façon "instantanée" du point de vue du souscripteur. Si elle prend plusieurs secondes, en particulier pour les souscriptions internet, le taux de devis terminés est susceptible de chuter considérablement.

5. Proportion de contrats qui ont été acceptés sur le nombre de contrats proposés.
6. À noter que l'ordre des questions ou les propositions des modalités influent sur les réponses.

Contrainte 3: Informations adaptées

Le nombre de questions et leur complexité doivent permettre à un prospect d'y répondre de manière pertinente dans un temps imparti.

La bonne compréhension de la calculette tarifaire s'impose comme une difficulté supplémentaire. Si aucune obligation légale astreint la tarification à être compréhensible dans sa totalité par le souscripteur, elle se doit d'être sensée et cohérente. La prime doit augmenter avec la taille du logement, le nombre de dépendances, la franchise ou le nombre de garanties/options. De plus, les renseignements demandés doivent rester en lien avec l'objet du contrat d'assurance et ne pas paraître intrusif. Par exemple, la valeur des voitures possédées ou le salaire ne peuvent pas être des critères tarifants pour l'assurance MRH bien qu'elles soient significatives actuariellement [7].

Contrainte 4: Informations cohérentes

La prime et les variables tarifaires doivent être cohérentes avec l'assurance souscrite.

Une ultime contrainte opérationnelle de souscription porte sur le montant de la prime sous l'angle du prospect. Mise en concurrence, le montant ne peut être excessif et se doit être proportionné au bien assuré. En effet, rien n'empêche le souscripteur d'expérimenter différents critères tarifaires. De plus, elle se doit d'être géographiquement cohérente, c'est-à-dire qu'entre deux biens contiguës possédant les mêmes caractéristiques, les montants des primes se doivent d'être proches. L'ajout d'une option en plus doit faire augmenter la prime.

Contrainte 5: Prime Proportionnée

Une prime d'assurance doit être concurrentielle, proportionnée au bien protégé, aux attributs du souscripteur et aux options supplémentaires souscrites. Elle se doit aussi d'être géographiquement cohérente.

1.1.1.d La vie du contrat et les indemnisations

Un produit d'assurance MRH est un contrat reconductible tacitement et pouvant être rompu. En d'autres termes, l'assuré et l'assureur peuvent en mentionnant l'autre, rompre le contrat ou l'annuler [8]. Dans le cas contraire, le contrat est reconduit automatiquement. L'assuré devra payer sa prime et l'assureur assurer le risque.

Outre la création d'une structure tarifaire pour les A.N., la tarification a pour objectif la revalorisation des cotisations des assurés en portefeuille. Ces cotisations varient à la hausse s'il y a une augmentation du risque ou à la baisse dans le cas contraire. Pour les revalorisations, le tarif nécessite de modéliser le taux de rétention. Puisque chaque assureur a sa propre clientèle qui ne réagit pas de la même façon à une variation de la cotisation, cette élasticité aux prix peut être plus ou moins modélisée. L'exemple le plus marquant apparait en assurance automobile entre Direct Assurances qui va attirer une population très élastique au prix et d'autres assureurs qui vont avoir une clientèle stable, multiéquipée plus sensible au niveau de la prestation. Le canal de distribution est grandement lié à l'élasticité au prix.

Une multitude d'indicateurs permet d'orienter la tarification : le ratio S/P Sinistre sur Primes [9], le **taux de rétention**, l'exposition à certains risques ou dans certaines zones, le **taux de transformation** des devis [10] - le pourcentage de nouveaux clients souscrivant un contrat d'assurance, le chiffre d'affaires (**CA**), la mutualisation inter-âge...

Pour prendre en compte ces contraintes, les entreprises procèdent à la **mutualisation des tarifs** ou/et à de l'**optimisation tarifaire** sur des critères qui leur sont propres pour ajuster la structure tarifaire.

7. C'est un effet de corrélation avec la valeur d'une maison qui elle est très significative. Plus les voitures valent cher, plus les occupants de l'habitation ont les moyens d'acheter une grande maison (en moyenne et toutes choses égales par ailleurs dans le modèle).

8. Il existe un certain nombre de conditions et de procédures que l'on peut trouver dans les codes suivants : Code des assurances, articles L113-1 à L113-17 et articles R113-1 à R113-14 ainsi que le Code de la consommation : articles L215-1 à L215-5.

9. Un S/P égal à 1 veut dire que les primes reçues sont égales aux sinistres payés et leur gestion. Il est nécessaire d'obtenir un S/P inférieur pour qu'un produit soit rentable. Le S/P des A.N. est généralement moins bon que les affaires en PTF.

10. En fonction du mode de souscription le taux de transformation est différent.

- La mutualisation d'un tarif est le procédé de réajuster la prime sur un critère tarifaire soit en diminuant son impact, soit en la supprimant. Par exemple, la segmentation par l'âge est supprimée par la majorité des mutuelles.
- L'optimisation tarifaire est un procédé permettant d'adapter le niveau de la prime pour maximiser la transformation des devis ou le taux de rétention [11]. Certains acteurs évaluent **l'élasticité aux prix** d'un contrat et adaptent leur optimisation tarifaire. Dans certains cas, les primes des concurrents, les caractéristiques des assurés ainsi que la sinistralité passée sont utilisées. L'optimisation tarifaire la plus simpliste est de diminuer un tarif pour qu'il soit plus faible qu'un concurrent. L'optimisation tarifaire est différente si cela concerne les affaires nouvelles (**A.N.**) ou les contrats déjà souscrits (dit en "portefeuille").

Pour finir, un produit d'assurances englobe les notions de gestions de sinistres, d'indemnisations qui sont la base du service rendu, mais aussi de fraudes, des notions de *prévention* directe (réduction du risque) ou indirecte comme la **réassurance** ou **coassurance**, obligatoire (la CCR et le régime CatNat) ou non. Les produits d'assurances (inter-produits) et leurs processus se juxtaposent. Sur ces différents points, les impacts sur les travaux de tarifications seront mentionnés quand nécessaire. Les notions d'indemnisation et de gestion de sinistres ou des clients sont de loin la partie la plus importante de la chaîne de valeur assurantielle.

1.1.1.e La base contrat

Une fois une A.N. concrétisée, celle-ci est incluse dans le portefeuille d'assurés / contrats de l'assureur. Un contrat est associé à une date d'effet, à une date fin, à un certain nombre de garanties, d'options et à une vision des caractéristiques du contrat. Une photographie des caractéristiques à une date donnée est appelée une **image**. Ces caractéristiques évoluent dans le temps. Un assuré peut déclarer avoir un enfant de plus, ajouter une option pour se couvrir contre un nouveau risque, rectifier/mettre à jour certaines informations. Pour chaque changement, il est créé une nouvelle **image** à laquelle est associée une date de début. Une base contrat peut être vue sous plusieurs angles en fonction de la date d'observation voulue. L'exemple 1.4 montre un cas simple d'évolution de la donnée.

Id Contrat	Id client	Produit	Date de souscription	Date d'effet	Date de fin	Image	Date début image	Date fin image	Information de gestion	Caractéristiques
1133456	153467	MRH 2019 maison	10/02/20	24/02/20	24/02/21	1	15/02/20	19/06/20	Informations de souscription	Informations de souscription
1133456	153467	MRH 2019 maison	10/02/20	24/02/20	24/02/21	2	19/06/20	01/09/20	Informations de souscription	Informations modifiées (I)
1133456	153467	MRH 2019 maison	10/02/20	24/02/20	24/02/21	3	01/09/20	**01/12/20**	Informations modifiées (I)	Informations modifiées (I)

Informations sur les cotisations payées ou non,
Type de paiement : trimestriel/semestriel/annuel, ...

FIGURE 1.4 – Un exemple de base contrat. Voici l'exemple d'un client ayant souscrit le 10 février 2020 une assurance habitation prenant effet le 24 février 2020. Une image (1) de l'assuré est créé avec les informations de souscription. Dans cet exemple, il déclare un changement comme l'ajout d'une dépendance le 19 juin et en septembre, la gestion de l'assureur change une information le concernant. Ces changements provoquent une nouvelle image à chaque fois ((2) puis (3)). On remarque que l'image (3) s'arrête le 1 décembre qui pourrait être la date d'observation de la base.

Il est important de savoir que les caractéristiques peuvent être modifiées par l'assureur que cela soit par l'ajout de données externes ou à cause à l'intervention d'un expert suite à un sinistre. Pour certain assureur, des variables sont ajustées comme le nombre de pièces, ou des informations y sont ajoutées comme le type de matériaux de la maison ou la valeur de reconstruction d'une maison. Ces évolutions temporelles impactent la modélisation (cf : l'article du chapitre 7).

11. Le pourcentage de personnes ne résiliant pas leur contrat. En MRH, le taux de rétention est entre 92 à 85 % selon les assureurs et les catégories considérées.

1.1.1.f La base d'indemnisations / sinistralités

Lors des déclarations des sinistres, des échanges d'informations ont lieu. À chaque nouvelle déclaration, le service de gestion crée une nouvelle ligne dans la base d'indemnisation. Ensuite, ce service va interroger l'assuré et la base contrat pour déterminer le type de sinistres, si l'assuré est couvert, le montant potentiel du dommage et l'indemnisation une fois la franchise et le plafond déduits. Cet ensemble d'informations est sauvegardé dans une base de données tabulaire ayant pour identifiant l'ID de l'assuré/client, le numéro du contrat associé, la date déclarée du sinistre, le type de dommage et à quelle garantie les dommages sont associés. Cette première étape amène des différences entre les assureurs. En effet, certains types de sinistres vont être associés à une garantie plutôt qu'à une autre ou être répartis sur plusieurs garanties. Le point le plus différenciant sont les évènements climatiques non déclarés comme Catastrophes Naturelles. Certains assureurs considèrent les inondations dans la garantie Dégâts Des Eaux **(DDE)**, d'autres les isolent alors que certains créent une garantie évènements climatiques.

Id Contrat	Id client	Produit	Date déclaration	Date de survenance	Paiements	Provisions	Recours	(Ré)Ouvert	Garanties	Descriptions	Canal
1133456	153467	neoMRH	24/02/20	24/02/20	6580	700	0	Oui	DDE	Inondation sous-sol, Peinture, mobilier	Internet
1133456	153467	neoMRH	04/01/21	27/12/20	0	600	0	Oui	DDE	Fuite salle de bain	Internet
245688	178077	PART 2015	25/10/21	24/10/21	654	0	0	Non	INC	Feu - cheminée.	Mail
245688	178077	neoMRH	06/07/21	06/07/21	0	0	0	Non	PJ	-	Assistance téléphonique

FIGURE 1.5 – Un exemple de base sinistre. Seuls les clients ayant eu des sinistres sont affichés. Le premier contrat a eu deux sinistres. Les canaux de déclarations des sinistres peuvent être différents tout comme les délais entre la survenance et la déclaration. Cette base évolue dans le temps avec les paiements, les déclarations, l'intervention d'expert sur place ...

Une autre problématique apparait lors de la mise à jour de la base sinistre, le coût du sinistre n'est pas connu dès le départ. Le processus d'indemnisation passe par plusieurs phases comme le montre le graphique 1.5.

1.1.1.g Le provisionnement

Le processus d'indemnisation s'initialise par une provision d'ouverture, c'est-à-dire le sinistre est ouvert par un montant forfaitaire : par exemple, le sinistre est créé dans la base de sinistralités et une provision de 1000 € lui est attribuée. Quand plus d'informations sur le sinistre sont données, une nouvelle provision lui est associée avec un montant adapté au sinistre. Une fois les factures fournies, le sinistre va être indemnisé en une ou plusieurs fois. Informatiquement, la colonne paiement est incrémentée par l'indemnisation et la provision va être réévaluée. Si le sinistre est cloturé, la provision associée à ce sinistre est annulée. La durée d'ouverture peut s'étendre de quelques jours comme de quelques années pour des sinistres "graves" avec des montants importants comme pour la sécheresse. Un sinistre peut être ré-ouvert à l'émergence de nouveaux problèmes issus du sinistre. Mathématiquement, le coût brut d'un sinistre à la date t est égal à :

$$\begin{cases} Paiement(t) + Provision(t) - Recours(t) & \text{si le sinistre est (ré)ouvert,} \\ Paiement(t) - Recours(t) & \text{si le sinistre est clos.} \end{cases}$$

Des sinistres dits "sans suites" se manifestent fréquemment. Ce sont des sinistres déclarés ne donnant pas lieu à une indemnisation. Plusieurs cas apparaissent : la personne n'était pas assurée pour le risque, les démarches n'ont pas abouti, elle a bénéficié d'un conseil de la part de l'assureur (en protection juridique par exemple). Le taux de sans suites est de l'ordre de 2 à 3% en MRH et varie selon les garanties. Ces sinistres ont nécessité le travail des services de gestions. Ils ont, alors, un coût opérationnel.

La tarification, parmi ses objectifs, doit modéliser le coût final d'un contrat, c'est-à-dire le coût final de tous les sinistres qui seront déclarés et assurés par le contrat. Comme le coût d'un sinistre (ré)ouvert est rarement le coût total du sinistre, il va être nécessaire d'estimer le montant final indemnisé - la vision à l'ultime du sinistre. La vision/date de référence d'extraction de la base est donc très importante et modifie la donnée.

Le provisionnement a aussi un aspect comptable. La durée de vie du contrat est plus longue que l'exposition au risque. En effet, les contrats prennent en compte l'ensemble des sinistres intervenus lors de l'exposition au risque. La déclaration de ceux-ci peut être faite ou actualisée/développée après la période de couverture du risque (par exemple : le second sinistre du graphique 1.5). À cet effet, des notions de provisionnement comptable (montant comptable pour faire face aux engagements présents pour des paiements futurs) sont obligatoires et propres aux assurances. Les provisions comptables sont des montants comptables calculés pour la clôture des comptes. Elles ont pour objectif d'évaluer les engagements futurs pour que le résultat comptable soit le plus proche de la réalité. En d'autres termes, les provisions obligent les assureurs à conserver les montants pour lesquels ils se sont engagés.

En quoi ces notions de provisions comptables sont-elles importantes pour la tarification actuarielle ? La tarification se base sur des données de sinistralités pour évaluer la prime commerciale et le coût moyen annuel (**Prime pure**). Elle utilise directement les données de la gestion, mais ces données ne sont pas complètes et pas toujours assez fines. Pour ajuster ces informations, il va être nécessaire de projeter/prendre en compte les sinistres non encore déclarés/fermés. Étant donné que ce travail est à produire annuellement pour chaque inventaire, l'utilisation d'une même nomenclature semble naturelle.

Dans cette thèse, il est mention de quelques provisions élémentaires [12] :

— **PSAP** (Provisions pour sinistres à payer) - R331-6, alinéa 4 : cette provision (aussi appelé **IBNR** (Incurred But Not Reported)) regroupe l'estimation des sinistres dossier/dossier, les sinistres non encore déclarés (**IBNYR** (Incurred But Not Yet Reported)) dit "tardifs" et ceux non encore payés dans leur totalité (**IBNER** (Incurred But Not Enough Reported)) ;

— **PANE** (resp. **PPNA**) Primes Acquises Non Emises (resp. Primes Payés non acquises) : ces provisions correspondent aux primes acquises ou émises durant l'année comptable associées aux engagements futurs ;

— **PM** (Provisions mathématiques) : Provisions pour les rentes .

D'après l'article A344-2, modifié par Arrêté du 26 décembre 2019 - Article 9, ces provisions se calculent par catégories ministérielles comme les Dommages aux Biens des Particuliers, les Catastrophes Naturelles, la Responsabilité Civile générale, la Protection juridique, les Pertes Pécuniaires Diverses... De manière générale, ces provisions sont calculées plus finement par produit et par type de risque (Inondations, Sécheresse ...). En fonction des garanties, il va être nécessaire de prendre en compte certaines provisions et pas d'autres.

1.1.1.h La vie du contrat : les bases d'études actuarielles

La majorité des études actuarielles s'adosse sur la jointure entre la base indemnisation et la base de contrat, appelée **la base de modélisation**. Selon l'objectif, il est nécessaire d'extraire une base adaptée pour l'étude. Pour estimer le montant total des sinistres pour refondre un tarif, il va être nécessaire de choisir une base sinistre avec la vision la plus récente possible. Pour une visée de provisionnement, il est nécessaire d'extraire plusieurs fois la base à différentes dates d'observations pour créer des triangles de paiements dans le temps, afin d'estimer des cadences de règlement ou le montant à l'ultime des sinistres ouverts. Pour des études de valeurs clients, il peut être utile de ne considérer que les données fournies à la souscription. Pour certaines études, des jointures avec des données externes s'effectuent. Dans ce cadre, les jointures se font traditionnellement par communes ou départements (ou bien pour l'assurance automobile le type de voiture pour les données SRA (Sécurité et Réparation Automobile)).

En actuariat, la modélisation s'appuie principalement sur des données tabulaires à partir de données numériques comme l'âge, la surface habitable ou des variables factorielles comme la qualité de l'occupant, la catégorie socio-professionnelle. Des variables continues sont discrétisées comme la date de construction du bâtiment. Selon les assureurs, les colonnes ont plus ou moins de valeurs manquantes. Par exemple, lorsqu'un assureur a un processus de souscription s'adaptant dynamiquement aux réponses précédentes, de nombreuses valeurs sont non renseignées, car elles n'ont jamais été demandées à l'assuré. Par exemple, pour un contrat des propriétaires non occupant, la valeur des biens garantis est peu demandé. En pratique, il est impossible de demander une information à un assuré qui n'était pas dans le questionnaire. **La mise à jour des données et les informations manquantes sont des problèmes indissolublement liés à la tarification.**

Les données de contrats et de sinistralités peuvent être étudiées individuellement par produit ou dans leur ensemble. Ces questionnaires sont différents suivant les produits et leur date de commercialisation. De nouvelles questions peuvent être demandées ou supprimées tout comme certaines caractéristiques

12. Il existe différentes nominations propres à chaque organisme d'assurances moins comptable comme **RAE** et **RAP** (Reste à Encaisser/Payer) et d'autres provisions plus annexes comme la **PREC** : Provision pour Risque En Cours.

peuvent voir leur définition être modifiée. Un organisme d'assurance fait une refonte en MRH tous les 3 à 5 ans en moyenne. Pour calculer des nouvelles structures tarifaires, le processus de création du produit va nécessairement devoir utiliser l'historique du précédent produit et les réponses aux questions des précédentes gammes.

Contrainte 6: Continuités et discontinuités du questionnaire

Le processus de tarification doit permettre la prise en compte de nouvelles variables tarifaires non utilisées jusqu'ici. De plus, les questions demandées à la souscription doivent être un minimum contiguë entre une gamme récente et la précédente.

La modélisation de la prime pure doit aussi respecter quelques contraintes législatives. En MRH, il existe différentes taxes en fonction des garanties (Article 1001-1° du Code Générale des Impôts). La garantie incendie est taxée à hauteur de 30%, la protection juridique à 13.4% alors que les autres garanties sont à 9%. De plus la garantie CatNat est calculée comme 12% du montant des garanties dommages (pour faire simple toutes les garanties hors attentat et les garanties corporelles). Ainsi, cette maille est la maille minimale de calcul. Les montants des primes entre ces différentes garanties doivent être bien proportionnés par rapport à la sinistralité observée [13].

Contrainte 7: Contrainte législative

La cotisation doit être calculable à une maille au moins aussi fine que celle demandée par le législateur.

La base d'étude principale en tarification est la jointure entre la base des contrats et celle des sinistres. En fonction de leur utilisation, certaines bases regroupent directement le nombre de sinistres par garanties, le coût total associé par image ou conservent individuellement les détails des sinistres. Pour modéliser la fréquence, il faut compter le nombre de sinistres par image. La majeure partie des bases rencontrées scinde les contrats par an, mais il est possible de conserver des lignes avec une exposition associée à plusieurs années.

Une exposition aux risques est calculée par garantie et prend une valeur entre 0 et 1 si les lignes sont annuelles. Un grand nombre de contrats ont des valeurs d'exposition égale à 1 (environ 60%) et d'autres ont une exposition entre 0 et 1. La répartition est quasiment uniforme. L'exposition est de loin l'information la plus importante et la plus tarifante. Elle nécessite un traitement spécial de la part de l'assureur. En effet, dans le cadre d'une rupture de contrat par l'une ou l'autre des parties, l'assureur doit rembourser *prorata temporis* la cotisation.

Contrainte 8: Exposition et primes

La prime doit être proportionnelle à l'exposition.

La tarification des **A.N.** s'appuie sur les données de la sinistralité passée sous l'hypothèse que les clients/devis et sinistres à venir sont similaires à ceux du passé. Cependant, certains biais temporels et l'évolution de la sinistralité obligent les actuaires à ajuster leurs modèles. De plus, lors de la création d'un nouveau produit, il est nécessaire de pouvoir ajouter de nouveaux critères facilement.

Contrainte 9: Tarifs adaptables

La tarification se doit d'être ajustable pour pouvoir prendre en compte les changements de causalités, de relations entre la sinistralité et les variables tarifaires ou des incohérences de modélisation.

1.1.2 Méthodologie de calcul d'une structure tarifaire

Le calcul d'une prime d'assurance fait l'objet d'un processus bien défini. Après avoir introduit les notations dans la partie 1.1.2.a, un graphique 1.6 résume le procédé générique de calcul d'une prime d'assurance.

13. Sinon cela peut être considéré comme de l'optimisation fiscale.

1.1.2.a Notations

Éléments contractuels :	
c	Contrat d'un assuré ;
N	Nombre de contrats dans le portefeuille de l'assureur ;
$\Pi^{comm}(c)$	Prime commerciale proposée au contrat c ;
$\Pi^{comm\ HT}(c)$	Prime commerciale hors taxes proposée au contrat c ;
$\Pi^{conc_i}(c)$	Prime commerciale proposée au contrat c du concurrent i ;
$\Pi^{charge\ TTC}(c)$	Prime chargée avec taxes associée au contrat c ;
$\Pi^{pure}(c)$	Prime pure associée au contrat c ;
$\Pi^{nette}(c)$	Prime nette associée au contrat c ;
$\Pi_{garantie\ g}(c)$	Prime nette d'une garantie g associée au contrat c ;
Base de données et modèles :	
$\mathcal{D}_{global}$	Base de données tabulaire disponible ;
$\mathcal{D}_{train}$	Base de données tabulaire utilisée pour la modélisation ;
$\mathcal{D}_{test}$	Base de données tabulaire utilisée pour les tests ;
$\mathcal{X}_{train}$	Base d'entrainement d'un modèle ;
$\mathcal{X}_{validation}$	Base de validation des hyper-paramètres d'un modèle ;
$\mathcal{X}_{test}$	Base de test d'un modèle ;
p	Nombre de variables de la base $\mathcal{X}_{train}$;
n	Nombre de lignes dans la base $\mathcal{X}_{train}$;
$\mathcal{M}$	Ensemble de modèles (m un élément de cet ensemble) ;
Variables :	
Y	Variable à modéliser (Nombre de sinistres, coût moyen...) associée aux observations $(Y_1, ..., Y_n)$;
$\hat{Y}$	Variable estimée associée aux prédictions $(\hat{Y}_1, ..., \hat{Y}_n)$;
$\mathbf{X}$	Ensemble des caractéristiques des assurés $(X_1, ..., X_p)$;
V ou v	Exposition au risque du contrat ;
S	Sévérité d'un sinistre ;
N	Nombre de sinistres pendant une exposition donnée v ;
L	Coût total historique pendant une exposition donnée v ;
s_{grave}	Montant à partir duquel un sinistre est considéré comme grave ;
$e_{prix}(c)$	Élasticité au prix du contrat c ;
$Freq$	Fréquence d'un sinistre avec $Freq_\Omega$,$Freq_{S<s_{grave}}$, $Freq_{S>s_{grave}}$ respectivement la fréquence de tous les sinistres, seulement les attritionnels et seulement les graves ;
CM	Coût moyen d'un sinistre ;
w_i	Poids associé à un individu (élément) i ;
I	Variable(s) intensité(s) d'un événement e_I ;
$P(I)$	Ensemble des événements considérés par I (e_I un élément de cet ensemble) ;
Outils mathématiques :	
$Cost(\hat{Y}, Y\|\mathcal{X})$ $Cost(m, \mathcal{X})$	Évaluation d'une métrique à optimiser sur la base $\mathcal{X}$ entre les valeurs observées Y et les valeurs estimées $\hat{Y}$ ou de façon équivalente d'un modèle m sur la base $\mathcal{X}$;
$\mathcal{L}(\hat{Y}\|Y)$	Vraisemblance de $\hat{Y}$ pour une distribution donnée par rapport à des valeurs observées Y ;
$Dev(\hat{Y}\|Y)$	Déviance pour une distribution donnée égale à $log(\mathcal{L}(\hat{Y}\|Y)) - log(\mathcal{L}(Y\|Y))$;
$Gini$	Gini : métrique mesurant la segmentation d'un modèle par rapport aux valeurs observées (Voir l'annexe D).
AUC	Area Under the Curve, aire sous la courbe de Lorenz (Voir l'annexe D) ;
LM	Modèle Linéaire ;
GLM	Modèle Linéaire Généralisé ;
β	Coefficients $\beta_0, ..., \beta_p$ associés aux modèles GLMs et LMs ;
RF	Random Forest (méthode dite "Machine Learning" - ML) ;
$CART$	Classification And Regression Trees (ML) ;
$XGBoost$	eXtrem Gradient Boosting (ML) ;

Notations des graphiques : $\{x_i, y_i\}$ représente une graphique de nuage de points avec les coordonnées (x_i, y_i) pour tout i.

Tout d'abord, la sous-section 1.1.2.b détaille les garanties usuelles de l'assurance MRH. Pour chacune de ses garanties, la modélisation suit un schéma similaire. La modélisation Fréquence × Coût Moyen est de loin préférée - sous-section 1.1.2.c - mais une distinction en fonction de la gravité des sinistres est faite - sous-section 1.1.2.d. Cependant, certaines garanties climatiques ne peuvent être modélisées de cette manière - sous-section 1.1.2.e et certaines ne sont modélisées que pour des besoins de souscription et non de tarification. En consolidant ces informations, la sous-partie 1.1.2.f détaille la création des primes d'assurances. Dans la section 1.1.3 suivante, les notions d'hyper-individualisation, de segmentations, d'anti-sélection et d'aléa moral complètent les principes et les contraintes associés à la tarification en assurance.

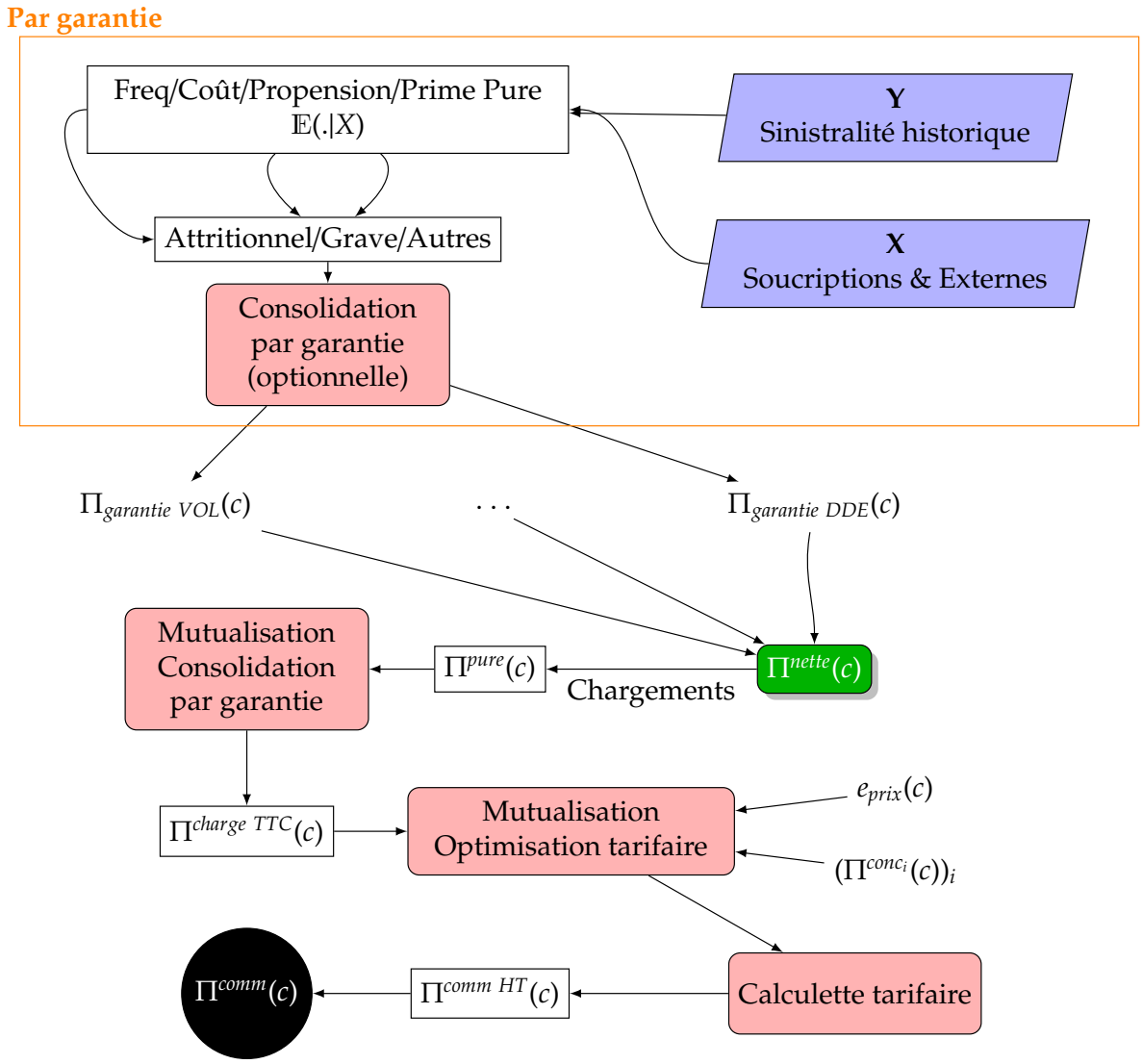


FIGURE 1.6 – Schéma du processus générique de calcul d'une prime d'assurance. Dans cette thèse, nous nous restreignons aux calculs de la prime nette, $\Pi^{nette}(c)$. Les squircles rouges correspondent aux notions actuarielles et les trapèzes bleus aux bases de données.

1.1.2.b Les garanties incluses dans l'assurance MRH

En pratique, le calcul d'un tarif d'un produit MRH se réalise à la maille la plus fine possible. Pour manipuler des modèles statistiques, l'hypothèse principale est de modéliser des variables aléatoires indépendantes identiquement distribuées (*i.i.d.*), c'est-à-dire qu'il est nécessaire de regrouper les mêmes types de risques. Les garanties s'imposent comme une maille pertinente répondant aux différentes contraintes.

L'assurance **MRH** enveloppe différentes assurances de dommages et assurances de personnes détaillées dans le tableau 1.2. En montant et en fréquence, les garanties DDE, VOL, INC, BDG, CATNAT, TGN et RC sont les garanties prépondérantes. Chaque garantie a ses propres variables tarifaires et particularités. Les montants et la fréquence de survenance moyens des sinistres pour des garanties de dommages sont expliqués majoritairement par les caractéristiques du bâtiment, tout particulièrement par la surface habitable ou le nombre de pièces. Les garanties de personnes comme la responsabilité civile sont liées aux nombres d'habitants, d'enfants ou aux catégories socio-professionnelles (CSP). Finalement, les garanties TGN et EVECLIM "climatiques" sont spécifiques et les variables tarifaires diffèrent en fonction du fait générateur du sinistre.

Table 1.2 – Ensemble des garanties généralement disponibles en MRH.

Type	Garantie	Acr.	Définition	Particularité
Dommage	Incendie	INC	Dommages résultant d'un incendie	De nombreux sinistres dits graves.
	Dégâts des eaux	DDE	Dommages suite à un dégât des eaux ou remontée des nappes phréatiques ou des inondations de faibles ampleurs	Sinistres de haute fréquence pour les biens collectifs et une partie climatique.
	Vol	VOL	Sinistres suite à un vol.	L'impact spatial est très important.
	Dommage électrique	ELEC	Sinistres suite à un accident électrique.	Majoritairement causés par la foudre.
	Panne	PAN	Option d'extension des garanties des équipements électroménager.	-
	Catastrophes naturelles	CatNat.	Sinistres associés à un évènement déclaré comme catastrophe naturelle dans la commune	Le calcul est réglementaire : 12 % de la somme des montants des autres garanties dommages.
	Tempête grêle neige	TGN	Sinistres suite à une tempête, à la grêle ou à la neige.	Sinistralités de hautes fréquences et spatialement corrélées.
	Évènements climatiques	EVE CLIM	Tous les sinistres climatiques hors CatNat et TGN couvert.	Elle n'est pas toujours existante car incluse dans certaines garanties.
	Bris de glace	BDG	Couvre les parties vitrées.	Sinistres bris de glace non associés aux autres garanties.
Personne	Responsabilité civile	RC	Couvrant les dommages à autrui par les personnes sous la responsabilité du souscripteur.	Obligatoire.
		RC VP	RC vie privée couvrant la personne assurée et les personnes sous sa garde.	-

		RC Spé	RC couvrant certaines spécificités comme certains chiens dangereux, activité professionnelle.	Optionnelle
		SCOL	RC scolaire	La Sinistralité moyenne est proportionnelle au nombre d'enfants.
	Défense et recours pénal	DPR	Frais pénaux associés à des dégâts liés aux garanties couvertes.	Compléments quasi-automatique de la RC.
	Protection juridique	PJ	Option permettant un conseil juridique ou de l'assistance d'un avocat lors d'une procédure judiciaire.	Optionnelle
	Attentats	ATT	Dommages matériels lors d'un attentat ou d'un acte de terrorisme	Une taxe d'un montant réglementaire de 5,90 € est à ajouter.

1.1.2.c La modélisation : Fréquence X Coût moyen

En assurance non-vie, le contrat d'assurance couvre un bien pendant une période donnée. Dès lors, plusieurs sinistres peuvent survenir pendant ce laps de temps. La grande majorité des sinistres sont indemnisés dans leur totalité hors franchises. Les survenances et les montants des sinistres n'impactent pas automatiquement les franchises futures (sauf pour certains sinistres CatNat avant 2022). Dans ce cadre, les montants et la fréquence ne sont pas structurellement dépendants. C'est pourquoi les actuaires utilisent l'approche Fréquence × Coût moyen. Dans d'autres cas, le coût total L est modélisé par l'approche des "modèles de prime pure".

① *Fréquence × Coût moyen*

La méthode de modélisation du coût d'une garantie la plus usitée est celle de modéliser séparément la fréquence et le coût moyen et de les assembler comme suit :

$$\begin{aligned}\mathbb{E}\Big(\frac{L_i}{v_i}|\mathbf{X}_i\Big) &= \mathbb{E}\Big(\frac{N_i S_i}{v_i}|\mathbf{X}_i\Big) \\ &= \underbrace{\mathbb{E}\Big(\frac{N_i}{v_i}|\mathbf{X}_i\Big)}_{Freq_i} \times \underbrace{\mathbb{E}\Big(S_i|\mathbf{X}_i\Big)}_{CM_i},\end{aligned} \tag{1.1}$$

pour un assuré i avec les caractéristiques $\mathbf{X}_i$. L'équation 1.1 suppose l'indépendance du montant des sinistres avec leur survenance. Pour certaines garanties (souvent climatiques - partie 1.1.2.e), cette hypothèse n'est pas raisonnable. Pour les sinistres graves et attritionnels de la majorité des garanties, cette hypothèse est plausible et acceptée.

En tarification, les actuaires cherchent à évaluer l'espérance $\mathbb{E}(S_i|\mathbf{X}_i)$ dit *Coût Moyen* et non les montants des sinistres S_i, l'espérance $\mathbb{E}\Big(\frac{N_i}{v_i}|\mathbf{X}_i\Big)$ dit *Fréquence* et non la survenance des sinistres N_i.

♠ *Modéliser la Fréquence*

Pour estimer la fréquence, il est nécessaire de comprendre les spécificités de la base de modélisation sous-jacente. Cette base résulte de l'assemblage des caractéristiques du contrat et de sa sinistralité. La

clef de jointure se crée à l'aide de la période des images des contrats et la date de survenance des sinistres et non de déclaration. La base de sinistralités utilisée est la plus récente possible pour que le nombre de sinistres "tardifs" **IBNYR** soit faible. Malgré tout, le nombre observé de sinistres sera vraisemblablement sous-estimé, ce biais est retraité suite aux modélisations.

Pour chacune des garanties avec le sous-découpage attritionnelle et grave, une base de modélisation est créée avec l'exposition associée à la garantie dans laquelle chaque ligne correspond à une image d'un contrat. L'exposition d'une image v_i est toujours inférieure à la durée de l'image. L'exposition v_i est une valeur numérique normalisée annuellement. En effet, l'actuaire cherche à calculer la sinistralité annuelle pour un assuré d'une image i correspondant à $\mathbb{E}\big(L_i|\mathbf{X}_i\big) = \mathbb{E}\big(\frac{L_i}{v_i}|\mathbf{X}_i\big)$ quand $v_i = 1$. Par exemple, quand $v_i = 0.5$, cela exprime que le contrat a été protégé et exposé au risque pendant une durée de 6 mois pendant l'image.

La base de données assemble trois types d'éléments nécessaires : les informations de l'image du contrat, l'exposition du contrat à la garantie pendant l'image et les informations sur les sinistres (ouvert, fermé, recours en cours, le nombre, les caractéristiques).

Certaines déclarations ne donnent pas lieu à des indemnisations. Pour la modélisation Fréquence × Coût, elles ne doivent idéalement pas être prises en compte. Dans le cas contraire, un ajustement doit être fait en nivelant la prime pure. Néanmoins, certaines déclarations ne donnant pas lieu à une indemnisation entrainent des frais de gestion de sinistres (frais d'experts, conseil juridique...). Idéalement, elles devraient être modélisées indépendamment. Communément, ces dernières transparaissent dans les chargements des sinistres *a posteriori* et sont réparties proportionnellement au niveau de la prime pure. De plus, certains sinistres donnent lieu à des indemnisations forfaitaires. Habituellement en MRH, ils ne sont pas distingués des autres sinistres.

♠ *Modéliser le Coût Moyen*

Pour l'estimation de CM_i, la base de modélisation exploitée est constituée des images des contrats ayant eu un sinistre avec une indemnisation non nulle. Plusieurs types de bases cohabitent : soit chaque ligne correspond à un sinistre, soit chaque ligne correspond à la somme des sinistres pendant une image. Dans ce dernier cas, le nombre de sinistres survenus N_i doit être disponibles dans la base.

Un détail sur le coût du sinistre est éventuellement présent : les paiements, les provisions et le montant des recours. Chaque assureur possède des processus d'indemnisations différents qui doivent être considérés (ex : frais des délégataires pour l'indemnisation). Afin de modéliser le coût moyen, la vision finale des indemnisations dite la vision à l'ultime, est évaluée. Pour les sinistres clos, le montant du paiement correspond à l'ultime. Pour les sinistres encore ouverts, les provisions ne sont pas toujours la meilleure estimation du restant à payer et sont peu informatives des sinistres à venir. Conséquemment, il est nécessaire de projeter à l'ultime en utilisant les informations des paiements uniquement. Cependant, pour les sinistres avec un développement long comme la sécheresse ou la RC corporelle, le paiement des premières années n'est pas assez représentatif du coût final et la projection à l'ultime est plus pertinente en incorporant les estimations du reste à charge (**IBNER** ou **RAP** ou **PM**) provenant d'experts ou d'actuaires.

Certaines spécificités peuvent surgir à travers cette base comme l'apparition de montants forfaitaires (en BDG par exemple) ou de montants négatifs de sinistres ou les recours. La plupart du temps, ces deux derniers cas sont rarement considérés.

② *Les modèles de primes pures*

Les modèles de primes pures sont de deux catégories. La première est d'estimer directement $\mathbb{E}\big(\frac{L_i}{v_i}|\mathbf{X}_i\big)$ à partir de la sinistralité observée. La seconde est de s'appuyer sur les résultats des modèles de fréquence et de coût moyen pour modéliser les effets appris. Dans les deux cas, chaque ligne correspond à une image et la grandeur à mesurer est le coût total historique ou estimé durant une image. Comme pour les bases de fréquences, il y a une exposition associée à chacune des garanties.

♠ *Estimer la prime pure à partir de la sinistralité :*
Dans certains cas, les informations disponibles ne permettent pas de dégager une structure tarifaire statistiquement pertinente, d'obtenir un volume de données suffisamment conséquent ou alors le coût moyen et la fréquence ne sont pas suffisamment indépendants. En pratique, rares sont les garanties concernées. Le plus souvent, cette approche est pertinente pour des garanties avec un nombre faible de

sinistres.

À la différence de l'approche Fréquence × Coût, les effets des variables sont moins linéaires. Conséquemment, la sélection des variables est plus ardue et la performance des modèles obtenus souvent moindre. La base de modélisation dans ce cas est finalement la combinaison des bases précédentes vues à l'ultime.

Par exemple, *CM* et *Freq* croissent généralement avec la surface habitable. Si cette variable peut ne pas être significative en Fréquence × Coût à cause du volume de données, en modélisant directement la sinistralité, cette variable peut le devenir.

♠ *Estimer la prime pure à partir de prédictions des modèles de coût et de fréquence :*

Cette modélisation est nécessaire pour **la consolidation des modèles**. La base de modélisation utilise les prédictions des fréquences estimées et les coûts moyens par garanties. Il n'y a plus de volatilité inhérente aux sinistres puisque que le modèle apprend sur *Freq* et *CM* et non *N* et *S*. L'objectif est d'obtenir un modèle récapitulatif simple et interprétable par garanties. Ce type de modèle permet aussi de comparer les résultats à la sinistralité observée ; voir s'il n'y a pas eu du sur-apprentissage ou si les effets marginaux sont explicables et cohérents par garantie. Ces tests permettent de vérifier en amont principalement les contraintes 4 et 5.

1.1.2.d Les sinistres attritionnels et graves

Pour modéliser des variables aléatoires *i.i.d* concernant une même garantie, les sinistres sont scindés en sous-groupes - en MRH, généralement en deux groupes : les sinistres attritionnels et les sinistres graves [14]. Dans certaines entreprises, une attention particulière est faite pour les indemnisations forfaitaires. Les sinistres graves sont définis comme des sinistres élevés avec une fréquence de survenance faible et les sinistres attritionnels sont fréquents, mais avec des montants d'indemnisations modiques.

♠ *Comment définir les sinistres graves et attritionnels ?*

La conjecture est que les sinistres graves et attritionnels sont de natures différentes. Plusieurs méthodes sont disponibles et le choix de l'une d'entre-elles dépend des hypothèses sous-jacentes. Pour les modéliser, deux questions sont primordiales :

— À quel point les faits générateurs entre sinistres graves et attritionnels sont-ils comparables ?
— Quelles sont les différentes caractéristiques influençant les montants de l'indemnisation ?

①*Modèle de sur-crêtes*

La première méthode, la plus utilisée, est celle des **sur-crêtes** ou d'écrêtements. La sur-crête est la partie d'un sinistre dépassant un seuil s_{grave}. Son agrégation se réalise comme suit :

$$\Pi = Freq_{\Omega} \times \Big(CM_{min(S,s_{grave})} + \mathbb{P}(S > s_{grave}) \times CM_{sur-crête} \Big), \tag{1.2}$$

où Π est la prime nette, $Freq_{\Omega}$ est la fréquence de la survenance de sinistres graves ou attritionnels, $CM_{min(S,s_{grave})}$ coût moyen des sinistrés capés, $\mathbb{P}(S > s_{grave})$ est la probabilité qu'un sinistre survenu est dit "grave" et $CM_{sur-crête}$ est le coût moyen des excès. Le modèle mesurant $\mathbb{P}(S > s_{grave})$ en fonction des caractéristiques de l'assuré s'appelle un **modèle de propension**.

L'équation 1.2 des modèles de sur-crêtes repose sur deux hypothèses : les faits générateurs des sinistres graves et attritionnels sont similaires et les biens qui sont indemnisés lors de sinistres graves ou attritionnels sont analogues.

Exemple 1.1: Garantie ELEC

La garantie ELEC peut être adaptée à l'approche de sur-crêtes. Les faits générateurs potentiels d'un sinistre sont les surtensions, les courts-circuits ou la foudre par exemple. La fréquence grave est donc générée de la même manière que celle attritionnelle. De la même façon, les objets endommagés

14. En auto, certaines garanties peuvent se modéliser en trois groupes : attritionnels, sinistres graves et sinistres extrêmes. Pour l'ensemble des travaux menés et d'assureurs rencontrés en MRH, je n'ai pas rencontré de tels cas. On peut néanmoins faire autant de groupes que possible en théorie. La contrainte est plus pratique que théorique.

ne changent pas ; ce sont les objets immobiliers et mobiliers. L'indemnisation des sinistres graves et attritionnels reposent sur les mêmes bases.

② *Modèle au premier euro*

La seconde méthode, dit "modèles au premier euro", considère un modèle grave et un modèle attritionnel indépendamment. Son agrégation s'effectue comme suit :

$$\Pi = Freq_{S<s_{grave}} \times CM_{attri} + Freq_{S>s_{grave}} \times CM_{grave}, \tag{1.3}$$

où $Freq_{S<s_{grave}}$(resp $Freq_{S>s_{grave}}$) est la fréquence des sinistres attritionnels, CM_{attri} (resp $CM_{S>s_{grave}}$) le coût moyen des sinistres attritionnels (respectivement pour les graves). L'équation 1.3 des modèles au premier euro repose sur l'hypothèse que les faits générateurs et les critères de la sévérité sont différents.

Exemple 1.2: Garantie DDE

La garantie DDE peut être adaptée à l'approche des modèles au premier euro quand les événements inondations non déclarés comme catastrophes naturelles y sont présents. Les faits générateurs sont différents ; l'inondation provenant d'évènements météorologiques (50% en fréquence chiffres 2022) et les DDE attritionnels provenant plus sur la vétusté et le comportement des assurés. Les parties endommagées du bien sont différentes.

②*Modèle de coûts indépendants*

Une troisième méthode estime un modèle de coût moyen grave indépendant et est une méthode à la convergence des deux précédentes méthodes (1.2) et (1.3),

$$\Pi = Freq_{\Omega} \times \Big((1 - \mathbb{P}(S > s_{grave}))CM_{attri} + \mathbb{P}(S > s_{grave}) \times CM_{grave}\Big). \tag{1.4}$$

Cette équation 1.4 repose sur l'hypothèse que les faits générateurs sont analogues, mais que les différences sur critères impactant le coût moyen sont significatives.

Exemple 1.3: Garantie INC

La garantie INC peut être un exemple valable. Même si l'aggravation de la sévérité de l'incendie peut dépendre de facteurs externes, les déclarations d'incendies sont provoquées par les mêmes faits générateurs. Cependant, le coût d'un incendie "grave" dépend de la valeur du bien, du toit ou des matériaux alors que pour les incendies à montant modiques, ce sont les valeurs du mobilier ou de peintures qui sont les plus importantes.

D'autres formes de prise en compte des sinistres graves peuvent être proposées. Par ailleurs, les fréquences des sinistres graves et attritionnels peuvent être dépendantes. En effet, avoir un sinistre grave peut faire diminuer la probabilité de déclarer un sinistre attritionnel (si le bien est quasiment détruit à la suite de l'incident ou réparer à neuf) ou l'augmenter (en fragilisant le bien). Dans ce cadre, un statisticien pourrait y ajouter une structure de dépendance ce qui modifierait la structure d'agrégation. Néanmoins, cela complexifierait grandement le tarif et entrerait en contradiction avec la contrainte 9. Les pratiques opérationnelles pour prendre en compte ce problème sont de deux types :

* Mettre en place un coefficient d'ajustement pour la garantie ;
* Modifier l'exposition au risque associé dans de rares cas.

♠ *Comment déterminer la distinction ou les seuils entre les sinistres graves et attritionnels ?*

De mon point de vue, des informations sur le contexte du sinistre et les faits générateurs pourraient permettre de classer le sinistre automatiquement en grave ou attritionnel. Cependant, ces informations ne sont pas toujours disponibles, de bonnes qualités ou faciles à déterminer.

Pour déterminer si un sinistre est grave ou attritionnel, l'idée est de détecter un changement de distribution probabiliste sous-jacent en travaillant sur les montants des sinistres. En pratique, un seuil par garanties [15] est déterminé au-dessus duquel on considère les sinistres comme graves. Il existe un panel de méthodes statistiques disponibles (voir l'annexe C pour le détail théorique et des exemples).

♠ *Les modèles utilisés*

Comme mentionné dans les précédents paragraphes, différentes grandeurs sont modélisées :

1. $\mathbb{P}(S > s_{grave})$ à l'aide d'un modèle de propension,
2. $Freq_{S>s_{grave}}$ à l'aide d'un modèle de fréquence,
3. CM_{grave} ou $CM_{sur-crête}$ à l'aide d'un modèle de grave ou de sur-crêtes.

Toutefois, la caractéristique principale des sinistres graves est d'avoir une faible fréquence et d'avoir une sévérité volatile. En conséquence, les estimateurs les plus simples sont aussi les plus utilisés, *c.-à-d.* $\mathbb{P}(S > s_{grave})$ est estimée par $\frac{Nb_{S>s_{grave}}}{Nb_S}$ ou bien $CM_{sur-crête}$ par $\frac{\sum_{S>s_{grave}} S - s_{grave}}{Nb_{S>s_{grave}}}$. Pour certaines garanties comme l'incendie, les renseignements transmettent suffisamment d'information pour permettre une modélisation de la sinistralité grave avec 3 à 5 critères tarifaires. Toujours est-il, les modèles de graves ont moins de critères tarifaires que les modèles attritionnels du fait de leur plus faible fréquence et leur plus grande volatilité.

1.1.2.e Les sinistres climatiques

Certaines catégories de sinistres ne peuvent être modélisées par l'approche Fréquence × Coût Moyen. Les risques climatiques comme CatNat et TGN en font partie. Comme l'explique Alexandre Mornet [32] dans sa thèse, les approches possibles se décomposent en différentes topologies comme les modèles physiques, les modèles de statistiques sur des données simulées et ceux sur des données brutes. Chacune des méthodes possèdent des avantages et des inconvénients explicités dans le chapitre 7.

Le principe de la modélisation climatique

Les sinistres dit "climatiques" et "CatNat" ont des propriétés différentes de celles des sinistres attritionnels ou graves. Les différences sont les suivantes.

— Les sinistres sont interdépendants spatialement. En effet, si votre voisin est impacté par des inondations, il est extrêmement probable que votre maison soit impactée par un événement d'une même intensité. En conséquence, il est judicieux de parler d'évènements "climatiques" (ou CatNat). Il est important de prendre en compte ces dépendances spatiales même lointaines pour les événements de faibles probabilités comme l'explique Nguyen, 2020 [33]. La dépendance spatiale s'examine rigoureusement pour les risques dits CatNat de faible fréquence et d'intensité importante.

— La fréquence d'un évènement climatique et son impact monétaire émanent de l'intensité de l'évènement. En conséquence, l'indépendance entre la fréquence, l'évènement climatique et son coût est fondé si l'évènement est conditionné pour une intensité donnée. Ici, on entend par intensité d'un évènement climatique l'ensemble des indicateurs physiques - climatiques - d'un évènement (ex : la durée d'une inondation, la hauteur d'eau, la force d'écoulement, la zone d'impact).

En conséquence, notons I une variable aléatoire représentant l'intensité d'évènements climatiques. Rappelons que pour les sinistres attritionnels, le coût d'un ensemble de contrat L peut être évalué tel que :

$$\begin{aligned}\mathbb{E}\Big(\frac{L}{v}|\mathbf{X}\Big) &= \mathbb{E}\Big(\mathbb{E}\Big(\frac{NS}{v}|\mathbf{X}, N\Big)|\mathbf{X}\Big) \\ &= \mathbb{E}\Big(\frac{N}{v}\mathbb{E}(S|\mathbf{X}, N)|\mathbf{X}\Big) \\ \text{Par indépendance } &= \mathbb{E}\Big(\frac{N}{v}|\mathbf{X}\Big)\mathbb{E}(S|\mathbf{X}).\end{aligned} \tag{1.5}$$

15. Ces seuils peuvent être aussi fonction des caractéristiques du bien. Néanmoins, les actuaires sont assujettis à une contrainte d'interprétabilité des résultats dans le temps. En d'autres termes, un trop grand nombre de seuils ou les changements de seuils trop fréquents rendent difficile l'analyse des résultats ou des changements.

Au contraire pour les risques climatiques, il est nécessaire de conditionner au minimum par l'intensité pour obtenir une indépendance.

$$\begin{aligned}\mathbb{E}\Big(\frac{L}{v}|\mathbf{X}\Big) =& \mathbb{E}\Big(\mathbb{E}\Big(\frac{NS}{v}|\mathbf{X},I\Big)|\mathbf{X}\Big)\\ \text{Par indépendance } =& \mathbb{E}\Big(\mathbb{E}\Big(\frac{N}{v}\Big|\mathbf{X},I\Big)\mathbb{E}(S|\mathbf{X},N,I)|\mathbf{X}\Big)\\ =& \mathbb{E}\Big(\mathbb{E}\Big(\frac{N}{v}\Big|\mathbf{X},I\Big)\mathbb{E}(S|\mathbf{X},I)|\mathbf{X}\Big).\end{aligned} \tag{1.6}$$

Z Du fait de l'indépendance entre les individus, $\mathbb{E}\Big(\frac{N}{v}|\mathbf{X}\Big)$ est estimable à l'aide de techniques statistiques usuelles. Au contraire, pour un évènement avec une intensité I donnée, $\mathbb{E}\Big(\frac{N}{v}\Big|\mathbf{X},I\Big)$ il n'y a pas d'indépendance entre les individus. Cette hypothèse d'indépendance est plus raisonnable pour l'évaluation $\mathbb{E}(S|\mathbf{X},I)$. En effet, le coût des dommages dépend majoritairement des caractéristiques des biens assurés de la maison. L'évaluation de la fréquence et de l'intensité I s'élabore de manière générale par un modèle d'aléa, l'évaluation de $\mathbb{E}\Big(\frac{N}{v}\Big|\mathbf{X},I\Big)$ par un modèle de vulnérabilité et $\mathbb{E}(S|\mathbf{X},I)$ soit dans le module de vulnérabilité, soit dans le module de financier. L'ensemble des calculs s'élabore à différentes mailles : maille individuelle (ex : modèles marché), communale ou départementale, par grilles... Le graphique 1.7 illustre la structure des risques climatiques par un réseau bayésien.

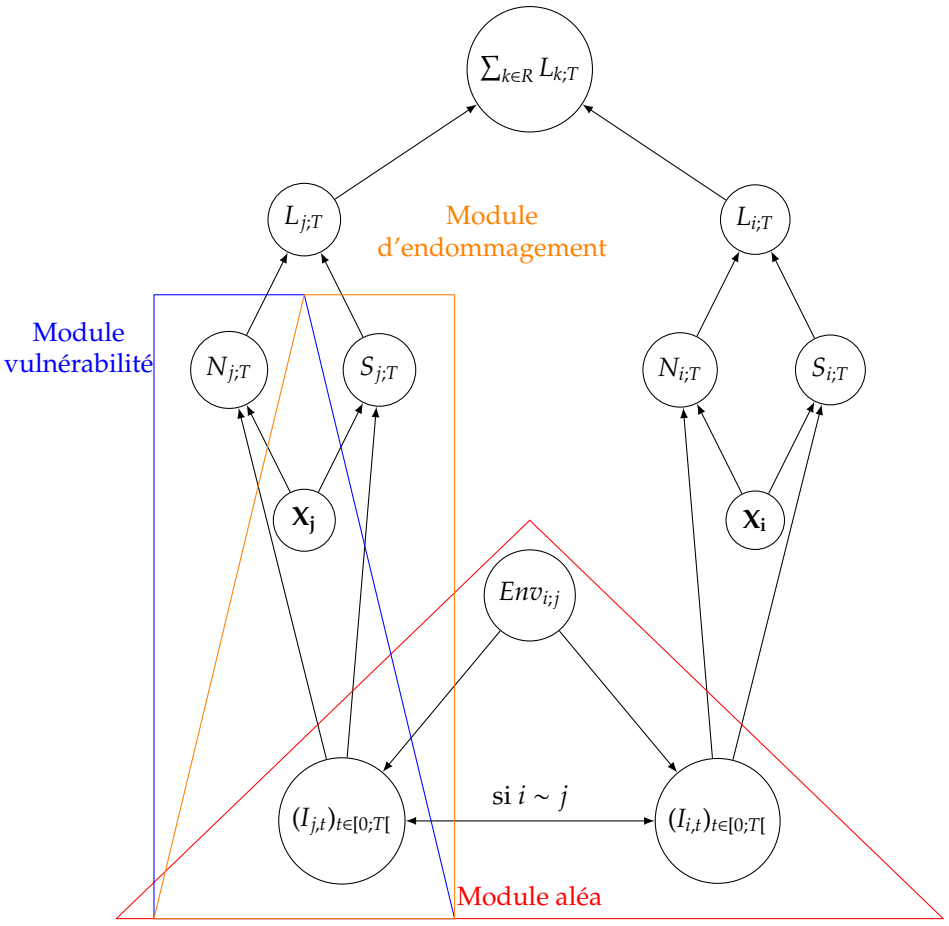


FIGURE 1.7 – Vision bayésienne des risques climatiques. Comme les risques climatiques sont des évènements temporels, il est nécessaire de considérer une période $[0;T[$ d'intensité I en amont pour la sinistralité à la date T. L'intensité I dépend des caractéristiques de l'environnement $Env_{i;j}$ liées aux individus i et j. Rigoureusement, $Env_{i;j}$ peut être vu comme un sous-ensemble d'information de $\mathbf{X}$.

Remarque : L'indépendance des individus pour l'évaluation $\mathbb{E}(S|I, \mathbf{X})$ peut être discutée ; des plans de préventions ou des actions de la municipalité peuvent contredire cette hypothèse d'indépendance. Néanmoins, cela semble négligeable au regard de la sinistralité. Toutefois, (Erhardt, 2020) met en avant à travers des études historiques qu'une dépendance entre des groupes de contrats apparait du fait du réchauffement climatique.

Les modèles pour estimer les risques climatiques sont divisés en modules :

1. Module d'aléa modélisant I ;
2. Module de vulnérabilité ou d'exposition [16] modélisant $\mathbb{E}(N|I, \mathbf{X})$;
3. Module d'endommagement (ou de dommages) modélisant $\mathbb{E}(S|I, \mathbf{X})$;
4. (Hors scope) Module financier calculant les coûts après les traités de réassurance et dépendances spatiales.

Différentes méthodologies et particularités sont détaillées dans le chapitre 7 avec une application aux risques de subsidences. Si pour les garanties TGN, ELEC pour la foudre ou DDE pour l'inondation non CatNat, certains assureurs segmentent leur niveau de primes. Pour les CatNats, le tarif est réglementaire et est égal à 12% des primes de dommages. Cette méthodologie pour les risques non CatNats est trop complexe par rapport à la contrainte 7 sur la capacité d'adapter les tarifs facilement. En conséquence, les assureurs simplifient les modèles en les regroupant dans d'autres garanties ou en déterminant des zoniers à larges zones. À travers une mutualisation large, les travaux de tarifications sont simplifiés pour ne considérer que $\mathbb{E}(N)$ et $\mathbb{E}(S)$ et proposer une tarification facilement adaptable et compréhensible.

En un mot, les modèles climatiques pour la tarification sont souvent assez simples. En revanche, les travaux pour l'acceptation du risque sous ces diverses formes nécessitent souvent une précision plus fine et moins contraignante opérationnellement. En effet, si la structure tarifaire est réglementaire, l'acceptation d'une souscription ne l'est pas. À partir d'indicateurs, un assureur pourrait refuser des biens dont le montant espéré des sinistres climatiques est au-dessus d'un certain seuil. Cependant, ce type de règles d'acceptation est souvent trop abrupte [17] et les assureurs préfèrent maîtriser la souscription plus finement. Par exemple, chaque commercial a le droit de faire souscrire un nombre maximum d'habitations assujetties à un risque climatique fort. Une autre méthode consiste à diminuer le montant des éventuelles ristournes en fonction de l'exposition au risque climatique de l'assureur.

1.1.2.f Les primes d'assurances

Le processus final du calcul de la prime commerciale Π^{comm} est détaillé dans les prochains paragraphes. Pour déterminer cette prime, différentes étapes sont faites et chacune d'entre-elles permet de calculer un type de "primes".

Les noms des primes ne font pas toujours consensus entre les différentes entreprises. Dans cette thèse, Π^{comm} sera la prime proposée à l'assuré. La prime HT commerciale $\Pi^{comm\ HT}$ correspond au montant perçu par la compagnie pour régler les sinistres, les frais de courtage, payer ses employés, augmenter ses fonds propres... Cette prime permet de calculer des ratios S/P ou le chiffre d'affaires HT. La prime chargée avec taxes $\Pi^{charge\ TTC}$ est la prime qui aurait été proposée s'il n'y avait aucune concurrence. La prime pure Π^{pure} est la meilleure estimation du coût d'un contrat en prenant en compte les indemnisations, les frais de gestion de sinistres... Le coût d'un contrat net de chargements est appelé Π^{nette}, la prime nette par garantie g est notée $\Pi_{garantie\ g}$.

Prime nette et pure : Dans cette thèse, la prime pure est définie comme la meilleure estimation des coûts d'un contrat y compris les frais de gestion de sinistres. Dans certaines entreprises, la prime nette est appelée prime pure ou prime pure technique. L'idée est que pour certaines garanties (PJ, CatNat sécheresse), les frais annexes sont trop importants pour ne pas être considéré dans le coût total. De plus, si le coût des actions de préventions, de conseils ou de sensibilisations est associé à la prime pure, la notion de prime "nette" me semble plus adaptée pour être la prime nette de ces éléments. Comme la législation tend vers plus de conseils et de préventions de la part des assureurs, ces coûts ne sont plus globaux ni anecdotiques.

16. Dans certain cas, le module de vulnérabilité regroupe aussi l'endommagement
17. Il faut se rappeler le rôle de la **BCT**.

❶ *Prime pure et prime nette*

La première étape est d'établir la prime nette $\Pi_{garantie\ g}(c)$ par garantie, en d'autres termes de réunir l'estimation de la sinistralité grave, attritionnelle et les autres types de sinistres. Certains assureurs consolident les modèles pour simplifier la structure tarifaire de $\Pi_{garantie\ g}(c)$.

Entre les garanties d'un produit MRH, les assureurs somment les primes nettes modélisées. En effet, comme l'espérance est linéaire, la prime nette finale d'un contrat c se calcule comme suit,

$$\Pi^{nette}(c) = \sum_{garantie\ g} \Pi_{garantie\ g}(c), \tag{1.7}$$

où $\Pi^{nette}(c)$ représente la meilleure estimation du coût de la sinistralité d'un contrat c.

Remarque : Des dépendances inter-garanties existent sur les sinistres graves. En effet, un sinistre grave réduit occasionnellement la probabilité de déclaration d'un sinistre attritionnel. Généralement, cette problématique faiblement impactante n'est pas considérée. Pour certains sinistres graves, le temps de réparation du bien peut être soustrait à la réelle exposition pour l'ensemble des garanties. Cela permet d'ajuster en amont la fréquence en prenant en compte la dépendance inter-garantie. Les impacts sont à la marge. Je n'ai observé qu'une seule fois ce type de modification.

Ensuite, il est nécessaire d'ajouter les chargements : les frais de gestion de sinistres, les frais de gestion administratifs, les frais de réassurances, de courtage, d'IT au sens large et règles législatives. Cette prime appelée *pure* est finalement la meilleure estimation du coût d'un contrat. Elle est très souvent calculée à l'aide d'un facteur de chargement multiplicatif $\theta(c)$:

$$\Pi^{pure}(c) = (1 + \theta(c))\Pi^{nette}(c). \tag{1.8}$$

$\Pi^{pure}(c)$ est la meilleure estimation du coût d'un contrat c. Elle est égale à un facteur multiplicatif près à $\Pi^{nette}(c)$.

Dans certaines assurances, ce chargement multiplicatif est occasionnellement différentié entre les garanties. Cette différentiation est faite à l'aide de clef de répartition souvent grossière.

❷ *Prime commerciale*

Une fois la prime pure calculée, une *marge* sur le produit est appliqué, englobant un chargement de sécurité pour faire face à des risques de dérive de la sinistralité, des besoins prudentiels de fonds... Il est nécessaire d'ajouter les taxes et cotisations au divers fonds $\%_{taxes\ garantie}$. La prime "commerciale" hors concurrence est recalculée par garantie :

$$\begin{aligned}\Pi(c)^{charge\ TTC} = frais + \sum_{garantie\ g} \Pi_{garantie\ g}(c) \times (1 + \%_{taxes\ garantie}) \\ \times (1 + \theta(c)) \times (1 + marge),\end{aligned} \tag{1.9}$$

pour simplifier, les primes réglementaires pour les garanties attentats et Catastrophes naturelles sont aussi notées $\Pi_{garantie\ g}(c)$. Les frais sont en partie proportionnels à $\Pi^{comm}(c)$ (comme les frais de distributions).

Comme $\Pi(c)^{charge\ TTC}$ n'est pas compétitive, la concurrence des autres assureurs doit être prise en compte. Le calcul Π^{comm} nécessite de créer un modèle optimisant la **transformations des devis** ou le **taux de rétentions**. Des modèles ont été faits dans le cadre de la tarification à l'adresse, mais ne seront pas détaillés ici. Complexes, ils sont propres à chaque assureur et avec de nombreux paramètres. La méthodologie pour déterminer la prime commerciale, notée $\Pi^{comm}(c)$, dépend de la $\Pi(c)^{chargeTTC}$ estimée pour le contrat c, $\Pi^{pure}(c)$, des primes commerciales des autres concurrents $(\Pi^{conc_i}(c))_{i\in\mathbb{N}}$ si disponibles, des chargements opérationnelles $\theta(c)$ et de l'élasticité au prix $e_{prix}(c)$ du potentiel souscripteur. Finalement, la prime $\Pi^{comm}(c)$ est normalement déterminée avant le calcul de la prime $\Pi^{comm\ HT}(c)$. En faisant le rapport, le coefficient de marge est obtenu $marge_{comm}(c) = \frac{\Pi^{comm}(c) - frais}{\Pi(c)^{charge\ TTC} - frais} - 1$.

La prime $\Pi^{comm\ HT}(c)$ est finalement égale à :

$$\Pi^{comm\ HT}(c) = \Pi^{comm}(c) - frais$$
$$- \sum_{garantie\ g} \Pi_{garantie\ g}(c) \times \%_{taxes\ garantie\ g} \times (1 + \theta(c)) \times (1 + marge_{comm}(c)),$$

où $frais$ se rapporte aux frais directement prélevés par exemple les frais d'acquisitions. La prime $\Pi^{comm\ HT}(c)$ correspond au montant disponible pour la gestion du contrat c.

❸ *Calculette tarifaire*

En réalité, cette prime commerciale n'est pas implémentable facilement. Il va être nécessaire de la faire transparaitre dans une grille tarifaire appelée **calculette tarifaire**. Il existe des variations dans la manière d'implémenter cette calculette, mais les variantes ont une structure très similaire. De manière très générale, une calculette tarifaire est composée de trois volets :

① Calcul de la prime technique commerciale avec ou sans marge par garanties $\Pi_{garantie}(c) \times (1 + marge_{comm}(c))$;
② Calcul de la prime chargée ;
③ Calcul de la prime TTC.

La forme de la calculette est multiplicative pour le calcul de $\Pi_{garantie}(c)$ *c-à-d* $\Pi_{garantie}(c) = M_0 \prod_{\forall h \in 1,\dots p} M(X_h(c))$ avec M_0 la prime de référence et $M(X_h(c))$ le multiplieur ou coefficient multiplicatif associé à la valeur de la variable $X_h(c)$. Une grille comporte environ 200 multiplieurs par garantie. Je propose au lecteur intéressé de consulter l'annexe E pour un exemple de grille tarifaire. Cette grille permet de vérifier l'ensemble de contraintes mentionnées précédemment :

- **Contrainte de proportionnalité 5 :** Comme le calcul de chaque garantie est fait par une simple multiplication, il est aisé de vérifier que les multiplieurs ne soient pas extrêmes. Par exemple, pour vérifier la cohérence spatiale, il suffit de regarder sur une carte les multiplieurs par zone. Les effets d'interactions ou structures complexes sont représentés directement en tant que variable tarifaire et associé à un multiplieur.
- **Contrainte de constance temporelle 1 et 8 :** Cette contrainte est aisément vérifiable. Il suffit de ne pas avoir de critères annuels et poser l'exposition du contrat égale à 1 ou 0.5 dans le cas de prime semestrielle. L'impact de l'exposition doit être égal à 1 [18].
- **Contrainte de borné inférieurement 2 :** La rentabilité se vérifie en comparant la prime nette et la prime commerciale HT. Dans le cadre de la grille tarifaire, il suffit de vérifier la différence des primes sur l'ensemble des contrats. Souvent, cette étape est déjà implicitement prévue dans le calcul de la marge. Néanmoins, le côté multiplicatif permet de s'assurer qu'il n'y ait pas un contrat avec des critères spécifiques ayant une prime proposée anormale.
- **Informations adaptées 3 :** Il suffit de vérifier que les critères soient compréhensibles et que la variabilité des multiplieurs pour une même variable ne soit pas trop important. Le temps de réponse est rapide ; avec 15 garanties et 20 de critères tarifaires, le nombre de calculs est faible et peu complexe (addition et multiplication). Ainsi, il est instantanément calculé à l'aide de n'importe quel outil même le plus basique.
- **Informations cohérentes 4 :** Le calcul de la prime est multiplicatif. Pour vérifier que les options prises augmentent bien la prime, ou que l'augmentation des franchises diminuent la prime, il faut contrôler que les multiplieurs soient au-dessus ou en dessous de 1. Dans le cas où plusieurs garanties utilisent le même critère, il suffit de prendre un individu de référence et de le comparer avec ou sans l'option. Comme la prime est multiplicative, l'évolution est similaire pour tous les autres contrats. Pour vérifier facilement la cohérence pour les variables continues surtout pour les effets non linéaires, il est usuel de discrétiser la variable par intervalle.
- **Contraintes législatives et continuités 6 et 7 :** Les contraintes sont vérifiées par la méthode de construction de la prime.
- **Tarifs adaptables 9 :** Si une des contraintes précédentes n'est pas vérifiée et/ou il est nécessaire d'ajouter une option, la grille tarifaire permet aisément de modifier l'impact de tels ou tels critères en changeant/ajoutant un multiplieur. Tous les impacts sont aisément calculables. Cette contrainte est vérifiée s'il n'y a pas trop de variables à étudier. Pour des besoins de mutualisations, il est aisé d'adapter les coefficients en supprimant les multiplieurs pour le critère de l'âge ou réduisant les multiplieurs de certaines zones géographiques.

18. En pratique, la prime annuelle est calculée. Ensuite, elle est divisée par 1, 2, 4, 12 en fonction du choix de la périodicité.

Il faut retenir pour le reste de la thèse que peu importe les modèles implémentés pour la tarification et les variables utilisées, la structure tarifaire est une grille. Pour déterminer cette grille, il sera nécessaire de consolider les modèles par garanties en des modèles multiplicatifs, cette étape est appelée consolidation.

La consolidation est le principe de poser une structure multiplicative sur une garantie. De manière générale, elle est faite une fois pour la prime nette et une seconde fois pour la prime commerciale pour prendre en compte les coefficients de marges, de mutualisation ou toutes autres évolutions tarifaires. Ainsi, les modèles sous-jacents doivent rester simples car les modèles à structure trop complexe ne peuvent pas apporter plus d'informations que peut contenir une grille. En pratique, la prime commerciale n'est plus la meilleure estimation du coût d'un contrat à la marge près.

En conclusion, la prime commerciale Π^{comm} est approximativement proportionnelle à la prime nette Π^{nette} pour un contrat donné. C'est pourquoi les métriques de segmentations et de performances calculées sur Π^{nette} permettent d'estimer les gains financiers, de transformations ou de rétentions. La prime nette sera étudiée dans le reste de la thèse. Cependant, l'optimisation tarifaire permet d'accentuer les effets de transformations ou de marge à partir de la connaissance du bien.

1.1.3 Hyper-individualisation, mutualisation, aléa moral et anti-sélection

Comme la tarification à l'adresse modifiera les parcours de souscriptions et les éléments de calcul d'une prime en profondeur, je propose de revenir sur les concepts et l'environnement concernant l'assurance MRH.

1.1.3.a Les différents principes

Les risques que le processus de tarification essaye d'évaluer ne sont pas exogènes. Les différentes parties liées par le contrat d'assurance ont un comportement différentié modifiant les risques et leur intensité. La raison du biais comportemental est l'asymétrie d'information. L'asymétrie d'information est le fait que chacun des parties connaissent des éléments du contrat que les autres ne connaissent pas. Cette dissymétrie induit deux phénomènes : l'aléa moral et l'anti-sélection (sélection adverse). Pour faire face à ces deux phénomènes, le marché de l'assurance oscille entre mutualisation et individualisation de la prime d'assurance.

❇ L'aléa moral

Un contrat d'assurance découle de l'aversion aux risques de l'assuré qui est prêt à payer pour réduire son risque. De cette aversion aux risques, le prix Nobel Arrow en 1963 en explique la notion d'aléa moral. Après la souscription d'un contrat, l'assuré aura tendance à avoir un comportement plus à risque. L'aléa moral est un biais comportemental où un agent modifie son comportement après la souscription d'un contrat d'assurance.

En tarification, l'évaluation des risques s'appuie sur une base historique de sinistralités. L'aléa moral est observable en moyenne pour la structure tarifaire passée. En partie pour ces raisons, des notions de franchises et de revalorisations/résiliations ou de bonus-malus permettent de contrôler l'aléa moral. Ces différents éléments diminuent les déclarations de sinistres à faibles coûts et à hautes fréquences. Ainsi en assurance habitation, les assurés ont intérêt à déclarer uniquement les sinistres importants. Ces derniers (ex : Incendies, Inondations, Vols) ont des impacts sur la qualité de vie suffisamment importants pour que l'assuré n'ait pas trop d'intérêt à modifier son comportement. Une prime doit donc être adaptable (contrainte 9) pour ajuster ces caractéristiques afin de limiter l'aléa moral.

❇ L'anti-sélection

Arrow (1963) comme Rothschild et Stiglitz (1976) ont étudié le principe d'anti-sélection. En assurance, l'anti-sélection provient du fait que l'information supplémentaire connue par l'assuré lui permet de sélectionner un assureur proposant la prime la plus basse.

L'assurance habitation en France est un produit avec un fort taux de rétentions car l'arrêt d'un contrat habitation est majoritairement liée à un changement d'habitation (achat, changement de locataires,

travaux ...), une augmentation significative de la prime ou suite d'une mauvaise gestion de sinistres de la part de l'assureur. De plus, elle est quasiment systématiquement souscrite et même obligatoire pour les locataires. Ainsi une mauvaise sélection impacte sur le long terme l'assureur et se contrôle majoritairement lors de la tarification.

En assurance, une solution est d'accroitre la segmentation des primes pour limiter ce biais comme démontré en assurance vie (1860, [24]) historiquement et modélisé par Rothschild et Stiglitz (1976). L'assureur trie et sélectionne ses assurés par sa structure tarifaire en choisissant son niveau de segmentation et de mutualisation.

❇ Segmentation et mutualisation

Si la prime est trop élevée par rapport aux risques, les bons risques (assurés) ont moins d'intérêt de souscrire une assurance et réciproquement avec les sous-tarifés. Pour restreindre cette anti-sélection, il est nécessaire de segmenter. La segmentation est le simple fait de proposer des primes différentes en fonction d'un ou plusieurs critères objectifs. (Frezal et Barry 2020, [29]) expliquent que la segmentation en tarification est *"not unfairly discriminatory"* ou *"actuarial fairness"* car les personnes aux caractéristiques similaires sont tarifées de la même façon. Par exemple, la prime doit être proportionnée aux risques protégés (contrainte 5) ; l'assurance d'une chambre de bonne doit être plus faible que celle d'un duplex de 300 mètres carrés.

Historiquement avec les informations disponibles du fait de la contrainte 3 de pertinence et de temps, la segmentation se faisait sur des grands groupes homogènes. Cependant, avec l'ajout de données externes (comme dans le cadre de cette thèse) il n'est plus possible de parler de groupes homogènes au niveau de la tarification, mais plutôt d'individualisation. En effet, avec une vingtaine de critères tarifaires, les groupes homogènes ne sont de l'ordre que d'une poignée de personnes [19].

Cependant, la notion de segmentation n'est pas antinomique à la notion de mutualisation et d'assurances. La mutualisation d'un produit d'assurance s'applique sur un portefeuille d'assurés sur deux angles : la mutualisation purement probabiliste (De Wit et Van Eeghen (1984)[28]) et une répartition de la marge des contrats (solidarité).

❇ Mutualisation en variance

Grâce au théorème centrale limite, la moyenne des réalisations pour les variables indépendantes mais non identiquement distribuées convergence sous conditions (Voir Lyapunov, 1901 [31]) [20].

Théorème 1.1.1

Soient $(X_k)_{k\in\mathbb{N}^*}$ un ensemble de variables indépendantes d'espérance μ_i et de variance σ_i finies. Si pour δ positif, $|X_k|$ possède un moment d'ordre $2+\delta$ et la condition de Lyapunoov est vérifiée :

$$\lim_{n\longrightarrow\infty} \frac{1}{(\sum_{i=1}^{n}\sigma^2)_i^{1+\delta/2}} \sum_{i=1}^{n} \mathbb{E}(|X_i-\mu_i|_i^{2+\delta}) = 0,$$

la somme des variables standardisées converge en distribution vers une loi normale centrée réduite quand n tend vers ∞ :

$$\frac{1}{(\sum_{i=1}^{n}\sigma_i^2)^{1/2}} \sum_{i=1}^{n} X_i-\mu_i \xrightarrow{d} \mathcal{N}(0,1).$$

En MRH, posons X la sinistralité. La variance σ_i et l'espérance μ_i pour tout individu i sont finies et peuvent être supposées bornées. De plus, pour $\delta = 1$, on peut aisément supposer la condition de Lyapunoov vérifiée. Comme σ_i et μ_i sont bornées pour tout i, $\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$ et $\frac{1}{n}\sum_{i=1}^{n} X_i-\mu_i$ convergent et l'écart type du risque moyen assuré est égale à :

$$\frac{1}{n}\Big(\sum_{i=1}^{n}\sigma_i^2\Big)^{1/2}. \tag{1.10}$$

19. Sans même parler de la date de souscription et du canal de souscription qui va influencer la souscription
20. Il existe des conditions plus faibles comme celle de Lindeberg, 1926.

Cette réduction apparait peu importe la segmentation. Cette mutualisation en variance permet de mutualiser les coûts de gestions des sinistres et les fonds prudentiels. Concrètement, le nombre de sinistres déclarés annuellement est connu en amont ce qui permet aux équipes de gestion d'être proportionnées et de mettre en place des processus industrialisés. Néanmoins, le seul bémol est que tous les sinistres ne sont pas indépendants comme ceux liés aux évènements CatNat. La diversification et la mutualisation de la variance ne sont donc pas parfaites.

❇ **Mutualisation en marge**

L'autre type de mutualisation est celle des primes Π^{comm} ou de la marge des contrats. Son absence correspond à l'individualisation du tarif. La mutualisation des tarifs a plusieurs objectifs, en particulier de limiter les biais de type II et de type III (Barry et Charpentier 2022, [25]). Le biais de type I sont les variables qui n'ont pas d'effets significatifs (ou prouvés). Le biais de type II est celui des critères tarifaires qui ont une réalité statistique avérée, mais non causale (ex : Homme/Femme en automobile). Le biais de type III est celui des critères tarifaires qui ont une réalité statistique, mais qui relève du cadre de discriminations sociales. En jouant sur ces critères, les assureurs jouent un rôle social en mutualisant. Selon les assureurs, la notion de discriminations n'est pas la même. Parmi les biais qui font l'objet de mutualisation par l'ensemble des assureurs sont le sexe/genre (administratif), les personnes seules avec enfants et certaines catégories socio-professionnelles. Pour la majorité des mutuelles, l'âge en MRH est souvent retiré de la structure tarifaire. Pour éviter certaines tentations de segmentations, des lois interdisent la collection de certaines informations en amont comme en Europe les croyances religieuses, l'orientation sexuelle, l'engagement syndical...

Dans le cadre de la tarification à l'adresse, je peux dire sans détours que les données à l'adresse permettent une tarification hyper-individualisée sans difficulté et sans même considérer la qualité des données. Cependant, les informations accessibles sont essentiellement sur le bâtiment. Ainsi les biais de type III sont plus faibles pour ce type de données. Pour finir, dans certain cas, inclure des critères des biais de type II et III pour ensuite les mutualiser permet d'éviter que les tarifs apprennent indirectement des biais par corrélation à l'aide d'autres critères.

1.1.3.b Un monde concurrentiel et la tarification inter-produits

L'anti-sélection est un phénomène assurantiel qui est amplifié par la concurrence entre les assureurs. Comme les assureurs n'utilisent pas les mêmes critères tarifaires, cela complexifie les impacts des choix tarifaires. De façon générale, un tarif plus segmenté que celui des concurrents permet d'augmenter leur phénomène anti-sélection tout en diminuant le sien. Cependant, la souscription d'une assurance n'est pas déterminée uniquement par le montant de la prime. Les travaux actuariels sur le taux de transformation des devis ou l'impact commercial montrent l'hétérogénéité des souscriptions. L'acceptation des contrats dépend aussi des remises commerciales, des mois de souscription, des produits proposés en même temps... De plus, les prospects testent un nombre limité d'assureurs.

L'assurance MRH s'inscrit dans des actions commerciales avec d'autres produits. Les personnes ayant plusieurs contrats d'assurances sont appelés les multi-équipés. Trois exemples seront mentionnés dans cette thèse :

- ✳ Une forte dépendance existe avec l'assurance automobile qui sert aussi de produit d'appel en attirant les prospects pour souscrire une assurance habitation en plus. Dans ce cadre, les marges des contrats sont souvent plus faibles ;
- ✳ Chez les banques-assureurs, l'assurance habitation a aussi un lien étroit avec l'obtention de prêts immobiliers ou l'assurance emprunteur en particulier. Dans ce cadre, la souscription d'une MRH est secondaire pour le commercial ;
- ✳ L'assurance habitation est aussi un produit d'appel pour des produits comme la PJ ou certaines assurances spécifiques (animaux de compagnies par exemple).

En fonction des situations, le tarif doit s'adapter à la valeur du client soit en se simplifiant, soit en étant plus compétitif, soit en étant plus rapide. Dans cet objectif, les caractéristiques supplémentaires permettent de mettre en avant des options à promouvoir (ex : panneaux solaires) et de proposer intelligemment une assurance habitation au client intéressé (appelé **lead qualifié**). L'objectif est de minimiser le temps associé des commerciaux pour le consacrer à d'autres produits et accompagnements.

1.1.3.c L'assurabilité d'un produit MRH

La notion d'assurabilité est un marronnier en assurance. Avec l'augmentation de la segmentation à travers des données et l'évolution des sinistres climatiques, cette thèse s'y intéresse.

L'article de Gollier 1997 [30] définit l'inassurabilité et l'inassurabilité partielle comme :

"a risk [is] un-insurable if, given the economic environment, no mutually advantageous risk transfer can be exploited by the consumer and the suppliers of insurance. Partial un-insurability occurs when the parties can exploit only part of the mutually advantageous transfer of risk." [21]

L'article (Charpentier, 2008 [26]) résume quelques critères nécessaires pour l'assurabilité d'un risque.

- Un contrat d'assurance doit posséder des aspects d'aléa pour être assurable légalement.
- Il doit comporter des risques diversifiables (Loi des grands nombres) pour être assurable actuariellement.
- Les parties concernées doivent se mettre d'accord sur un prix.
- De façon plus secondaire, le coût moyen doit être quantifiable et l'assureur se doit d'être solvable pour tous les évènements/risques couverts.

La majorité des risques couverts en l'assurance habitation sont assurables : diversifiables, aléatoires et solvables.

En France, la notion d'inassurabilité a des implications légales à travers la loi du 13 juillet 1982 - L. 125-1 du code des assurances. Celle-ci défini une catastrophe naturelle comme les *"dommages matériels directs non assurables ayant eu pour cause déterminante l'intensité anormale d'un agent naturel, lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises"*. Ces catastrophes sont prises en charge de façon spécifique par une protection de l'état. L'assurance MRH est très largement le premier produit associé en volume et en charge. Cette loi a eu pour objectif de rendre assurable des risques dit "inassurable".

En effet, le risque climatique, en particulier du fait des évènements les plus importants (CatNat), remet en question la diversification et la solvabilité à cause de la dépendance spatiale et temporelle. Les évènements CatNats sont des évènements de faibles occurrences avec des pertes associées importantes. L'évaluation des sinistralités n'est pas aisée car volatile avec des temps de développements longs. Par exemple, un évènement de sécheresses impacte un grand nombre d'habitations en même temps. Les assureurs/réassureurs doivent débourser des montants importants en même temps mais dont le montant final est connu 3 à 5 ans après l'évènement. Certains risques comme l'inondation, sont très localisés. Ainsi la notion d'aléa est diminuée. Une habitation au bord d'un fleuve sera très probablement impactée par au moins une inondation sur une vingtaine d'années. L'article de Charpentier and Le Maux, 2014 [27] met en avant une modélisation sur la théorie de l'utilité. Concrètement, si l'État agit pour prendre en charge les risques climatiques, les assurés ont peu d'intérêt à se couvrir contre les risques climatiques, car ce sont eux par les taxes qui payent les sinistres des non-assurés. Plus les risques sont corrélés, plus cet effet est important. Comme les assureurs prennent les pertes dans leur limite de solvabilité, le reste des pertes revient aux assurés. Pour répondre à cette problématique, les assurances CatNat en France sont obligatoirement attachées aux produits habitations, ont une prime réglementaire et une protection étatique par la CCR permet de diminuer les pertes.

Le dernier chapitre discutera de cette notion d'assurabilité pour le risque sécheresse.

Références

[24] (1860). *Railway Times*. Number vol. 23.

[25] Barry, L. and Charpentier, A. (2022). L'équité de l'apprentissage machine en assurance.

[26] Charpentier, A. (2008). Insurability of climate risks. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 33(1) :91–109.

[27] Charpentier, A. and Le Maux, B. (2014). Natural catastrophe insurance : How should the government intervene ? *Journal of Public Economics*, 115 :1–17.

21. Fr : un risque [est] non assurable si, compte tenu de l'environnement économique, aucun transfert de risque mutuellement avantageux ne peut être exploité par le consommateur et les assureurs. Il y a non-assurabilité partielle lorsque les parties ne peuvent exploiter qu'une partie du transfert de risque mutuellement avantageux.

[28] De Wit, G. and Eeghen, J. v. (1984). Rate making and society's sense of fairness. In *Premium Calculation in Insurance*, pages 151–169. Springer.

[29] Frezal, S. and Barry, L. (2020). Fairness in uncertainty : Some limits and misinterpretations of actuarial fairness. *Journal of Business Ethics*, 167(1) :127–136.

[30] Gollier, C. (1997). About the insurability of catastrophic risks. *Geneva Papers on Risk and Insurance. Issues and Practice*, pages 177–186.

[31] Lyapunov, A. M. (1901). A general proposition of probability theory. *CR Acad. Sci. Paris*, 132 :814–815.

[32] Mornet, A. (2015). *Contributions à l'évaluation des risques en assurance tempête et automobile*. Theses, Université Claude Bernard - Lyon I.

[33] Nguyen, V. D., Metin, A. D., Alfieri, L., Vorogushyn, S., and Merz, B. (2020). Biases in national and continental flood risk assessments by ignoring spatial dependence. *Scientific reports*, 10(1) :1–8.

1.2 Les outils statistiques et méthodologies utilisés

La structure particulière des données $\mathcal{D}_{train}$ à disposition permet d'utiliser différents outils mathématiques. Principalement, les méthodes se rapportent aux modèles linéaires généralisés (GLMs).

Les modèles linéaires généralisés (GLMs) ont été introduits par Nelder et Wedderburn (1972) [56]. Ce sont les modèles standards pour la tarification dans les compagnies d'assurances. Cependant, de plus en plus de méthodes dites "Machines Learning" sont utilisées (Dalonzo 2011 [40], Henckaert et al. 2018, [48]). Du boosting comme AdaBoost (Liu et al. 2014 [53]) ou XGBoost (Pesantez-Narvaez et al., 2019 [57]), à des modèles basées sur des algorithmes de forêts pour les modèles de sévérité (Wuthrich et al. 2019, [62]) ou de fréquence (Yang et al. 2018 [63]), c'est leur versatilité et leur performance qui ont accru leur popularité. Néanmoins, ces méthodes dépendent beaucoup plus des données et ce sont de plus des boites noires [22]. Comme les modèles utilisés en assurance sont régulés notamment par *European Union's General Data Protection Regulation* (GDPR ou RGPD, 2018 [39]), la transparence, la compréhension des algorithmes envers un prospect et l'analyse de la sélection adverse doivent être considérées (Cf Henckaerts et al. 2020 [49]).

1.2.1 Les modèles linéaires généralisés (GLM)

Cette partie parcours la théorie des GLMs et discute leur utilisation opérationnelle. La théorie reprend les notations du livre *Generalized Linear Models* de Mc Cullagh et J.A. Nelder.

1.2.1.a Généralité sur la théorie sur les GLMs

Le cadre théorique suppose que Y suit une distribution de la famille exponentielle, c'est-à-dire que la densité s'écrit en y :

$$f_Y(y, \theta, \phi) = exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\}, \tag{1.11}$$

avec $a(.)$, $b(.)$ et $c(.,.)$ des fonctions définies avec les paramètres reéls θ et le paramètre de dispersion réel ϕ. En actuariat, les distributions usuelles sont la distribution de Poisson et Bernoulli pour les valeurs discrètes, les distributions Gaussienne et gamma pour les valeurs continues et pour les variables hybrides la distribution Tweedie.

La base pour la modélisation $\mathcal{D}_{train}$ en amont est divisée $\mathcal{X}_{train}$ et $\mathcal{X}_{test}$. Il n'y a pas de base de validation $\mathcal{X}_{validation}$. Pour trouver les paramètres, l'algorithme des GLMs minimise la déviance sur $\mathcal{X}_{train}$ (ou maximise la vraisemblance du modèle) sous l'indépendance des observations,

$$D(\hat{\theta}_i, y) = \sum_{i=1}^{n} w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))\}, \tag{1.12}$$

22. en EN : *Black Box c-à-d* la structure finale des modèles n'est pas humainement pleine comprise par le practicien.

avec $\hat{\theta}_i$ l'estimateur θ du modèle et $\tilde{\theta}_i$ celui du modèle saturé[23]. w_i est le poids associé à chaque observation et toujours égal à 1 en actuariat sauf exception.

Les GLMs dépendent d'une structure linéaire sur l'espérance de Y :

$$\mu(\theta) = \mathbb{E}(Y|\theta), \quad \mathbf{X}\beta = g(\mu) \tag{1.13}$$

avec g une fonction de lien à choisir. Les fonctions usuelles de liens en actuariat sont la fonction identité pour la distribution gaussienne, la fonction log pour les lois de Poisson et de gamma et la fonction $\mu \to log(\mu/(1-\mu))$ pour la distribution Bernoulli. De façon équivalente, les déviances s'écrivent aussi en fonction de $\mathbf{y}$ et $\hat{\mu}$. La prédiction $\hat{\mu}$ dépend de la somme linéaire des variables $\mathbf{X}\beta$. Au contraire pour les modèles GAM, CART, RF et autres, $\hat{\mu}$ dépend d'un paramètre de $\mathbf{X}$ non linéaire. Dans certains cas, il est nécessaire de calibrer β d'une variable manuellement *c-à-d* la variable est incluse dans un modèle, mais le coefficient est prédéfini. Cette variable est appelé un **offset**. Il peut y en avoir plusieurs. Les offsets sont importants pour prendre en compte l'exposition.

Résumé : L'objectif du GLM est de trouver des coefficients β sous la contrainte que $\mathbf{Y}$ suit une distribution donnée de la famille exponentielle en maximisant la vraisemblance. L'espérance de $\mathbf{Y}$ est liée par une fonction définie avec une structure linéaire en $\mathbf{X}$.

❋ *Interpréter un modèle GLM*

Les GLMs ont une structure stable et explicative. En effet, il suffit de connaitre β pour déterminer l'augmentation exacte d'une variable sur le résultat dans le modèle. Par exemple, si on ajoute une unité à la valeur X_1, le résultat final sera $\mu = g^{-1}(\mathbf{X}\beta + 1 \times \beta_1)$. Pour certaines distributions associées avec certaines fonctions de liens, les modèles sont plus simples à expliquer. Le choix de la distribution et de la fonction de lien sont sous les contraintes (3, 5 et 9) de temps, de proportionnalité et d'adaptabilité des coefficients.

⊷ **Exemple additif - Distribution normale et lien identité :**Ce cadre permet d'obtenir une structure additive sur les prédictions, dans le sens que si on ajoute une unité à la valeur X_1, l'évolution est égale à

$$\mu(X_1+1, X_2, ..., X_p) - \mu(X_1, X_2, ..., X_p) = 1 \times \beta_1.$$

⊷ **Exemple multiplicatif - Distribution gamma et lien log :** Pour les GLM log-Poisson et log-gamma, c'est la structure multiplicative qui est importante. Si on ajoute une unité à la valeur X_1, l'évolution est égale à

$$\frac{\mu(X_1+1, X_2, ..., X_p)}{\mu(X_1, X_2, ..., X_p)} = exp(1 \times \beta_1).$$

L'autre avantage de la structure multiplicative est qu'une variable V telle que $log(V)$ est un offset avec $\beta = 1$ a un effet proportionnel sur μ. Ainsi dans les modèles log-Poissons, le log de l'exposition est toujours en offset pour vérifier la contrainte 1 sur l'exposition.

Définition : Dans les GLM Log-Poisson et Log-Gamma, $exp(1 \times \beta_1)$ s'appelle le multiplieur de la variable X_1.

❋ *Évaluer la performance d'un modèle*

Pour évaluer la performance, différentes métriques sont utilisées. Les plus connues et utilisées sont résumées dans le tableau 1.3. Pour comparer des modèles, les métriques ne mesurent pas le même phénomène.

Les métriques "cibles" dépendent de la distribution sous-jacente du modèle. Si deux modèles supposent différentes distributions alors chacun des modèles sera souvent plus performant sur sa vraisemblance. Par exemple, généralement un GLM log-Poisson sera plus performant qu'un GLM logit sur un EDR poisson, mais moins performant sur un EDR Bernoulli. Pour les comparer, il serait nécessaire au minimum de vérifier lequel des modèles est meilleur sur l'AIC, le BIC et l'écart à la moyenne. Souvent, le

23. Modèle dont les prédictions sont égales aux observations.

gain significatif en performance se détecte sur les métriques que j'appelle "secondaire", MSE et MAE. Par ailleurs, la segmentation est un aspect important en actuariat d'où l'étude de métriques de segmentation comme le Gini.

	Métriques	Formules	Cible	Commentaire
Cibles	Log vraisemblance	$log\mathcal{L}$	Max.	La vraisemblance de la distribution. Ref [43]
	Déviance	$log\mathcal{L}$ -$log\mathcal{L}_{sat}$	Min.	Équivalent à la vraisemblance Ref [56]
	PseudoR	$1 - \frac{log\mathcal{L}}{log\mathcal{L}_0}$	Max.	Ici, celui de McFadden. Ref [55]. Voir l'article [50] pour d'autres PseudoR.
	EDR	$1 - \frac{Dev.}{Dev._0}$	Max.	Pourcentage de la déviance résiduelle apprise.
Perf.	$Ecart_{moy.}$	$\lvert\frac{1}{n}\sum_{i=1}^{n} y - \hat{y}\rvert$	Min.	
	AIC	$2(p+1) - log\mathcal{L}$	Min.	Estimateur "historique" qui se base sur la distance Kullback-Leibler. Ref [34]
	BIC	$ln(n)(p+1) - log\mathcal{L}$	Min.	Pénalise plus fortement le nombre de variables que l'AIC. Ref [59]
Second.	MSE	$\sum_{i=1}^{n}(y-\hat{y})^2$	Min.	Suppose une distribution gaussienne
	R^2	$1 - \frac{MSE}{MSE_0}$	Max.	Suppose une distribution gaussienne
	MAE	$\sum_{i=1}^{n}\lvert y-\hat{y}\rvert$	Min.	Métrique d'adéquation à la médiane
Segmentation	*AUC*	Aire sous la courbe	Max.	Voir l'annexe D pour le calcul des courbes possibles.
	Gini	$2AUC - 1$	Max.	Valeur entre -1 et 1. Doit être supérieur à 0.
	$Gini_{norm}$	$\frac{Gini}{Gini_{sat}}$	Max.	Proportion de segmentation maximum apprise

TABLE 1.3 – Les indices 0 et *sat* représentent les métriques évaluées sur les sorties respectivement un modèle avec un *intercepte* et un modèle saturé retournant $\hat{y} = y$.

Pour le reste de cette thèse, les termes suivants de gradation seront utilisés.

❒ **Meilleur** : Un modèle m_1 est meilleur qu'un modèle m_2 si sur un ensemble des métriques comportant des métriques cibles, de performance, secondaire et de segmentation, le modèle m_1 est préférable à m_2 pour chacune des métriques.

❒ **Performant** : m_1 est plus performant que m_2, si sur les métriques de performances communes aux deux modèles, le modèle m_1 est préférable à m_2.

❒ **Segmentant** : m_1 est davantage segmentant que m_2, si sur les métriques de segmentations communes aux deux modèles, le modèle m_1 est préférable à m_2.

Dans le cas où les métriques de performances sont différentes, c'est-à-dire quand la distribution sous-jacente est différente, une partie de la littérature considère que la comparaison de l'AIC ou le BIC permet de dire qu'un modèle est meilleur qu'un autre. Ce cas ne se présentant pas dans cette thèse, je ne discuterais pas de ce sujet complexe.

1.2.1.b GLM et modèles de coût moyen

Pour la modélisation du coût moyen, les modèles log-gamma sont très largement utilisés. Pour les modèles graves, il est préférable de faire des modèles log-Normaux car la queue de distribution est plus lourde. Pour de rares risques, les modèles linéaires peuvent être utilisés.

Comme la partie 1.1.2.c le dépeint, deux variantes existent pour les bases de modélisation ; soit une ligne correspond à un unique sinistre, soit une ligne correspond à la somme de l'ensemble des sinistres survenus durant une image. Dans le premier cas, un GLM log-gamma sans offset peut être fait. Dans le second cas, il est nécessaire poser $\omega_i = N_i$ comme le montre l'annexe B. Rappelons que des sinistres avec des montants négatifs ou nuls apparaissent dans la base. Ceux-ci sont écartés de la modélisation.

Pour le coût moyen, les métriques en assurances sont assez faibles et volatiles car le coût d'un sinistre est par nature aléatoire. Rappelons que les modèles prédisent un montant moyen pour un contrat et non une estimation d'une indemnisation en particulier. Cela se traduit par des EDR et des Ginis normalisés faibles. Ils diffèrent en fonction de la nature du bien. Le tableau 1.4 répertorie des ordres de grandeurs pour des modèles combinés appartement et maison.

Garanties	DDE	INC	VOL	BDG	ELEC	TGN	CATNAT
EDR Gamma (%)	5.5 - 6	7-8	5-7	3.5 -4.5	5.5-8	4-8	5-16
Gini Norm. (%)	25	24-30	26-30	20-23	23-30	21-32	22-34

TABLE 1.4 – EDR moyen pour les sinistres attritionnels pour des modèles *CM* combinés appartement et maison. Les valeurs dépendent aussi de la période historique prise en compte. Ici les métriques sont des résultats de simples GLMs-Gamma.

Dans un modèle de coût moyen, il y a entre 2 à 5 variables avec 3 à 4 modalités par variables, une variable sur la taille du bien en général et assez rarement des zoniers. En effet, le nombre d'observations limite la qualité des modèles et la stabilité des variables comme le zonier. Ce petit nombre est le produit entre la fréquence moyenne de la garantie et le nombre d'assurés : soit 3% pour le DDE en appartement ou 0.1% en BDG par exemple.

1.2.1.c GLM et modèles de Fréquences

Les modèles de fréquences sont souvent modélisés par des modèles GLMs log Poisson ou logit. Dans de très rares cas, des modèles binomiaux négatifs sont faits pour des garanties comme le BDG ou CatNat par exemple. Pour des modélisations d'évènements ne pouvant donner lieu qu'à un seul sinistre, il est préférable d'utiliser des modèles Bernoulli comme logit ou probit. Pour des probabilités faibles (inférieures à 1%), la littérature met en avant que les résultats entre modèles log-Poisson et logit sont quasiment équivalents.

Les modèles de fréquences en actuariat doivent toujours prendre en compte l'exposition V dans la modélisation de façon linéaire (contrainte 2). Automatiquement, pour les modèles log-Poisson ou probit, le $log(V)$ est rajouté en offset avec le β associé égal à 1. Tout comme les modèles de coût moyen, les métriques sont faibles comme le résume le tableau 1.5. En effet, il est impossible de prédire le nombre précis et le type de sinistres d'un bien assuré.

Garantie	DDE	INC	VOL	BDG	ELEC
EDR Poisson (%)	3	1.8	2.5-4	2.5	9-10
Gini Norm.(%)	25-30	28-30	30-40	30	60

Garantie	TGN-A	TGN-M	CATNAT-A	CATNAT-M
EDR Poisson (%)	2-6	3-8	1-7	3-12
Gini Norm.(%)	60	30-35	26-60	40-60

TABLE 1.5 – EDR moyen pour les sinistres attritionnels pour des modèles combinés Appartement(A) et Maison(M) en fréquence observée sur plusieurs assureurs. Les valeurs dépendent aussi de la période historique prise en compte. Ici les métriques sont des résultats de simples GLMs.

Dans un modèle de fréquence attritionnelle, le nombre de variables peut être important avec un nombre de modalités aussi important jusqu'à une cinquantaine de coefficients. De façon générale, le zonier s'ajoute aux variables. Il est à noter qu'en MRH, les modèles sont à la maille individuelle. Pour des flottes ou des contrats regroupant un grand nombre de bâtiments, les modèles sont calibrés légèrement différemment.

Le cas des modèles de propension ou de transformations/rétentions

Les modèles de propensions ou de transformations/rétentions ont pour objectif de prédire une variable booléenne avec 0 et 1. Ce sont des modèles logits ou probits qui sont utilisés. De préférence, en actuariat les modèles logits sont utilisés, car les *odds-ratios* s'interprètent de façon quasi-similaire aux multiplieurs des modèles log-Poisson dans le cas des faibles probabilités.

1.2.1.d Les modèles de prime pure

Dans certains cas, il est intéressant de modéliser directement la sinistralité $\mathbb{E}(L_i|\mathbf{X}_i)$. Pour cela, les modèles Tweedie sont utilisés.

La distribution Tweedie (Tweedie, 1984 [61]) est une généralisation de la famille exponentielle et possède deux paramètres p et α. Dans le cas particulier où $p = 1$, la distribution Tweedie est équivalente à celle de Poisson. Dans le cas $p = 2$, elle est égale à une distribution Gamma. En assurance, p est compris entre 1 et 2, est souvent proche de 1.5 en tarification assurantielle (Jorgensen et Paes De Souza, 2014 [51] et Smyth et Jorgensen, 2002 [60] et un exemple récent Delong et al. 2021 [42]).

Le désavantage de ce type de modélisation est que certaines variables ont des impacts marginaux différents entre le coût moyen et la fréquence. Ainsi ces variables risquent d'avoir des effets marginaux complexes et non monotones sur la prime pure. De plus, les observés sont plus volatils par nature. Néanmoins, les modèles de primes pures sont utiles pour les petites bases où la taille de l'échantillon ne permet pas d'obtenir des effets significatifs pour les modèles de fréquence et de cout moyen. De plus, ils ne supposent pas une indépendance du coût moyen et de la fréquence.

Le cas de la consolidation : L'autre approche des modèles de primes pures concerne la consolidation. L'objectif est de retrouver la prime pure et les coefficients à l'aide d'un seul modèle : quasiment exclusivement un GLM log-gamma pour son caractère multiplicatif. Pour cela, ce modèle se base sur les résultats des modèles de fréquence, de propension et de coût moyen attritionnels et graves. Le modèle résume la prime pure résultant d'opération de plusieurs modèles complexes en une grille multiplicative.

La particularité de cette approche est que la volatilité inhérente des sinistres disparait, car on essaye de modéliser une prime à partir de prédictions. De plus, les variables significatives ont déjà été choisies et sont connues. En revanche, il y a souvent des effets non linéaires et non monotones. C'est lors de la consolidation qu'il est nécessaire de vérifier l'impact final des variables pour vérifier les contraintes 5 et 9. Par exemple : une augmentation de la franchise d'une garantie doit faire diminuer le montant de la garantie.

1.2.1.e L'ajout d'interaction non linéaire

Interactions : Pour ajouter des effets non linéaires dans les GLMs, plusieurs démarches existent. L'idée simple est de créer des nouvelles variantes de variables. Les trois principales approches sont les interactions, les splines et l'ajout de zoniers. Ce dernier permet de prendre en compte des données externes à la souscription et la méthodologie en sera détaillée.

① Interactions et splines

Interactions : Il existe des effets d'interactions entre les variables tarifaires qui peuvent être rajoutés dans un GLM.

Soit deux variables X_1 et X_2 et une fonction $F(Y)$. Une fonction $F(Y)$ exhibe une interaction entre X_1 et X_2 si :

$$\mathbb{E}\Big(\frac{\delta^2 F(Y)}{\delta X_1 \delta X_2}\Big)^2 \geq 0, \tag{1.14}$$

(Condition de Friedman et Popescue, 2008 [46]) .

En actuariat, la méthode pour incorporer une interaction est de créer une nouvelle variable $X_{1\times 2} = X_1 \times X_2$ et de rechercher son coefficient β dans un GLM. Traditionnellement, X_1 ou/et X_2 sont des variables booléennes simplifiant la prise en compte d'interactions. La recherche des interactions peut se faire en connaissance de cause, en testant les interactions directement dans le GLM ou en s'aidant de

méthodes ML [24]. Ces interactions ne sont pas très contraignantes, mais ont l'inconvénient de démultiplier le nombre de coefficients.

En MRH, les interactions usuellement utilisées et significatives apparaissent entre les types d'habitations, le nombre de pièces et la qualité de l'occupant. En effet, l'effet marginal de l'augmentation du nombre de pièces est souvent différent si l'assuré est un locataire ou un propriétaire non occupant. L'autre approche pour considérer des interactions est d'effectuer un modèle par catégorie. Par exemple : un modèle pour les **LOC** et un modèle pour les **PNO** ou un modèle pour les appartements et un modèle pour les maisons.

Splines : Pour de nombreuse variables continues, les impacts sur la sinistralité ne sont pas linéaires. Il est donc nécessaire de transformer les variables en amont. Une méthode usuelle est d'utiliser des splines. Un spline est une application transformant une variable continue en une ou plusieurs fonctions polynomiales continues. Il existe différents types de splines (B-splines, splines cubiques, ...).

Dans cette thèse, l'ajout de splines double en moyenne le nombre de coefficients associés à la variable à estimer. De manière générale, en MRH, l'âge et les surfaces habitables font éventuellement l'objet de splines (Voir la figure 1.9 pour un exemple).

L'utilisation de spline complexifie la transposition dans une grille tarifaire. Ainsi, lors de la consolidation, les variables qui ont fait l'objet de splines sont discrétisées pour recueillir de simples multiplieurs interprétables.

Pourquoi utilise-t-on des splines pour discrétiser ensuite ? Les modèles *Freq* et *CM* apprennent leurs effets sur N et S respectivement qui sont de nature volatile. De ce fait, si une discrétisation est faite lors de la confection des modèles, les effets marginaux entre deux intervalles contiguës peuvent être contradictoires avec des sauts significatifs entre des valeurs proches de la variable. Dans ce cas, le modèle sur-apprend et entre en contradiction avec la contrainte de proportionnalité de la prime (contrainte 5). Au contraire, pour les modèles de consolidations, les modèles viennent réapprendre des effets déjà appris basés sur *Freq* et *CM*. Les multiplieurs finaux approximent les effets des splines.

② Les zoniers

Traditionnellement, les variables utilisées **X** proviennent directement du questionnaire de souscriptions. Une information nécessite des retraitements importants pour pouvoir être utilisée : **L'information géographique**. Il est évident que la position du bien assuré a une influence sur la sinistralité du fait de son climat (précipitations méditerranéennes, montagnes), du voisinage (densité, urbanisation) mais aussi de sa sinistralité historique (le meilleur exemple étant la garantie Vol). L'objectif d'un zonier est de regrouper par zone les communes proches avec des caractéristiques similaires. Le même principe se retrouve en assurance automobile avec le zonier géographique, mais aussi dans le véhiculier où l'objectif est de regrouper les types de voitures en fonction de leurs caractéristiques.

La compréhension de la fabrication d'un zonier est indispensable pour bien comprendre la tarification à l'adresse. Ces zoniers peuvent être incorporés aux modèles de *Freq*, de *CM* ou de prime pure en fonction de leur significativité.

Trois types d'approches pour la confection de zoniers existent et chacune est utilisée par différents assureurs :

- les zoniers basés sur les zones urbaines ;
- les zoniers basés uniquement sur la crédibilité des résidus géographiques ;
- les zoniers expliquant les résidus géographiques par des données externes.

Zoniers basés sur les zones urbaines : Ce sont des zoniers qui s'appuient sur le pôle géographique en fonction de la zone d'emploi et de la caractéristique méga-urbaine, urbaine, rurale ou isolée de la commune - graphique 1.8. Les groupes entre les communes se constituent avec les zones de mêmes expositions et caractéristiques et la significativité des coefficients associées. Cette approche a l'avantage d'être simple, interprétable et de segmenter des natures de risques différents entre les grandes villes comme Paris ou Marseille. De plus, la caractéristique urbaine est la plus informative géographiquement pour de nombreuses garanties en MRH. Cependant, elle a le défaut d'imposer des zones déjà délimitées. En conséquence, certaines garanties, par exemple VOL ou TGN, nécessitant un découpage plus fin ou différent, induisent un zonier peu stable. Pour éviter cet écueil, certaines assurances y ajoutent une interaction avec le département complexifiant la structure tarifaire.

Zoniers basés sur les résidus : Pour adapter les zones à la sinistralité, deux approches similaires se basent sur les résidus du GLM et en les agrégeant à la maille de la commune. La première approche

24. Voir le mémoire d'actuaire de Silvia BUCCI, Étude et implémentation de techniques d'analyse de sensibilité dans les modèles de tarification Non-Vie. Application à la tarification à l'adresse. 2021

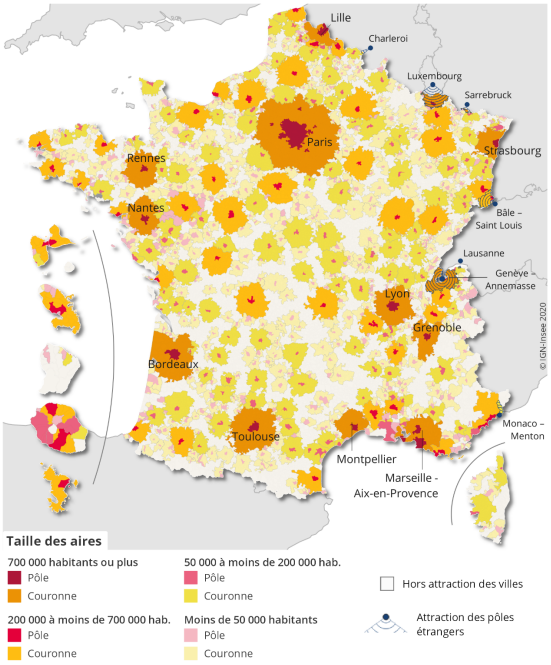


FIGURE 1.8 – Données open sources fournies par l'Insee. En France, neuf personnes sur dix vivent dans l'aire d'attraction d'une ville – Insee Focus – 211

lisse les résidus à l'aide d'approches de crédibilité ou de processus gaussien. Pour des raisons de performances et d'interprétabilité des méthodes, la seconde approche modélise les résidus à l'aide de données externes.

Le processus est le même pour les deux approches :

❶ Regrouper les prédictions des GLMs sans zoniers et les observations à la maille communale ou INSEE ;
❷ Calculer des résidus d'Anscombes (voir [35] et [54]) de préférence pour obtenir une distribution normale des résidus. En effet, des résidus non normaux comme les résidus Poissons sont difficiles à modéliser contrairement à une distribution normale ;
❸ Lisser ou modéliser des résidus pour réduire la volatilité de la survenance du risque et obtenir une prédiction indépendamment de l'exposition du portefeuille : par exemple, des méthodes de crédibilités peuvent être utilisées pour la première approche. Pour la seconde, des méthodes de statistiques (GLMs) ou Machine Learning (ML) permettent de modéliser les résidus à l'aide de données externes ;
❹ Regrouper les communes par score de risques : ce regroupement se fait généralement par Classification en Ascendantes Hiérarchiques ou similaire à partir des résidus lissés ou modélisés et peut considérer des informations géographiques/urbaines pour conserver une cohérence spatiale du zonier ;
❺ Réintégrer la variable zonier catégorielle dans le GLM. Pour valider le zonier, les résidus finaux obtenus doivent s'apparenter à un bruit blanc sur la carte de France.

Il existe des variantes équivalentes comme avec les variables non géographiques en offset. Les données externes utilisées sont souvent des données de populations présentes dans la commune, du type d'activités et une ou deux variables météorologiques comme les précipitations ou la température. Concrètement, les données externes sont utilisées en moyenne à la commune. Puisque les communes sont associées finalement à une zone, les données externes n'apparaissent pas dans la calculette tarifaire.

Quelles remarques sur le zonier
Retirer les variables géographiques : Il est souvent conseillé de ne pas ajouter les variables géographiques lors de la conception du premier GLM sans le zonier. Théoriquement, rien n'empêche d'ajouter de telles variables. En revanche, cela complexifie les analyses et les modélisations, car les variables géographiques initialement dans le GLM sont corrélées avec la variable de zonier. De plus, l'hypothèse sous-jacente d'un zonier est que les effets géographiques sont spatialement *smooth*, c'est-à-dire continues et lisses. L'ajout en amont d'une variable géographique risque d'entrainer des effets discontinus.

Catégoriser en zones : Il est possible d'ajouter le résidu modélisé directement comme un indicateur de risques. Si les modèles sont plus sujets à du sur-apprentissage, la contrainte est avant tout opérationnelle. En effet, les effets des résidus ne seront pas linéaires, obligeant à l'utilisation de splines et nécessitant de faire des calculs supplémentaires dans la calculette tarifaire (un par zonier). Finalement, lors de la ré-estimation d'un tarif ou de changement de critères tarifaires les valeurs des résidus évoluent. En conséquent, les comparaisons et évolutions de tarifs sont très complexes à analyser.

1.2.2 La sélection des variables

Pour cette thèse, la sélection de variables par GLMs est importante, mais pas innovante. Il est important de rappeler que les variables dans un modèle se choisissent "intelligemment" en tarification. Contrairement à certaines pratiques ML, la contrainte de cohérence 4 oblige les actuaires à ne pas standardiser les variables tarifaires. De façon générale, il y a peu de variables traditionnelles continues, car les questions sont souvent à choix multiples ou avec des réponses discrètes (Nombre d'enfants ou nombre de pièces) avec un nombre limité de modalités.

Au contraire, dans le projet de la tarification à l'adresse, les variables continues sont en grand nombre. Il est nécessaire de récupérer l'ensemble de définition de chaque variable continue sur $\mathcal{X}_{train}$ et d'appliquer ces bornes à $\mathcal{X}_{validation}$ et $\mathcal{X}_{test}$.

Pour les variables discrètes, il est préférable de conserver l'aspect linéaire-continue plutôt que de discrétiser. Néanmoins, sur le nombre de pièces ou l'âge, certains assureurs préfèrent mettre la variable de façon catégorielle augmentant significativement le nombre de coefficients, risquant d'avoir de la multicolinéarité, des effets marginaux incohérents et le sur-apprentissage. Dans l'outil EMBLEM (un outil marché de pricing), les variables continues sont catégorisées et ensuite une interpolation est réapprise sur les coefficients. Si les pratiques sont théoriquement valides, l'aspect continu est très souvent favorisée dans cette thèse au vu du nombre de nouvelles variables.

Pour les variables catégorielles, des regroupements entre les modalités peuvent être faits. Ces regroupements doivent avoir tout d'abord une justification actuarielle et logique. L'idée est d'obtenir des coefficients pour chacune des modalités explicables, éthiques et les plus robustes possibles.

Il existe un panel de stratégies différentes pour la sélection de variables. Les GLMs se basent essentiellement sur des tests de significativité sur les coefficients et du rapport de déviances entre les modèles. Un lecteur intéressé pourra se référer au livre de Nelder et Mc Cullagh. Les approches sont les tests de significativité, les méthodes *stepwises* ou bien s'aider de l'importance des variables fournies par d'autres méthodes.

1.2.2.a La significativité des variables et sélection des variables dans un GLM

Pour choisir entre deux modèles GLMs emboités [25], le test le plus commun est celui du rapport de vraisemblance où $D_{m_1} - D_{m_2} \sim \chi^2_k$ avec χ^2_k une loi khi-deux centrée de paramètre k et k est le nombre de coefficients en plus du modèle m_1 par rapport au modèle m_2. Ce test vérifie que le modèle m_1 est significativement plus performant que le modèle m_2. De façon similaire, des tests de Wald peuvent être fait directement sur les coefficients pour étudier leur significativité, c'est-à-dire que leur effet marginal est bien différent de 0.

Une approche naïve de sélection consiste à tester exhaustivement l'ensemble des combinaisons des variables. Pour p variable, cela revient à tester 2^p modèles. Pour être plus efficace, des méthodes *stepwise*

25. C'est-à-dire toutes les variables d'un des modèles sont présents dans l'autre.

- *forward*, *backward* et combinée (bidirectional) *elimination* - ont été mises en place et consistent à tester dans un certain ordre les variables.

Une première approche est la méthode *forward* où à partir d'un modèle avec un *intercepte* toutes les variables sont testées une à une. La variable minimisant le plus l'AIC et significative pour un test du type 3 est conservé. On réitère ce processus par la suite autant de fois que nécessaire.

Une seconde approche est de faire du *backward* en partant d'un ensemble de variables indépendantes. Les variables les moins significatives sont retirées une à une jusqu'à toutes les variables soient significatives. En actuariat, le nombre de variables et le fait qu'il y a de nombreuses colinéarités rend impraticable cette méthode.

La méthode la plus utilisée est de procéder un *forward* puis un *backward*. Dans le cadre de variables plutôt indépendantes, cette méthode est plutôt efficace en temps et permet d'avoir des performances proches de la méthode exhaustive. Cependant, en pratique, il est nécessaire de procéder avec plus de précautions du fait de nombreuses corrélations et des transformations des variables (splines, regroupements, interactions...). Ainsi, une fois la méthode *stepwise* utilisée, l'actuaire doit simplifier son modèle en ajoutant ou enlevant des variables en fonction de leur significativité et de leur intérêt commercial.

En théorie, il est nécessaire de regarder les résidus pour valider le modèle. Il en existe différentes sortes : les valeurs des résidus de Pearson standardisés, non standardisés ou studentisés, les résidus de déviance et les résidus d'Anscombes. Théoriquement, Cameron et Trivedi (1998) proposent de représenter les résidus en fonction des variables prédites pour voir l'ajustement du modèle. McCullagh et Nelder (1989) [54] proposent d'utiliser des variables endogènes pour déterminer s'il y a besoin de les ajouter. À partir de ces graphiques, il est nécessaire de regarder si l'hétéroscédasticité des résidus [26] est vérifiée. De manière générale, ces analyses graphiques sont peu faites et remplacées par les graphiques mentionnés dans la partie suivante. La raison est simple ; si l'hétéroscédasticité n'est pas vérifiée par exemple, il serait nécessaire de mettre en œuvre des modèles plus complexes dont le gain final est assez négligeable face aux coûts de développement.

1.2.2.b Les analyses graphiques

Pour la sélection de variables, l'outil graphique principal d'un actuaire en tarification est le graphique Observé Vs Prédit - Figure 1.9.

Ce graphique peut être complété en partitionnant par an la sinistralité observée pour vérifier la cohérence temporelle. La validation d'une variable se fait aussi en ajoutant les prédictions d'un autre modèle utilisant les mêmes variables sauf la variable observée pour déterminer le gain en adéquation à la courbe des observés. D'un point de vue personnel, ce graphique analysé sur X_{test} est nécessaire pour valider tous les modèles de tarifications et les effets des variables capturées. À partir de ces graphiques, les segmentations des variables, les splines, le choix des bornes sont plus aisées. Ce type de graphique a plusieurs atouts. Il permet de :

- Déterminer l'ensemble de définition d'une variable tarifaire et sa distribution rapidement. Ainsi, on comprend la volatilité de certains segments s'il y a peu d'exposition et les caractéristiques principales (ex : portefeuille jeune ou âgé).
- Remarquer les écarts entre les prédictions et les observées et d'adapter les variables (pour borner, catégoriser).
- Comparer les effets marginaux des effets observés. En effet, par effet de corrélation, les effets peuvent être différents. De ce fait, justifier l'utilisation d'une variable. On peut aussi ajouter des intervalles de confiance sur les multiplieurs.
- Remplacer les graphiques sur les résidus proposés par McCullagh et Nelder ou Cameron et Trivedi.
- Vérifier en moyenne par segment que le modèle est bien calibré.

Le défaut de ce graphique est de mal capturer les effets de corrélations multidimensionnelles par exemple la corrélation spatiale. Pour cela, il faut analyser graphiquement les résidus sur une carte pour vérifier s'ils ont bien homogènes sur la France. Lors de l'ajout de données externes, une analyse graphique de la distribution des variables est plus que nécessaire. En effet, certaines caractéristiques des variables se trouvent dans une zone particulière et viennent brouiller l'analyse des effets marginaux (Voir le chapitre 2).

26. C'est-à-dire qu'il n'y a pas de variance différente en fonction des valeurs des variables

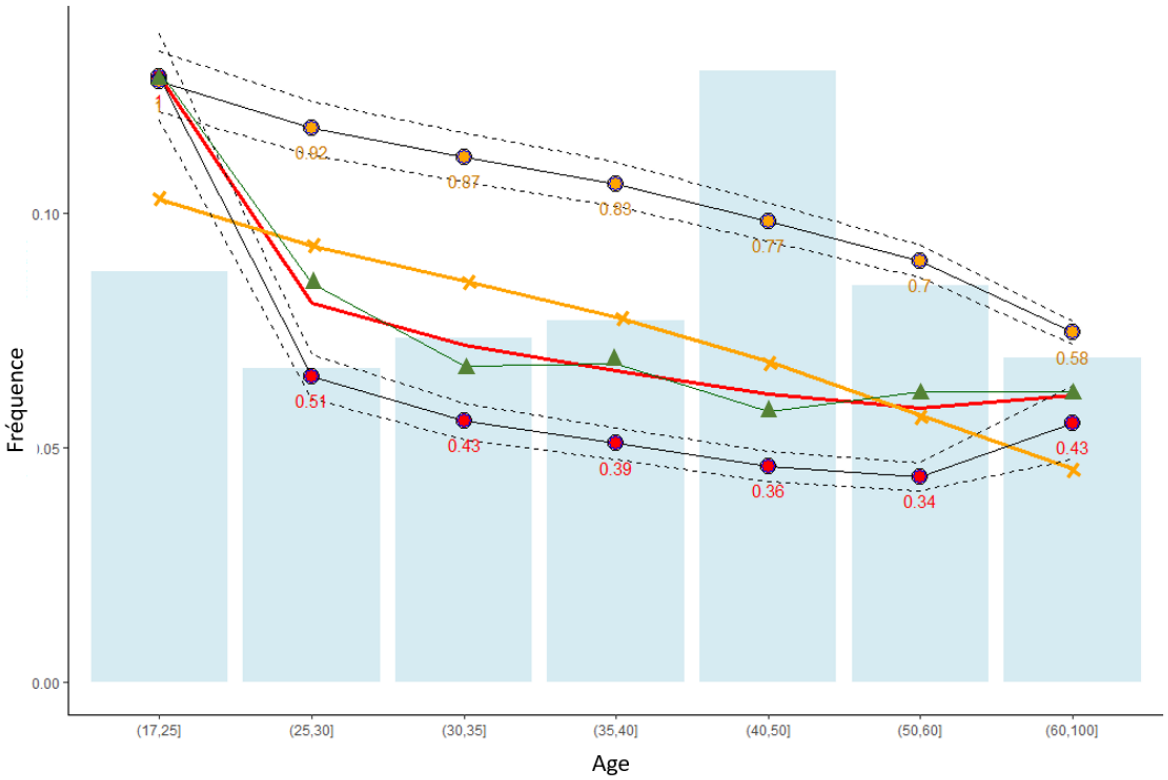


FIGURE 1.9 – Sur la base $\mathcal{X}_{train}$, ▲ correspond à la moyenne du nombre de sinistres observé, × à la moyenne des prédictions du modèle GLM sans prise en compte de spline, ● à la moyenne des prédictions du modèle GLM avec un spline de la base $\mathcal{X}_{train}$. ❍ correspond à l'effet marginal par rapport à une valeur donnée (ici à la modalité moyenne 17-25). La variable *Age* est discrétisée uniquement pour le graphique et ne l'est pas dans les GLMs.

1.2.3 Les modèles non linéaires

Depuis une vingtaine d'années, les méthodes de Machines Learning (ML) sont de plus en plus populaires, en particulier dans le secteur industriel. L'état de l'art de l'utilisation de méthodes ML pour combattre le changement climatique *Tackling climate change* Rolnick et al. (2019) [58] répertorie leur intérêt : la performance, le pouvoir d'agrégation, la rapidité de calcul, la versatilité ...

Dans le secteur de l'assurance non-vie, ces méthodes pourraient être en concurrence avec des méthodes statistiques comme les Modèles Linéaires Généralisés (GLM) introduits par Nelder et Wedderburn (1972) [56] ou modèles additifs généralisés introduits par Hastie en 1990 [47]. En 2022, la quasi-totalité des assureurs français utilise les GLMs pour la tarification. Les réseaux de neurones, les modèles de forêts aléatoires ou les modèles de boosting comme les XGBoost doivent surmonter un nombre important d'obstacles avant d'être utilisées. Par exemple, les concepts (détaillé par David 2015 [41]) (anti-sélection, biais moral, asymétrie d'informations ...) et les contraintes précédemment mentionnées sont mal anticipés avec l'utilisation de modèles ML. Au contraire, les méthodes statistiques par leurs aspects pratiques et théoriques sont plutôt bien comprises et étudiées (Voir Kaas et al. (2008) [52] ou Frees (2010) [44]).

Ainsi les méthodes statistiques ont deux atouts principaux qui expliquent la popularité en cours face aux modèles ML. Le premier atout est leur structure multiplicative simple qui permet une rapide interprétation, des contrôles et des modifications simples (comme mentionnée précédemment). Le second atout est lié à la nature aléatoire des contrats d'assurances. Contrairement à de l'analyse d'images, de biens endommagés, de la gestion d'e-mails, les performances des modèles en tarification sont limitées par l'incertitude inhérente du risque. Ainsi, les gains en performance des méthodes ML sont faibles et monétairement sont inférieurs aux coûts engendrés sur la maintenance, le contrôle des modèles par rapport aux GLMs.

En revanche, les modèles ML sont très largement utilisés pour sélectionner des variables en amont, déterminer des potentielles interactions et évaluer l'importance des variables. Habituellement, elles

sont utilisées pour valider des modèles statistiques, compléter les valeurs manquantes ou construire un zonier. De manière générale, elles ne sont pas utilisées aux endroits où il est nécessaire d'industrialiser une méthode ou de mettre en œuvre un protocole sans le contrôle humain.

Il existe plusieurs types de modèles MLs : CARTs, ANN, SVM, XGBoost ou RF. Seules les deux dernières ont été utilisées dans le cadre de cette thèse. Les deux méthodes reposent respectivement sur les principes de bootsting et de bagging et seront détaillées rapidement dans la partie suivante. Pour des détails théoriques, le lecteur peut se référer aux ouvrages mentionnés ci-dessous.

La base pour la modélisation $\mathcal{D}_{train}$ en amont est divisée $\mathcal{X}_{train}$, $\mathcal{X}_{validation}$ et $\mathcal{X}_{test}$.

1.2.3.a Les modèles de boosting : l'XGBoost

Le principe du boosting dans le cadre d'une régression est d'entrainer des modèles de faibles performances *"weak learner"* itérativement sur les erreurs de précédentes pour les sommer et obtenir un modèle robuste et sans structure particulière. Tout d'abord développé par Robert Schapire et Yoav Freund, c'est la méthode Adaboost [45] développé conjointement par les deux auteurs qui ont popularisé le concept.

L'extension XGBoost ([37]) est utilisée dans cette thèse pour la régression. Il a un avantage de combiner performance, gain en temps en limitant son utilisation de la mémoire. L'unique inconvénient rencontré est qu'il converge plutôt mal lorsqu'il y a des valeurs extrêmes ou aberrantes comme pour le coût moyen en sécheresse ou les premières bases de tarification à l'adresse de qualité assez faibles. Cette extension est utilisée dans le chapitre 7 pour la modélisation d'arrêté de catastrophes naturelles en sécheresse.

1.2.3.b Les modèles baggings : les forêts aléatoires

Le bagging est le nom court de *Bootstrap Aggregation*. Proposé par Breiman en 1996 [36], l'idée simple est de créer plusieurs bases de données b de taille n en prenant avec remise des lignes de la base de référence, puis d'entrainer un modèle sur chacune des bases, M_b. Le résultat pour un individu i avec les caractéristiques $\mathbf{X}_i$ est la moyenne des prédictions $\hat{y}_i^b$ ou la classe la plus représentée. Parce que les méthodes de Bagging sont généralement moins performantes que les méthodes de Boosting, Breiman en 2001 ([70]) a proposé à la suite un nouveau type de modèles : *RF* . S'appuyant sur la structure des arbres CARTs [38], un *RF* est une méthode de Bagging avec un côté aléatoire en plus : pour chaque nœud d'un modèle CART, $mtry \in \mathbb{N}$ - un hyper-paramètre - variables sont aléatoirement choisies.

Pour un usage actuariel, les modèles de Bagging ont l'avantage d'être performants et particulièrement robustes aux bruits, mais ont l'inconvénient d'être difficilement implémentables opérationnellement et coûteux en mémoire et en calcul.

Il existe une extension très intéressante actuariellement s'inspirant de la théorie de la crédibilité (voir [48]). Celle-ci est utilisée en particulier dans les chapitres 3 et 7 pour la modélisation de la fréquence des sinistres attritionnels et pour la fréquence conditionnelle pour le risque sécheresse.

1.2.3.c Les différences entre les sélections des variables

Les modèles de Machine Learning ont la particularité de ne pas informer sur la significativité ou non d'une variable sur l'output. Seules les métriques de performance et le graphique ci-dessus peuvent vérifier la bonne convergence d'un modèle. Cependant, différents outils graphiques et mesures permettent de mesurer une relative importance et effet des variables (Voir chapitre 3.). Comme pour la plupart des méthodes de sélection de variables, la corrélation inter-variables doit être regardée de près et peut biaiser l'importance accordée à une variable.

De manière générale, il est courant de faire une première sélection par ML des variables pour les tester dans un GLM. Il est cependant nécessaire de faire attention à certains points. Par nature, les GLMs ne sélectionnent que les variables avec des effets continus et non complexes alors que les modèles ML peuvent prendre en compte des interactions complexes et des effets discontinues. Une variable utilisable en machine learning ne l'est pas toujours pour les GLMs.

1.2.4 La sélection et correction de la base de modélisation

1.2.4.a Pourquoi ?

La définition de la base de modélisation est très importante et dépend de l'objectif. En souscription, l'objectif est que les résultats des modèles représentent le coût espéré d'un prospect. Pour cela, il faut que la base de modélisation ressemble le plus possible aux données sur lesquelles les modèles seront utilisés. Cependant, les assurés sont différents des prospects sur plusieurs tableaux. Les caractéristiques comme l'âge, les propriétaires locataires ainsi que les parcours de souscriptions influent sur le risque, les choix et même les réponses du questionnaire. Comme les grandeurs modélisées sont par essence volatile, la comparaison entre la base de modélisation et les données sur lesquelles les modèles utilisés, est difficile. Techniquement, il existe plusieurs méthodes pour contrôler l'adéquation, la cross-validation, la séparation entre train-test, le contrôle des valeurs influentes ou aberrantes... L'idée est qu'il est nécessaire de contrôler les effets appris et de limiter le sur-apprentissage.

1.2.4.b La sélection des individus

Lors du choix de la base de modélisation, des lignes sont écartées comme les biens avec des caractéristiques particuliers : château, manoir, maison de retraite... En effet, ce sont des biens qui passent par des canaux de souscriptions différents avec une tarification différente. De plus, certaines lignes dont les montants de sinistres sont trop extrêmes sont mis de côté pour éviter une mauvaise convergence de modèles. Par exemple : un sinistre provoquant un désamiantage doit être mis de côté ; les montants sont très élevés et le produit peut posséder maintenant une exclusion.

La question se pose aussi quand plusieurs produits co-existent. Les nouveaux produits avec peu d'individus peuvent se baser sur les données d'un ou plusieurs produits en Run-Off, soit en prenant directement le portefeuille complet pour choisir ou robustifier l'estimation des coefficients tarifiant de certains segments, soit en adaptant les coefficients des GLMs directement. Dans certains cas, il peut manquer des caractéristiques tarifantes. Habituellement, les contrôles de ces ensembles de filtres se font surtout en comparant les coefficients obtenus des modèles avec les différentes bases. Il est cependant important aussi de comparer les prédictions, car les structures de corrélations peuvent être différents d'un portefeuille à un autre ou d'une année à une autre.

Cette sélection doit toujours être en lien avec la finalité. La sélection d'individus jeunes pour la création d'un produit d'assurance centrée seulement sur les jeunes est évidente. Mais la sélection de maisons uniquement bien géolocalisées, ou la non-prise en compte d'individus spécifiques ou qui ne feront plus parties de prospects du produit peut être discuté. Les processus de souscriptions et de tarifications sont évidemment interdépendants et nécessite de bien comprendre la finalité avant de modéliser.

1.2.4.c Quels historiques prendre pour les modélisations ?

L'historique utilisé est important et doit refléter le risque à venir. Pour les modélisations des garanties non climatiques usuelles, 3 à 5 ans d'historique sont suffisants. En effet, la fréquence de ces garanties évolue entre 0.01 à 0.3 points ces dernières années [27]. Cependant, pour des portefeuilles ayant des structures de risques changeants ou un renouvellement important, cet intervalle historique peut être réduit. Dans l'autre sens, l'historique utilisable est borné par les évolutions importantes du produit : nouvelles gammes, garanties, nouvelles questions... Le point limitant est la non-stationnarité dans le temps.

L'estimation de la sinistralité est en réalité biaisée par les différentes années. En effet, une tendance baissière de la fréquence provoque évidemment une surestimation de la fréquence à venir. Dans ce cadre, la prise en compte des années dans la modélisation est faite pour projeter la tendance de chacune des garanties. Similairement, certaines années doivent être traitées différemment. Par exemple, le cas le plus récent est les impacts du confinement ou de la Covid qui ont mécaniquement diminué la fréquence de la RC Scol ou VOL par exemple. Idem, pour la garantie DDE, certaines années très pluvieuses entrainent une augmentation de la fréquence. Il est donc nécessaire d'adapter les poids des années où les expositions directement de ces années pour représenter au mieux la sinistralité à venir.

27. Chiffre FFA 2021

Au contraire, pour les risques rares ou climatiques, la volatilité des évènements et des années obligent à utiliser un historique le plus long possible. Un long historique stabilise l'estimation moyenne de la sinistralité et permet d'avoir un panel de scénarios possibles représentatifs. Il est à noter que la non-stationnarité dans le temps est toujours problématique, mais il est très difficile de la prendre en compte lorsque les années ne sont pas identiquement distribuées.

1.2.4.d Les Primes as if ou Primes à euro constant

Les derniers paramètres importants à prendre en compte sont l'inflation et les tendances. L'inflation se traite souvent dès le départ en dé-inflatant les montants des sinistres par des indicateurs comme des indices FFB. En fonction de la nature du risque, les indices peuvent être adaptés. Sur le coût moyen des sinistres attritionnels, il est aussi possible d'utiliser l'année d'exposition en tant que variable dans les modèles. Même si elle peut être corrélée avec les autres caractéristiques du portefeuille, cela permet d'avoir une estimation *data based* de l'inflation. Dans les deux cas, il va être nécessaire d'extrapoler la tendance de l'inflation capturée pour l'année à venir. Pour les sinistres graves, c'est souvent la tendance attritionnelle qui est utilisée. En effet, le développement long des sinistres graves biaise les estimations. Sur la prime, l'indice des prix de l'assurance habitation de l'Insee fait référence et reflète l'évolution des primes du marché. Il est nécessaire de comprendre que les évolutions des primes des A.N. et les revalorisations ne vont pas toujours dans le même sens.

Pour l'évolution de la fréquence, l'évaluation des tendances se fait soit en amont en corrigeant directement les expositions (ex : Impact des confinements) ou en ajoutant l'année d'exposition en tant que variable ou offset. Dans les deux cas, une extrapolation va être mise en place. La difficulté supplémentaire doit être l'appréhension des sinistres provoqués par des événements climatiques (DDE, ELEC TGN majoritairement) dont les périodes de retours doivent être bien comprises pour catégoriser les années.

Le recalcul des coefficients se fait généralement annuellement ou semestriellement. Cette nouvelle tarification en l'année t s'appelle une version t d'un tarif. Les A.N sont regroupés par année de souscription appelée "génération". Ainsi pour une même génération, la version d'un tarif est la même que cela soit pour tarifier la prime proposée en A.N. ou les différentes revalorisations. Dans cette thèse, nous nous limitons temporellement à l'année d'étude et calculons qu'une seule version du tarif.

Conclusion

Le processus de tarification est très étroitement lié à la souscription, au processus de revalorisation ainsi qu'à toute la chaine de valeurs assurantielle. J'ai voulu dans cette partie être le plus généraliste possible et d'expliquer les tenants et aboutissants techniques et méthodologiques des choix des actuaires. Si la réglementation des produits d'assurances impacte les choix et les méthodes des actuaires, ces derniers ont aussi façonné et guidé la réglementation assurantielle sur des principes de proportionnalités, de mutualisations ou d'indemnisations. En conséquence, les évolutions des produits d'assurances dépendent de la réglementation en place.

Dans le cadre de la réflexion sur la modification des processus de souscription, la responsabilité et la provenance des données tarifaires, les méthodes de tarification ne peuvent être foncièrement modifiées. Dans un premier temps, les méthodes traditionnelles GLMs et calculettes tarifaires se doivent être similaires pour permettre de comparer les produits.

Références

[34] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6) :716–723, 1974.

[35] FJ Anscombe. Contribution to the discussion of h. hotelling's paper. *JR Stat. Soc B*, 15 :229–230, 1953.

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

[37] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[38] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93(443) :935–948, 1998.

[39] European Commission. Reform of eu data protection rules, 2018.

[40] Andrea Dal Pozzolo. Comparison of data mining techniques for insurance claim prediction. ..., 2011.

[41] Mihaela David. A review of theoretical concepts and empirical literature of non-life insurance pricing. *Procedia Economics and Finance*, 20 :157–162, 2015.

[42] Łukasz Delong, Mathias Lindholm, and Mario V Wüthrich. Making tweedie's compound poisson model more accessible. *European Actuarial Journal*, 11(1) :185–226, 2021.

[43] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*. 1925.

[44] Edward W Frees. *Regression modeling with actuarial and financial applications*. Cambridge University Press, 2010.

[45] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1) :119–139, 1997.

[46] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The annals of applied statistics*, pages 916–954, 2008.

[47] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

[48] Roel Henckaerts, Katrien Antonio, Maxime Clijsters, and Roel Verbelen. A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, 2018(8) :681–705, 2018.

[49] Roel Henckaerts, Marie-Pier Côté, Katrien Antonio, and Roel Verbelen. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, pages 1–31, 2020.

[50] Bo Hu, Jun Shao, and Mari Palta. Pseudo-r 2 in logistic regression model. *Statistica Sinica*, pages 847–860, 2006.

[51] Bent Jørgensen and Marta C Paes De Souza. Fitting tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1) :69–93, 1994.

[52] Rob Kaas, Marc Goovaerts, Jan Dhaene, and Michel Denuit. *Modern actuarial risk theory : using R*, volume 128. Springer Science & Business Media, 2008.

[53] Yue Liu, Bing-Jie Wang, and Shao-Gao Lv. Using multi-class adaboost tree for prediction frequency of auto insurance. *Journal of Applied Finance and Banking*, 4(5) :45, 2014.

[54] Peter McCullagh and John A Nelder. *Generalized linear models*. Routledge, 2019.

[55] Daniel McFadden et al. Conditional logit analysis of qualitative choice behavior. 1973.

[56] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society : Series A (General)*, 135(3) :370–384, 1972.

[57] Jessica Pesantez-Narvaez, Montserrat Guillen, and Manuela Alcañiz. Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks*, 7(2) :70, 2019.

[58] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv :1906.05433*, 2019.

[59] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[60] Gordon K Smyth and Bent Jørgensen. Fitting tweedie's compound poisson model to insurance claims data : dispersion modelling. *ASTIN Bulletin : The Journal of the IAA*, 32(1) :143–157, 2002.

[61] Maurice CK Tweedie et al. An index which distinguishes between some important exponential families. In *Statistics : Applications and new directions : Proc. Indian statistical institute golden Jubilee International conference*, volume 579, pages 579–604, 1984.

[62] Mario V Wuthrich and Christoph Buser. Data analytics for non-life insurance pricing. *Swiss Finance Institute Research Paper*, (16-68), 2019.

[63] Yi Yang, Wei Qian, and Hui Zou. Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. *Journal of Business & Economic Statistics*, 36(3) :456–470, 2018.

Chapitre 2

Géolocalisation et les données à l'adresse

2.0.1 Les autorisations légales et attribution des travaux.

Pour ces travaux de thèse, l'ensemble des données utilisées émanent de deux assureurs français. Les autorisations constituent l'objet de contrats préalablement établis pour une diffusion anonyme dans le cadre de cette thèse. Des informations plus générales résultent du partenariat Addactis France® et *nam.R* lors de la création du projet **SHoP** et de sa commercialisation.

L'ensemble des résultats et graphiques proviennent directement de mes travaux et de mes codes. Dans le cas contraire, cela sera mentionné. Les exceptions sont les suivantes : les travaux de géolocalisation et de créations de données ont été exclusivement produits par l'entreprise *nam.R*. J'ai contribué dans un premier temps puis avec Bilal et Mederick au choix des éléments à améliorer par des retours d'expériences. Ayant travaillé dans un premier temps seul sur la partie technique de la modélisation de **SHoP** pendant un an et demi, l'ensemble des codes et des méthodologies ont été développés à cette occasion. Toutefois, les équipes d'Addactis P&A ont très largement repris, vérifié, adapté la méthodologie et les processus pour l'industrialiser sur d'autres portefeuilles d'assureurs. Cela a permis de valider la stabilité et la robustesse des résultats et des méthodologies mises en place. Les validations des méthodes développées dans cette thèse, la vérification de la stabilité des résultats de cette thèse ont été fait sur quatre autres portefeuilles d'assureurs par l'ensemble des équipes, par binôme ou plus : Silvia avec Quang, Victoria avec Bilal, Marie avec Cedric et Franck, Linda avec Bruno, Yasser et Arshak.

Je ne me permets pas de m'attribuer l'ensemble de ces travaux de R&D, car chaque suivi et échanges avec les équipes d'Addactis ® ou des assureurs ont permis d'amender mes erreurs tout en adaptant ces travaux pour de nouvelles applications opérationnelles.

Les parties suivantes examinent l'ensemble des données provenant d'un partenariat avec une entreprise fournisseuse de données *nam.R*. Les données sont étudiées sur des expositions aux risques d'assureurs avec leur consentement. Dans certains cas, les cartographies sont génériques provenant des données sources directement. L'ensemble de ces éléments sont des données dont les règles de copyright ont été respectées. Dans cette thèse, les données respectent le **RGPD** et le contrat par lesquelles elles sont liées. L'ensemble des travaux répond à l'engagement de l'assureur pour mieux évaluer et faire vivre les contrats d'assurances. Les valeurs manquantes se réfèrent soit à une absence de données sources, soit à une absence d'exposition de la part de l'assureur.

2.1 La géolocalisation à partir d'une adresse

Obtenir les coordonnées géographiques des biens assurés à partir d'un contrat d'assurance en **MRH** est un processus complexe. À la souscription, rares sont les assureurs demandant la géoposition du bien assuré. Parmi les rares assureurs, LUKO®, un acteur récent fondé en 2016 utilise le cadastre pour sa souscription ; les prospects choisissent leur parcelle en défilant une carte. De plus, la grande majorité des assureurs ne géolocalise pas leurs contrats. Il existe cependant des initiatives de préventions ou de marketing comme ceux de Générali® géolocalisant leurs bâtiments pour déterminer des scores de risques climatiques et technologiques. Les limites rencontrées sont souvent les mêmes. Il est à noter que pour l'assurance MRP, des contrôles pour l'exposition aux risques climatiques et technologiques sont

souvent faits sur place. Les limites rencontrées sont souvent les mêmes.

Dans les bases de modélisations, la géolocalisation d'un contrat ou des biens assurés est indisponible à la maille du bâtiment. Seule l'adresse permet la géolocalisation du bien. Celle-ci avait pour premier objectif d'envoyer des informations (échéances, modification des CGs ...) par voie postale. Pour la tarification, la précision de la maille communale associée à l'adresse est pour l'instant suffisante pour la création de zonier - voir le chapitre 1. Néanmoins, certains acteurs, surtout en réassurance comme la CCR, font le travail de géolocaliser à partir de l'adresse pour affiner les modélisations climatiques à l'échelle de l'adresse ou de la rue. Les différentes approches utilisent les mêmes bases détaillées dans la partie 2.1.1. Néanmoins, des complications apparaissent lors du processus - partie 2.1.2 expliquant pourquoi de nombreux acteurs arrêtent la géolocalisation à l'adresse ou même à la rue. Ici, le fournisseur de données géolocalise le bâtiment. Cependant, en affinant la précision, de nouvelles limites apparaissent et sont détaillées dans la partie 2.1.3.

2.1.1 Les référentiels nécessaires et le processus de géolocalisation

Pour géolocaliser des bâtiments à partir d'une adresse, trois référentiels sont nécessaires :

- **Le Référentiel de bâtiments géolocalisés et délimités** : La **BD TOPO** est la base open data de référence pour les bâtiments en France. Celle-ci peut-être complétée par d'autres bases comme **BDNB** (Base de Données Nationale des Bâtiments) ou bien par des analyses en computer vision pour détecter les bâtiments (surtout les annexes) non présents dans le référentiel.
- **Le Référentiel d'adresses reliées à un ou plusieurs bâtiments** : La base de référence est la **BAN** (Base Adresse Nationale). Celle-ci est améliorée et corrigée par des règles simples comme l'emplacement des numéros pairs ou impairs. Elle est aussi complétée par des informations externes.
- **Référentiel des parcelles/cadastres** : La base cadastre **Etalab** est utilisée majoritairement. Elle regroupe deux bases connues, celle de la France hors Alsace-Lorraine et celle de l'Alsace-Lorraine.

À l'intérieur de chacun des référentiels, un détail plus fin est fait. Par exemple, chaque bâtiment est associé à une topologie d'utilisation ; ex : résidentielle, individuelle, collectif, école ... De plus, un bâtiment peut contenir plusieurs parties résidentielles ou plusieurs types de bâtiments différents, par exemple : grange agricole et logements. Dans les récentes évolutions du géocoding, une sous-maille pour les parties résidentielles a été créée. Pour la base adresse, de façon similaire, une adresse peut être liée uniquement à une sous-partie résidentielle d'un bâtiment ou à un bâtiment. De plus, une adresse est reliée de façon plus macroscopique à un iris, à une commune, à un département...

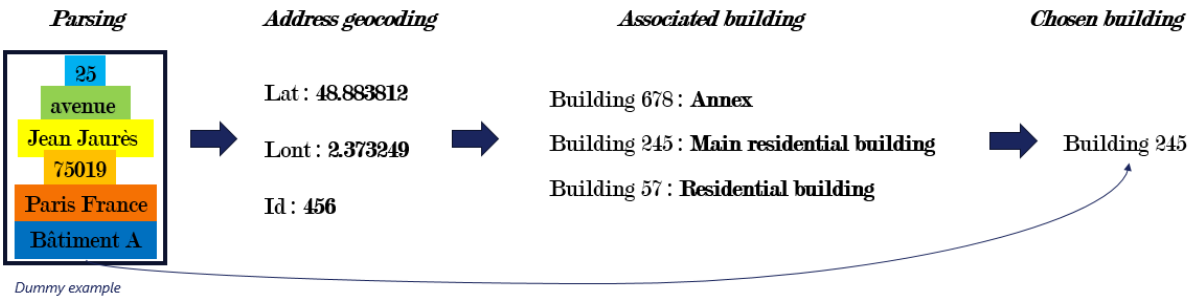


FIGURE 2.1 – Exemple simple pour comprendre le processus de géolocalisation d'une adresse.

L'information de base sur la géolocalisation est l'adresse. Une adresse française se subdivise en plusieurs parties qui devraient idéalement suivre les règles suivantes :

- **Un numéro de rue :** Un nombre entier entre 0 et 9999. Il associe chaque bâtiment à un nombre unique pour chaque type de rue ou lotissement. Il peut être suivi de "Bis", "Ter" ...
- **Type de route :** La dénomination recommandée doit être adaptée à l'environnement de la rue et doit être standardisée ; *Fr : Allée, avenue, boulevard, chemin, cours, impasse, passage, périphérique, place, quai, route, rue, ruelle, square.* Elle peut être en langue régionale ;
- **Nom de la route :** Il y a peu de règles sur la dénomination pour les routes. Ex : Celui-ci peut être en langue étrangère avec des chiffres ou être une date ;
- **Code postal :** C'est une séquence de cinq chiffres créer pour faciliter la délivrance postale qui permet en premier lieu à la poste d'identifier la commune pour la livraison de la lettre - ex : 75019 ;

- **Nom de la commune :** ex : Paris ;
- **Complément d'adresse :**(*optionnel*) Un complément (ou supplément) d'adresse permet aux postiers de mieux délivrer des lettres pour les immeubles ou les lotissements avec une structure un peu complexe. Cela peut concerner l'étage, le numéro de l'escalier ou la personne - ex : "to M. Smith". Cette information est souvent peu utile pour la géolocalisation.
- **Pays :**(*optionnel*) - ex : France.

Le processus de géolocalisation est plutôt simple et est décrit dans la figure 2.1. Tous les bâtiments sont détectés en amont de chaque géolocalisation. L'adresse envoyée est reliée à une adresse de la base référentielle déjà géolocalisée. Cette jointure se fait à l'aide d'un score d'appariement se basant sur la commune et les caractères des adresses disponibles. Ensuite, à partir de cette géolocalisation, plusieurs bâtiments potentiels sont détectés à partir du référentiel bâtiment. Ici, le bâtiment choisi est celui qui est "principal". Dans le cas où il y a plusieurs bâtiments principaux, l'un d'entre eux est choisi en fonction des autres adresses, de la distance, de la taille...

Un processus similaire avait été aussi testé à l'aide des parcelles en maillon intermédiaire. Toutefois, des bâtiments peuvent être positionnés sur plusieurs parcelles et inversement. Au début du projet, les deux méthodes coexistaient. Cependant, la géolocalisation par la parcelle fut abandonnée pour des raisons de coûts, mais aussi parce qu'il était difficile de choisir quelles méthodes prendre quand les bâtiments choisis étaient différents. Néanmoins, ce référentiel des cadastres est nécessaire. En effet, il permet de relier les annexes à chaque bâtiment, relier les piscines, déterminer les zones communes de copropriété suite à la géolocalisation des bâtiments...

Le résultat du processus est retourné sous forme d'une géométrie d'un bâtiment ou une partie d'un bâtiment et sa géolocalisation (longitude et latitude). Cette géométrie est utile du point de vue création de nouvelles variables, du processus de souscriptions ou pour des aspects graphiques. De manière générale, un point longitude et latitude est fourni et il a été convenu qu'il correspond aux centroïdes du bâtiment dans la majorité des cas[1]. La nature du bâtiment est aussi fournie, mais reste à une topologie assez large. En effet, si les bâtiments associés à des cimetières, des églises ou des hypermarchés ne changent pas ou peu, la présence de magasins ou leurs types évoluent trop rapidement dans le temps.

Un certain nombre d'indicateurs sont fournis pour évaluer la qualité de chaque étape du processus du géocodage. Par exemple, l'adresse appariée à l'adresse envoyée, un score de confiance entre les différents liens, le nombre de bâtiments reliés à l'adresse, le nombre d'adresses reliées au bâtiment choisi, la distance entre le point adresse et le centroïde du bâtiment. Ces indicateurs permettent d'évaluer la qualité individuelle du géocoding - voir la partie 3.2.3.

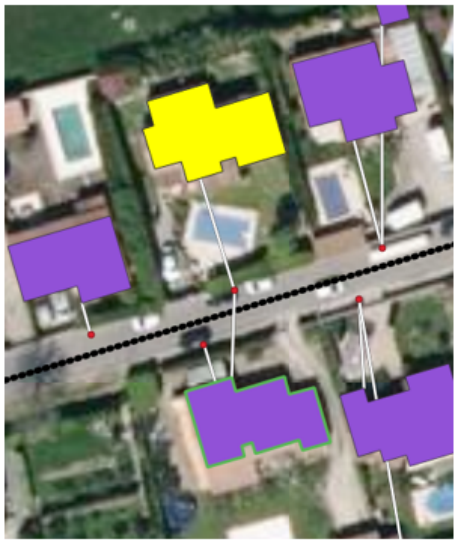


FIGURE 2.2 – Exemple de géolocalisation. Les points de géolocalisation de la BAN ne sont pas parfaits et suffisants. D'autre part, la notion de parcelle est indispensable pour le rattachement des piscines.

2.1.2 Les complications récurrentes

La difficulté réside dans l'hétérogénéité des situations et des cas pratiques. En effet, certaines adresses ont encore des stigmates de l'histoire. Si la norme *AFNOR NF Z 10-011* a été mise en place en 2013 pour standardiser les adresses avec les contraintes susmentionnées comme un nom de route de moins de 38

1. Il subsiste des formes de bâtiments problématiques dans certaines situations, mais cela reste très à la marge pour notre niveau de précision.

caractères, cette norme est obligatoire uniquement à Paris. De plus, seulement les communes de plus de 2000 habitants sont, depuis 1994, obligées de recenser leurs rues et les numéros de rues. Comme il n'y a pas (ou peu) de contraintes réglementaires et que les petites communes sont nombreuses, un nombre non négligeable d'adresses et de rues ne sont pas à jour ou inconnues.

Au fur et à mesure des années, des changements de noms ou de types de voies comme une impasse devenant une rue apparaissent. Plus à la marge, il existe aussi des rues ayant les mêmes noms dans une même commune. Le graphique 2.3 fourni par le data provideur en juin 2020 met bien avant les difficultés des liens adresses-bâtiments en zone rurale.

Ainsi, toutes les adresses d'une rue ne peuvent pas toujours être géolocalisées comme le montre un exemple dans la section 7.2 sur la subsidence. Dans les zones rurales en particulier, le contrat ou les adresses sont liées à plusieurs bâtiments. Cela diminue la probabilité de sélectionner le bon bâtiment. En moyenne, plus de 20 % des bâtiments sont choisis avec ambiguïtés dont une parte du fait du nombre de bâtiments sur une parcelle - graphique 2.4. [2].

L'exemple 2.1 de la partie 2.1.1 témoigne que même si l'adresse choisie est la bonne et bien géocodée, il y a encore de l'incertitude sur le bâtiment à choisir. Les contrats ne concernant que des annexes, ne sont pas à géolocaliser.

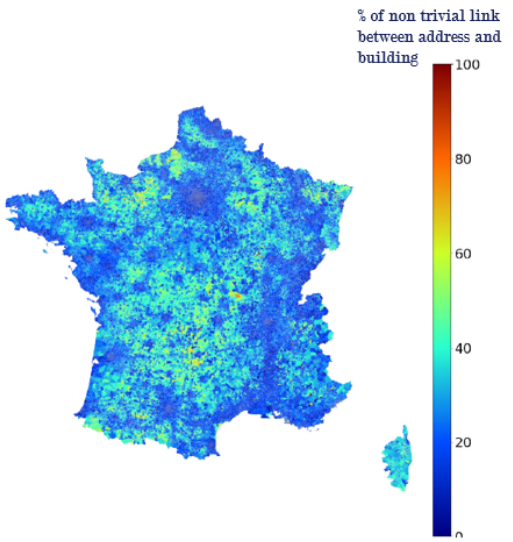


FIGURE 2.3 – Cartographie du pourcentage des liens non triviaux entre les adresses et les bâtiments (Carte *nam.R*).

Du reste, les compagnies d'assurances n'ont pas toujours mis en place des contrôles sur la qualité des adresses. Les travaux sur quatre portefeuilles d'assureurs ont montré qu'environ 20% ne sont pas géolocalisables (surtout avec des numéros de rues absents). Des problèmes systématiques sont inhérents aux lieudits français - environ 5% des adresses.

Des erreurs topographiques, des abréviations non génériques, des routes non existantes ou des adresses partielles à cause des changements de **SI** [3] sont les causes les plus récurrentes. L'adresse de l'assuré fait toujours priorité sur l'adresse du référentiel lors des A.N., cependant la question du stock se pose. Lors du changement d'un nom d'impasse à route, la géolocalisation du bâtiment ne change pas, mais son adresse oui. Dans certaines communes, des adresses respectant la norme AFNOR ont été attribuées pour des besoins administratifs comme la fibre cependant les habitants utilisent toujours l'ancienne dénomination, n'ayant pas connaissance ou oublié la nouvelle adresse.

2. Cela ne veut pas dire que les erreurs sont de l'ordre de 20%. Il y a un travail pour choisir le bâtiment principal qui permet de contrôler l'ambiguïté.
3. Système d'information.

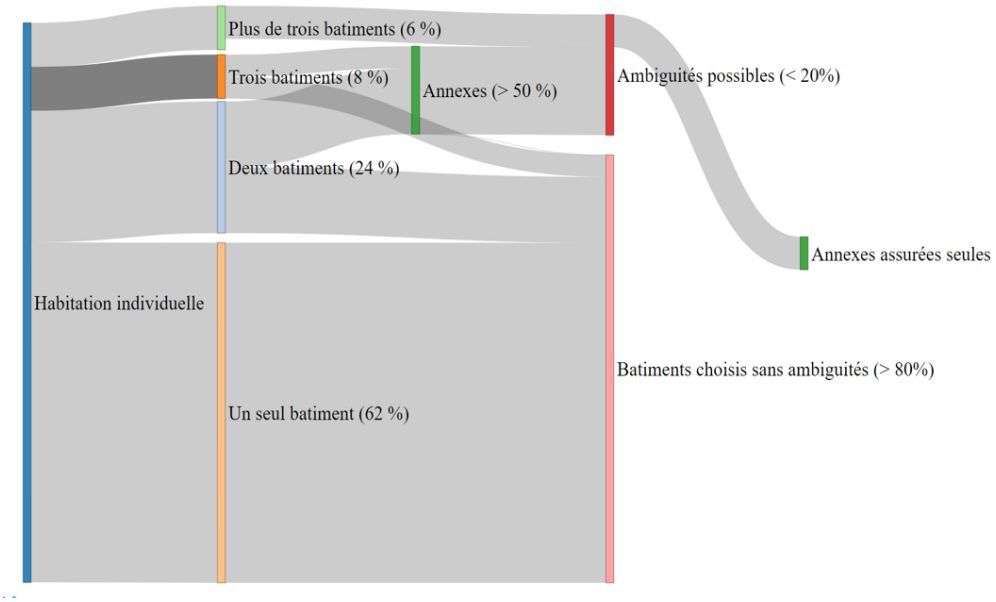


FIGURE 2.4 – Nombre de bâtiments par parcelle sur le parc résidentiel français.

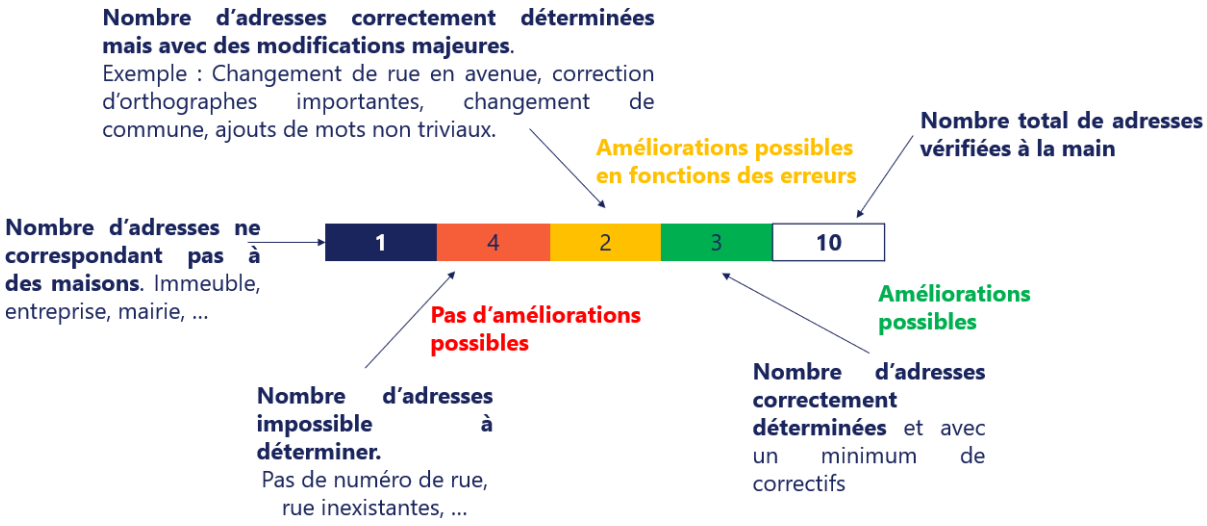


FIGURE 2.5 – Processus manuel mis en place pour le contrôle des adresses non géocodées.

Résultat du géocoding - juin 2020 : Pour évaluer ces erreurs, les adresses non géolocalisées ont été étudiées. La méthode d'évaluation est manuelle : en utilisant Google Maps, Street View et les images aériennes, j'ai vérifié si l'adresse est géolocalisable. Chaque résultat des adresses est catégorisé décrit dans la figure 2.5. La figure 2.6 résume les résultats.

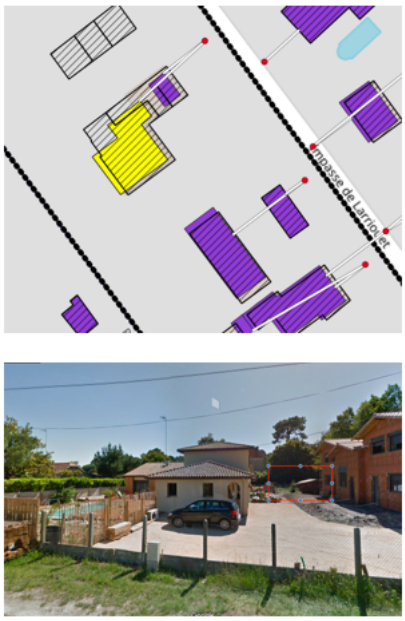


FIGURE 2.7 – Un exemple d'annexe qui est en réalité la maison principale d'un locataire. De nombreux bâtiments sans délimitation claire l'entourent avec un bâtiment en construction à côté.

Environ 75% des adresses ne sont pas géocodables manuellement sans bien connaitre les environs. Même si l'adresse est bien géocodée, trop de cas spécifiques existent et limitent opérationnellement les améliorations du process. Néanmoins, la méthode de géolocalisation des adresses est perfectible ; environ 7% des adresses pourraient être géocodées facilement.
Ces résultats ne sont pas nouveaux et sont cohérents avec le taux de géocodage de la CCR : 20 % des adresses ne sont même pas géocodées à la rue. La CCR [a] géolocalise :

- 52% des habitations à l'adresse ;
- 76% des habitations à la rue ;
- 99% des habitations à la commune.

a. en 2020

Ces chiffres se retrouvent à travers d'autres sociétés de livraisons. Les sociétés des distributions évaluent entre 5% à 10% le taux d'erreur lié à l'adresse. La Poste l'évalue proche de 3% du fait de la connaissance du terrain de leurs facteurs. Dans la documentation récente de la base IGN version 3.0, il est estimé que 6% des adresses dont les noms de voies étaient erronées et que l'exhaustivité des bâtiments de plus de 20 m^2 est de l'ordre de 98%. Si la notion dynamique et temporelle des adresses et des bâtiments y est ajoutée, le taux de non-géocodage et d'erreurs s'explique aisément. Les différents travaux d'imageries aériennes permettent de mettre à jour ce référentiel régulièrement (tous les trois ans en théorie). En pratique, cela correspond à 400 000 nouveaux bâtiments construits annuellement qui ne peuvent être directement référencés. Il est à noter qu'en France les nouveaux bâtiments sont déjà en partie couverts par la garantie décennale des constructeurs et des normes de constructions et ainsi le risque reste ainsi limité au début pour les garanties DDE, INC, BDG ou sécheresse. Il est à espérer que le logement soit répertorié dans les années qui suivent sa construction, pour mettre à jour les données manquantes au bâtiment .

En bref, la géolocalisation à partir de l'adresse ne peut être parfaite et des erreurs vont apparaître. Les seules personnes capables d'identifier le ou les bâtiments assurés sans erreurs sont les assurés eux-mêmes. Pour cela, des visions aériennes et des façades doivent être disponibles pour tous les bâtiments. Étant donné que la souscription est sous la contrainte du temps (contrainte 3), l'idée est de présélectionner le bâtiment en amont pour accélérer le processus. Cependant, la notion de bâtiments demeure aussi imprécise. En effet, les bâtiments évoluent aussi dans le temps.

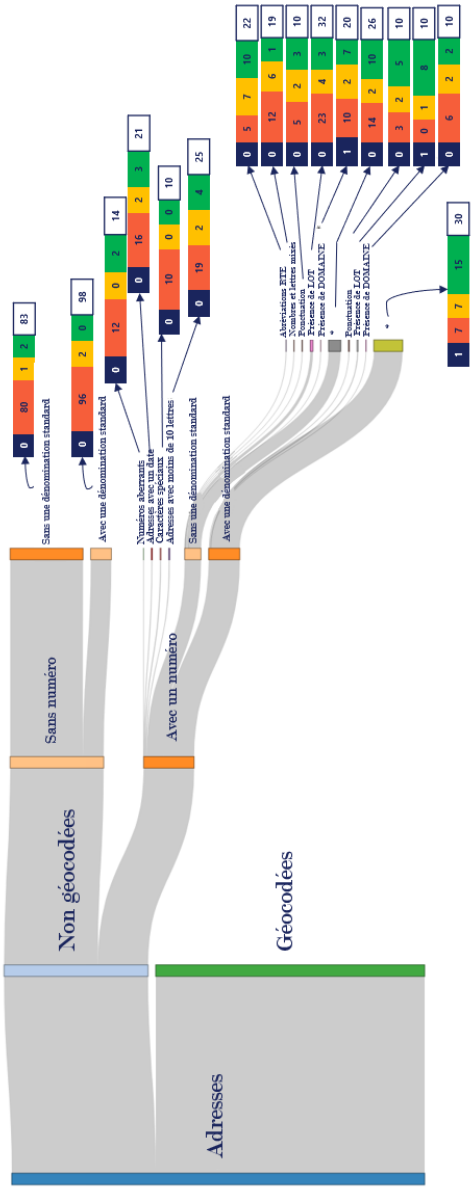


FIGURE 2.6 – Résultats sur un exemple stratifié sur les adresses non géocodées utilisant la méthode expliquée figure 2.5. Environ 76.1%, 6.5%, 15.5% et inférieur à 1% de contrats appartiennent respectivement aux catégories rouge, orange, verte et noire. Voir le tableau 2.1 en annexe.

2.1.3 Les limites de la géolocalisation

Des limites évidentes du géocoding portent sur la nature du bien, sur l'exhaustivité des bases, mais des contraintes apparaissent aussi de l'observation à un instant t d'un référentiel évoluant dans le temps.

FIGURE 2.8 – Maison avec deux boites aux lettres. D'après les données assureurs, il y a un studio au-dessus du garage.

Une question qui s'est rapidement posée concerne la définition d'un appartement ou d'une maison. Pour le data provideur, il existe une différence entre logement individuel et collectif. Si plusieurs boites aux lettres sont reliées à un bâtiment, alors c'est un bâtiment collectif. Cependant, il existe des maisons qui ont été séparées en deux avec une partie studio. La distinction est d'autant plus difficile en zone urbaine. Est-ce qu'une boulangerie avec un seul logement au premier étage et un jardin doit être considéré comme un appartement ou une maison ? Pour une résidence partagée, la question se pose aussi. Suite à des travaux, une maison individuelle peut devenir un habitat collectif. Ainsi, lors du processus de géolocalisation, il n'est pas toujours aisé de déterminer la partie résidentielle et de la caractériser.
Cette problématique amène la question du découpage d'un bâtiment par un agent autre que le fournisseur de données. Qui a raison et existe-t-il des règles simples pour définir la nature d'un bien ? L'ambivalence des concepts en cours amène dans ce cadre à se questionner sur des objets élémentaires d'un contrat d'assurances. Par les données, des différences apparaissent entre les maisons de villes, de faubourg, de banlieue ou de campagne, mais aussi entre les appartements ruraux et les appartements citadins. Ces notions étaient souvent implicitement prises en compte à la commune dans les zoniers.

Cette limite questionne sur l'individualisation de la prime. Est-ce qu'il y a des maisons ou des appartements qui changent complètement d'estimations de leurs risques lorsque leurs caractéristiques sont regardées à la loupe ? À partir des observations que j'ai pu mener, il est constaté que les habitations avec une nature équivoque (entre maisons et appartement) permettent de lisser ces concepts d'assurances individuelles et collectives dans le bon sens. Les primes avec les données à l'adresse ne sont pas orthogonales par rapport aux primes traditionnelles au sujet de la nature du bien.

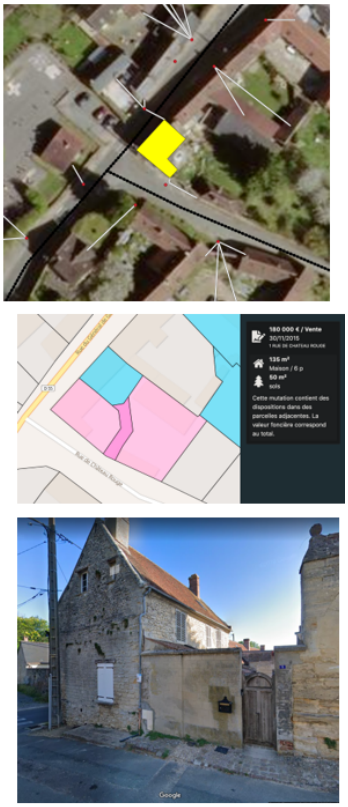


FIGURE 2.9 – À l'aide de la base **DVF**, il est visible que le bâtiment a été subdivisé. De plus, le dernier étage, l'attique n'est pas aménagé.

De plus, les bâtiments et leurs découpages évoluent dans le temps. De manière générale, les bâtiments sont observés avec 2 à 3 ans de retard, le temps que les bases se mettent à jour et que l'information soit récupérée. Les nouveaux bâtiments venant ou étant en cours de construction ne sont pas géolocalisables ou difficilement à travers l'exemple 2.7 (géocodeur juin 2020). Les bâtiments peuvent être également agrandis avec extension et/ou annexes. Ces modifications ne sont pas instantanément prises en compte, pas toujours visibles ni mêmes déclarées. Des questions apparaissent sur les informations au bâti, par exemple la période de construction doit-elle être différentiée entre le bâtiment principal et l'agrandissement ? Plus problématique, certains bâtiments se subdivisent en deux ou sont regroupés ensemble suite à des successions ou des ventes. Dans ces cas-là, il est plus difficile de mettre à jour le référentiel bâtiment. Il est évident que l'amélioration de la précision peut amener à des études de cas précis. Cependant, en pratique, la prise en compte individuelle de ces cas n'est ni utile ni impactant lors de la tarification.

Finalement, les données ne sont jamais exhaustives et propres. Il y a de nombreux bâtiments qui ne sont pas répertoriés (hangars, serres ...). De plus, des bâtiments individuels sont de temps en temps fusionnés entre eux ou mal découpés. Quasiment invisible de façon aérienne, une solution serait de pouvoir analyser les façades, ce qui n'est opérationnellement pas possible pour l'instant (disponibilité, qualité et poids des images). Dans ces cas précis, il est nécessaire lors de la souscription de permettre aux agents de modifier et d'ajouter de l'information. Ces erreurs sont souvent liées aux zones rurales et sont déjà partiellement prises en compte à l'aide d'autres caractéristiques du contrat.

La responsabilité des données : Dès lors la question de la responsabilité et de la capacité à modifier l'information repose sur l'assureur. Dans notre cas, les données sont mises à jour au moins bi-annuellement. Cependant, les bases seront toujours actualisées avec un temps de retard sur la réalité. L'exemple des bâtiments en construction peut poser un problème.

Ainsi, le mauvais découpage d'un bâtiment est maintenant de la responsabilité de l'assureur. Comme l'assureur propose le découpage des bâtiments, l'assuré n'a pas automatiquement l'obligation de déclarer les évolutions de son bâtiment. Le premier impact est de produire de l'anti-sélection, les assurés profitant des erreurs et ne souscrivant pas quand l'erreur leur est défavorable. De plus, l'assuré pourrait demander une modification des données. L'autre problématique est lors de la mise à jour des informations dévoilant une construction d'annexes ou une extension du bâtiment. Comment faut-il considérer les évolutions sur le bâti ? Est-ce que l'assureur peut mettre à jour une prime ou le bien protégé sans l'aval de l'assuré ? Ces évolutions doivent être renseignées dans les bases. Cependant, les corrections des bases ne doivent pas altérer l'information du découpage initiale des bâtiments.

Au bout du compte, l'amélioration de la précision du bien impose un changement de paradigme où l'assureur en améliorant sa connaissance du bâtiment prend à sa charge les erreurs et les évolutions de la donnée. Comme la relation B&C n'est active principalement que lors de la souscription et des sinistres, tous les échanges et la communication reposeraient sur l'initiative des assureurs. En bref, ces difficultés ne sont pas fondamentalement gênantes, mais traduisent un niveau de complexité supplémentaire. Cette complexité impacte fortement l'IT d'un assureur et à un coût opérationnel très important pour développer des outils de souscriptions et de former les agents à de nouveaux processus.

La partie suivante étudie le niveau d'information suivant apportée par la tarification à l'adresse - les variables externes.

Résultats du contrôle des adresses non géocodées

Types d'adresses		Nb de Contrats	Catégorie rouge	Catégorie orange	Catégorie verte	Autres
Sans Chiffres	Dénomination incorrect (Den.Inc)	531 441	Proportion moyenne : 0.9639 (+/- 0.0329) Nb moyenne : 512 256 (+/- 17 484)	Prop. moy. : 0.0241 (+/-0.027965) Nb moy. : 12 808 (+/-14 862)	Prop. moy. : 0.012 (+/-0.02) Nb Moy. : 6 377 (+/-10491)	-
Sans Chiffres	Dénomination correct (Den.Corr)	151 499	Prop. moy. : 0.9796 (+/-0.023) Nb Moy. : 148 408 (+/-3489)	Prop. moy. : 0.0204 (+/-0.023) Nb Moy. : 3 091 (+/-3489)	-	-
Avec Chiffres	Den.Corr - *	210 834	Prop. moy. : 0.2333 (+/-0.127) Nb Moy. : 49 188 (+/-26705)	Prop. moy. : 0.2333 (+/-0.127) Nb Moy. : 49 188 (+/-26705)	Prop. moy. : 0.5 (+/-0.15) Nb Moy. : 105 417 (+/-31561)	Prop. moy. : 0.0333 (+/-0.054) Nb Moy. : 7 021 (+/-11445)
Avec Chiffres	Den. Inc - *	89 394	Prop. moy. : 0.5385 (+/-0.161) Nb Moy. : 48 139 (+/-14411)	Prop. moy. : 0.0769 (+/-0.086) Nb Moy. : 6 874(+/-7647)	Prop. moy. : 0.3846 (+/-0.156) Nb Moy. : 34 381 (+/-13970)	-
Avec Chiffres	Den.Corr – Chiffres-lettres	5 699	Prop. moy. : 0.3 (+/-0.239) Nb Moy. : 1710 (+/-1359)	Prop. moy. : 0.2 (+/-0.207) Nb Moy. : 1140 (+/-1181)	Prop. moy. : 0.5 (+/-0.26) Nb Moy. : 2850 (+/-1481)	-
Avec Chiffres	Den. Inc – Chiffres-lettres	1 846	Prop. moy. : 0.6316 (+/-0.183) Nb Moy. : 1166 (+/-337)	Prop. moy. : 0.3158 (+/-0.176) Nb Moy. : 583 (+/-325)	Prop. moy. : 0.1053 (+/-0.115) Nb Moy. : 194 (+/-213)	-
Avec Chiffres	Den.Corr – LOT	1 596	Prop. moy. : 0.6 (+/-0.255) Nb Moy. : 958 (+/-407)	Prop. moy. : 0.2 (+/-0.207) Nb Moy. : 319 (+/-331)	Prop. moy. : 0.2 (+/-0.207) Nb Moy. : 319 (+/-331)	-
Avec Chiffres	Den. Inc – LOT	20 227	Prop. moy. : 0.7812 (+/-0.12) Nb Moy. : 15801(+/-2429)	Prop. moy. : 0.125 (+/-0.095) Nb Moy. : 2528 (+/-1930)	Prop. moy. : 0.0938 (+/-0.086) Nb Moy. : 1897 (+/-1730)	-
Avec Chiffres	Den. Corr – Ponctuation	8 956	-	Prop. moy. : 0.1 (+/-0.156) Nb Moy. : 896(+/-1400)	Prop. moy. : 0.8(+/-0.207) Nb Moy. : 7165(+/-1856)	Prop. moy. : 0.1(+/-0.156) Nb Moy. : 896(+/-1400)
Avec Chiffres	Den. Inc – Ponctuation	4 543	Prop. moy. : 0.5 (+/-0.26) Nb Moy. : 2272 (+/-1181)	Prop. moy. : 0.3 (+/-0.239) Nb Moy. : 1363 (+/-1084)	Prop. moy. : 0.2 (+/-0.207) Nb Moy. : 909 (+/-942)	-
Avec Chiffres	Den. InCorr – RTE	1 679	Prop. moy. : 0.2273 (+/-0.146) Nb Moy. : 382 (+/-246)	Prop. moy. : 0.3182 (+/-0.163) Nb Moy. : 534 (+/-273)	Prop. moy. : 0.4545 (+/-0.174) Nb Moy. : 763 (+/-293)	-
Avec Chiffres	Den. InCorr – DOMAINE	1 250	Prop. moy. : 0.5 (+/-0.184) Nb Moy. : 625 (+/-230)	Prop. moy. : 0.1 (+/-0.11) Nb Moy. : 125 (+/-138)	Prop. moy. : 0.35 (+/-0.176) Nb Moy. : 438 (+/-220)	Prop. moy. : 0.05 (+/-0.081) Nb Moy. : 62 (+/-101)
Avec Chiffres	Moins de 10 caractères	6 834	Prop. moy. : 0.76 (+/-0.14) Nb Moy. : 5194 (+/-956)	Prop. moy. : 0.08 (+/-0.089) Nb Moy. : 547 (+/-607)	Prop. moy. : 0.16 (+/-0.12) Nb Moy. : 1093 (+/-821)	-
Avec Chiffres	Caractères spéciaux	8 070	Prop. moy. : 1 (+/- 0) Nb Moy. : 8070 (+/-0)	-	-	-
Avec Chiffres	Chiffres aberrants	133	Prop. moy. : 0.8571(+/-0.155 Nb Moy. : 114(+/-21)	-	Prop. moy. : 0.1429 (+/-0.155) Nb Moy. : 19(+/-21)	-
Avec Chiffres	Adresses avec des dates	7 311	Prop. moy. : 0.7619 (+/-0.153) Nb Moy. : 5 570 (+/-1118)	Prop. moy. : 0.0952 (+/-0.105) Nb Moy. : 696 (+/-770)	Prop. moy. : 0.1429 (+/-0.125) Nb Moy. : 1045 (+/-914)	-
Total estimation *		1 051 312	~799 853	~67 884	~162 867	~7 979
Proportion (en %)		100%	~76,1 %	~6,5%	~15,5%	~ > 1%

TABLE 2.1 – Mars 2020 - contrôles des adresses non géocodees.

2.2 Les informations externes

De nombreuses bases et sources de données ont été utilisées pour ce projet. Chaque variable a des spécificités qui lui sont propres. Dans cette thèse, plusieurs types de variables seront mentionnés dans la partie 2.2.1. Ensuite, des exemples de variables seront présentés sur les données à l'INSEE (2.2.2), sur les données à l'adresse (2.2.3) et sur les données météorologiques (2.2.4).

2.2.1 La typologie des variables disponibles

Les données dites "externes" sont des données qui ne sont pas récupérées lors de la souscription ou pour la gestion de l'indemnisation. Ces données peuvent être réparties par typologie. En fonction de la méthode de collection des données et des caractéristiques des sources, les variables sont plus ou moins robustes. Le coût d'une variable dépend du coût d'achat des données, de la création de l'infrastructure, des jointures, de la maintenance des sources, des modèles et de l'IT.

Voici une liste non exhaustive des types de variables disponibles :

— **Open Data** : Information à l'échelle de l'iris : chômage, nombre d'employés, nombre d'hôpitaux dans un iris, nombre de gares ...
— **Variable de mesures** : Nombre de bâtiments dans une parcelle, surface de la parcelle, nombre de bâtiments dans un rayon de 100 mètres ...
— **Variable de distance** : Distance à un cours d'eau, altitude, différentiel d'altitude avec le cours d'eau le plus proche, distance à la caserne de pompier la plus proche, distance à la rampe d'autoroute la plus proche ...
— **Variables computationnelles** : Surface d'empreinte au sol, type de toit, présence de velux ...
— **Variables apprises** : Surface habitable, consommation théorique énergétique, panneaux solaires, type de matériaux du toit ...
— **Variables spécifiques** : Variables météorologiques, type de chauffage, la période de construction ...

Les variables du premier type concernent les données accessibles librement avec peu de travail sur la donnée - voir la section 3.1.2.c. Souvent, ce sont des variables avec une bonne qualité de données. Le second type correspond à des variables qui sont déduites des référentiels bâtiments ou parcelles. La qualité dépend uniquement de la qualité du référentiel, les méthodes de mesures étant robustes et simples. La qualité de ces variables dépend grandement de la géolocalisation (quasi-uniquement). Une troisième typologie concerne les variables de distance, nécessitant une géolocalisation précise d'un certain nombre de bâtiments. La qualité de ces variables est plus dépendante de l'exhaustivité de la position des éléments que de la géolocalisation précise. En effet, pour la plupart des variables de distance, un écart d'une 50 mètres impacte peu la qualité relativement à l'évaluation du risque assurantiel. Un avant-dernier type de variables peut être mentionné : les variables computationnelles. Ce sont des variables nécessitant un calcul non trivial sur la géométrie du référentiel bâtiment. Finalement, un dernier type de variables concerne les variables nécessitant des traitements spatiaux spécifiques et non triviaux. Par exemple, les variables météorologiques doivent être interpolées et corrigées, les variables comme la période de construction ou le type de chauffage qui sont souvent disponibles à la maille du quartier et doivent être transformées à l'adresse.

Les variables les plus complexes sont les variables computationnelles et les variables apprises, car elles nécessitent un nombre de labels (représentatifs) pour être déterminées. De plus, elles dépendent de la qualité des autres éléments et des autres variables.

Globalement, une variable appartient à plusieurs types en même temps. Provenant de différentes sources, les méthodes de jointures entre les informations disponibles et le référentiel bâtiment sont souvent différentes. Quand les informations ne sont pas disponibles, une complétion est faite à partir des autres informations sur le bâtiment et en utilisant les bâtiments déjà labellisés.

Dans les sections suivantes, des variables disponibles seront montrées par des cartographies. De mon point de vue, lors d'utilisations de données externes en MRH, les cartographie spatiales et leurs analyses sont indispensables. L'analyse cartographique est ici visuelle, car suffisamment efficace pour notre utilisation. Elle est plus rapide que l'utilisation de diagramme de Moran, d'indice de Moran (Moran, 1950 [68]) ou d'indice de Geary (Geary, 1954 [65]) pour estimer l'existence d'auto-corrélation spatiale. En revanche, des indicateurs comme Moran I local (Anselin 1995 [64]) combinés avec la méthode de Holm (Holm, 1979 [66]) seraient complémentaires aux analyses pour détecter des clusters locaux et seraient très intéressants à mettre en oeuvre, mais couteux en calculs.

L'objectif des parties suivantes est de mettre en avant les pièges lors de l'utilisation de telles variables. Toutes les cartographies sont basées sur des portefeuilles maison d'assureurs. Les zones en blanc corres-

pondent à une non-exposition. Pour la quasi-totalité des variables, il n'y a pas de changements majeurs pour les appartements et dans l'analyse spatiale. Chaque variable a un indice associé de qualité qui prend les modalités "Very Low", "Low", "Medium", "High" et "Very High".

2.2.2 Les données à la maille INSEE

Les données à l'INSEE ou à la municipalité correspondent à des variables démographiques, ou sur l'infrastructure présente dans une commune. Ces données sont usuellement utilisées dans des zoniers utilisant des données externes (voir le chapitre 1). Les cinq exemples sont représentatifs des types de variables disponibles.

La variable *zones urbaines* - carte 2.10 - provenant directement de la cartographie 1.8 est une variable qui est informative et très usuelle. Utilisée pour le zonier, elle est mise en interaction avec les variables des départements. Il est nécessaire de faire très attention aux corrélations spatiales de cette variable avec les autres variables. En effet, cette variable va être très corrélée avec la densité urbaine et ainsi que les infrastructures et du nombre présents dans l'iris. D'une qualité parfaite, souvent, elle ne doit pas être utilisée avec des variables de distance (voir le chapitre 3 pour une discussion à ce sujet) .

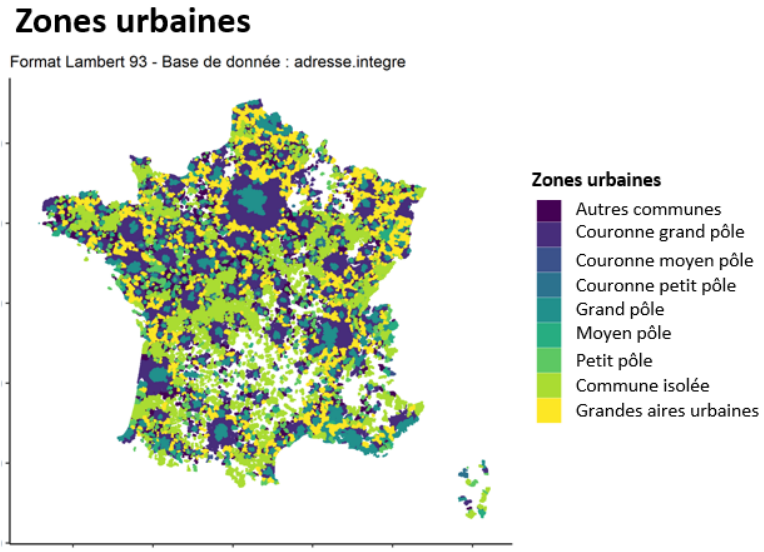


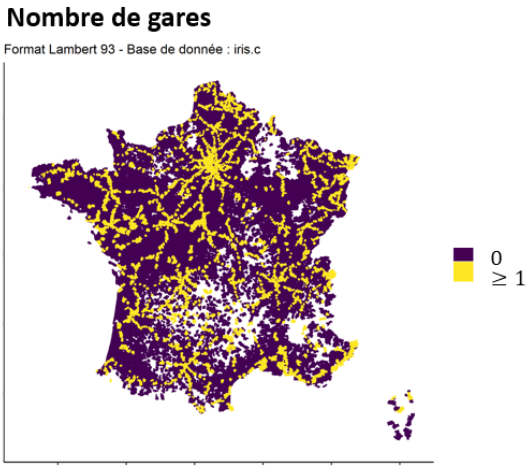
FIGURE 2.10 – Cartographies des zones urbaines

La présence de gares dans l'iris - carte 2.11a est une variable de qualité parfaite. Une spécificité est qu'elle est très corrélée à la variable urbaine ou de densité. De ce fait, cette variable va souvent ressortir significative par corrélation spatiale et non par un effet causal. De façon similaire, le nombre d'écoles - 2.11b - par communes est aussi une variable informative très corrélée avec la densité urbaine. Néanmoins, pour la garantie RC, elle permet de capturer la sinistralité scolaire où le nombre d'enfants induit une fréquence BDG plus important par exemple. Un panel d'informations est accessible comme le nombre de médecins, d'avocats, de commissariats... Figées dans le temps, ces variables sont plutôt très stables temporellement.

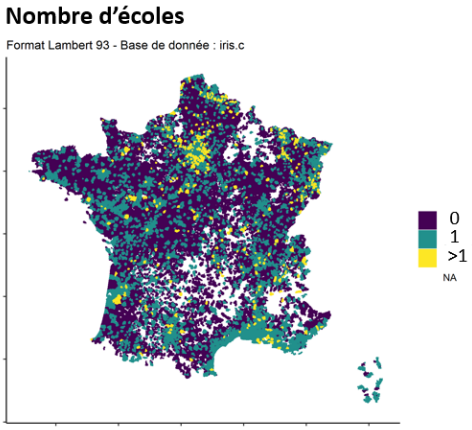
Des variables composites sont disponibles comme *le nombre de cadres* par iris ou *le nombre d'habitants* comme la carte 2.12.

Un ensemble de variables est dit composite si elles sont construites à partir d'une ou deux variables interdépendantes (Levy, 1982 [67]). Dans la majorité de nos cas, les variables composites rencontrées $Z_1, ..., Z_h, P$ sont positives telles que $\sum_{i=1,...,h} Z_i = P$ avec $h \in \mathbb{N}$.

En d'autres termes, par rapport au nombre d'habitants, il est certain que la somme du nombre de cadres, d'ouvriers, de salariés... est toujours égale au nombre d'actifs dans la commune. Ainsi, les variables composites sont difficiles à intégrer dans une structure tarifaire. En effet, les corrélations ne sont pas linéaires. Elles brouillent aussi l'interprétation des tendances ; exemple : est-ce qu'un faible nombre d'employés traduit un nombre plus important de cadres ou bien une densité plus faible ?



(a) Nombre de gares par commune



(b) Nombre d'écoles par commune

FIGURE 2.11 – Des variables à la maille iris qui sont de qualité quasi-parfaite. Elles sont temporellement figées en 2015.

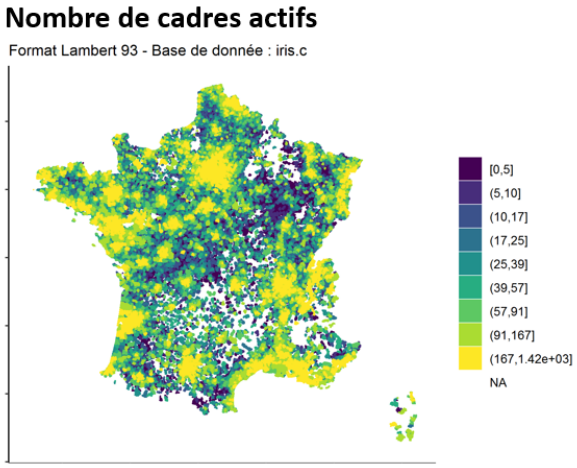


FIGURE 2.12 – Nombre de cadres en 2015 dans la commune.

D'autres variables plus spécifiques comme le *zone d'exposition de l'aléa sismique* - carte 2.13 - sont disponibles. Il est nécessaire de remarquer que cette variable permet de délimiter des zones en particulier. Les zones les plus exposées (en jaune) correspondent aux zones montagneuses tout particulièrement. Ainsi, un effet marginal capturé reflète généralement une corrélation spatiale avec l'altitude et non une fragilisation du bâti du fait de séisme fréquent. Étant donné que les dommages liés à une activité sismique sont compris dans la Garantie CatNat (non tarifé), elle n'est pas utilisée pour la tarification. Cependant, elle peut servir dans la simulation de dommages sismiques pour des besoins de réassurance ou de simulation de sinistralités.

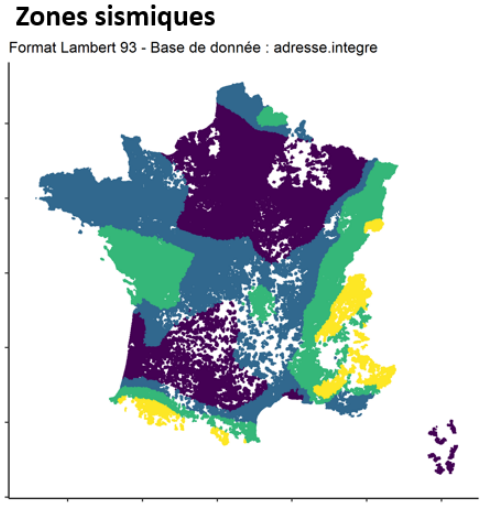


FIGURE 2.13 – Cartographie du zonage sismique de la France. Cette variable est de bonne qualité. (Du risque élevé à faible : Jaune à Bleu)

Ces variables à la maille de l'INSEE ont l'avantage d'être de bonne qualité, mais d'être de façon générale très corrélées à la densité. Ainsi, il n'est pas intéressant de regarder les indices de qualité des variables. L'inconvénient principal est qu'elles sont à une maille assez large et ne permettent pas de segmenter plus qu'un zonier classique. En revanche, elles permettent une extrapolation justifiable et pertinent dans les zones à faible exposition.

2.2.3 Les données à l'adresse et au bâtiment

Les variables les plus intéressantes et segmentantes se trouvent à l'adresse. La qualité des variables de ces variables est très dépendante de la précision de la géolocalisation, mais aussi de l'auto-corrélation spatiale et des bases de données sources. Les cartographies sont représentées comme la moyenne de l'ensemble des valeurs des bâtiments présents dans la commune ou le plus fréquent pour les variables catégorielles.

Les pendants continus et à l'adresse de la variable de zones urbaines sont les variables comptant *le nombre de bâtiments résidentiels dans un rayon de 50 mètres ou 100 mètres*. Se fondant sur le référentiel bâtiment, elle permet à l'intérieur même d'une commune de différentier les bâtiments isolés des autres. De façon complémentaire, des variables comme *la distance au bâtiment résidentiel le plus proche* ou *la mitoyenneté* permettent d'affiner l'information d'urbanisation.

De manière générale, les effets marginaux sont importants et la qualité des variables très bonnes. Cela est d'autant plus vrai que la précision n'impacte que très peu la sinistralité. En revanche, pour des raisons de stabilité des données, la variable *nombre de bâtiments dans un rayon de 100 mètres* est plus robuste que celle à 50 mètres pour quasiment la même information. De plus, il est nécessaire de ne pas mettre en même temps une variable *zonier urbain* avec *nombre de bâtiments dans un rayon de 100 mètres*. En revanche, la combinaison de cette dernière variable avec la distance à un bâtiment résidentiel (très corrélée) est souvent pertinente et informative. Ces informations ne sont pas captées par les zoniers traditionnels.

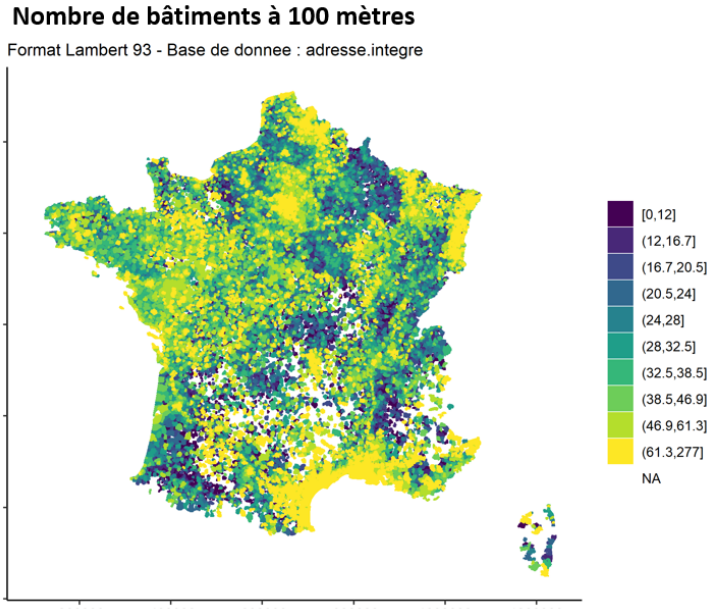
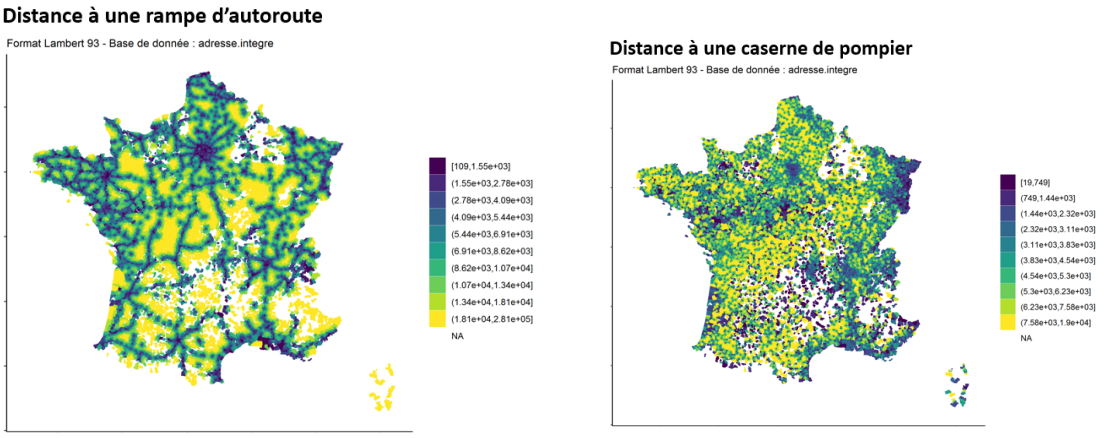


FIGURE 2.14 – Cartographie de la variable du nombre de bâtiments dans un rayon de 100 mètres. Même si la variable est de bonne qualité, il est nécessaire de remarquer que certaines zones semblent sous-estimées. En effet, dans le nord-est, les bâtiments sont moins souvent bien découpés. Ce type de problème peut être résolu en considérant les parties résidentielles par la suite.

Il existe d'autres variables à l'adresse de très bonne qualité fondées sur des calculs de distances comme les cartes 2.15a et 2.15b. Les variables de distances sont très fortement corrélées spatialement aux zones urbaines à cause de leur caractère clairsemé en zone rural. Il serait plus intéressant de calculer une distance par route ou une durée de trajet en voiture entre deux points. Cependant, de telles variables sont beaucoup plus difficiles à calculer et à massifier [4], induisant un coup de développement beaucoup plus élevé pour des effets marginaux. Souvent, ces informations sont très reliées à la distance à un centre-ville, en particulier pour la variable *distance à une caserne de pompier* rendant redondante une majeure partie de l'information.

4. Définition *massifier* : calculer les valeurs d'une variable sur l'ensemble des bâtiments français.



(a) Distance à une bretelle d'autoroute.

(b) Distance à une caserne de pompier.

FIGURE 2.15 – Des variables de distances calculées à vols d'oiseaux.

Une variable fondamentale est la *valeur de la maison* - carte 2.16. Se basant sur la base **DVF**[a] majoritairement, une extrapolation avec les caractéristiques des bâtiments est appliquée pour tous les biens n'ayant pas eu d'échanges depuis 2014. C'est une des variables les plus importantes et significatives. De qualité très bonne à la maille communale et correcte à l'adresse, la précision pour la modélisation **MRH** est de l'ordre de plusieurs centaines d'euros. Par nature la valeur d'une maison est volatile et change assez significativement dans le temps. Les données anciennes sont donc à prendre avec plus de précautions. De ce fait, la qualité du géocodage impacte peu la qualité de la variable.

a. Demandes de valeurs foncières

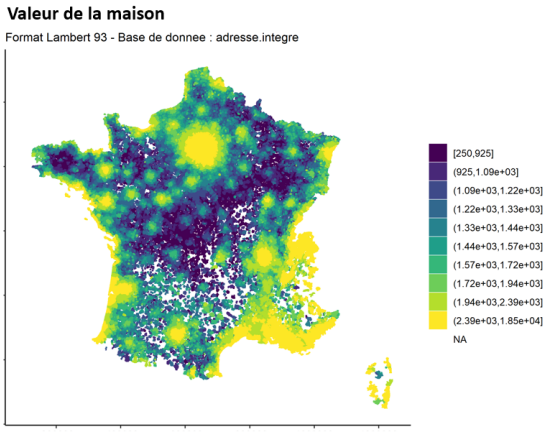
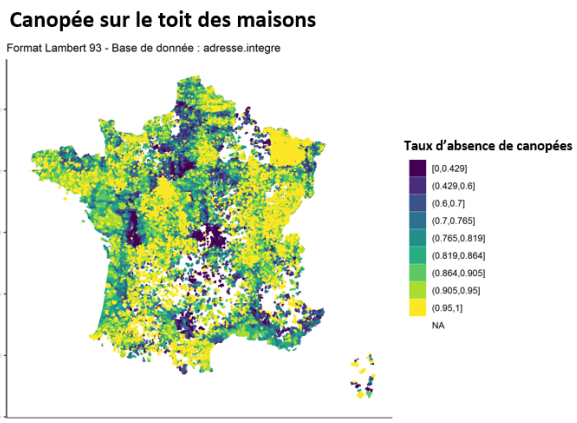


FIGURE 2.16 – Cartographie de la variable *valeur de la maison*. Il y a une forte corrélation avec des variables d'urbanisation et des variables de richesses, *le nombre de cadres* par exemple.

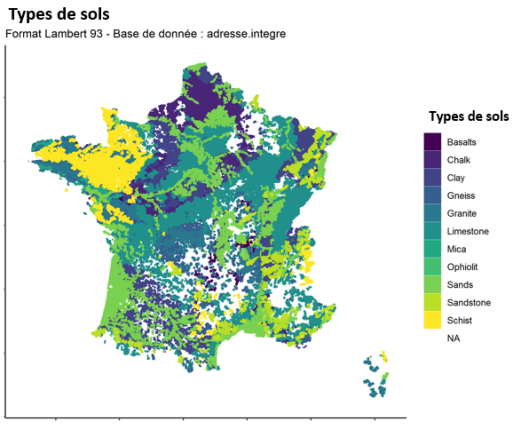
Les variables de *computer vision* comme *la présence de canopée sur le toit* 2.17a sont les variables les plus complexes à appréhender. La qualité de ces variables dépend des images, des éléments à détecter, le type de toit et la géolocalisation du bâtiment. Par exemple, l'auto-corrélation spatiale des panneaux solaires est très faible à la maille des bâtiments. Par conséquent, la qualité de la variable est très impactée par la géolocalisation. Au contraire, le type de matériaux du toit (voir l'exemple dans le chapitre 6) est très autocorrélé spatialement. De ce fait, une erreur de géolocalisation en Bretagne est très peu impactante, car tous les bâtiments utilisent l'ardoise. Cette dernière information sur les matériaux du toit est aussi une information importante pour estimer la qualité des variables de computer vision : un élément noir sur un toit noir est difficile à analyser. En plus de nécessiter une étude spatiale de la variable, il est nécessaire de faire attention à la qualité spatiale de la variable à travers les indices de qualité.

Un exemple de corrélation spatiale complexe est celui de la variable *type de sol* - carte 2.17b. La variable provient du **BRGM**[5] et est de bonne qualité et robuste aux erreurs de géolocalisation. L'auto-corrélation spatiale est aisée à observer. Lors des différents travaux, la modalité "Schist" est très souvent ressortie significative en VOL. En effet, elle venait capturer la sinistralité particulière de la Bretagne. De même, dans les premières modélisations, elle ressort toujours significative lors d'utilisation de variable de computer vision car la qualité des variables de computer vision en Bretagne était très faible à cause de l'ardoise et cette modalité venait réajuster les erreurs.

5. Bureau de Recherches Géologiques et Minières.

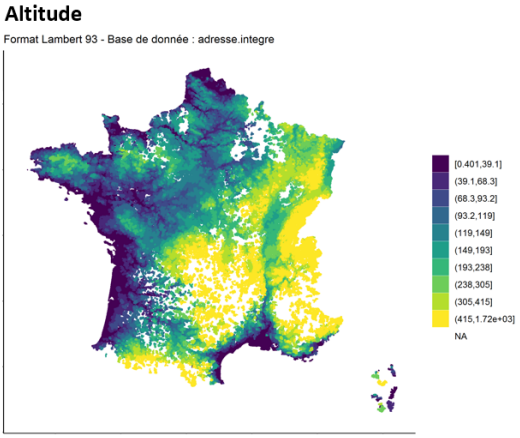


(a) Cartographie de la variable *présence de canopée*.



(b) Cartographie du *type de sol* majoritaire associé au bâtiment.

FIGURE 2.17 – De nombreuses auto-corrélations spatiales sont à prendre en compte lors des modélisations.



Le *type de sol* est une variable souvent significative pour les garanties climatiques (sécheresse, inondation, remontée de nappes phréatiques) mais aussi à cause de corrélations avec *l'altitude* - carte 2.18. Cette dernière, de très bonne qualité, est l'une des variables qui représente le mieux les difficultés provenant des corrélations entre les variables météorologiques et les caractéristiques des bâtiments. La difficulté majeure réside dans l'apparition de zones locales où la corrélation spatiale entre les variables diminuent la significativité des variables à l'adresse.

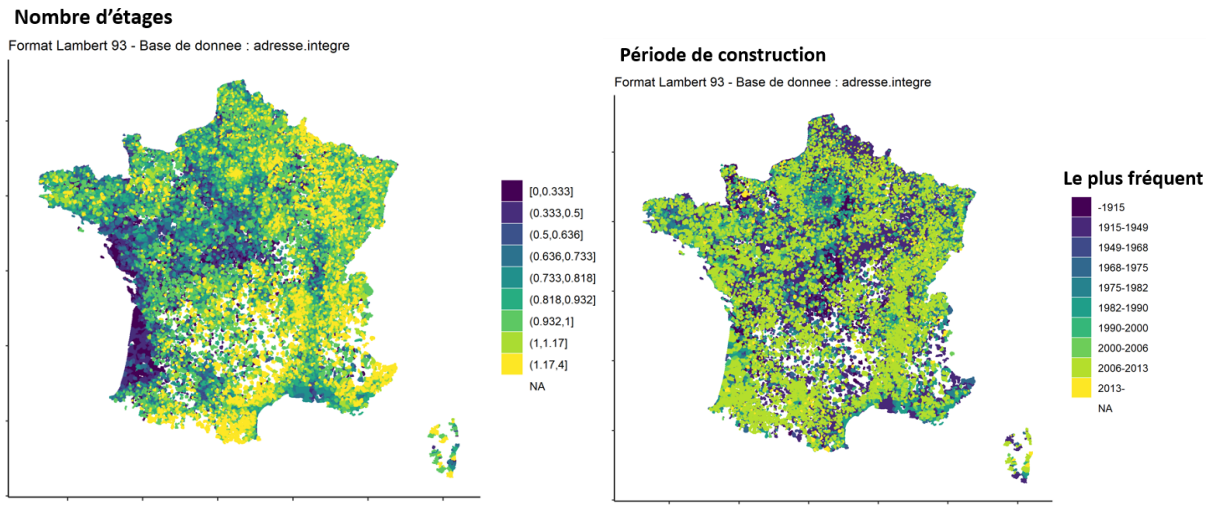
FIGURE 2.18 – Cartographie de *l'altitude des maisons*.

Les variables aux bâtiments sont les plus impactées par la qualité du géocoding comme *le nombre d'étages* et *la période de construction* (2.19a et 2.19b). Bien sûr, il existe des dépendances spatiales locales mais elles sont plutôt faibles par rapport aux autres variables comme *la surface habitable* ou *le nombre de pièces*. Ces variables ont des effets marginaux majeurs sur la sinistralité, mais sont aussi très corrélées avec l'urbanisation. L'urbanisation est peu impactée par la géolocalisation. La corrélation avec les variables et leurs qualités induit des effets complexes à analyser (Voir la partie sur les effets la qualité de la donnée liée à la géolocalisation sur les coefficients des GLMs 6.3.1). Les variables sur le bâtiment sont souvent complétées/corrigées (jusqu'à 80 % des bâtiments sont modifiés).

Souvent, plusieurs sources coexistent à différentes mailles et des méthodes d'agrégations doivent être faites. La variable *période de construction* en est l'exemple majeur. D'un côté, des données à la maille Iris (ou plus fin) comptant le nombre de bâtiments construit entre telles et telles périodes existent. D'un autre côté, la base ADEME (sur les rapports obligatoires suites à des travaux) renseigne à la maille du bâtiment l'année de construction. En conséquence, la variable *période de construction* est construite de façon assez complexe et sa qualité est robustifiée à la maille de l'iris. Il est à noter qu'il existe des biais dans la donnée initiale. Les dates de constructions peuvent être fausses ou approximatives. *nam.R* a démontré qu'il y avait une sur-présence de bâtiments construits avant 1949. En effet, à une époque, il y avait moins de démarches administratives à réaliser pour les bâtiments anciens. Que faire si une extension au bâtiment a été ajouté ? Cette même problématique se retrouve pour les **DPE** [6] où il semble que beaucoup de classes E seraient en réalité catégorisés en classe D. De plus en fonction de la date

6. Diagnostic de performance énergétique.

du DPE, les catégories ne correspondent pas à la même chose. Ce n'est en réalité pas un problème actuariellement parlant car la donnée est stable. Néanmoins, la question se pose sur l'autorisation d'utiliser des informations différentes de celles déclarées.



(a) Cartographie du *nombre d'étages*.

(b) Cartographie de *la période de construction*.

FIGURE 2.19 – Variables à l'adresse à la maille du bâtiment.

Cette problématique est d'autant plus frappante pour *la présence de piscine*. Provenant majoritairement des informations fiscales, il est connu qu'elle fait l'objet de fraude/non-déclaration assez récurrente. Est-ce qu'un tarif peut utiliser le fait que fiscalement la maison n'a pas déclaré de piscine alors que des méthodes de computer vision en détectent une ?

Certaines variables aux bâtiments ne sont pas disponibles pour les appartements. La surface habitable d'un appartement est quasiment impossible à déterminer ou à approximer. En effet, pour une même adresse, pour un même étage, une grande diversité d'appartements différents en surface et en nombre de pièces co-existent.

Finalement, à l'adresse, des informations légales et communales sont disponibles comme les *PPRN* (Plan de Prévention des Risques Naturels), les variables de *retraits gonflements argiles*. L'impact du géocoding sur cette dernière n'est pas très important. En effet, ce sont des variables très autocorrélées spatialement à la maille du quartier/commune et dont la précision est de l'ordre du quartier. Cela est d'autant plus vrai pour la variable *retrait gonflement argile*[7]. En effet, il est impossible de faire des prélèvements piézométriques dans tous les quartiers. Dès lors la précision est globalement d'une centaine de mètres près. D'autres variables sont impactées par les décisions communales comme la carte 2.20 montre. Certaines communes n'ayant pas fait de *PPRN* mutualiseront un peu par défaut le risque appris contrairement aux autres communes.

Indices de qualités : Les indices de qualité sont disponibles et pertinents pour les variables à l'adresse. L'observation spatiale de ces indices permet de détecter certain biais. Cependant, les indices sont utilisés plus à titre indicatif et informatif que pour la modélisation. Les indicateurs de qualités de données ne prennent pas en compte la problématique de la qualité du géocoding suite au rattachement d'une adresse à un bâtiment. Après un grand nombre d'échanges avec *nam.R*, les indices de qualités représentent en premier lieu la crédibilité des sources et de leurs jointures avec le référentiel bâtiment. Ensuite, les méthodes utilisées (complétion, interpolation, correction) ainsi que les autres variables associées permettent d'individualiser la notion de qualité. Cette étape apporte une information de crédibilité de l'observation, mais aussi de son imprécision. La combinaison de ces deux éléments pour faire l'indice final est restrictive[8]. En bref, les indices de qualité pour les variables à l'adresse sont des indices de crédibilité et d'imprécision. Bien qu'ils soient imparfaits, ils restent pratiques pour guider les modélisations. Pour chacune des variables, un détail bien précis de la construction est disponible. La création des indices de qualité dépend des sources, des variables et ne sera pas plus discuté que cela dans cette thèse.

7. Voir chapitre 7.
8. Si la qualité de la source est *"Medium"* et la qualité de la jointure est *"Low"*, l'indice de qualité final est *"Low"*.

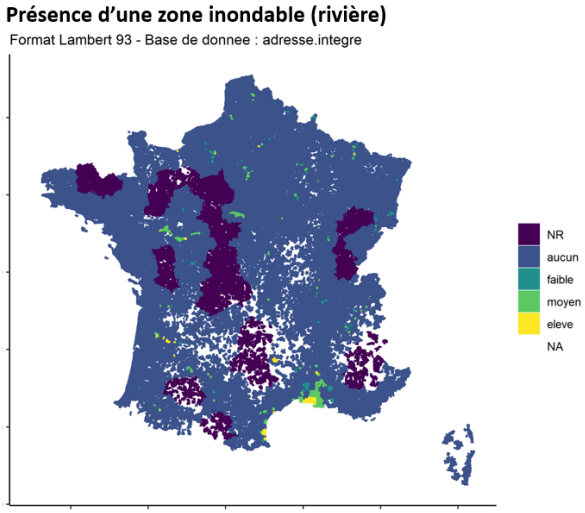
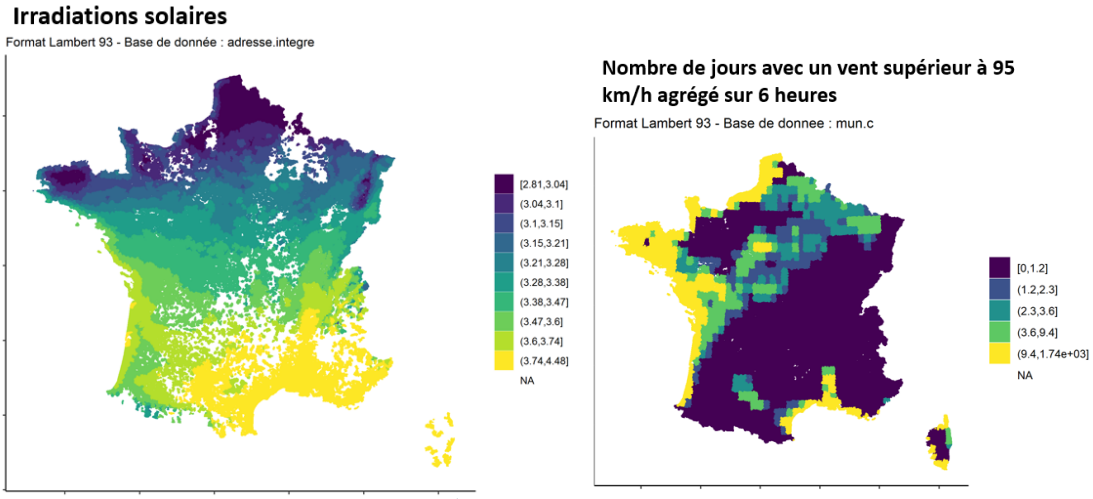


FIGURE 2.20 – Cartographie du type de risque inondation donné par les PPRN. Les variables PPRNs sont de bonne qualité. L'intérêt de la cartographie à la commune est faible car les distinctions se font à l'intérieur d'une commune.

2.2.4 Les variables météorologiques



(a) Cartographie l'irradiation solaire moyenne sur un historique de plus de 20 ans.

(b) Cartographie d'une variable sur les niveaux de vent : moyenne de 2010 à 2018.

FIGURE 2.21 – La précision des variables est souvent entre 20 à 50 km en fonction de l'interpolation dans les premières versions des données. Dans les versions les plus récentes, le niveau de précision a été augmenté.

Les variables météorologiques sont importantes en **MRH**. En effet, contrairement à l'assurance automobile, le bien assuré ne se mue pas spatialement et subit donc tous les aléas météorologiques de la commune. Plusieurs types de variables météorologiques sont disponibles. Le nombre de variables disponibles est de l'ordre d'une centaine (précipitations, gels, neiges, températures, vents).

Les **variables en moyennes historiques** sont la moyenne sur plusieurs années d'indicateurs annuels ou saisonnaux. Elles renseignent surtout sur le type de climat : par exemple, *l'irradiation solaire* - 2.21a ou la vitesse du vent. Cependant, certains **indicateurs "quantiles" extrêmes**, comme les rafales de vents ne sont pas extrêmement robustes et informatifs sous forme de moyenne. En effet, une moyenne ne permet pas d'appréhender les effets d'un évènement climatique sur la sinistralité. C'est pourquoi, pour toutes les années à partir de 2010, quasiment l'ensemble des indicateurs météorologiques sont aussi observées

annuellement comme le *nombre de jours de gel* - carte 2.22.

Il est notable que la France métropolitaine a des climats facilement identifiables. En d'autres termes, à partir de trois indicateurs météorologiques, il est aisé de caractériser l'emplacement des communes. Lors de modélisations, il est donc souvent recommandé d'utiliser au maximum deux variables météorologiques pour éviter le sur-apprentissage. Cette remarque est d'autant plus important dans les modèles RF ou XGBoost lorsque la qualité des variables à l'adresse n'est pas parfaite.

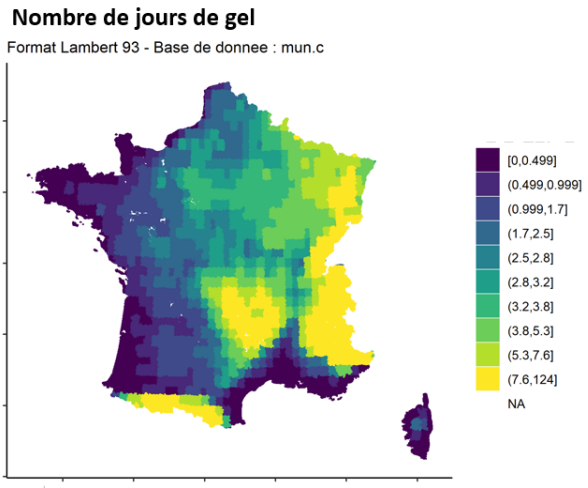


FIGURE 2.22 – Cartographie d'une variable météorologique sur *le nombre de jours de gel*.

Pour appréhender les risques spécifiques comme la sécheresse, des indices de sécheresses (comme *SPI*, *RDI*, *SSWI*) captent des informations à travers plusieurs caractéristiques météorologiques (Cf l'article sur la sécheresse pour plus de détails - chapitre 7). Ces indicateurs sont complexes à utiliser, car très spécifiques. La problématique de corrélation spatiale et d'auto-corrélations est fréquente comme le montre la carte de la duration du SPI ; moyenne sur 10 ans - carte 2.23. Ces indicateurs sont majoritairement utilisés qu'annuellement.

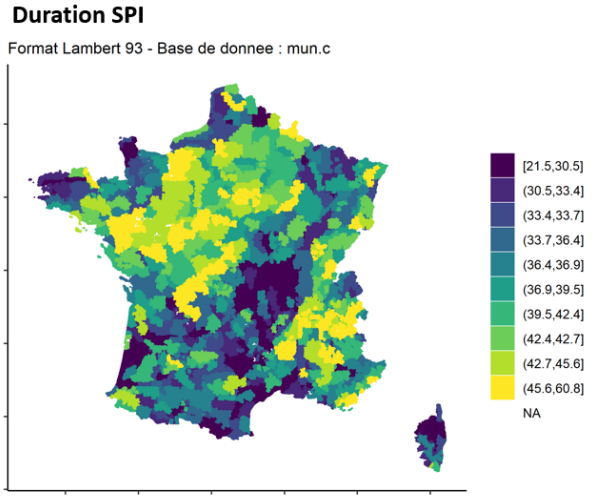


FIGURE 2.23 – Un exemple d'indicateurs météorologiques.

De manière générale, ces variables sont interpolées sur une grille 20 km × 20 km. Elles proviennent de source de données bien connues et retraitées comme **ERA5** ou de Météo France. Le choix de la source est important et modifie les valeurs pour une même variable. Pour un usage assurantiel, il est à noter que la précision nécessaire des variables météorologiques est moyennement élevée et elles sont d'excellentes qualités. Au contraire, les indices de qualités des variables météorologiques mesurent l'imprécision à un niveau très fin. En conséquence, ces indices de qualités de données apportent alors une information inutile, car trop précise et ne sont pas utilisés.

Conclusion

Les données à l'adresse se sont construites dans le temps. Il faut mentionner que c'est grâce à la politique de transparence et d'open data de la France que ce projet a pu voir le jour. Tout d'abord, l'adressage des bâtiments mis en place pour la Poste rend la géolocalisation des bâtiments possible. Dans d'autres pays comme en Amérique centrale, la notion d'adresse n'est pas figée et potentiellement ne le sera jamais avec le déclin des envois postaux. De la même façon, c'est à travers les législations, les études étatiques, les sondages qui sont mis en œuvre en France depuis des décennies que les données ont pu être récupérées, comparées et étudiées.

En revanche, la mise à disposition des données ne doit pas occulter les aspects éthiques et RGPD dont les assurés et les prospects bénéficient et auxquels les assureurs doivent se contraindre. S'il y a une contradiction, une incohérence ou une erreur dans la donnée, quelles règles de proportionnalités doivent être mises en place et quels droits un assuré peut opposer à un assureur ? Ce sont des questions légales qui vont se poser de plus en plus sans que des réponses simples ni d'orientations ne peuvent être proposées.

Finalement, la mise à jour de la donnée est une nouvelle problématique dont les processus restent à définir. L'amélioration de la qualité d'une variable pendant un processus de souscription ou la mise à jour d'informations vont influer les revalorisations et les primes proposées.

Références

[64] Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical analysis*, 27(2) :93–115.

[65] Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3) :115–146.

[66] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

[67] Ley, P. (1972). Quantitative aspects of psychological assessment : An introduction.

[68] Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2) :17–23.

Chapitre 3

Produit MRH à partir de la tarification à l'adresse

Ce chapitre rapporte le travail de recherche et de développement pour déterminer si les données à l'adresse étaient opérationnellement viables et informatives.

L'ensemble de ces méthodologies ont été validées par les équipes techniques de l'assureur en question. Ces derniers ont utilisé directement les mêmes variables et les mêmes filtres que j'ai développé. Ils ont ajouté leur composante zonier commerciale dans les modèles. Les résultats obtenus sont les mêmes que ceux que j'ai trouvés sur des métriques AUC/Gini et Pseudo-R de Faddens.

Par la suite, au moins cinq assureurs (certains avec un portefeuille de 200 000 d'assurés et d'autres d'un 1 million d'assurés) ont aussi fait l'objet d'études approfondies sur toutes les garanties et les résultats obtenus valident ceux présentés dans ce chapitre. Environ 35 % du marché français en terme d'exposition a été testé.

De ce fait, ce chapitre se réfère à l'année 2020. La base du projet est bien implantée dans le territoire français comme le montre la cartographie 3.1. Il est à noter que grâce à l'amélioration des variables et la géolocalisation, les performances des projets suivants sont largement meilleures.

L'objectif de ce développement était de valider la faisabilité technique du projet. Dans ce cadre, la méthodologie vise à créer des modèles utilisables opérationnellement, mais aussi à déterminer les points faibles et les points d'améliorations des données. Les résultats présentés dans cette section ont été pris de la sixième modélisation. Durant ce travail, la base de donnée a été mise à jour six fois dont trois mises à jour ont fait l'objet de changement de géocodeurs.

Ces travaux ont nécessité un travail approfondi sur quatre verrous techniques. Le travail en grande dimension est le premier. Avec plus de 200 variables au début du projet, une méthode de sélection de variables a été développée en utilisant des méthodes supervisées et non supervisées pour une utilisation de GLM. Le second verrou est d'évaluer la qualité de la donnée et quantifier ses impacts et sa dynamique. Pour aller plus loin, il sera nécessaire d'évaluer la performance lors de la substitution de variables et de mettre en œuvre des méthodes pour la comparaison des modèles non emboités. Finalement, le dernier verrou est de l'ordre du produit, avec la nécessiter d'évaluer les impacts éthiques, économiques et souscriptions apparaissant lors du changement de méthodes de souscription.

3.1 Le contexte du projet

Traditionnellement, les assureurs récupèrent des informations sur le(s) bâtiment(s) et le bénéficiaire à travers des questions lors de la souscription. En tarification, une meilleure segmentation (plus de variables ou mieux définies) que les concurrents permet de mieux contrôler l'anti-sélection. Dans ce cadre, les modèles utilisés se doivent d'être les plus performants possibles sur l'ensemble des risques. Néanmoins, les résultats de la tarification proviennent aussi du processus de souscriptions et un nombre important de questions réduit le nombre de devis terminés. En France en 2020, 19 questions sont demandées en moyenne lors du questionnaire de souscriptions [1]. Une disparité existe entre les entreprises : les banques-assureurs demandent moins de questions que les assureurs traditionnels. De façon générale, une douzaine de questions concernent le(s) bâtiment(s) et une demi-douzaine portent sur des informations personnelles. Le reste des questions portent sur le contrat d'assurance souscrit. Le nombre de

1. Travaux de Silvia BUCCI et Linda KROLIKOWSKI pour Addactis FRANCE

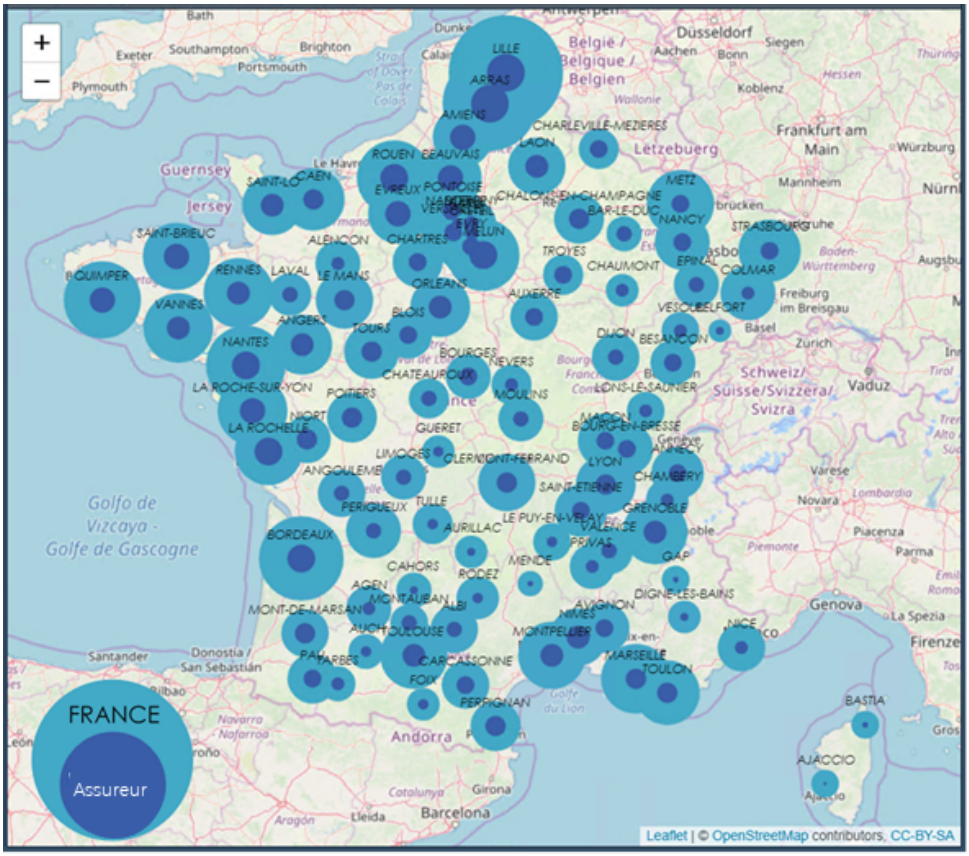


FIGURE 3.1 – Implantation du portefeuille des maisons individuelles dans le territoire français. (Cartographie comparant les contrats de 2016-2018 aux nombres d'habitations individuelles en 2016 crée par Victoria Delavaud).

variables utilisées lors de la détermination du tarif provient très largement de ces réponses. Ainsi, le nombre de variables utilisables est limité et la complexité des variables est adaptée. Par exemple, peu d'assureurs demandent la taille précise de la surface habitable d'une annexe. En effet, le prospect peut ne pas connaitre la réponse précise. Le fait de poser ce type de questions peut avoir un effet négatif, surtout pour les devis en ligne où le prospect va tester différents assureurs simultanément avec un temps limité ou bien le commercial se doit de relancer le prospect impactant son travail. En conséquence, un trop grand nombre de questions ou trop complexes diminuent le nombre de devis finis ou la qualité des réponses.

C'est dans ce contexte que la tarification à l'adresse a été étudiée en premier lieu pour augmenter le nombre de variables sans avoir l'inconvénient d'accroitre la durée de souscription avec les méthodes qui sont usuellement utilisées en tarification : les **GLM**s.

3.1.1 L'apport des données externes

La géolocalisation des adresses permet d'accéder à un grand nombre d'informations grâce à l'agrégation de données Open source, des données non structurées externes, grâce à de l'analyse d'images (*computer vision*) ou de la complétion de données manquantes. Classée quatrième en 2020 par l'indice Global Open Data, la législation française évolue rapidement vers encore plus d'accessibilité envers des données de natures diverses. Plus particulièrement, de plus en plus de données sont disponibles sur le thème de l'énergie, des bâtiments et des risques climatiques. De plus, en France, l'**IGN** a photographié l'ensemble du territoire français et remet à jour les images tous les trois ans idéalement. Dans ce projet, les méthodes de computer vision et **ML** permettent de détecter la présence de velux sur les toits ou de compléter certains variables comme la surface dite habitable. De plus, les méthodes **ML** ont largement accéléré la sélection de variables ou l'étude des données.

Avec l'utilisation extensive de données externes, deux questions se posaient au départ :

— *Peut-on tarifer correctement l'assurance habitation individuelle en utilisant uniquement l'âge et la géolo-*

calisation du bâtiment assuré ?
— *Qu'apportent les nouvelles variables à l'adresse à l'assurance habitation individuelle ?*

3.1.1.a L'amélioration de performance des modèles

L'augmentation de la segmentation semble être limitée par la base de données disponibles, de sa qualité, mais aussi de la souscription. Dans le cadre de ce projet, il est opportun de parler d'hyper-segmentation/hyper-individualisation du tarif. Les limites sur nombre de variables et sur leur complexité sont plus faibles, cependant elles ont un coût de création. L'équivalent direct de la tarification à l'adresse est la télématique pour l'assurance automobile. Sur ce sujet, Barry et Charpentier, 2020 [69] ont étudié l'individualisation du risque sur l'assurance deux roues. Lemaire et al. (2016) [87] explique que l'individualisation du risque s'arrête quand *(trad :) "le coût d'ajouter plus de facteur de risques est plus élevée que le profit qu'apporte la classification du risque".* Le coût de maintenance de ces variables externes est plus élevé que celui de la souscription. Dans notre cas, seul le coût des variables de computer vision est plus élevé que le profit engendré. C'est pourquoi un fournisseur de données édifie les bases de données. Ainsi, en mutualisant les coûts sur l'ensemble des acteurs assurantiels mais aussi d'autres secteurs, la solution devient viable économiquement.

3.1.1.b Les changements du parcours de souscriptions

Le processus de souscription est complètement modifié par l'apport de la géolocalisation. Contrairement à une souscription traditionnelle, l'évaluation de la prime dépend du choix du bâtiment. Au lieu ou en plus de répondre à un questionnaire, le prospect doit choisir le ou les bâtiments qu'il veut assurer - voir la section 3.2. Cette approche digitale et la réduction du temps passée lors de la souscription peuvent augmenter le nombre de devis, l'expérience client et donc la réduction des coûts. En agences, ce temps gagné peut être consacré pour discuter d'autres sujets, d'autres produits et des explications sur le contrat, du conseil (DDA). Finalement, puisque tous les bâtiments du marché français sont connus, une évaluation complète peut être fait pour comparer une exposition à celui du marché ou un niveau de prime.

3.1.1.c Les approches considérées

Les modèles attritionnels suivants sont considérés :
— **Modèle Traditionnel (M1) :** Une tarification traditionnelle utilisant comme variables segmentantes les variables de souscriptions ;
— **Modèle Souscription rapide OD (M2) :** Une tarification optimisée pour que le temps de souscription soit le plus rapide possible *c.-à-d.* seules les variables dites *Open Data (OD)*, l'âge de l'occupant et la formule sont utilisées.
— **Modèle Souscription rapide (M3) :** Une tarification optimisée pour que le temps de souscription soit le plus rapide possible *c.-à-d.* seules les variables dites externes, l'âge de l'occupant et la formule sont utilisées.
— **Modèle Performant OD (M4) :** Une tarification optimisée en performance et en segmentation en ajoutant aux variables de souscription les variables Open Data (OD).
— **Modèle Performant (M5) :** Une tarification optimisée en performance et en segmentation en ajoutant aux variables de souscription toutes les variables externes disponibles.

En utilisant uniquement les données externes, l'intérêt est de limiter le nombre de questions. Pour que les modèles (M2) et (M3) soient utilisables, il est souhaitable qu'ils soient au moins équivalents au modèle (M1) - traditionnel - c'est pourquoi ils seront aussi nommés équivalent. Les modèles (M4) et (M5) améliorent toujours la performance par construction et sont appelés performant. Ils le sont par construction car ils utilisent le meilleur sous-ensemble de toutes les variables disponibles.

3.1.1.d Processus de modélisations

Le processus de modélisations pour la tarification actuarielle comme pour d'autres secteurs semblent avoir atteint un consensus. Dans ce travail, la sélection de variables sera très largement étudier et discuter surtout sous la vision de la qualité de la donnée. Le processus de création de la donnée et ses spécificités seront détaillés dans la section 3.1.2. Le choix de la base de modélisations est ainsi très importante, c'est-à-dire la sélection des lignes. En effet, cette étape est très souvent implicite lors de la création de la base de modélisation traditionnelle, les erreurs sont faibles en nombre et l'impact marginal sur la modélisation. Ici, cette sélection est différente en fonction des garanties, des choix des variables et de la géolocalisation. Les résultats sont détaillés dans la section 3.4.

3.1.2 La base de modélisations pour étudier la faisabilité du projet

Dans cette étude, les données proviennent de trois acteurs : un assureur, un fournisseur de données et moi-même. L'assureur fournit les données de sinistralités attritionnels et de contrats sur une historique allant de 2016 à 2018. À partir de la base adresse des contrats, le fournisseur de données a relié chaque adresse à un bâtiment et associé les données externes à disposition. Le graphique 3.2 explique le processus de création de la base de données avec les données externes géolocalisées.

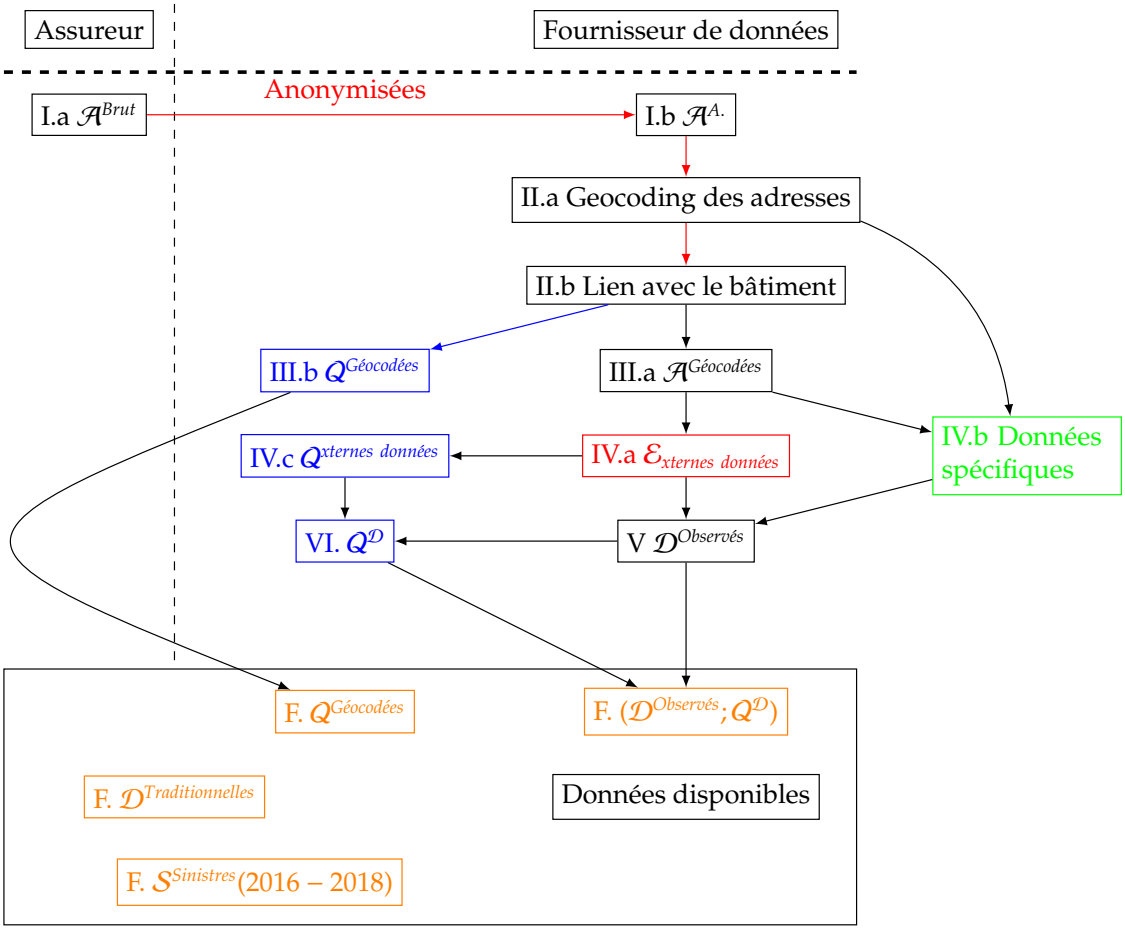


FIGURE 3.2 – Le processus de création de la base de données avec les données externes géolocalisées. Les éléments en rouge représentent les processus qui diminuent le plus la qualité de la donnée, en bleu les processus d'évaluation de la qualité de la donnée. et en vert ceux dont la qualité est très bonne (sans perte de qualité).

(I.) Les informations obtenues $\mathcal{A}^{Brut}$ pendant la souscription comportent des informations sur le bâtiment assuré **(I.a)**, en particulier sa location sous forme d'un code INSEE, du nom de la rue, le nom de la commune, le numéro de rue et d'autres compléments. Ces données sont anonymisées, $\mathcal{A}^{A.}$ et envoyées au fournisseur de données **(I.b)**.

(II.) L'adresse est géocodée **(II.a)**, puis elle est reliée à un bâtiment **(II.b)**. Cette étape induit une perte d'information. En effet, la géolocalisation de l'adresse peut être fausse ou incorrecte à cause de la base adresses ou de problématiques de géolocalisation (voir la section 3.2). Puis, de l'incertitude apparait lors de l'étape **(II.b)** car la jonction bâtiment-adresses n'est ni surjective ni injective.

(III.) Deux bases de données sont ensuite calculées : une base avec l'information de géolocalisation pour chaque adresse $\mathcal{A}^{Géocodées}$ et une base de qualité sur la géolocalisation $Q^{Géocodées}$ qui contient des indices de qualité de données pour chaque bâtiment et pour chaque variable (**(II.a)** and **(II.b)**).

(IV.) À partir du référentiel bâtiment, toutes les informations sont compilées en données tabulaires pour chaque variable disponibles $\mathcal{E}_{xternes\ données}$ **(IV.a)** et pour les données spécifiques **(IV.b)** - données météorologiques et iris. Ces dernières ne nécessitent pas une géolocalisation précise du bâtiment. De plus, une base de données avec tous les indices de qualité de données $Q^{xternes\ données}$ est disponible **(IV.c)**.

(V.) et (VI.) Une base de données avec l'ensemble des caractéristiques des bâtiments $\mathcal{D}^{Observés}$ est calculée en amont (Valeurs manquantes, meilleures observations ...). Chaque modèle et données utilisées en entrée créent de l'incertitude qui est évaluée à l'aide des autres observations. Ainsi, pour chaque variable, des indices de qualités individualisées sont fournis dans $Q^{\mathcal{D}}$ (VI.).

(F.) Finalement, le fournisseur de données fournit ($\mathcal{D}^{Observés}$;$Q^{\mathcal{D}}$) et $Q^{Géocodées}$. L'assureur transmet la base de contrats $\mathcal{D}^{Traditionnelles}$ de souscriptions et $\mathcal{S}^{Sinistres}$ (2016 – 2018) la base historique de sinistres de 2016 à 2018.

Le fournisseur n'a pas accès aux informations du contrat qui pourraient lui permettre d'améliorer la géolocalisation.

La $\mathcal{D}_{global}$ obtenue est la jointure entre les quatre bases de données.

3.1.2.a Les bases de données de l'assureur

Base de sinistralités : Cette base contient uniquement la sinistralité attritionnelle en MRH habitations individuelles pour toutes les garanties hors CatNat et TGN. La définition des graves a été faite par l'assureur avec ses seuils historiques. Les garanties considérées sont : DDE, INC, VOL, ELEC, BDG et RC. D'autres garanties étaient aussi disponibles mais avec de trop faibles volumes. Les sinistres sont vus à l'ultime.

Base assureur dit traditionnelle : Cette base correspond aux informations fournies durant la souscription et la vie du contrat. Le chapitre 1.1 fournit un détail sur les informations disponibles. Il est néanmoins important de noter que beaucoup de questions précises sont posées. L'assureur associé au projet a un questionnaire étoffé, en particulier sur le bâtiment : la surface des panneaux solaires, la surface des annexes et le nombre de pièces. Les données sont de très bonnes qualités.

3.1.2.b Les données nécessitant la géolocalisation de l'adresse

Ce sont les informations mentionnées dans le chapitre 2. Cette étude étant la première, les évolutions et améliorations de la géolocalisation et des variables ont été importantes. Les résultats finaux sont basés sur une géolocalisation qui relie l'adresse à un complexe de bâtiments puis à un bâtiment. Au début, deux géocodeurs co-existaient, l'un utilisant la parcelle comme étape intermédiaire, l'autre, vraisemblablement, ressemblaient à celui actuel. Il y a eu aussi des améliorations entre le choix des bâtiments à partir des adresses. Sur les variables, les données météorologiques à l'année n'étaient pas disponibles, ni même les indices de sécheresse. Et de nombreuses évolutions sont apparues pour certaines variables comme la période de constructions, le type de chauffage, la surface habitable, la valeur de la maison et les panneaux solaires. Une variable intéressante existait mais n'est maintenant plus disponible : la présence de velux.

3.1.2.c Les données dit Open Data (OD)

Il est intéressant de comparer le modèle traditionnel (M1) et celui amélioré par les données Open Data - (M4). L'idée est de comprendre si sans la géolocalisation ni trop d'investissements, il est possible d'améliorer les modèles de façon significatives. Ici, l'OD est définie comme des données disponibles en deux, trois clics, ne nécessitant ni plusieurs bases de données, ni une infrastructure dédiée. En résumé, elles vérifient les trois règles suivantes :

— La variable ne doit pas nécessiter la géolocalisation de l'adresse ;
— La variable doit être disponible sous format .csv/.json avec presque aucune modification ;
— La variable doit être disponible sur tout le territoire français.

Environ 40 variables sont concernées comme le type d'environnement (*Corine land cover*) et des informations à l'iris comme le nombre de cadres ou le nombre de magasins... De manière générale, ce sont des variables de très bonnes qualités, robustes provenant de sources gouvernementales. Elles sont d'ailleurs souvent utilisées dans les zoniers.

3.2 La problématique de la géolocalisation et le choix de la base de modélisation

L'emploi des modèles définit la base de modélisations à considérer. La création de la base suit un schéma classique - voir la partie 1.1.2. Les sinistres sont subdivisés par garanties et en attritionnels, graves et climatiques. Cependant, la géolocalisation influence le choix des bases.

Avec une base de données longue comme large, la sélection des variables dépend de la base sélectionnée. De façon similaire, le meilleur sous-ensemble de la base représentant au mieux les risques étudiés est recherché. La première partie 3.2.1 résume dans un premier temps les résultats de géolocalisation en comparant les changements dans le processus de souscription. Dans un second temps, la partie 3.2.2 décrit les filtres d'intégrités. Enfin, la partie 3.2.3 détaille le principe des filtres de modélisation pour choisir le sous-ensemble représentatif de la base des données produites lors de l'industrialisation de la souscription. Le nettoyage de la base sinistre ne sera pas discuté.

3.2.1 La géolocalisation et processus de souscriptions

À la date de juin 2020, 24 millions d'adresses avaient été géocodées. La plupart d'entre elles provient de la **BAN**. Pour le référentiel bâtiments, 35 millions ont été géocodés avec 84.1% de résidentiel dont 80.1% de bâtiments individuels. 58.7% d'entre eux sont considérés comme bâtiment principal. De nombreux indicateurs sont disponibles pour évaluer la crédibilité du géocoding. Pour ce projet, les estimations de la précision du géocoding sont disponibles dans le tableau 3.1. Ces estimations ne considèrent pas la probabilité qu'il y ait plusieurs bâtiments assurés et sont basées sur un contrôle manuel. Le chapitre 2 sur la géolocalisation explique plus en détail les différentes problématiques.

Qualité	Very High	High	Medium	Low	Very Low
Répartition	68 %	3.3%	18.1%	9.2%	1.4%
Précision	99%	95 - 90 %	$\approx$ 70%	$\approx$ 50%	-

TABLE 3.1 – L'évaluation de la qualité s'est faite en utilisant un échantillon stratifié comparant le géocoding et le véritable emplacement en utilisant Google Maps.

3.2.1.a Processus de souscripteur

Le schéma 3.3 montre les grandes lignes du processus de souscriptions.

Pour respecter le **RGPD**, le fournisseur de données ne reçoit pas d'informations personnelles : l'adresse seule n'est pas une information personnelle. En effet, seul l'assureur pour son modèle de tarification reçoit les réponses **(IV.a)**. Il est aussi le seul à savoir si la transformation du prospect en assuré a été réussie ou pas **(VI.a)**.

Contrairement à la géolocalisation du portefeuille, le prospect peut vérifier et modifier le choix du ou des bâtiments. Cette correction **(V)** amène le fournisseur de données à renvoyer les bonnes informations et diminue les erreurs de géolocalisations. Ainsi, le processus de souscription ne produit pas d'erreurs de géolocalisation. En effet, le processus suppose que la connaissance du prospect est quasiment parfaite.

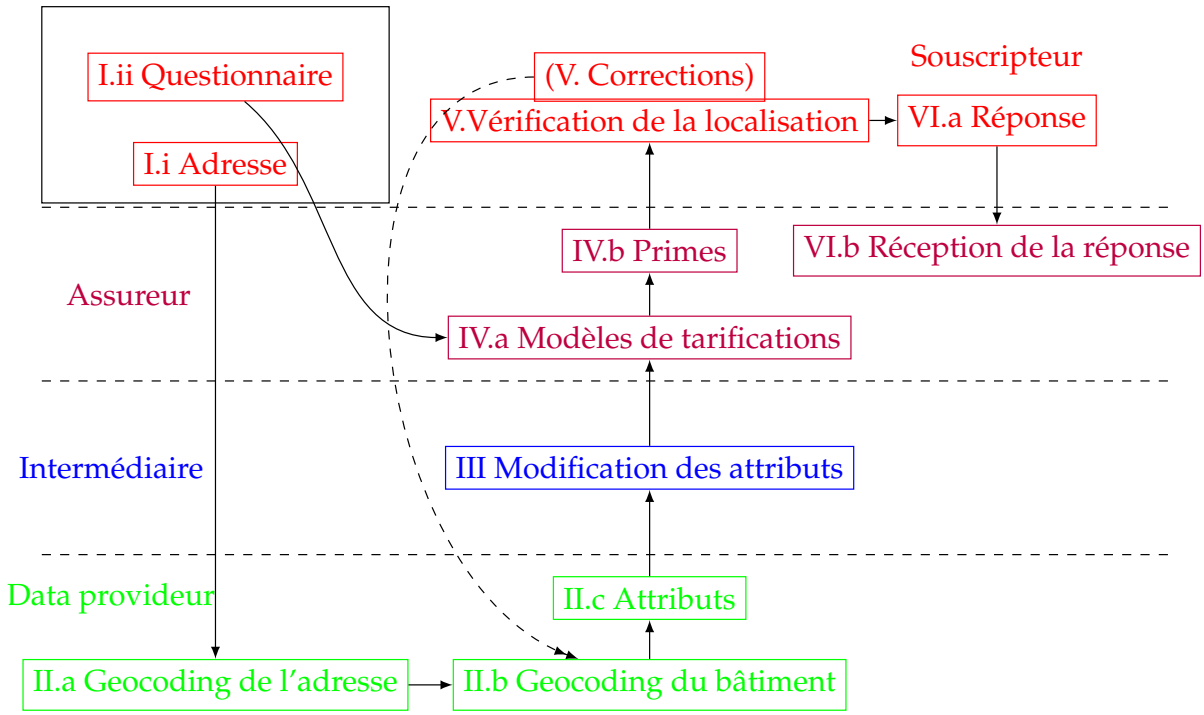


FIGURE 3.3 – Exemple de processus de souscription. La proposition de la prime peut se faire après la vérification de la géolocalisation. L'intermédiaire, un agent, un outil ou une **API** adapte les attributs à un usage tarifaire.

◆**Les contraintes de souscriptions :**

Huit contraintes de souscriptions avaient été déterminées dans le chapitre 1. Les contraintes sur la constance annuelle de la prime, qu'elle doit être bornée inférieurement, vérifier les contraintes législatives, et celle de linéarité de l'exposition (respectivement les contraintes 1, 2, 7 et 8) ne sont pas impactées par le changement d'approche pour la souscription.

La contrainte de la continuité et discontinuité du questionnaire (contrainte 6) est moins contraignante pour ce nouveau processus de souscriptions. En effet, les critères tarifaires comme les variables externes sont accessibles en dehors de la relation B&C. Ainsi la continuité du questionnaire s'applique sur un nombre moins important de variables. Dans le cadre du modèle (M3), les questions sur l'âge et la formule choisie seront toujours demandées. Dans ce cas, il n'y a plus cette contrainte à vérifier.

La contrainte de proportionnalité (contrainte 5) est toujours valide en partie. Si le prospect ne connait pas nécessairement les variables externes utilisées, celui-ci pourrait plus aisément examiner les primes proposées pour les bâtiments voisins. Dans le cadre des modèles (M3), comme aucune question sur le bâtiment n'est posée, le prospect peut comparer les primes proposées juste en cliquant sur le bâtit voisin. Ainsi, si la proportionnalité des variables externes complexes ou inconnues pour le prospect est moins importante, celle de la composante géographique risque de prendre plus d'importance.

La contrainte d'adaptabilité du tarif (contrainte 8) doit toujours être vérifiée. En effet, des effets d'aléas moraux ou d'anti-sélection peuvent apparaître lors de l'ajout de variables externes. Même si les variables pouvant être associées à des biais de type I ou II (partie 1.1.3) sont en plus faible nombre, la mutualisation des primes est toujours valide. L'utilisation de GLMs est donc nécessaire.

La contrainte d'une adaptation des informations tarifaires (contrainte 3) est complètement altérée. La complexité des variables n'est plus assujettie à la connaissance du prospect. Cependant, il est nécessaire que l'accessibilité des données externes par la calculette tarifaire soit quasi instantanée et à tout instant. C'est pourquoi les variables sont calculées en amont et le taux de corrections des variables se doit d'être faible.

3.2.2 L'intégrité de la base de données

Comme les modèles doivent être calibrés sur des données les plus semblables à celles utilisées en pratique, les modèles de tarification (IV.a) se fittent sur des données avec le moins possible d'erreurs de

géolocalisation. Mais la qualité des variables, elle, n'est pas impactée par les erreurs de géolocalisations durant le processus de souscription 3.3, il n'y a alors pas lieu d'adapter la base à la qualité des variables. La méthodologie procède en deux phases. La première correspond à l'intégrité de la base de données - cette phase retire tous les bâtiments non résidentiels ou les appartements. La seconde correspond aux filtres de modélisations.

Durant la phase de géocoding, des erreurs apparaissent. Certaines adresses peuvent être reliées à des bâtiments non résidentiels (serres, entrepôts ...) ou des appartements (voir chapitre 2). Dans cette base, nous avions même une gare SNCF qui avait été choisie. Cette dernière avait été bien géocodée, mais il existait des logements incorporés dans cette gare. Ces erreurs ou particularités ne doivent pas être incluses dans la base d'étude.

Les contraintes que doit vérifier chaque ligne de notre base, sont définies pour vérifier l'intégrité de la base. Ramakrishnan 2003, [91] définit la notion d'intégrité à l'aide d'un ensemble de contraintes. Pour cette thèse, les notations seront allégées.

Définition 1

Un ensemble de contraintes TC est une application telle que :

$$\begin{aligned} TC : &\mathcal{D} \longrightarrow \mathcal{D} \\ &D_x \mapsto D_y \end{aligned} \tag{3.1}$$

où $\mathcal{D}$ est l'ensemble des bases de données, D_x et D_y sont des éléments $\mathcal{D}$ tels que $D_y \subseteq D_x$, en d'autres termes que D_y est une sous-base de D_x.

Définition 2

Une base de données **D** est cohérente à un filtre TI si $TI(\mathbf{D}) = \mathbf{D}$. On dit que TI est vérifié par **D**.

3.2.2.a Les contraintes d'intégrités sur le type de bâtiment

La contrainte d'intégrité à vérifier est que tous les bâtiments soient des bâtiments individuels et résidentiels.

Pour cet objectif, le premier ensemble de contraintes concerne l'adresse ; par exemple : le complément d'adresse ne doit pas contenir les mots suivants "*Appartement*", "*APP*", "*APPT*" ou "*Immeuble*". Ce sont des erreurs provenant des contrats de l'assureur.

Ensuite, les contraintes suivantes sont choisies en regardant la proportion d'adresses non conformes. Quatre autres contraintes sont appliquées pour améliorer l'intégrité de la base :

- **Surface habitable non manquante NA** : Quand la valeur n'est pas donnée, cela correspond à un bâtiment non résidentiel selon le fournisseur de données.
- **0 <Nombre de pièces < 20** : Un nombre de pièces égale à 0 ou supérieur à 20 réfère à un bâtiment particulier, complexe ou mal défini (Manoir, hôtel particulier, maisons de retraite ...).
- **Le nombre de codes NAF < 5** : Quand le nombre de codes **NAF** lié à l'adresse [2] est important, très souvent cela correspond à des bâtiments associés à des entreprises, des immeubles ou des bâtiments publiques.
- **Surface au sol < 2000** et **surface de parcelle < 25 000** : Ces contraintes suppriment les bâtiments trop vastes pour être des maisons. Dans ce processus, certaines maisons sont supprimées, mais le gain apporté par le filtre compense largement les pertes de lignes. Traditionnellement, ces maisons font l'objet d'une tarification différente.
- **Nombre d'étages ≤ 4** : Quasiment l'ensemble des bâtiments concernés étaient des appartements.

Contraintes d'intégrités sur les erreurs de découpage : Certaines des contraintes précédentes permettent aussi de supprimer les erreurs de découpages des bâtiments. En effet, certains bâtiments adjacents sont considérés comme un unique bâtiment. Pour ces bâtiments, les variables comme la surface habitable sont fortement impactées.

Contraintes sur les attributs : Le dernier ensemble de contraintes porte sur la complétude et l'ensemble de définition des variables. Quasiment tous les attributs vérifient ces contraintes du premier

2. Le code **NAF** correspond aux activités déclarées reliées à l'adresse.

abord. En effet, le fournisseur de données a déjà corrigé les erreurs et complété les variables. Néanmoins, certaines variables ne sont, par définition, pas complétées. Par exemple : la notion de parcelles n'existent pas, n'est pas disponible ou est mal définie dans certains départements.

3.2.2.b L'impact des contraintes d'intégrité

Les filtres d'intégrité *TI* ont plusieurs impacts non négligeables. La volumétrie de la base de données est diminuée d'environ 5%. Certaines contraintes ont un impact spatial : certaines municipalités, la plupart du temps rurales, ne sont plus représentées dans la base de modélisation. À cause des filtres, des biais peuvent apparaitre et réduire la robustesse des modèles. La définition des filtres est lacunaire car il est impossible d'ajouter certains filtres. Par exemple le filtre sur le nombre d'étages ne peut être plus restrictif sinon de nombreuses maisons en Île-de-France seraient supprimées. Ainsi, même si les filtres d'intégrité sont vérifiés, il reste des bâtiments non résidentiels.

En outre, les bâtiments choisis ne sont pas toujours les bons (Voir le chapitre 2). Une partie importante des erreurs de géocodages sont traitées par ces filtres d'intégrité, mais il est nécessaire d'ajouter des filtres de modélisations pour réduire les erreurs de géolocalisation.

3.2.3 La base de modélisation

Comme les filtres d'intégrité n'ont pas pour objectif de réduire les erreurs de géolocalisations, des filtres dits "de modélisations" sont appliqués.

3.2.3.a L'impact des erreurs de géolocalisation

La géolocalisation des bâtiments est imparfaite pour plusieurs raisons. L'une d'entre elles est que l'adresse n'est pas assez précise pour choisir le(s) bâtiment(s) assurés (voir chapitre 2). En 2020, j'estimais un taux d'erreur de l'ordre de 10 à 15% [3].

Rappelons que ces erreurs n'impactent pas toutes les variables de la même façon. Par exemple : les variables à l'iris ne sont pas impactées si le bâtiment choisi est celui du voisin. Les variables avec une dépendance spatiale forte sont moins impactées par exemple les types de toits, le nombre d'étages ou bien la taille des maisons. Au contraire, les variables comme la surface habitable, la présence de panneaux solaires, sont fortement impactées.

Dans le cadre de les **LM**s ou les **GLM**s, les chapitres 4, 5 et 6 contribuent théoriquement à la problématique de la crédibilité de la donnée. À l'aide de cette théorie, j'ai pu évaluer et détecter les améliorations essentielles pour la continuité du projet. En effet, avec un grand nombre de variables, avec la qualité de chacune et du géocodage, il est difficile d'expliquer pourquoi les coefficients évoluent de telle façon suite à un changement de base (nouvelles variables, géocodeur mais avec une même sinistralité et adresses).

En résumé, le taux de géolocalisations fait diminuer proportionnellement les coefficients des autres variables quand il n'y a pas trop de corrélations inter-variables. Pour une variable avec un coefficient estimé sur des données parfaitement géolocalisées, si le coefficient initial est égal à 1 et si on ré-estime le même coefficient sur une base dans laquelle l'on se trompe une fois sur dix de valeurs, le coefficient estimé sera en moyenne égale à 0.9 (sans corrélation). Aussi, le modèle aura une performance bien moindre que le véritable modèle. Par conséquent, il est nécessaire de travailler avec une base avec la géolocalisation la plus précise possible. La partie 6.3.1 examine l'influence de l'évolution des géocodeurs sur les évolutions des coefficients.

3.2.3.b Les contraintes de modélisation

Notons *TM* l'ensemble des contraintes de modélisations. Dans notre cas, ces contraintes ne sont basées que sur les indices de qualité mesurant la qualité du géocodage, *c-à-d*

- La qualité entre le lien de l'adresse fournie et de l'adresse géolocalisée ;
- La qualité entre le lien du bâtiment choisi et de l'adresse géolocalisée ;
- Si le bâtiment choisi est considéré comme bâtiment principal ;
- Le nombre de bâtiments reliés à l'adresse géolocalisée ;
- La distance entre le point adresse géolocalisée et le bâtiment ;
- Le nombre d'adresses géolocalisées et reliées au bâtiment choisi.

3. qui me semble présentement sous-estimé.

Plus les filtres sont restrictifs, moins il y a d'erreurs de géolocalisations. Néanmoins, ces filtres impactent très fortement la volumétrie des bases de données. Si pour certaines garanties (ex : Fréquence DDE), la volumétrie des données est importante, pour d'autres (ex : Coût Moyen en RC), les contraintes ne peuvent pas être trop restrictives. D'ailleurs, pour la RC, quasiment aucune variable sur le bâtiment n'est utilisée. Si les filtres de modélisations doivent être les mêmes pour les bases de modélisations de *Freq* et de *CM* pour une même garantie, elles doivent être différentes et plus ou moins restrictives en fonction des garanties.

L'idée est donc de rechercher l'ensemble optimal de contraintes pour chaque garantie.

Définition 3

Un ensemble de contrainte TM est dit optimal pour une métrique de performance $Cost$ et un modèle $m_{\mathcal{X}_{\text{train}}}$ entrainé sur une base de donnée $\mathcal{X}_{\textbf{train}}$ évalué sur une base de test $\mathcal{X}_{\textbf{test}}$ si TM est la solution du programme de minimisation suivant :

$$\underset{TM}{Argmin}\ \ Cost(\ m_{TM(\mathcal{X}_{\textbf{train}})},\ TM(\mathcal{X}_{\textbf{test}})). \tag{3.2}$$

Une solution optimale est notée $TM_{opti}(Cost, m_{\mathcal{X}_{\text{train}}}, \mathcal{X}_{\textbf{test}})$.

3.2.3.c Le processus final d'optimisation

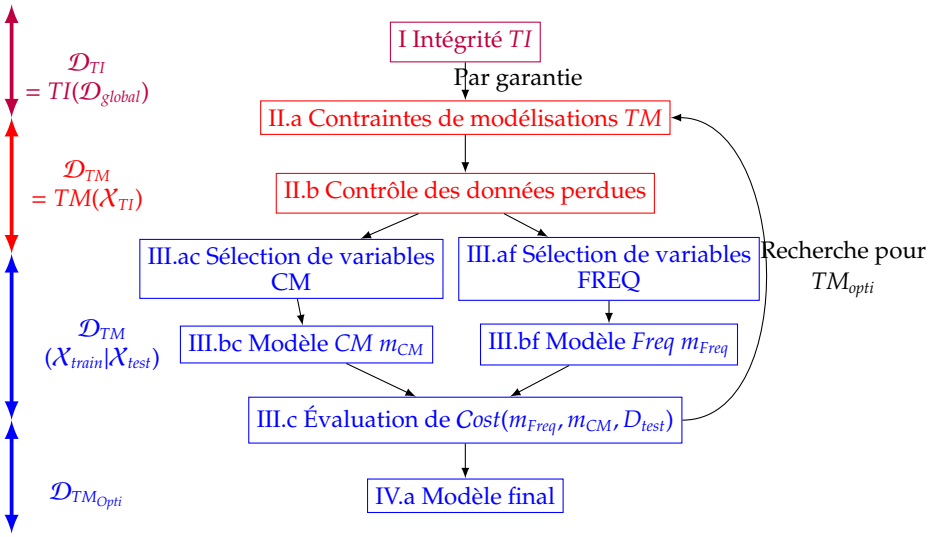


FIGURE 3.4 – Processus de modélisation pour une base $\mathcal{D}$.

La procédure suivie - graphique 3.4 - est simple à mettre en oeuvre, mais longue en pratique. La phase d'intégrité **(I)** n'est faite qu'une seule fois. Mais le choix des contraintes de modélisations TM est fait plusieurs fois. Pour chaque risque à modéliser, la phase **(II.a)**, choix de TM, doit être faite plusieurs fois car les phases **(II.af)** et **(III.ac)** changent de façon significative en fonction des filtres. TM_{opti} est choisi après plusieurs itérations. Pour la consolidation, les contraintes TM_{opti} les plus restrictives sont utilisées.

Un contrôle sur les lignes perdues **(II.b)** est fait (exemple 3.5). Dans certains cas, le risque est complètement biaisé. Pour vérifier le biais du risque, les coefficients du modèle (M1) sont utilisés. En effet, le modèle (M1) n'est pas impacté par la géolocalisation. La métrique de performance, $Cost$ sert aussi de métrique de contrôles. Ici, la déviance moyenne de Tweedie avec $p = 1.5$ a été choisi. La comparaison s'effectue entre $\mathcal{D}_{TI}$ et $\mathcal{D}_{TM}$. De grandes différences entre $\mathcal{D}_{TM}$ et $\mathcal{D}_{TI}$ reflètent les biais. De manière générale, les filtres utilisés biaisent les risques dans les zones rurales essentiellement.

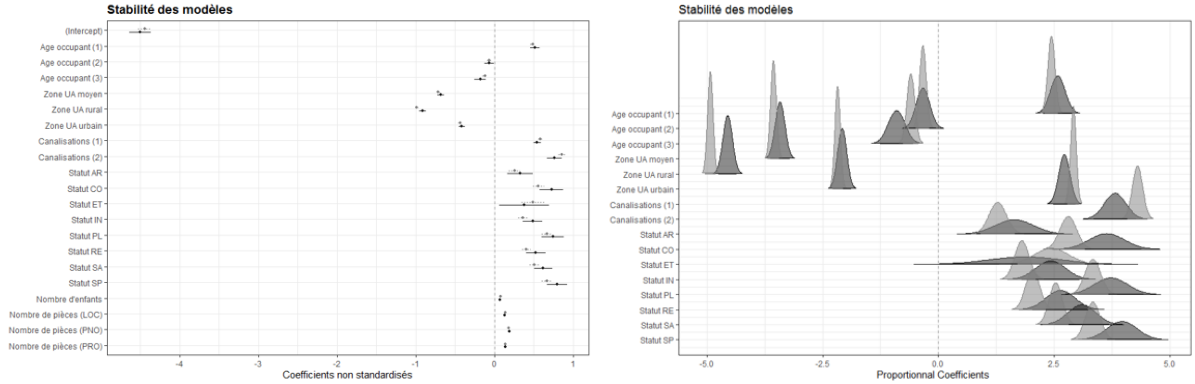


FIGURE 3.5 – Comparaison du modèle de fréquence DDE entre la base complète et la base géolocalisée $D_{TM_{Opti}}$ ($3^{\text{ème}}$ ou $4^{\text{ème}}$ base de données). Ici, seul le coefficient de ruralité est significativement impacté.

3.3 La sélection des variables et des modèles

La sélection des variables est plus complexe que pour la base de données traditionnelle des assureurs. La première difficulté est d'utiliser de nouvelles variables encore inconnues. La seconde est liée au nombre de variables induisant des problèmes informatiques de temps de calculs et de mémoires. La troisième problématique provient du fait que ces variables n'ont pas été crées initialement pour un objectif de tarification. Les indices de qualité de données peuvent ne pas être adaptés (comme trop restrictifs pour les variables météorologiques) ou les intervalles des variables ne sont pas optimisées. Par exemple, pour le nombre de pièces, les assureurs définissent des règles métiers et comptent les grandes pièces comme deux ou trois pièces. La surface habitable calculée est plus proche d'un calcul des surfaces chauffées comprenant les garages. De plus, le processus de création de chaque variable est très différent avec des niveaux de précisions hétérogènes.

Beaucoup de ces informations peuvent être redondantes, de faibles qualités et imprécises. Néanmoins, l'objectif de cette sélection de variables permet de choisir les variables à améliorer et qui apportent une information encore inconnue sur le risque. La correction des valeurs est difficile car le fléau de la dimension, ((Indyk et Motwani, 1998) [85] ou (Friedman 1997) [78]) rend complexe la détection des valeurs aberrantes dans un premier temps.

3.3.1 Le processus de sélection de variables

Au commencement du projet, environ 300 variables étaient disponibles. La première étape préliminaire est de supprimer les variables qui ne peuvent pas être utilisées sur l'ensemble du périmètre. Une qualité trop faible ou des informations pas assez différenciantes provoque la suppression de la variable. Ensuite, le processus détaillé dans le schéma 3.6 est appliqué et débute par une sélection non supervisée. Les objectifs de cette **phase 1** sont :

- ➢ D'analyser les informations apportées par les données externes ;
- ➢ Comprendre les dépendances entre les variables ;
- ➢ Éliminer les redondances d'informations ;
- ➢ Réduire le nombre de variables.

L'élimination de la redondance d'information permet à tous les modèles d'être plus interprétables. De nombreuses méthodes d'interprétation nécessitent de minimiser cette redondance par corrélation univariée mais aussi multivariée. Par exemple, avec trois variables météorologiques, il est aisé de retrouver le département et même la commune. Ainsi, moins de trois variables météorologiques sont conservées.

Dans cette étude, une analyse de corrélation est faite et combinée avec une analyse en composante principale pour comprendre les effets multivariés-linéaires. Ensuite, une sélection supervisée avec un **RF** (Breiman, 2001 [71]) combiné avec une analyse de GLM Lasso et Ridge est appliquée lors de la **Phase 2**. En utilisant des indicateurs d'importance de variables et de *partiales dépendances plots* (Friedman, 2001 [79]) appliqués a un **RF**, les variables choisies peuvent être sélectionnées et préparées pour les GLMs. L'objectif principal est de minimiser le temps de modélisation, car le processus est réitéré plusieurs fois.

Ce processus est long et fastidieux. Par la suite, il a été très largement simplifié. Néanmoins, au vu de la qualité initiale des variables, il est très robuste et a permis une compréhension vaste et minutieuse

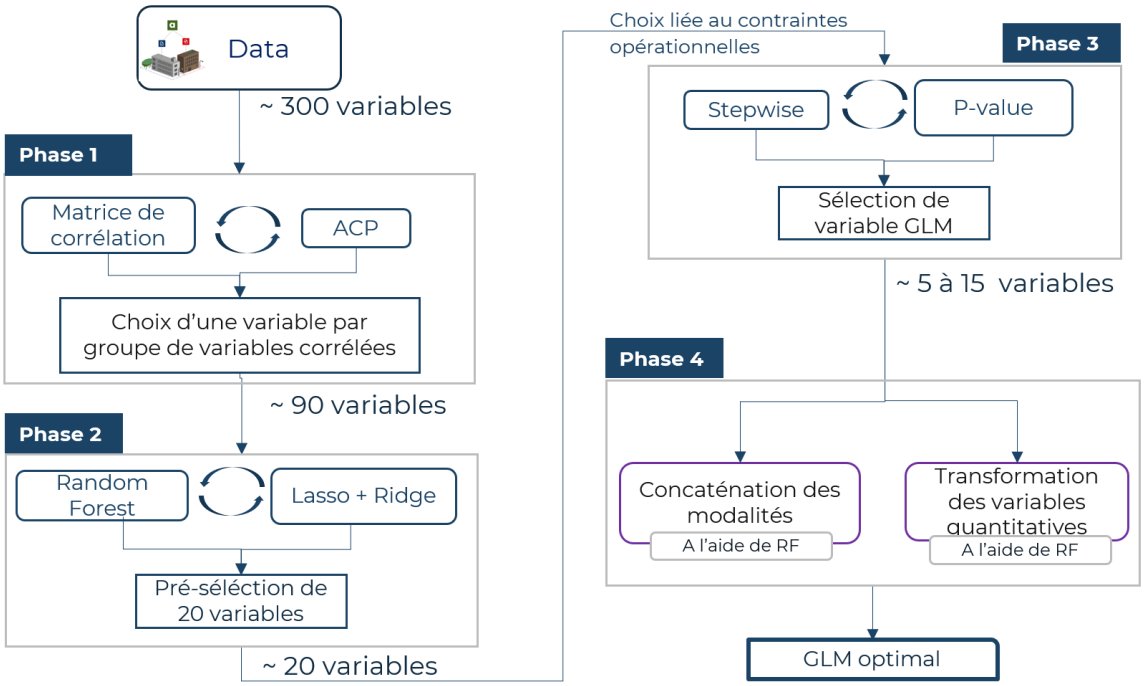


FIGURE 3.6 – Le processus de sélection avec un grand nombre de variables de qualités hétérogènes.

de la donnée. Les GLM Lasso et Ridge ont permis de capturer les variables à effets linéaires faibles, mais non captée par les autres variables, alors que le RF, coûteux en temps, dissimule certaines variables par les interactions apprises.

3.3.2 La sélection non supervisée

La première sélection de variables est une filtration. La **Phase 1** utilise l'analyse des corrélations par les graphes et les composantes principales de la base. La principale difficulté est le nombre de variables.

3.3.2.a L'analyse de la matrice de corrélation

La base de données contient des variables continues et catégorielles. Très peu de variables ordonnées et composites sont présentes : les **DPE**, la période de construction et des variables à l'iris. Celles-ci seront traitées avec précautions et toujours testées dans nos modèles finaux. Comme la base de données mélange les types de variables, une matrice de "corrélation" est calculée pour les variables continues, booléennes et catégorielles [4]. Même si les coefficients ne sont pas strictement comparables entre les coefficients de corrélation de Pearson et les coefficients polychoriques, l'objectif est d'avoir une vision macro des dépendances entre les variables [5]. Les variables avec plus de 8 modalités, sont considérées comme continues (Seulement 2 variables sont concernées). Cependant, la matrice de corrélation n'est pas lisible, car trop de coefficients sont affichés comme le présente la figure 3.7.

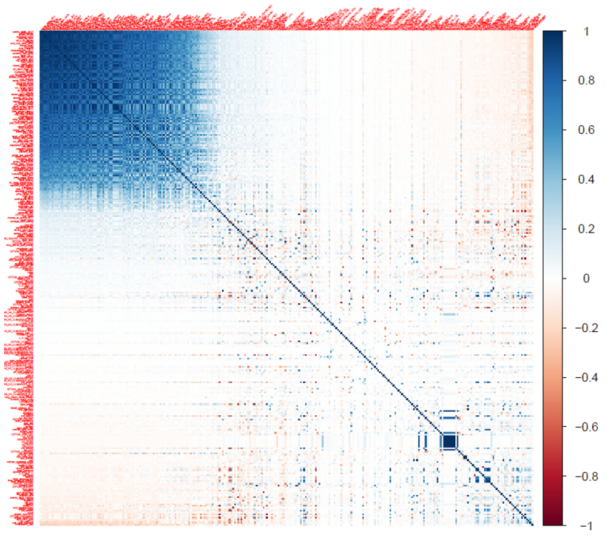


FIGURE 3.7 – Une partie de la matrice de corrélation calculée avec *mixed.Cor* du package *psych*.

Pour obtenir des informations interprétables, la théorie des graphes permet d'obtenir une visualisation par réseau comme le montre 3.8. Ce graphe se fonde sur une matrice d'adjacence à partir de la matrice de corrélation :

$$Coeff_{i;j}^{Adjacent} = \mathbf{1}_{|cor_{i;j}|>s} cor_{i;j} \tag{3.3}$$

avec X_i,X_j deux variables et un seuil $s \in]0,1[$ appliqué au coefficient de corrélation $cor_{i;j}$ entre les deux variables.

3.3.3 L'étude du graphe de corrélation

L'algorithme utilisé est celui de (Fruchterman, 1991 [81]). Le seuil s permet de contrôler l'éclatement du réseau pour une meilleure lisibilité.

La figure 3.8 a été déterminée sur la base complète (DDE) avec un seuil $s = 0.4$. La nature des informations apportées par les données externes est bien identifiée. Une dissociation importante entre les données de souscription et les données externes est visible. En effet, ces dernières n'apportent aucune information personnelle comme sur l'âge du souscripteur ou sa catégorie socio-professionnelle. Les liens apparaissant entre les données assureurs et externes sont l'*option piscine* et la *présence d'une piscine* et les *panneaux solaires*.

4. à l'aide du package *R psych* [92].
5. Il est à mentionner qu'au début des échanges avec le fournisseur de données, les variables utilisées pour telles ou telles variables étaient inconnues pour des raisons politiques.

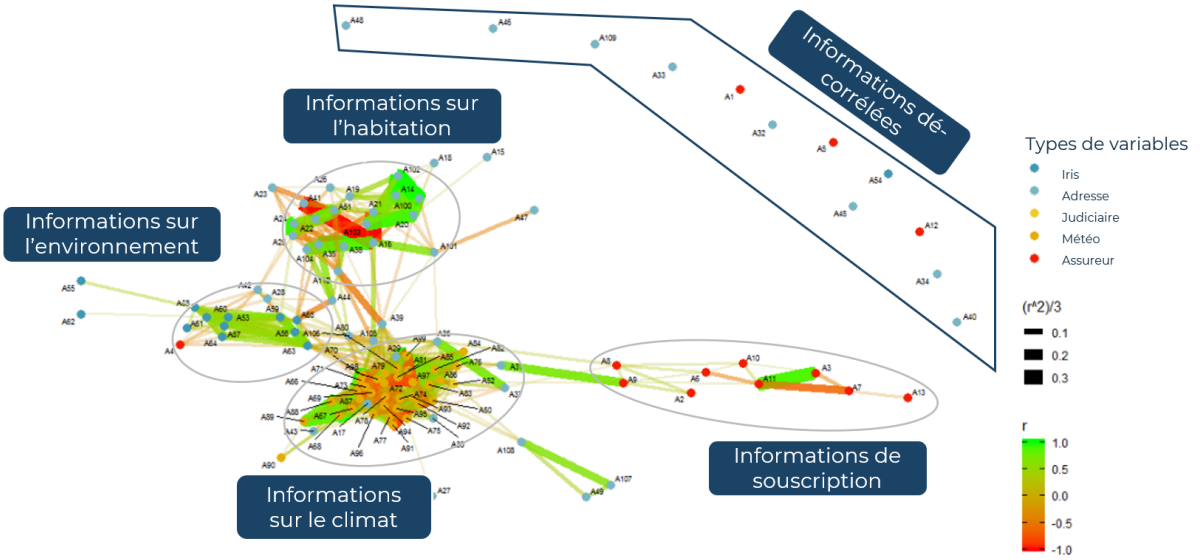


FIGURE 3.8 – Exemple d'éclatement de la matrice de corrélation avec $s = 0.4$. (Pas de filtres de modélisations, $2^{éme}$ base du projet)

De plus, les données externes apportent trois types d'informations principales : les informations météorologiques, sur l'environnement du bâtiment et sur ses caractéristiques. Il existe des variables non fortement corrélées comme le type de chauffage que l'on veut conserver. Cette figure 3.8 peut être appliquée à chaque cluster pour mieux les comprendre. Sur les données météorologiques, un cluster sur le vent, la température ou les valeurs extrêmes apparaissent. Ce graphique est très dépendant du seuil, mais le choix initial des points est peu influent. Ici, la corrélation nombre de pièces et la surface habitable n'apparait pas. Comme le temps de calcul de la matrice n'est pas négligeable, on applique aussi une analyse en composantes principales pour éliminer les corrélations linéaires multivariées. L'objectif est d'avoir le moins de liens pour un seuil donné. Pour cela, cette phase est réappliquée autant de fois que nécessaire pour conserver au maximum 90 variables.

Remarque La qualité des variables et le géocoding influencent fortement le graphique.

3.3.3.a Les analyses des variables disponibles

Pour gérer les problèmes de corrélation et de qualité, l'analyse peut-être faite sur des bases mieux géocodées [6]. La faible qualité influence grandement les variables comme la surface habitable qui, elle, est essentielle. À partir du graphique 3.9, les sous-ensembles du réseau permettent de mieux comprendre les données - figures 3.10, 3.11, 3.12, 3.13 et 3.14.

Après un premier filtre, le graphe 3.9 peut être affiché avec moins de variables pour un meilleur éclatement. Les 4 clusters de variables sont toujours présents, mais les dépendances inter et intra-groupes sont plus visibles. En haut à droite - figure 3.10 - les informations à l'iris sont affichées. Les variables en haut correspondent à des variables indépendantes des autres - figure 3.11. À l'intérieur du graphe principal, les trois zones précédemment mentionnées : météorologiques - figure 3.12, sur le contrat et la personne - figure 3.13 - et des informations sur le bâtiment - figure 3.14.

6. Ce graphique ressemble grandement au graphique sur la base entière des dernières versions pour d'autres assureurs.

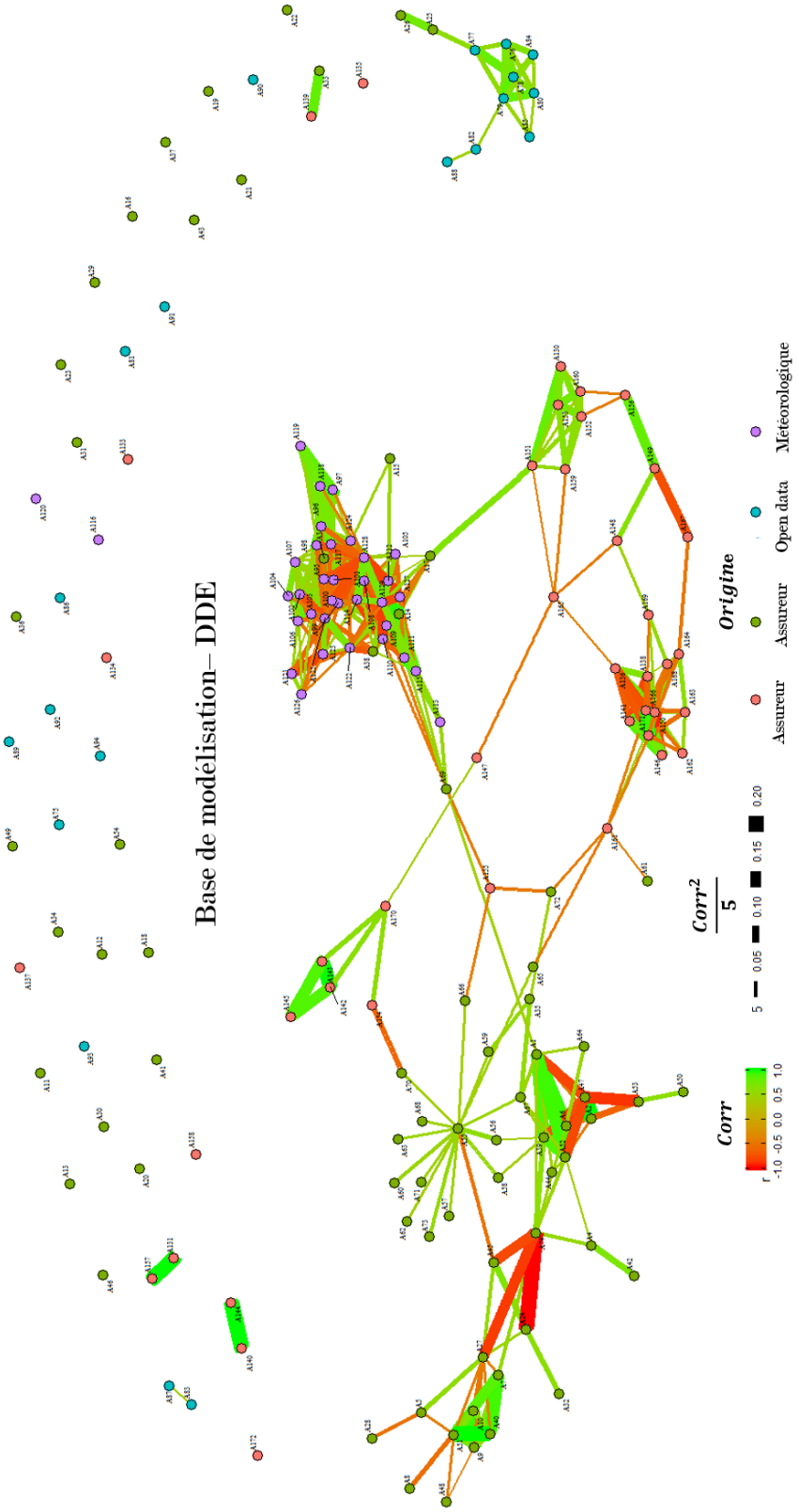


FIGURE 3.9 – Graphique de corrélation pour la base DDE (4ème base de modélisation).

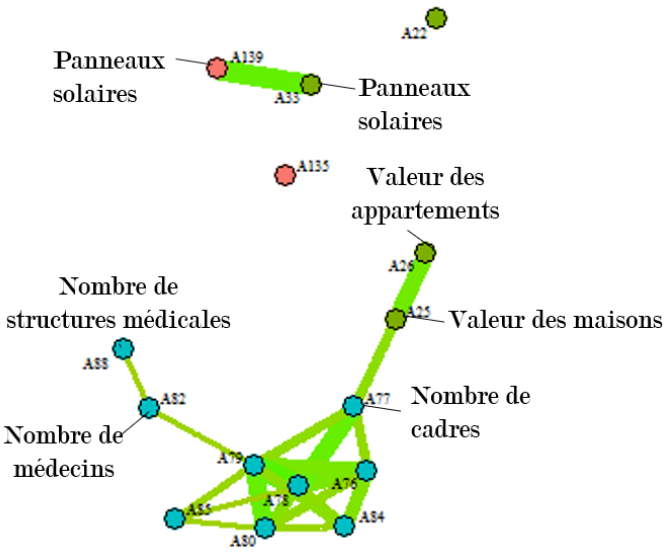


FIGURE 3.10 – Graphique de corrélation pour la base DDE. Focus sur les informations à l'iris.

Analyse de la figure 3.10 La déclaration de panneaux solaires et la détection d'un panneau solaire par analyse d'images sont très fortement corrélées. Des variables composites sont aussi très corrélées à la densité : le nombre de cadres actifs, d'ouvriers par iris, le nombre de médecins... Il y a des corrélations économiques par exemple spatiales : la valeur des maisons et le nombre de cadres d'une commune. Pour éviter de telles redondances, il est nécessaire de choisir l'une d'entre elles ou de créer de nouvelles variables comme le rapport entre ouvriers et cadres. Dans une logique de coût de maintenance des variables, la première option est privilégiée.

Analyse de la figure 3.11 Dans d'autres cas, certaines variables ne sont pas corrélées (ou peu). On y retrouve des variables de zoniers, une dépendance entre le capital assuré et l'option de protection des objets précieux. Néanmoins, les variables ne sont pas réellement indépendantes. Les variables de distance comme la distance à un cours d'eau, à une **ICPE** ou de zoniers sont dépendants, mais pas linéairement. C'est pourquoi une ACP a été ajoutée. De plus, une variable aléatoire et indépendante a été ajoutée à l'ensemble des variables. Elle permet de contrôler la convergence des modèles ML par la suite.

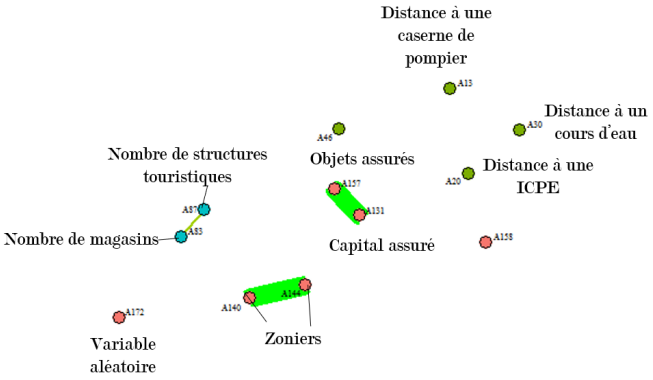


FIGURE 3.11 – Variables décorrélées apportant des informations décorrélées des autres variables

Analyse de la figure 3.12 Les informations sur le contrat et la personne assurée proviennent majoritairement de la souscription. Deux clusters se distinguent :

— Information sur le souscripteur : son âge, le nombre d'enfants, sa profession...

— Les options spécifiques : Option piscine, équestre...

Des informations sur le bâtiment sont distants des deux groupes comme le nombre de pièces. Un petit clin d'oeil est fait entre la corrélation négative entre l'option dommage électrique et le fait de déclarer une activité professionnelle en lien avec des travaux électriques. La corrélation entre les variables de piscine est aussi visible. Leur cas met en exergue les difficultés de l'exercice. Elles ne sont pas corrélées avec les mêmes informations. En effet, la variable externe à cause de sa provenance et de la qualité du géocodage est plus décorrélée avec d'autres variables. De plus, elle n'a aucun lien avec le contrat (on peut avoir une piscine sans prendre d'option) et elle est basée sur la déclaration communale de piscine pour des raisons fiscales. La variable assureur est corrélée avec les autres options. En effet, les assurés ont tendance à prendre plusieurs options en même temps. De plus, il semble qu'il y a un biais spatial sur la prise des options piscines, toutes choses égales par ailleurs les options piscines ne sont pas fréquemment prises dans le sud.

Analyse de la figure 3.13 Les variables météorologiques sont très liées. En 2020, c'étaient des variables moyennes annuelles qui étaient fournies. Empiriquement, ces variables avec la valeur des maisons vont en partie remplacer les zoniers DDE, BDG et INC. Dans cette première base météorologique, trois

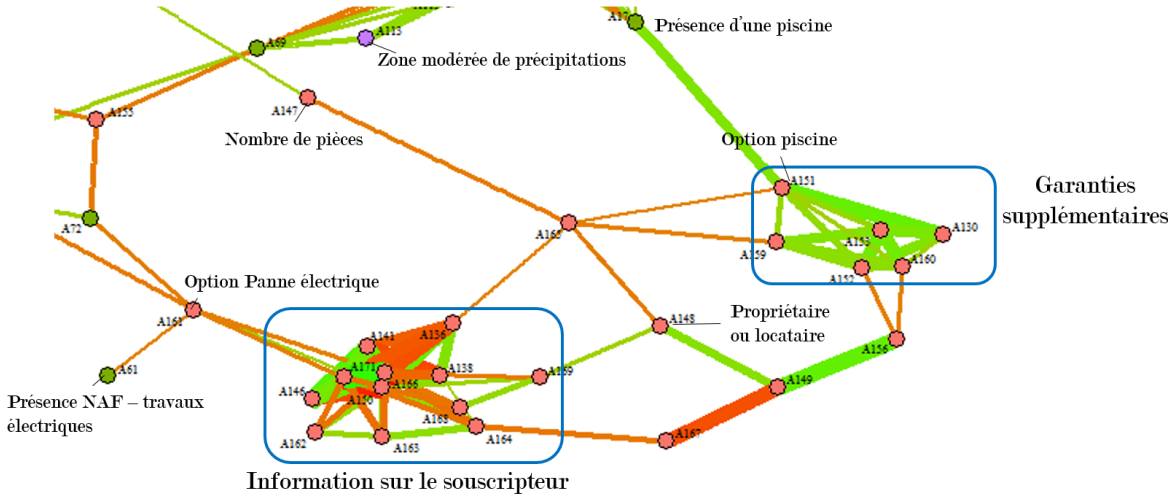


FIGURE 3.12 – Graphique de corrélation pour la base DDE. Focus sur les informations sur les variables assureurs

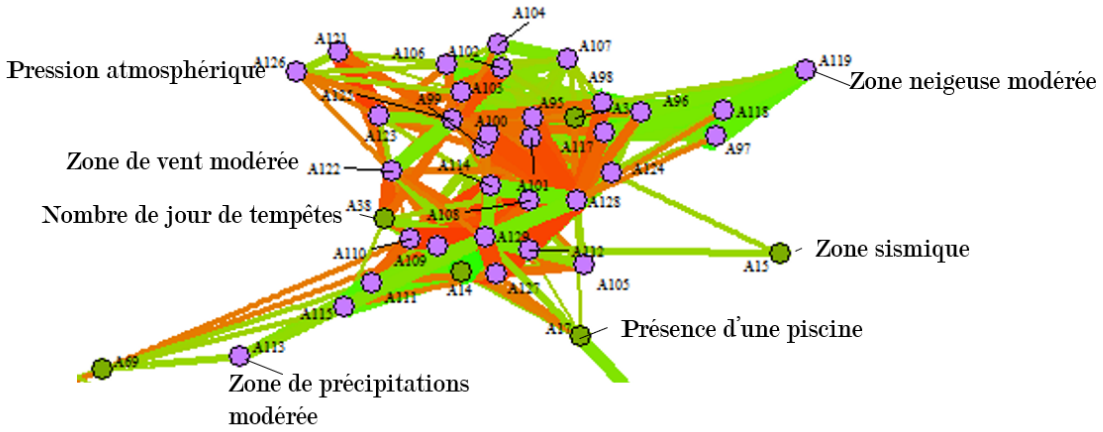


FIGURE 3.13 – Graphique de corrélation pour la base DDE. Les corrélations des variables météorologiques.

groupes apparaissent : les variables de température et de précipitations, les attributs sur la vitesse du vent et les variables "extrêmes" le nombre de jours tels que la température dépasse de 5 dégrées la moyenne saisonnale. Empiriquement, seulement deux variables maximums peuvent être mises dans un GLM. Sinon, les effets deviennent impossibles à analyser à cause des corrélations spatiales.

Analyse de la figure 3.14 Trois clusters peuvent être mentionnés :

— Les informations sur l'environnement - ex : *Nombre de bâtiments dans un rayon de 50 mètres.*

— Les informations sur le bâtiment lui-même - ex : *Nombres d'étages, surface des annexes ...*

— Le code NAF et les informations corrélées à l'environnement autour comme *la période de construction* du bâtiment.

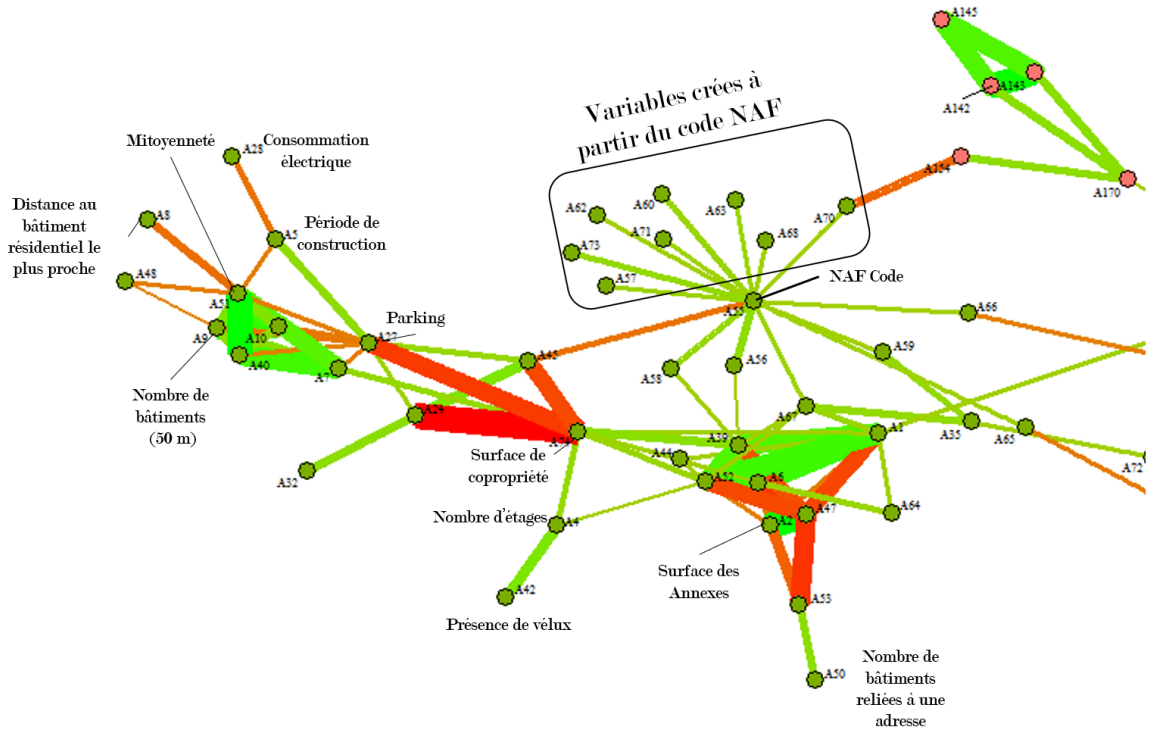


FIGURE 3.14 – Information sur le bâtiment

3.3.4 L'analyse en composantes principales

Pour accélérer et observer les dépendances multivariées, une ACP (Analyse en composantes principales) développée par Pearson, 1901 [80] mais popularisée par Hastie a été ajoutée à ce premier filtre. Ici l'ACP n'est pas utilisée comme une méthode de réduction de dimension. Les composantes principales ne sont pas uniquement utilisées dans le but graphique. En effet, les variables se doivent d'avoir un côté explicable dans une tarification et l'objectif final est de réduire le nombre de variables à améliorer, à maintenir et à fournir. De plus, il est très probable que, dans certains cas, certaines variables viennent à être manquantes.

La normalisation des variables impacte l'ACP mais aussi l'ordre dont lesquelles les composantes principales représentent l'inertie du nuage de points. Pour notre objectif de filtres, une ACP standardisée et une brute est appliquée. Sans s'attarder sur les résultats, ces analyses mettent en avant les corrélations rurales des variables et aussi la dépendance entre les variables de distance et certaines régions.

L'utilisation qui est faite de l'**ACP** est de réduire par composantes principales le nombre de variables. En cela, l'utilisation de l'ACP est la même que le précédent graphe. À une exception, les variables factorielles sont transformées en plusieurs variables booléennes permettant d'étudier plus finement les liens entre les modalités.

3.3.5 La sélection supervisée

La stratégie de modélisation est l'approche $Freq \times CM$. Pour rappel, ici ne sont considérés que les sinistres attritionnels. Si pour la **phase 1**, les bases de *train* et de *test* sont inutiles. Pour les phases suivantes, il est nécessaire de subdiviser la base complète. Pour la consolidation finale, la séparation est faite de la même façon, c'est-à-dire que pour chaque garantie *Gar.* associée au filtre $TM_{Gar.}$, $TM_{Gar.}(\mathcal{D}_{train}) \subseteq (\mathcal{D}_{train})$ et idem pour $\mathcal{D}_{test}$.

De la même façon, la sélection de variables sera une filtration pour la phase 2. L'objectif est de sélectionner une vingtaine de variables.

3.3.5.a La filtration des variables externes

La première étape est d'entrainer un **RF** à partir des 90 variables qui ont été déterminées dans la **phase 1**. Les variables assureurs qui sont trop corrélées aux variables externes sont mises de côté. À cause du nombre de variables et de leur qualité, la performance du **RF** doit toujours être comparé à celle du GLM (M1). Une fois que le **RF** converge vers une solution suffisamment performante, plusieurs analyses permettent d'interpréter le modèle pour son analyse :

— Le graphique de dépendance partielle (Friedman, 2001 [79]) ;
— Les indices d'importance des variables : Il en existe plusieurs types (Carolin and al. 20018 [96]).
— Les méthodes de sensibilité comme Sobol 1993, 2003 [94], [95] ou du Shapley Sensibility analysis (provenant de la théorie des jeux (Shapley, 1952 [93]) qui ont été utilisé pour l'analyse de sensibilité comme dans (Owen, 2014 [90]).

Les graphiques de dépendances partielles (Friedman, 2001 [79]) - extension de Becker and Cleveland (1996) - montrent les effets "marginaux" appris par le modèle RF [7]. Pour utiliser la variable dans le GLM, il faut que les variations soient continues. Basée sur le principe de Monte-Carlo, ce graphique repose sur l'hypothèse d'indépendances des variables avec la variable étudiée. Pour la plupart des variables externes, cette hypothèse est vérifiée.

Sous les mêmes hypothèses, l'importance des variables (Breiman, Friedman, Olshen and Stone (1983) [72]) nécessite une certaine indépendance entre les variables pour être interprétée. Genuer et al. 2010 [83] explique cette problématique dans *Sensitivity to highly correlated predictors*. De plus, l'importance des variables est rarement égale à 0. C'est particulièrement contraignant lors de notre phase de sélection. Par ailleurs, une valeur seuil peut supprimer certaines variables de faibles qualités et d'autres, par effet d'interaction, peuvent avoir une importance élevée.

Les méthodes de sensibilité sont très coûteuses en temps et en calcul. Le nombre d'itérations augmentent exponentiellement avec le nombre de variables et la taille de la base. Pour cet objectif de filtration, ces méthodes sont trop complexes et elles ne sont pas appliquées. Néanmoins, j'ai eu l'honneur de suivre les travaux de Silvia Bucci dans le cadre de son mémoire. Elle a appliqué les méthodes de SOBOL, SHAPLEY sur des modèles XGBoost et RF pour déterminer des interactions à ajouter dans un GLM. Plusieurs points sont à noter sur sa démarche. Elle a eu beaucoup de difficulté à faire converger ces estimateurs de Sobol et de Shapley. En effet, il lui a été nécessaire de travailler sur des bases avec un volume important. Avec du recul et comparant aux différents travaux qui ont été faits par la suite, je pense que cela était dû à la qualité de la base de données. Cette dernière augmentait la dimension de la donnée et donc le nombre de lignes nécessaires pour être une bonne représentation de la base de données initiale. De plus, il faut aussi remarquer que pour les modèles de **RF**, le paramètre *mtry* est faible (1) ou pour le XGBoost le *maxdepth* est faible (7). Cela provient de la combinaison de deux facteurs : la taille de la base de données réduites et la qualité du géocodage impactant différemment les variables. Ainsi, les modèles ont tendance à apprendre des "mauvais" effets introuvables dans la base de test. Les résultats de Silvia Bucci sont cohérents et mettent en avant que la variable *valeur de la maison* et les variables de pluviométries ont des interactions pertinentes sur la garantie DDE. Deux problèmes empêchent l'utilisation de ces interactions : Est-ce que les effets appris sont des effets causaux et pas uniquement spatiaux ? De plus, les deux variables sont continues et créent des effets marginaux assez complexes. Ces interactions remplacent complètement la variable zonier (en particulier *valeur de la maison* + *âge* + *précipitations*). Dans notre cas, ce zonier sera pris en compte uniquement par d'autres variables externes.

7. s'il n'y a pas trop d'interactions apprises dans le modèle.

3.3.5.b L'utilisation de Random forest pour la sélection des variables

Pourquoi utiliser un RF (Breiman 2001 [70]) et non un RNN ou un XGboost ? Des essais ont été faits avec des ANNs. Néanmoins, l'optimisation d'un ANN est coûteuse en temps et les modèles n'ont pas convergé. Très probablement qu'à cause de la qualité de la donnée combinée avec le nombre de variables, les modèles auraient nécessité une volumétrie plus importante d'observations. La préférence du RF au XGBoost provient de la capacité d'arrêter un RF au milieu d'une itération sans amoindrir les performances. L'importance des variables est empiriquement plus robuste que les XGBoosts. Cependant, avec du recul, les XGBoosts performent mieux, sont plus rapides à entrainer et moins couteux en mémoire sur des données assurantielles. Pour la distribution Poisson appliquée aux RF, l'hyper-paramètre de crédibilité proposée par Fritz Bichsel en 1960 et expliqué par Bühlmann and Gisler (2005) [73] a permis d'améliorer la performance. Pour vérifier la convergence du RF vers une structure correcte, la déviance moyenne du modèle est comparée avec la déviance moyenne du GLM (M1) sur $\mathcal{X}_{test}$.

3.3.5.c Quelques propos techniques sur la complexité en temps et en mémoire

Pour cette étude, le package `h2o`([86]) et `distRforest` ont été utilisés avec `R`.

L'idéal pour les RF est le package `distRforest` car il permet de prendre en compte différentes distributions et l'exposition. Néanmoins, la complexité en temps du fait de la non-parallélisation et la complexité en mémoire sont beaucoup plus importantes et l'hyper-paramètre de crédibilité amplifie le temps de calcul.

D'autre part, le package `h2O` permet de paralléliser, mais n'autorise que la distribution gaussienne sans la possibilité d'ajouter un offset. Ainsi, pour les garanties avec un nombre de lignes importantes, la fréquence est modélisée à l'aide `h2o`. Pour les modèles de *CM* ou pour RC, le package `distRforest` est utilisé. Avec le package `h2O`, Pour éviter des problèmes de sur-apprentissage avec que des valeurs prédites soient égales à 0, le *max depth* paramètres est fixé à 15 [8].

Dans les premières modélisations, un RF final `distRforest` était entrainé sur les variables sélectionnées pour les GLMs, mais faute de temps et de résultats concrets, cette étape n'a pas été réitéré.

Une fois, cette sélection faite, le nombre de variables choisies est de l'ordre de 30 à 40 variables. Comme discuter précédemment, aucune étude d'interactions par Shapley ou graphiques de dépendances partielles n'ont été fait systématiquement. Pour vérifier leurs effets marginaux, des modèles GLM-LASSO et RIDGE avec respectivement les distributions Gamma et Poisson ont été entrainés. Ces derniers permettent d'étudier les données de façon supervisée sous un autre angle.

3.3.6 La pénalisation LASSO and RIDGE pour la sélection de variables

Pourquoi utiliser un GLM-LASSO et un RIDGE pour la sélection de variables ? Premièrement, les variables considérées comme influentes par un RF ne sont pas toujours adaptées pour un GLM. Deuxièmement, le RF peut avoir appris des effets inhérents à la qualité des données. Cela entraine des biais de sélections. Finalement, le GLM-RIDGE en particulier permet de comprendre les effets de corrélations entre les variables plus précisément.

La première étape est d'entrainer un **GLM-LASSO** pour enlever les variables les moins importantes. Ce modèle classique (Tibshirani, 1996 [97]) pénalise la fonction de coût avec une norme L_1 sur les coefficients de régression. Cette méthode est intéressante pour des bases avec nombreuses de variables (Meinshausen, Nicolai and Bühlmann 2006 [89]). La faible densité des estimateurs des paramètres permettent de sélectionner les variables les plus influentes. Un **GLM-LASSO** Poisson est utilisé pour la sélection de variables de fréquence (même distribution, même structure multiplicative et même fonction de lien).

La seconde problématique de multicolinéarité est traitée par la régression RIDGE (Hoer, 1970 [84]). Grâce au *Ridge-path*, les corrélations entre les variables peuvent s'expliquer plus facilement. L'objectif qualitatif est d'obtenir un ensemble de variables les moins corrélées possibles pour des raisons de corrélation spatial, mais aussi de qualité.

La combinaison des trois modèles : GLM-LASSO, GLM-RIDGE et RF permettent de sélectionner une vingtaine de variables.

→*Deux raisons de limiter la colinéarité des variables.*

8. Dans le papier Breiman, 2001[70], seuls deux paramètres existent : le nombre d'arbres *ntree* et le nombre de variables tirées au sort *mtry*. Néanmoins, il existe des variantes avec nombreux différents autres paramètres nécessaires pour accélérer et simplifier les calculs.

La première raison opérationnelle est de limiter le nombre de variables. Chaque variable conservée induit un coût de maintenance, un coût pour améliorations et un coût de distributions. La qualité globale des variables pâtit d'un trop grand nombre de variables conservées.

La seconde raison est que la qualité de la donnée est complexifiée par la corrélation (voir chapitre 6). À cause de la qualité, certains coefficients peuvent avoir des effets marginaux contre-intuitifs et même opposés à ce qui était attendu. Ce fut le cas pour la variable *surface habitable* lorsqu'elle était utilisée au début avec *le nombre de bâtiments dans un rayon de 50 mètres* ou *la surface d'empreinte au sol*.

3.3.7 La création d'une prime nette par GLM.

Rappelons que pour le calcul de la prime nette, les GLMs sont de loin la méthode statistique la plus utilisée. Ici, la méthode de sélection de variables utilisées est exactement la même que celle mentionnée lors du chapitre 1.1. À cause de la problématique de qualité sous-jacente, la compréhension du modèle est d'autant plus importante pour vérifier que les modèles ne sont pas aberrants. L'approche de modélisation est celle de $Freq \times CM$ (voir la partie 1.1.2.c) sans considérer de dépendance (De Jong et al., 2008 [75] ou Fress et al. 2014, [76]).

À partir des 20 à 30 variables sélectionnées lors de la **phase 2**, la **phase 3** correspond à une sélection de variable usuelle pour les GLMs. Par habitude, un GLM log-Poisson est utilisé pour la fréquence et un GLM log-gamma pour le coût moyen. Il est néanmoins intéressant de noter que pour le BDG fréquence un GLM binomial négatif était plus adapté pour le modèle (M5) alors que pour le modèle (M1) le modèle le plus adapté était un GLM log-Poisson. Néanmoins, ce dernier a été préféré pour les deux. En minimisant à iso-base, la déviance et l'AIC, un ensemble de variables a été déterminé pour tous les modèles. La validation temporelle a été validée et de rares splines ont été utilisés, en particulier sur l'âge de l'occupant. En comparant les modèles et les prédictions, des graphes similaires comme la figure 3.15 permettent de vérifier que les variables apportent bien une information nouvelle.

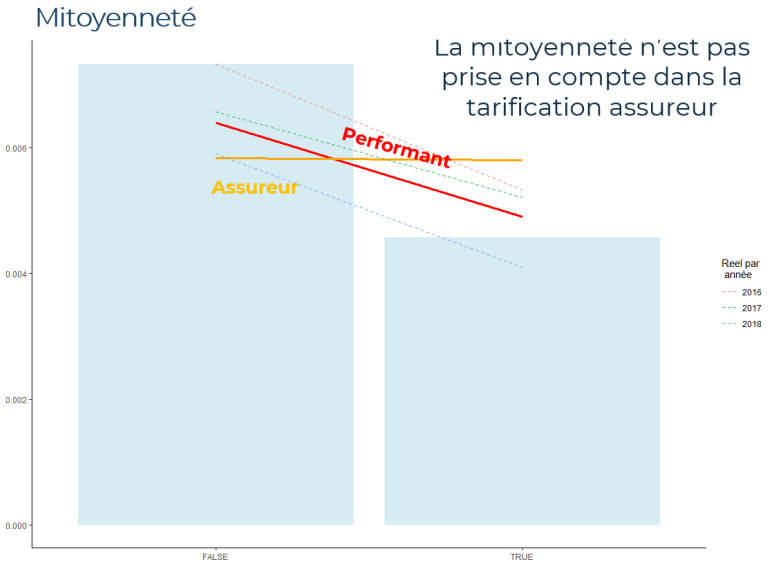


FIGURE 3.15 – Exemple d'effet non capturé par le modèle (M1), mais capturé par la suite. L'étude est faite sur la première base de modélisation DDE en fréquence et sur $\mathcal{X}_{test}$ (3 ème base).

Entre la fréquence et le coût moyen, le nombre de variables supplémentaires pourrait permettre de prendre en compte une dépendance en conditionnant pour l'une d'entre elles. En effet, une habitation avec des dégâts des eaux fréquents peut entraîner un montant plus élevé en coût moyen (Garrido et al. 2016 [82]). Dans notre cadre déjà complexe, sans sinistres graves et avec des modèles non emboités, l'indépendance est supposée et l'approche $Freq \times CM$ est conservée.

Finalement, la **phase 4** a pour objectif d'améliorer les modèle GLMs en le simplifiant, en ajoutant des interactions pertinentes et aussi en essayant d'utiliser les indices de qualité des données. Seule la confiance en le nombre d'étages aurait pu être utilisée à l'aide d'interactions. C'était le seul indicateur

pertinent avec une volumétrie suffisante par modalité et des effets interprétables. Cependant, il n'a pas été ajouté aux modèles, ne pouvant justifier la stabilité de l'indicateur et complexifiant outre-mesure les modèles.

Cette **phase 4** peut-être considérée comme une phase de consolidation ou de mutualisation.

3.3.7.a Les métriques utilisées pour comparer les modèles

La partie 1.2.1.a présente les différentes métriques utilisés dans cette thèse. Si la déviance et l'AIC sont utilisées pour la création du modèle, des métriques moins sensibles au nombre de lignes de $\mathcal{X}_{train}$ comme le Gini et le Pseudo R de Mc Fadden sont utilisés pour comparer les modèles entre eux.

Pour mesurer la performance d'un modèle, le Pseudo R de Mc Fadden [88] est calculé pour chacun des modèles sur la même base. Plus la métrique est proche de 1, plus le modèle est performant. Le Pseudo R permet de comparer les différents modèles non emboités d'une façon macro [9]. Le choix de la distribution du Pseudo R dépend des modèles utilisés. Ici, la distribution sous-jacente du GLM est utilisée pour que Pseudo R mesure directement la performance du modèle GLM.

Pour mesurer la segmentation d'un modèle, l'article de Frees et al., 2014 [77] utilise l'indice de Gini basé sur la courbe de Lorenz ou des concordances pour comparer les primes pures et la sinistralité historique. Sur le graphique, la ligne d'égalité (partant de l'origine avec un angle de 45 degrés) est tracée et représente le cas où une unique prime est proposée à tout le monde. L'indice de *Gini* est défini comme deux fois l'aire entre la courbe et la ligne d'égalité. De façon équivalente, il est possible de regarder uniquement l'aire sous la courbe, l'AUC, qui sera la métrique de segmentation. Il existe plein de nombreuses variantes des Ginis qui seront détaillés en annexe D. Le Gini et ses équivalents représentent la segmentation des primes. En d'autres termes, plus le Gini est élevé, plus les personnes avec une prime plus élevée que d'autres ont une sinistralité plus élevée. Ce n'est pas une mesure de performance, car le Gini est invariant à certaines translations et multiplications. Les écarts à la moyenne des différents modèles ont aussi été calculés. Dans le cadre de GLMs, les écarts calculés sur $\mathcal{X}_{train}$ ou $\mathcal{X}_{test}$ sont quasiment nuls et n'apportent aucune information intéressante pour comparer les modèles (M1), (M2), (M3), (M4) et (M5).

3.3.7.b L'indice de Gini et profitabilité

Selon Frees et al. (2014) [77], une structure tarifaire ayant un $Gini^{frees}$ plus élevé qu'une autre, résultera vraisemblablement en un portfolio plus profitable puisqu'elle fournit une meilleure distinction entre les mauvais risques et les bons risques. Plus le $Gini^{frees}$, qui s'en déduit, est élevée, plus le tarif est segmentant et plus le profit est important. L'idée est de montrer que la segmentation permet de faire des gains de marge. Même s'il n'y existe pas de relation directe entre l'indice de Gini et celui de $Gini^{frees}$, une augmentation de la segmentation induit une augmentation des deux Ginis. Cependant, cette proposition ne tient que dans la mesure où les modèles sont emboités dans un monde compétitif. Dans le cadre de modèles non emboités, les modèles peuvent être très largement concernés par la sélection adverse.

→ *Contre-exemple pour les modèles non emboités.*

Prenons le concept de Akerlof 1970 et reprenons l'exemple de Charpentier et al. (2015) [74]. Supposons un monde à coût moyen constant et une prime pure technique égale à $\mathbb{E}(S) = 1000 \times \mathbb{E}(N)$. Donnons-nous une répartition et une fréquence espérée pour chacun des segments représentés dans la table 3.2

	Jeune (J)	Actif (E)	Sénior (S)	Total
Urbain (V)	12 % (500)	9 % (2000)	12 % (500)	9.5% (3000)
Rural (C)	8% (500)	6.67% (1000)	4% (500)	6,33% (2 000)
Total	10% (1 000)	8.22% (3 000)	6.5% (1 000)	8,23 % (5000)

TABLE 3.2 – Exemple fictif de Charpentier et al. (2015). La fréquence annuelle $\mathbb{E}(N)$ est différentiée par deux segments (l'age de l'occupant et la location) avec le nombre de personnes assurées entre crochet.

Charpentier et al. (2015) [74] considère une compagnie segmentant une variable de plus que l'autre compagnie. Dans ce modèle compétitif, le prospect choisit toujours la prime la plus basse. Quand la

9. D'autres Pseudo R existent comme celui pénalisant le nombre de coefficients appris. Plus complexes à calculer, ils n'étaient pas plus pertinents.

compagnie B segmente mieux que la compagnie A comme le montre les tables 3.3 et 3.4, son ratio technique $\frac{S}{P}$ est égale à 100 % et vient dégrader le ratio de son concurrent. Avec de l'optimisation tarifaire, la compagnie A peut aisément dégager une marge, au détriment d'une part de marché plus faible. Dans ce cadre, la conjecture de Frees et al. (2014) [77] est vérifiée. En effet, il est facile de montrer que le Gini de la compagnie B est plus élevé.

	Companie A - None	Companie B - Localisation	Marché
V (3 000)	82,3	95	82,3
C (2 000)	82,3	63,3	63,3
Primes	247	126,67	373,67
Sinistres	285	126,67	411,67
S/P	115,4 %	100,0 %	110,2 %
Part du marché	66.1 %	33,9%	

TABLE 3.3 – Exemple fictif de Charpentier et al. (2015). Comparaison entre une compagnie segmentant ni sur l'âge ou la localisation et une autre ne segmentant sur que sur la localisation.

	Companie A - None	Companie B - Age	Marché
J (1 000)	82,3	100	82,3
E (3 000)	82,3	82,2	82,2
S (1 000)	82,3	65	65
Primes	82,33	311,67	394
Sinistres	100	311,67	411,67
S/P	121,5 %	100,0 %	104,5 %
Part du marché	20.9 %	79,1%	

TABLE 3.4 – Exemple fictif de Charpentier et al. (2015). Comparaison entre une compagnie segmentant ni sur l'âge ni sur la localisation et une autre ne segmentant que sur la localisation.

Montrons un contre-exemple avec des modèles non emboités. La compagnie A utilise l'âge et la compagnie B uniquement la localisation. Cet exemple simple 3.5 montre qu'aucun des modèles n'est profitable. En effet, un seul segment reste profitable. Dans ce cadre, si la compagnie A a le choix entre segmenter selon l'age ou la localisation, elle choisira la localisation comme son concurrent même si ce dernier à un modèle avec un Gini plus faible. Les deux compagnies sont impactées par une double sélection d'adverse. La conséquence est une diminution importante des primes de toute part.

	A Companie - Localisation	Companie B - Age	Marché	Sinitralité
J-V (500)	95	100	82,3	120
J-C (500)	63,3	100	63,32	80
E-V (2 000)	95	82,2	82,2	90
E-C (1 000)	63,3	82,2	63,3	66.7
S-V (500)	95	65	65	120
S-C (500)	63,3	65	63,3	40
Primes	142,45	196,90	339.35	
Sinistres	166.7	240,00	406.70	
S/P	117,5 %	121,89 %	119,8 %	
Part du marché	40 %	60%		

TABLE 3.5 – Comparaison entre une compagnie segmentant sur l'âge et une autre ne segmentant que sur la localisation.

Dans notre cas, seuls les modèles (M4) et (M5) sont emboités dans le modèle (M1). Même si le modèle (M3) a un meilleur Gini que celui (M1), le modèle (M1) pourrait être préférable si les concurrents utilisent (M1).

3.4 Les résultats sur l'ajout de données externes

3.4.1 Les comparaisons des modèles

Tous les résultats sont divisés par la métrique du modèle (M1) ou soustraits pour l'AUC. Pour chaque métrique, plus la métrique est élevée, plus le modèle est performant. La distribution affichée a été très largement arrondie pour être anonymisés. Les résultats ont été mis à jour sur la 6ème livraison de la base de modélisations.

Garanties	Distribution	M1	M2	M3	M4	M5
DDE	30 %	100 %	76,8 %	101,8 %	109,5 %	127,7 %
VOL	10 %	100 %	110,0 %	118,3 %	120,0 %	138,7 %
INC	20 %	100 %	68,3 %	102,5 %	103,7 %	119,3 %
RC	15 %	100 %	57,0 %	79,0 %	100,8 %	112,1 %
ELEC	15 %	100 %	88,6 %	117,4 %	104,7 %	132,3 %
BDG	10 %	100 %	74,4 %	87,4%	102,3 %	107,3 %

TABLE 3.6 – Évolution des Pseudo R de Fadden pour les modèles de fréquence sur $\mathcal{X}_{test}$. Le ratio est fait entre le modèle et la métrique du modèle (M1).

Garanties	Distribution	M1	M2	M3	M4	M5
DDE	30 %	0 pts	- 1 pts	+ 0,3 pts	+ 0,4 pts	+ 1,2 pts
VOL	10 %	0 pts	+ 0,6 pts	+ 1,1 pts	+ 1,1 pts	+ 2 pts
INC	20 %	0 pts	- 1,2 pts	+ 0,3 pts	+ 0,2 pts	+ 0,7 pts
RC	15 %	0 pts	- 1,5 pts	- 1 pts	- 0,1 pts	+ 0,1 pts
ELEC	15 %	0 pts	- 0,6 pts	+ 0,6 pts	+ 0,3 pts	+ 2 pts
BDG	10 %	0 pts	- 1,1 pts	- 0,3 pts	+ 0,3 pts	+ 0,3 pts

TABLE 3.7 – Évolution de l'AUC pour les modèles de fréquence sur $\mathcal{X}_{test}$. Les chiffres représentent la différence entre l'AUC des modèles et celui de (M1) .

Garanties	Distribution	M1	M2	M3	M4	M5
DDE	20%	100 %	89,7 %	104,4 %	105,5 %	114,7%
VOL	40 %	100 %	89,9 %	107,4%	105,3 %	117,9 %
INC	20 %	100 %	88,9 %	115,0 %	100,6 %	129,8 %
RC	10 %	100 %	12,4 %	92,8 %	106,3 %	140,9 %
ELEC	5 %	100 %	90,5 %	110,3 %	109,5 %	131,4 %
BDG	5 %	100 %	92 %	114,1 %	108,4 %	123,2 %

TABLE 3.8 – Évolution des Pseudo R de Fadden pour les modèles de coût moyen sur $\mathcal{X}_{test}$. Le ratio est fait entre le modèle et la métrique du modèle (M1).

Garanties	Distribution	M1	M2	M3	M4	M5
DDE	20 %	0 pts	- 1,8 pts	+ 0,1 pts	+ 0 pts	+ 0,3 pts
VOL	40 %	0 pts	- 0,6 pts	+ 0,4 pts	+ 0,2 pts	+ 0,6 pts
INC	20 %	0 pts	- 1 pts	+ 0,6 pts	0 pts	+ 1,1 pts
RC	10 %	0 pts	- 3 pts	- 0,6 pts	- 0,1 pts	+ 0,6 pts
ELEC	5 %	0 pts	- 0,5 pts	+ 0,3 pts	-0,1 pts	+ 0,6 pts
BDG	5 %	0 pts	- 0,1 pts	+ 0,6 pts	+ 0,3 pts	+ 0,9 pts

TABLE 3.9 – Évolution de l'AUC pour les modèles de coût moyen sur $\mathcal{X}_{test}$. Les chiffres représentent la différence entre l'AUC des modèles et celui de (M1) .

Remarque : L'AUC pour les modèles de sévérité est volatile. Même si les résultats semblent bons pour le modèle performant, en bootstrappant les données, l'intervalle de confiance à 90 % met en avant une volatilité de ±0,2 à 0,5. Pour la fréquence, Les résultats sont très stables autour de ±0,1 pts, a l'exception pour la RC qui a un intervalle autour de 0.4.

De manière générale, les tables 3.6, 3.7, 3.8 et 3.9 montrent que l'ajout d'Open Data apporte significativement moins d'informations que les données à l'adresse. En performance, les modèles (M3) pour une souscription rapide sont équivalents avec les modèles (M1) traditionnels. Cependant, il y a des différences importantes entre les garanties. Quand la sinistralité dépend grandement des informations sur l'individu comme le nombre d'enfants, les performances sont moindres. C'est le cas en BDG et en RC. Au contraire, en VOL ou en ELEC, l'apport des données est très significatif. Pour les modèles (M3),

le gain en performance est très largement supérieur aux apports des zoniers. Les modèles (M5) et (M4) sont significativement meilleurs que les modèles (M1) et les modèles (M5) sont meilleurs que ceux de (M4).

3.4.2 La comparaison de la prime nette

Les résultats se font par garanties et sont représentés table 3.10.

Garantie	Distribution	M1	M2	M3	M4	M5
DDE	20 %	0 pts	- 1,5 pts	- 0,2 pts	+ 1,4 pts	+ 2,1 pts
VOL	40 %	0 pts	- 0,6 pts	+ 1,8 pts	+ 1,5 pts	+ 2,8 pts
INC	20 %	0 pts	- 1,4 pts	- 0,3 pts	0,1 pts	+ 0,7 pts
RC	10 %	0 pts	- 3 pts	- 2,2 pts	0 pts	- 0,3 pts
ELEC	5 %	0 pts	- 0,3 pts	+ 1,2 pts	+0,1 pts	+ 1,7 pts
BDG	5 %	0 pts	- 0,9 pts	- 0,7 pts	+ 0 pts	+ 0,5 pts

TABLE 3.10 – Évolution de l'AUC pour les modèles combinés sur $\mathcal{X}_{test}$. Les chiffres représentent la différence entre l'AUC des modèles et celui de (M1).

Globalement, l'AUC montre que les modèles (M1) et (M3) sont équivalents en performance. Néanmoins, le modèle (M3) est moins robuste à cause des problèmes de géolocalisation. Cela veut aussi dire qu'il a d'autant plus de potentiel d'amélioration et de pouvoir prédictif. Lors des mises jour suivantes, en particulier des éléments de géolocalisation, une amélioration très significative de la performance a été observée tout comme une meilleure stabilité des multiplieurs. D'un autre côté, le modèle (M2) avec l'open data n'apporte pas assez d'informations pour être utilisable. Les modèles (M4) peuvent être considérés comme les modèles marchés les plus performants du marché utilisés par certains acteurs.

Remarque : Le modèle (M5) est moins segmentant sur la garantie *RC* que le modèle (M1). Une unique variable à l'adresse avait été ajoutée et dont la tendance avait été validée par des graphes *observés VS prédits*. Plusieurs explications sont plausibles. La méthode de calcul de l'AUC n'était pas encore adaptée pour les valeurs égales induisant une volatilité des AUCs. Il est possible que la variable externe n'ait pas été bien bornée (erreur de ma part) ou que la base de test ne soit pas assez représentative (erreur de ma part). Dans les nouveaux modèles, en utilisant les mêmes variables, le modèle (M5) améliore légèrement, mais significativement la segmentation.

	M1	M2	M3	M4	M5
Prime attritionnelle	0 pts	- 1,1 pts	- 0,1 pts	+ 0,4 pts	+ 0,9 pts
Prime sans RC	0 pts	-1 pts	+ 0,1 pts	+ 0,4 pts	+ 1 pts

TABLE 3.11 – Évolution de l'AUC pour les modèles combinés sur $\mathcal{X}_{test}$. Les chiffres représentent la différence entre l'AUC des modèles et celui de (M1).

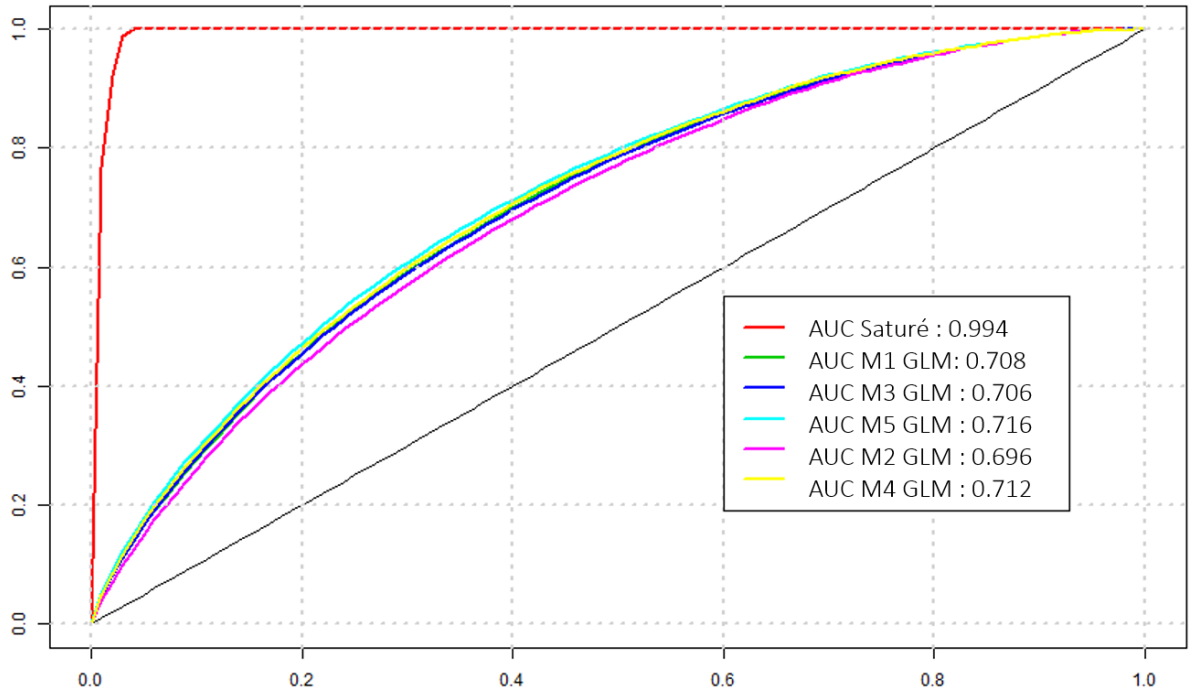


FIGURE 3.16 – Courbe de coordonnance pour la prime attritionnelle globale sur $\mathcal{D}_{TM_{opti}}$.

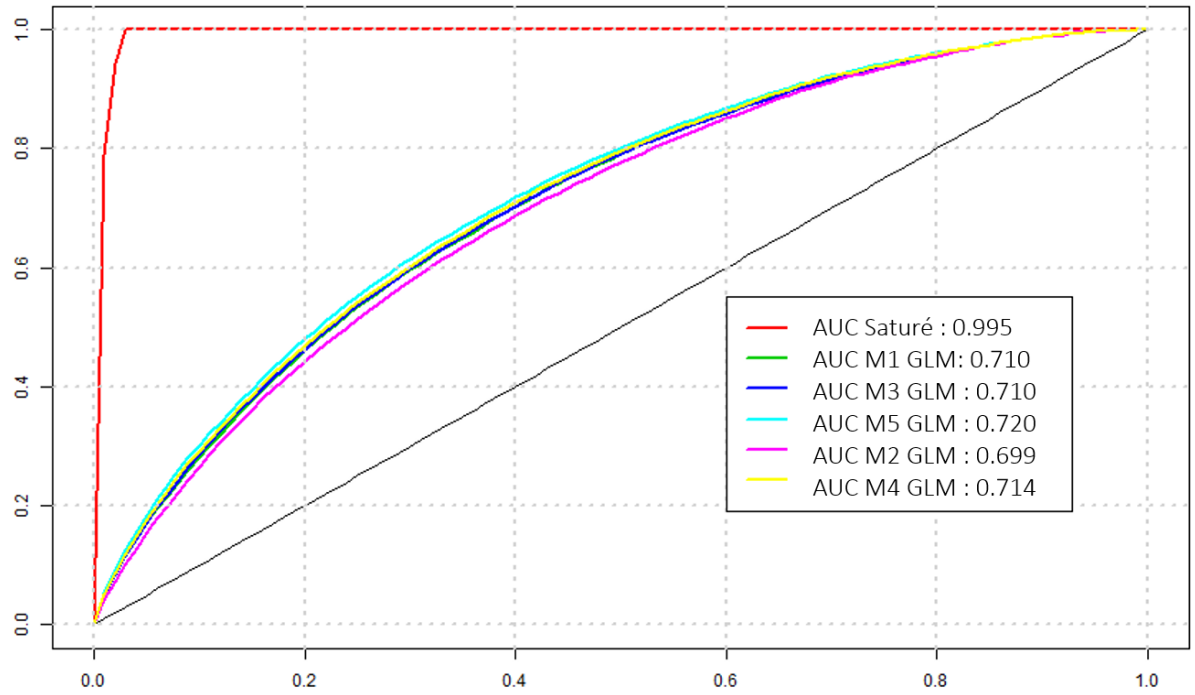


FIGURE 3.17 – Courbe de coordonnance pour la prime de dommages attritionnelle globale sur $\mathcal{D}_{TM_{opti}}$

3.4.3 La comparaison des modèles non emboités par la prime nette

Graphique sur la déviance de Tweedie. Pour comparer les primes nettes, différentes méthodes existent. Il n'existe pas un équivalent du Pseudo R qui serait une fonction des deux Pseudo R de *Freq* et de *CM*. L'indice de Gini peut être appliqué, mais pour les modèles non-emboités, les résultats sont plus compliqués à analyser - voir la partie 3.3.7.a. Cependant, la comparaison des vraisemblances peut être utilisée pour comparer deux modèles.

Prenons deux ensembles de primes Π_1, Π_2 proposées par les modèles m_1 et m_2 respectivement ainsi que la sinistralité observée $\mathbf{Y}$. En affichant le graphique suivant 3.18, qui trace pour tout $\alpha \in [0;1]$:

$$\Big\{\alpha;\ log(\mathcal{L}(\alpha\Pi_1 + (1-\alpha)\Pi_2; \mathbf{Y}|\mathcal{X}_{test}))\Big\}$$

où la log-vraisemblance choisie est celle de la distribution de Tweedie [98] pour paramètre $p = 1.5$ [10]. Plusieurs modèles peuvent être testés. Pour les modèles non emboités, la meilleure prime - barycentre linéaire des deux autres primes, apparait toujours entre les deux modèles. Pour les modèles emboités, la meilleure prime selon ce critère est l'une ou l'autre des approches. Les graphiques 3.18 et 3.19 comparent les modèles (M1) et (M3) aux autres.

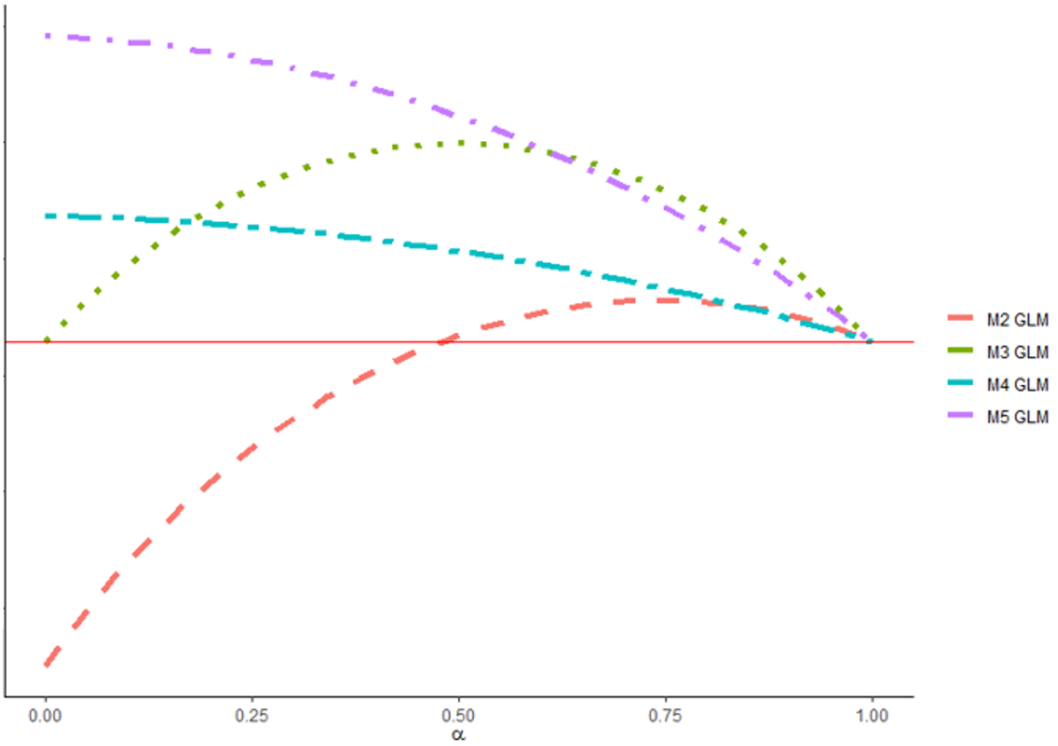


FIGURE 3.18 – Le graphique de la déviance de Tweedie comparant le modèle M1 aux autres $\mathcal{X}_{test}$. La ligne rouge horizontale correspond à la valeur de la vraisemblance du modèle (M1).

Même si la log-vraisemblance des modèles (M1) et (M3) sont proches, l'aspect très bombé de la courbe montre que les deux modèles ne sont absolument pas équivalents. De plus, la meilleure prime se trouve pour $\alpha = 0.5$. Ces graphiques permettent de déduire que les deux modèles sont équivalents en terme d'apport d'information sur la sinistralité. Il est intéressant de remarquer que le modèle (M4) a une performance en deçà de la meilleure combinaison des modèles (M1) et (M3). Si la performance de la combinaison des deux modèles avait été égale au modèle (M5), on aurait pu parler d'une orthogonalité des informations non emboitées des deux modèles. Néanmoins, des corrélations spatiales subsistant par exemple entre le nombre de pièces et les variables météorologiques, viennent rendre non orthogonales les informations.

Les mêmes remarques apparaissent lors de l'étude de la figure 3.19. Le modèle (M2) est complètement inclu dans le modèle (M3) et le modèle (M5) n'est pas complètement emboité dans le modèle (M3). Lors

10. Le choix $p = 1.5$ est quasi arbitraire. En effet, empiriquement, la valeur de p tourne autour de 1.5 et 1.7 pour un certain nombre d'études. La sensibilité de la courbe aux valeurs de p est en réalité négligeable et peu pertinente pour notre analyse.

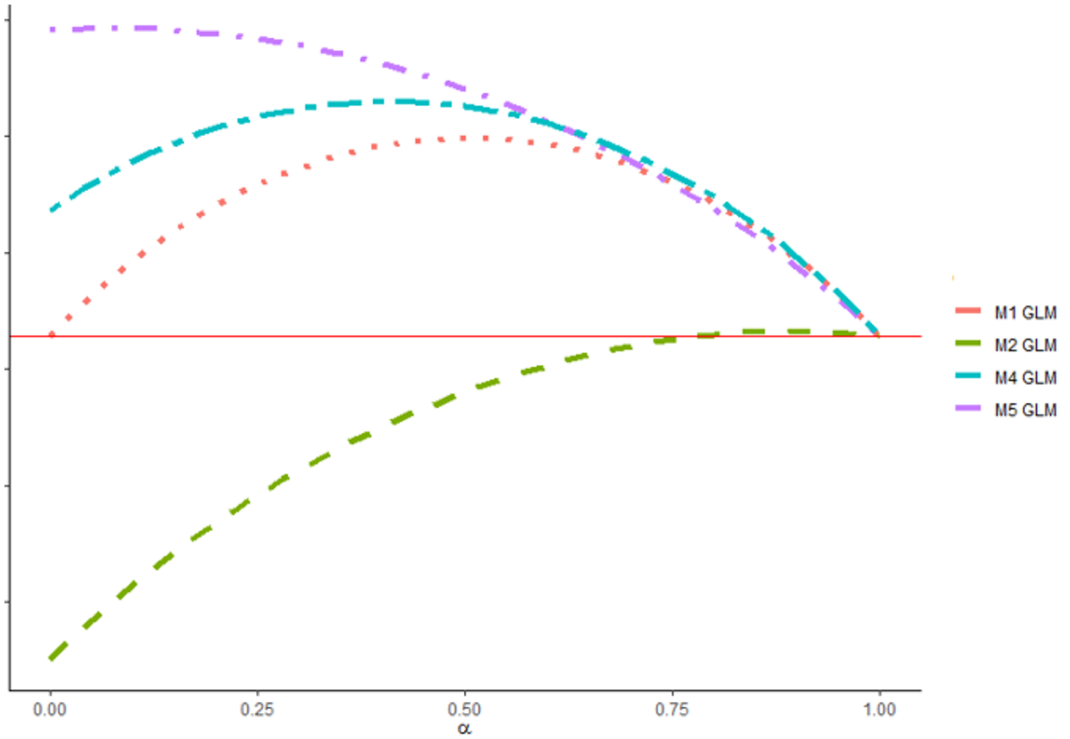


FIGURE 3.19 – Le graphique de la déviance de Tweedie comparant le modèle M3 aux autres. La ligne rouge horizontale correspond à la valeur de la vraisemblance du modèle (M3).

d'une simulation dans le cadre d'un marché utilisant uniquement le modèle (M1), le modèle (M3) devrait provoquer une perte en profitabilité pour les modèles (M2) et (M3) et des gains pour les modèles (M4) et (M5). Au contraire, dans un marché utilisant uniquement le modèle (M3), l'utilisation d'un modèle (M1) ou (M2) serait désastreux. Même pour le modèle (M4) plus performant, la sélection adverse se ferait ressentir. Les meilleures solutions seraient d'utiliser le modèle (M5) ou le modèle (M3).

Conclusion

Ce premier travail démontre que les données externes utilisant la géolocalisation améliore grandement les performances des modèles. En termes de performance, il serait possible de se passer des informations de souscriptions pour mettre en œuvre un processus de souscription accélérer en posant le moins de questions possible. Cependant, la sélection adverse rend difficile l'utilisation de tels modèles. Au contraire, les modèles de performances permettent de gagner en segmentation de façon substantielle.

Le choix de tel ou tel modèle est en réalité illusoire. Dans la pratique, c'est toujours un modèle hybride qui sera choisi permettant de réduire la taille des questionnaires tout en gagnant en performance. Aujourd'hui, si la tarification à l'adresse reste encore un objectif, le projet se développe tout d'abord pour le suivi du portefeuille des contrats comme la revalorisation, le suivi des contrats avant de s'intéresser aux nouvelles affaires. En effet, la modification d'un processus de souscription est un risque opérationnel important. Je voudrais tout de même mentionner que ces données peuvent servir dans la cadre de la RSE (Responsabilités Sociales et Environnementales) où tout mène à croire que l'actuariat devra évaluer le coût en CO2/GES de son portefeuille en se basant sur la note DPE ou GES par exemple, dans un futur assez proche.

Références

[69] Barry, L. and Charpentier, A. (2020). Personalization as a promise : Can big data change the practice of insurance ? *Big Data & Society*, 7(1) :2053951720935143.

[70] Breiman, L. (2001a). Random forests. *Machine learning*, 45(1) :5–32.

[71] Breiman, L. (2001b). Statistical modeling : The two cultures. *Statistical science*, 16(3) :199–231.

[72] Breiman, L. F., Friedman, J., Olshen, S., and Stone, C. (1983). Cj, 1984. classification and regression trees. *Pacific Grove, Kalifornien*.

[73] Bühlmann, H. and Gisler, A. (2006). *A course in credibility theory and its applications*. Springer Science & Business Media.

[74] Charpentier, A., Denuit, M., and Elie, R. (2015). Segmentation et mutualisation, les deux faces d'une même pièce ? *Risques*, 103 :19–23.

[75] De Jong, P. and Heller, G. Z. (2008). Generalized linear models for insurance data. *Cambridge Books*.

[76] Frees, E. W., Derrig, R. A., and Meyers, G. (2014a). Predictive modeling in actuarial science. *Predictive modeling applications in actuarial science*, 1(1).

[77] Frees, E. W., Meyers, G., and Cummings, A. D. (2014b). Insurance ratemaking and a gini index. *Journal of Risk and Insurance*, 81(2) :335–366.

[78] Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1) :55–77.

[79] Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[80] F.R.S., K. P. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11) :559–572.

[81] Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software : Practice and experience*, 21(11) :1129–1164.

[82] Garrido, J., Genest, C., and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance : Mathematics and Economics*, 70 :205–215.

[83] Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters*, 31(14) :2225–2236.

[84] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67.

[85] Indyk, P. and Motwani, R. (1998). *Approximate nearest neighbors : towards removing the curse of dimensionality*.

[86] LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., and Malohlava, M. (2019). *h2o : R Interface for 'H2O'*. R package version 3.26.0.11.

[87] Lemaire, J., Park, S. C., and Wang, K. (2016). The use of annual mileage as a rating variable. *Astin Bulletin : The Journal of the IAA*, 46(1) :39.

[88] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, p.(ed.) : Frontiers in econometrics. *Frontiers in econometrics*, pages 104–142.

[89] Meinshausen, N., Bühlmann, P., et al. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3) :1436–1462.

[90] Owen, A. B. (2014). Sobol'indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1) :245–251.

[91] Ramakrishnan, R., Gehrke, J., and Gehrke, J. (2003). *Database management systems*, volume 3. McGraw-Hill New York.

[92] Revelle, W. (2019). *psych : Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.9.12.

[93] Shapley, L. S. and Snow, R. (1952). Basic solutions of discrete games. *Contributions to the Theory of Games*, 1 :27–35.

[94] Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical modelling and computational experiments*, 1(4) :407–414.

[95] Sobol, I. M. (2003). Theorems and examples on high dimensional model representation. *Reliability Engineering & System Safety*, 79(2) :187–193.

[96] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1) :307.

[97] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, 58(1) :267–288.

[98] Tweedie, M. C. (1984). *An index which distinguishes between some important exponential families*, volume 579. Proc. Indian statistical institute golden Jubilee International conference.

Chapitre 4

Modèles linéaires et la qualité de données

Lors de la première réception des données à l'adresse, chaque observation possédait un indicateur de sa propre qualité. La première idée était de l'incorporer dans les modèles. Mais de quelle façon incorpore-t-on des indices de qualité de données ? De quelles qualités parle-t-on ? Quels changements induisent une évolution de la qualité de données ?

Après une étude approfondie de la littérature, aucunes méthodes découvertes permettaient de répondre à ces problématiques. Ainsi le chapitre 4, à travers l'article *Estimation and prediction with data quality indexes in linear regressions*, propose une structure permettant d'étudier la qualité sur des modèles linéaires. Cependant, les modèles utilisées en tarification sont très largement des GLMs et la structure qui a été posée sur la qualité de données est trop simplistes par rapport aux données de la tarification à l'adresse. Le chapitre 5 étend les précédents résultats aux GLMs.

Finalement, le chapitre 6 enrichit les apports théoriques et rapproche les résultats à la tarification à l'adresse. Dans les sous-sections 6.1.1 et 6.1.2 qu'une baisse de qualité fait diminuer la performance des modèles. La sous-section 6.2.1 veut généraliser cette structure en mettant en avant des propriétés intéressantes de la structure de la qualité de la donnée développée. Par ailleurs, la sous-section 6.3.1 permet de comprendre l'évolution des coefficients lorsque les indices de qualités évoluent. C'est d'ailleurs grâce à cette structure que les choix des améliorations des données ont été permis pour la tarification à l'adresse. La sous-section 6.3.2 considère l'effet de la qualité de données sur l'anti-sélection. L'utilisation de modèles emboités avec une problématique de qualité de données est étudiée.

Ce chapitre reprend l'article Estimation and prediction with data quality indexes in linear regressions co-écrit avec Xavier Milhaud soumis à Computational Statistics.

Abstract

Despite many statistical applications brush the question of data quality aside, it is a fundamental concern inherent to external data collection. In this paper, data quality relates to the confidence one can have about the covariate values in a regression framework. More precisely, we study how to integrate the information of data quality given by a $(n \times p)$-matrix, with n the number of individuals and p the number of explanatory variables. In this view, we suggest a latent variable model that drives the generation of the covariate values, and introduce a new algorithm that takes all these information into account for prediction. Our approach provides unbiased estimators of the regression coefficients, and allows to make predictions adapted to some given quality pattern. The usefulness of our procedure is illustrated through simulations and real-life applications.

keywords : credibility & quality index & regression.

4.1 Introduction

This paper is motivated by very new real-life applications. Indeed, more and more data providers [1] have recently proposed additional services related to the measure of data quality, in addition to the data itself. The need for such measures comes from the fact that the reliability of information gathered from different external sources may vary from one observation to another, due to imprecise merges and time inconsistency most of time. In this context, data quality is often embodied by individualized quality indexes provided with the original dataset. This means that one receives not only the database containing the information about the individuals and their characteristics, but also some indexes giving the confidence one can have in such values. This is made through a $(n \times p)$-matrix of data quality indexes, where n denotes the number of observations and p is the number of associated characteristics. Statistically speaking, a natural question is then : "*how can such individual quality indexes be used for estimation and prediction ?*". This question, which has not yet been explored, is nowadays of crucial importance for practical applications. In this view, we try in this paper to make our contribution into the linear regression framework.

More generally, the topic of data quality has a myriad of applications and the literature has reached a consensus of multiple dimensions analysis, see (Todoran et al., 2014[112]). These dimensions are most of time layered, as explained by Wang et al. 1995 [116]. In the definition of data quality, the keystone is the integrity of data, which is assumed to be verified in our paper. Straightforwardly, the integrity is satisfied if all the observations are plausible (Ramakrishnan and Gehrke, 2000 [108]). Moreover, in the linear regression's setting, four dimensions are usually considered as crucial (Rogova and Bosse, 2010 [109]) : i) completeness, or how to deal with missing data ; ii) imprecision (fuzziness, consistency, accuracy), or how to integrate the accuracy's influence in modelling ; iii) uncertainty (probability, credibility, reliability) ; and iv) timeliness (the older the data the more uncertain). In their paper, Decker and Martinenghi, 2009 [100] highlight that time impacts all other dimensions.

We focus hereafter on uncertainty, knowing that numerous methods have already been developed to deal with completeness and imprecision. About completeness, these methods are mostly based on assumptions like MCAR (Missing Completely At Random), MAR (Missing At Random), or MNAR (Missing Not At Random) ; see for instance Van Buuren 2018 [114] or Little and Rubin 2019 [105]. Concerning imprecision, observations are seen as a shift from the *real* observations, as in Errors-In-Variable (EIV) models (VanHuffel and Lemmerling, 2013 [115]) or Regression Dilution (Berglund et al., 2008 [99]). In such models, the estimated coefficient is often lower than the *real* one (in absolute value). This is the so-called *attenuation* phenomenon, which has been extensively described in the literature (see Hausman (2001)[103], Fuller (2009), [102]). For us, the observation's uncertainty is quantified and is called quality index. We assume that the response is fully observed, and that the quality index is perfectly measured. The latter represents the probability for the observed value to be the right one. Recent works dealing with uncertain data considered a mismeasurement approach, using hyperparameters (as in EIV models, see Tami et al., 2018 [111] ; or Trabelsi et al., 2016 [113] for decision trees). Without taking into account data quality directly, the RANSAC (RANdom SAmple Consensus, Fischler and Bolles,

1. See for instance namR at `https://namr.com/en/`.

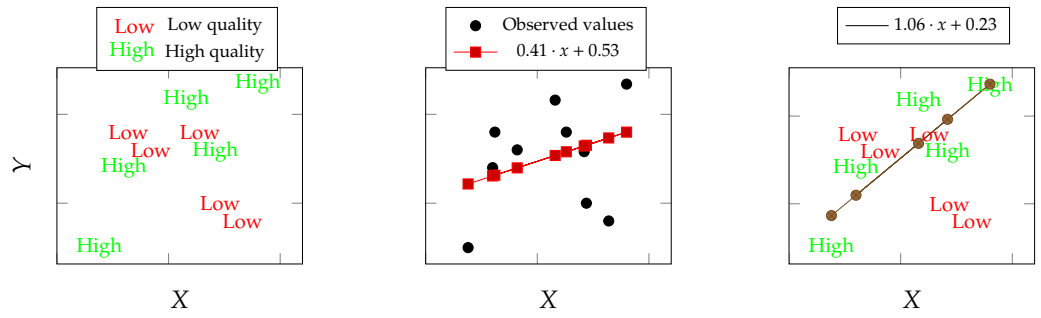
1981 [101]) algorithm and its different extensions have been developed to use only a subsample of the *real* observations. In the sequel, we argue that given a multivariate dataset and its corresponding quality, it is possible to find the best regression coefficients to be used for individual predictions with respect to specific data quality indexes. This work thus stands out from EIV models customized for mismeasurement approach. However, mismeasurement is overshadowed by the credibility of the data.

The paper is built as follows : Section 4.2 introduces the general idea. Section 4.3 describes the estimation method, and introduces the algorithm that allows to make predictions based on individual quality indexes. In Section 4.4, we give the theoretical results underlying the bias correction of regression coefficients. We then perform in Section 4.5 a simulation study showing the validity of these results. Finally, Section 4.6 is devoted to a real-life application, also making the link with missing data.

4.2 Integration of data quality

4.2.1 Example

Assume that the quality index is divided into “low” and “high” quality data, as in Figure 4.1a. Without having any prior information on the quality, a classical regression can easily be fitted (see Figure 4.1b). However, keeping in mind that the quality differs depending on observations, it makes sense to consider that this model should be adjusted. For instance, Figure 4.1c shows that fitting a model only on high quality data strongly modifies the previously obtained regression coefficients. Indeed, the slope increases from 0.41 to 1.06. Intuitively, quality should therefore be considered to avoid loosing some crucial information. Also, how should data quality modify the predicted outcome? For instance, if two individuals have the same feature X with different qualities (one is "high" and the other one is "low"), the prediction should also differ.



(a) Scatter plot of all observations, whatever the quality.

(b) Regression without inserting quality information .

(c) Regression using only “high quality” observations.

FIGURE 4.1 – Different regression models depending on data quality.

4.2.2 Explanatory variables : a latent variable model integrating quality

To integrate the information on data quality in linear regressions, a primary idea could be to weigh the observations. However, this is not feasible in a multivariate setting. In the univariate case, the use of the quality index as a weight will neither correct the quality impact, nor provide a way to adapt predictions to the data quality. In our setting, we would like to benefit from more comprehensive information provided by individualized quality indexes, referring to the confidence one has about the i-*th* observation of the j-*th* covariate.

In this view, we introduce the following latent variable model :

$$\mathbf{X} = \mathbf{X}^{real} \circ \mathbf{\Omega} + \mathbf{Z} \circ (J_{n,p+1} - \mathbf{\Omega}) \tag{4.1}$$

where $\circ$ corresponds to the Hadamard product, $J_{n,p+1}$ is the $n \times (p+1)$-identity matrix under Hadamard multiplication, $\mathbf{X} = (X_{ij}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ are the observed covariates, $\mathbf{X}^{real} = (X_{ij}^{real}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ are the *real* covariates, $\mathbf{Z} = (Z_{ij}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ are the *wrong* covariates (with same distribution as $\mathbf{X}^{real}$), and $\mathbf{\Omega} = (\omega_{ij}) \in \mathcal{M}_{n\times(p+1)}(0,1)$ is a binary mask indicating whether the i-*th* observation of the j-*th* covariate X_{ij} is perfectly observed or not. In other words, $\mathbf{\Omega}$ tells us if one observes the real observation or not, and ω_{ij}

can be seen as a Bernoulli random variable. In practice, the data at disposal is made of individual quality indexes through some matrix $\mathbf{Q} = (1, Q_j)_{j=1,...,p} = (Q_{ij}) \in \mathcal{M}_{n\times(p+1)}([0,1])$, together with n *iid* replications $(Y_i, \mathbf{X}_i)_{i=1,..,n}$ where $Y_i \in \mathbb{R}$ and $\mathbf{X}_i = (1, X_{i1}, ..., X_{ip}) \in \mathbb{R}^{p+1}$. Each element Q_{ij} of the matrix $\mathbf{Q}$ informs us on the quality related to the observed covariate value X_{ij}. We use $\mathbf{Q}$ as the expectation of $\mathbf{\Omega}$, leading to define the quality index as a credibility index. This means that for all $i = 1, ..., n$ and $j = 1, ..., p$, we have :

$$\mathbb{E}(\omega_{ij}) = \mathbb{P}(\omega_{ij} = 1) = Q_{ij}. \tag{4.2}$$

The variance is thus obtained straightforwardly : $Var(\mathbf{\Omega}) = \mathbf{Q} \circ (J_{n,p+1} - \mathbf{Q})$. Let us introduce the mean quality for covariate X_j, i.e. $\bar{Q}_j = (1/n)\sum_{i=1}^{n} Q_{ij}$. We assume that for each covariate there is at least one individual strictly positive observed quality index, i.e. for $j = 1, ..., p$, $\{i \mid Q_{ij} \neq 0\} \neq \emptyset$ (hence $\bar{Q}_j \neq 0$). This assumption is not restrictive in practice, because such covariates would be removed from the data.

4.2.3 Regression model under consideration

We now consider the multivariate linear regression framework with the intercept, i.e. the dependent response $Y = (Y_1, ..., Y_n)$, the explanatory variables $\mathbf{X} = (1, \mathbf{X}_1, ..., \mathbf{X}_n)$, and the relationship

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = \mathbf{x}\beta, \tag{4.3}$$

with $Y \sim \mathcal{N}(\mathbf{x}\beta, \sigma^2)$ and $\beta = (\beta_0, \beta_1, ..., \beta_p)^T \in \mathbb{R}^{p+1}$ a vector of regression coefficients.

Remind that in this work, $\mathbf{X}$ is not fully observed since it is governed by the latent variable model (4.1). This has obvious consequences on the estimation of the regression coefficients. However, the novelty lies in that additional information given by the individual quality indexes $(\mathbf{Q}_1, ..., \mathbf{Q}_n)$, where $\mathbf{Q}_i = (1, Q_{i1}, ..., Q_{ip})$ can be used to compensate this incomplete information. Each given $\mathbf{Q}_i$ defines a *quality pattern*, denoted further K for ease of notation. For instance, using Example 4.2.1 with two similar covariates, four quality patterns $K = (Q_1, Q_2)$ exist. These are given by (low,low), (low,high), (high,low), (high,high).

4.2.4 Interactions in the latent variable model

In the generating process (see Equation (4.1)), several correlation structures between $\mathbf{X}^{real}$, $\mathbf{Z}$ and $\mathbf{\Omega}$ can be studied. They are connected to Rubin's nomenclature (Rubin (1976)[110]), namely the MCAR, MAR and MNAR settings. In what follows, we focus on an extension of MCAR (Heitjan and Basu (1996),[104]), in that MCAR would correspond to our setting where the quality indexes would only take values 0 or 1. More formally, for one given j, the real observations X_j^{real}, the wrong values Z_j and the quality variable Ω_j are independent. For all $j \neq k$, the random variables Ω_j and Ω_k are independent, and Ω_k is independent from X_j^{real} and Z_j. Finally, X_j^{real} is also independent from Z_k. In a nutschell, the sole potentially correlated variables are X_j^{real} with X_k^{real}, and Z_j with Z_k. In practice, the independence between the quality variables Ω_j and Ω_k $(j \neq k)$ suggests that all observed covariates come from unrelated sources.

4.3 Estimation process with quality indexes

4.3.1 Intuition : reducing the predictive error by mitigating on quality patterns

In the linear regression context, the solution $\hat{\beta}$ minimizes the Residual Mean Squared Error (RMSE), given by $RMSE(\beta|\mathbf{X}, Y) = (1/n)\sum_i (Y_i - \mathbf{X}_i\beta)^2$. In our framework, we can put together two individuals with same quality indexes, *i.e.* $\mathbf{Q}_i = \mathbf{Q}_{i'}$, and define a corresponding quality pattern K. Therefore, looking at all individuals sharing the same quality pattern K, introduce the coefficient $\hat{\beta}^K$ minimizing $\sum_{i\backslash\mathbf{Q}_i=K} RMSE(\beta|\mathbf{X}_i, Y)$. The latter coefficient is of course different from $\hat{\beta}$, and has a better predictive power for observations with quality pattern K. Denote $P(\mathbf{Q})$ the set of all observed quality patterns. The cost metric can thus be improved, since

$$RMSE(\hat{\beta}|\mathbf{X}, Y) = (1/n)\sum_{K\in P(\mathbf{Q})}\sum_{i\backslash\mathbf{Q}_i=K}(Y_i - \mathbf{X}_i\hat{\beta})^2 \geq (1/n)\sum_{K\in P(\mathbf{Q})}\sum_{i\backslash\mathbf{Q}_i=K}(Y_i - \mathbf{X}_i\hat{\beta}^K)^2 \tag{4.4}$$

where $(1/n)\sum_{K\in P(\mathbf{Q})}\sum_{i\backslash\mathbf{Q}_i=K}(Y_i - \mathbf{X}_i\hat{\beta}^K)^2 = \sum_{K\in P(\mathbf{Q})} RMSE(\hat{\beta}^K|\mathbf{X}, Y, \mathbf{Q} = K)$.
Here, the estimated coefficient $\hat{\beta}^K$ is the solution for individuals having quality pattern K. To seek such

coefficients would lead to fit as many models as quality patterns, which looks impossible. In Section 4.3.3, we suggest an algorithmic-based strategy to overcome this issue. This strategy relies on theoretical results provided in Section 4.4.

4.3.2 Benchmark and models under study

When fitting a linear regression in a classical context, the explanatory variables are considered as perfectly observed. This means that we consider the so-called *real model*,

$$\mathbb{E}[Y \mid \mathbf{X}^{real}] = \mathbf{X}^{real}\beta,$$

where the coefficients are estimated based on the real (fully observed) dataset $\mathbf{X}^{real}$, leading to the estimator $\hat{\beta}$. Hereafter, we need to figure out the impact of the covariate generating process. To this aim, we introduce the three following models :

— M_1 **(*perfect quality* model)** : fitted on the observed dataset $\mathbf{X}$ in the case when the covariates are perfectly observed, i.e. all the quality indexes Q_{ij} equal 1 :

$$\mathbb{E}[Y \mid \mathbf{X}, \mathbf{Q} = J_{n,p+1}] = \mathbf{X}\beta^{M_1}.$$

Note that $\hat{\beta}^{M_1}$ differs from the classical estimator $\hat{\beta}$ (see also Section 4.4.3.a).

— M_2 **(*naive* model)** : fitted on the observed dataset $\mathbf{X}$ without taking into account the associated quality :

$$\mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}\beta^{M_2}.$$

— M_3 **(*pattern-adjusted* model)** : based on $\mathbf{X}$ and $\mathbf{Q}$, obtained from Algorithm 1 (see Section 4.3.3) :

$$\mathbb{E}[Y \mid \mathbf{X}, \mathbf{Q}] = \mathbf{X}\beta^{M_3}.$$

When $\mathbf{Q} = J_{n,p+1}$ (or $Q_{ij} = 1$ for all i, j), model M_3 is similar to M_1 and $\hat{\beta}^{M_3} = \hat{\beta}^{M_1}$.

Basically, model M_2 is adapted to the observed dataset but does not incorporate quality. Model M_1 is not adapted to the observed dataset (collected with unperfect quality), and model M_3 represents an individualized version of M_2. Model M_2 leads to biased coefficients, due to the imperfect quality that has not been considered. Model M_1 is not biased, but it cannot be used in real life! Finally, M_3 takes into account the quality of each observed value. Note that $\hat{\beta}^{M_3}$ consists of individualized regression coefficients in practice, since it is based on each quality pattern. This means that for two individuals having different quality patterns $K = \mathbf{Q}_i$ and $K' = \mathbf{Q}_{i'}$, the vector $\hat{\beta}^{M_3}$ will differ. To simplify the notations, $\hat{\beta}^{M_3}$ is denoted $\hat{\beta}^K$ in the sequel. Before switching to the theoretical results that show how to infer these regression coefficients, one presents the algorithm that allows to make adjusted predictions depending on the quality pattern.

4.3.3 Algorithmic prediction using data quality indexes

Algorithm 1, leading to the pattern-adjusted model M_3, allows to take into account data quality **without increasing the number of estimated parameters**. It is based on results provided in Section 4.4.

To begin with, one estimates M_2 from $\mathbf{X}$. Then, considering an estimate of the observed covariance matrix Σ (given by $\hat{\Sigma} = (1/n)\ {}^t\mathbf{X}\mathbf{X}$, where ${}^t\mathbf{X}$ denotes the matrix transpose) and the quality indexes $\mathbf{Q}$, the covariance matrix Σ^{real} related to $\mathbf{X}^{real}$ is estimated (steps [2] and [3]). From $\hat{\Sigma}^{real}$, $\hat{\beta}^{M_2}$ and Y, the coefficients $\hat{\beta}^{M_1}$ are inferred in step [4], using the mean quality $\bar{\mathbf{Q}} = (1, \bar{Q}_1, ..., \bar{Q}_p)$. This estimator is obviously unadapted since it eclipses the heterogeneity of individual quality indexes. Finally, one thus has to adjust the latter coefficients to take it into account. To summarize, Algorithm 1 seeks the hidden *real* coefficients based on the observed covariates and their associated qualities, and then deduce the set of coefficients adapted to each quality pattern.

4.4 Theoretical results

Recall that $X_j = X_j^{real} \circ \Omega_j + X_j \circ (J_{1,n} - \Omega_j)$ with $\Omega_j \sim \text{Bernoulli}(Q_j)$, and that $\bar{X} = (1/n)\ \sum_{i=1}^{n} X_i$ refers to the empirical mean of any random variable X. Without loss of generality, we assume that X_j^{real} is centered ($j = 1, ..., p$). Moreover, we denote Σ (respectively Σ^{real}) the observed (resp. real) covariance matrix of the covariates. Given two covariates X_j and X_k (resp. X_j^{real} and X_k^{real}), the terms $Cov_{jk} = Cov(X_j, X_k)$ (resp. $Cov_{jk}^{real} = Cov(X_j^{real}, X_k^{real})$) are the non-diagonal elements of Σ (resp. Σ^{real}).

Algorithm 1: Quality-based prediction

Data: $\mathbf{X}$ with corresponding quality matrix $\mathbf{Q}$, and response Y.
Result: Prediction of Y_i given $(X_{i1}, \ldots, X_{ip})$ and $(Q_{i1}, \ldots, Q_{ip})$.
begin
  [1] Estimate $\hat{\beta}^{M_2}$ from $\mathbf{X}$ and Y ;
  [2] Estimate the covariance matrix $\hat{\Sigma}$ from $\mathbf{X}$;
  [3] Estimate $\hat{\Sigma}^{real}$ from $\hat{\Sigma}$ and $\mathbf{Q}$ doing : (Lemma 2) ;
  for $j = 1, ..., p$ *and* $k = 1, ..., p$ **do**
    Estimate $Cov(X_j^{real}, X_k^{real})$ (Lemma 1);
  end
  [4] Estimate $\hat{\beta}^{M_1}$ from $\hat{\beta}^{M_2}$, $\hat{\Sigma}^{real}$ and Y (Theorems 1 and 2);
  [5] Prediction of Y;
  for *each quality pattern* $K \in P(\mathbf{Q})$ **do**
    Estimate $\hat{\beta}^K$ from $\hat{\beta}^{M_1}$, $\hat{\Sigma}^{real}$ and K (Corollaries 1 and 2) ;
    for *For each individual i such that* $\mathbf{Q}_i = K$ **do**
      $\hat{Y}_i = \mathbf{X}_i \hat{\beta}^K$;
    end
  end
end

4.4.1 Assumptions about the correlation structure in the covariate generating process

First of all, in the same spirit as Muzellec et al. (2020) [106], our main assumption is that the *wrong* observations Z_j have the same distribution as the *real* ones X_j^{real} $(j = 1, ..., p)$. Note that Z_j is thus centered, and assume also that the distribution of each covariate has a finite second-order moment. We discuss now four assumptions underlying the correlation structure between the components of $\mathbf{X}^{real}$ and those of $\mathbf{Z}$:

(X-A1) All the random variables X_j^{real} $(j = 1, ..., p)$ are mutually independent.
(X-A2) Each variable X_j^{real} is at most correlated with another variable X_k^{real} $(j \neq k)$.
(Z-A1) All the random variables Z_j and Z_k are independent.
(Z-A2) The vector (Z_j, Z_k) has the same correlation structure than (X_j^{real}, X_k^{real}), $j \neq k$.

Note that under (X-A1), considering either (Z-A1) and (Z-A2) is similar. Under *(X-A2)*, the covariance Σ^{real} is made of submatrices Σ_{jk}^{real} such that $\Sigma_{jk}^{real} \in \mathcal{M}_{2\times 2}(\mathbb{R}^*)$. Trivially, under (X-A1) and (Z-A1), submatrices Σ_{jk}^{real} are diagonal matrices. We consider pairwise correlated covariates at most, knowing that our results could be extended to other correlation structures. However, the additional complexity in such contexts would lead to unreadable formulas. Notice that in EIV-models, the assumption (X-A1) is often supposed. To make the link with real-life applications, these assumptions are connected to how the covariates are collected. For instance, if they come from the same database extraction (and are thus based on the same underlying key), the correlation between Z_j and Z_k should be similar to the one between X_j^{real} and X_k^{real}. In this case, *(X-A2)* and *(Z-A2)* would be appropriate. When collecting data from completely different sources, assumption (Z-A1) should logically be used.

4.4.2 Observed and real covariance matrices

4.4.2.a Relationship between covariances

Given a dataset with two covariates and their joint quality $(X_{ij}, Q_{ij})_{i=1,...,n}$, $(X_{ik}, Q_{ik})_{i=1,...,n}$ with $j \neq k$, we now state the relation between the observed and real covariances. Note that under *(X-A1)*, all terms obviously equal zero. We thus focus on the different cases under (X-A2).

Lemma 1

Under (X-A2), the relation yields :

$$\text{Under (Z-A1)} : Cov_{jk} = Q_j Q_k \, Cov_{jk}^{real}. \tag{4.5}$$

$$\text{Under (Z-A2)} : Cov_{jk} = (1 + 2Q_j Q_k - Q_j - Q_k) \, Cov_{jk}^{real}. \tag{4.6}$$

See the proof 1 at section 4.7

Such results could be extended to other correlation structures between the components of $\mathbf{Z}$ and $\mathbf{X}^{real}$. However, one would need to specify this correlation structure. Note that if X_j^{real} and Z_j have the same distribution, then their standard deviations are similar. As a result, the same relationship exists for Pearson's correlation.

Because the response Y is supposed to be perfectly observed, we also have

$$Cov(X_j, Y) = Q_j Cov(X_j^{real}, Y), \qquad \text{for } j = 1, ..., p.$$

4.4.2.b Link between the information matrices under (Z-A1)

In linear regression, the information matrix (inverse of the empirical estimate of the covariance matrix) is a milestone to derive asymptotic results. We thus aim to link the real information matrix to the observed one. Remind that for two covariates X_j and X_k with corresponding covariance matrix Σ_{jk}, the latter can be consistenly estimated using the estimator

$$\hat{\Sigma}_{jk} = (1/n) \, {}^t(X_j, X_k)(X_j, X_k).$$

Lemma 2

Assume that the covariance matrix Σ_{jk}^{real} is not singular.
Under (X-A1) and (Z-A1), $\Sigma_{jk} = \Sigma_{jk}^{real}$.
Under (X-A2) and (Z-A1),

$$\Sigma_{jk}^{-1} = \frac{(1 - (\rho_{jk}^{real})^2)}{(1 - (\rho_{jk}^{real})^2 Q_j^2 Q_k^2)} (\Sigma_{jk}^{real})^{-1} \circ \begin{bmatrix} 1 & Q_j Q_k \\ Q_j Q_k & 1 \end{bmatrix}, \tag{4.7}$$

$$\tag{4.8}$$

with ρ_{jk}^{real} the Pearson correlation between X_j^{real} and X_k^{real} such that $|\rho_{jk}^{real}| \neq 1$.

See the proof 2 at section 3.

Under (X-A2) and (Z-A2), one would need to replace $Q_j Q_k$ by $(1 + 2Q_j Q_k - Q_j - Q_k)$. The proof is trivially the same.

Using the strong law of large number (SLLN), we have

$$\hat{\Sigma}^{real} \xrightarrow{a.s.} \Sigma^{real} \quad \text{and} \quad \hat{\Sigma} \xrightarrow{a.s.} \Sigma, \quad \text{as } n \to +\infty. \tag{4.9}$$

We now focus on the relationship between the estimators $\hat{\beta}^{M_2}$ (obtained when fitting M_2 based on $\mathbf{X}$) and $\hat{\beta}$ (obtained when fitting the classical linear regression model, based on $\mathbf{X}^{real}$).

4.4.3 Impact on regression coefficients, case of independent covariates

As already mentioned, to consider (Z-A1) or (Z-A2) under (X-A1) makes no difference. Remind that M_2 is fitted on an unperfect dataset (observed covariates), contrary to M_1 which would estimate the coefficients from the real dataset (thus fictive since not observed). Here, we quantify the bias and variance of estimators coming from the fitting of M_1. The coming results thus refer to steps [3]-[4] of Algorithm 1 in Section 4.3.3.

4.4.3.a Relation between M_1 and M_2 coefficients

We first study how the estimator of the regression coefficient is modified when going from model M_2 to model M_1. Remind that for $j = 1, ..., p$, $\bar{Q}_j = (1/n) \sum_{i=1,...n} Q_{ij}$.

Theorem 1

Under (X-A1) and (Z-A1) or (Z-A2), we have for $j = 1, ..., p$:

$$(1 / \bar{Q}_j)\, \hat{\beta}_j^{M_2} \xrightarrow{a.s.} \beta_j, \quad \text{as } n \to +\infty. \tag{4.10}$$

The intercept remains unchanged, *i.e.* $\hat{\beta}_0^{M_2} \xrightarrow{a.s.} \beta_0, \quad \text{as } n \to +\infty.$

See the proof 3 at section 4.7.

Since model M_1 incorporates the data quality through the matrix $\mathbf{Q}$ (like M_2), the coefficient $(\hat{\beta}_j^{M_2}/\bar{Q}_j)$ **is assimilated to** $\hat{\beta}_j^{M_1}$ **(**$j = 1, ..., p$**), and** $\hat{\beta}_0^{M_1} = \hat{\beta}_0^{M_2}$. Notice that $\hat{\beta}_j^{M_1}$ differs from $\hat{\beta}_j$, as already mentioned in Section 4.3.2. Indeed, $\hat{\beta}^{M_1}$ and $\hat{\beta}$ are not based on the same dataset, meaning that the estimators have different variances (despite they have the same expectation, see the covariate generating process).

If the covariates were not centered, they would be a shift in the intercept (see Appendix 3). The latter would equal to $\sum_{j=1}^{p} \mathbb{E}(X_j)\beta_j$. Then, the expected difference would be

$$\hat{\beta}_0^{M_2} \xrightarrow{a.s.} \beta_0 - \sum_{j=1}^{p} \mathbb{E}(X_j) \frac{\beta_j^{M_2}(1 - Q_j)}{Q_j}. \tag{4.11}$$

By definition, $\mathbb{E}(\Omega_j) = Q_j$. This means that when the mean quality lowers, the model gives more credit to the global mean than to the covariate value.

Remark : The variance of $\hat{\beta}^{M_2}$ is well known, given by

$$\begin{aligned} &Var(\hat{\beta}^{M_2}) = \hat{\sigma}^2 \ ((1/n)\, {}^t\mathbf{X}\mathbf{X})^{-1}, \\ &Var(\hat{\beta}^{M_1}) = Var(\hat{\beta}^{M_2} \circ D) = \hat{\sigma}^2((1/n)\, {}^t\mathbf{X}\mathbf{X})^{-1} \circ D^2 \geq Var(\hat{\beta}^{M_2}), \end{aligned} \tag{4.12}$$

where D is a diagonal matrix in which the j-*th* ($j > 1$) term $D_j = (1/\bar{Q}_j)$, and $D_1 = 1$. The estimator $\hat{\beta}^{M_1}$ has therefore a higher variance than $\hat{\beta}^{M_2}$.

4.4.3.b Deduce $\hat{\beta}^K$ involved in the predictive model M_3

Remind that K denotes a quality pattern. We have seen in Section 4.3.2 that the vector $\hat{\beta}^K$ exactly matches $\hat{\beta}^{M_1}$ when all individual quality indexes equal 1, i.e. when $K = J_{1,p+1}$. In full generality, when $K = \mathbf{Q}_i$ is made of terms $Q_{ij} \neq 1$, the vector $\hat{\beta}^K$ of individualized regression coefficients need to be calculated. They thus provide an approximation of some correction of β^{M_2} needed to take into account the individual quality pattern $\mathbf{Q}_i = K$. This way, they minimize $RMSE(\hat{\beta}^K|\mathbf{X}, Y)$ for a given quality pattern K, as required by Equation (4.4).

Corollary 1

From Theorem 1, for any quality pattern $K = (Q_{ij})_{j=1,...,p}$ where $Q_{ij} \in [0,1]$:

$$\hat{\beta}_j^K = Q_{ij}\,\hat{\beta}_j^{M_1} = \frac{Q_{ij}}{\bar{Q}_j}\,\hat{\beta}_j^{M_2}, \qquad j = 1, ..., p. \tag{4.13}$$

For the intercept, we have $\hat{\beta}_0^K = \hat{\beta}_0^{M_1} = \hat{\beta}_0^{M_2}$.

4.4.4 Pairwise correlated covariates

The case of total independence between covariates is often not realistic in real-life. One thus needs to extend our results to the setting where some correlation between X_j^{real} and X_k^{real} exists. In this view, we study the case of pairwise correlated explanatory variables. For a brief discussion on more general cases, the reader is referred to Appendix 4.8. As previously, we first detail the modification on the regression coefficients themselves.

4.4.4.a Relation between M_1 and M_2 coefficients

The following theorem enables to make the link between the estimators obtained when considering the quality indexes or not in the regression model. To lighten the formulas, let's write $\rho = \rho_{jk}^{real}$.

Theorem 2

For all $j \neq k$, if $|\rho| = |\rho_{jk}^{real}| \neq 1$, we have under (X-A2) and as $n \to +\infty$:

Under (Z-A1) :

$$\frac{1}{1-\rho^2}\left(\frac{\hat{\beta}_j^{M_2}}{\bar{Q}_j}(1-\rho^2\,\bar{Q}_j\bar{Q}_k) + \sqrt{\frac{Var(X_k)}{Var(X_j)}}\frac{\hat{\beta}_k^{M_2}}{\bar{Q}_k}\rho\,(\bar{Q}_j\bar{Q}_k - 1)\right) \xrightarrow{a.s.} \beta_j,$$

Under (Z-A2) :

$$\frac{1}{1-\rho^2}\left(\frac{\hat{\beta}_j^{M_2}}{\bar{Q}_j}(1-\rho^2\,(1+2\bar{Q}_j\bar{Q}_k - \bar{Q}_j - \bar{Q}_k)) + \sqrt{\frac{Var(X_k)}{Var(X_j)}}\frac{\hat{\beta}_k^{M_2}}{\bar{Q}_k}\rho\,(2\bar{Q}_j\bar{Q}_k - \bar{Q}_j - \bar{Q}_k)\right) \xrightarrow{a.s.} \beta_j. \tag{4.14}$$

See the proof 4 at the section 4.7.

The relation thus depends on both the correlation between the two covariates and their respective qualities. The shift of the intercept remains the same as under (X-A1). In the same spirit as previously, $\hat{\beta}^{M_1}$ is actually used instead of $\hat{\beta}$ in practice.

Remark : Under (X-A1) and (X-A2), we know that the regression coefficients $\hat{\beta}^{M_2}$ are asymptotically unbiased. Their variance is well known, given by

$$Var(\hat{\beta}^{M_2}) = \sigma^2((1/n)\mathbf{X}^T\mathbf{X})^{-1}.$$

$$\text{Then, each block of matrix tends to :} \quad \hat{\sigma}^2\frac{(1-\rho^2)}{(1-\rho^2 Q_j Q_k)}(\Sigma_{jk}^{real})^{-1} \circ \begin{bmatrix} 1 & Q_jQ_k \\ Q_jQ_k & 1 \end{bmatrix}.$$

By using (4.22) in 4, knowing ρ, we get

$$Var(\hat{\beta}_j^{M_1}) = \frac{\hat{\sigma}^2}{(1-\rho^2)}\Bigg(\frac{(1-\rho^2\bar{Q}_j\bar{Q}_k)}{\bar{Q}_j^2}(\Sigma_{j;j}^{real})^{-1} + \frac{Var(X_k)}{Var(X_j)}\frac{(\bar{Q}_j\bar{Q}_k - 1)^2\rho^2(1-\rho^2\bar{Q}_j\bar{Q}_k)}{\bar{Q}_k^2}(\Sigma_{k;k}^{real})^{-1} + 2(\bar{Q}_j\bar{Q}_k - 1)\rho\sqrt{\frac{Var(X_k)}{Var(X_j)}}(\Sigma_{j;k}^{real})^{-1}\Bigg).$$

where $(\Sigma_{j;k}^{real})^{-1}$ is the element of j^{th} row and k^{th} column of the matrix $(\Sigma^{real})^{-1}$. Here, the interpretation is slightly more complex. Indeed, it depends on the correlation structure and the quality relation between the correlated covariates.

4.4.4.b Deduce $\hat{\beta}^K$ to make adjusted predictions in M_3

Remind that X_j^{real} and X_k^{real} are correlated covariates. In the same spirit as in Section 4.4.3.b, we now derive the best coefficients $\hat{\beta}^K$, adapted to the quality pattern $K = \mathbf{Q}_i = (Q_{ij})_{j=1,\dots,p}$.

Corollary 2

Under (X-A2) and (Z-A1) and from Theorem 2, we have

$$\hat{\beta}_j^K = \frac{Q_{ij}}{1 - Q_{ij}^2 Q_{ik}^2 \rho^2} \left(\hat{\beta}_j^{M_1} (1 - Q_{ik}^2 \rho^2) + \sqrt{\frac{Var(X_k)}{Var(X_j)}} \hat{\beta}_k^{M_1} \rho (1 - Q_{ik}^2) \right). \tag{4.15}$$

4.5 Simulations with data quality inputs

We aim here to check our theoretical results. More precisely, we would like to confirm the correction factors needed to unbias the estimation of the regression coefficients when taking into account data quality. In this view, all the simulated examples are designed using a common matrix $\mathbf{Q}$ representing the individual quality indexes, where $\mathbf{Q}$ is drawn once for all from some specified distribution (in real-life, $\mathbf{Q}$ would be given as input). Each simulated dataset is now built as follows :
Step 1 : $\mathbf{X}^{real}$ is simulated given the marginals and its corresponding correlation structure (either under assumption (X-A1) or (X-A2)) ;
Step 2 : $\mathbf{Z}$ is simulated given the marginal distributions of $\mathbf{X}^{real}$ and its correlation structure (either under assumption (Z-A1) or (Z-A2)) ;
Step 3 : $\mathbf{\Omega}$ is simulated from $\mathbf{Q}$ through Bernoulli trials ;
Step 4 : $\mathbf{X}$ is derived from the latent variable model, see Equation (4.1) ;
Step 5 : Y is deduced from the specified linear relationship with $\mathbf{X}^{real}$.

All the study has been performed using R open source statistical software, see R Core Team, 2019 [107].

4.5.1 Estimators properties, case of correlated covariates

Consider the theoretical model

$$Y = 50 + 4X_1^{real} + 4X_2^{real} + \epsilon, \tag{4.16}$$

with ϵ such that $\epsilon \sim \mathcal{N}(0, \sqrt{5})$, and where X_1^{real} and X_2^{real} are correlated covariates respectively following $\mathcal{N}(2, \sqrt{5})$ and $\Gamma(2, 3)$. The correlation structure is given by a Gaussian copula with parameter $\rho = 0.2$. Given that X_1^{real} and X_2^{real} are known since they are simulated (1000 times), the theoretical model (4.16) and the corresponding regression coefficients ($\beta = (\beta_0, \beta_1, \beta_2) = (50, 4, 4)$) can thus be estimated, leading to the vector $\hat{\beta}$. In real-life, X_1^{real} and X_2^{real} are obviously unknown. We only observe n *iid* replications of X_1 and X_2, *i.e.* $X_1 = (X_{i1})_{i=1,\dots,n}$ and $X_2 = (X_{i2})_{i=1,\dots,n}$, where each observation is built from $X_{ij} = \omega_{ij} X_{ij}^{real} + (1 - \omega_{ij}) Z_{ij}$ with $\omega_{ij} \sim \mathcal{B}ernoulli(Q_{ij})$ $(j = 1, 2)$.

In this illustration, the $(Q_{ij})_{i=1,\dots,n;j=1,2}$'s are drawn from independent discrete uniform distributions on {0,0.5,1}. A standard regression model, fitted on $(Y, \mathbf{X})$, should thus come up with estimators impacted by the quality. As can be seen in Table 4.1 where a focus is made on the estimator of β_1 (knowing that similar results apply to β_2), the results show that the model M_2 strongly underestimates the real impact of X_1 on the response. Indeed, over the 1000 simulated datasets, the mean of the estimator $\hat{\beta}_1^{M_2}$ is close to two times lower than expected (about 2.2 instead of 4). On the contrary, the coefficient $\hat{\beta}^{M_1}$ has been rightly corrected using Theorem 2. The bigger the sample, the lower the difference between the estimator $\hat{\beta}_1^{M_1}$ and β_1, showing that the asymptotic regime has to be reached to guarantee good approximations. Figure 4.2 represents the estimators of β_1 in the different models, where one can see that means and medians are overlapping, and where the 5% and 95%-quantiles are plotted to represent the variability of the estimators. Once again, the asymptotic properties reveal obvious as the confidence intervals narrow when the sample size increases. Not surprisingly (see Remark 4.4.3.a), Figure 4.2 also shows that the estimators in model M_1 have higher variances than others, which was already numerically

observed in Table 4.1. Moreover, Figure 4.3 confirms that the maximum likelihood properties still seem to hold (asymptotically gaussian unbiased coefficients), which was expected since translations applied to Gaussian distributions do not change their behaviour.

We are now interested in the importance of the level of correlation between the covariates. The same regression coefficients have thus been inferred with different correlation levels, corresponding to parameters of the Gaussian copula varying from -0.8 to 0.8. From Table 4.2, note that the main lesson to be learnt is that the variance of $\hat{\beta}^{M_1}$ seems to be significantly increasing when the correlation between covariates gets high (especially from $|\rho| = 0.8$). Concerning the results in expectation for the random variables $\hat{\beta}_1^{M_1}$ and $\hat{\beta}_1^{M_2}$, we observe the same conclusions as previously.

4.5.2 Interpretation and prediction

Closer to a real-life context, we now simulate a unique dataset $\mathbf{X}$ and use bootstrap resampling (with 500 bootstrap samples) to get the distribution of the estimators. Once again, $\mathbf{Q}$ remains unchanged. Each bootstrap sample is splitted into two independent subsamples (learning : 70%, and test : 30%) in order to evaluate the predictive power of models.

4.5.2.a Very strongly correlated variables : impact on the estimation of coefficients

Let be given the model

$$Y = 10 + X_1^{real} + X_2^{real} + \epsilon, \tag{4.17}$$

with ϵ such that $\epsilon \sim \mathcal{N}(0, \sqrt{5})$, and where X_1^{real} and X_2^{real} are correlated random variables respectively following $\Gamma(2,1)$ and $\Gamma(2,3)$ distributions. The correlation structure between X_1^{real} and X_2^{real} is defined by a Gaussian copula, with correlation ρ. To begin with, we set $\rho = 0.7$. Moreover, assume that the qualities Q_1 and Q_2 follow independent continuous uniform distributions, such that $Q_1 \sim Q_2 \sim \mathcal{U}(0.7, 1)$.

We follow the same estimation procedure to get the estimators of interest, more precisely their bootstrapped distribution. Figure 4.4 shows the two bootstrapped distribution of the regression coefficients $\hat{\beta}_1^{M_2}$ and $\hat{\beta}_2^{M_2}$ (obtained fitting model M_2, dotted area), as well as those of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ (plain area). Here, because the covariates are strongly correlated, the lost information on X_1^{real} due to imperfect quality is partially offset by X_2. Therefore, the regression coefficients have not necessarily decreased, contrary to what one could anticipate (remember the discussion about the *attenuation* phenomenon, see Section 4.1). Indeed, notice that the expectation of $\hat{\beta}_1^{M_2}$ roughly equals 0.8 whereas that of $\hat{\beta}_2^{M_2}$ is close to 1.4 (remind that $(\beta_1, \beta_2) = (1, 1)$). This is an important message to avoid wrong conclusions in terms of interpretation.

4.5.2.b Predictive power using low quality observations

Secondly, we aim to check the predictive power of the proposed approach, see Algorithm 1 in Section 4.3.3. Here, we go one step further since the predictive power is assessed through the comparison

	Estimator mean			Estimator variance		
n	$\hat{\beta}_1$	$\hat{\beta}_1^{M_2}$	$\hat{\beta}_1^{M_1}$	$\hat{\beta}_1$	$\hat{\beta}_1^{M_2}$	$\hat{\beta}_1^{M_1}$
100	3.993	2.572	4.149	0.1216	0.6724	1.2472
500	4.001	2.124	3.933	0.0603	0.3069	0.6743
1000	4.001	2.213	4.021	0.0419	0.2124	0.4466
5000	4.000	2.229	4.068	0.0173	0.0998	0.2060
7500	4.001	2.232	3.990	0.0133	0.0810	0.1675

TABLE 4.1 – Estimators in models M_1 and M_2. Means and variances are obtained from 1000 simulations.

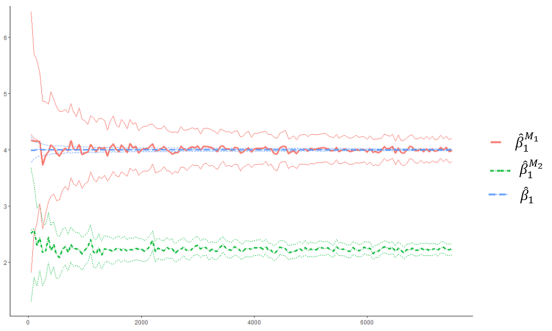


FIGURE 4.2 – Asymptotic consistency of estimators in M_1 and M_2 models (1000 simulations), with 90%-confidence bands.

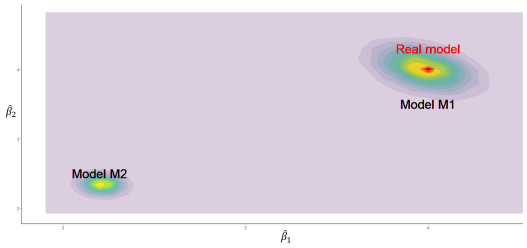


FIGURE 4.3 – Density of the estimators of β_1 and β_2 in models M_1 and M_2 (n = 7500 with 1,000,000 simulations).

	Estimator mean			Estimator variance		
ρ	$\hat{\beta}_1$	$\hat{\beta}_1^{M_2}$	$\hat{\beta}_1^{M_1}$	$\hat{\beta}_1$	$\hat{\beta}_1^{M_2}$	$\hat{\beta}_1^{M_1}$
-0.8	4.000	0.771	3.967	0.0188	0.0390	0.2026
-0.6	4.000	1.132	3.972	0.0150	0.0475	0.1474
-0.4	4.000	1.452	3.978	0.0134	0.0542	0.1266
-0.2	4.000	1.736	3.981	0.0127	0.0598	0.1208
0	4.000	1.985	3.970	0.0121	0.0602	0.1204
0.2	4.000	2.211	3.978	0.0123	0.0649	0.1363
0.4	4.000	2.412	3.978	0.0130	0.0699	0.1761
0.6	4.000	2.593	3.976	0.0162	0.0756	0.2621
0.8	4.000	2.752	3.964	0.0184	0.0822	0.4980

TABLE 4.2 – Estimators of β_1. Empirical means and variances obtained from 1000 simulations (with n = 10 000).

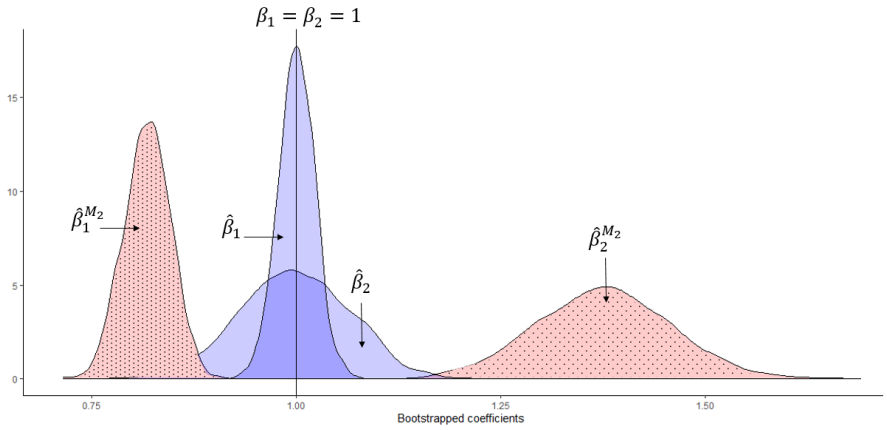


FIGURE 4.4 – Distribution of bootstrapped coefficients in the classical model (based on $\mathbf{X}^{real}$ and in M_2 (based on $\mathbf{X}$). Q_j is drawn from a $\mathcal{U}niform(0.7, 1)$ for $j \in 1, 2$, and $\rho_{1,2} = 0.7$.

between predicted and observed responses on the validation set. This means that the individualized regression coefficients $\hat{\beta}^K$, associated to the quality pattern K, are computed and used for prediction. Suppose that $\rho = 0.3$ and that $\mathbf{Q}$ follows a Uniform discrete distribution, with $Q_{ij} \sim \mathcal{U}(\{0, 0.3, 0.7, 1\})$. Denote by $RMSE_k$ the Residual Mean Squared Error (RMSE) of model M_k, and $RMSE_{real}$ the one of the real model.

Figure 4.5 illustrates the bootstrapped RMSE evaluated on each model under study. On the one hand, the estimator of the theoretical model (4.17) - fitted on $\mathbf{X}^{real}$ - is of course the best one according to the RMSE metric. On the other hand, model M_1 - using $\hat{\beta}^{M_1}$ - is fitted on the observed dataset (M_2 is first fitted, and then corrected to find M_1), but the RMSE is still computed using $\mathbf{X}^{real}$ to ensure comparability (mandatory to use $\hat{\beta}$ and $\hat{\beta}^{M_1}$). Not surprisingly, Figure 4.5 confirms that the real model (using $\hat{\beta}$ evaluated thanks to $\mathbf{X}^{real}$) is the best one. Nonetheless, because there is not a significant gap between $\hat{\beta}$ and $\hat{\beta}^{M_1}$, $RMSE_1$ is very close to $RMSE_{real}$ ($(RMSE_1 - RMSE_{real})/RMSE_{real} \leq 0.5\%$). Indeed, both are estimators of β. In practice, as $\mathbf{X}^{real}$ is unknown, the estimator $\hat{\beta}$ cannot be used for prediction on $\mathbf{X}$. $RMSE_2$ and $RMSE_3$ (using $\mathbf{X}$) are significantly higher than $RMSE_1$, due to the imperfectly observed dataset. More precisely, the loss of RMSE due to the quality is the difference between $RMSE_{real} = RMSE(\hat{\beta}|\mathbf{X}^{real})$ and $RMSE_2 = RMSE(\hat{\beta}^{M_2}|\mathbf{X})$. In this simulation, $RMSE_2$ is 94% higher than $RMSE_{real}$. Using $\mathbf{Q}$ helps to recover a significant amount of information corresponding to the difference between $RMSE_2$ and $RMSE_3$. In this example, this amounts to 21% ($(RMSE_2 - RMSE_3)/RMSE_{real}$). Moreover, this improvement depends on the quality pattern K for each individual. Indeed, model M_3 provides each individual with the best estimator corresponding to its quality indexes, as illustrated by Figure 4.6. In Figure 4.6, the comparison between observed and predicted responses for each individual are wrapped by quality patterns. As expected, model M_2 is invariant to the quality pattern, and model M_3 converges to the real model when the vector (Q_1, Q_2) gets closer to $(1, 1)$. Also, M_3 converges to a constant when (Q_1, Q_2) gets closer to $(0, 0)$.

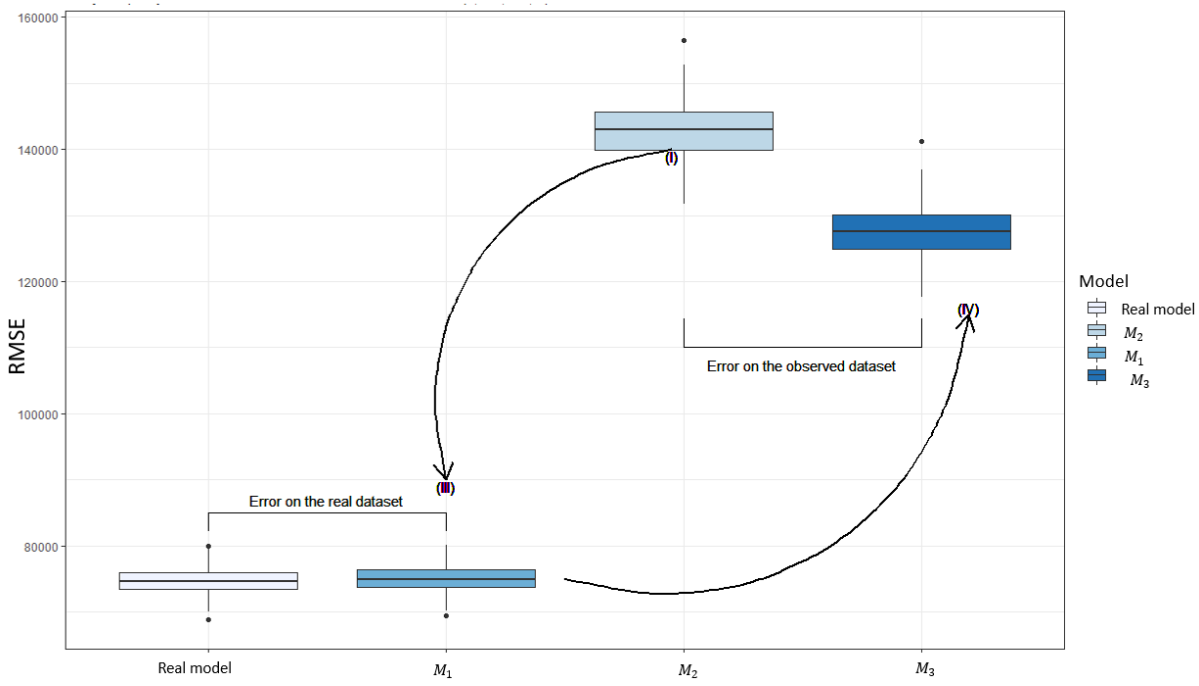


FIGURE 4.5 – Dispersion of RMSE (500 bootstrap samples, $n = 7000$, $Q_{ij} \sim \mathcal{U}(\{0, 0.3, 0.7, 1\})$.

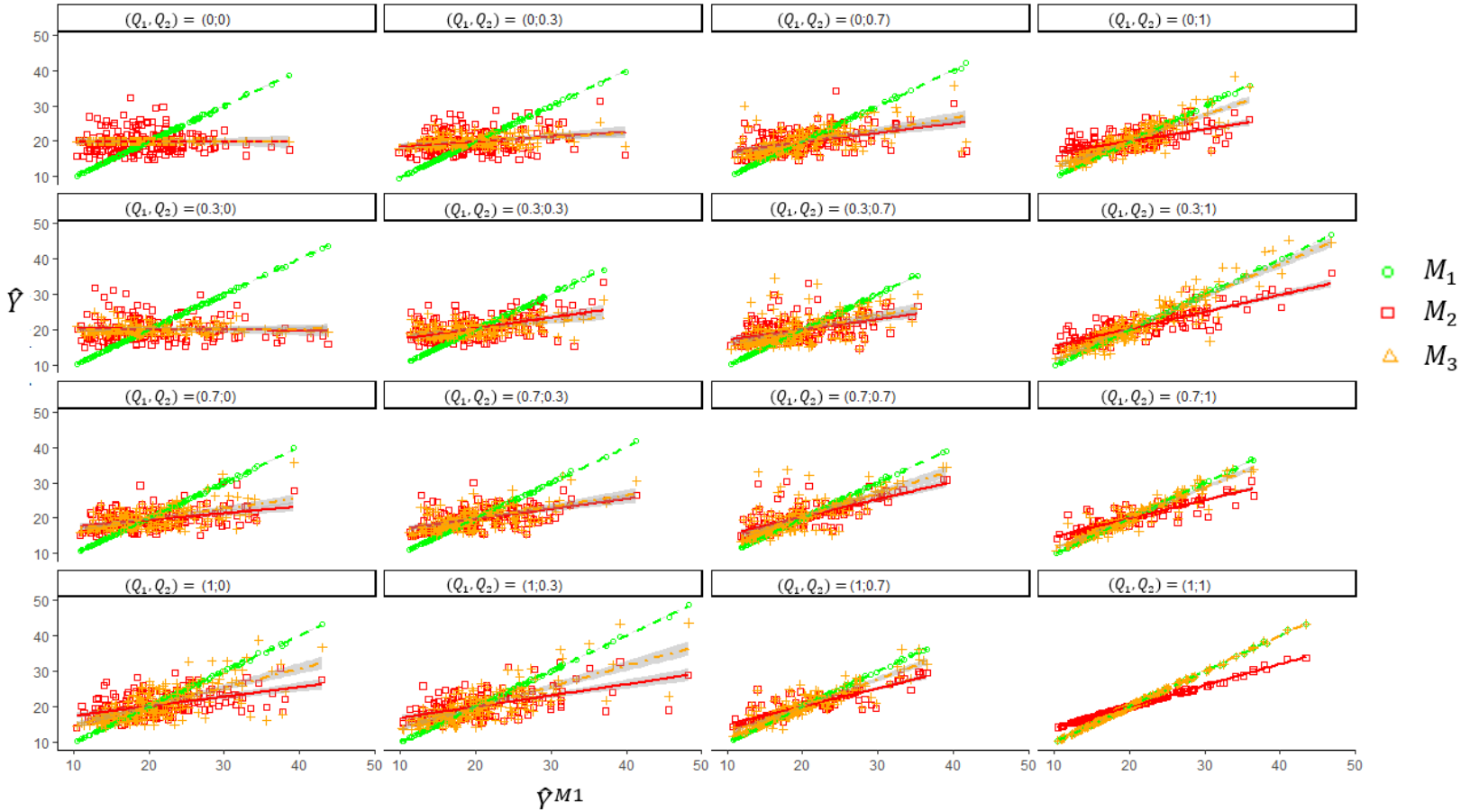


FIGURE 4.6 – Predictions by quality pattern. The quality is taken from a uniform sampling between (0, 0.3, 0.7, 1). ρ is taken equal to 0.3 in order to show the amelioration by pattern of quality. The green points correspond to the predicted value using M_1 with $\mathbf{X}^{real}$. The red squares correspond to the predicted value using M_2 with $\mathbf{X}$. The orange crosses correspond to the predicted value using M_3 with $\mathbf{X}$.

4.6 Real-life applications

4.6.1 Imputation and the link to missing values

Let us now consider a more complex model, given by

$$Y = 10 + 4X_1^{real} + 2X_2^{real} + 3X_3^{real} + 4X_4^{real} + \epsilon,$$

where ϵ is a Gaussian random noise such that $\epsilon \sim \mathcal{N}(0, \sqrt{5})$, and $X_1^{real}, X_2^{real}, X_3^{real}$ and X_4^{real} are correlated random variables respectively following $\mathcal{N}(2, \sqrt{5})$, $\mathcal{E}(2)$, $\Gamma(2,1)$ and $\Gamma(2,3)$ distributions. The correlation structure is defined by a Gaussian copula, with correlation matrix

$$\begin{bmatrix} 1 & 0.3 & 0 & 0 \\ 0.3 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.6 \\ 0 & 0 & 0.6 & 1 \end{bmatrix}.$$

The corresponding quality indexes Q_1, Q_2, Q_3 and Q_4 follow independent discrete uniform distributions in $\{0;1\}$. When $Q_{ij} = 0$, it is similar to a missing value replaced by a random observation drawn from the empirical distribution of X_j^{real}. Algorithm 1 initially determines M_1 regression coefficients, and then adapt these coefficients to the different quality patterns, as shown in Figure 4.7. Little by little, the predictions deteriorate when the number of missing values increase. If all the covariates have missing values, the model M_3 predicts the mean of Y, which is indeed the best estimator without any other information than the response. In line with Section 4.5.2.b, Figure 4.8 confirms that the RMSE is lower using M_3 than using M_2, whatever the quality pattern considered. However the lower the quality, the lower the model performance.

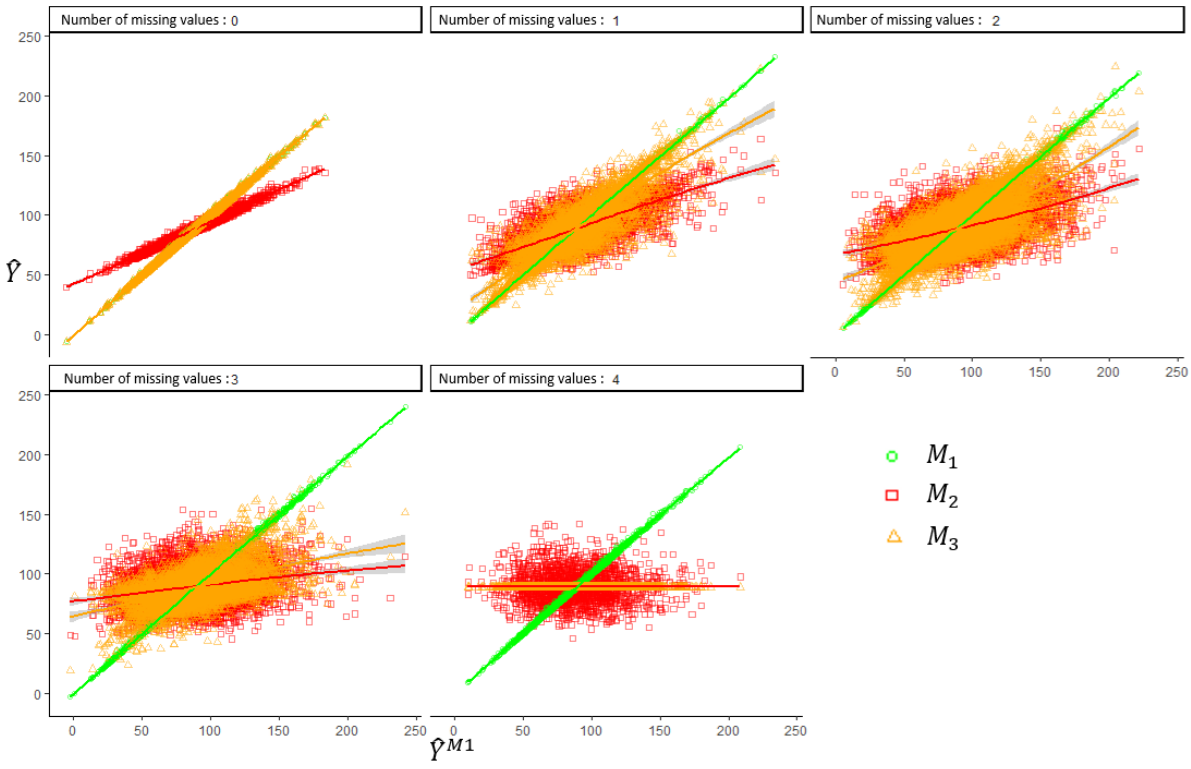


FIGURE 4.7 – Predictions wrapped by number of missing values under (X-A2). Predicted values by M_2 (red squares) and M_1 model using $\hat{\beta}^{M_1}$ (green circles) against the true values ones. Orange triangles are the predictions using the model M_3.

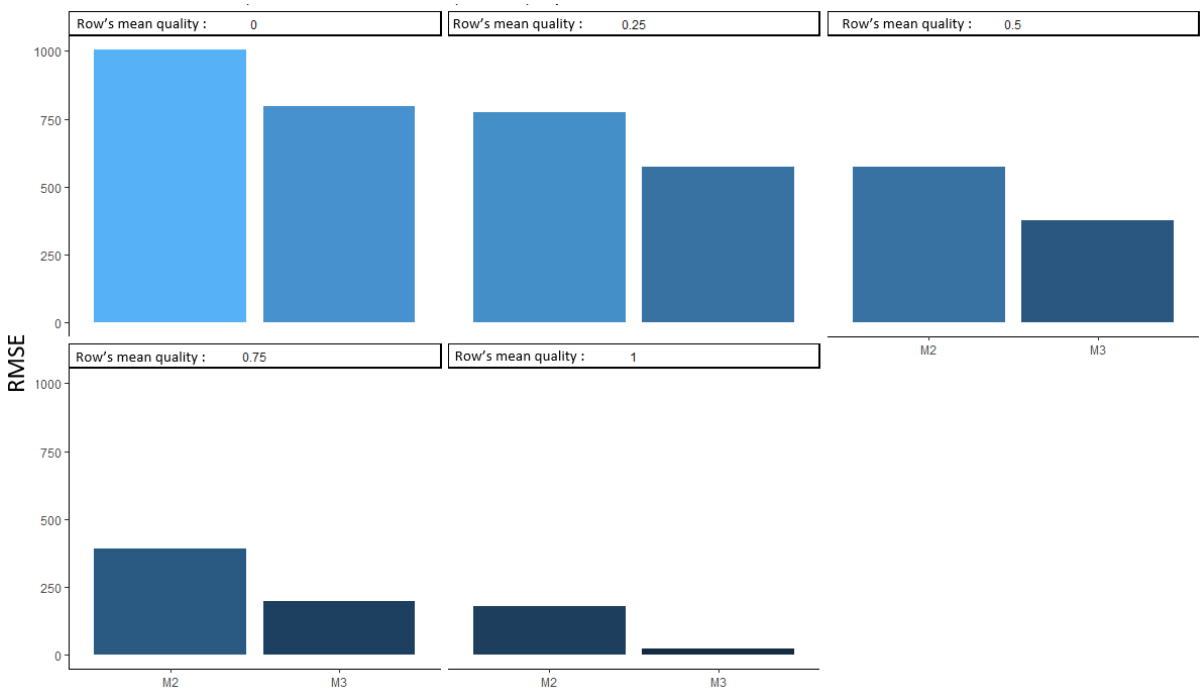


FIGURE 4.8 – RMSE plot wrapped according to $(1/p)\sum_{j=1}^{p} Q_{ij}$ for each individual i. $RMSE_3$ is always lower than $RMSE_2$ whatever the quality pattern.

4.6.2 Household insurance : imputation at the underwriting process

We consider hereafter a real-life portfolio with more than 500 000 insurance contracts. On the French household insurance market, the premium is typically based on the living surface. We aim here to recover the living surface (represented by $X_{liv.surf.}$) for policyholders underwriting household insurance contracts, using both the information about the heated surface ($X_{heat.surf.}$) and the number of rooms (X_{rooms}). That is to say,

$$\mathbb{E}[X_{liv.surf.} \,|\, X_{heat.surf.}, X_{rooms}] = \beta_0 + \beta_1 X_{heat.surf.} + \beta_2 X_{rooms},$$

where X_{rooms} and $X_{heat.surf.}$ are subjected to unperfect quality, and are generated from $X^{real}_{rooms}, Z_{rooms}, X^{real}_{heat.surf.}$ and $Z_{heat.surf.}$. The two covariates X^{real}_{rooms} and $X^{real}_{heat.surf.}$ are correlated, which justifies (X-A2) assumption. This correlation is estimated to 0.28. Conversely, X_{rooms} and $X_{heat.surf.}$ are collected from unrelated sources, leading to Assumption (Z-A1).

On the one hand, X_{rooms} is asked at the underwriting process. On the other hand, $X_{heat.surf.}$ is inferred from external sources related to house geolocalization. The living surface and the heated surface for each policyholder are known in practice, but the number of rooms is missing for 50% of the observations. This means that the variable X_{rooms} is associated to a quality variable Q_{rooms}, which can take values 0 or 1. Concerning the heated surface, we know that this information is not perfect for many reasons ; including wrong geocoding and the fact that it is built from several inputs like number of floors and footprint surface. Therefore, $X_{heat.surf.}$ comes together with a quality variable $Q_{heat.surf.}$, where $Q_{heat.surf.}$ can take values "Very High", "High", and "Medium" in our context. At the end, we will be able to compare the predicted values of the living surface to the actual ones that were observed.

First, we need to associate the modalities of $Q_{heat.surf.}$ to numeric values using Remark 1. Setting $Q_{heat.surf.} = 1$ for cases when $Q_{heat.surf.}$ equals “Very High", the values corresponding to other modalities can be derived. For instance, considering that Y is perfectly observed,

$$\frac{Cov(X_{heat.surf.}, Y \,|\, Q_{heat.surf.} = "High")}{Cov(X_{heat.surf.}, Y \,|\, Q_{heat.surf.} = "VeryHigh")} = \frac{q \times Cov(X^{real}_{heat.surf.}, Y)}{1 \times Cov(X^{real}_{heat.surf.}, Y)} = q.$$

The value q thus corresponds to the numerical value associated to the modality "High" of $Q_{heat.surf.}$. The variable Y should represent the living surface. However, to avoid overfitting, we cannot recover q like

Models	$\beta_0^{K=0.81}$	$\beta_0^{K=0.91}$	$\beta_0^{K=1}$	β_0	$\beta_1^{K=0.81}$	$\beta_1^{K=0.91}$	$\beta_1^{K=1}$	β_1
M_3^{bis}	58.8499 ***	51.2914 ***	45.0096 ***	-	0.5851 ***	0.6619 ***	0.7216 ***	-
M_2	-	-	-	50.4782***	-	-	-	0.67024***
M_1	-	-	-	45.1524	-	-	-	0.7346
M_3	56.6961	50.6204	45.1524	-	0.5950	0.6680	0.7346	-

TABLE 4.3 – Comparison of obtained regression coefficients depending on the modelling under consideration. (***) = p.value < 0.05 (no p.value for M_1 or M_3 models).

this (since q is an input for the future linear model explaining the living surface). Anyway, any variable with perfect quality can be used. For example, Y could stand for the number of rooms, since it has a perfect quality when available. Finally the modalities of $Q_{heat.surf.}$ ("Very High", "High", "Medium") are associated respectively to values (1, 0.91, 0.81).

Beforehand, only 429 607 contracts of the initial database were kept to match the assumption that $X^{real}_{heat.surf.}$ and $Z_{heat.surf.}$ have the same distribution. To do so, we ensured that the distribution of $X_{heat.surf.}$ remains similar whatever the selected subpopulation corresponding to the different modalities of $Q_{heat.surf.}$. In practice, note that this leads to keep houses where $X_{heat.surf.}$ lies in between 30 and 100 square meters ($\bar{X}_{heat.surf.} = 80.67$), and where $X_{liv.surf.}$ is lower than 200 square meters ($\bar{X}_{liv.surf.} = 105.3$).

Firstly, we consider the simplified modelling

$$\mathbb{E}[X_{liv.surf.} \mid X_{rooms}, X_{heat.surf.}] = \beta_0 + \beta_1 X_{heat.surf.}.$$

We fit M_2 on a training sample (322 412 observations), and use the algorithm M_3. The remaining 107 195 observations (test set) are used for generalization performance metrics, as the well-known R^2_{test} coefficient representing the percentage of explained variance. To check that our procedure works well, consider the model M_3^{bis}, given by

$$\begin{aligned} X_{liv.surf.} = {} & (\beta_0^{K=1} + \beta_1^{K=1} X_{heat.surf.})\mathbf{1}_{Q_{heat.surf.}=1} + (\beta_0^{K=0.91} + \beta_1^{K=0.91} X_{heat.surf.})\mathbf{1}_{Q_{heat.surf.}=0.91} \\ & + (\beta_0^{K=0.81} + \beta_1^{K=0.81} X_{heat.surf.})\mathbf{1}_{Q_{heat.surf.}=0.81} + \epsilon. \end{aligned}$$

This model minimizes the MSE for each quality pattern, and should estimate the same regression coefficients as those obtained using model M_3. Nevertheless, unlike M_3, model M_3^{bis} requires to fit as many models as the number of quality patterns.

Table 4.3 shows the results obtained when fitting M_3^{bis}. Notice that the obtained coefficients of M_3^{bis} and M_3 are very similar, as expected. The estimated coefficients of model M_1 are given, but cannot be used in practice (recall that $\mathbf{X}^{real}$ is unknown). In terms of prediction performance on the test set, the model M_3 is better than M_2. Indeed, the coefficient R^2_{test} is higher (0.06485 against 0.06464). The coefficient R^2_{test} can also be computed on subpopulations, according to their quality pattern. Respectively for "Very High", "High" and "Medium" quality index, the coefficient R^2_{test} of M_2 equals 0.08626, 0.061016, 0.04216 (against 0.08708, 0.061016, 0.04254 in M_3). The further the quality index from the mean quality ($\bar{Q}_{heat.surf.} \approx 0.91$), the better the improvement.

Let us now move to the full model, given by

$$\mathbb{E}[X_{liv.surf.} \mid X_{rooms}, X_{heat.surf.}] = \beta_0 + \beta_1 X_{heat.surf.} + \beta_2 X_{rooms}.$$

Table 4.4 shows the significant improvement of the indicator R^2_{test} by quality pattern, between models M_2 and M_3. To improve the performance for the entire data set, the model M_2 lowers the performance of each pattern quality leading to negative R^2_{test} for some quality patterns. M_3 corrects the bias and increases the R^2_{test} about 0,10 for all quality patterns. In the same spirit as previously, we introduce the model M_3^{bis} which is optimal for the R^2_{test} metric by quality pattern (when the number of observations is sufficient). However, the smaller the data set, the less robust M_3^{bis}. Indeed, model M_3^{bis} estimates 18 coefficients, and not 3 coefficients as for M_2 or M_3. For smaller datasets, the results are less volatile using M_3. To illustrate this, each model is trained on 5000 houses that define the training sample. This training set is bootstrapped 5000 times. The distribution of R^2_{test} is then obtained (always considering the same test set), and Figure 4.9 confirms our previous statement.

Finally, contrary to M_3^{bis}, M_3 can extrapolate to new pattern of quality. Suppose that new data are provided but with some observations associated with "Medium" quality or "Low" quality. Using the previous dataset, Algorithm M_3 is still able to extrapolate the best coefficient for the "Low" quality data. On the contrary, M_3^{bis} is unable to do so. In this example, authors had to filter low-quality data, which are

	$Q_{heat.surf.}$ - 1		$Q_{heat.surf.}$ - 0.91		$Q_{heat.surf.}$ - 0.81	
	Q_{rooms} - 1	Q_{rooms} - 0	Q_{rooms} - 1	Q_{rooms} - 0	Q_{rooms} - 1	Q_{rooms} - 0
R^2_{test} - M_2	0.3642	-0.0225	0.3484	-0.0507	0.3383	-0.0522
R^2_{test} - M_3	0.4743	0.0743	0.4537	0.0592	0.4678	0.03960
R^2_{test} - M_3^{bis}	0.4743	0.0743	0.4554	0.0593	0.4740	0.04399
Nb. obs. (test)	10774	10759	36182	36082	8518	8516

TABLE 4.4 – Indicator R^2_{test} by quality pattern.

very often linked to rural areas. Indeed, the living surface depends on the population density, therefore the exogenous information of the modality "Very Low" is correlated with $X_{liv.surf.}$.

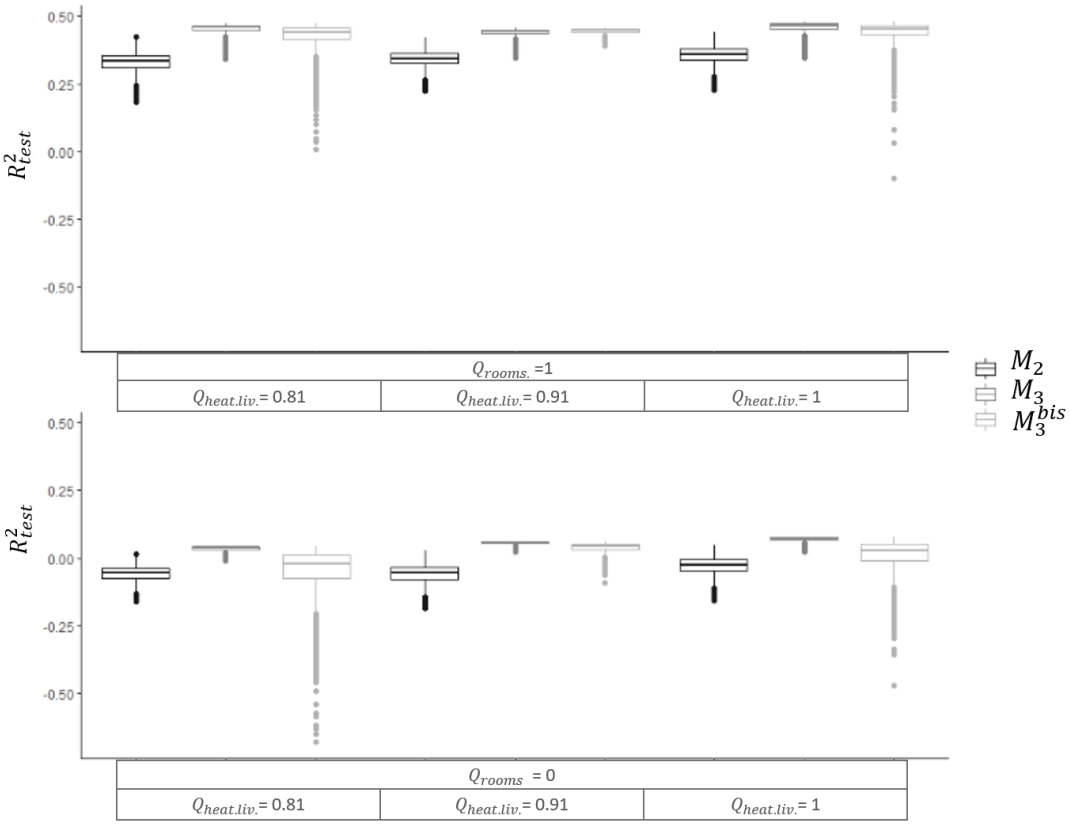


FIGURE 4.9 – Comparison between M_2, M_3 and M_3^{bis} on a smaller training dataset. The R^2_{test} metric is evaluated by quality pattern.

Conclusion

In this work, we introduce a new method to take into account individualized quality indexes in linear regressions. These indexes can be either qualitative or quantitative indicators, including the case of missing values. Our paper shows in an easy way how to deal with such information, relying on theoretical results. Basically, one modifies the original estimated regression coefficient by applying a correction factor that integrates the quality information. These results are very likely to be useful for practitioners, knowing that more and more studies use open data.

Further developments would be needed. First, some of our assumptions could be relaxed. Second, one could add dependency between the errors, or apply this framework to generalized linear models. Other considerations like the case of regularized regressions (Ridge, LASSO) lead to wide open questions, and would be worth to be explored. These are left for future research.

Acknowledgements The authors would like to thank the firm and the data provider for the data used in Section 4.6.2.

4.7 Proofs

4.7.1 Proof of Lemma 1

Proof 1

Thanks to the equation (4.1), the expected value remains unchanged. Consider two covariates and their joint quality (X_j, Q_j) , (X_k, Q_k) $(j \neq k)$. The qualities Q_j and Q_k corresponds to the expectation of the quality variables Ω_j and Ω_k, where Ω_j and Ω_k are Bernoulli random variables. Introduce Ω_{jk} = (Ω_j, Ω_k) and denote w_m the realization of Ω_{jk}.

Clearly, $w_m \in \{(0,0),(0,1),(1,0),(1,1)\}$. We can therefore write :

$$Cov_{jk} = \sum_{w_m} \Big[Cov(X_j, X_k|\Omega_{jk} = w_m) + \mathbb{E}(X_j|\Omega_{jk} = w_m)\mathbb{E}(X_k|\Omega_{jk} = w_m)\Big]\mathbb{P}(\Omega_{jk} = w_m) - \underbrace{\mathbb{E}(X_j)}_{=0}\underbrace{\mathbb{E}(X_k)}_{=0},$$

$$\begin{aligned}
Cov_{jk} &= \mathbb{P}(\Omega_{jk} = (0,0))\Big(Cov(X_j, X_k|\Omega_{jk} = (0,0)) + \mathbb{E}(X_j|\Omega_{jk} = (0,0))\mathbb{E}(X_k|\Omega_{jk} = (0,0))\Big) \\
&\quad + \mathbb{P}(\Omega_{jk} = (1,0))\Big(Cov(X_j, X_k|\Omega_{jk} = (1,0)) + \mathbb{E}(X_j|\Omega_{jk} = (1,0))\mathbb{E}(X_k|\Omega_{jk} = (1,0))\Big) \\
&\quad + \mathbb{P}(\Omega_{jk} = (0,1))\Big(Cov(X_j, X_k|\Omega_{jk} = (0,1)) + \mathbb{E}(X_j|\Omega_{jk} = (0,1))\mathbb{E}(X_k|\Omega_{jk} = (0,1))\Big) \\
&\quad + \mathbb{P}(\Omega_{jk} = (1,1))\Big(Cov(X_j, X_k|\Omega_{jk} = (1,1)) + \mathbb{E}(X_j|\Omega_{jk} = (1,1))\mathbb{E}(X_k|\Omega_{jk} = (1,1))\Big) \\
&= \mathbb{P}(\Omega_j = 0)\ \mathbb{P}(\Omega_k = 0)\ Cov(Z_j, Z_k) + \mathbb{P}(\Omega_j = 1)\ \mathbb{P}(\Omega_k = 0)\ Cov(X_j^{real}, Z_k) \\
&\quad + \mathbb{P}(\Omega_j = 0)\ \mathbb{P}(\Omega_k = 1)\ Cov(Z_j, X_k^{real}) + \mathbb{P}(\Omega_j = 1)\ \mathbb{P}(\Omega_k = 1)\ Cov(X_j^{real}, X_k^{real}).
\end{aligned}$$

The final expression depends on the assumptions about the correlation structure between the covariates,

$$\begin{aligned}
Cov_{jk} &= Q_jQ_kCov_{jk}^{real} + Q_j(1 - Q_k)Cov(X_j^{real}, Z_k) \\
&\quad + (1 - Q_j)Q_kCov(X_k^{real}, Z_j) + (1 - Q_j)(1 - Q_k)Cov(Z_k, Z_j).
\end{aligned}$$

Under (X-A2) and (Z-A1), $Cov(Z_j, Z_k) = 0$. We always have in our framework $Cov(X_j^{real}, Z_k) = 0$. We thus obtain :

$$Cov_{jk} = Q_jQ_kCov_{jk}^{real}.$$

Similarly, under (X-A2) and (Z-A2), the covariance can be written,

$$Cov_{jk} = (1 + 2Q_jQ_k - Q_j - Q_k)\ Cov_{jk}^{real}.$$

Using the Kolmogorov's strong law of large number, $\bar{Q}_j$ and $\bar{Q}_k$ converges almost surely to Q_j and Q_k. □

4.7.2 Proof of Lemma 2

Proof 2

For each covariate X_j ($j = 1, ..., p$), Z_j has the same distribution as X_j^{real}. Given the latent variable model (equation 4.1) and under (X-A1) and (Z-A1), it is straightforward to show that :

$$\mathbb{E}(X_j) = \mathbb{E}(X_j^{real}),$$
$$Var(X_j) = Var(X_j^{real}).$$
$$Cov(X_j, X_k) = Cov(X_j^{real}, X_k^{real}) = 0.$$

Then, the covariance matrix Σ_{jk} equals to :

$$\Sigma_{jk} = \begin{bmatrix} Var(X_j) & 0 \\ 0 & Var(X_k) \end{bmatrix} = \begin{bmatrix} Var(X_j^{real}) & 0 \\ 0 & Var(X_k^{real}) \end{bmatrix} = \Sigma_{jk}^{real}.$$

Let's focus now on the covariance matrix Σ_{jk} under (X-A2) and (Z-A1). As mentioned before, the variance remains unchanged. To lighten the equations, let write $Var(X_j^{real}) = Var_j^{real}$. However, we know from Lemma 1 that the covariance is changing proportionally to the mean quality under (Z-A1) :

$$Cov_{jk} = Q_j Q_k \, Cov_{jk}^{real}.$$

Recall that the quality Q_j corresponds to $\mathbb{E}(\Omega_j)$ for $j = 1, ..., p$, where Ω_j is a Bernoulli random variable. We write $cor(X_k^{real}, Y)$ the Pearson correlation here between X_k^{real} and Y. We assumed the non singularity of the real information matrix, i.e. $|cor(X_j^{real}, X_k^{real})| \neq 1$ (and thus $|cor(X_j, X_k)| \neq 1$) and notice that we state

$$(\Sigma_{jk}^{real})^{-1} = \frac{1}{Var_j^{real} Var_k^{real} - (Cov_{jk}^{real})^2} \begin{bmatrix} Var_j^{real} & -Cov_{jk}^{real} \\ -Cov_{jk}^{real} & Var_k^{real} \end{bmatrix},$$

$$\Sigma_{jk}^{-1} = \frac{1}{Var(X_j)Var(X_k) - Q_j^2 Q_k^2 (Cov_{jk}^{real})^2} \begin{bmatrix} Var(X_j) & -Q_j Q_k Cov_{jk}^{real} \\ -Q_j Q_k Cov_{jk}^{real} & Var(X_k) \end{bmatrix}.$$

To lighten the notation, we denote ρ_{jk}^{real} the Pearson correlation between the two covariates X_j^{real} and X_k^{real}. To find a relation between $(\Sigma_{jk}^{real})^{-1}$ and Σ_{jk}^{-1}, denote $(\Sigma^{-1})_{kk}$ the k^{th} diagonal term and $(\Sigma^{-1})_{jk}$ the element on the j^{th} row and k^{th} column. We can now easily state the following relations :

$$(\Sigma^{-1})_{kk} - (\Sigma^{real})_{kk}^{-1} = -(\Sigma^{real})_{kk}^{-1} \times \frac{(\rho_{jk}^{real})^2 (1 - Q_j^2 Q_k^2)}{(1 - (\rho_{jk}^{real})^2 Q_j^2 Q_k^2)},$$

$$(\Sigma^{-1})_{jk} - (\Sigma^{real})_{jk}^{-1} = -(\Sigma^{real})_{jk}^{-1} \times \Big(1 - \frac{Q_j^2 Q_k^2 \big(1 - (\rho_{jk}^{real})^2\big)}{(1 - (\rho_{jk}^{real})^2 Q_j^2 Q_k^2)}\Big),$$

leading to :

$$\Sigma_{jk}^{-1} = (\Sigma_{jk}^{real})^{-1} \circ \begin{bmatrix} \frac{(1-(\rho_{jk}^{real})^2)}{(1-(\rho_{jk}^{real})^2 Q_j^2 Q_k^2)} & Q_j Q_k \frac{(1-(\rho_{jk}^{real})^2)}{(1-(\rho_{jk}^{real})^2 Q_j^2 Q_k^2)} \\ Q_j Q_k \frac{(1-(\rho_{jk}^{real})^2)}{(1-(\rho_{jk}^{real})^2 Q_j^2 Q_k^2)} & \frac{(1-(\rho_{jk}^{real})^2)}{(1-(\rho_{jk}^{real})^2 Q_j^2 Q_k^2)} \end{bmatrix}$$
$$= \frac{(1 - (\rho_{jk}^{real})^2)}{(1 - (\rho_{jk}^{real})^2 Q_j^2 Q_k^2)} (\Sigma_{jk}^{real})^{-1} \circ \begin{bmatrix} 1 & Q_j Q_k \\ Q_j Q_k & 1 \end{bmatrix}.$$

□

4.7.3 Proof of Theorem 1

Proof 3

Recall that the covariates are supposed centered and that we are under (X-A1). Using our notations, the classical OLS regression coefficient $\hat{\beta}$ satisfies :

$$ {}^t\mathbf{X}^{real}\mathbf{X}^{real}\hat{\beta} = {}^t\mathbf{X}^{real}Y \iff (1/n)\,{}^t\mathbf{X}^{real}\mathbf{X}^{real}\hat{\beta} = (1/n)\,{}^t\mathbf{X}^{real}Y. $$

As the sample size tends to infinity, β is the solution of

$$ \begin{aligned} \Sigma\beta &= {}^tCov(\mathbf{X}^{real}, Y), \\ \beta &= \frac{Cov(\mathbf{X}^{real}, Y)}{Var(\mathbf{X}^{real})}. \end{aligned} $$

When focusing on M_2 , we have for $j = 1, ..., p$:

$$ \beta_j^{M_2} = \frac{1}{Var(X_j)} \times Cov(X_j, Y), \forall j \in 1, ..., p. \tag{4.18} $$

According to Lemma 2, $\Sigma = \Sigma^{real}$, i.e, $Var(X_j) = Var(X_j^{real})$. Moreover, using the Lemma 1,

$$ Cov(X_j, Y) = Q_j Cov(X_j^{real}, Y) = Q_j\beta_j Var(X_j^{real}). $$

Recall that the quality Q_j corresponds to $\mathbb{E}(\Omega_j)$, where Ω_j is a Bernoulli random variable. Finally, the difference easily follows :

$$ \beta_j^{M_2} - \beta_j = \beta_j \times (Q_j - 1) \iff \beta_j^{M_2} = \beta_j Q_j. \tag{4.19} $$

If X_j^{real} is not centered, the intercept changes by $-\mathbb{E}(X_j)\frac{\beta_j^{M_2}(1-Q_j)}{Q_j}$ for each covariate X_j. The other coefficients stay unchanged by centering. Indeed, without loss of generality, consider $\mathbb{E}(X_1) \neq 0$ and $X_1^c = X_1 - \mathbb{E}(X_1)$. The shift is easily found

$$ \begin{aligned} \mathbb{E}(Y|\mathbf{X}) &= \beta_0 + \beta_1 X_1 + \sum_{2\leq j\leq p} X_j\beta_j, \\ \mathbb{E}(Y|\mathbf{X}) &= \beta_0 + \underbrace{\mathbb{E}(X_1)\beta_1}_{\perp\!\!\!\perp X_j, \forall j} + \beta_1 X_1^c + \sum_{2\leq j\leq p} X_j\beta_j. \end{aligned} $$

Therefore, only the intercept in the centered case shifts (by $\mathbb{E}(X_1)\beta_1$) due to X_1 centering. We will first center the variable and uncenter it afterward. We center first the variable X_1 for the model M_2. The intercept shifts by $\mathbb{E}(X_1)\beta_1^{M_2}$ then we can apply the previous results in the centered case. Finally, we recenter the variable X_1 for the real model, with a shift of $\mathbb{E}(X_1)\beta_1$. Therefore, the global shift is equal to $\mathbb{E}(X_1)\beta_1^{M_2} - \mathbb{E}(X_1)\beta_1$. This part of proof for the intercept ends, replacing β_1 by $\frac{\beta_1^{M_2}}{Q_j}$.

To end the proof, $\bar{Q}_j$ and $\hat{\beta}_j$ converges almost surely to β_j and Q_j using SLLN and the maximum likelihood properties. The $\hat{\beta}^{M_2}$ calculated in Equation (4.18) with the empircal estimators converges almost surely. Indeed, the Kolmogorov's SLLN ensures the converge *a.s.* of each estimators : variance and covariance. The continuous mapping theorem and the fact that the product of series converging *a.s.* converges *a.s.* ends to achieve the proof of the convergence *a.s.*. Therefore,

$$ \hat{\beta}_j^{M_2}/\bar{Q}_j \xrightarrow{a.s.} \beta_j^{M_2}/Q_j = \beta_j, \tag{4.20} $$

Remark that no gaussian properties on $\hat{\beta}_j^{M_2}$ can be state, the errors not being gaussian. □

4.7.4 Proof of Theorem 2

Proof 4

Recall that the covariates are supposed centered. The proof with uncentered covariates would only modify the intercept values (see the previous proof 3). The ordinary OLS regression coefficient $\hat{\beta}$ satisfies :

$$ {}^t\mathbf{X}\mathbf{X}\hat{\beta} = {}^t\mathbf{X}Y \iff \frac{1}{n}\,{}^t\mathbf{X}\mathbf{X}\hat{\beta} = \frac{1}{n}\,{}^t\mathbf{X}Y. $$

As n tends to infinity, the regression coefficient β is the solution of

$$ \Sigma\beta = {}^t\mathbb{C}ov(X, Y). $$

To lighten the notation, for two correlated, covariates X_k, X_j, $k \neq j$ lets denote $\rho = \rho_{jk}^{real}$. Then using Gauss-Jordan elimination, for two correlated covariates, $|\rho| \neq 1$, X_k, X_j, $k \neq j$ in linear regression, we can state :

$$ \begin{aligned} \beta_k &= \frac{cor(X_k^{real}, Y) - \rho\, cor(X_j^{real}, Y)}{1-\rho^2} \times \sqrt{\frac{Var(Y)}{Var_k^{real}}}, \\ \beta_j &= \frac{cor(X_j^{real}, Y) - \rho\, cor(X_k^{real}, Y)}{1-\rho^2} \times \sqrt{\frac{Var(Y)}{Var_j^{real}}}. \end{aligned} $$

Thanks to the Lemma 1, under the assumption (X-A2) and (Z-A1), we can write

$$ \begin{aligned} \beta_k^{M_2} &= \frac{Q_k cor(X_k^{real}, Y) - Q_j^2 Q_k \rho\, cor(X_j^{real}, Y)}{1-Q_j^2Q_k^2\rho^2} \times \sqrt{\frac{Var(Y)}{Var_k^{real}}}, \\ \beta_j^{M_2} &= \frac{Q_j cor(X_j^{real}, Y) - Q_k^2 Q_j \rho\, cor(X_k^{real}, Y)}{1-Q_j^2Q_k^2\rho^2} \times \sqrt{\frac{Var(Y)}{Var_j^{real}}}. \end{aligned} $$

Recall that the quality Q_j corresponds to $\mathbb{E}(\Omega_j)$ for $j = 1, ..., p$, where Ω_j is a Bernoulli random variable. Using the Cramer system and $D \neq 0$ due to the assumption $\{i|q_{ij} \neq 0\} \neq \emptyset$ for $j = 1, ..., n$, the system can be solved,

$$ \begin{aligned} cor(X_k^{real}, Y) &= \frac{1}{D}\,\mathrm{Det}\begin{vmatrix} \gamma \times b_k^{M_2} & -Q_jQ_k^2\rho \\ \gamma \times b_j^{M_2} & Q_j \end{vmatrix}, \\ cor(X_j^{real}, Y) &= \frac{1}{D}\,\mathrm{Det}\begin{vmatrix} Q_k & \gamma \times b_k^{M_2} \\ -Q_kQ_j^2\rho & \gamma \times b_j^{M_2} \end{vmatrix}, \end{aligned} $$

where $\gamma = 1 - \rho^2Q_j^2Q_k^2$, $b_k^{M_2} = \sqrt{\frac{Var_k^{real}}{Var(Y)}}\beta_k^{M_2}$ and $D = \mathrm{Det}\begin{vmatrix} Q_k & -Q_j^2Q_k\rho \\ -Q_k^2Q_j\rho & Q_j \end{vmatrix}$. By simpliying,

$$ cor(X_k^{real}, Y) = \frac{b_k^{M_2}}{Q_k} + \rho\, Q_jQ_k\, \frac{b_j^{M_2}}{Q_j}. \tag{4.21} $$

A relation between β and β^{M_2} immediately follows :

$$\begin{aligned}\beta_k &= \frac{1}{1-\rho^2}\left(\frac{b_k^{M_2}}{Q_k} + \rho\, Q_jQ_k\,\frac{b_j^{M_2}}{Q_j} - \rho\frac{b_j^{M_2}}{Q_j} - \rho^2\, Q_jQ_k\,\frac{b_k^{M_2}}{Q_k}\right)\sqrt{\frac{Var(Y)}{Var_k^{real}}}\\ &= \frac{1}{1-\rho^2}\left(\frac{b_k^{M_2}}{Q_k}(1-\rho^2\, Q_jQ_k) + \frac{b_j^{M_2}}{Q_j}\rho\,(Q_jQ_k-1)\right)\sqrt{\frac{Var(Y)}{Var_k^{real}}}\\ &= \frac{1}{1-\rho^2}\left(\frac{\beta_k^{M_2}}{Q_k}(1-\rho^2\, Q_jQ_k) + \sqrt{\frac{Var_j^{real}}{Var_k^{real}}}\frac{\beta_j^{M_2}}{Q_j}\rho\,(Q_jQ_k-1)\right).\end{aligned} \tag{4.22}$$

The relation between β and β^{M_2} depends on the correlation between the two variables. The shift of the intercept is the same as in the (X - A1) statement. By replacing the values with the corresponding estimator, (SLLN and the continuous mapping theorem ensuring the convergence almost surely),

$$\frac{1}{1-\rho^2}\left(\frac{\hat{\beta}_j^{M_2}}{\bar{Q}_j}(1-\rho^2\,\bar{Q}_j\bar{Q}_k) + \sqrt{\frac{Var(X_k)}{Var(X_j)}}\frac{\hat{\beta}_k^{M_2}}{\bar{Q}_k}\rho\,(\bar{Q}_j\bar{Q}_k-1)\right)\xrightarrow{a.s.}\beta_j$$

which ends the proof. (The proof under (Z-A2) is done in the same way.) In the other way round for the Corollary 4.15, for given mean quality indexes, (Q_k, Q_j), we can find β^{M_2} according to the β.

$$\beta_k^{M_2} = \frac{Q_k}{1-Q_j^2Q_k^2\rho^2}(\beta_k(1-Q_j^2\rho^2) + \sqrt{\frac{Var_j^{real}}{Var_k^{real}}}\beta_j\rho(1-Q_j^2)).$$

Indeed, in the same way than Equation (4.21), with similar notations,

$$cor(X_k, Y) = b_k + \rho\, b_j.$$

$$\text{Then : } \beta_k^{M_2} = \left(\frac{Q_k\, b_k + Q_k\,\rho\, b_j}{1-Q_j^2Q_k^2\rho^2} - Q_k^2Q_j\rho\; cor(X_k^{real}, Y)\frac{b_j+\rho\, b_k}{1-Q_j^2Q_k^2\rho^2}\right)\times\sqrt{\frac{Var(Y)}{Var_k^{real}}}. \tag{4.23}$$

$$\beta_k^{M_2} = \frac{Q_k}{1-Q_j^2Q_k^2\rho^2}\left(\beta_k\times(1-Q_j^2\rho^2) + \sqrt{\frac{Var_j^{real}}{Var_k^{real}}}\,\beta_k\,\times\rho(1-Q_j^2)\right). \tag{4.24}$$

(4.25)

To end the proof, the different values are replaced by their empirical estimator. SLLN and the continuous mapping theorem ensure the convergence almost surely of the estimator of last equation . □

4.8 Multivariate case

Until now, we have studied the case of pairwise correlated covariates. The ordinary OLS regression coefficient $\hat{\beta}$ follows :

$${}^t\mathbf{X}\mathbf{X}\hat{\beta} = {}^t\mathbf{X}Y \iff (1/n)\,{}^t\mathbf{X}\mathbf{X}\hat{\beta} = (1/n)\,{}^t\mathbf{X}Y.$$

As n tends to ∞, β is the solution of

$$\Sigma\beta = {}^tCov(X, Y).$$

If Σ is invertible, different methods exist as the Gauss-Jordan elimination to find a solution. However, one could remark that the relation between $\beta_k^{M_2}$ and β_k depends only on the Pearson correlation ρ_{jk}^{real} and Q_j and Q_k for all covariates k correlated to covariate k. The different proofs on the OLS coefficient could be extended with this method.

Bibliography

[99] Berglund, L., Garmo, H., Lindbäck, J., Svärdsudd, K., and Zethelius, B. (2008). Maximum likelihood estimation of correction for dilution bias in simple linear regression using replicates from subjects with extreme first measurements. *Statistics in Medicine*, 27(22) :4397–4407.

[100] Decker, H. and Martinenghi, D. (2009). Modeling, measuring and monitoring the quality of information. In *International Conference on Conceptual Modeling*, pages 212–221. Springer.

[101] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6) :381–395.

[102] Fuller, W. A. (2009). *Measurement error models*, volume 305. John Wiley & Sons.

[103] Hausman, J. (2001). Mismeasured variables in econometric analysis : problems from the right and problems from the left. *Journal of Economic perspectives*, 15(4) :57–67.

[104] Heitjan, D. F. and Basu, S. (1996). Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3) :207–213.

[105] Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

[106] Muzellec, B., Josse, J., Boyer, C., and Cuturi, M. (2020). Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR.

[107] R Core Team (2019). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[108] Ramakrishnan, R. and Gehrke, J. (2000). *Database management systems*. McGraw Hill.

[109] Rogova, G. L. and Bosse, E. (2010). Information quality in information fusion. In *2010 13th International Conference on Information Fusion*, pages 1–8. IEEE.

[110] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3) :581–592.

[111] Tami, M., Clausel, M., Devijver, E., Dulac, A., Gaussier, E., Janaqi, S., and Chebre, M. (2018). Uncertain trees : Dealing with uncertain inputs in regression trees. *arXiv preprint arXiv :1810.11698*.

[112] Todoran, I.-G., Lecornu, L., Khenchaf, A., and Le Caillec, J.-M. (2014). Toward the quality evaluation of complex information systems. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*, volume 9091, page 90910N. International Society for Optics and Photonics.

[113] Trabelsi, A., Elouedi, Z., and Lefevre, E. (2016). Handling uncertain attribute values in decision tree classifier using the belief function theory. In *International conference on artificial intelligence : Methodology, systems, and applications*, pages 26–35. Springer.

[114] Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

[115] Van Huffel, S. and Lemmerling, P. (2013). *Total least squares and errors-in-variables modeling : analysis, algorithms and applications*. Springer Science & Business Media.

[116] Wang, R. Y., Reddy, M. P., and Kon, H. B. (1995). Toward quality data : An attribute-based approach. *Decision support systems*, 13(3-4) :349–372.

Chapitre 5

Modèles linéaires généralisés et la qualité de données

Ce chapitre reprend un article Integrating data quality into GLM for insurance pricing co-écrit avec Stéphane Loisel.

Abstract Pricing models and regulatory work done by actuaries incorporate more and more external data provided by data providers. The reliability of these external data needs to be investigated since all aspects of regression are impacted by data quality. Therefore, actuaries need to address this notion of quality. In this paper, the impact of data credibility on GLMs is studied. The latter is measured by an exogenous and individualized quality index. Under the hypothesis that inconsistent data have the same distribution as consistent data, a method to find the true impact of the variables on the predictor is proposed. Under several assumptions, we use this method to adapt the prediction depending on each data quality value. Operational remarks and actuarial applications illustrate the creation and use of quality indexes.
keywords : Credibility, Quality index, Generalized Linear Model

5.1 Introduction

Traditionally, actuarial pricing was limited by the number of variables used and their complexity. Indeed, variables available to actuaries originate from the underwriting process. Potential customers have limited knowledge and time to complete a questionnaire. To offset this problem and to improve risk knowledge, insurance companies use external data, for which the reliability is debatable. External data come from a third party, and the users cannot interfere in the data creation process. The data reliability depends on each observation within the same variable. Indeed, gathering processes often aggregate data sets from various sources with heterogeneous quality. Legally, insurers' entities are responsible for data quality (articles 219, 237, 244, 245, 247 from Solvency II Commission Delegated Regulation (EU) 2015/35). Their work must assess and justify the data quality even if the data are coming from a third party : *'Data used in the internal model obtained from a third party shall not be considered to be appropriate unless the insurance or reinsurance undertaking is able to demonstrate a detailed understanding of those data, including their limitations'*, article 237. In France, the French Prudential Supervision and Resolution Authority, ACPR, [117, ACPR 2011] states that 10% of the data come from external parties for economic capital calculations or for pricing purposes. The data quality does not have a negligible impact, as illustrated by Campbell [119, Campbell et al., 2006] relating several actuarial examples. Therefore, to assess and account for the data quality issue, actions must be initiated. These various notions of quality have already been discussed for actuarial purposes in exploratory cases on the North American side ([123, Francis, 2005]) or on the UK side, [119, Campbell et al., 2006] for instance. To the best of our knowledge, advice to consider data quality is still very qualitative ([124, GCASB, 2014]) with such basic recommendations as : deleting, imputing, or correcting the problem. These solutions are discussed but are not sufficient. Models depend on observations' quality, and the latter is reflected by quality indexes given by the data provider. How can an individualized and exogenous quality index be used for prediction ?

The literature suggests a multiple dimension analysis to evaluate data quality ([135, todoran, 2014]). For instance, the completeness dimension of data is a research field where numerous methods have been developed to address missing values ([137, Van Buuren, 2018], [130, Little and Rubin, 2019]). These methods are based on assumptions such as MCAR (Missing Completely At Random), MAR (Missing At Random), or MNAR (Missing Not At Random). In the present paper, the credibility dimension is studied further. On the mismeasurement side of the uncertainty dimension, some research exist that uses a tree algorithm ([136, Trabelsi et al., 2016], [134, Tami et al. 2018]) or the EIV-mismeasurement ([138, Van Huffel and Lemmerling, 2013]) framework. On the credibility side of the uncertainty dimension, robust estimation theory such as the RANSAC (RANdom SAmple Consensus, [122, Fischler and Bolles, 1981]) algorithm and its different extensions such as KALMANSAC [139, Vedaldi et al., 2005] deal with outliers and inliers that appear in computer vision research. The drawback of these methods is the left-aside observations for which no prediction can be made. It is operationally unthinkable that some contracts are not priced.

In our framework, the credibility of observations is quantified and called the quality index. It is assumed to be perfectly measured. Observational uncertainty is modeled by a latent variable model. In this work, quality indexes are exogenous, individualized, and equal to the probability that the observation is the true one. Indeed, this framework derives from work with a data provider. In different investigations, the data provider delivers data and quality indexes associated with it. For example, a goal was to price household insurance contracts by using building geolocation and external data. During this work, it was clear that the given quality indexes evaluated each observation's credibility more than its precision. These quality indexes are exogenous (given by the data provider). The framework and assumptions developed in this paper arise from this case.

The main assumption is that wrong observations have the same distribution as the empirical ones. Under this strong, but necessary assumption, [120, Chatelain and Milhaud, 2021] considered the case of a basic linear regression and the associated correlation matrices. Since GLMs are preferred in the insurance industry, GLM cases are studied through likelihood. The goal is to give a precise answer to the following question. Given an individualized quality index (here based on the credibility dimension), how can this quality index be used in a multivariate **GLM**? How can actuaries set up a pricing model with a variable that has quality problems?

Contributions : This paper presents two main contributions. First, it shows how to consider quality indexes in a GLM regression. After explaining why basic recommendations do not work, we provide methods and theoretical results for most GLM frameworks commonly used in actuarial science. These results are summarized in Table 2 in Subsection 3.6 and proved in the Appendix. We also provide a new algorithm to improve prediction by mitigating data quality patterns in three different operational cases. Second, some operational and practical notes are given to help actuaries create and use the quality indexes in practice, from a general point of view and thanks to some illustrations with real-world household insurance dataset and geolocalisation data.

Outline of the paper : The paper is structured as follows : section 5.2 introduces the general framework and the notation. We specify how uncertainty is integrated into the covariate generating process. Section 5.3 gives the main algorithms and theoretical results to find the model that is unbiased by the data quality. Section 5.4 explains adapting the prediction to the data quality. Hereafter, a simulation study illustrates the results in Section 5.5. Section 5.6 demonstrates the various assumptions for actuarial uses. In detail, subsections 5.6.4 and 5.6.3 discuss the use of quality indexes and the case of imperfect data quality indexes. All these remarks are illustrated on household insurance for which this framework was originally created. As the paper already contains a lot of notation due to the nature of our problem, we relegate some technical aspects regarding regularity conditions, the form and the optimization of the log-likelihood as well as convergence results in section 5.7. Various operational remarks are given afterwards, focusing on log-Gaussian case. Proofs of main results can be found in the last section.

5.2 Data problems and imputation

5.2.1 Notations

The set of all $n \times m$ matrices, where all the elements are in the interval I, is denoted $\mathcal{M}_{n\times m}(I)$. p represents the number of variables without the intercept, and n represents the number of rows. Data are important :

— $\mathbf{X} = (X_{ij}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$: the data set available with data quality problems *i.e.* observed covariates ;

— $\mathbf{X}^{real} = (X_{ij}^{real}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$: the data set, in practice not available, corresponding to the "real" observations.

We want to take advantage of the exogenous information provided by an *individualized* quality index related to the confidence we can have about the $i-th$ observation of the $j-th$ covariate, further denoted Q_{ij}.

In this view, we introduce the following latent variable model :

$$\mathbf{X} = \mathbf{X}^{real} \circ \mathbf{\Omega} + \mathbf{Z} \circ (J_{n,(p+1)} - \mathbf{\Omega}), \tag{5.1}$$

where :

— $\circ$ corresponds to the Hadamard product,

— $J_{n,(p+1)}$ is the $n \times (p+1)$-identity matrix under Hadamard multiplication,

— $\mathbf{Z} = (Z_{ij}) \in \mathcal{M}_{n\times(p+1)}(\mathbb{R})$ are considered the "wrong" covariate values having the same distribution as $\mathbf{X}^{real}$,

— $\mathbf{\Omega} = (\omega_{ij}) \in \mathcal{M}_{n\times(p+1)}(\{0,1\})$ is a binary mask indicating whether the $i-th$ observation of the $j-th$ covariate X_{ij} is perfectly observed or not. In other words, $\mathbf{\Omega}$ tells us if one observes the "real" observation or not. We assume that the covariate distribution has a finite second moment.

In practice, the data at hand are made of individualized quality indexes through some matrix $Q = (Q_{ij}) \in \mathcal{M}_{n\times(p+1)}([0,1])$, together with n *i.i.d* replications $(Y_i, X_i)_{i=1,..,n}$ of $(Y, \mathbf{X})$, where $Y_i \in \mathbb{R}$ and $\mathbf{X}_i = (1, X_{i1}, ..., X_{ip}) \in \mathbb{R}^{p+1}$. The vector of quality indexes of the i-th row is written as $\mathbf{Q}_i = (1, Q_{i1}, ..., Q_{in})$.

A vector of specific values of quality indexes is called a *quality pattern* associated with the notation K. Each element Q_{ij} of the matrix Q informs us of the quality related to the observed covariate value X_{ij}. Equation 5.1 can be adapted for j in $1, ..., n$,

$$\mathbf{X_j} = X_j^{real} \times \Omega_j + Z_j \times (1 - \Omega_j). \tag{5.2}$$

We consider as Q the expectation of $\mathbf{\Omega}$, which leads to defining the quality index as a credibility index. This means that for all $i = 1, ..., n$, $j = 1, ..., p$, the quality index Q_{ij} is equal to :

$$Q_{ij} = \mathbb{E}(\omega_{ij}) = \mathbb{P}(\omega_{ij} = 1) = \mathbb{P}(X_{ij} = X_{ij}^{real}) - \mathbb{P}(Z_j = X_{ij}^{real}). \tag{5.3}$$

The quality index corresponds to the probability to have taken not the "correct" but the "real" observation. The part "$-\mathbb{P}(Z_j = X_{ij}^{real})$" corresponds to the probability of obtaining the true value randomly. In other words, $Q_{ij} = 0$ means that the value X_{ij} is not informative of the risk of i. We denote for the rest of the paper ($j = 1, ..., p$), $\bar{Q}_j = \frac{1}{n}\sum_{i=1}^{n} Q_{ij}$ and assume that $\bar{Q}_j \neq 0$. This assumption is not limiting, especially for real-world applications where such covariates are simply removed from the data. However, this assumption does not mean that an individual having all quality indexes null is nonexistent.

In this framework, the singularity is that $\mathbf{X}^{real}$ is not fully observed, which has consequences on the estimation of the regression coefficients.

5.2.2 Inapplicability of basic recommendations

The basic recommendations proposed by various actuarial investigations on deleting and imputing new values on "wrong observations" are not viable solutions for this framework.

Imputating : We consider the strategy to impute new values on outliers or low-quality observations [1]. Defining outliers in the multivariate case when the other covariates are not of good quality is difficult. This is even more true for actuarial pricing, where the risks for modeling have intrinsic variability : claim cost, claim frequency, retention rate, and so on. Without considering exogenous information, robust estimation theory such as the RANSAC (RANdom SAmple Consensus, [122, Fischler and Bolles, 1981]) algorithm and its various extensions have been developed that use only a subsample of "real observation" (inliers) in modeling. Straightforwardly, in our framework, the data quality influences the definition of outliers for regressions, as shown in Figure 5.1. Indeed, the quality of the data set biases the outlier detection. In that situation, some perfectly observed observations are defined as outliers. The goal is also to predict values for individuals with wrong observation(s). In the multivariate case and with an increasing variance outcome, the definition of an outlier is operationally even more complex to address. For instance, in our framework, if $(X_{i,1}, X_{i,2})$ is defined as an outlier, is $X_{i,1}$ or $X_{i,2}$ or both wrong ?

1. Outlier detection and influential values have been studied, for instance, by [125, Hadi, 1991]) or [121, Cook, 1977].

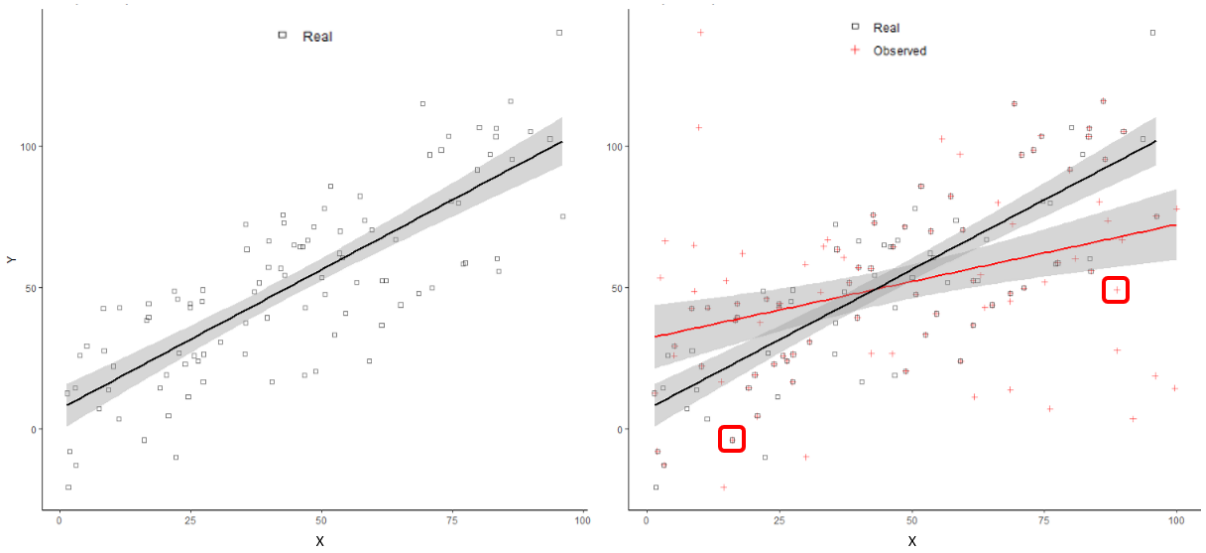


FIGURE 5.1 – A univariate example where black squares are the real observations and crosses are the observed values. This graph is based on simulated data with $X \sim \Gamma(1,2)$ and $Y = 10 + 1X$, and Q follows a uniform distribution between 0.2 and 1. The data set $\mathbf{X}$ is created with the previous framework defined in Subsection 5.2.1. Two linear regressions are fitted : one on the real data set $\mathbf{X}^{real}$ - the square points, and the other on the observed data set $\mathbf{X}$ - the red cross points. Logically, in the second figure, when a cross is not within a square it means that is a wrong observation. Two points are highlighted : a real point that could be considered an outlier and an incorrect value that could be considered an inlier.

Deleting : Given a data set and its joint quality index, a naive workaround is to delete low-quality observations. An easy way is to select a threshold on the quality indexes and to remove individuals having one of their quality indexes below. This solution can hardly be run on some low-quality data or for high-dimensional data sets. Indeed, this latter issue was illustrated by [140, Zhu and al., 2019]. With an independent probability of a value missing equal to 0.05 and 300 covariates, this deleting approach suppresses 95% of the data set.

For our framework, let us assume assumptions similar to Zhu et al. 2019 [140], *i.e.*, in the case of complete independence of quality and observations [2]. We assume that all the quality indexes are independently distributed as a $Uniform(0.4, 0.8)$. Not only does the low quality of the data imply a small threshold, but the various observations can range broadly around the mean value. For a threshold of 0.5 and 10 variables defined as earlier, only 6 % rows on average have all their covariates exceeding the threshold. In addition, errors can be correlated spatially, and this filtering process may bias portfolio risks. For open data used in household insurance, this is particularly true for urban area zones : covariates are often of lower quality in rural areas. In short, filtering strategies are not optimal. Finally, neither imputing nor deleting correct the impact of quality on models.

5.2.3 An illustrative example

Exposure	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	Q_3	Y	Premium
0.6	45	2	454	0.8	350	?
1	30	3	1000	0.6	0	?
1	43	2	2500	0	2450	?
0.2	61	6	245	0.7	0	?
1	53	3	723	1	-	?
1	53	3	723	0.5	-	?

TABLE 5.1 – The four first lines exemplifies a training data set while the two last lines represent a testing data set.

2. From the notations used in this paper : (C1) with the assumptions (X-A1) and (Z-A1).

We consider a simple example : the explanatory variables, (X_1, X_2, X_3) and $\mathbf{Y}$. Here, only the last variable X_3 has an associated individualized quality index Q_3 where $Q_3 \in (0, 1)$. Each $X_{i,3}$ observation has an associated quality index $Q_{i,3}$, which is between 0 and 1 - 1 being an observation of perfect quality and 0 the worst one. Table 5.1 represents a dummy example. Here, X_1 refers to occupant age in years, X_2 to the number of rooms and X_3 to the house value in £ per m^2. Arbitrarily, Y could be the annual claims amount. From a training data set with an imperfect variable, how can actuaries predict the future mean claim cost knowing a value perfectly, or more generally, knowing a value imperfectly ? Here, both the last individuals have the same observed characteristics but not the same quality index. How should the premium differ ?

The **naive way** is to use $(\beta_0, \beta_1, \beta_2, \beta_3)$ such $\mathbb{E}[Y \mid \mathbf{X}] = g^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)$ for g a function. However, to truly understand the underlying relation, one wants to use Q_3 to find the **"Real" model** $\mathbb{E}[Y \mid \mathbf{X}^{real}] = f(\beta_0 + \beta_1 X_1^{real} + \beta_2 X_2^{real} + \beta_3 X_3^{real})$, which can called the model for data of **"perfect quality"**.

First, the index cannot be used as a weight in multivariate regressions. Indeed, the use of weights may bias the regression and does not correct the impact of quality. Second, Table 5.1 displays another problem. If an actuary fits a model with medium-quality observations, how should she or he adapt its prediction for observations for which the covariate value is perfectly known or unknown ?

In our framework, quality indexes are associated with values between 0 and 1, as shown in Example 5.1. In real-life applications, quality indexes are exogenous information given by the data provider and take qualitative values such as "very high", "high", "medium", "low", and "very low".To overcome this issue, the last section of this paper shows how to associate a value to a quality index modality when using our theoretical framework.

5.2.4 Frameworks under study

Several assumptions are examined throughout this section. They are linked with Rubin's nomenclature [131, Rubin, 1976], yet contested by [133, Seaman, 2013]. From Equation (5.1), various cases can be investigated depending on the correlation structure of $(\mathbf{X}^{real}, \mathbf{Z}, \mathbf{\Omega})$. Let us consider the two following situations summarized in Figures 5.2a and 5.2b. These cases depend only on the type of collection of each variable. We suppose that the information brought to the predictor from $\mathbf{Z}$ is not distinct from $\mathbf{X}^{real}$; $\mathbf{Z}$ is informative only through its correlation with $\mathbf{X}^{real}$ on $\mathbf{Y}$.

A discrete variable is considered a sum of Boolean variables in regressions. In between these Boolean variables, the quality variables are equal. Hence, the case (C2) with fully correlated quality variables is a necessary assumption.

5.3 Frameworks studied

5.3.1 Models under study

In this section, we study the following GLM given by

$$\mathbb{E}[Y \mid \mathbf{X}^{real}] = g^{-1}(\mathbf{X}^{real}\beta),$$

and the associated likelihood $\mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{real})$ by using the real data set $\mathbf{X}^{real}$ [3].

In most cases, the previous model is unknown in our framework. Hereafter, this model is called the **"Real" model**.

We denote the following naming :

— M_2 **("Naive" model)** : Model fitted on the observed data set $\mathbf{X}$:

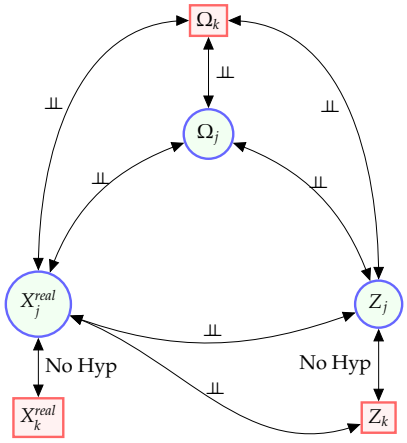
$$\mathbb{E}[Y|\mathbf{X}] = g^{-1}(\mathbf{X}\beta^{M_2}),$$

where $\hat{\beta}^{M_2}$ is the solution of $Argmax_\beta \mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})$. When $\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})$ is estimated by using $\mathbf{X}^{real}$ and $\mathbf{Q}$, we denote it as $\mathcal{L}^{M2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q})$. Let $\hat{\beta}^{M_2|\mathbf{X}^{real},\mathbf{Q}}$ be the solution of $Argmax_\beta \mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q})$.

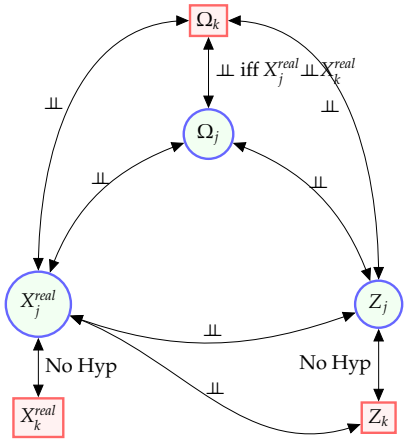
— M_1 **("perfect quality" model)** : Model fitted on the observed data set $\mathbf{X}$, which estimates the coefficient of the real model, β :

$$\mathbb{E}[Y|\mathbf{X}, \mathbf{Q} = J_{n,p+1}] = g^{-1}(\mathbf{X}\beta^{M_1}).$$

3. See Subsection 5.3.3 and the appendices.



(a) Case (C1) - Total uncertainty ($j \neq k$). (No Hyp) means No hypothesis.



(b) (C2) - Local imprecision with unrelated errors.

FIGURE 5.2 – Cases studied.

In our framework, we denote the solution $\hat{\beta}$ as the solution of $Argmax_\beta \mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{real})$ and $\hat{\beta}^{M_1}$ as the solution of $Argmax_\beta \mathcal{L}^{M_1}(\beta; \mathbf{Y}|\mathbf{X}, \mathbf{Q})$ defined in Section 5.3.4. $\mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{real})$ cannot be determined in practice since $\mathbf{X}^{real}$ is not fully observed ;

For all the proofs, the variables are supposed to be centered.

5.3.2 Assumptions under study

We assume that each covariate distribution has a finite second-order moment, and we recall that $Z_j \sim X_j^{real}$ for $j = 1, ..., p$. Here, the discussion concerns the assumptions underlying the correlation structure between the covariates $\mathbf{X}^{real}$, as well as for the random variables $\mathbf{Z}$. Thus, we define the following five assumptions covering various alternative possibilities.

(X-A1) All the random variables X_j^{real} ($j = 1, ..., p$) are independent.
(X-A2) Each variable X_j^{real} is correlated with only one variable X_k^{real} ($j \neq k$).
(X-A3) For all $k \neq p$, the variable X_k^{real} is independent of X_p^{real} and $\bar{Q}_k = 1$.
(Z-A1) All the random variables Z_j and Z_k are independent.
(Z-A2)(Z_j, Z_k) has the same correlation structure as (X_j^{real}, X_k^{real}), $j \neq k$.

For GLM, correlations between imperfectly observed covariates, such as (X-A2), are not considered. However, for linear regression, (X-A2) was considered in [120]. When assumption (X-A3) is studied, we write $\mathbf{X}_{(*p)} = (1, X_1; \cdots; X_{p-1})$ and its observed sample as $\mathbf{X}_{i;(*p)}$. In the same way, $\beta^{(*p)}$ refers to $(\beta_0, \cdots, \beta_{p-1})$.

Recursively, the theorem can be found easily under (X-A1) in the same way as (X-A3). However, the explicit formulas are complex without any particularity.

The choice of the correlation structure of $\mathbf{Z}$ depends only on the data. Based on the same extraction and on the same key (*e.g.*, geocoding), the correlation between two Z_i, Z_j is similar to X_i^{real}, X_j^{real} ones for $i \neq j$ and $i, j \in \{1, ..., p\}$. In this case, (Z-A2) is more appropriate. For errors that are completely independent, (Z-A1) is preferred. In some other cases, the correlation structure might also differ, leading to different assumptions on the $\mathbf{Z}$ dependency structure.

5.3.3 The likelihood of the model with quality index

For actuarial pricing, most of the models used are GLMs. In the GLM case, the set of coefficients $\beta = (\beta_0, ..., \beta_p)^T$ is found by maximizing the likelihood or log-likelihood (ML-Maximum likelihood) [4];

$$\underset{\beta \in \mathbb{R}^p}{Argmax}\ \mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}) = \underset{\beta \in \mathbb{R}^p}{Argmax} \sum_{i=1}^{n} log(f_Y(Y_i|\mathbf{X}_i; \beta)), \tag{5.4}$$

where $\mathcal{L}$ is the likelihood function of the outcome $\mathbf{Y}$, given $\mathbf{X}$ and β and f_Y is the density function of Y.

Because the observations are independent and identically distributed, the previous log likelihood is the sample analog of $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$. We assume mild regularity conditions (see Appendix 5.7.3) for the proper convergence of our models. In our framework, these regularity conditions lead to the existence of the moment-generating function for each imperfectly observed covariate. The appendix contains more details on the theoretical aspect.

5.3.4 Deriving β^{M_2}

We recall that the vector β^{M_2} exactly matches β^{M_1} when all individualized quality indexes equal 1, *i.e.*, when $K = J_{1,p+1}$. For any distribution and link function, it is possible to estimate the expected M_2 log-likelihood for a given $\mathbf{Q}$ by using $\mathbf{X}^{real}$ in the univariate case.

Theorem 1

Let $(\mathbf{Y}, \mathbf{X}, \mathbf{X}^{real}, \mathbf{Q})$ be the data sets as defined by Equation (5.1). We consider assumption (C1) in the univariate case $p = 1$. We also consider mild regularity assumptions, especially $\int_{\mathbb{R}^2} |log(f_Y(y|z; \beta))| dF_{Z_1}(z) dF_Y(y) < \infty$ for any value of β. Knowing $(\mathbf{Y}, \mathbf{X}^{real}, \mathbf{Q})$, a sample estimator of $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$ is

$$\begin{aligned} &\bar{Q}_1 \sum_{i=1}^{n} log(f_Y(Y_i|X_{i1}^{real}; \beta)) \\ &+ (1 - \bar{Q}_1) \times \sum_{i=1}^{n} \frac{1}{n} \sum_{h=1}^{n} log(f_Y(Y_i|X_{h1}^{real}; \beta)). \end{aligned} \tag{5.5}$$

This estimator converges almost surely and is written as $log(\mathcal{L}^{M_2}(\beta; \mathbf{Y}|X^{real}, \mathbf{Q}))$. The associated maximum likelihood estimator $\hat{\beta}^{M_2|X^{real},\mathbf{Q}}$ converges in probability to β^{M_2}, *i.e.*

$$\hat{\beta}^{M_2|X^{real},\mathbf{Q}} \xrightarrow{\mathbb{P}} \beta^{M_2}.$$

The theorem can be easily extended to multivariate hypotheses (X-A3) and (Z-A1).

Theorem 2

Under the assumptions (X-A3) and (Z-A1) and the same hypothesis as in the univariate case, the

4. or equivalently minimizing the deviance

sample analog of $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta)))$

$$\begin{aligned}&\bar{Q}_p \sum_{i=1}^{n} log(f_Y(Y_i|\mathbf{X}^{real}_{i;(*p)}, X_{i;p} = X^{real}_{i;p};\beta)\\ &+ (1-\bar{Q}_p)\sum_{i=1}^{n}\frac{1}{n}\sum_{h=1}^{n} log(f_Y(Y_i|\mathbf{X}^{real}_{i;(*p)}, X_{i;p} = X^{real}_{h;p};\beta)),\end{aligned} \tag{5.6}$$

is consistent. The associated maximum likelihood estimator $\hat{\beta}^{M_2|X^{real},\mathbf{Q}}$ converges in probability to β^{M_2}, *i.e.*

$$\hat{\beta}^{M_2|X^{real},\mathbf{Q}} \xrightarrow{\mathbb{P}} \beta^{M_2}.$$

In fact, for any correlation structure between $\mathbf{X}^{real}$, $\mathbf{\Omega}$, and $\mathbf{Z}$, an estimate of the expected likelihood of M_2 can be found easily through simulations. The only constraints needed are that the mild regularity conditions must be verified under the chosen correlation structure or by recurrence. In this paper, we consider only (X-A3) for clarity purposes.

A downside of these methods is that $\mathbf{X}^{real}$ must be known, which is not always the case. Nonetheless, if $\mathbf{X}$ has the same correlation structure as $\mathbf{X}^{real}$ [5], a possible solution is to simulate Y^{new} from $\mathbf{X}$ by using $\hat{\beta}$ and therefore be able to calculate the previous estimator.

Both theorems permit the estimation of β^{M_2} through $\mathbf{X}^{real}$ for any $\mathbf{Q}$.

5.3.5 Deduce β^{M_1}

By using Theorem 2, β^{M_1} can be found for several distributions. Because log-Poisson GMLs are the most commonly used for actuarial modeling, this part focuses on log-Poisson GLM under (X-A3) and (Z-A1). Estimators for other distributions or assumptions can be created exactly in the same way.

We denote $V = (v_i)_{i=1,\dots,n}$ as the exposure for more traditional notation for count distributions. The exposure is supposed to be perfectly observed.

We recall that only $\mathbf{X}_p$ has a heterogeneous quality. By using Equation (5.36) in Appendix 5.7.5, an estimator of $log(\mathcal{L}(\hat{\beta};\mathbf{Y}|\mathbf{X}^{real},\mathbf{Q}))$ can be found as follows :

$$\begin{aligned}log(\mathcal{L}^{M_1}(\hat{\beta};\mathbf{Y}|\mathbf{X},\mathbf{Q})) = \frac{1}{\bar{Q}_p}\Big[&log(\mathcal{L}^{M2}(\hat{\beta};\mathbf{Y}|\mathbf{X})\\ &- (1-\bar{Q}_p)\times log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}^{real}_{(*p)}))\\ &- (1-\bar{Q}_p)\times\sum_{i=1}^{n} v_i e^{\hat{\beta}^{*p}\mathbf{X}^{real}_{i,(*p)}}(1-M_{X_p}(\hat{\beta}_p))\Big].\end{aligned} \tag{5.7}$$

All the terms on the right are known and can be evaluated. Indeed,

— $log(\mathcal{L}^{M_2}(\hat{\beta};\mathbf{Y}|\mathbf{X}))$ is the M_2 model log-likelihood when using all the covariates;

— $M_{X_p}(\hat{\beta}_p)$ can be estimated, or for particular distributions, given the distribution parameters, the moment generating function is explicitly known;

— $log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}^{real}_{(*p)}))$ is the M_2 model log-likelihood when using all the covariables except for $\mathbf{X}_p$; under the assumption (X-A3), $log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}^{real}_{(*p)}))$ is equal to $log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}_{(*p)}))$.

In the same spirit, another estimator can be put forward as a sum of the previous estimator conditioned by the pattern of quality K_p :

$$\begin{aligned}log(\mathcal{L}^{M_1}(\hat{\beta};\mathbf{Y}|\mathbf{X},\mathbf{Q})) = \sum_{K_p\in P(Q_p),K_p\neq 0}\frac{1}{K_p}\Big[&log(\mathcal{L}^{M2}(\hat{\beta};\mathbf{Y}|\mathbf{X}_{Q_p=K_p})\\ &- (1-K_p)\times log(\mathcal{L}(\hat{\beta}^{*p};\mathbf{Y}|\mathbf{X}^{real}_{(*p);Q_p=K_p}))\\ &- (1-K_p)\times\sum_{i=1}^{n} v_i e^{\hat{\beta}^{*p}\mathbf{X}^{real}_{i;(*p);Q_p=K_p}}(1-M_{X_p}(\hat{\beta}_p))\Big].\end{aligned} \tag{5.8}$$

5. e.g., in the (C1) case under (X-A1) and (Z-A1) or in the (C2) case with fully correlated quality variables and (Z-A2).

where $\mathbf{X}_{Q=K_p}$ represents the data set in which only individual i such as $Q_{i;p} = K_p$ are kept. The second estimator $log(\mathcal{L}^{M_1}(\hat{\beta}; \mathbf{Y}|\mathbf{X}, \mathbf{Q}))$ from Equation (5.8) is often more precise by construction than Equation (5.7). Individuals with a null quality index are not considered. Therefore, in the following part, the second estimator is used. These two estimators converge in probability to $log(\mathcal{L}(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q}))$. In the same way, the solution of the maximum likelihood converges in probability.

Optimization program : In contrast to the classical optimization method, the iterative weighted least square algorithm used to fit GLM parameters cannot be used. Empirically, the Nelder-Mean optimization from the *optim* function used the *stats* package (R software) seems to have a more stable convergence than the Newton-Raphson algorithm.

Indeed, for some distributions, the moment generating function may not exist or has extremely high values for some values of $\hat{\beta}_p$. In this case, the estimated derivative may be important. For these reasons, the Newton-Raphson method can lead to critical starting oscillations depending on the $\hat{\beta}_p$ and $\mathbf{X}_p$ distributions. This is why Nelder-Mean optimization is here preferred and starting at $\hat{\beta}_p = 0$.

5.3.6 Estimators proprieties for other distributions

Table 5.2 shows the different results for various GLMs and assumptions. For the most common GLMs used in nonlife pricing (log-Gaussian, log-Poisson, and log-gamma GLMs), some interesting results can be found thanks to the additive or multiplicative structure. However, for probit or Inv-gamma GLMs, no explicit formulas can be found without approximation.

Log-Gaussian GLM : the log-Gaussian GLM structure leads to an explicit relation between β and β^{M2}. Therefore, the M_1 log-likelihood does not need to be calculated. Because log-Gaussian GLM and linear regression are equivalent, the same results can be stated. It is important to note that β_j^{M2} depends only on Q_j and β_k and Q_j for all X_k correlated to X_j in the log-Gaussian case. In other words, the coefficient of a variable is not altered by the quality of variables not correlated to it . In case (C2) with a fully correlated quality variable under (Z-A2), *i.e.* $\Omega_j = \Omega_k$ for all j and k, log-Gaussian coefficients $\beta_j^{M_2}$ have a simple affine relationship with $\beta_j^{M_1}$ for $j = 1, ..., p$.

Multiplicative structures : In case (C1), the log-Poisson and log-gamma GLMs multiplicative structure provide the ability to find $log(\mathcal{L}^{M1}(\beta; \mathbf{Y}|\mathbf{X}, \mathbf{Q}))$. However, in the multivariate case, under (X-A3) and (Z-A1), $\beta_p^{M_2}$ depends on Q_p, the moment generating function $M_{X_p}(t)$ and β_j for $j = 1, ..., p-1$. The main difference is that $\beta_j^{M_2}$ depends on the distribution of X_p. For log-Poisson models, $\beta_j^{M_2} = \beta_j$ and $\beta_p^{M_2}$ converge in probability on the interval $[0, \beta_p]$. However, for log-gamma GLMs, this property is not true, and the other coefficients, β_j^{M2}, are also impacted by Q_p. In case (C2), regrettably, no properties of the estimator can be stated for log-gamma and log-Poisson GLM.

Technical details about this part of the paper can be found in section 5.7, and proofs of main results are given in Appendix for the sake of concision.

5.4 Adapting the prediction to data quality

5.4.1 Reducing the error by mitigating the quality pattern

In this work, $\mathbf{X}$ is governed by the underlying process generating the covariates, as in Equation (5.1). In linear regression, the solution $\hat{\beta}$ minimizes the residual squared error (RSE) calculated on the data set $\mathbf{X}$. In GLM regression ([56, Nelder and Wedderburn, 1972]), it is the mean deviance $(1/n)Dev(\hat{\beta}|\mathbf{X}, Y)$ calculated on the data set $\mathbf{X}$ that is minimized (or equivalently the maximization of the likelihood). Our particular framework enables the grouping of two individuals i and i' having the same quality indexes (*i.e.* $\mathbf{Q}_i = \mathbf{Q}_{i'}$), which defines a quality pattern. We denote $P(\mathbf{Q})$ as the set of all quality patterns present in the data. By considering this approach, the cost metric can be improved since

$$(1/n)Dev(\hat{\beta}|\mathbf{X}, Y) \geq (1/n) \sum_{K \in P(\mathbf{Q})} \sum_{i \backslash \mathbf{Q}_i = K} Dev(\hat{\beta}^K|\mathbf{X}_i, Y_i), \tag{5.9}$$

where $\hat{\beta}^K$ is the solution found on a subset of the data with quality pattern K. The strategies to calculate these various coefficients are introduced in Section 5.3.1. For some cases, one could want a model to be used for the best prediction for a given set of quality indexes called the quality pattern K. In full generality, when $K = \mathbf{Q}_i$ is made of terms $Q_{ij} \neq 1$, the coefficients $\hat{\beta}^K$ need to be calculated. $\hat{\beta}^{K=\mathbf{Q}_i}$ is an estimator of β^{M_2} when the model is fitted on data set $\mathbf{X}$ but in the case $\mathbf{Q} = J_{n,1}\ \mathbf{Q}_i$. Therefore, the coefficient $\hat{\beta}^K$ is the one minimizing the mean $Dev(\hat{\beta}^K|\mathbf{X}, \mathbf{Y})$ for a given pattern of quality K as desired (see Equation (5.9)).

The next subsection introduces this algorithm, denoted M_3. The idea is to individualize the prediction. Depending on the situation, several strategies may be interesting :

A.) The modeler may have access to enough data of perfect quality $\mathbf{X}^{\mathbf{real}}$ and wants to find the best prediction for a data set $\mathbf{X}$ and $\mathbf{Q}$;

B.) The modeler may have access to enough data $\mathbf{X}$ and $\mathbf{Q}$ for each quality pattern and wants to find the best prediction for a $\mathbf{X}$ for each quality pattern;

C.) The modeler may have access to only data $\mathbf{X}$ and $\mathbf{Q}$ where the data volume for each pattern is not sufficient or for which they want to find the best model for a new quality pattern.

Model GLM	Case		Hyp $\mathbf{X}^{real}$	Hyp $\mathbf{Z}$	Estimator convergence	Log-likelihood (M_1)
Id-Gaussian	(C1)-(C2)	$p = 1$	No Hyp	No Hyp	$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$ and $\frac{\hat{\beta}_1^{M2}}{\bar{Q}_1} \xrightarrow{\mathbb{P}.} \beta_1$	-
	(C1)	$p > 1$	X-A1	Z-A1	$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$, $\frac{\hat{\beta}_j^{M2}}{\bar{Q}_j} \xrightarrow{\mathbb{P}.} \beta_p, \quad j = 1, ..., p.$	-
	(C1)	$p > 1$	X-A3	Z-A1	$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$, $\frac{\hat{\beta}_j^{M2}}{\bar{Q}_j} \xrightarrow{\mathbb{P}.} \beta_p, \quad j = 1, ..., p.$	-
	(C2) $\Omega_j = \Omega_k$	$p > 1$	No Hyp	Z-A2	$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$, $\frac{\hat{\beta}_j^{M2}}{\bar{Q}_j} \xrightarrow{\mathbb{P}.} \beta_p, \quad j = 1, ..., p.$	-
Log-Poisson	(C1)-(C2)	$p = 1$	No Hyp	No Hyp	$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$ and $\hat{\beta}_1^{M_2} \xrightarrow{\mathbb{P}.} [0; \beta_1]$	Yes
	(C1)	$p > 1$	X-A3	Z-A1	$\hat{\beta}_j^{M_2} \xrightarrow{\mathbb{P}.} \beta_j, \quad j = 0, ..., p-1$ $\hat{\beta}_p^{M_2} \xrightarrow{\mathbb{P}.} [0; \beta_p]$.	Yes
	(C2) $\Omega_j = \Omega_k$	$p > 1$	No Hyp	Z-A2	-	Yes
Log-gamma	(C1)-(C2)	$p = 1$	No Hyp	No Hyp	$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0$ and $\hat{\beta}_1^{M_2} \xrightarrow{\mathbb{P}.} [0; \beta_1]$	Yes
	(C1)	$p > 1$	X-A3	Z-A1	-	Yes
	(C2) $\Omega_j = \Omega_k$	$p > 1$	No Hyp	Z-A2	-	Yes
Inv-gamma	(C1)	$p > 1$	X-A3	Z-A1	-	No
Probit	(C1)	$p > 1$	X-A3	Z-A1	-	No

TABLE 5.2 – Summary of the results on the most common GLMs.

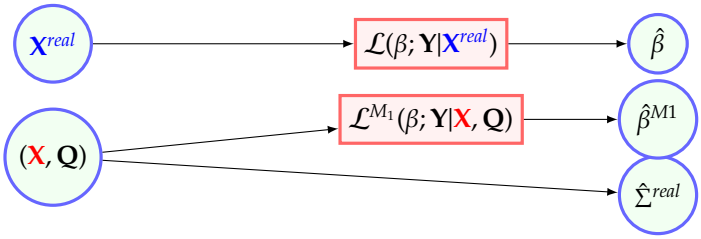
5.4.2 Algorithm 5.3 for GLM

M_3 **("Pattern-adjusted" models)** are based on $\mathbf{X}$ and Q, obtained from Algorithm 1. The models depend on each quality pattern :

$$\mathbb{E}[Y_i \mid \mathbf{X}_i,\ K = (Q_{ij})_{1\leq j\leq p}] = g^{-1}(\mathbf{X}_i\beta^K),$$

where K denotes the quality pattern associated with individual i. In this work, when $\mathbf{Q} = J_{n,1}K$, $\hat{\beta}^{M_3}$ estimates β^K.

I. Estimation of the real model coefficients :



II. For given a quality pattern $\mathbf{Q}_i = K \in P(Q)$, each individual i, $\mathbf{X}_i$:

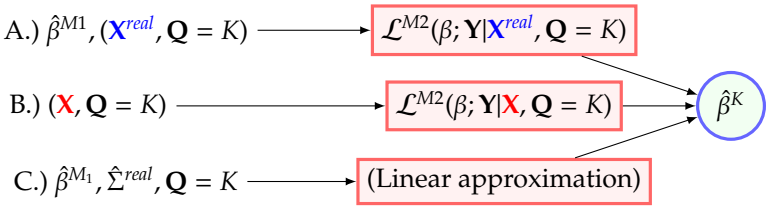


FIGURE 5.3 – Process to take into account the quality index for GLM. In this way, this process adjusts the coefficient to each quality pattern.

For GLM regression, a similar algorithm as for linear regression (see [120]) is suggested. To that end, the likelihood is studied instead of the correlation matrix. However, the algorithm M_3 cannot be applied as easily. No closed formula exists to link β^{M_2} with β. Therefore, in this work we propose finding β^{M_1}, an estimator of β by maximizing an estimator of the real model likelihood by using $\mathbf{Q}$ and $\mathbf{X}$ (see Section 5.3.4) or directly from $\mathbf{X}^{real}$.

In the event that $\mathbf{X}^{real}$ is known, i.e., a sufficiently large sample $\mathbf{X}$ is perfectly observed (step A.) or is of the same quality pattern K (step B.), $\hat{\beta}^K$ can be directly estimated from the maximization of the likelihood $\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X}^{real}, \mathbf{Q} = J_{n,1}K)$ or $\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X}, \mathbf{Q} = J_{n,1}K)$ (see Appendix 5.7.3). If the correlation structure of $\mathbf{X}^{real}$ is the same as that of $\mathbf{X}$, another solution is to simulate a new Y^{new} by using $\mathbf{X}$ and $\hat{\beta}^{M1}$ to apply an estimator proposed in Subsection 5.3.4.

The most recent strategy when $\mathbf{X}$ and $\mathbf{Q}$ are available with insufficient volume for each quality pattern is the following. Once $\hat{\beta}^{M_1}$ is determined, we propose to use a linear correction (step C.) to estimate $\hat{\beta}^K$. This approximation works well for small values of β (see Section 5.5.2).

Another strategy is to estimate $\mathcal{L}^{M_2}$ with estimators $\mathcal{L}^{M_1}$ and by using Equation 5.30 by recurrence. If the estimator converges in probability, the solution of the maximum likelihood converges in probability, but the estimator may have several local maxima.

5.5 Simulation study - M1 estimator

In this section, we verify our theoretical results on the estimator properties for log-Poisson GLM. In this scenario, all the simulated examples are created by using the following steps involving all the aforementioned quantities required to generate the right data :
Step 1 : Q is given in practice. For the simulation, it is randomly generated ;
Step 2 : $\mathbf{X}^{real}$ is simulated given the marginals and the correlation structure ;

Step 3 : $\mathbf{Z} = (Z_1, ..., Z_p)$ is simulated given $\mathbf{X}^{real}$ and the assumptions;
Step 4 : Y is simulated from its relationship with $\mathbf{X}^{real}$;
Step 5 : Ω is simulated from Q through Bernoulli trials;
Step 6 : $\mathbf{X}$ is derived thanks to Equation (5.1).

The study was performed using R ([107]) statistical software.

5.5.1 Find $\hat{\beta}^{M_1}$ coefficients

Let $\mathbb{E}(Y|\mathbf{X}^{real}) = exp(1 + 0.4X_1^{real} + 0.5X_2^{real} + 0.6X_3^{real} + 0.07X_4^{real})$ with $X_1 \sim \Gamma(2,1)$, $X_2^{real} \sim \mathcal{N}(0,1)$, $X_3 \sim \mathcal{P}ois(2)$, $X_4 \sim \mathcal{N}(0,10)$ and Y follow a Poisson distribution. The quality index follows an independent discrete distribution on the values $(0.5; 0.75; 1)$ with probabilities $(0.25; 0.25; 0.5)$ for Q_4. We assume all the other covariates to be perfectly observed, *e.g.*, $Q_{i,j} = 1$ for all $i \in 1, ..., n$ and $j \in \{1, 2, 3\}$.

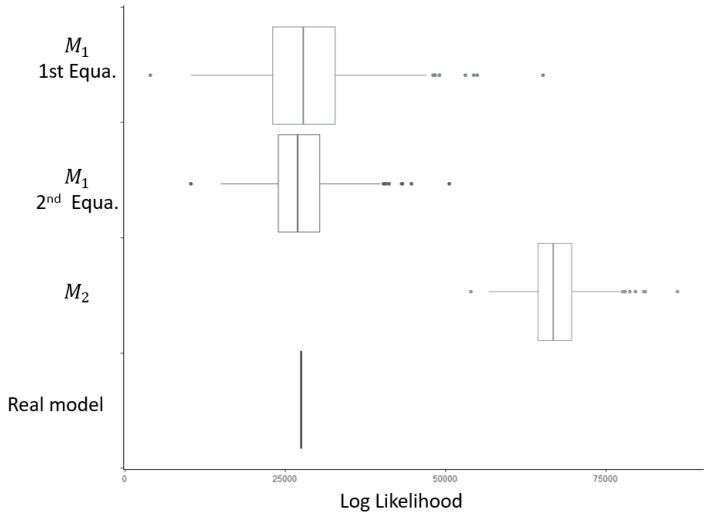


FIGURE 5.4 – Estimation of the M_1 log-likelihood for the log-Poisson GLM by using Equation (5.7) for a given $\mathbf{X}$ and $\mathbf{Q}$. The moment function is estimated by using its empirical estimator. The true function leads to the same graph but with a smaller variance. 2000 simulations are performed for a given $\mathbf{X}^{real}$ and $\mathbf{Q}$.

Using the preceding result, M_1 likelihood can be estimated as shown in Figure 5.4. The use of an imperfectly observed data set implies a wider variance of the estimator M_1 than that of the real model. Here, the first estimator has a larger variance than that of the second estimator. As shown by Equation (5.39), the coefficients β_1, β_2 and β_3 did not change due to the independence in between the variables - Figure 5.5 - and the coefficient associated with X_4 is corrected - Figure 5.6.

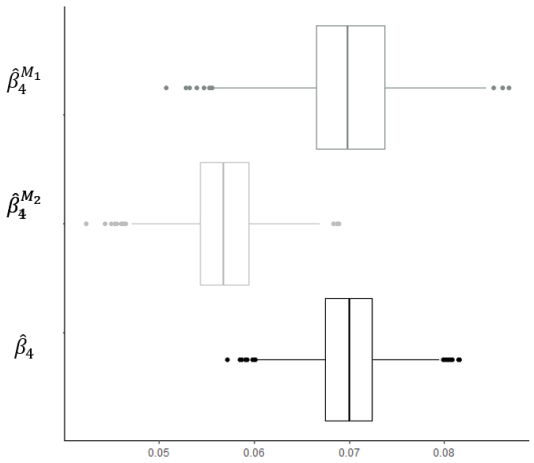


FIGURE 5.5 – $\hat{\beta}_4^{M_2}$ is smaller than $\hat{\beta}$ because of the quality of the variable. $\hat{\beta}_4^{M_1}$ is unbiased but has a larger variance than that of the real coefficient.

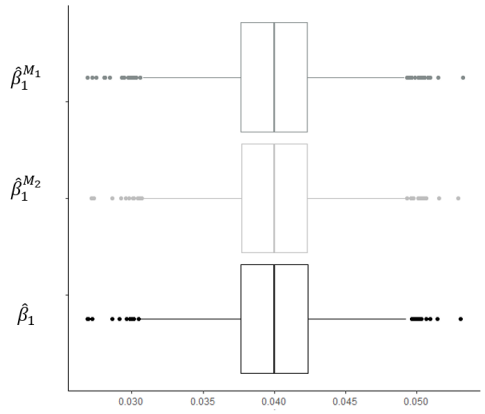


FIGURE 5.6 – The quality of the variable X_4 does not impact the estimation of β_1, β_2, β_3; here, highlighted by β_1 with a X_1^{real} standard normal distribution and Y following a Poisson distribution. Other distributions of X_1^{real} have also been evaluated and lead to the same results. 2000 simulations are performed for a given $\mathbf{Q}$.

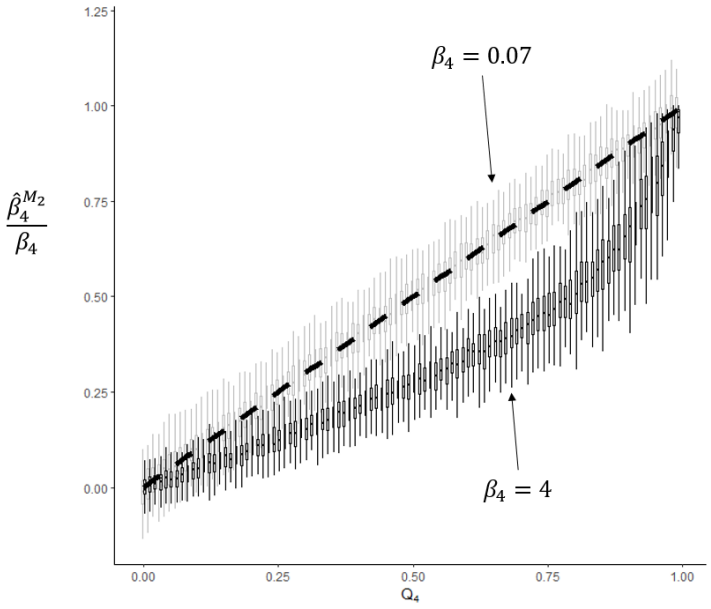


FIGURE 5.7 – 1000 simulation for each quality. For linear regressions, a linear evolution through the quality can be seen for low coefficients; however, for higher values, the relationship is not proportional to the quality.

5.5.2 Adapt the coefficient to the quality

Unlike linear regression, no explicit relation exists between β and β^{M_2} or β^K as a function of the quality. The coefficient is a barycenter of $\hat{\beta}^{M_1}$ and 0. Moreover, $\hat{\beta}_p^{M_2}$ converges to 0 when Q_p tends to 0. We suggest using the linear approximation, *i.e.*, $\hat{\beta}_p^{K=Q_p} = Q_p \times \hat{\beta}_p^{M_1}$. Indeed, as shown in Figure 5.7, for small values of β_4 (≈ 0.07), the impact of the moment on the likelihood is lower than that for higher values of $\beta_4 = 4$. Therefore, the coefficient can be estimated linearly only for small $\mathbb{E}(Y)$ but overestimates the coefficient for higher values. In household insurance for individuals, this assumption is adapted for the low annual frequency of claims.

5.6 Discussion

In this discussion, the adequacy of the various assumptions and hypotheses are evaluated. The following example comes straight up from a project on household pricing when using geolocated addresses to add external data. First, Subsection 5.6.1 gives proper examples encountered and justifies the various hypotheses needed in our framework in Subsection 5.6.2. In our example, issues remain, such as imperfect quality indexes, their evaluation, and correlations between variables. Sections 5.6.3 and 5.6.4 emphasize the limits and present some solutions using interactions.

5.6.1 Examples

This section discusses various cases in the context of pricing by using residential geolocation. Here, the goal is to model the frequency or claim cost of household insurance by using only the address and external data. As explained in the introduction, our particular framework is adapted to this problem. The first step in finding the various covariates associated with a house's characteristics is to associate the address with its geocoding and then to link the geocoding to the correct parcel of land or with the building. Then, by geolocating external data and calculating characteristics from picture analysis or other prediction methods, a database is created. The variable to model is given by insurers' departments. It corresponds to the frequency or claims cost and is supposed to be perfectly observed.

Here, the quality of the collected data is mainly examined through the credibility dimension. If the geocoding is wrong, all the observations could be taken from another building. The consistency of the variable and the way it is collected also change the data quality. Moreover, the reliability of the predicted characteristics also depends on the reliability of the covariables used in the predictive model.

The various assumptions are discussed through the pricing of home insurance when using geocoding.

Example of case C1 : The collection of the variables, the presence of a pool, and the presence of solar panels can fit the description. We suppose that the pool variable collection uses a governmental data set based on inhabitants' declarations and the solar panels variable uses house geocoding to determine which pictures to analyze. The collection of the two variables is not correlated. The case (C1) and the assumption (Z-A1) are appropriate. Indeed, if one is wrongly observed, it does not induce the other one to be, and the errors are not linked with the variable value, *i.e.*, $\mathbf{Q}$, $\mathbf{X}^{real}$ and $\mathbf{Z}$ are independent.

Example of case C2 : The living surface, the number of rooms, and the footprint are globally some of the most segmenting features in household pricing. Various data sets and methods are available in France to collect them, such as DVF[6]. This database geolocates parcels and contains various features such as property values, the number of rooms, the surface of the parcel, and the living surface, among others. The database is created from all properties transferred since 2015. With respect to the uncertainty dimension, errors arise from the connection between geocoding and an address or between an address and a building. Each of these steps affects the data quality depending on the feature. An incorrect geocoding could imply that the observations are taken from another building. For all these variables, case (C2) and assumption (Z-A2) are appropriate since they are collected from the same building.

When the wrong values are informative : The previous example also acts with respect to the mismeasurement dimension. Data quality, impacted by the consistency of the collection in the database, interferes on it due to the timeless dimension ; houses might have changed since the last property transfer. Indeed, the precision of the house's size may be biased after an expansion of a house if the database is not updated in the meantime. Moreover, correlations between $\mathbf{X}^{real}$ and $\mathbf{Z}$ also come from the way that variables are collected ; the best example is spatial correlation. For instance, let us look into a variable informing on the number of floors being collected from pictures analysis. The impact of geocoding uncertainty is not globally the same as before. Indeed, neighboring houses often have the same height or number of floors. Then, even if the collection of the data is done on the wrong building, $\mathbf{Z}$ are correlated with $\mathbf{X}^{real}$.

6. This database comes from a certified public service documenting the property values declared during property transfers : it is available as open data

When the quality indexes are informative : All of the above variables can fall into this category due to spatial correlation. In our study, it is rare that the quality variable does not depend on whether the building is in a rural area or urban area. Moreover, if it is in a megalopolis the detection of the house size may be difficult due to building density, a globally smaller systematic uncertainty could appear on this variable for urban houses. Then, $\mathbf{Z}$ could be automatically correlated with $\mathbf{X}^{real}$ but also with $\mathbf{Q}$. The same analysis can be performed on tall buildings, *e.g.*, for the number of floors.

One of the most difficult cases is when the quality depends on other variables, for instance, the material of the roof and the analysis of a roof to detect a window - see Figures 5.8 and 5.9. In this case, the modality of dark slate informs the risk, not because dark slate changes the risk but due to the low quality of the variable roof windows associated with it.

FIGURE 5.8 – The detection of a window on a roof is immediate from the IGN cartography (47.183722, -1.812768) - © IGN 2018.

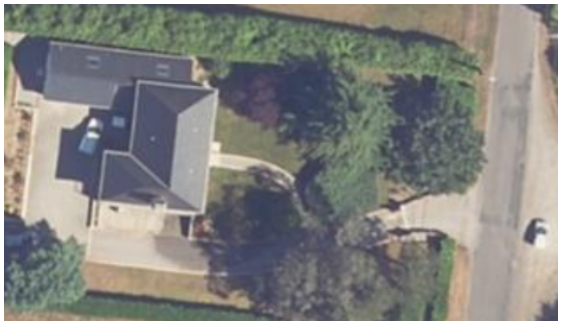


FIGURE 5.9 – The detection of a window on a roof is harder because of the dark slate roof from the IGN cartography (47.179068, -1.814216) - © IGN 2018

5.6.2 Actuarial justification of this framework assumptions

Integrity of the data set and the assumption on Z : In all examples discussed above, the wrong observations $\mathbf{Z}$ are coming from real individuals; it justifies the main assumption that ""wrong" values $\mathbf{Z}_j$ follow the same distribution as $\mathbf{X}_j^{real}$ (Equation (5.1)). However, the assumption is true only if the integrity of the data set is valid. Indeed, for instance, if some incorrect observations are taken from commercial buildings or flats when pricing residential household insurance, this assumption is not supported.

Assumption (X-A3) : The assumption (X-A3) is a very restrictive assumption. Nonetheless, it can be appropriate for underwriting. First, the use of several imperfectly observed covariates is not recommended and not adapted when aiming for a stable model. In addition, the traditional covariates used are known to be of high quality; in practice, up to one or two variables of heterogeneous quality are integrated. Adding some imperfect variables correlated to others also distorts the coefficients of these perfectly observed variables.

Use of the linear approximation to find an adapted model : As shown in Section 5.5, a linear approximation can be good for small values of the coefficients. In other words, the approximation can be valid when claim count modeling is performed at the individual scale. Indeed, in household insurance, the mean damage frequency is approximately 1 % (for instance, water damage or fire damage coverage). The other benefit is that only one model is fitted.

Add a new variable in pricing : Last, our framework can be used to estimate β for a new covariate. Without a data set and claims associated with it, the observations of this new variable have to be determined by using external data or models. Indeed, it is impossible to request new information once the contract is signed. However, a question can be added to the underwriting questionnaire during a quote, and therefore, the covariate can be used in the new pricing model. Logically, information from the underwriting questionnaire is much better quality and is often supposed to be perfectly observed (for most of the variables). Therefore, pricing models must use β, adapted to perfectly observed variables, and not β^{M2}.

5.6.3 Use interactions with quality indexes

The various results also help to understand how to deal with a finite number of quality groups within a variable. One could propose to use interactions instead of the framework in this paper. Indeed, the quality effect could be considered by adding an interaction between Q_j and X_k, $k \neq j$. We consider the following log-Gaussian GLM :

$$\mathbb{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{5.10}$$

and ρ the correlation between the two covariates. We suppose that the data set has another variable Q_1 with two modalities (high and low) informing on the quality of X_1. From the results earlier, adding some interactions between X_1 and Q_1 only, i.e.,

$$\mathbb{E}[Y|\mathbf{X}, Q_1] = \mathbf{1}_{Q_1=L}(\beta_0^{Q_1=L} + \beta_1^{Q_1=L} X_1) + \mathbf{1}_{Q_1=H}(\beta_0^{Q_1=H} + \beta_1^{Q_1=H} X_1) + \beta_2 X_2 \tag{5.11}$$

is the best option only if they are not correlated. The interaction should be on both variables :

$$\mathbb{E}[Y|\mathbf{X}, Q_1] = \mathbf{1}_{Q_1=H}(\beta_0^{Q_1=H} + \beta_1^{Q_1=H} X_1 + \beta_2^{Q_1=H} X_2) + \mathbf{1}_{Q_1=L}(\beta_0^{Q_1=L} + \beta_1^{Q_1=L} X_1 + \beta_2^{Q_1=L} X_2). \tag{5.12}$$

Obviously, with more covariates and quality indexes, many more parameters are needed to fit exactly $n \times 2^{h-p}$, where h is the sum of the modalities' number of each quality index. Moreover, the coefficients $\hat{\beta}_2^{Q_1=H}$ and $\hat{\beta}_2^{Q_1=L}$ could have different signs (see [120] or the appendix). For other distributions, the whole issue is much more complex. Therefore, in such a case, limiting the correlation between variables should be the priority.

5.6.4 Determine quality indexes and the impact of imperfect quality indexes

In a pricing data set studied, the quality index was given as an ordered variable with the following modality ("very low", "low", "medium", "high", "very high"). Is it possible to determine the equivalent quality index by modality ?

By evaluating a model by modality, quality indexes can be easily found given baseline coefficients - for example, β (known or evaluated thanks to the best quality points). The difficulty resides in the way that the quality is given. By fitting a univariate linear model with variables centered and an interaction between X_1 and the variable representing the quality $K(X_1)$ with $M \in \mathbb{N}$ modalities,

$$E[Y|\mathbf{X}, K(X_1)] = \beta_0^{M2} + \sum_{m=1,\dots,M} \beta_1^{M2,K(X_1)=m} X_1 \mathbb{1}_{K(X_1)=m}, \tag{5.13}$$

each quality index modality can be evaluated. Indeed, we recall that the modality $K(X_1) = 1$ corresponds to perfect quality observations, and the quality index value of the modality m is equal to $Q_m = \beta_1^{K(X_1)=m} / \beta_1^{K(X_1)=1}$.

Figure 5.10 shows a real example of a quality index assessment. The model is a univariate log-Poisson GLM that uses only the living surface to predict the water damage frequency. The values of the living surface come at first from labels by using DVF by associating a building with a property sale. To complete the missing information, prediction methods are performed by using the house characteristics. If the confidence in the database geocoding is perfect, the confidence associated is "very high". Otherwise, the confidence is degraded depending on the reliability of the geocoding of the property sales database. On the other hand, predicted values are associated with a maximum of "high" (in the majority "medium"). The credibility is degraded depending on the quality of the covariates and the score associated with each result. Two filters are considered for the geolocation of the addresses to link the claims and these characteristics : a filter keeping all the buildings considered as the main one on the parcel and a second keeping only the building if it is linked only to one address. Figure 5.10 helps to evaluate the quality index values. We suppose that $\beta^{High-One\ address} = \beta$ is perfect. The value of each Q can be approximated by $\frac{\hat{\beta}^Q}{\beta}$ by using a linear approximation. We recall that the annual frequency of water damage is low, approximately 3 per 100.

Then, the "medium" quality value was estimated at 0.6, and the "low" quality value was estimated at 0.5. However, the coefficient of very low-quality values has the opposite sign. Very low-quality values are linked to a rural zone. Case (C4) is the most appropriate, and our evaluation method cannot be used.

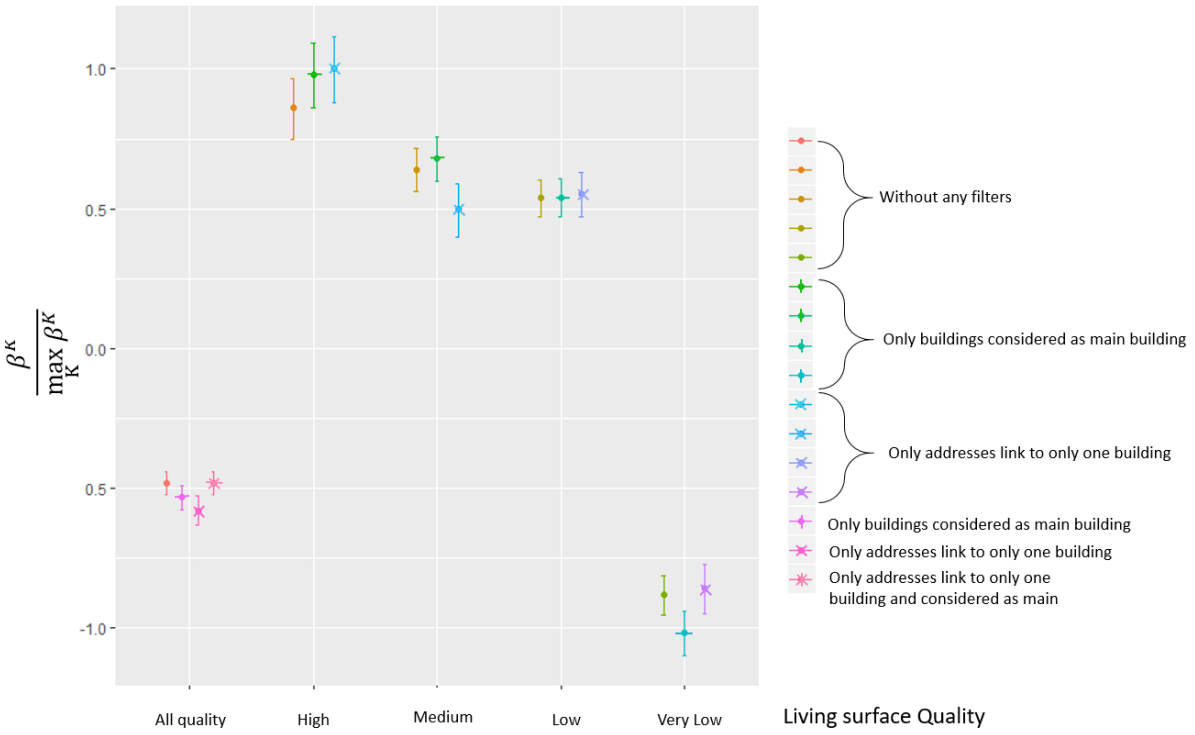


FIGURE 5.10 – Ratio of the living surface coefficient for the log-Poisson. Because there were too few "very high" quality observations, they were regrouped with "high" quality observations. Different filters based on building geolocation information are applied to the data set to challenge the quality index.

In the same way, "low" quality values are also more linked to rural density than "medium" or "high"[7]. Consequently, the associated value to medium quality 0.5 can be debated. Indeed, the "low" and "very low" quality are correlated with other characteristics impacting the risks. The coefficients calculated on the database are therefore impacted. This is an important limitation of our framework. In such cases, the use of a threshold to set aside "very low" quality observations is recommended so that the data set satisfies our assumptions. The different filters on geocoding show that leaner detail could be added within a value of the quality index. In this case, a modality may regroup different levels of quality. In other words, quality indexes are not perfectly determined.

In practice, a modality might regroup observations of different quality. This part considers the case of a modality regrouping two types of observations with distinct qualities. We denote m as a modality of n observations which regroups n_α and n_κ number observations with the quality $\bar{Q}_\alpha$ and $\bar{Q}_\kappa$, respectively ($n_\alpha + n_\kappa = n$). The difference between the coefficients of the model and the real model coefficients can be expressed as a barycenter of the sum of the group's quality under (X-A1) :

$$\beta_1^{M2,\bar{Q}_m} - \beta_1 = \frac{n_\alpha}{n}(\bar{Q}_\alpha - 1)\beta_1 + \frac{n_\kappa}{n}(\bar{Q}_\kappa - 1)\beta_1. \tag{5.14}$$

Equation (5.14) can be easily extended to higher dimensions. If groups of differing quality are mixed together and are given the same quality index value, the best one should be the weighted mean of each quality in a context of linear regression with the assumption (X-A1). However, under (X-A2) (with correlation), the aggregation of the quality influences the coefficient values of other correlated covariates.

Proposition 1 For log-Gaussian GLM, under assumptions (X-A2) and (Z-A1), given k and j such as $\rho_{kj}^{real} = \rho$, $Q_k \neq 0$ and $Q_j \neq 0$, if $\rho\beta_k^{M_1} \geq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j^{M_1}$,

$$\begin{aligned} \beta_k^{M_2} :]0,1] &\to \mathbb{R} \\ Q_k &\mapsto \beta_k^{M_2}(Q_k|Q_j). \end{aligned} \tag{5.15}$$

7. Here, the mismeasurement side of the data quality is set aside.

is an increasing convex function. Otherwise, it is decreasing concave.

Therefore, the weighted mean of the quality is a biased approximation. Indeed, according to Proposition 1, if $\rho\beta_k \geq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j$, for $i \neq j$:

$$\forall Q_\alpha, Q_\kappa \in [0,1], \quad \beta_k^{M_2}(\frac{n_\alpha}{n}\bar{Q}_\alpha + \frac{n_\kappa}{n}\bar{Q}_\kappa) \leq \frac{n_\alpha}{n}\beta_k^{M_2}(\bar{Q}_\alpha) + \frac{n_\kappa}{n}\beta_k^{M_2}(\bar{Q}_\kappa). \tag{5.16}$$

Consequently, regrouping two groups of different qualities biases the coefficient accordingly to the correlation. The equivalent quality index in linear regression under this assumption should be lower than the weighted mean of the quality. Because the convexity depends on the correlation, the weighted mean of the quality may be a fine approximation with a low correlation between covariates.

Conclusion

This investigation extends a method to consider a quality index on the credibility dimension for GLM regression. In pricing, the quality index corresponds to an external score when open/external data are added to a traditional data set. Moreover, for Rubin's nomenclature, various cases exist depending on the correlation structure between quality indexes, real observations, and incorrect observations. Relaxing the various assumptions, especially some hypotheses between a variable and its quality, is the next step. These results are especially useful for actuaries that are in charge of the data quality they use and the resulting models. The various cases have been discussed under one real pricing scheme when using geolocated addresses to find external information. Finally, actuaries should keep in mind that they are answerable for the data quality they use. Therefore, in this work we suggest a method to evaluate data quality and propose recommendations with data quality indexes for use.

To use data quality indexes with correlated covariates, further research is ongoing to adapt decision trees for this purpose and to relax assumptions between quality variables and the true values. Several issues remain when trying to generalize our approach for penalized likelihood optimization and in quality index evaluation.

Acknowledgements : The authors thank the firm and the data provider who have allowed them to use the portfolio, to geolocate it and to create this data set for this paper. The authors would like to thank the various referees who helped improve this paper.

Conflict of interest

The authors declare no conflict of interest.

5.7 Theoretical framework

5.7.1 The covariance impacted by quality index

Given a data set with two covariates and their joint quality (X_j, Q_j), (X_k, Q_k), $j \neq k$ as in Equation (5.1), the following lemma states the relationship between the real covariance Cov_{jk}^{real} and the observed covariance Cov_{jk}^{real} under various assumptions :

Lemma 1

In the case (C1), the relation yields

$$\textbf{Under (Z-A1)}\ Cov_{jk} = Q_jQ_k \times Cov_{jk}^{real}. \tag{5.17}$$

$$\textbf{Under(Z-A2)}\ Cov_{jk} = (1 + 2Q_jQ_k - Q_j - Q_k) \times Cov_{jk}^{real}. \tag{5.18}$$

In case (C2), if X_j^{real} and X_k^{real} are independent, both (5.17) and (5.18) hold true. Otherwise, if the

joint quality is completely and positively dependent, then the following results hold :

$$\textbf{Under (Z-A1)}\ Cov_{jk} = Q_j \times Cov_{jk}^{real}, \tag{5.19}$$

$$\textbf{Under (Z-A2)}\ Cov_{jk} = Cov_{jk}^{real}. \tag{5.20}$$

The proof of the case (C1) is available in [120, Chatelain and Milhaud, 2021]. The proof of case (C2) is a trivial extension of the previous one. When Z is correlated with X^{real}, an additional term corresponding to the correlation between the "wrong" value and the "right" value appears. The results could therefore be extended to such cases both under (Z-A1) or (Z-A2), but one needs to specify the correlation structure between $\mathbf{X}^{real}$ and $\mathbf{Z}$. Because each covariate $\mathbf{X}^{real}$ and $\mathbf{Z}$ have the same distribution, $Var(X_j) = Var(X_j^{real}) = Var(Z_j)$. Therefore, the same relationship with Pearson's correlation is also true. Thanks to Lemma 1, Σ^{real} can be evaluated from $\mathbf{Q}$ and Σ.

5.7.2 Regression model under consideration

Given the independent variables $(Y_1, ..., Y_n)$, the corresponding explanatory variables $(\mathbf{X}_1, ..., \mathbf{X}_n)$, and individualized quality indexes $(\mathbf{Q}_1, ..., \mathbf{Q}_n)$ where $\mathbf{Q}_i = (Q_{i1}, ..., Q_{ip})$, this part reviews the generalized linear model (GLMs). A GLM is defined by three components : the distribution's response variable Y, which is from the exponential family, a linear predictor $\mathbf{X}\beta$ and a link function g defined as

$$\mu = E[Y|\mathbf{X} = \mathbf{x}] = g^{-1}(\mathbf{x}\beta), \tag{5.21}$$

where $\mathbf{X}$ is the vector of covariates including a constant (see Section 5.2.1), and $\beta = (\beta_0, \beta_1, ..., \beta_p) \in \mathbb{R}^{p+1}$ is the vector of regression coefficients. β is found through maximum likelihood optimization. The classical linear regression model is a particular case of a GLM where $Y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2)$ and the link g is the identity function. The following sections link $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$ and $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta)))$.

5.7.3 Univariate analysis in GLM

This section focuses on the univariate case (p = 1). In cases (C1) and (C2), the quality variable $\mathbf{\Omega}$ is independent from the other variables. For β in $\mathbb{R}^2$, the model M_2 maximizes the following log-likelihood :

$$\begin{aligned}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta))) = \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta))|\Omega = 1) \times \mathbb{E}(\Omega_1 = 1) \\ + \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z}; \beta))|\Omega = 0) \times \mathbb{E}(\Omega_1 = 0).\end{aligned} \tag{5.22}$$

In Equation (5.22), the expected value of $\Omega = 0$, equal to Q_1, is known, and when $\Omega = 0$, X_1^{real} is not observed and Z_1 is given.

In cases (C1) and (C2), the quality variable $\mathbf{\Omega}$ is independent from the other variables, which leads to :

$$\begin{aligned}\mathbb{E}(log(f_Y(Y|\mathbf{X}^{real}; \beta)))|\Omega = 1) = \mathbb{E}(log(f_Y(Y|\mathbf{X}^{real}; \beta))), \\ \mathbb{E}(log(f_Y(Y|\mathbf{Z}; \beta))|\Omega = 0) = \mathbb{E}(log(f_Y(Y|\mathbf{Z}; \beta))).\end{aligned} \tag{5.23}$$

We recall that one main regularity condition is mandatory for the MLE convergence of the exponential family-based $\mathbf{X}^{real}$ (see Section 6.2 of [129, Lehmann and Casella, 1998] for all the conditions needed) :

Assumption 1 We assume that for every β, $log(f_Y(Y|\mathbf{X}^{real}; \beta))$ is integrable i.e, $\mathbb{E}(|log(f_Y(Y|\mathbf{X}^{real}; \beta))|) < +\infty$.

Assumption 5.7.3 and the other ones for the exponential family lead to the *Bartlett identities*, and both derivatives can be passed under the integral sign ([118, Barndorff-Nielsen, 1978]). The Bartlett identities are :

$$\begin{aligned}\mathbb{E}\Big(\frac{\delta}{\delta\beta} log(f_Y(Y|\mathbf{X}^{real}; \beta))\Big) = 0, \\ Var\Big(\frac{\delta}{\delta\beta} log(f_Y(Y|\mathbf{X}^{real}; \beta))\Big) = -\mathbb{E}\Big(\frac{\delta^2}{\delta\beta^2} log(f_Y(Y|\mathbf{X}^{real}; \beta))\Big).\end{aligned} \tag{5.24}$$

Moreover, the same assumptions on $\mathbf{Z}$ are also needed for the MLE convergence of the exponential family based on $\mathbf{X}$ in particular :

Assumption 2 We assume that for every β, $log(f_Y(y|z;\beta))$ is integrable i.e $\mathbb{E}(|log(f_Y(Y|Z;\beta))|) < +\infty$.

Because Z_j has the same marginal distribution as X_j^{real}, the regularity conditions 5.7.3 and 5.7.3 highly overlap. The main difference is the independence of $\mathbf{Z}$ from Y. Therefore, the *Bartlett identities* are still satisfied.

Remark : In the univariate case, the condition $\int_{\mathbb{R}^2} |log(f_Y(y|z;\beta))|dF_{Z_1}(z)dF_Y(y) < \infty$ implies $\int_{\mathbb{R}} |log(f_Y(y|\ z;\beta))\ |dF_{Z_1}(z) < \infty$ for any value of y and β. We recall that the $\mathbf{Z}_1$ distribution has the same distribution as $\mathbf{X}_1^{real}$. Hereafter, the canonical link function is used. In this case, the log-likelihood maximized can be written as follows :

$$Y_i(\beta_0 + \beta_1 X_1) - b(\beta_0 + \beta_1 X_1) + C^{st}, \tag{5.25}$$

where C^{st} is a constant independent of β and X and b(.) a real function. In the Bernoulli case assuming $\beta_1 \geq 0$ without loss of generality, the case $\beta_1 = 0$ being trivial, the condition

$$\begin{aligned}\int_{\mathbb{R}} |log(f_Y(y|z;\beta))|dF_{Z_1}(z) &\underset{\text{Triangle ineq.}}{\leq} \int_{\mathbb{R}} y_i|\beta_0 + \beta_1 z_1|dF_{Z_1}(z) + \int_{\mathbb{R}} |log(1 + exp(\beta_0 + \beta_1 z_1))|dF_{Z_1}(z),\\ &\underset{\frac{x}{1+x} \leq log(x) \leq x}{\leq} \int_{\mathbb{R}} y|\beta_0 + \beta_1 z_1|dF_{Z_1}(z) + \int_{\mathbb{R}} exp(\beta_0 + \beta_1 z_1)dF_{Z_1}(z),\end{aligned} \tag{5.26}$$

can be fulfilled with a condition on the $\mathbf{Z}_1$ moment generating function's existence for all β and with $\mathbb{E}(|Z_1|) < +\infty$. In the Poisson case, the same sufficient condition can easily be shown :

$$\int_{\mathbb{R}} |y(\beta_0 + \beta_1 x_1) - exp(\beta_0 + \beta_1 x_1)|dF_{Z_1}(z) \underset{\text{Triangle ineq.}}{\leq} \int_{\mathbb{R}} y|\beta_0 + \beta_1 z_1|dF_{Z_1}(z) + \int_{\mathbb{R}} exp(\beta_0 + \beta_1 z_1)dF_{Z_1}(z). \tag{5.27}$$

As the second moment existence was needed for linear regression convergence, the moment function's existence is also needed for proper maximum likelihood convergence.

How can we estimate $\mathbb{E}(log(f_Y(\mathbf{Y}|Z_1;\beta)))$ with X_1^{real}? In the multivariate case, the value $Z_{i,1}$ can be estimated by using the other covariables depending on the case. If X_1^{real} and Z_1 are correlated or dependent, a function g could exist such that $g(X_1^{real})$ is a good estimator of Z_1. Under case (C1), none of these solutions can be applied. Indeed, the quality index $\mathbf{Q}_1$, the real data-set X_1^{real} and the wrong values Z_1 are completely independent.

The following estimator,

$$\bar{Q}_1 \sum_{i=1}^{n} log(f_Y(y_i|X_{i,1}^{real};\hat{\beta})) + (1 - \bar{Q}_1) \times \sum_{i=1}^{n} \frac{1}{n} \sum_{h=1}^{n} log(f_Y(y_i|X_{h,1}^{real};\hat{\beta})). \tag{5.28}$$

converges almost surely to $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\hat{\beta})))$ (see the proof in Appendix 5.9.1).

In the multivariate case, the previous assumptions can easily be extended depending on the correlation structure. Under assumptions (X-A3) and (Z-A1), we recall the notation $\mathbf{X}_{(*p)} = (1, X_1; \cdots; X_{p-1})$ and its observed sample $X_{i;(*p)}$. In the same way, β^{*p} refers to $(\beta_0, \cdots, \beta_{p-1})$. The expected likelihood

$$\begin{aligned}\mathbb{E}(log(f_Y(\beta;\mathbf{Y}|\mathbf{X}))) = Q_p\ &\mathbb{E}(log(f_Y(Y|\mathbf{X}_{(*p)}^{real}, \mathbf{X}_p = \mathbf{X}_p^{real};\beta)))\\ &+ (1 - Q_p) \int_{\mathbb{R}} \mathbb{E}(log(f_Y(Y|\mathbf{X}_{(*p)}^{real}, \mathbf{X}_p = z;\beta)))f_{Z_p}(z)dz).\end{aligned} \tag{5.29}$$

can be written in a similar way as in the univariate case under mild regulatory conditions.

Under these assumptions, estimators $\hat{\beta}$ and $\hat{\beta}^{M2}$ converge in probability to β and β^{M2}, respectively.

Generalization could be easily done by recurrence under (X-A1) and (Z-A1), leading to similar formulas that are less readable by using the following relation :

$$\begin{aligned}\mathbb{E}(log(f_Y(\beta;\mathbf{Y}|\mathbf{X}))) = Q_p\ &\mathbb{E}(log(f_Y(Y|\mathbf{X}_{(*p)}, \mathbf{X}_p = \mathbf{X}_p^{real};\beta)))\\ &+ (1 - Q_p) \int_{\mathbb{R}} \mathbb{E}(log(f_Y(Y|\mathbf{X}_{(*p)}, \mathbf{X}_p = z;\beta)))f_{Z_p}(z)dz).\end{aligned} \tag{5.30}$$

For instance, for $p = 2$,

$$\begin{aligned}\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_1Q_2\ \mathbb{E}(log(f_Y(Y|\mathbf{X}^{real}; \beta))) \\ &+ (1-Q_1)Q_2 \int_{\mathbb{R}} \mathbb{E}(log(f_Y(Y|\mathbf{X}_1^{real} = z, \mathbf{X}_2^{real}; \beta)))f_{Z_p}(z)dz) \\ &+ Q_1(1-Q_2) \int_{\mathbb{R}} \mathbb{E}(log(f_Y(Y|\mathbf{X}_1^{real}, \mathbf{X}_2^{real} = z; \beta)))f_{Z_2}(z)dz) \\ &+ (1-Q_1)(1-Q_2) \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{E}(log(f_Y(Y|\mathbf{X}_1 = z_1, \mathbf{X}_2 = z_2; \beta)))f_{Z_1}(z_1)f_{Z_2}(z_2)dz_1dz_2).\end{aligned} \tag{5.31}$$

5.7.4 Example 1 : Identity-Gaussian GLM

In this section, the focus is on the Identity-Gaussian GLM. We recall that the covariates are centered (see the paper [120, Chatelain and Milhaud, 2021] for the uncentred case).

Let $\beta \in \mathbb{R}^{p+1}$. The likelihood to optimize is :

$$log(\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})) \propto \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i\beta)^2, \tag{5.32}$$

where for notation $\mathbf{X}_i = {}^t(1, \mathbf{X}_{i;1})$ and $\beta = {}^t(\beta_0, \beta_1)$.

In the univariate case, the Bartlett identities give the same results as the OLS :

$$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0, \frac{\hat{\beta}_1^{M2}}{\bar{Q}_1} \xrightarrow{\mathbb{P}.} \beta_1. \tag{5.33}$$

Indeed, linear regressions and log-Gaussian GLMs are equivalent. This proof can be easily extended in the multivariate case under assumptions (X-A1) and (Z-A1).

Under assumptions (X-A3) and (Z-A1), only β_p is impacted by

$$\hat{\beta}_j^{M_2} \xrightarrow{\mathbb{P}.} \beta_j, \frac{\hat{\beta}_p^{M2}}{\bar{Q}_p} \xrightarrow{\mathbb{P}.} \beta_p, \quad j = 0, ..., p-1. \tag{5.34}$$

In the particular case (C2) when the quality variables are fully correlated, i.e., $\Omega_j = \Omega_k \rightarrow Q_j = Q_k$ ($j \neq k$), under assumption (Z-A2) without any assumption on the correlation structure of $\mathbf{X}^{real}$, the following can be shown :

$$\hat{\beta}_0^{M_2} \xrightarrow{\mathbb{P}.} \beta_0, \frac{\hat{\beta}_j^{M2}}{\bar{Q}_j} \xrightarrow{\mathbb{P}.} \beta_p, \quad j = 0, ..., p. \tag{5.35}$$

The proofs of these results are in Appendix 5.9.2. The paper [120, Chatelain and Milhaud, 2021] proved the theorem 2 under (X-A2), where ρ represents the Pearson correlation of two variables.

5.7.5 Example 2 : log-Poisson GLM

In the Poisson case, the additive structure simplifies some calculus. Under assumptions (X-A3) and (Z-A1), the existence of the moment generating function $M_{X_p^{real}}(t) = M_{X_p}(t) = M_{Z_p}(t)$ for all $t \in \mathbb{R}$ and its derivatives' existence are ensured by the mild regularity condition 5.7.3. We denote V as the exposure. To keep the notation simple, let us omit the exposure V in the expected likelihood $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}; \beta)))$. Let $\beta \in \mathbb{R}^{p+1}$. The sample estimator of the expected likelihood is equal to

$$\begin{aligned}log(\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})) &= \bar{Q}_p\ log(\mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{\mathbf{real}}) + (1-\bar{Q}_p)\ log(\mathcal{L}(\beta^{*p}; \mathbf{Y}, \mathbf{X}^{\mathbf{real}}_{(*\mathbf{p})})) \\ &+ (1-\bar{Q}_p) \sum_{i=1}^{n} V_i e^{\beta^{*p}\mathbf{X}^{real}_{i;(*p)}} (1 - M_{X_p^{real}}(\beta_p)).\end{aligned} \tag{5.36}$$

Under (X-A3) and (Z-A1), $X^{real}_{(*p)} = X_{(*p)}$ allows us to evaluate the M_1 likelihood by using only $X_{(*p)}$ and $\mathbf{Q}$,

$$\begin{aligned} log(\mathcal{L}^{M_1}(\hat{\beta}; \mathbf{Y}|\mathbf{X}, \mathbf{Q})) = \frac{1}{\bar{Q}_p}(log(\mathcal{L}^{M2}(\hat{\beta}; \mathbf{Y}|\mathbf{X}) - (1-\bar{Q}_p) \times log(\mathcal{L}(\hat{\beta}^{*p}; \mathbf{Y}|X_{(*p)})) \\ - (1-\bar{Q}_p) \times \sum_{i=1}^{n} V_i e^{\hat{\beta}^{*p} \mathbf{X}^{real}_{i;(*p)}} (1 - M_{X_p}(\hat{\beta}_p))). \end{aligned} \tag{5.37}$$

The expected likelihood can be bounded :

$$\begin{aligned} Q_p \; log(\mathcal{L}(\beta; \mathbf{Y}|\mathbf{X}^{real})) + (1-Q_p) \; log(\mathcal{L}(\beta^{*p}; \mathbf{Y}, \mathbf{X}^{\mathbf{real}}_{(*\mathbf{p})})) \\ + (1-Q_p) \sum_{i=1}^{n} V_i e^{\beta^{*p} X^{real}_{(i);(*p)}} \left(1 - \exp\left(\frac{\beta_p^2 \mathbb{V}(\mathbf{X}^{\mathbf{real}}_{\mathbf{p}})}{2}\right)\right) \\ \leq log(\mathcal{L}^{M_2}(\beta; \mathbf{Y}|\mathbf{X})) \leq Q_p \; log(\mathcal{L}(\beta; \mathbf{Y}|X^{real})) + (1-Q_p) \; log(\mathcal{L}(\beta^{*p}; \mathbf{Y}, \mathbf{X}^{\mathbf{real}}_{(*\mathbf{p})})). \end{aligned} \tag{5.38}$$

By introducing the normalized coefficient $b_p = \frac{\beta_p}{\sqrt{\mathbb{V}(X^{real}_p)}}$, one can see that a small, normalized coefficient implies a narrower interval. In other words, the impact of variable quality on the likelihood logically depends on the normalize coefficient.

See Appendix 5.

Lemma 2

Let $\beta \in \mathbb{R}^{p+1}$. Under assumption (X-A3), the derivatives of the M_2 log-likelihood for j in $\{0, ..., p-1\}$ are equal to :

$$\begin{aligned} \frac{\delta}{\delta \beta_p} \mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) = Q_p \; d_p(\beta) \\ - (1-Q_p) \int_{\mathbb{R}^{p-1}} v e^{\beta^{*p} x^{real}_{(*p)}} dF_{\mathbf{X}^{real}_{(*p)}}(x^{real}_{(*p)}) \; M'_{X_p}(\hat{\beta}_p); \\ \frac{\delta}{\delta \beta_j} \mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) = d_j(\beta), \end{aligned} \tag{5.39}$$

where d_i is the derivative according to β_i of $\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}^{real})))$.

Unlike the log-Gaussian case, the difference between $\beta_p^{M_2}$ and β_p depends on the distribution of X_p.

When $\hat{\beta}_p \to 0$, $M'_{X_p}(\hat{\beta}_p) \to 0$. The derivatives in Equation (5.39) are a constant function of the mean quality. Therefore, the following proposition 1 can be deduced.

Proposition 1

We suppose the framework of this paper has a log-Poisson distribution. Under assumptions (X-A3), *i.e.* $Q_j = 1$ for $j \in \{1, \cdots, p-1\}$ and $Q_p \in (0, 1)$,

$$\begin{aligned} \beta_p^{M_2} : [0, 1] &\to \mathbb{R} \\ Q_p &\mapsto \beta_p^{M_2}(Q_p) \end{aligned} \tag{5.40}$$

is a monotonic function of the quality.

By using Lemma 2, it is straightforward to show the following theorem.

Theorem 3

Under Assumptions (X-A3) and (Z-A1),

$$\begin{aligned} &\hat{\beta}_j^{M_2} \xrightarrow{\mathbb{P}} \beta_j, \quad j \in 0, ..., p-1, \\ &\hat{\beta}_p^{M_2} \xrightarrow{\mathbb{P}} [0; \beta_p]. \end{aligned} \tag{5.41}$$

A particular application of this theorem would be under the univariate case $p = 1$. Remark that in the univariate case, (C1) and (C2) are equal. In the multivariate case, under (Z-A2) and (C2) with fully correlated quality variable and without any assumption on the structure of $\mathbf{X}^{real}$, the expected log-likelihood can be written only using $\mathbf{X}$ and the quality index $\mathbf{Q}$, (see Appendix 8) :

$$\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta))) = \frac{1}{Q_1}\Big(\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}, \beta)) - (1 - Q_1)\Big(\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0)) + \ Vexp(\beta_0)\ (1 - M_{\mathbf{X}^{real}}(\beta_*))\Big)\Big), \tag{5.42}$$

where $M_{\mathbf{X}^{real}}(\beta_*)$ is the multivariate generating function of $X_1^{real}, ..., X_p^{real}$ and $\beta_* = {}^t(\beta_1, ..., \beta_p)$. Unfortunately, no explicit bounds can be stated. The paper [128, Kuwarannancharoen and Sundaram, 2018]provides an upper bound on the location of the local minimum of the sum of two strongly convex functions under the assumption of a bounded gradient. The difficulty is that the log-likelihood exponential family is almost strongly convex, *i.e.* strongly convex in the neighborhood of β, as proven in [127, Kakade et al. 2010].

5.7.6 Example 3 : log-gamma GLM

The expected log-likelihood of log-gamma $Y \sim \Gamma(\mu, \nu)$ can be written as follows :

$$\mathbb{E}(log(f_Y(Y|\mathbf{X}, \beta))) = \int_{\mathbb{R}^{p+1}} \nu(-y\ exp(-\mathbf{x}\beta) - \mathbf{x}\beta + (\nu - 1)log(y) - log(\Gamma(\nu))dF_{X_1,...,X_p,Y}(\mathbf{x}, y). \tag{5.43}$$

Here, the only interest is to maximize the log likelihood according to β for a known ν. Therefore, the expected log-likelihood is studied,

$$\mathbb{E}(log(f_Y(Y|\mathbf{X}, \beta))) \propto \int_{\mathbb{R}^{p+1}} -y\ exp(-\mathbf{x}\beta) - \mathbf{x}\beta dF_{X_1,...,X_p,Y}(\mathbf{x}, y). \tag{5.44}$$

Under (X-A3) and (Z-A1), the expected log likelihood $\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X})))$ is equal to

$$Q_p\ \mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) + (1 - Q_p)\mathbb{E}(log(f_Y(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}_{(*p)}^{real})))M_{X_p}(-\hat{\beta}_p). \tag{5.45}$$

The M_1 estimator can be calculated as

$$\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) = \frac{1}{Q_p}\Big(\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}))) - (1 - Q_p)\mathbb{E}(log(f_Y(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}_{(*p)}^{real})))M_{X_p}(-\hat{\beta}_p)\Big). \tag{5.46}$$

Lemma 3

Let $\beta \in \mathbb{R}^{p+1}$. Under the assumption (X-A3), the derivative of the M_2 log-likelihood for j in $\{0, ..., p-1\}$ is equal to :

$$\begin{aligned} \frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_p\ d_j(\beta) \\ &\quad + (1 - Q_p)\ \frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\beta^{*\mathbf{p}}; \mathbf{Y}|\mathbf{X}_{(*p)}^{real})))M_{X_p}(\beta_p), \\ \frac{\delta}{\delta\beta_p}\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_p\ d_p(\beta) \\ &\quad + (1 - Q_p)\ \mathbb{E}(log(f_Y(\beta^{*p}; \mathbf{Y}|\mathbf{X}_{(*p)}^{real})))M'_{X_p}(\beta_p), \end{aligned} \tag{5.47}$$

where d_j is the derivative according to β_j of $\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}^{real})))$.

Lemma 3 cannot lead to a theorem such as in the log-Poisson case. The minimization of a sum of concave functions in $\mathbb{R}^{p+1}$ does not necessarily lead to $\beta_j^{M2} \in [min(\beta_j^{-p}, \beta_j), max(\beta_j^{-p}, \beta_j)]$, where β_j^{-p} is the maximum likelihood estimator $\mathbb{E}(log(f_Y(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}^{real}_{(*p)})))$ and $\beta_p^{-p} = 0$. Therefore, the log-gamma coefficients are not easily bounded in the general case. Nevertheless, the β_p^{M2} are still continuous according to the quality.

Under (C2) and (Z-A2), $\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X})))$ is equal :

$$Q_p \ \mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) + (1 - Q_p) M_{\mathbf{X}^{real}}(-\beta) \mathbb{E}(log(f_Y(\hat{\beta}_0; \mathbf{Y}))). \tag{5.48}$$

The proof is no different from the preceding. Nevertheless, no bound can be stated when $p > 2$.

5.7.7 Without multiplicative properties : Inv-gamma GLM and Probit GLM

The expected log-likelihood of Inv-gamma $Y \sim \Gamma(\mu, \nu)$ will be maximized for a known ν. The maximum likelihood estimator will maximize the sample analog of

$$\mathbb{E}(log(f_Y(Y|\mathbf{X}, \beta))) \propto \int_{\mathbb{R}^{p+1}} -y \ \mathbf{x}\beta + log(\mathbf{x}\beta) dF_{X_1,...,X_p,Y}(\mathbf{x}, y). \tag{5.49}$$

The expected log likelihood $\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X})))$ is equal to

$$Q_p \ \mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) + (1 - Q_p) \int_{\mathbb{R}^{p+1}} log(\mathbf{x}^{real}\beta^{*p} + z_p\beta_p) dF_{X_1^{real},...,X_{p-1}^{real},Y}(\mathbf{x}^{real}, y) dF_{Z_p}(z_p). \tag{5.50}$$

Because $log(\mathbf{X}^{real}\beta^{*p} + Z_p\beta_p)$, the sample analog cannot be estimated by using only $\mathbf{X}^{real}$ and $\mathbf{X}$, which does not allow us to find a relation between the likelihood when using only these two data sets.

For the Bernoulli distribution when using its canonical link function, the expected log-likelihood :

$$\mathbb{E}(log(f_Y(Y|\mathbf{X}, \beta))) \propto \int_{\mathbb{R}^{p+1}} -y \ \mathbf{x}\beta + log(1 + exp(\mathbf{x}\beta)) dF_{X_1,...,X_p,y}(\mathbf{x}, y), \tag{5.51}$$

cannot be calculated by using only $\mathbf{X}^{real}$. Looking for an approximation, $log(1+exp(x)) = log(2+exp(x)-1) = log(2) + log(1 + (exp(x) - 1)/2) \sim log(2) + (exp(x) - 1)/2 + o((exp(x) - 1))$ when $exp(x)$ is close to 1. Therefore, when $\mathbf{x}\beta$ is close to 0, a fine approximation with a multiplicative structure can be stated.

5.8 Various operational remarks

In this section, the focus is on the simplest case of GLM : log-Gaussian.

5.8.1 Quality impact and attenuation

The different results show that the "attenuation" [8] on $\hat{\beta}$ due to data quality can be explained. However, the quality impacts might not always decrease the coefficients, as shown in [120, Chatelain and Milhaud, 2021]. The quality of a variable impacts all coefficients related to other correlated variables in the uncentered case. Figure 5.11 under (X-A2) and (Z-A1) in linear regression shows that even in simple cases, the "attenuation" is not always true. With some correlation, the coefficient can be higher than the usual one (the true coefficient is equal to 1 and is represented by the line on Figures 5.11 and 5.12) and it might even change sign.

This is especially detrimental to insurance pricing, where covariate choices must be justified by their impacts. Indeed, some coefficients may seem counterintuitive due to quality impacts. Figures 5.11 and 5.12 provide an illustration of the impact of Q_1 depending on Q_2 (β_1, β_2 always equal to 1). The coefficients' evolution is not linear with the correlation. Figure 5.11 shows that if $\rho < 0$, $\beta_1^{M_2}$ can be negative even if $\beta_1 > 0$. Another point is that the coefficients can be considered null even if the variable's quality is not low. For instance, for $Q_2 = 0.7$ and $\rho \approx -0.4$, $\beta_1^{M_2} \approx 0$. In this case, dropping the variable X_1

8. As called in the econometric literature.

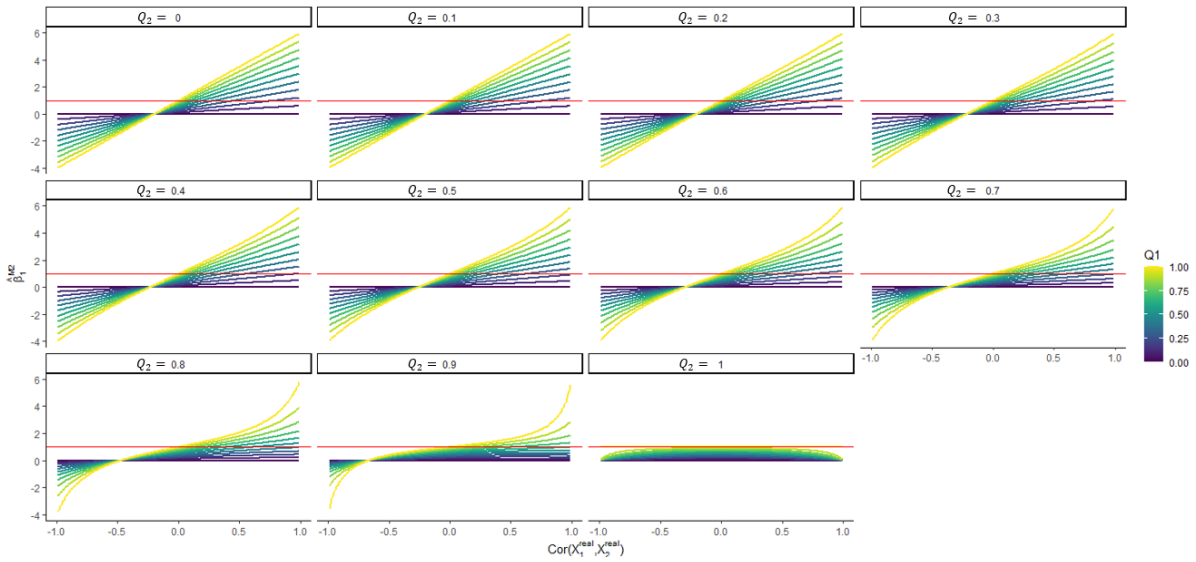


FIGURE 5.11 – Log-Gaussian GLM : Value of $\beta_1^{M_2}$ wrapped by Q_2 and grouped by Q_1. The coefficients β are all equal to 1, and the ratio of the standard deviation $\sqrt{Var(X_1^{real})/Var(X_2^{real})}$ equals 6. The red straight line represents, β_1 which is equal to 1.

does not have any impact on $\beta_2^{M_2}$ even if the true coefficient is different from 0. Moreover, by finding the β^{M_1} - thanks to **X** and **Q**, the modeler can find the "real" impact of a variable in models, thus justifying it.

> Here, the discussion is under the simplest hypothesis for case (C1) and for the Gaussian distribution where the variable quality does not impact the coefficients of the other independent variables. For other distributions, the quality impacts could further complicate the whole issue.

5.8.2 Missing data

The case of missing values can be seen as a particular case of this framework, where missing values are observations with a null quality. In the case of linear regression under (C1), (X-A1) and (Z-A1), the mean imputation is equivalent to the process explained in this paper. We denote the following model $E[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and ρ as the correlation between the two covariates. First, we suppose $\rho = 0$ and the individual i having its $X_{i,1}$ missing ; by using a simple mean imputation, the predicted value of Y_i is

$$\hat{Y}_i = \beta_0 + \beta_1 \mathbb{E}(X_1^{real}) + \beta_2 X_{i,2}^{real}. \tag{5.52}$$

Under (X-A1) and (Z-A1), the predicted value of y_i would be

$$\hat{Y}_i = \beta_0^K + \beta_1^K Z_{i,1} + \beta_2^K X_{i,2}^{real}, \tag{5.53}$$

where K is the pattern of the quality ; here $K = (0, 1)$, $\hat{\beta}_j^K$ is the best estimator, $j \in \{0; 1; 2; 3\}$ and $Z_{i,1}$ is a value drawn randomly from the empiric distribution of X_1^{real}. Due to the various assumptions and $K = (Q_1 = 0, Q_2 = 1)$, the coefficients can be written as

$$(\beta_0^K, \beta_1^K, \beta_2^K) = (\beta_0 + \beta_1 \mathbb{E}(X_1^{real}), 0, \beta_2), \tag{5.54}$$

showing the equivalence between the two methods. However, in correlated cases, for instance, under (X-A2) and (Z-A1), the coefficients equal :

$$(\beta_0^K, \beta_1^K, \beta_2^K) = (\beta_0 + \beta_1 \mathbb{E}(X_1^{real}) - \sqrt{\frac{Var(X_2^{real})}{Var(X_1^{real})}} \beta_1 \rho \mathbb{E}(X_2^{real}), 0, \beta_2 + \sqrt{\frac{Var(X_2^{real})}{Var(X_1^{real})}} \beta_1 \rho). \tag{5.55}$$

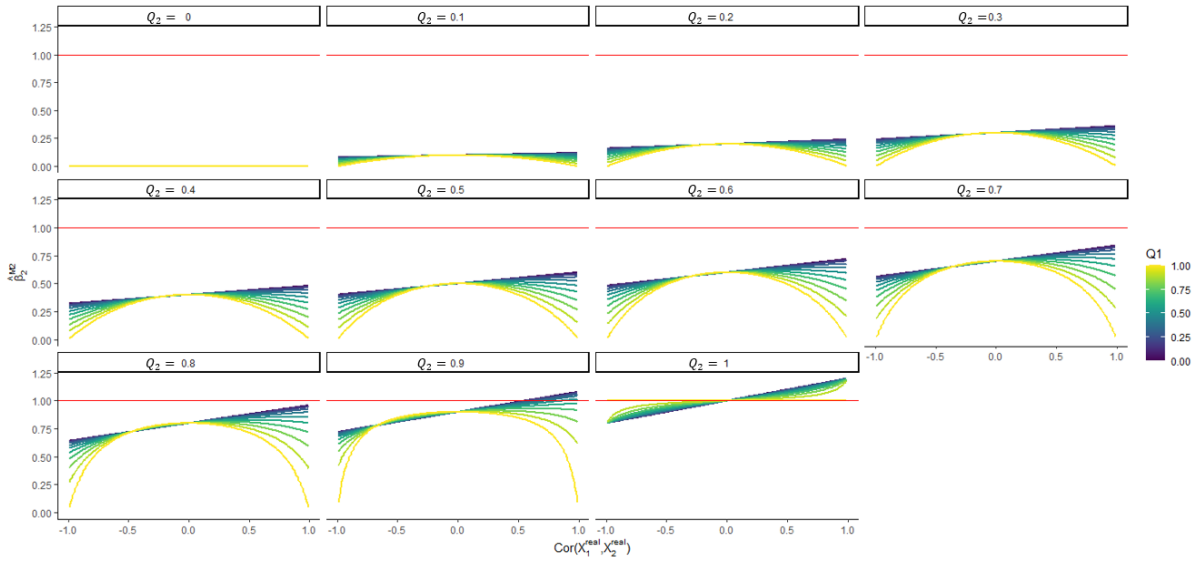


FIGURE 5.12 – Log-Gaussian GLM : Value of $\beta_2^{M_2}$ wrapped by Q_2 and grouped by Q_1. The coefficients β are all equal to 1, and the ratio of the standard deviation $\sqrt{Var(X_1^{real})/Var(X_2^{real})}$ equals 6.

Thus, the equivalent imputation here for $X_{i;1}^{real}$ is $\mathbb{E}(X_1^{real}) - \sqrt{\frac{Var(X_2^{real})}{Var(X_1^{real})}}\rho(\mathbb{E}(X_2^{real}) - X_{i;2})$. This imputation corresponds to the result of a linear regression to predict X_1^{real} by using only X_2^{real} for the individual, which is the best one according to the linear regression. In fact, $\sqrt{\frac{Var(X_2^{real})}{Var(X_1^{real})}}\beta_1^{M_1}\rho\mathbb{E}(X_2^{real})$ corresponds to the linear part of the information X_1^{real} already taken into account by X_2^{real}.

Multivariate case By extrapolating these results, it seems that for one missing observation, the equivalent imputation should be the prediction of the linear regression of the other covariates. However, this statement does not consider other covariate quality issues. Beyond case (C1), the reliability of the other covariates can also be correlated with the fact that the values are missing. This remark is close to the analysis of Seaman 2014 [132, Seaman and White, 2014] about how to impute with fully conditional specifications.

For the log-Poisson GLM, it has been shown that under (X-A3) and (Z-A1), the mean imputation is equivalent, which is not always true for other assumptions. For instance, in the log-gamma model, the other coefficients are also impacted. To find the best solution, one should find β^K and modify all the other coefficients.

5.9 Proofs

5.9.1 Proof of the Theorem 1

Proof 1

Let $(\mathbf{Y}, \mathbf{X}, \mathbf{X}^{real}, \mathbf{Q})$ be the data sets as defined by Equation (5.1). In the univariate case $p = 1$, the expected log-likelihood of the model M_2 depends on the quality index,

$$\begin{aligned}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta))) = \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))|\Omega = 1) \times \mathbb{E}(\Omega = 1) \\ + \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z};\beta))|\Omega = 0) \times \mathbb{E}(\Omega = 0),\end{aligned} \tag{5.56}$$

thanks to the independence between Ω and, respectively $\mathbf{X}^{real}$ and $\mathbf{Z}$. The first term is known, and because $\mathbf{Z}$ is independent of Y, the second can be rewritten by using Fubini's theorem :

$$\mathbb{E}(log(f_Y(Y|\mathbf{Z};\hat{\beta}))) = \mathbb{E}_Y \int_{\mathbb{R}} log(f_Y(Y|z;\beta))dF_{Z_1}(z). \tag{5.57}$$

Because Z_1 has the same distribution as X_1^{real}, X_1^{real} can be used to estimate the density f_{Z_1}, so $dF_{Z_1}(s) = dF_{X_1^{real}}(s)$. Finally, the previous equation can be estimated by the mean sample. Because $\{X_{1;1}^{real}, ..., X_{n;1}^{real}\}$ are *i.i.d* observations, the sample estimator is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{h=1}^{n} log(f(y_i|X_{h,1}^{real};\hat{\beta})). \tag{5.58}$$

With $\mathbb{E}(|log(f_Y(Y|\ \mathbf{Z},\beta)|) < \infty$ by using the strong law of large numbers, this sample estimator converges almost surely. The sample estimator $\sum log(f_Y(y_i|X_{i;1}^{real};\hat{\beta}))$ converges almost surely to $\mathbb{E}(log(f_Y(Y|\mathbf{X}^{real};\hat{\beta}))$. By using Kolmogorov's strong law of large numbers, $\bar{Q}_1$ converges in probability toward Q_1. Thus,

$$\bar{Q}_1\sum_{i=1}^{n} log(f_Y(y_i|X_{i,1}^{real};\hat{\beta})) + (1-\bar{Q}_1)\times\sum_{i=1}^{n}\frac{1}{n}\sum_{h=1}^{n} log(f_Y(y_i|X_{h,1}^{real};\hat{\beta})), \tag{5.59}$$

converges almost surely to $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\hat{\beta})))$. We denote this estimator $log(\mathcal{L}^{M_2}(\beta|\mathbf{Y},\mathbf{X}^{real},\mathbf{Q}))$.

Finally, the following points are true :

— observations are i.i.d and the density is Lebesgue measurable ;

— the parameter space of β is compact and open ;

— the previous estimator is concave as the sum of concave functions and is differentiable according to β ;

— Identifiability : the estimator function is a smooth function of β and converges in probability for all β towards $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\hat{\beta})))$ which has an unique solution.

Therefore, by using the Cramer-Rao conditions - Collorary 3.8 of [129, Lehmann and Casella, 1998], the global maximum exists, is unique and converges in probability to β^{M_2}, *i.e.*

$$\hat{\beta}^{M_2|X^{real},\mathbf{Q}} \xrightarrow{\mathbb{P}} \beta^{M_2},$$

meaning that the estimator is consistent.

5.9.2 Identity-Gaussian proofs

5.9.2.a Proof of Equation (5.33)

Proof 2

We recall that X_1^{real} is supposed to be centered. The likelihood maximization solution can be found as the solution when the derivative is set equal to 0. Taking derivatives with respect to β, the derived sample estimator can be written as follows :

$$\begin{aligned}\frac{\delta\mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\hat{\beta}^{M2}))}{\delta\beta} =& \frac{\delta\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta)))}{\delta\beta}\times Q_1 + \frac{\delta\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z};\beta)))}{\delta\beta}\times(1-Q_1)\\ =& Q_1\frac{\delta}{\delta\beta}\int_{\mathbb{R}^2}(y-x\beta_1+\beta_0)^2 dF_{X_1^{real},Y}(x,y)\\ &+(1-Q_1)\frac{\delta}{\delta\beta}\int_{\mathbb{R}}\int_{\mathbb{R}}(y-z\beta_1-\beta_0)^2 dF_{Z_1}(z)dF_Y(y)=0.\end{aligned} \tag{5.60}$$

We recall that the MLE solution exists and is unique. It is a well-known fact that if the identity $\int f_Y(y;\beta)d(y)=1$ is twice differentiable with respect to β, both derivatives can be passed under the integral sign. Therefore, the theorem of a differential under the integral can be applied and leads to the first *Bartlett identities*,

$$\begin{aligned}\frac{\delta\mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\hat{\beta}^{M2}))}{\delta\beta_0} &= -2\,Q_1\int_{\mathbb{R}^2}(y-x\beta_1-\beta_0)dF_{X_1^{real},Y}(x,y)\\ &\quad -2\,(1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}}(y-z\beta_1-\beta_0)dF_{Z_1}(z)dF_Y(y)=0,\\ \frac{\delta\mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\hat{\beta}^{M2}))}{\delta\beta_0} &= -2\,Q_1\int_{\mathbb{R}^2}x(y-x\beta_1-\beta_0)dF_{X_1^{real},Y}(x,y)\\ &\quad -2\,(1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}}z(y-z\beta_1-\beta_0)dF_{Z_1}(z)dF_Y(y)=0.\end{aligned} \tag{5.61}$$

Recall that $\int_{\mathbb{R}} z dF_{Z_1}(z)=0=\int_{\mathbb{R}} x dF_{X_1^{real}}(x)$. Therefore, the solutions of the precedent equations are :

$$\begin{aligned}\beta_0^{M2} &= Q_1\int_{\mathbb{R}^2} y dF_{X_1^{real},Y}(x,y)+(1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}} y dF_{Z_1}(z)dF_Y(y),\\ &= \int_{\mathbb{R}} y dF_Y(y)=\beta_0,\\ \beta_1^{M2} &= \frac{Q_1\int_{\mathbb{R}^2}xy dF_{X_1^{real},Y}(x,y)}{Q_1\int_{\mathbb{R}^2}x^2 dF_{X_1^{real},Y}(x,y)+(1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}}z^2 dF_{Z_1}(z)dF_Y(y)}=Q_1\beta_1.\end{aligned} \tag{5.62}$$

We recall that $\int_{\mathbb{R}} z dF_{Z_1}(z)=0=\int_{\mathbb{R}} x dF_{X_1^{real}}(x)$. The proof can be generalized exactly in the same way under (X-A1) and (Z-A1) for $p>1$.

5.9.2.b Proof of Equation (5.34)

Proof 3

The first *Bartlett identities* under assumption (X-A3) are equal, for $j = 1, ..., p-1$:

$$\begin{aligned}
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X}, \beta^{M_2}))}{\delta \beta_0} &= -2\, Q_1 \int_{\mathbb{R}^{p+1}} (y - \mathbf{x}_{*p}^{real} \beta^{*p;M_2} - x_p^{real} \beta_p^{M_2}) dF_{X_1^{real},...,X_p^{real},Y}(\mathbf{x}_{*p}^{real}, x_p^{real}, y) \\
&\quad - 2\,(1 - Q_1) \int_{\mathbb{R}^p} \int_{\mathbb{R}} (y - \mathbf{x}_{*p}^{real} \beta^{*p;M2} - z_p \beta_p^{M2}) dF_{Z_p}(z_p) dF_{X_1^{real},...,X_{p-1}^{real},Y}(\mathbf{x}_{*p}^{real}, y) = 0, \\
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X}, \beta^{M2}))}{\delta \beta_j} &= -2\, Q_1 \int_{\mathbb{R}^{p+1}} x_j^{real} (y - \mathbf{x}_{*p}^{real} \beta^{*p;M_2} - x_p^{real} \beta_p^{M_2}) dF_{X_1^{real},...,X_p^{real},Y}(\mathbf{x}_{*p}^{real}, x_p^{real}, y) \\
&\quad - 2\,(1 - Q_1) \int_{\mathbb{R}} \int_{\mathbb{R}^p} x_j^{real} (y - \mathbf{x}_{*p}^{real} \beta^{*p;M_2} - z_p \beta_p^{M_2}) dF_{Z_p}(z_p) dF_{X_1^{real},...,X_{p-1}^{real},Y}(\mathbf{x}_{*p}^{real}, y) = 0, \\
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X}, \beta^{M2}))}{\delta \beta_p} &= -2\, Q_1 \int_{\mathbb{R}^{p+1}} x_p^{real} (y - \mathbf{x}_{*p}^{real} \beta^{*p;M_2} - x_p^{real} \beta_p^{M_2}) dF_{X_1^{real},...,X_p^{real},Y}(\mathbf{x}_{*p}^{real}, x_p^{real}, y) \\
&\quad - 2\,(1 - Q_1) \int_{\mathbb{R}^p} \int_{\mathbb{R}} z_p (y - \mathbf{x}_{*p}^{real} \beta^{*p;M_2} - z_p \beta_p^{M_2}) dF_{Z_p}(z_p) dF_{X_1^{real},...,X_{p-1}^{real},Y}(\mathbf{x}_{*p}^{real}, y) = 0.
\end{aligned} \tag{5.63}$$

We recall that $\int_{\mathbb{R}} z_j dF_{Z_j}(z_j) = 0 = \int_{\mathbb{R}} x_j^{real} dF_{X_j^{real}}(x_j^{real})$ for $j = 1, ..., p$. Therefore, the solutions of the precedent equations are :

$$\beta_0^{M2} = \beta_0, \beta_j^{M2} = \beta_j, \beta_p^{M2} = Q_p \beta_p, \quad j = 1, ..., n. \tag{5.64}$$

Let us end the proof by replacing β^{M_2} and Q_1 with their estimators. Each of them converges in probability to $\hat{\beta}^{M_2}$ thanks to asymptotic MLE properties and $\bar{Q}_1$ by using Kolmogorov's strong law of large numbers.

5.9.2.c Proof of Equation (5.35)

Proof 4

For this proof, we write $\beta_* = (\beta_1, ..., \beta_p)$. In case (C2) with a perfectly correlated quality variable, *i.e.*, , $\Omega_j = \Omega_k \rightarrow Q_j = Q_k$ $(j \neq k)$, the following equation is obtained :

$$\begin{aligned}
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X}, \beta^{M_2}))}{\delta \beta} &= 0 \\
&= \frac{\delta \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real}; \beta^{M_2})))}{\delta \beta} \times Q_1 + \frac{\delta \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z}; \beta^{M_2})))}{\delta \beta} \times (1 - Q_1) \\
&= Q_1 \frac{\delta}{\delta \beta} \int_{\mathbb{R}^{p+1}} (y - x^{real} \beta_*^{M_2} - \beta_0^{M_2})^2 dF_{X_1^{real},...,X_p^{real},Y}(x, y) \\
&\quad + (1 - Q_1) \frac{\delta}{\delta \beta} \int_{\mathbb{R}} \int_{\mathbb{R}^p} (y - \mathbf{z} \beta_*^{M_2} - \beta_0^{M_2})^2 dF_{Z_1,...,Z_p}(\mathbf{z}) dF_Y(y).
\end{aligned} \tag{5.65}$$

The first *Bartlett identity* under assumption (Z-A2) is equal :

$$\begin{aligned}
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X}, \beta^{M_2}))}{\delta \beta_0} &= -2\, Q_1 \int_{\mathbb{R}^{p+1}} (y - \mathbf{x}^{real}\beta_*^{M_2} - \beta_0^{M_2}) dF_{X_1^{real},\dots,X_p^{real},Y}(\mathbf{x}^{real}, y) \\
&\quad - 2\,(1 - Q_1) \int_{\mathbb{R}} \int_{\mathbb{R}^p} (y - \mathbf{z}\beta_*^{M_2} - \beta_0^{M_2}) dF_{Z_1,\dots,Z_p}(\mathbf{z}) dF_Y(y) = 0, \\
\frac{\delta \mathbb{E}(log(f(\mathbf{Y}|\mathbf{X}, \beta^{M_2}))}{\delta \beta_j} &= -2\, Q_1 \int_{\mathbb{R}^{p+1}} \mathbf{x}_j^{real}(y - \mathbf{x}^{real}\beta_*^{M_2} - \beta_0^{M_2}) dF_{X_1^{real},\dots,X_p^{real},Y}(\mathbf{x}^{real}, y) \\
&\quad - 2\,(1 - Q_1) \int_{\mathbb{R}} \int_{\mathbb{R}^p} z_j(y - \mathbf{z}\beta_*^{M_2} - \beta_0^{M_2}) dF_{Z_1,\dots,Z_p}(\mathbf{z}) dF_Y(y) = 0.
\end{aligned} \tag{5.66}$$

Recall that $\int_{\mathbb{R}} z_j dF_{Z_j}(z_j) = 0 = \int_{\mathbb{R}} x_j^{real} dF_{X_j^{real}}(x_j^{real})$ for $j = 1, \dots, p$. Under the assumption (Z-A2), $\int_{\mathbb{R}^p} z_j \mathbf{z} dF_{Z_1,\dots,Z_p}(\mathbf{z}) = \int_{\mathbb{R}^p} \int_{\mathbb{R}} x_j^{real} \mathbf{x}^{real} dF_{X_1^{real},\dots,X_p^{real}}(\mathbf{x}^{real})$. Therefore, the solutions of the precedent equations are :

$$\beta_0^{M2} = \beta_0, \beta_j^{M2} = Q_j \beta_j, \quad j = 1, \dots, n. \tag{5.67}$$

Let end the proof by replacing the β^{M_2} and Q_1 by their estimators. Each of them converges in probability ; $\hat{\beta}^{M_2}$ thanks to asymptotic MLE proprieties and $\bar{Q}_1$ using the Kolmogorov's strong law of large numbers.

5.9.3 Proof for the GLM Log-Poisson

5.9.3.a Proof of Equations (5.36) and (5.38)

Proof 5

To keep the notation simple, I omit the exposure V in $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta)))$. Under the assumption (X-A3), using the Fubini's theorem, the expected likelihood (without the constant part) is equal to

$$\begin{aligned}
\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta))) \propto & Q_p \mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^{p+1}} -ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z} + n(\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z) dF_{X^{real}_1,\dots X^{real}_{p-1},V,Y}(\mathbf{x}^{real}_{(*p)},v,y)dF_{Z_p}(z) \\
\propto & Q_p \mathbb{E}(log(f_Y(Y|\mathbf{X}^{real};\beta))) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^{p+1}} +ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}} - ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z} dF_{X^{real}_1,\dots X^{real}_{p-1},V,Y}(\mathbf{x}^{real}_{(*p)},v,y)dF_{Z_p}(z) \\
& + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^{p+1}} -ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}} + y(\beta^{*p}\mathbf{x}^{real}_{(*p)}+\beta_p z) dF_{X^{real}_1,\dots,X^{real}_{p-1},V,Y}(\mathbf{x}^{real}_{(*p)},v,y)dF_{Z_p}(z) \\
\propto & Q_p \mathbb{E}(log(f_Y(Y|\mathbf{X}^{real};\beta))) + (1-Q_p)\mathbb{E}(log(f_Y(Y|\mathbf{X}^{real}_{(*p)};\beta^{*p}))) \\
& + (1-Q_p)\int_{\mathbb{R}^{p+1}} ve^{\beta^{*p}\mathbf{x}^{real}_{(*p)}} dF_{X^{real}_1,\dots X^{real}_{p-1},V,Y}(\mathbf{x}^{real}_{(*p)},v,y)(1-M_{X_p}(\beta_p)).
\end{aligned} \tag{5.68}$$

Because all the inputs are centered, the last term of the integral is null. Moreover, the moment generating function $M_{\mathbf{X}^{real}_p}(t)$ exists for all $t \in \mathbb{R}$, and the expected likelihood has a sample analog by using only $\mathbf{X}^{real}$

$$\begin{aligned}
& \bar{Q}_p\ log(\mathcal{L}(\beta|\mathbf{Y},\mathbf{X}^{real})) + (1-\bar{Q}_p)\ log(\mathcal{L}(\beta^{*p}|\mathbf{Y},\mathbf{X}^{real}_{*p})) \\
& + (1-\bar{Q}_p)\sum_{i=1}^{n} v_i e^{\beta^{*p}\mathbf{X}^{real}_{i;(*p)}}(1-M_{X_p}(\beta_p)).
\end{aligned} \tag{5.69}$$

If X_p is bounded, the Hoeffding Lemma gives us a proper upper bound, and Jensen's inequality gives us the lower bound. Indeed,

$$exp\left(\beta\mathbb{E}(X)\right) \leq \mathbb{E}(e^{\beta X}) \leq exp\left(\beta\mathbb{E}(X) + \frac{\beta^2(max(X)-min(X))^2}{8}\right).$$

With the Hoeffding inequality, another bound can be deduced without needing a bounded variable [a] :

$$exp\left(\beta\mathbb{E}(X)\right) \leq \mathbb{E}(e^{\beta X}) \leq exp\left(\beta\mathbb{E}(X) + \frac{\beta^2\mathbb{V}(X)}{2}\right).$$

These inequalities lead to Equation (5.38).

a. We recall that X_p is assumed to possess a second moment through the mild regularity conditions 5.7.3 -5.7.3.

5.9.3.b Proof of the Lemma 2

Proof 6

For j in $0, ..., p$, the gradient

$$\begin{aligned}\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta))) = Q_p\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(N|\mathbf{X}^{real};\beta)))\\ + (1-Q_p)\int_{\mathbb{R}}\int_{\mathbb{R}^p} -v\mathbf{x}_j^{real}e^{\beta^{*p}\mathbf{x}_{(*p)}^{real}+\beta_p z} + n\mathbf{x}_j^{real}dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real},n)dF_{Z_p}(z)\end{aligned} \tag{5.70}$$

can be separated into two part. We recall that $dF_{Z_p}(z) = dF_{X_p^{real}}(x)$ because Z_p and X_p^{real} have the same distribution. Under assumption (X-A3), $dF_{X_1^{real},...X_p^{real}}(\mathbf{x}^{real}) = dF_{X_1^{real},...X_{p-1}^{real}}(\mathbf{x}_{(*p)}^{real})dF_{X_p^{real}}(x_p^{real})$ for $j = 1, ..., p-1$. By replacing these values in the previous equation, we have :

$$\begin{aligned}\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X};\beta))) &= Q_p\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(N|\mathbf{X}^{real};\beta)))\\ &\quad + (1-Q_p)\int_{\mathbb{R}^{p+1}} -vx_j^{real}e^{\beta\mathbf{x}^{real}} + n\mathbf{x}_j^{real}dF_{X_1^{real},...X_p^{real},N}(\mathbf{x}^{real},n)\\ &= Q_p\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(N|\mathbf{X}^{real};\beta)))\\ &\quad + (1-Q_p)\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(N|\mathbf{X}^{real};\beta))) = d_j(\beta).\end{aligned} \tag{5.71}$$

The derivative with respect to β_p is calculated thanks to Equation (5.69) without difficulty. This end the proof for Equation (5.39).

5.9.3.c Proof of the theorem 3

Proof 7

The solution (MLE) β^{M_2} exists and is unique. Moreover, the solution β^{M_2} is a global maximum. Therefore, the solution β^{M_2} nullifies the partial derivatives, $d_j^{M_2}$, for $j = 0, ..., p.$, i.e., $d_j^{M_2}(\beta^{M_2}) = 0$. In the same way, $d_j(\beta) = 0$. We note that

$$\begin{aligned} d_j(\beta) &= \int_{\mathbb{R}} \int_{\mathbb{R}^p} -v x_j^{real} e^{\beta_{*p} \mathbf{x}_{*p}^{real} + \beta_p x^{real}} dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real}, n) dF_{X_p^{real}}(x^{real}) \\ &= \int_{\mathbb{R}^p} -v x_j^{real} e^{\beta_{*p} \mathbf{x}_{*p}^{real} + \beta_p x^{real}} dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real}, n) \underbrace{\int_{\mathbb{R}} e^{\beta_p x^{real}} dF_{X_p^{real}}(x^{real})}_{>0} = 0, \end{aligned} \tag{5.72}$$

which leads to $\int_{\mathbb{R}^p} -v x_j^{real} e^{\beta_{*p} \mathbf{x}_{*p}^{real} + \beta_p x^{real}} dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real}, n) = 0$.

We denote b as a set of coefficients such as $b_{*p} = \beta_{*p}$ and $b_p \in \mathbb{R}^*$, $j = 1, ..., p-1$. The derivative $d_j^{M_2}(\beta^{M_2})$,

$$\begin{aligned} d_j^{M_2}(b) = d_j(b) &= \int_{\mathbb{R}} \int_{\mathbb{R}^p} -v x_j^{real} e^{\beta_{*p} \mathbf{x}_{(*p)}^{real} + b_p x^{real}} dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real}, n) dF_{X_p^{real}}(x^{real}) \\ &= \underbrace{\int_{\mathbb{R}^p} -v x_j^{real} e^{\beta_{*p} \mathbf{x}_{*p}^{real}} dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real}, n)}_{=0} \int_{\mathbb{R}} e^{b_p x^{real}} dF_{X_p^{real}}(x^{real}), \end{aligned} \tag{5.73}$$

is null for $j = 1, ..., p-1$. Differentiating with respect to β_p, the derivative equals to

$$d_p^{M_2}(b) = Q_p \, d_p(b) - (1 - Q_p) \int_{\mathbb{R}^p} v e^{\beta_{*p} X_{i;*p}^{real}} dF_{X_1^{real},...X_{p-1}^{real},N}(\mathbf{x}_{(*p)}^{real}, n) M_{X_p}^{'}(b_p). \tag{5.74}$$

If $b_p > \beta_p$, then $d_p(b) > 0$, and if $b_p < 0$, then $d_p(b) < 0$. In the same way, if $b_p > \beta_p$, then $-M_{X_p}^{'}(b_p) > 0$, and if $b_p < 0$, then $-M_{X_p}^{'}(b_p) < 0$. These inequalities lead to the following : if $b_p > \beta_p$, then $d_p^{M_2}(b) < 0$, and if $b_p < 0$, then $d_p^{M_2}(b) > 0$.

Because $b \mapsto d_p^{M_2}(b)$ is a continuous function, a $b_p \in [0, \beta_p^{M_1}]$ exists, such as $d_p^{M_2}(b) = 0$.

We have proven that there exists b with the following characteristics $b_{*p} = \beta_{*p}$ and $b_p \in [0, \beta_p^{M_1}]$:

$$d_j^{M_2}(b) = 0, d_p^{M_2}(b) = 0. \tag{5.75}$$

Because the solution of M_2 log-likelihood maximization is unique, the previous solution is the global maximum β^{M_2}.

For $j = 1, ..., p$, we end the proof by replacing β_j, Q_1 and β_j with their estimators. Each of them converges in probability to $\hat{\beta}_0$ and $\hat{\beta}_p$, respectively, following the asymptotics MLE proprieties and $\bar{Q}_1$ by using the strong law of large number,

$$\hat{\beta}_j^{M_2} \xrightarrow{\mathbb{P}} \beta_j, \hat{\beta}_p^{M_2} \xrightarrow{\mathbb{P}} [0; \beta_p], \quad j \in 0, ..., p-1. \tag{5.76}$$

5.9.3.d Proof of the log-Poisson results for (C2) assumption

Proof 8

Denote $\beta_* = (\beta_1, ..., \beta_p)$. In the case (C2) with perfectly correlated quality variables, *i.e.* $\Omega_j = \Omega_k \rightarrow$

$Q_j = Q_k$ $(j \neq k)$, the equation under (Z-A2) can be written :

$$\begin{aligned}\mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\beta)) =& Q_1\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) + (1-Q_1)\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{Z};\beta)))\\ =& Q_1\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) + (1-Q_1)\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0)))\\ &+ (1-Q_1)\int_{\mathbb{R}}\int_{\mathbb{R}^p} -ve^{\beta_0+\beta_*\mathbf{z}} + n(\beta_0+\beta_*\mathbf{z})dF_{Z_1,\dots Z_p}(\mathbf{z})dF_N(n)\\ =& Q_1\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) + (1-Q_1)\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0)))\\ &+ (1-Q_1)\ vexp(\beta_0)\ M_{\mathbf{Z}}(\beta_*),\end{aligned} \tag{5.77}$$

where $M_{\mathbf{Z}}(\beta_*)$ is the multivariate generating function of $Z_1, ..., Z_p$ and under (Z-A2) is equals to $M_{\mathbf{X}^{real}}(\beta_*)$. The first of *Bartlett identities*,

$$\begin{aligned}\frac{\delta\mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\beta^{M2}))}{\delta\beta} =& Q_1\frac{\delta}{\delta\beta}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta^{M_2}))) + (1-Q_1)\frac{\delta}{\delta\beta}\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0^{M_2})))\\ &+ (1-Q_1)V\frac{\delta}{\delta\beta}\Big(exp(\beta_0^{M_2})(1-M_{\mathbf{X}^{real}}(\beta_*^{M_2}))\Big) = 0,\end{aligned} \tag{5.78}$$

does not permit to find a bound on β^{M_2} (see the remark for log-gamma GLM). However, $\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta)))$ can be calculated using only $\mathbf{X}$,

$$\begin{aligned}\mathbb{E}(log(f_Y(\mathbf{Y}|\mathbf{X}^{real};\beta))) =& \frac{1}{Q_1}\Big(\mathbb{E}(log(f(\mathbf{Y}|\mathbf{X},\beta)) - (1-Q_1)\mathbb{E}(log(f_Y(\mathbf{Y}|\beta_0)))\\ &- (1-Q_1)\ V\ exp(\beta_0)\ M_{\mathbf{X}^{real}}(\beta_*)\Big).\end{aligned} \tag{5.79}$$

5.9.4 Proof for log-gamma GLM

5.9.4.a Proof of the Lemma 3

Proof 9

The expected log-likelihood to maximize is equivalent to :

$$\int_{\mathbb{R}^{p+1}} -y\ exp(-\mathbf{x}\beta) - \mathbf{x}\beta dF_{X_1,...,X_p,Y}(\mathbf{x}, y). \tag{5.80}$$

The expected log likelihood $\mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X})))$ is equal to

$$Q_p\ \mathbb{E}(log(f_Y(\hat{\beta}; \mathbf{Y}|\mathbf{X}^{real}))) + (1 - Q_p)\mathbb{E}(log(f_Y(\hat{\beta}^{*p}; \mathbf{Y}|\mathbf{X}^{real})))M_{X_p}(-\hat{\beta}_p). \tag{5.81}$$

Under Assumption (X-A3), the derivative of the M_2 log-likelihood for j in $\{0, ..., p-1\}$ is equal to

$$\begin{aligned}
\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_p\ d_j(\beta) + (1 - Q_p) \\
&\frac{\delta}{\delta\beta_j}\int_{\mathbb{R}}\int_{\mathbb{R}^p} -y\ exp(-\mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p) - \mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p dF_{X_1^{real},...,X_{p-1},Y}(\mathbf{x}_{*p}^{real}, y)dF_{Z_p}(\mathbf{z}_p) \\
&= Q_p\ d_j(\hat{\beta}) + (1 - Q_p) \\
&\int_{\mathbb{R}}\int_{\mathbb{R}^p} -yx_j^{real}exp(-\mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p)dF_{X_1^{real},...,X_{p-1},Y}(\mathbf{x}_{*p}^{real}, y)dF_{Z_p}(\mathbf{z}_p) \\
&= Q_p\ d_j(\beta) + (1 - Q_p)\frac{\delta}{\delta\beta_j}\mathbb{E}(log(f_Y(\beta^{*p}; \mathbf{Y}|\mathbf{X}^{real})))M_{X_p}(-\beta_p), \\
\frac{\delta}{\delta\beta_p}\mathbb{E}(log(f_Y(\beta; \mathbf{Y}|\mathbf{X}))) &= Q_p\ d_p(\beta) + (1 - Q_p) \\
&\int_{\mathbb{R}}\int_{\mathbb{R}^p} -yz_p exp(-\mathbf{x}_{*p}^{real}\beta_{*p} - \mathbf{z}_p\beta_p)dF_{X_1^{real},...,X_{p-1},Y}(\mathbf{x}_{*p}^{real}, y)dF_{Z_p}(\mathbf{z}_p) \\
&= Q_p\ d_j(\beta) - (1 - Q_p)\mathbb{E}(log(f_Y(\beta^{*p}; \mathbf{Y}|\mathbf{X}^{real})))M_{X_p}^{'}(-\beta_p).
\end{aligned} \tag{5.82}$$

5.9.5 Convexity : Propositions 1

Proof 10

We denote the covariates X_j, X_k $(i \neq j)$ with a Pearson correlation ρ for which $|\rho| \neq 1$ and we suppose β_k and β_j nonnull. Using the corollary of [120, Chatelain and Milhaud, 2021], the following derivatives are found :

$$\begin{aligned}
\frac{\delta\beta_k^{M_2}(Q_k|Q_j)}{\delta Q_k} &= A \times \frac{1 + Q_j^2Q_k^2\rho^2}{(1 - Q_j^2Q_k^2\rho^2)^2}, \\
\frac{\delta^2\beta_k^{M_2}(Q_k|Q_j)}{\delta Q_k^2} &= A \times \frac{2Q_j^2Q_k\rho^2}{(1 - Q_j^2Q_k^2\rho^2)^3}(3 + Q_k^2Q_j^2\rho^2),
\end{aligned} \tag{5.83}$$

with $A = \beta_k(1 - Q_j^2\rho^2) + \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j\rho(1 - Q_j^2)$.

A is positive only if $\rho\beta_k > -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j$. Indeed,

$$
\begin{aligned}
0 \leq & \beta_k(1-Q_j^2\rho^2) + \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j^{M_1}\rho(1-Q_j^2) \\
0 \leq & \beta_k + \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j\rho - Q_j^2(\rho^2\beta_k - \sqrt{\frac{Var(X_j)}{Var(X_k)}}\beta_j\rho), \qquad (5.84)
\end{aligned}
$$

$$
\text{if } \rho \geq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\frac{\beta_j}{\beta_k} \text{ and } \beta_k \geq 0 \text{ or } \rho \leq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\frac{\beta_j}{\beta_k} \text{ and } \beta_k \leq 0.
$$

Then $\beta_k^{M_2}(Q_k|Q_j))$ is convex if $\rho \geq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\frac{\beta_j}{\beta_k}$ and $\beta_k \geq 0$ or $\rho \leq -\sqrt{\frac{Var(X_j)}{Var(X_k)}}\frac{\beta_j}{\beta_k}$ and $\beta_k \leq 0$ and concave in the two other cases. If Q_1, Q_2 or ρ are null, $\frac{\delta^2\beta_k^{M_2}(Q_k)}{\delta Q_k^2} = 0$ which ends the proof.

Bibliography

[117] Autorité de contrôle prudentiel, ACPR (2021). Synthèse de l'enquête déclarative de 2019 sur la gestion des données alimentant les calculs prudentiels des organismes d'assurance. Technical Report 119, ACPR.

[118] Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of statistics*, pages 151–157.

[119] Campbell, R., Francis, L., Prevosto, V., Rothwell, M., and Sheaf, S. (2006). Report of the data quality working party. Technical report.

[120] Chatelain, P. and Milhaud, X. (2021). Linear regression and data quality through individualized credibility index. preprint.

[121] Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1) :15–18.

[122] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6) :381–395.

[123] Francis, L. A. (2005). Dancing with dirty data methods for exploring and cleaning data.

[124] General Committee of the Actuarial Standards Board and Applies to All Practice Areas, GCASB (2014). Data quality - revised edition. Technical Report 23, Actuarial Standard of Practice.

[125] Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society : Series B (Methodological)*, 54(3) :761–771.

[126] Heitjan, D. F. and Basu, S. (1996). Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3) :207–213.

[127] Kakade, S., Shamir, O., Sindharan, K., and Tewari, A. (2010). Learning exponential families in high-dimensions : Strong convexity and sparsity. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 381–388. JMLR Workshop and Conference Proceedings.

[128] Kuwarannancharoen, K. and Sundaram, S. (2018). On the location of the minimizer of the sum of two strongly convex functions. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 1769–1774. IEEE.

[129] Lehmann, E. and Casella, G. (1998). Unbiasedness. *Theory of Point Estimation*, pages 83–146.

[130] Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

[131] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3) :581–592.

[132] Seaman, S. and White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods*, 43(16) :3499–3515.

[133] Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3) :278–295.

[134] Tami, M., Clausel, M., Devijver, E., Dulac, A., Gaussier, E., Janaqi, S., and Chebre, M. (2018). Uncertain trees : Dealing with uncertain inputs in regression trees. *arXiv preprint arXiv :1810.11698*.

[135] Todoran, I.-G., Lecornu, L., Khenchaf, A., and Le Caillec, J.-M. (2014). Toward the quality evaluation of complex information systems. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*, volume 9091, page 90910N. International Society for Optics and Photonics.

[136] Trabelsi, A., Elouedi, Z., and Lefevre, E. (2016). Handling uncertain attribute values in decision tree classifier using the belief function theory. In *International conference on artificial intelligence : Methodology, systems, and applications*, pages 26–35. Springer.

[137] Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

[138] Van Huffel, S. and Lemmerling, P. (2013). *Total least squares and errors-in-variables modeling : analysis, algorithms and applications*. Springer Science & Business Media.

[139] Vedaldi, A., Jin, H., Favaro, P., and Soatto, S. (2005). Kalmansac : Robust filtering by consensus. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 633–640. IEEE.

[140] Zhu, Z., Wang, T., and Samworth, R. J. (2019). High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv :1906.12125*.

Chapitre 6

Compléments sur la qualité de données et le lien avec la tarification à l'adresse

6.1 Des compléments sur la qualité de données

Une question immédiate se pose : *Quelle influence a la qualité de données dans les métriques évaluant la performance de la régression linéaire ?* Les deux sous-sections suivantes 6.1.1 montrent que la qualité de données détériore presque sûrement le coefficient de détermination R^2 quand le nombre d'observations tend vers l'infini (preuve dans la sous-section 6.1.2). Ces conjectures paraissent évidentes, mais la question se pose sur les modifications minimes de la qualité de données. En biaisant les coefficients, certains algorithmes (Lasso, Ridge) améliorent les métriques telles que le RMSE.

6.1.1 L'évolution des métriques de références

La relation entre les MSSE des deux modèles peut être explicitée. Considérons les métriques habituelles suivantes :

$$\hat{R}^2 = \frac{\left\|\hat{Y} - \mathbf{1}\bar{y}\right\|^2}{\left\|Y - \mathbf{1}\bar{y}\right\|^2} \tag{6.1}$$

où $\|.\|$ représente la norme euclidienne avec $\hat{R^2}$ le coefficient de détermination du modèle réel et $(\hat{R^2})^{M_2}$ celui du modèle M_2 sur une base donnée. Dénotons R^2 la limite quand n tend vers $+\infty$. La métrique considérée pour la preuve est la moyenne de SSE (la somme de la variance expliquée par la régression) noté $MSSE$:

$$MSSE^{M2} = \frac{1}{n}\left\|\hat{Y}^{M2} - \mathbf{1}\bar{y}\right\|^2 \text{ et } MSSE = \frac{1}{n}\left\|\hat{Y} - \mathbf{1}\bar{y}\right\|^2. \tag{6.2}$$

Trivialement, $MSSE > MSSE^{M2}$ induit $\hat{R}^2 > (\hat{R}^2)^{M2}$.

Théorème 6.1.1

Dans le cas (C1), sous (Z-A1) ou (Z-A2), et (X-A1), les relations asymptotiques entre les métriques des modèles linéaires sont :

$$MSSE \geq MSSE^{M2}, \quad \text{et } R^2 \geq (R^2)^{M2}, \qquad \textbf{p.s ,} \tag{6.3}$$

quand n tends $+\infty$. Les inégalités sont strictes s'il existe au moins un $j \in 1, ..., p$ tel que $Q_j \neq 1$ et $\beta_j \neq 0$.

Théorème 6.1.2

Sous les hypothèses (X-A2) et (Z-A1) et le cas (C1), les relations asymptotiques entre les métriques

usuelles des modèles linéaires sont :

$$MSSE \geq MSSE^{M_2}, \quad \text{et } R^2 \geq (R^2)^{M_2}, \qquad \textbf{p.s ,} \tag{6.4}$$

quand n tend vers ∞.

Les hypothèses sur $\mathbf{Z}$ impactent la performance des modèles. Cependant, l'inégalité est fausse pour l'hypothèse (Z-A2). En effet, pour les cas avec une faible corrélation ou une qualité faible, ces inégalités semblent triviales. Cependant, dans le cas d'une corrélation entre des variables proches de 1 et une qualité de la donnée proche de 1, le biais des coefficients induit par la qualité compense la perte de l'information à cause de l'hypothèse (Z-A2).

6.1.2 Les preuves des théorèmes 6.1.1 et 6.1.2

Preuve 1: Théorème 6.1.1

En utilisant les mêmes notations que dans le chapitre 4, calculons le MSSE du modèle M_2 pour n observations. Pour simplifier les équations, considérons Y centrée, c'est-à-dire que $\bar{y} = 0$ quand n tends vers $+\infty$. Cette manipulation ne modifie que la valeur de l'intercepte et rappelons que toutes les variables sont supposées centrées. Ainsi $MSSE^{M_2} = \frac{1}{n}\left\|\hat{Y}^{M_2} - \mathbf{1}\bar{y}\right\|^2$ tend vers $\frac{1}{n}\left\|\hat{Y}^{M_2}\right\|^2$.

$$\begin{aligned}\frac{1}{n}\left\|\hat{Y}^{M_2}\right\|^2 &= \frac{1}{n}\left\|X\beta^{M_2}\right\|^2 \\ &= \frac{1}{n}\,{}^t(X\beta^{M_2})(X\beta^{M_2}) = \frac{1}{n}\,{}^t\beta^{M_2}\,{}^tXX\beta^{M_2} \\ \frac{1}{n}\left\|\hat{Y}\right\|^2 &= \frac{1}{n}\,{}^t\beta\,{}^tX^{real}X^{real}\beta\end{aligned}$$

Sous (X-A1), les matrices de corrélation empirique $\frac{1}{n}\,{}^tX^{real}X^{real}$ sont diagonales et tendent presque surement vers Σ^{real} (SLLN).

Par abus de notation dans la suite de la preuve $\lim_{n\to+\infty} MSSE^{M2}$ est noté $MSSE^{M2}$. Le MSSE du modèle M_2 converge presque sûrement vers :

$$\begin{aligned}MSSE^{M2} - MSSE &= {}^t\beta^{M_2}\Sigma^{real}\beta^{M_2} - {}^t\beta\Sigma^{real}\beta, \\ MSSE^{M_2} &= \underbrace{MSSE}_{\text{Variance inhérent de la modélisation}} \\ &\quad - \underbrace{\sum_{j=1}^{p}(\beta_j)^2 Var(X_j^{real})(1-Q_j)}_{\text{Variance perdue à cause de la qualité de la donnée}}\quad .\end{aligned} \tag{6.5}$$

Comme $\sum_{j=1}^{p}(\beta_j)^2 Var(X_j^{real})(1-Q_j) \geq 0$, l'inégalité $MSSE \geq MSSE^{M2}$ s'en déduit aisément. S'il existe au moins un j tel que $\beta_j \neq 0$ et $Q_j \neq 1$ alors $\sum_{j=1}^{p}(\beta_j)^2 Var(X_j^{real})(1-Q_j) > 0$ et $MSSE > MSSE^{M2}$.

Preuve 2: Théorème 6.1.2 et contre-exemple pour (Z-A2).

Sous (Z-A2) ou (Z-A1) et (X-A2), le $MSSE^{M_2}$ s'écrit de façon plus complexe. En posant $\bar{\Omega} = (J_{n;(p+1)} - \Omega)$, $nMSSE^{M_2} = \left\|\hat{Y}^{M_2} - \mathbf{1}\bar{y}\right\|^2$ tend vers $\left\|\hat{Y}^{M_2}\right\|^2$ qui est égal à :

$$
\begin{aligned}
&\left\|X\beta^{M_2}\right\|^2 = \left\|X^{real}\beta^{M_2} + \left((Z \circ \bar{\Omega})\beta^{M_2} - (X^{real} \circ \bar{\Omega})\beta^{M_2}\right)\right\|^2 \\
&= \left\|X^{real}\beta^{M_2}\right\|^2 + \left\|\left((Z - X^{real}) \circ \bar{\Omega}\right)\beta^{M_2}\right\|^2 + 2\,{}^t(X^{real}\beta^{M_2})\left((Z - X^{real}) \circ \bar{\Omega}\right)\beta^{M_2} \\
&= \left\|X^{real}\beta\right\|^2 + \left\|X^{real}\beta^{M_2} - X^{real}\beta\right\|^2 + 2\,{}^t(X^{real}\beta)(X^{real}\beta^{M_2} - X^{real}\beta) \\
&+ 2\,{}^t(X^{real}\beta^{M_2})\left((Z - X^{real}) \circ \bar{\Omega}\right)\beta^{M_2} + \left\|\left((Z - X^{real}) \circ \bar{\Omega}\right)\beta^{M_2}\right\|^2 \\
&= \left\|X^{real}\beta\right\|^2 + {}^t(X^{real}\beta^{M_2} - X^{real}\beta)(X^{real}\beta^{M_2} - X^{real}\beta) \\
&+ 2\,{}^t(X^{real}\beta)(X^{real}\beta^{M_2} - X^{real}\beta) + 2\,{}^t(X^{real}\beta^{M_2})\left((Z - X^{real}) \circ \bar{\Omega}\right)\beta^{M_2} \\
&+ {}^t\beta^{M_2}\,{}^t(X^{real} \circ \bar{\Omega})(X^{real} \circ \bar{\Omega})\beta^{M_2} - 2\,{}^t\beta^{M_2}\,{}^t(X^{real} \circ \bar{\Omega})(Z \circ \bar{\Omega})\beta^{M_2} \\
&+ {}^t\beta^{M_2}\,{}^t(Z \circ \bar{\Omega})(Z \circ \bar{\Omega})\beta^{M_2} \\
&= \left\|X^{real}\beta\right\|^2 + {}^t\beta^{M_2}\,{}^t(X^{real} \circ \bar{\Omega})(X^{real} \circ \bar{\Omega})\beta^{M_2} - 2\,{}^t\beta^{M_2}\,{}^t(X^{real} \circ \bar{\Omega})(Z \circ \bar{\Omega})\beta^{M_2} \\
&+ {}^t\beta^{M_2}\,{}^t(Z \circ \bar{\Omega})(Z \circ \bar{\Omega})\beta^{M_2} + {}^t\beta^{M_2}\,{}^tX^{real}X^{real}\beta^{M_2} - {}^t\beta\,{}^tX^{real}X^{real}\beta \\
&+ 2\,{}^t\beta^{M_2}\,{}^tX^{real}Z \circ \bar{\Omega}\beta^{M_2} - 2\,{}^t\beta^{M_2}\,{}^tX^{real}X^{real} \circ \bar{\Omega}\beta^{M_2}.
\end{aligned}
\tag{6.6}
$$

Les matrices sont diagonales par blocs deux à deux et convergent presque surement (SLLN). Sous (Z-A2), les limites sont égales :

$$
\begin{aligned}
\lim_{n\to+\infty} \frac{1}{n}\,{}^t(Z \circ \bar{\Omega})(Z \circ \bar{\Omega}) &= \lim_{n\to+\infty} \frac{1}{n}\,{}^t(X^{real} \circ \bar{\Omega})(X^{real} \circ \bar{\Omega}) \\
&= \lim_{n\to+\infty} \frac{1}{n}\,{}^tX^{real}(X^{real} \circ \bar{\Omega}).
\end{aligned}
$$

Cependant, sous l'hypothèse (Z-A1), la covariance entre les variables est différente,

$$
\lim_{n\to+\infty} \frac{1}{n}\,{}^t(Z \circ \bar{\Omega})(Z \circ \bar{\Omega}) \neq \lim_{n\to+\infty} \frac{1}{n}\,{}^t(X^{real} \circ \bar{\Omega})(X^{real} \circ \bar{\Omega}).
$$

Ainsi la différence est égale à :

$$
-2 \sum_{k,j\in\{1,\ldots,p\},k<j} Cov_{kj}^{real}(1 - Q_kQ_j).
$$

Utilisant l'équation 6.6, $MSSE^{M_2}$ sous (X-A2) et (Z-A2) est égale à :

$$
\begin{aligned}
MSSE^{M_2} = MSSE &+ {}^t\beta^{M_2}\Sigma^{real}\beta^{M_2} - {}^t\beta\Sigma^{real}\beta \\
&+ 2\lim_{n\to+\infty} \frac{1}{n} \underbrace{{}^t\beta^{M_2}\,{}^tX^{real}Z \circ \bar{Q}\beta^{M_2}}_{\substack{=0 \\ \text{Indépendance (C1) et} \\ \text{les variables sont centrées}}}
\end{aligned}
$$

Posons I l'ensemble des indices des variables indépendantes de toutes les autres variables et KJ

l'ensemble des variables corrélées deux à deux.

$$\begin{aligned} MSSE^{M_2} = & \underbrace{MSSE}_{\text{Variance inhérente du modèle}} \\ & - \sum_{j \in I}^{p} (\beta_j)^2 Var(X_j^{real})(1 - Q_j) \\ & + \sum_{k,j \in KJ} Var(X_k^{real}) \left((\beta_k^{M_2})^2 - (\beta_k)^2 \right) \\ & + Var(X_j^{real}) \left((\beta_j^{M_2})^2 - (\beta_j)^2 \right) \\ & + 2Cov_{jk}^{real} \times \left(\beta_j^{M_2} \beta_k^{M_2} - \beta_j \beta_k \right). \end{aligned} \tag{6.7}$$

Le cas où $p = 2$, avec $\beta_1 = \beta_2/20 = 1$, $Var(X_1) = Var(X_2) = 1$ et $Q_1 = 0.95$, $Q_2 = 1$ et $\rho > 0,99$, la différence $MSSE^{M_2} - MSSE$ est positive. Ainsi dans certains cas, $MSSE^{M_2} > MSSE$ c'est-à-dire qu'une perte en qualité peut améliorer les performances du modèle linéaire.

Utilisant l'équation 6.6, $MSSE^{M_2}$ sous (X-A2) et (Z-A1) est égale à :

$$\begin{aligned} MSSE^{M_2} = & MSSE \\ & + {}^t\beta^{M_2} \Sigma^{real} \beta^{M_2} - {}^t\beta \Sigma^{real} \beta \\ & + 2 \lim_{n \to +\infty} \frac{1}{n} \underbrace{{}^t\beta^{M_2} {}^tX^{real} Z \circ \bar{Q} \beta^{M_2}}_{\substack{=0 \\ \text{Indépendance (C1) et} \\ \text{les variables sont centrées}}} \\ & + \lim_{n \to +\infty} \frac{1}{n} {}^t(Z \circ \bar{Q})(Z \circ \bar{Q}) - \lim_{n \to +\infty} \frac{1}{n} {}^t(X^{real} \circ \bar{Q})(X^{real} \circ \bar{Q}). \end{aligned}$$

Posons I l'ensemble des indices des variables indépendantes de toutes les autres variables et KJ l'ensemble des variables corrélées deux à deux.

$$\begin{aligned} MSSE^{M_2} = & \underbrace{MSSE}_{\text{Variance inhérente du modèle}} \\ & - \sum_{j \in I}^{p} (\beta_j)^2 Var(X_j^{real})(1 - Q_j) \\ & + \sum_{k,j \in KJ} Var(X_k^{real}) \left((\beta_k^{M_2})^2 - (\beta_k)^2 \right) \\ & + Var(X_j^{real}) \left((\beta_j^{M_2})^2 - (\beta_j)^2 \right) \\ & + 2Cov_{jk}^{real} \times \left(\beta_j^{M_2} \beta_k^{M_2} - \beta_j \beta_k \right) \\ & - 2Cov_{kj}^{real} \times (1 - Q_k Q_j) \beta_j^{M_2} \beta_k^{M_2}. \end{aligned} \tag{6.8}$$

Posons $\sigma_{X_j^{real}} = \sqrt{Var(X_j^{real})}$. L'équation précédente peut être écrite comme suit :

$$\begin{aligned} MSSE^{M_2} = & \underbrace{MSSE}_{\text{Variance inhérente du modèle}} \\ & - \sum_{j \in I}^{p} (\beta_j)^2 Var(X_j^{real})(1 - Q_j) \\ & + \sum_{k,j \in KJ} (\beta_k^{M_2}; \beta_j^{M_2}; \beta_k; \beta_j) H \, {}^t(\beta_k^{M_2}; \beta_j^{M_2}; \beta_k; \beta_j) \end{aligned}$$

où H est la matrice par blocs, définie comme suit :

$$H = \left(\begin{array}{cc|cc} \sigma^2_{X_j^{real}} & \rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} & 0 & 0 \\ \rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} & \sigma^2_{X_k^{real}} & 0 & 0 \\ \hline 0 & 0 & -\sigma^2_{X_j^{real}} & -\rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} \\ 0 & 0 & -\rho\sigma_{X_j^{real}}\sigma_{X_k^{real}} & -\sigma^2_{X_k^{real}} \end{array}\right)$$
$$- \left(\begin{array}{cc|cc} 0 & (1-Q_kQ_j)\rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} & 0 & 0 \\ (1-Q_kQ_j)\rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$
$$= \left(\begin{array}{cc|cc} \sigma^2_{X_j^{real}} & Q_kQ_j\rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} & 0 & 0 \\ Q_kQ_j\rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} & \sigma^2_{X_k^{real}} & 0 & 0 \\ \hline 0 & 0 & -\sigma^2_{X_j^{real}} & -\rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} \\ 0 & 0 & -\rho\sigma_{X_j^{real}}\sigma_{X_k^{real}} & -\sigma^2_{X_k^{real}} \end{array}\right).$$

La matrice H n'est pas définie positive. En utilisant le théorème sur (X-A2) du précédent article, la relation suivante peut être écrite :

$$\begin{pmatrix} \beta_k^{M_2} \\ \beta_j^{M_2} \end{pmatrix} = \underbrace{\frac{1}{1-Q_k^2Q_j^2\rho^2} \circ \begin{pmatrix} Q_k\,(1-Q_j^2\rho^2) & Q_k\,\rho(1-Q_j^2)\frac{\sigma_{X_j^{real}}}{\sigma_{X_k^{real}}} \\ Q_j\,\rho(1-Q_j^2)\frac{\sigma_{X_k^{real}}}{\sigma_{X_j^{real}}} & Q_j\,(1-Q_k^2\rho^2) \end{pmatrix}}_{\text{Noté } P_1} \begin{pmatrix} \beta_k \\ \beta_j \end{pmatrix}.$$

La partie supérieure de la matrice H se réécrit comme suit :

$$\begin{aligned} {}^tP_1 &\begin{pmatrix} \sigma^2_{X_j^{real}} & Q_kQ_j\rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} \\ Q_kQ_j\rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} & \sigma^2_{X_k^{real}} \end{pmatrix} P_1 \\ &= \frac{1}{F}\begin{pmatrix} \sigma^2_{X_j^{real}} & \rho\sigma_{X_k^{real}}\sigma_{X_j^{real}} \\ \rho\sigma_{X_j^{real}}\sigma_{X_k^{real}} & \sigma^2_{X_k^{real}} \end{pmatrix} \circ \begin{pmatrix} A & B \\ B & C \end{pmatrix}, \end{aligned}$$

où :

$$\begin{aligned} F =& (1-Q_j^2Q_k^2\rho^2)^2 \\ A =& Q_k^2(1-Q_j^2\rho^2)^2 \\ &+ 2Q_k^2Q_j^2(1-Q_j^2\rho^2)(1-Q_k^2)\rho^2 \\ &+ Q_j^2(1-Q_k^2)^2\rho^2, \\ C =& Q_j^2(1-Q_k^2\rho^2)^2 \\ &+ 2Q_j^2Q_k^2(1-Q_k^2\rho^2)(1-Q_j^2)\rho^2 \\ &+ Q_k^2(1-Q_j^2)^2\rho^2, \\ B =& Q_j^2(1-Q_k^2\rho^2)(1-Q_k^2) \\ &+ Q_k^2Q_j^2(1-Q_k^2\rho^2)(1-Q_j^2\rho^2) \\ &+ Q_k^2(1-Q_j^2\rho^2)(1-Q_j^2) \\ &+ Q_k^2Q_j^2(1-Q_k^2)(1-Q_j^2). \end{aligned}$$

Donc, la matrice 4×4 H peut être réécrite comme une matrice hermitienne 2×2,

$$\begin{pmatrix} \beta_k^{M_2} \\ \beta_j^{M_2} \\ \beta_k \\ \beta_j \end{pmatrix}^t H \begin{pmatrix} \beta_k^{M_2} \\ \beta_j^{M_2} \\ \beta_k \\ \beta_j \end{pmatrix} \propto (-1)^t \begin{pmatrix} \beta_k \\ \beta_j \end{pmatrix} \left(\begin{pmatrix} \sigma^2_{X_j^{real}} & \sigma_{X_k^{real}} \sigma_{X_j^{real}} \\ \sigma_{X_j^{real}} \sigma_{X_k^{real}} & \sigma^2_{X_k^{real}} \end{pmatrix} \circ \underbrace{\begin{pmatrix} F-A & \rho(F-B) \\ \rho(F-B) & F-C \end{pmatrix}}_{\text{Noté } M} \right) \begin{pmatrix} \beta_k \\ \beta_j \end{pmatrix}. \tag{6.9}$$

Mettons de côté le cas $Q_k, Q_j = \{1, 1\}$ qui correspond au cas trivial de la qualité parfaite où les deux modèles sont égaux et donc $MSSE = MSSE^{M_2}$. Comme Q_k sont Q_j interchangeables, nous allons étudier $Q_k \in]0, 1]$,$Q_j \in]0, 1[$.

Proposition 1

Wolkowicz et Styan, 1980 ont démontré qu'une matrice hermitienne carrée M est strictement définie positive si et seulement si $\frac{Tr(M)^2}{Tr(M^2)} > 1$ et $Tr(M) > 0$ où Tr est la trace de la matrice.

Avec les notations précédentes, les conditions peuvent se réécrire,

$$Tr(M) = (F - A) + (F - C) > 0,$$

et

$$\frac{Tr(M)^2}{Tr(M^2)} = \frac{((F-A)+(F-C))^2}{(F-A)^2 + \rho^2(F-B)^2 + (F-C)^2 + \rho^2(F-B)^2} > 1$$
$$\Leftarrow \rho^2(F-B)^2 < (F-A)(F-C).$$

Remarque : Dans cette partie, nous allons utiliser la proposition suivante pour prouver les deux conditions.

Proposition 2

Si un polynôme quadratique réel P(.) a son coefficient quadratique positif et qu'il existe $a, b, c \in \mathbb{R}$ avec $a \neq b \neq c, a \leq b$ et $c \notin]a, b[$ tel que :
— $P(a) \geq 0$; $P(b) \geq 0$; $P(c) < 0$,
alors la fonction polynomiale est strictement positive sur $]a, b[$.

La preuve de cette proposition est simple en remarquant qu'un polynôme quadratique réel à coefficient directeur positif est négatif au maximum sur un seul intervalle. Ainsi, nous allons montrer qu'il existe un ϵ strictement positif tel que $P(1 + \epsilon)$ est négatif et avec a et b égal respectivement 0 et 1.

Pour prouver $Tr(M) > 0$, Posons $K = Q_k^2$, $J = Q_j^2$, $P = \rho^2$. L'intervalle d'étude est $J \in]0, 1[$, $K \in]0, 1]$, $P \in]0, 1[$.

Pour $J \in]0, 1[$, $P \in]0, 1[$ et $K \in]0, 1]$, pour le polynôme $(F - A)$,

$$\rightarrow \text{Polynôme en K} : \mathbf{F - A = K^2 J\, P(1 - JP) - K(1 - J^2P^2) + 1 \times (1 - JP)}$$
$$\text{Quand K} = 0 : F - A = 1 - JP > 0,$$
$$\text{Quand K} = 1 : F - A = JP - J^2P^2 - 1 + J^2P^2 + 1 - JP = 0$$
$$\text{Quand K} = 1 + \epsilon : F - A = (2\epsilon + \epsilon^2)\, JP(1 - JP) - \epsilon(1 - J^2P^2)$$
$$= \epsilon((2 + \epsilon)\, JP(1 - JP) - (1 - J^2P^2)).$$

Si $\epsilon \in]0, \frac{1-JP}{JP}[$ alors $F - A < 0$. En utilisant la proposition précédente, en remarquant que le coefficient quadratique est positif, $I - A$ est strictement positif pour $J \in]0, 1[$, $P \in]0, 1[$ et $K \in]0, 1[$ et nulle si $K = 1$.

De façon similaire, pour $(F-C)$,

$$\begin{aligned}
&\rightarrow \text{Polynôme en KP} : \mathbf{F-C = K^2\ P^2 J(1-J) - KP(1-J^2) + 1\times(1-J)}\\
&\text{Quand KP} = 0 : F-C = 1-J > 0,\\
&\text{Quand KP} = 1 : F-C = J-J^2-1+J^2+1-J = 0,\\
&\text{Quand KP} = 1+\epsilon : F-C = (2\epsilon+\epsilon^2)J(1-J) - \epsilon(1-J^2)\\
&\qquad = \epsilon((2+\epsilon)J(1-J) - (1-J^2)).
\end{aligned}$$

Si $\epsilon \in]0, \frac{1-J}{J}[$ et $J \neq 0$ alors pour $KP = 1+\epsilon$, $F-C < 0$. Le coefficient quadratique est positif. Le polynôme $F-C$ est strictement positif pour toutes valeurs de $KP \in]0,1[$ impliquant pour toutes valeurs de $P \in]0,1[$ et $K \in]0,1]$ et pour $J \in]0,1[$.

Les deux résultats démontrent que $2F-A-C > 0$ et $Tr(M) > 0$ pour $J \in]0,1[$, $P \in]0,1[$ et $K \in]0,1]$.

Pour prouver que $\frac{Tr(M)^2}{Tr(M^2)} > 0$, l'inégalité suivante est suffisante

$$\rho^2(F-B)^2 < (F-A)(F-C).$$

Avec les notations précédentes, pour le polynôme $(F-B)$,

$$\begin{aligned}
&\rightarrow \text{Polynôme en K} : \mathbf{(F-B) = K^2 J(1-J) - K(1-J^2) + 1\times(1-J)} :\\
&\text{Quand K= } 0 : F-B = 1-J > 0,\\
&\text{Quand K} = 1 : F-B = 0,\\
&\text{Quand K} = 1+\epsilon : F-B = \ (2\epsilon+\epsilon^2)J(1-J) - \epsilon(1-J^2).\\
&\qquad = \epsilon((2+\epsilon)J(1-J) - (1-J^2))
\end{aligned}$$

Similairement à $(F-A)$, $(F-B)$ est positif sur l'intervalle considéré. En réutilisant les polynômes précédents :

$$\begin{aligned}
(F-A) &= K^2 J\ P(1-JP) - K(1-J^2P^2) + 1\times(1-JP),\\
(F-C) &= K^2\ P^2 J(1-J) - KP(1-J^2) + 1\times(1-J).
\end{aligned}$$

remarquons que :

— pour $K = 0$, $(F-A) > P(F-B)$ et $(F-C) = (F-B)$,

— pour K = 1, $(F-C) > (F-A) = (F-B)$.

$(F-B)-(F-C)$ est un polynôme quadratique en K et a un coefficient quadratique positif. Comme le polynôme est nul pour K = 0, et négatif pour K = 1, cela implique $(F-C) > (F-B)$ sur $K \in]0,1]$.

$(F-A) - P(F-B)$ a un coefficient quadratique positif. En utilisant le résultat précédent pour $\epsilon \in]0, \frac{1-P}{P}[$, $(F-A) - P(F-B)$ est strictement négatif pour $K = 1+\epsilon$. A l'aide de la proposition, l'inégalité $(F-A) > P(F-B)$ est prouvée sur l'ensemble de l'intervalle.

Rappelons que pour $J \in]0,1[$, $P \in]0,1[$ et $K \in]0,1[$, $I-A$, $I-B$ et $I-C$ sont strictement positifs. À l'aide des inégalités précédentes $(F-C) > (F-B)$ et $(F-A) > P(F-B)$, l'inégalité est finalement démontrée :

$$P(F-B)^2 < (F-A)(F-C) \Longrightarrow \frac{Tr(M)^2}{Tr(M^2)} > 1.$$

Ainsi la matrice M est définie positive. D'après le théorème du produit de Schur, si deux matrices hermitiennes 2×2, M_1, M_2 sont définies positives, alors $M_1 \circ M_2$ est défini positif. Ces théorèmes peuvent être étendus à une matrice M_1 semi-définie positive si tous ses termes sont non nuls et à M_2 définie positive.

Cependant, pour le cas K = 1, $\frac{Tr(M)^2}{Tr(M^2)} = 1$. Dans ce cadre, le calcul est simplifié car les polynômes $(F-A)$ et $(F-B)$ sont nuls :

$$
{}^{t}\begin{pmatrix} \beta_k^{M_2} \\ \beta_j^{M_2} \\ \beta_k \\ \beta_j \end{pmatrix} H \begin{pmatrix} \beta_k^{M_2} \\ \beta_j^{M_2} \\ \beta_k \\ \beta_j \end{pmatrix} \propto (-1) \; {}^{t}\begin{pmatrix} \beta_k \\ \beta_j \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2_{X_k^{real}}(F-C) \end{pmatrix} \begin{pmatrix} \beta_k \\ \beta_j \end{pmatrix} = -\beta_j^2 \sigma^2_{X_k^{real}}(F-C) \leq 0,
$$

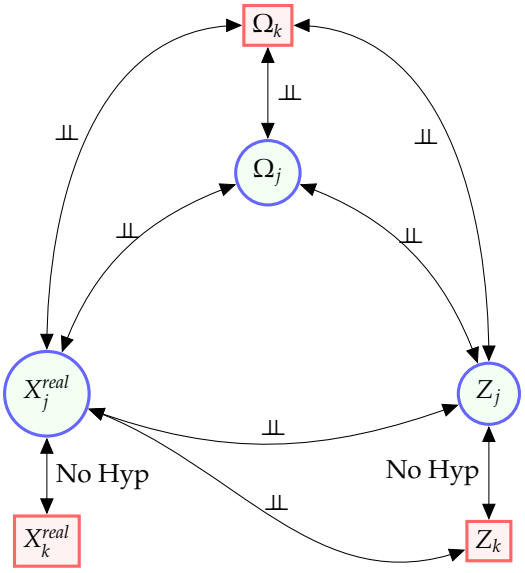
car (F-C) est positif pour K = 1. Dans l'équation 6.9, le premier terme est semi-défini positif avec tous ses termes non-nulles et M est défini positif. Ainsi, le produit d'Hadamard est défini positif.

Pour tout β_k et β_j, l'équation 6.9 est négative (strictement si β_k et β_j sont non nulles.). Utilisant l'équation 6.8, l'inégalité $MSSE \geq MSSE^{M_2}$ a été démontrée et la preuve est conclue en utilisant les relations entre les métriques (équations 6.1).

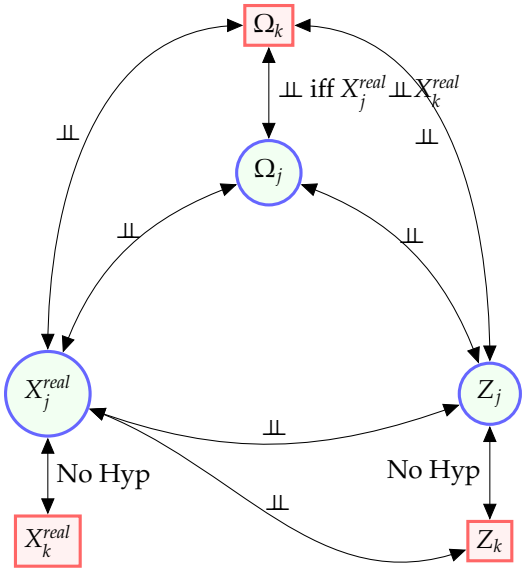
6.2 Le lien avec la théorie des valeurs manquantes

Comme il a été vu dans le chapitre 5, différentes hypothèses peuvent être prises pour la structure de qualité de donnée. Les subdivisions ont été très largement inspirées par Rubin [131, Rubin, 1976] et possèdent les mêmes limites. Différents cas peuvent être créés sur ($\mathbf{X}^{real}$, $\mathbf{Z}$, $\mathbf{\Omega}$). Considérons-les dans les figures 6.1a, 6.1b, 6.1c, 6.1d. Ces cas dépendent surtout du type de collection de variables et ils ont été inspirés largement des cas pratiques de la tarification à l'adresse. L'hypothèse usuelle est que l'information apportée sur la variable à prédire par $\mathbf{Z}$ n'est pas indépendante de $\mathbf{X}^{real}$; c'est-à-dire $\mathbf{Z}$ est informatif à travers la corrélation avec $\mathbf{X}^{real}$ sur $\mathbf{Y}$.

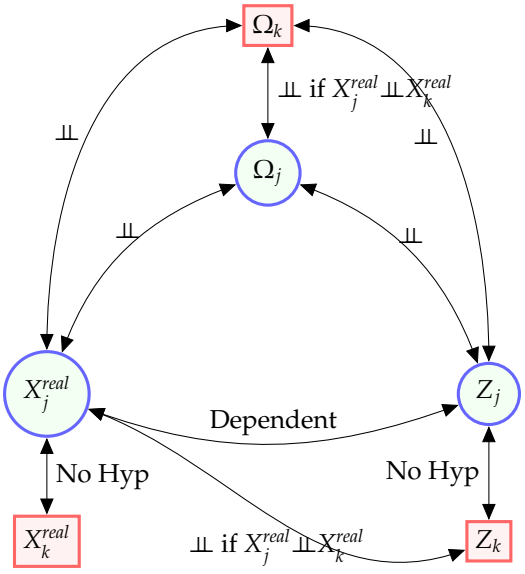
Par exemple, l'hypothèse MCAR (Missing Completly At Random) ([131, Rubins 1976], [126, Heitjan and Basu, 1996]) est un cas particulier de (C1) où les indices de qualités sont égaux uniquement à 0 et 1. L'indépendance multivariée entre les variables suppose que les erreurs sont indépendantes. En d'autres termes, chaque observation de chaque variable est recueillie à partir de sources différentes et



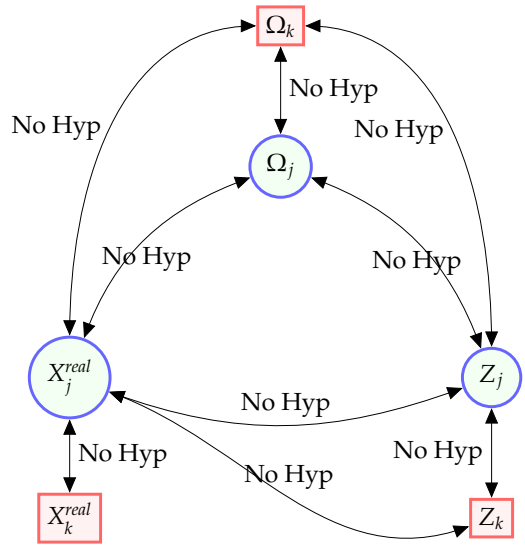
(a) Case (C1) - Incertitude totale ($j \neq k$). (No Hyp) Pas d'hypothèses.



(b) Cas (C2) - Incertitude locale avec des erreurs indépendantes.



(c) Case (C3) - Imprécision ($j \neq k$).



(d) Case (C4) ($j \neq k$).

FIGURE 6.1 – Les cas étudiés.

non liées ou avec des erreurs non liées. De la même manière, l'hypothèse MAR (Missing At Random) est un cas particulier de (C2) et (C3). En effet, elle correspond à une certaine dépendance entre les indices de qualité/observations manquantes. Dans le cas (C3), les valeurs erronées Z_j sont corrélées avec les valeurs réelles X_j^{real}. Un cas particulier est quand $(Z_j - X_j^{real})$ suit une distribution gaussienne et centrée. Le dernier cas (C4) est étroitement lié au cadre MNAR (Missing Not At Random), où une certaine dépendance existe entre chaque variable. Dans la plupart des cas rencontrés, Ω dépendent des valeurs de $\mathbf{X}^{real}$. Ainsi, les erreurs $\mathbf{Z}$ peuvent dépendre des vraies valeurs $\mathbf{X}^{real}$; les erreurs sont informatives rendant l'analyse plus complexe.

Dans la tarification à l'adresse, les variables sont souvent dans le cadre (C4) avec certaines dans le cas (C1) entre-elles, d'autres dans le cas (C2) et (C3). Entre deux variables dans une même base, certaines données peuvent être dans le cas (C1) provenant d'une même source alors que d'autres observations sont dans le cas (C2). Prenons par exemple la variable *surface habitable* et *nombre d'étages*. Lorsque la donnée la plus fiable provient du référentiel bâtiment, le cas (C2) est le plus appropriée, sinon la donnée de chacun provient de différentes sources (souvent départementales et notariat) et le cas (C1) est celui approprié. Dans le cas où la surface est inconnue et est calculée par un Light GBM Gamma à partir d'informations dont le *nombre d'étages*, le cas (C4) est obligatoire. Cet exemple est réel et fut présent lors de la réception de la cinquième base pour la modélisation des sinistres attritionnels.

Les différents cas (C1), (C2), (C3) et (C4) coexistant à l'intérieur d'une même variable explique le choix du terme *pattern* de qualité de donnée pour un individu. En effet, l'évaluation de l'indice de qualité de données d'une variable et son utilisation dépendent des autres variables et sont différents pour chaque individu. Pour $j \neq i$ même si les *pattern* sont égaux $\mathbf{Q}_i = \mathbf{Q}_j$, l'interprétation de la donnée dépend des hypothèses (C1), (C2), (C3) et (C4) entre les variables. Ainsi, le vecteur $\mathbf{Q}_i$ n'est pas suffisant pour qualifier la donnée. Il est nécessaire de comprendre le motif, la structure que forme la qualité de la donnée.

Dans le cadre linéaire, si l'identification individuelle des cas est faisable, cette prise en compte dans la modélisation n'est mathématiquement pas beaucoup plus difficile. Les résultats peuvent aussi être généralisés dans le cas (C3). Cependant, il est nécessaire de faire des hypothèses sur la dépendance sous forme de copules par exemple. La limite pratique souvent intervient dans la connaissance des données sources et la modélisation des dépendances. Dans l'exemple précédent, il est évident qu'il va être difficile d'obtenir un détail fin et de modéliser les dépendances sans étudier les modèles sous-jacentes.

Soit deux variables X_1 et X_2 telle qu'une proportion π de données soit dans le cadre (C1) et $1 - \pi$ dans le cadre (C2) avec $\Omega_1 = \Omega_2$ (Donc $Q_1 = Q_2$). Sans difficulté, la covariance observée $Cov_{1,2}$ s'écrit facilement en fonction de la covariance réelle $Cov_{1,2}^{real}$. Par exemple sous l'hypothèse (Z-A1) :

$$Cov_{1,2} = (\pi Q_1 Q_2 + (1 - \pi) Q_1) Cov_{1,2}^{real}.$$

La qualité de la donnée peut aussi être différente conditionnellement aux hypothèses. Tant que $\mathbf{\Omega}$ est indépendant de $\mathbf{X}^{real}$, les calculs se font facilement. L'idée est transposable pour les GLMs à travers la vraisemblance.

L'algorithme M_3, lui, ne change pas. En effet, le modèle réel est estimé par le modèle M1 en premier lieu puis il s'adapte au cas individuel de la qualité de données. Cela justifie aussi pourquoi nous avions choisi les étapes suivantes $\hat{\beta}^{M_2} \longrightarrow \hat{\beta}^{M_1} \longrightarrow \hat{\beta}^{K}$ et non $\hat{\beta}^{M_2} \longrightarrow \hat{\beta}^{K}$ directement. Rappelons aussi qu'en pratique, il est nécessaire de contrôler que les $\hat{\beta}^{M_1}$ sont cohérents.

6.2.1 La transitivité des indices de qualité

L'ensemble des théorèmes mentionnés dans les articles précédents compare une variable X avec une variable dite "idéale", X^{real} par rapport à une variable à prédire Y. Lorsque la notion de qualité est déterminée, il semble souhaitable d'introduire une notion d'ordre.

6.2.1.a La relation d'ordre des indices de qualité

Définition 1

Une relation d'ordre peut être définie dans le cadre (C1), c'est-à-dire que X est de meilleure qualité en termes de crédibilité X^{real} notée $X \leq_{Qc.Y.C1} X^{real}$ si on peut écrire :

$$\begin{aligned} X =& X^{real} \times \Omega + Z \times (1 - \Omega) \text{ et} \\ (Cov(X,Y))^2 \leq& (Cov(X^{real}, Y))^2 \end{aligned} \tag{6.10}$$

où $X \sim X^{real} \sim Z$ sont des variables aléatoires et indépendantes entre elles et avec Ω une variable booléenne. Ω et Z sont indépendants de Y.

En effet, les différentes propriétés de relation d'ordres sont vérifiées :

1. **Réflexivité :** $X \leq_{Qc.C1} X$ pour toutes variables X. En effet, c'est le cas particulier où Ω est toujours égale à 1 ;
2. **Antisymétrie :** $X_1 \leq_{Qc.Y.C1} X_2$ et $X_2 \leq_{Qc.Y.C1} X_1$ implique $X_1 = X_2$ pour toutes variables X_1 et X_2 ;
3. **Transitivité :** $X_1 \leq_{Qc.Y.C1} X_2$ et $X_2 \leq_{Qc.Y.C1} X_3$ implique $X_1 \leq_{Qc.Y.C1} X_3$.

Preuve 1: Antisymétrie

Soit X_2 et X_2 deux variables aléatoires telles que $X_1 \leq_{Qc.Y.C1} X_2$ et $X_2 \leq_{Qc.C1} X_1$. Il existe Ω_1,Ω_1,Z_1 et Z_2 tel que

$$X_1 = X_2 \times \Omega_2 + Z_2 \times (1 - \Omega_2) \tag{6.11}$$

Les relations d'ordres $X_1 \leq_{Qc.Y.C1} X_2$ et $X_2 \leq_{Qc.Y.C1} X_1$ impliquent que $Cov(X_1, Y)^2 = Cov(X_2, Y)^2$.

$$\begin{aligned} Cov(X_1, Y)^2 &= \Big(Cov(X_2 \times \Omega_2, Y) + Cov(Z_2 \times (1 - \Omega_2), Y)\Big)^2 \\ \text{(C1)} &= Cov(X_2 \times \Omega_2, Y)^2 \\ &= Cov(X_2, Y)^2 \mathbb{E}(\Omega_2)^2 = Cov(X_2, Y)^2. \end{aligned} \tag{6.12}$$

Cette équation est vraie dans l'unique cas où $\mathbb{E}(\Omega_2) = 1$. En d'autres termes que $\Omega_2 = 1$ presque sûrement et donc que $X_1 = X_2$.

Preuve 2: Transitivite

Soit X^{real} une variable aléatoire. Posons X_{Q_1} de qualité Q_1 par rapport X^{real}, c'est-à-dire que la variable X_{Q_1} est construite par le modèle latent suivant :

$$X_{Q_1} = X^{real} \times \Omega_{Q_1} + Z_{Q_1} \times (1 - \Omega_{Q_1}) \tag{6.13}$$

où $\mathbb{E}(\Omega_{Q_1}) = Q_1$. Posons de façon similaire X_{Q_2} par rapport à X_{Q_1},

$$X_{Q_2} = X_{Q_1} \times \Omega_{Q_2} + Z_{Q_2} \times (1 - \Omega_{Q_2}) \tag{6.14}$$

où $\mathbb{E}(\Omega_{Q_2}) = Q_2$. En appliquant successivement les deux équations avec les hypothèses d'indépen-

dances, on obtient :

$$\begin{aligned} X_{Q_2} =& X^{real} \times \Omega_{Q_1} \times \Omega_{Q_2} \\ &+ Z_{Q_2} \times \Omega_{Q_1} \times (1 - \Omega_{Q_2}) \\ &+ Z_{Q_1} \times (1 - \Omega_{Q_1}) \times \Omega_{Q_2} \\ &+ Z_{Q_2} \times (1 - \Omega_{Q_2}) \times (1 - \Omega_{Q_1}). \end{aligned} \tag{6.15}$$

Comme $X_{Q_1} \sim X^{real} \sim Z_{Q_1}$ et $X_{Q_2} \sim X_{Q_1} \sim Z_{Q_2}$, Z_{Q_1} a la même distribution que Z_{Q_2} et les mêmes structures de corrélation avec les autres variables. Comme les variables Ω sont des variables binaires, posons $\Omega_{Q_3} = \Omega_{Q_2} \times \Omega_{Q_1}$ et son espérance $Q_3 = \mathbb{E}(\Omega_{Q_3}) = Q_2 \times Q_1$. Il est donc possible d'écrire :

$$X_{Q_2} = X^{real} \times \Omega_{Q_3} + Z \times (1 - \Omega_{Q_3}). \tag{6.16}$$

où Z a la même distribution que Z_{Q_1} et que Z_{Q_2}. Pour finir il est nécessaire de vérifier que la variable X_{Q_2} est de qualité Q_3 par rapport à X^{real} dans le cadre (C1). Par l'hypothèse d'indépendance mutuelle, Z, Ω_{Q_3}, X^{real} sont indépendantes et Z, Ω_{Q_3} aussi avec Y. L'inégalité sur la corrélation se prouve directement avec le lemme du premier article.

Remarques :

— L'indépendance entre Ω_{Q_2} et Ω_{Q_1} n'est pas nécessaire, mais impactera la valeur de Q_3 ;

— Dans le cas (C2), la transitivité modifie les dépendances entre les variables Ω. Donc la transitivité pourrait être étendue sous conditions en particulier sur la corrélation entre les variables $\mathbf{Z}$. Ainsi la relation d'ordre serait dans le cadre de dépendance entre les erreurs et les vraies valeurs.

De plus, il est intéressant de comparer deux variables $\mathbf{X}_{Q_1}$ et $\mathbf{X}_{Q_2}$ de qualité respective Q_1 et Q_2 différent de 1. Cela permet d'étendre la notion d'ordre partiel pour tendre vers un ordre total sous l'hypothèse (C1) et une distribution donnée. Pour cela, démontrons la proposition suivante.

Enonce 2.1: Lien entre indices de qualité

Supposons que nous sommes dans le cadre (C1) et posons X^{real} une variable aléatoire et deux variables selon le modèle de variable latent avec des erreurs et des indices de qualité (X_{Q_1}, Q_1) et $(\mathbf{X}_{Q_2}, Q_2)$:

$$\begin{aligned} X_{Q_1} =& X^{real} \times \Omega_{Q_1} + Z_{Q_1} \times (1 - \Omega_{Q_1}), \\ X_{Q_2} =& X^{real} \times \Omega_{Q_2} + Z_{Q_2} \times (1 - \Omega_{Q_2}), \end{aligned}$$

où $\mathbb{E}(\Omega_{Q_1}) = Q_1$ et $\mathbb{E}(\Omega_{Q_2}) = Q_2$.

Si Z_{Q_2}, Z_{Q_1} sont indépendants et que l'ensemble des évènements $\{\Omega_{Q_1} = 1\}$ est inclus dans $\{\Omega_{Q_2} = 1\}$, la variable X_{Q_2} est de qualité Q_3 par rapport X_{Q_1} tel que Q_3 est égale à $\frac{Q_2}{Q_1} - 1$ dans le cadre (C1) ;

$$X_{Q_1} \leq_{Qc.Y.C1} X^{real}, X_{Q_2} \leq_{Qc.Y.C1} X^{real} \Longrightarrow X_{Q_1} \leq_{Qc.Y.C1} X_{Q_2}.$$

Preuve 3: Lien entre indices de qualité

Soit X^{real} une variable aléatoire.

Posons X_{Q_1} de qualité Q_1 par rapport X^{real}, c'est-à-dire que la variable X_{Q_1} est construite par le modèle latent suivant :

$$X_{Q_1} = X^{real} \times \Omega_{Q_1} + Z_{Q_1} \times (1 - \Omega_{Q_1}) \tag{6.17}$$

où $\mathbb{E}(\Omega_{Q_1}) = Q_1$.

Posons X_{Q_2} par rapport à X^{real} tel que

$$X_{Q_2} = X^{real} \times \Omega_{Q_2} + Z_{Q_2} \times (1 - \Omega_{Q_2}) \tag{6.18}$$

où $\mathbb{E}(\Omega_{Q_2}) = Q_2$. On peut réécrire de la manière suivante :

$$\begin{aligned} X_{Q_2} =& X^{real} \times \Omega_{Q_1} \times \Omega_{Q_2} \\ &+ Z_{Q_2} \times \Omega_{Q_1} \times (1 - \Omega_{Q_2}) \\ &+ X^{real} \times (1 - \Omega_{Q_1}) \times \Omega_{Q_2} \\ &+ Z_{Q_2} \times (1 - \Omega_{Q_2}) \times (1 - \Omega_{Q_1}). \end{aligned} \tag{6.19}$$

Notons Z_3 tel que

$$\begin{aligned} X_{Q_2} =& X_{Q_1} \times \Omega_{Q_1} \times \Omega_{Q_2} \\ &+ Z_3 \times \Omega_{Q_1} \times (1 - \Omega_{Q_2}) \\ &+ Z_3 \times (1 - \Omega_{Q_1}) \times \Omega_{Q_2} \\ &+ Z_3 \times (1 - \Omega_{Q_2}) \times (1 - \Omega_{Q_1}). \end{aligned} \tag{6.20}$$

Z_3 est indépendant de Y et de X_{Q_1} si et seulement si $P((1-\mathbf{\Omega}_{Q_1}) \times \Omega_{Q_2}) = 0) = 1$ ce qui est équivalent la condition de la proposition. Dans ce cas, X_{Q_2} est de qualité $E(\Omega_1\Omega_2) = \frac{Q_2}{Q_1} - 1$ par rapport X_{Q_1}. Les conditions sur les corrélations s'en déduisent facilement.

Pour finir, pour obtenir l'ordre total, il serait nécessaire d'exposer une relation d'ordre sous (C1), en liant tout X_1 et X_2 ayant une même distribution. Malheureusement, ce n'est toujours possible. Il suffit de créer un exemple sous (C3) pour s'en convaincre.

6.2.1.b Quelques extensions

La notion de qualité permet d'étendre différents théorèmes. La transitivité des indices de qualités permet d'étendre l'inégalité entre les R^2. En d'autres termes, le coefficient de détermination R^2 croît avec la qualité de donnée dans le cas (C1).

Dans cette section, le cas (C1) est supposé. Considérons la même structure pour un modèle LM que l'article 4 pour $(\mathbf{X}^{real}, \mathbf{Y}, \mathbf{\Omega}^1, \mathbf{X}^1)$ et pour $(\mathbf{X}^{real}, \mathbf{Y}, \mathbf{\Omega}^2, \mathbf{X}^2)$. Notons M_2^1 et M_2^2 les modèles associés.

Théorème 6.2.3

Sous l'hypothèse (X-A1) et (Z-A1) ou (Z-A2), si pour tout $j = 1, ..., n$,

$$X_j^1 \leq_{Qc.Y.C1} X_j^2$$

alors $(R^2)^{M_2^1} \leq (R^2)^{M_2^2}$.

Preuve 4: Preuve du théorème 6.2.3

L'inégalité $X_j^1 \leq_{Qc.Y.C1} X_j^2$ pour tout $j = 1, ..., n$ implique qu'il existe une multivariée variable booléenne $\mathbf{\Omega}^*$ tel que $(\mathbf{X}^2, \mathbf{Y}, \mathbf{\Omega}^*, \mathbf{X}^1)$ suit la structure de données sous $(C1)$. En appliquant le théorème 6.1.1, on a bien $(R^2)^{M_2^1} \leq (R^2)^{M_2^2}$.

L'extension s'applique aussi sous les hypothèses (X-A2) et (Z-A1) pour les modèles linéaires.

De la même façon, sous les hypothèses (X-A3) et (Z-A1), si $X_p^1 \leq_{Qc.Y.C1} X_p^2$ et $X_j^1 = X_j^2$ pour tout $j \in 0, ..., p-1$, alors pour un modèle log-Poisson,

$$\begin{aligned} &\beta_j^{M_2^1} = \beta_j^{M_2^2}, \quad j \in 0, ..., p-1 \\ &\beta_p^{M_2^1} \in [0, \beta_p^{M_2^2}]. \end{aligned}$$

En bref, la relation d'ordre sur la qualité permet d'étendre l'ensemble des théorèmes.

6.3 La qualité des données et la tarification à l'adresse

La structure sur la qualité de données bien que simpliste permet de comprendre les problématiques de la tarification à l'adresse. La partie 6.3.1 fait le lien avec le géocoding. La partie 6.3.2 regarde la qualité de données sous l'angle de l'anti-sélection.

6.3.1 L'évolution de la qualité du géocodage

Dans le cadre de la tarification à l'adresse, la qualité des variables a évolué en fonction de la qualité du géocodage. Ainsi entre deux livraisons de bases de données, des évolutions très importantes sur le géocodage faisaient changer l'ensemble de la qualité des variables et tous les coefficients en même temps. Dès lors, il semble souhaitable d'utiliser ce qui a été fait précédemment pour évaluer l'amélioration du géocodage mais aussi justifier les évolutions des variables.

D'un point de vue pratique, la théorie qui a été mise en place se limite à un nombre de variables corrélées égal à deux et sans prise en compte de splines. Supposons deux variables X_1 la surface habitable et X_2 le nombre de bâtiments à 50 mètres. La corrélation entre ces deux variables est égale à 0.3, *c-à-d* $\rho = 0.3$.

Supposons ainsi $\beta_1 = -\beta_2 = 1$, que les variances sont unitaires et les variables centrées. Posons τ_{geo} le taux de bon géocodage.

Le taux de géocodage impacte la qualité des deux variables et les relations suivantes sur la qualité de données peuvent être justifiées par la qualité de données :

$$Q_1(\tau_{geo}) = \tau_{geo}^2, \quad Q_2(\tau_{geo}) = 0.5 + \tau_{geo}/2. \tag{6.21}$$

En effet, le géocodage influe deux fois sur la surface habitable, la première fois pour relier la source de donnée au référentiel bâtiment et la seconde pour relier le bâtiment à l'adresse de l'assureur. Pour le nombre de bâtiments à 50 mètres, le taux de géocodage ne dépend que du lien adresses-bâtiments et à cause de l'auto-corrélation spatiale, l'impact du géocodage est moindre et borné inférieurement.

Pour la première base reçue, les qualités des variables peut-être estimées avec les valeurs comme suit $\bar{Q}_1 = 0.56$ et $\bar{Q}_2 = 0.875$ avec $\tau_{geo} = 0.75$. Supposons que la corrélation entre les erreurs sont sous l'hypothèse (Z-A1).

Dans ce cadre, les coefficients obtenus seraient égaux à :

$$\beta_1^{M_2} \approx 0.578, \quad \beta_2^{M_2} \approx 1.053. \tag{6.22}$$

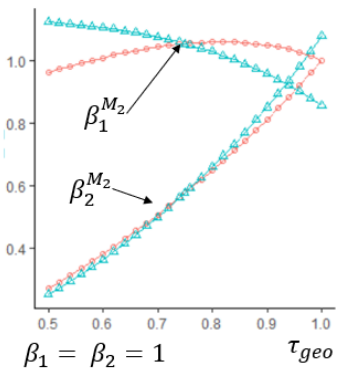
À partir des formules précédentes, l'impact de l'amélioration du géocodage est estimable. Donc si τ_{geo} augmente de 5 %, les coefficients attendus seraient :

$$\beta_1^{M_2} = 0.649, \quad \beta_2^{M_2} = 1.058. \tag{6.23}$$

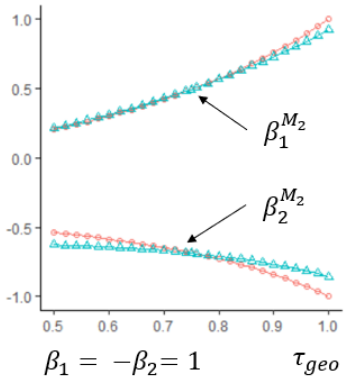
Néanmoins, l'amélioration du géocodage peut aussi être plus localisée par exemple en améliorant l'association des bons bâtiments et en détectant les bonnes dépendances. Ainsi, supposons que la qualité X_2 ne change pas et posons :

$$\begin{gathered} Q_1(\tau_{geo}) = \tau_{geo}^2, \quad Q_2(\tau_{geo}) = 0.875, \\ \text{Pour une augmentation de 5 points de } \tau_{geo}^2 : \\ \beta_1^{M_2} = 0.660, \quad \beta_2^{M_2} = 1.026. \end{gathered} \tag{6.24}$$

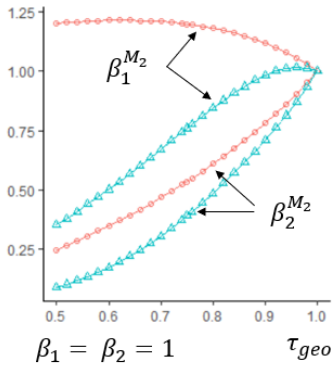
Cette structure de modélisation de la qualité permet de comprendre l'évolution des coefficients. En fonction de la structure les coefficients évoluent de façon radicalement différemment. Ainsi en fonction de l'évolution observée, il est possible de comprendre les évolutions du géocodage. Les figures 6.2a et 6.2b montrent que les évolutions attendues sont différentes selon la structure de la qualité des données, et aussi selon le signe des coefficients. La figure 6.2c montre qu'il est même possible d'avoir des augmentations simultanées linéaires même avec de la corrélation élevée 0.4. Ce type de structure correspond aux variables *nombre de pièces* et *surface* pour les zones rurales et l'autre en zone urbaine. En



(a) Exemple avec $\rho = 0.3$, les lignes en rouges et des points circulaires correspondent à $Q_1(\tau_{geo}) = \tau_{geo}^2$ et $Q_2(\tau_{geo}) = 0.5 + \tau_{geo}/2$, les lignes en bleues et des triangles à $Q_1(\tau_{geo}) = \tau_{geo}^2$ et $Q_2(\tau_{geo}) = 0.875$.



(b) Exemple avec $\rho = 0.3$, les lignes en rouges et des points circulaires correspondent à $Q_1(\tau_{geo}) = \tau_{geo}^2$ et $Q_2(\tau_{geo}) = 0.5 + \tau_{geo}/2$, les lignes en bleues et des triangles à $Q_1(\tau_{geo}) = \tau_{geo}^2$ et $Q_2(\tau_{geo}) = 0.875$.



(c) Exemple avec $\rho = 0.4$, les lignes en rouges et des points circulaires correspondent à $Q_1(\tau_{geo}) = \tau_{geo}^2$ et $Q_2(\tau_{geo}) = 0.75 + \tau_{geo}/4$, les lignes en bleues et des triangles à $Q_1(\tau_{geo}) = \tau_{geo}^2$ et $Q_2(\tau_{geo}) = \tau_{geo}^4$.

effet, le taux de géolocalisation impacte peu les surfaces habitables en zone rurale, car elles sont souvent équivalentes dans un voisinage proche contrairement en zone urbaine. En revanche, l'estimation du nombre de pièces est plus difficile et dépend beaucoup du type de maison. En résumé à l'intérieur d'une même variable, les dépendances de la qualité des données n'est pas la même.

Dans le cadre de la tarification à l'adresse, c'est le géocodage qui impacte le plus la qualité des variables. Les exemples sont bien sûr succincts et simples. Il faut ajouter de manière générale que l'évolution du géocodage n'est pas homogène sur le territoire, tout comme les corrélations et les coefficients. Ainsi pour mieux comprendre et expliquer les évolutions des coefficients, il est nécessaire de se restreindre sur de petits périmètres et un nombre de variables faibles.

6.3.2 L'impact de la qualité de données sur l'anti-sélection

La qualité d'une variable impacte les coefficients des autres variables qui lui sont corrélées. En utilisant un simple exemple linéaire, il est facile d'observer les effets de l'anti-sélection qu'induit la qualité de la donnée.

Posons $\Pi = Y$ une prime à déterminer par quatre variables X_1, X_2, X_3 et X_4. Elles suivent respectivement $\Gamma(3,3)$, $Exp(2)$, $\Gamma(2,2)$ et $Norm(0,10)$ et ont une structure de dépendance représentée par une copule gaussienne avec la matrice de corrélation suivante :

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0.2 & 1 \end{bmatrix}.$$

Supposons que la vraie structure du risque sous-jacent est linéaire :

$$E[Y|\mathbf{X}^{real} = \mathbf{x}^{real}] = \beta_0 + \beta_1 x_1^{real} + \beta_2 x_2^{real} + \beta_3 x_3^{real} + \beta_4 x_4^{real},$$

avec $\beta = (1000, 30, 30, 10, 5)$.

Comparons les tarifs des compagnies d'assurances, l'une ayant à disposition X_1^{real}, X_2^{real} et X_3^{real}, la seconde X_1^{real}, X_2^{real}, X_3^{real} et X_4. La variable X_4 est supposée comme étant l'unique variable ayant des problèmes de qualités (Hypothèse X-A3). Les trois autres variables sont supposées d'une qualité parfaite (provenant de souscription). Cela correspondrait à un ajout de données externes à une calculette tarifaire robuste.

Pour la simulation, l'indice de qualité Q_4 est distribué comme une variable discrète sur les valeurs (0.5, 0.8, 1) équi-probablement, c'est-à-dire $\bar{Q}_4 \approx 0.76$. La première compagnie d'assurance A utilise le modèle suivant :

$$\Pi^A = \hat{\beta}_0 + \hat{\beta}_1 X_1^{real} + \hat{\beta}_2 X_2^{real} + \hat{\beta}_3 X_3^{real} + \hat{\beta}_4 X_4.$$

La seconde B utilise $\Pi^B = \hat{\beta}_0 + \hat{\beta}_1 X_1^{real} + \hat{\beta}_2 X_2^{real} + \hat{\beta}_3 X_3^{real}$.

Le tableau 6.1 détaille le résultat des deux modèles linéaires. Pour la même base d'entrainement 7×10^5 lignes et de test 3×10^5 lignes, le R^2 ajusté est meilleur en utilisant les données externes. Cependant, même si la performance au global est améliorée, ce n'est pas toujours le cas à une maille individuelle.

Companie	R^2 Ajusté	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
A - réel	0.998	1000.00	30.02	29.99	10.00	4.996
A - M2	0.9236	995.48	30.16	29.95	10.79	3.77
A - M3(K=1)	Indisponible	996.74	30.16	29.95	10.45	4.7
B	0.821	993.05	30.18	29.96	11.88	0

TABLE 6.1 – Comparaison des coefficients entre les différents modèles entrainés sur les mêmes bases. Le modèle M_1 est inconnu en pratique.

En comparant la distribution des résidus - figure 6.4 - de chacun des modèles sur la base de test, les résidus de la compagnie A sont meilleures que ceux de la compagnie B mais toujours moins performants que ceux du modèle M_1. Sans utiliser X_4, les résidus de B sont orthogonaux aux résidus du modèle réel M_1 - figure 6.3.

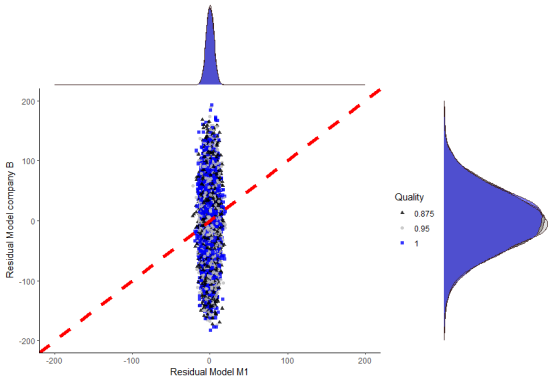


FIGURE 6.3 – Comparaison des résidus entre le modèle réel et le modèle de la compagnie B.

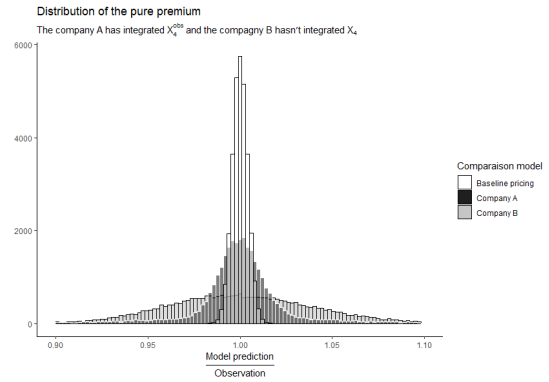


FIGURE 6.4 – Histogramme des résidus de chacun des modèles.

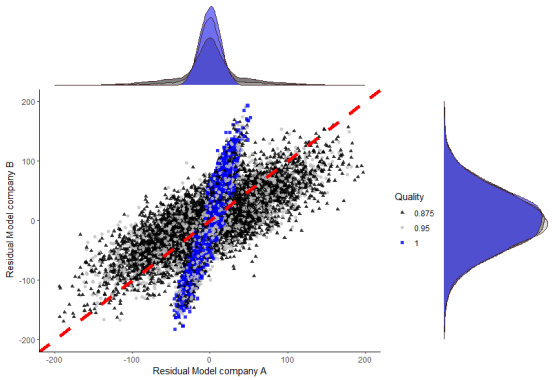


FIGURE 6.5 – Comparaison des résidus entre les modèles de la compagnie A – M_2 et B.

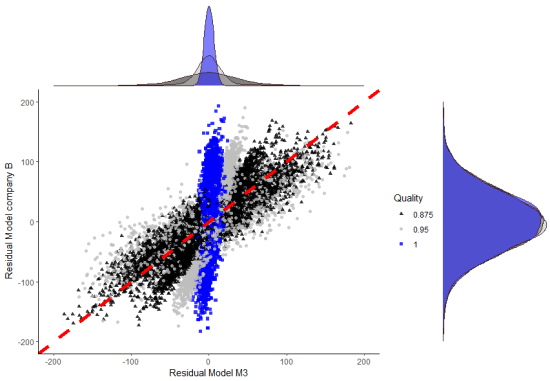


FIGURE 6.6 – Comparaison des résidus entre les modèles de la compagnie A – M_3 et B.

Le modèle de la compagnie A n'a pas la même distribution spatiale face aux résidus du modèle M_1 que la compagnie B - figure 6.5. Il est important de remarquer que les intervalles des erreurs sont identiques, mais que pour un individu $\mathbf{\Omega_4} = 1$ il n'y a pas la même distribution des résidus que pour le modèle réel. Ce biais est réduit avec l'algorithme M_3 comme le montre la figure 6.6. La distribution des résidus converge vers celle du modèle B ($M_3(Q_4 = 0)$) quand la qualité diminue.

Supposons que les prospects ne soient intéressés que par le montant de la cotisation/prime pour choisir la compagnie d'assurances. Ils vont comparer uniquement les compagnies A et B. Ici, on s'affranchit de la volatilité de la base d'entrainement en regardant quand n tend vers l'infini.

Enonce 3.1: Profit face au modèle réel ou $M1$

Supposons un duopole avec deux compagnies A and B utilisant $\mathbf{X}$,$\boldsymbol{X}^{real}$ $\mathbf{Q}$ et $\mathbf{Y}$ comme définis dans cette section. Supposons que les prospects soient intéressés que par le montant de la prime. Posons que $\beta_4 \neq 0$ et que X_4 n'est pas une distribution dégénérée. Si la compagnie A utilise le modèle réel ou M_1 (ce qui veut dire que la compagnie accède à X_4^{real} lors de la souscription) et la compagnie B utilise l'un des modèles M_2, M_3 ou n'utilise pas la variable X_4 dans son tarif (équivalent à $M_3(Q_4 = 0)$), la rentabilité des compagnies est connue :

$$\begin{aligned} \frac{C}{P}^{A-M1} &= 100\%, \\ \frac{C}{P}^{B} &> 100\%. \end{aligned} \tag{6.25}$$

Pour les modèles emboités comme pour le modèle M1 - de la compagnie A et contre la compagnie B, le modèle le plus segmentant implique $\frac{C}{P} = 100\%$ et les autres modèles obtiennent $\frac{C}{P} > 100\%$ à cause de l'anti-sélection. Ici, on suppose qu'il n'y a pas de chargement ni de modèle de marge. Le raisonnement ne se fait que sur la prime nette. En réalité, l'impact financier d'une meilleure segmentation s'optimise sur des aspects de transformations, de marges et de rétentions. Après cette optimisation, le ratio de la prime commerciale $\frac{C}{P}$ du modèle M_1 sera meilleur que tous les autres modèles.

Preuve 1: Preuve de l'énoncé 3.1

Définissons I_A, I_B et $I_{A=B}$ l'ensemble des individus ayant choisi respectivement les compagnies A, B ou ne pouvant choisir entre l'un ou l'autre.

Comme les individus ne choisissent qu'en fonction du prix, cela revient à dire que par exemple $\Pi_i^A < \Pi_i^B$ pour $i \in I_A$.

La compagnie A utilise M_1 et la compagnie B utilise soit M_2, soit M_3 ou, soit celui sans la variable X_4. Pour i appartenant I_A et $I_{A=B}$, la différence espérée entre la prime et la sinistralité est égale :

$$\begin{aligned} \mathbb{E}(\Pi_A(i) - C(i)) &= \beta_0^{M1} - \beta_0^{M1} + \underbrace{(\beta_1^{M1} - \beta_1^{M1})}_{=0} \; x_{i;1}^{real} \\ &+ (\beta_2^{M1} - \beta_2^{M1}) \; x_{i;2}^{real} + (\beta_3^{M1} - \beta_3^{M1}) \; x_{i;3}^{real} \\ &+ \underbrace{(\beta_4^{M1} - \beta_4^{M1})}_{X_4^{real}\text{Connu}=0} \; x_{i;4}^{real} + \mathbb{E}(\epsilon_i) \\ &= 0. \end{aligned} \tag{6.26}$$

Conséquemment, pour i in I_A,

$$\frac{C}{P}^{A-M1}(i) = 100\%.$$

Soit $i \in I_B$. On a $\mathbb{E}(\Pi_B(i) - C(i)) < \mathbb{E}(\Pi_A(i) - C(i)) = 0$. Comme il existe au moins i tel que $\mathbb{E}(\Pi_B(i) - C(i)) < 0$ car $\beta_4 \neq 0$ et la distribution X_4^{real} est non dégénérée alors $\frac{C}{P}^{B}(i) > 100\%$. CQFD.

Cette preuve ne nécessite comme uniques hypothèses que les distributions soient non dégénérées et que $\beta_4 \neq 0$.

Cependant, si on compare les modèles M_2, M_3 avec $M_3(Q_p = 0)$, les résultats sont plus complexes. La table 6.2 montre un exemple. Même si le modèle de la compagnie A segmente mieux que la compagnie B, la simulation montre que $\frac{C}{P}^{A} < \frac{C}{P}^{B}$ mais la qualité déséquilibre le ratio $\frac{C}{P}^{A-M3} \neq 100\%$. En effet, si la variable X_4 est fausse pour un individu i, la prime proposée correspond à une perturbation aléatoire d'impact : $\hat{\beta}_4 Z_{i;4}$ et la différence avec la compagnie B est égale à $\hat{\beta}_4^{Q_4=q_{i;4}} \times Z_{i;4} + (\hat{\beta}_3^{Q_4=0} - \hat{\beta}_3^{Q_4=q_{i;4}}) \times X_{i;3}$. Dans ce monde concurrentiel, le souscripteur i accepte l'assurance de la compagnie A si la marge est

Comparaison	Part du marché A (%)	C/P A (%)	C/P B (%)
Réel - Compagnie A	49.73 ± 0.56	100.00 ± 0.01	103.9 ± 0.01
M2 - Compagnie A	49,66 ± 0.57	99.99 ± 0.03	102,4 ± 0.01
M3 - Compagnie A	49,60 ± 0.57	99.85 ± 0.04	102,4 ± 0.01

TABLE 6.2 – Avec un intervalle de confiance à 95% sur la base de test. Le tarif de la compagnie B se base uniquement sur X_1, X_2 et X_3.

plus faible que la compagnie B. Ainsi, la prime proposée possède une composante aléatoire décorrélée du risque induisant une anti-sélection. Le modèle M_3 réduit l'impact de l'anti-sélection par *pattern K* de qualité de données. La transition d'une modèle M_2 vers un modèle M_3 réduit logiquement cet effet que pour la compagnie A (Dans la simulation). Ainsi $\frac{C}{P}^B$ n'est pas modifié comme le montre le tableau 6.2. Dans l'exemple, la copule gaussienne et les marginales utilisées impliquent que plus d'individus vont choisir la compagnie A avec le modèle M_2 que le modèle M_3 - comme le montre les figures 6.5 et 6.6.

Enonce 3.2: Profit face au modèle M_2

Supposons le duopole des compagnies A et B comme précédemment mentionnée sur les mêmes bases d'entrainements $\mathbf{X}$, $\mathbf{Q}$ et $\mathbf{Y}$. Les prospects sont intéressés uniquement au montant de la cotisation. Supposons que les variables X_3 et X_4 sont gaussiennes non dégénérées, centrées et jointes par une copule bivariée normale de paramètre ρ.

Si la compagnie A utilise le modèle M_2 et la compagnie B uniquement les trois variables X_1, X_2 et X_3,

$$\text{L'affirmation suivante est vraie : } \frac{C}{P}^B > 100\%.$$
$$\text{L'affirmation suivante peut être fausse : } \frac{C}{P}^{A-(M2)} < 100\%.$$

Preuve 2: Le profit du modèle M2 n'est pas toujours positif

En repartant des équations précédentes et posant la variance de X_3 et X_4 égale à respectivement σ_3, σ_4, le théorème de l'article 4, permet d'écrire pour un assuré i ayant les caractéristiques connues $(x_{i;3}^{real}, x_{i;4})$:

$$\begin{aligned}
\mathbb{E}(\Pi_A(i) - C(i) | (X_3^{real}, X_4) = (x_{i;3}^{real}, x_{i;4})) &= \beta_4 \frac{\sigma_3}{\sigma_4} \frac{\rho(1-\bar{Q_4}^2)}{(1-\rho^2\bar{Q_4}^2)} x_{i;3}^{real} \\
&+ \beta_4 \times (Q_4 \times x_{i;4} - \mathbb{E}(x_{i;4}^{real} | (X_3^{real}, X_4) = (x_{i;3}^{real}, x_{i;4})), \\
\mathbb{E}(\Pi_B(i) - C(i) | (X_3^{real}, X_4) = (x_{i;3}^{real}, x_{i;4})) &= \beta_4 \frac{\sigma_3}{\sigma_4} \rho x_{i;3}^{real} \\
&- \beta_4 \mathbb{E}(x_{i;4}^{real}) | (X_3^{real}, X_4) = (x_{i;3}^{real}, x_{i;4})), \\
\Pi_A(i) - \Pi_B(i) = \beta_4 \frac{\sigma_3}{\sigma_4} (\rho \underbrace{\frac{(1-\bar{Q_4}^2) - (1-\rho^2\bar{Q_4}^2)}{(1-\rho^2\bar{Q_4}^2)}}_{\text{Noté } H} x_{i;3}^{real} &+ \bar{Q}_4 x_{i,4}).
\end{aligned} \tag{6.27}$$

Posons dans un premier temps $\beta_4\rho > 0$. Comme pour certains assurés le profit espéré sera négatif et pour d'autres, il sera positif, il est donc nécessaire de regarder le profit moyen sur tous

les contrats. Sous l'hypothèse d'une copule bivariée gaussienne, avec f_i la fréquence d'un assuré i,

$$\begin{aligned}
&\sum_{i\in I_A\cup I_{A=B}} f_i \mathbb{E}(\Pi_A(i) - C(i) | (X_3^{real}, X_4) = (x_{i;3}^{real}, x_{i;4})) \\
&= \sum_{i\in I_A\cup I_{A=B}} f_i\beta_4 \frac{\sigma_3}{\sigma_4} \frac{\rho(1-\bar{Q}_4^{\,2})}{(1-\rho^2\bar{Q}_4^{\,2})} x_{i;3}^{real} \\
&+ f_i\beta_4(\bar{Q}_4 x_{i,4} - \underbrace{\mathbb{E}(X_{i;4}^{real} | X_3^{real} = x_{i;3}^{real})}_{\text{Var. multivariées gauss. : } x_{i;3}^{real}\rho}),
\end{aligned} \tag{6.28}$$

converge vers [a]

$$\beta_4 \int_{-\infty}^{\infty}\int_{-\infty}^{-sH^{-1}\rho^{-1}\bar{Q}_4} \frac{\sigma_3}{\sigma_4}\frac{\rho(1-\bar{Q}_4^{\,2})}{(1-\rho^2\bar{Q}_4^{\,2})} t + s\bar{Q}_4 - \rho \times t dF_{X_3^{real},X_4}(t,s). \tag{6.29}$$

Si $\Omega = 0$, le profit s'écrit, pour $\rho > 0$:

$$\begin{aligned}
&\beta_4\bar{Q}_4 \int_{-\infty}^{\infty}\int_{-\infty}^{-tH\rho\bar{Q}_4^{-1}} s dF_{X_3^{real}}(t) dF_{Z_4}(s) \\
&+\beta_4\frac{\sigma_3}{\sigma_4}\left(\frac{(1-\bar{Q}_4^{\,2})}{(1-\rho^2\bar{Q}_4^{\,2})} - \rho\right) \int_{-\infty}^{\infty}\int_{-\infty}^{-sH^{-1}\rho^{-1}\bar{Q}_4} t dF_{X_3^{real}}(t) dF_{Z_4}(s) \\
&= \underbrace{\beta_4\bar{Q}_4 \frac{1}{\sqrt{2\pi}} exp(-1-(H\rho\bar{Q}_4^{-1})^2)}_{\text{noté } C_1} \\
&\underbrace{-\beta_4\frac{\sigma_3}{\sigma_4}\left(\frac{(1-\bar{Q}_4^{\,2})}{(1-\rho^2\bar{Q}_4^{\,2})} - \rho\right)\frac{1}{\sqrt{2\pi}} exp(-1-(H^{-1}\rho^{-1}\bar{Q}_4)^2)}_{\text{noté } C_2}.
\end{aligned} \tag{6.30}$$

Si $\rho < 0$, le profit est égal à :

$$\begin{aligned}
&\beta_4\bar{Q}_4 \frac{1}{\sqrt{2\pi}}(-exp(-1-(H\rho\bar{Q}_4^{-1})^2)) \\
&+ \beta_4\frac{\sigma_3}{\sigma_4}\left(\frac{(1-\bar{Q}_4^{\,2})}{(1-\rho^2\bar{Q}_4^{\,2})} - \rho\right)\frac{1}{\sqrt{2\pi}} exp(-1-(H^{-1}\rho^{-1}\bar{Q}_4)^2).
\end{aligned} \tag{6.31}$$

Si $\Omega = 1$, le profit s'écrit $\rho > 0$:

$$\begin{aligned}
&\beta_4 \frac{\sigma_3}{\sigma_4}\left(\frac{(1-\bar{Q}_4^{\,2})}{(1-\rho^2\bar{Q}_4^{\,2})} - \rho\right) \int_{-\infty}^{\infty}\int_{-\infty}^{-sH^{-1}\rho^{-1}\bar{Q}_4} t dF_{X_3^{real},X_4^{real}}(t,s) \\
&\qquad + \beta_4 \bar{Q}_4 \int_{-\infty}^{\infty}\int_{-\infty}^{-tH\rho\bar{Q}_4^{-1}} s dF_{X_3^{real},X_4^{real}}(t,s). \\
&= \beta_4 \frac{\sigma_3}{\sigma_4}\left(\frac{(1-\bar{Q}_4^{\,2})}{(1-\rho^2\bar{Q}_4^{\,2})} - \rho\right) \int_{-\infty}^{\infty}\int_{-\infty}^{-tH\rho\bar{Q}_4^{-1}} t f_{X_3^{real}|X_4^{real}}(t)dt f_{X_4^{real}}(s)\, ds \\
&\qquad + \beta_4 \bar{Q}_4 \int_{-\infty}^{\infty}\int_{-\infty}^{sH^{-1}\rho^{-1}\bar{Q}_4} s f_{X_3^{real}|X_4^{real}}(t)dt f_{X_4^{real}}(s)\, ds. \\
&= \underbrace{-\beta_4 \frac{\sigma_3}{\sigma_4}\left(\frac{(1-\bar{Q}_4^{\,2})}{(1-\rho^2\bar{Q}_4^{\,2})} - \rho\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}} exp\left(-1 - \frac{(H^{-1}\rho^{-1}\bar{Q}_4 - \rho)^2}{1-\rho^2}\right)}_{\text{noté } C_3} \\
&\qquad \underbrace{+ \beta_4 \bar{Q}_4 \frac{1}{\sqrt{2\pi(1-\rho^2)}} exp\left(-1 - \frac{(H\rho\bar{Q}_4^{-1} - \rho)^2}{1-\rho^2}\right)}_{\text{noté } C_4}
\end{aligned} \tag{6.32}$$

ou si $\rho < 0$, est égal à :

$$\begin{aligned}
&\beta_4 \frac{\sigma_3}{\sigma_4}\left(\frac{(1-\bar{Q}_4^{\,2})}{(1-\rho^2\bar{Q}_4^{\,2})} - \rho\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}} exp\left(-1 - \frac{(H^{-1}\rho^{-1}\bar{Q}_4 - \rho)^2}{1-\rho^2}\right) \\
&\qquad -\beta_4 \bar{Q}_4 \frac{1}{\sqrt{2\pi(1-\rho^2)}} \left(1 - exp\left(-1 - \frac{(H\rho\bar{Q}_4^{-1} - \rho)^2}{1-\rho^2}\right)\right).
\end{aligned} \tag{6.33}$$

Le profit espéré s'écrit donc sous (C1) (indépendance de la qualité et des caractéristiques des assurés) :

$$\begin{aligned}
&\mathbb{E}(\Omega_4)(C_3 + C_4) + (1 - \mathbb{E}(\Omega_4))(C_1 + C_2) \\
&= \mathbb{E}(\Omega_4)(C_3 + C_4 - C_1 - C_2) + (C_1 + C_2).
\end{aligned} \tag{6.34}$$

En d'autres termes, si $(C_3 + C_4) \neq (C_1 + C_2)$, le profit attendu sera maximum soit pour $\Omega = 1$ ou $\Omega = 0$. Si on suppose que $\bar{Q}_4$ est égale à $\mathbb{E}(\Omega_4)$, c'est-à-dire que l'indice est parfaitement mesuré, le profit est égal à

$$Q_4(C_3 + C_4) + (1 - Q_4)(C_1 + C_2). \tag{6.35}$$

Le cas $\beta_4\rho < 0$ fait apparaitre la même formule pour le profit.

Pour un rapport $\frac{\sigma_3}{\sigma_4}$ est égal à 1 et $\beta_4 = 1$, il existe un choix pour (ρ, Q_4) tel que le profit est négatif comme $(-0.844, 0.949)$ pour un profit -1.22×10^{-6}. Plus précisément dans une zone avec $\rho, Q_4 \in [-0.990, -0.844] \times [0.8820, 1]$ le profit est négatif ou nul ce qui prouve la dernière partie de l'énoncé.

a. Sous l'hypothèse que f_i converge vers la distribution sous-jacente. N.B : cela est souvent faux en réalité. Les jeunes ou les locataires sont plus fréquents que les personnes agés ou les propriétaires par exemple.

Avant de continuer sur le concurrent B, il est intéressant de regarder de plus près le profit en fonction de différentes métriques. Le graphique 6.7 représente le profit en fonction de la corrélation et de la qualité. Logiquement plus il y a de la corrélation, plus la variable X_3^{real} a capturé l'information de X_4^{real}. Il est néanmoins important de remarquer que les cas où les variables sont corrélées à plus de 0.8 sont très rares en actuariat et que les variables seraient considérées comme quasiment les mêmes à la qualité de données près.

Comme mentionner précédemment, la figure 6.8a montre que pour une estimation de la qualité de données le profit est linéaire en la qualité de données sous-jacente. Ainsi le profit du modèle M_2 avec des qualités hétérogènes est égal au profit avec une qualité homogène égale à $\bar{Q}_4$. En d'autres termes,

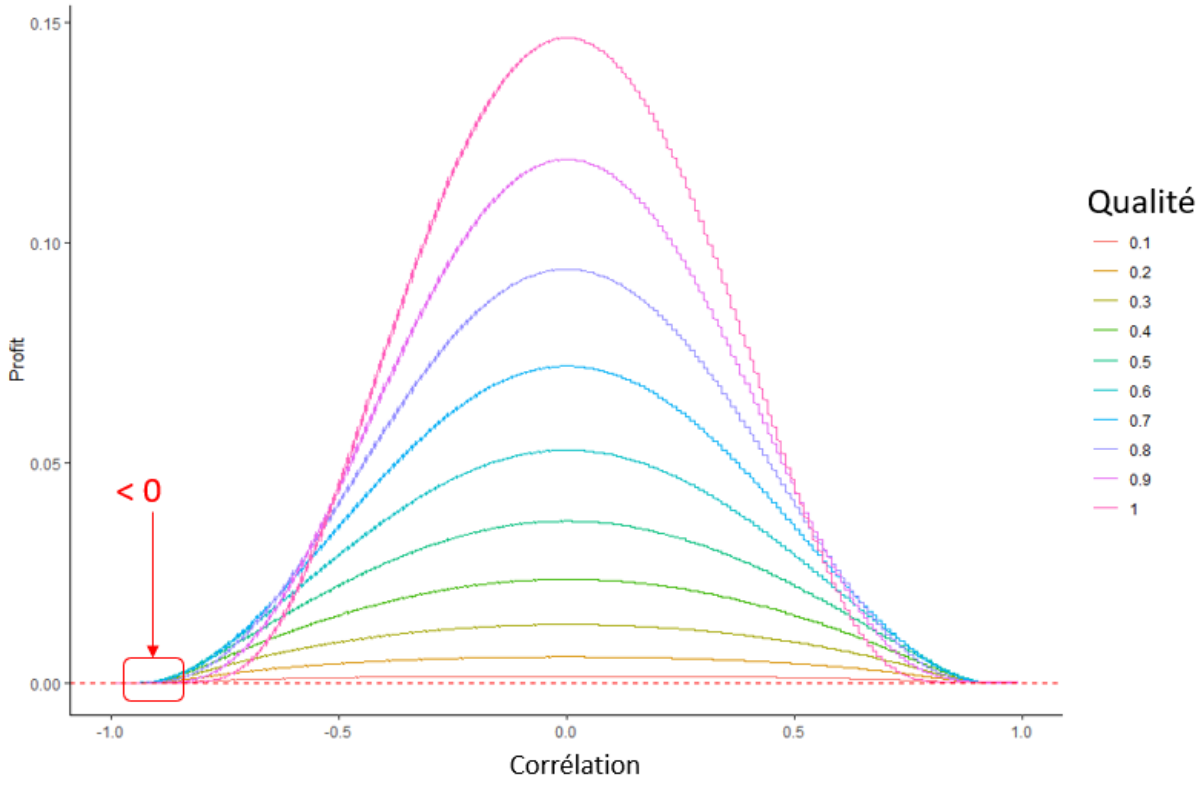


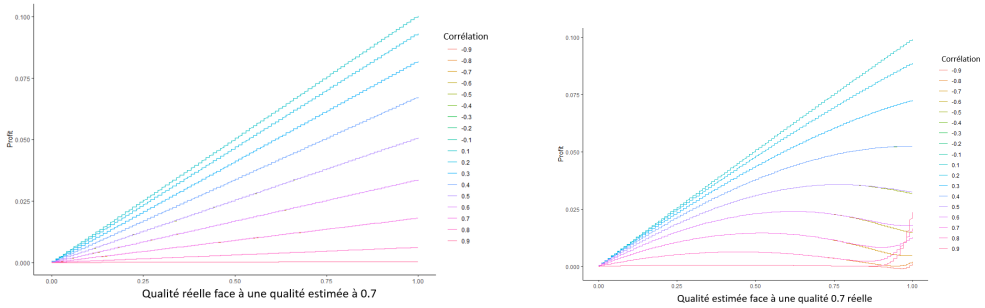
FIGURE 6.7 – Seulement pour des variables très corrélées négativement et une bonne qualité de données, le profit est négatif. La courbe n'est pas symétrique.

Le profit du modèle M_2 ne dépend pas de la distribution de la qualité sous-jacente [1].

La figure 6.8b montre que le profit optimum n'est pas toujours en la valeur Q réelle. Par exemple, pour une corrélation faible de 0.1, l'utilisation des coefficients M_3 avec $K = 1$ maximise le profit pour $\bar{Q}_4 = 0.7$. Pour une corrélation moyenne $\rho = 0.5$, le maximum du profit se trouve aux alentours de 0.7, donc avec l'utilisation des coefficients M_3 avec $K = 0.7$.

Finalement, si on suppose que la qualité estimée est égale à la qualité réelle, on obtient les profils de profits du graphe 6.9. Les courbes ne sont pas toujours convexes. Conséquemment, le modèle M_3 peut dans certains cas (avec de la corrélation importante) être moins profitable que le modèle M_2.

1. Sous les hypothèses de normalité.



(a) Évolution du profit lorsque l'erreur de la variable est estimée à de 0.7 quand elle vaut une autre valeur.

(b) Évolution du profit lorsque la qualité réelle de la variable est de 0.7 quand elle est estimé par une autre valeur.

FIGURE 6.8 – L'optimisation tarifaire et les erreurs de mesures montrent bien l'intérêt de simuler le profit sur une base test .

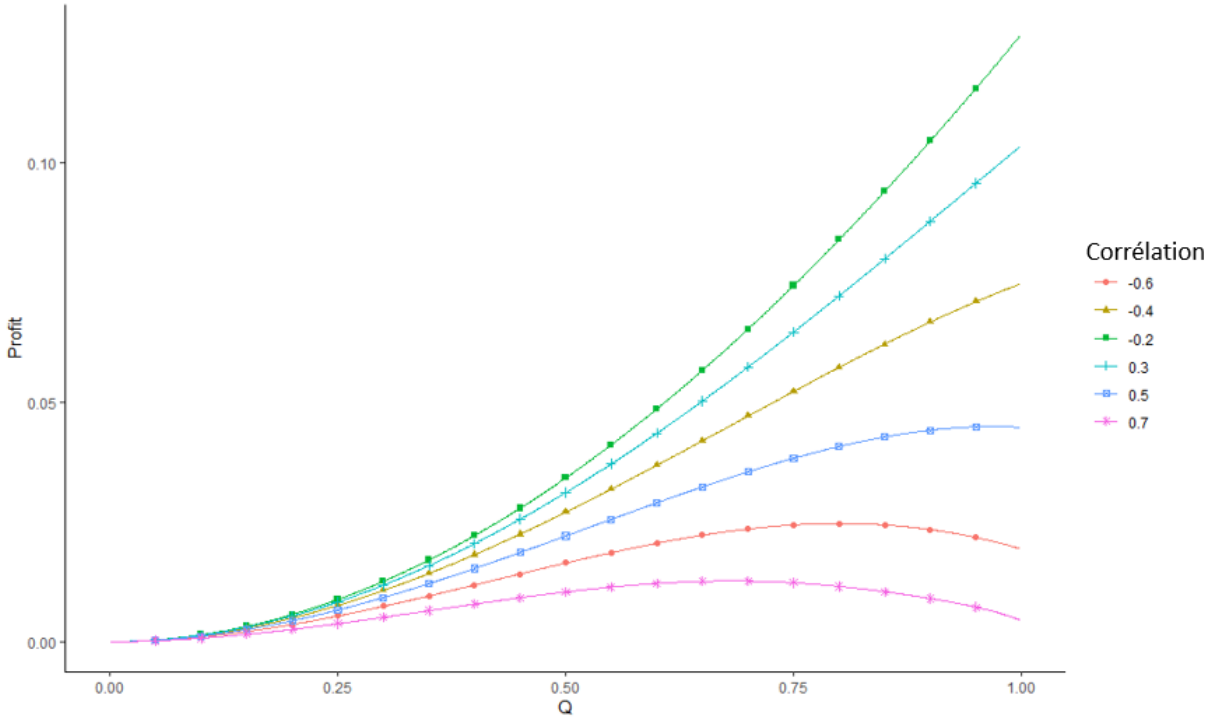


FIGURE 6.9 – La courbe des profits n'est pas toujours convexe. Pour des cas où la corrélation est faible, le modèle M_3 est plus profitable que M_2. Dans les autres cas, il est nécessaire de simuler ou recalculer les profits.

Preuve 3: Profit négatif pour la compagnie B

Pour le concurrent B n'utilisant pas l'information X_4, le profit se calcule de la même façon. Supposons dans un premier temps $\beta_4\rho > 0$. Sous l'hypothèse d'une copule bivariée gaussienne, avec f_i la fréquence d'un assuré i,

$$\begin{aligned}\sum_{i\in I_A\cup I_{A=B}} f_i\mathbb{E}(\Pi_B(i) - C(i)|(X_3^{real}, X_4) = (x_{i;3}^{real}, x_{i;4})) \\ = \sum_{i\in I_B\cup I_{A=B}} f_i\beta_4\frac{\sigma_3}{\sigma_4}\rho x_{i;3}^{real} + \beta_4 x_{i;3}^{real}\rho\end{aligned} \tag{6.36}$$

converge vers

$$\beta_4 \rho \frac{\sigma_3}{\sigma_4} \int_{-\infty}^{\infty} \int_{-sH^{-1}\rho^{-1}\bar{Q}_4}^{-\infty} t dF_{X_3^{real}, X_4}(t, s). \tag{6.37}$$

Ainsi le profit s'écrit simplement pour $\Omega = 0$:

$$-\beta_4 \rho \frac{1}{\sqrt{2\pi}} exp\Big(-1 - (H^{-1}\rho^{-1}\bar{Q}_4)^2 \Big), \tag{6.38}$$

et pour $\Omega = 1$:

$$-\beta_4 \rho \frac{1}{\sqrt{2\pi(1-\rho^2)}} exp\Big(-1 - \frac{(H^{-1}\rho^{-1}\bar{Q}_4 - \rho)^2}{1-\rho^2} \Big). \tag{6.39}$$

Dans le cadre $\beta_4 \rho < 0$, on obtient le même profit négatif sans difficulté. Parce qu'aucun des termes ne peuvent être nuls ou positifs pour $Q_4 \neq 0$. Il vient d'être démontré que $\frac{C}{P}^B > 100\%$.

En bref, la qualité de données impacte fortement l'estimation du profit. Sans optimisation tarifaire, les concurrents sont pénalisés par la segmentation. En revanche, dans certains cas la qualité aussi peut avoir des impacts négatifs sur la marge des modèles prenant en compte la qualité de données. Il faut garder en mémoire que les résultats sont dans le cadre gaussien avec des copules gaussiennes deux à deux corrélées. Contrairement à la tarification à l'adresse, il n'y a pas ici de corrélations entre les erreurs, la qualité de la donnée et les vraies valeurs. Ainsi même sous ces hypothèses simples, les métriques de segmentation globales des modèles ne sont pas suffisantes pour évaluer les gains de marges. de manière générale, il va être nécessaire de réévaluer par simulation les marges et d'ensuite optimiser les coefficients pour profiter du gain d'informations par rapport aux concurrents.

Chapitre 7

Données à l'adresse : application à la sécheresse

Les données à l'adresse permettent de modéliser plus simplement et précisément des risques qui sont peu modélisés dans les directions de tarification. La section 7.1 étudie chacun des principes marchés pour les modules aléa, de vulnérabilité et de dommage. Ces spécificités propres aux risques climatiques et les données à l'adresse permettent de mettre en œuvre une modélisation du risque de subsidence plus performante qui est détaillée dans l'article de la sécheresse - section 7.2. Ce dernier questionne l'assurabilité du risque et les évolutions récentes législatives à travers des règles de souscriptions, du provisionnement et de la segmentation que permettent les données à l'adresse. Nous rappelons que nous sommes dans le cas de la tarification pour des assureurs de tailles moyennes en général et non de réassureurs. Les contraintes comme les volumes de données sont différentes.

7.1 La modélisation des risques climatiques

La modélisation des risques climatiques est différente des autres garanties à cause de la dépendance spatiale, les auto-corrélations temporelles et la rareté des évènements. Comme mentionner dans le chapitre 1.1, l'objectif est de modéliser de la manière suivante :

$$\underbrace{\mathbb{E}}_{\text{Module aléa}} \Big(\overbrace{\mathbb{E}\Big(\frac{N}{e}\Big|I\Big)}^{\text{Module de vulnérabilité}} \times \underbrace{\mathbb{E}(S|I)}_{\text{Module de dommage}} \Big). \tag{7.1}$$

Dans de nombreux modèles, la notion de dommage est comprise dans le module de vulnérabilité. Un module financier permet de calculer les coûts des programmes de réassurance. Dans notre cas, nos modèles prennent en compte d'indemnisations payées et non les dommages réels. Deux approches existent les approches statistiques et les modèles dits "marchés" plus géophysiques.

7.1.1 Le module d'aléa

Pour évaluer $\mathbb{E}\Big(\mathbb{E}\Big(\frac{N}{e}\Big|I\Big)\mathbb{E}(S|I)\Big)$, l'estimateur le plus souvent considéré par les modèles marchés est :

$$\sum_{e_I \in P(I)} \mathbb{E}\Big(\frac{N}{e}\Big|I\Big)\mathbb{E}(S|I) w_{e_I}, \tag{7.2}$$

où $P(I)$ est l'ensemble des intensités des évènements climatiques représentatifs, w_{e_I} est le poids/la fréquence de l'évènement e_I. Il est impossible d'obtenir exhaustivement $P(I)$. En effet, historiquement certains évènements e_I probables n'ont jamais été observés, en particulier pour les évènements avec une période de retour les plus importantes. Pour ces derniers, il est intéressant d'extrapoler les résultats et les probabilités. Ainsi un sous-ensemble $P^*(I)$ représentatif de $P(I)$ est utilisé.

Selon (Charpentier, 2008 [143]), ces modélisations basées sur ces scénarios apparaissent dès les années 1970[1]. Les modèles physiques sont apparus fin des années 1980 (AIR en 1987, RMS en 1988). Pour les modèles de "marchés" (voir Mornet, 2014 [150]), le choix des évènements climatiques et leurs probabilités associées est déterminé dans un module aléa schématisé par le graphique 7.1.

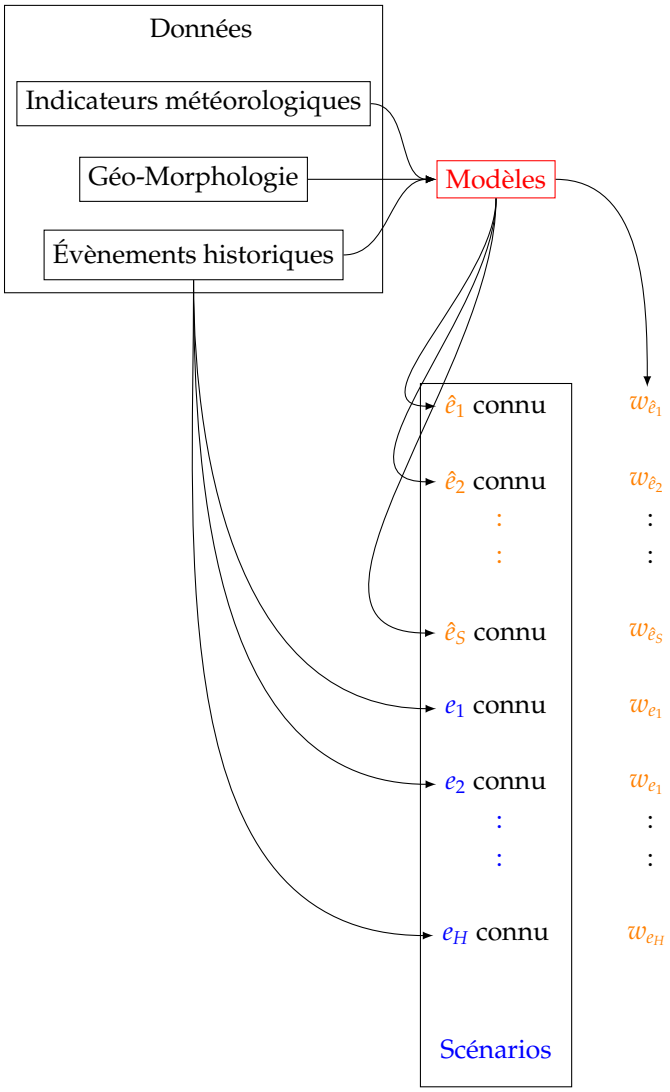


FIGURE 7.1 – Exemple simplifié d'un module aléa.

Le module aléa se base sur des scénarios déjà disponibles mais aussi complétés par certains scénarios mis aux points par des spécialistes "climatiques"(météorologues, hydrologues ...). Pour la création des scénarios, il existe de nombreux modèles différents. Par exemple, les modèles hydrologiques sont multiples comme des modèles dynamiques ISBA-MOCDOU ou bien statistiques - plus *data-driven* (ex : MODOR). De plus, ces modèles sont souvent la combinaison d'un nombre de modèles régionaux et temporels. Dayon et al., 2015 [145] met en lumière qu'il est difficile de considérer un même type de modèle pour les bassins français. De même, le compte rendu de Boe et al. 2018 [141] (DRIEE) expliquent que les scénarios "sécheresses" sont saisonniers et différents en fonction des régions. Les analyses mensuelles et même journalières sont préférées. La vision moyenne annuelle est donc à proscrire pour modéliser les intensités, d'autant plus que le réchauffement climatique vient exacerber les extrêmes temporellement. La thèse de Dayon, 2015 [145], détaille prospectivement. Selon le scénario RCP 8.5 avec le modèle GCM, le réchauffement climatique augmenterait de 10% de précipitations en hiver, mais diminuant les intempéries de 30% l'été.

Pour les modèles statistiques, différentes approches existent. Elles sont souvent à une maille large

1. Voir Gestion des risques naturels et changement climatique : les challenges des actuaires - Julien Tomas. la source de l'information initiale n'a pas été retrouvé.

comme celui des départements.Ces modèles souvent combinent les deux approches. Les approches diffèrent beaucoup en fonction des données disponibles et le choix de la précision temporelle reste important. La dépendance entre les périodes de retours ou *Joseph Effect - Husrt phenomena* peut être négligée annuellement selon Charpentier et Sibaï, 2009 [144]. Dans ce dernier, ils comparent des modèles ACD (Autoregressive conditional duration) et une application de la théorie des valeurs extrêmes. Ils concluent que l'indépendance est presque vraie à condition que les valeurs estimées soient des évènements avec des pas annuels pour des périodes de retours supérieurs à un an.

Pour le risque tempêtes, Mornet 2014 [150] détermine un indice de tempêtes pour ensuite déterminer un zonier (subdivisant la France en cinq zones) en utilisant les k-médoides. Ces zones sont déterminées en fonction d'une distance qui pondère la distance géographique et un indice de dépendance des extrêmes. Par zone, la théorie des valeurs extrêmes (voir Embrechts et al. 1997 [152]), ici une loi de Pareto généralisée, est utilisée pour modéliser les évènements tempêtes se basant sur des données journalières pour, par exemple, évaluer les périodes de retours des événements historiques (ex : la tempête de Lothar). Au contraire, Prettenthaler et al. 2012 [153] ne modélise pas la fréquence et les dépendances spatiales associées entre les régions autrichiennes. L'utilisation du MCMC (Markov Chain Monte Carlo) usant du bootstrapp non paramétrique sur des données empiriques permet de conserver la cohérence spatiale des différents scénarios possibles.

Pour le risque d'inondations, comme les débordements, les remontées de nappes ou les coulées de boues, de nombreuses études utilisent des distributions GEV. Les paramètres de la GEV pour les précipitations sont données en fonction de la région dans l'article Martins et al. 2000[149].

Il existe de nombreuses méthodes pour la modélisation des intensités des risques climatiques. Prenons l'exemple des précipitations. Les modèles les plus souvent utilisés sont des générateurs de temps : modèles faisant intervenir intensité et occurrence dont Gabriel et Neumann, 1962 [146] sont les précurseurs. Deux types de générateurs existent les modèles paramétriques (Katz 1977 [147], voir l'introduction de Kleiber et al. 2012 [148] pour d'autres exemples pertinents.) et non paramétriques (distributions empiriques (Trigo et Palutikof 1999 [157]), réseaux neuronaux (Rajagolapan et Lall 1999 [154]), estimateurs à noyaux (Semenov et Porter 1995 [155]), processus gaussiens (Kleiber et al. 2012 [148]) ...). L'avantage des modèles non paramétriques est de s'adapter facilement aux données, mais avec l'inconvénient majeur de ne pouvoir générer difficilement des séries temporelles différentes que sur lesquelles le modèle a été entraîné [2].

Les modèles doivent considérer la dépendance spatiale entre les stations de mesures. Pour une utilisation actuarielle, il semble usuel de déterminer des clusters pour regrouper les stations météorologiques (ex : Brito et al. 2017 [142], utilise la méthode de Wald, Baulin et al. 2016 [151] par classification) pour réduire la dimension d'études des séries temporelles.

Pour certains risques climatiques de faible envergure et avec une période de retour courte, l'approche historique est suffisante. Par exemple pour l'inondation non CatNat, la prime associée peut être calculée par des modèles Tweedies ou GLM combinés avec un zonier utilisant des données externes [3]. L'extrapolation des risques sur des zones qui n'ont pas été sinistrées se fait très bien à l'aide des informations des plans de préventions et autres informations spatiales.

• Application à la sécheresse

Dans notre cas, le module d'aléa choisi pour la sécheresse est extrêmement simple et consiste des événements annuels de 2010 à 2018 dont les poids sont supposés équiprobables. La difficulté est que nous n'avons pas d'assez de données historiques et une précision temporelle insuffisante pour apprendre des scénarios et les valider. Cependant, toutes zones potentielles ont été impactées au moins une fois durant ces 10 ans. Finalement, à cause du réchauffement climatique, il semble très hasardeux d'estimer une quelconque probabilité de retour sans modèle climatique. C'est pourquoi le modèle d'aléa est resté le plus simple possible.

De façon naturelle, les indicateurs hydrologiques sont plutôt indiqués pour évaluer certaines sécheresses mais pas toujours sa sévérité. Il en existe un grand nombre d'indicateurs possibles adaptés à chaque type de risques. L'analyse des sècheresses doit se faire sur plusieurs indices. En revanche, il ne faut pas *surinterpréter des changements des sécheresses hydrologiques pour l'appliquer sur les sécheresses agricoles* et vice-versa (citation de Boe et al. 2018, [141]). La période choisie est l'an pour éviter de considérer le Joseph Effect.

2. Différentes solutions existent comme utiliser des variables externes pour déterminer les scénarios possibles (Stehlik et Bárdossy 2002 [156]).

3. Voir des mémoires d'actuaires de (Gahbiche, 2017) chez AXA ou (Hia, 2020) chez Pacifica.

7.1.2 Le module de vulnérabilité

Le module de vulnérabilité [4] (ou d'exposition) a pour objectif de modéliser la distribution N sachant une intensité I ou un évènement climatique e_i. $N|I$ est très souvent censurée. En MRH, il est nécessaire de distinguer deux types de sinistres : les sinistres "Catastrophes Naturelles" qui rentrent dans le cadre du régime des Catastrophes Naturelles et les sinistres dits "climatiques" qui par exclusion concerne le reste des évènements climatiques.

Toutes les habitations d'une commune ayant subi un dommage provenant d'un évènement déclaré CatNat sont prises en charge par l'assureur. Dans le cas contraire, certains sinistres peuvent ne pas être pris en charge par l'assureur. Cette distinction est importante selon le type de sinistres. Pour les dégâts provoqués par une sécheresse non déclarée comme CatNat, sont rarement (jamais) pris en charge. À l'inverse, les dégâts suite à une inondation le sont souvent.

Dénotons $N^*|I$ la variable latente représentant le fait que l'habitation a été endommagée par l'évènement ou non. Pour un évènement e d'intensité I, la loi de N sachant cet évènement e est déterminé comme suit

$$N|e \sim Z_e N^*|e, \tag{7.3}$$

où Z_e est une variable aléatoire prenant 0 et 1 comme valeur représentant la prise en charge ou non du sinistre. Z_e est influencé par la déclaration ou non de l'évènement en catastrophes naturelles et aussi des moyens de prévention qui ont été mises en œuvre.

Il est important de remarquer que la déclaration e en CatNat dépend du nombre de personnes impactées donc l'intensité de l'évènement I mais aussi de l'action et la connaissance des communes du régime CatNat. Une intensité importante provoque très probablement une déclaration de Catastrophes Naturelles. Naulin et al. 2016 [151] explique qu'il est nécessaire d'évaluer cette probabilité de non-déclaration qu'ils ont observés dans leurs données ce qu'ils expliquent comme *"eventual protection of the risk (e.g. issues such as local elevation being more important, or the presence of natural or artificial protections, barriers) or to the low proportion of material damages in comparison to the deductible insurance"* [5]. Ainsi leur modèle détermine la probabilité individuelle comme par un GLM logistique :

$$\mathbb{E}(Z_e) = \frac{exp(a + bWD + c\ S)}{1 + exp(a + bWD(e) + c\ S)} d,$$

avec

— a,b,c et d des paramètres,

— WD la profondeur d'eau,

— S la différence entre la hauteur d'eau maximum et celle associée à une période de retour de 2 ans.

Dans la plupart des études statistiques comme la précédente, peu de données sur les caractéristiques individuelles sont disponibles ou fiables. La modélisation peut donc se faire au global en modélisant directement $\mathbb{E}\left(\frac{N}{v}\middle| I\right)\mathbb{E}(S|I)$. De ce fait, le module d'endommagement et de vulnérabilité est souvent considéré par un seul modèle (d'occurrence et d'intensité).

Pour les tempêtes à une maille départementale, Mornet et al. 2014 [150] détermine $N^*|I$ par une fonction linéaire de l'intensité pour les tempêtes les plus importantes.

Dans les modèles marchés ou physique, $N^*|e$ est moins probabiliste. En effet, pour un scénario e, il suffit de calculer l'intensité I de l'évènement à l'emplacement de l'habitation. Des règles déterministes permettent de savoir si $N^*|e$ est égale à 1 ou non. (Par exemple, s'il y a 20 centimètres d'eau dans l'habitation pendant 2 jours, la maison est sinistrée.)

Modélisation des modules d'aléa et de vulnérabilité simultanément : Si le module aléa est inclus en même temps que le module de vulnérabilité, il est important de prendre en compte la dépendance spatiale dans les modèles de vulnérabilité (fréquences) ou d'endommagement (sévérités). L'un des désavantages de cette méthode est de complexifier la bonne prise en compte de la dépendance spatiale en la prenant partiellement dans les modèles. La dépendance spatiale entre les régions peut se prendre en compte par la théorie des graphes (Erhardt, 2020) [6] pour la fréquence. À l'aide des régions hydrographiques, il entraîne un modèle auto-logistique centré sur des pertes mensuelles de 1979-2018,

$$log\Big(\frac{p_i}{1-p_i}\Big) = {}^t\mathbf{X_i}\beta + \sum_{i\sim j} \eta(Y_i - \mu_i)$$

4. Dans certaines définitions, la vulnérabilité prend aussi en compte le module dommage

5. Trad FR : Des protections du risque éventuel (c'est-à-dire les cas de sur-élevation ou de présence de barrières naturelles ou matérielles) ou la faible endommagement comparé à la franchise

6. Présentation orale à la C.A.S. en 2021.

où :

— p_i est la probabilité mensuelle qu'une inondation cause des dommages dans la région i ;
— η est un paramètre de dépendances spatiales $i \sim j$;
— μ_j représente la valeur espérée sous indépendance spatiale, c'est-à-dire $\mu_j = \mathbb{E}(Y_j|\eta = 0) = \frac{exp({}^t X_j \beta)}{1+exp({}^t X_j \beta)}$;
— ${}^t\mathbf{X_i}$ (resp β) correspondent aux variables (resp coefficients associés) tels les caractéristiques hydrologiques et météorologiques des régions (profondeur des nappes phréatiques, débits naturels...), des caractéristiques géographiques (altitudes, pentes, les types de sols et leur utilisation, la nature des sols...) et aussi des informations anthropologiques comme la densité de population, le taux d'urbanisation et la richesse individuelle.

Pour les risques tempêtes, la gestion de la dépendance spatiale est faite à partir d'un historique de tempêtes ou en fonction des indices sur des stations de mesures. Pour la sécheresse, Robert Erhardt [7] a subdivisé par une grille uniforme (par demi-degré) les États-Unis pour déterminer une structure de dépendance spatiale. La validation de ce type de modèle est extrêmement complexe et ne sera pas considérer dans notre article.

•Application à la sécheresse

Pour l'application aux retraits gonflement-argile, l'idée est de modéliser $N^* > 0|e$ comme la probabilité d'avoir une CatNat à la maille communale et $Z_e|N^* > 0$ comme la probabilité d'avoir le bâtiment endommagé lorsqu'il y a une déclaration CatNat. Comme les 5 ans d'historique assureur ne sont pas satisfaisantes pour estimer $N^* > 0$, le module CATNAT va donc estimer $N^* > 0|I$ depuis 2010 à partir de l'historique de la base GASPAR comportant l'ensemble des arrêtés jusqu'aux débuts du régime CatNat. À cause des évolutions de la législation et de la qualité de la base, il est difficile de remonter plus loin que 2010. Conditionnellement aux indicateurs météorologiques et d'urbanisations, les déclarations CatNats d'une même année entre les municipalités voisines sont supposées indépendantes.

Pour estimer $Z_e|N^* > 0$ ou $Z_e|CatNat$, l'un des avantages d'avoir des sécheresses à répétition ces dernières années est d'avoir une base de donnée de bâtiments dans des communes ayant déclarées une CatNat très importante et quasi exhaustive en termes de type d'habitations sur les cinq ans.

7.1.3 Le module d'endommagement ou de dommage

Le module d'endommagement ou de dommage a pour objectif de modéliser la distribution S sachant une intensité I. La difficulté réside dans le petit nombre d'observations disponibles. Généralement, les assureurs utilisent un historique important pour réduire l'incertitude de la modélisation. Toutefois, il est important de rappeler que l'évolution de la domotique, de la densité de la population et l'inflation associée est importante sur de longues périodes.

Dans les modèles marchés, ce sont des courbes de dommage qui sont utilisées. Une courbe de dommage est une grille reliant des niveaux d'intensités et des caractéristiques du bien à un dommage. Différentes courbes existent comme l'évaluation du taux de destruction par rapport à la somme assuré ou calcul direct des coûts dommages. Cependant, la véracité et l'adéquation à un portefeuille d'assureur est difficile à vérifier. Ils sont souvent mal adaptés aux marchés français par expérience et à des structures de portefeuilles d'assurés particuliers.

Des organismes français, le CEPRI par exemple, ont créé des courbes d'endommagements ; le modèle SIMUDOM le fait en fonction des caractéristiques des habitations et des biens à l'intérieur. L'avantage est de pouvoir adapter à l'inflation le prix des biens domotiques, mais avec l'inconvénient de ne pouvoir valider cette courbe d'endommagement sans demander un inventaire précis lors d'un sinistre.

Statistiquement, différentes méthodes existent. Du fait du peu d'informations sur le bien et de sinistralités qu'à un assureur, les modèles linéaires sont largement utilisés. Mornet 2014 ([150]) relie $\sum S$ à $exp(\gamma I)$ où γ est déterminé par Moindres Carrés Ordinaires. Naulin et al., 2016 [151] déterminent pour plusieurs classes de risques (ex : occupant, non occupant, propriétaire, locataire) un coefficient α tel que le taux de destruction T se détermine par $\alpha \sqrt{WD}$,WD étant la hauteur d'eau. La fonction de dommages pour les risques tempêtes de Prettenthaler et al. 2012 [153] associe le coût à :

$$L_{e,j} \sim \mathcal{N}(s\, W_{i;j} + I^{LM} + I_i^S + I_j^G) + S\, \Gamma(v, \beta),$$

7. Réponse à une question posée lors de la présentation

avec j représentant la région, e l'évènement de tempête, $W_{i;j}$ l'indice d'intensité de la tempête associée, I_i^S, I_j^G, I^{LM} des coefficients estimés par moyenne empirique. Cependant, d'autres études préfèrent une approche plus physique en reliant le coût des dommages à la vitesse cubique du vent. Finalement, c'est toujours le caractère explicable du module de dommage qui est le plus important et le faible nombre d'observations induits des modélisations simples avec peu de variables pour l'ensemble des risques climatiques.

Pour prendre en compte des dépendances dans les observations, les modèles les plus complexes traient par effet mixte cette dépendance. (Erhardt, 2020) utilise un modèle gaussien à effet mixte sur les pertes individuelles estimées (perte à la maille du département divisée par l'exposition) normalisé par la méthode Box-Lenkins,

$$\begin{aligned} L_i^* = \frac{L_i^\lambda - 1}{\lambda} &= \beta_{0,k} + X_i\beta + \epsilon_i, \\ \text{Avec } \epsilon_i &\sim \mathcal{N}(0, \sigma^2), \\ \beta_{0,k} &\sim \mathcal{N}(0, \tau^2). \end{aligned}$$

L'inconvénient est de perdre une partie d'explicative importante du modèle.

- **Application à la sécheresse**

Le modèle de dommage en sécheresse est de loin le plus difficile en termes de résultats, de volatilité et de contrôles. Contrairement, aux autres risques climatiques, il semble que l'intensité des évènements n'influe que peu sur le dommage engendré. Cela permet aussi d'éviter de complexifier le modèle final et de ne pas considérer les dépendances spatiales des évènements. Le module de dommage sera un module de coût moyen. Les taux de destructions ne fonctionnent pas du tout sur les risques de subsidence. L'estimation des IBNRs est nécessaire à cause du développement long des sinistres graves (souvent plus de 3 ans). Les modèles GLM log-gamma sont meilleurs que les modèles linéaires ou les approches Box-Lenkins.

Références

[141] Boé, J., Radojevic, M., Bonnet, R., Dayon, G., de France, I., and Habets, F. (2018). Scénarios sécheresse sur le bassin seine-normandie.

[142] Brito, T. T., Oliveira-Júnior, J. F., Lyra, G. B., Gois, G., and Zeri, M. (2017). Multivariate analysis applied to monthly rainfall over rio de janeiro state, brazil. *Meteorology and Atmospheric Physics*, 129(5) :469–478.

[143] Charpentier, A. (2008). Insurability of climate risks. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 33(1) :91–109.

[144] Charpentier, A. and Sibaï, D. (2009). Dynamic flood modeling : combining hurst and gumbel's approach. *Environmetrics : The official journal of the International Environmetrics Society*, 20(1) :32–52.

[145] Dayon, G. (2015). *Evolution du cycle hydrologique continental en France au cours des prochaines décennies*. PhD thesis, Université Paul Sabatier-Toulouse III.

[146] Gabriel, K. and Neumann, J. (1962). A markov chain model for daily rainfall occurrence at tel aviv. *Quarterly Journal of the Royal Meteorological Society*, 88(375) :90–95.

[147] Katz, R. W. (1977). Precipitation as a chain-dependent process. *Journal of Applied Meteorology*, 16(7) :671–676.

[148] Kleiber, W., Katz, R. W., and Rajagopalan, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed gaussian processes. *Water Resources Research*, 48(1).

[149] Martins, E. S. and Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3) :737–744.

[150] Mornet, A. (2015). *Contributions à l'évaluation des risques en assurance tempête et automobile*. Theses, Université Claude Bernard - Lyon I.

[151] Naulin, J. P., Moncoulon, D., Le Roy, S., Pedreros, R., Idier, D., and Oliveros, C. (2016). Estimation of insurance-related losses resulting from coastal flooding in france. *Natural Hazards and Earth System Sciences*, 16(1) :195–207.

[152] Paul Embrechts, Claudia Kluppelberg, T. M. (2008). *Modelling Extremal Events : for Insurance and Finance (Stochastic Modelling and Applied Probability)*. Corrected edition.

[153] Prettenthaler, F., Albrecher, H., Köberl, J., and Kortschak, D. (2012). Risk and insurability of storm damages to residential buildings in austria. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 37(2) :340–364.

[154] Rajagopalan, B. and Lall, U. (1999). A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water resources research*, 35(10) :3089–3101.

[155] Semenov, M. A. and Porter, J. (1995). Climatic variability and the modelling of crop yields. *Agricultural and forest meteorology*, 73(3-4) :265–283.

[156] Stehlık, J. and Bárdossy, A. (2002). Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation. *Journal of Hydrology*, 256(1-2) :120–141.

[157] Trigo, R. M. and Palutikof, J. P. (1999). Simulation of daily temperatures for climate change scenarios over portugal : a neural network model approach. *Climate Research*, 13(1) :45–59.

7.2 L'assurabilité des risques de subsidences

Cette section reprend des travaux et un article coécrit avec Stéphane Loisel non encore soumis.
Cette partie est inspirée de l'article Subsidence and household insurances in France : geolocated data and insurability.

Abstract

The insurability of natural disasters has always been an issue faced by insurers, states, and insured persons. In France, the insurer and the legislator are concerned with subsidence risks due to several consecutive dry years. More and more open data are available in France, which enables insurers to have better knowledge based on geolocation. This knowledge plus the increase in subsidence risks affect the insurability of the subsidence risk. Using mostly GLMs, the most common models used in France, this paper shows the improvement in the knowledge about subsidence risks. The results bring to the fore the importance of legislative control and the recently enforced CatNat program, leading the authors to question the CatNat fee stagnation.
Keyword : Subsidence, actuarial pricing, reserving.

Introduction

France has a specific reinsurance program for natural disasters, especially non-life insurance. In France, the law about climatic catastrophes - *CatNat* - 13 July 1982 - L. 125-1 of *Code Des Assurances*, defines the effect of natural disasters as *"the direct and uninsurable property damage being the main cause of the abnormal intensity of a natural agent when the usual measures to be taken to prevent such damage could not prevent their occurrence or could not be taken."* Natural disasters fitting (drought, floods, earthquake, hurricanes, etc.) this definition are considered Natural Catastrophes (CatNat). Once a municipality has declared CatNat, the claim's indemnities are compulsory for goods that are insured against fire damage. The legislature slightly modified the legislation in 28 December 2021, proving its long-term robustness. Developed first for floods, damages due to drought (such as clay shrinkage/subsidence) started to be taken into account in 1989. Flood damage corresponds to 46 % of the CatNat damage, and subsidence risks have increased in recent years (especially from 2016 to 2019). Insurers and the legislature are starting to seriously consider the increase in the frequency of subsidence (more globally, natural disasters) due to urbanization and climate change.

France and subsidence Subsidence damage to buildings arises from sinking of an area, mainly due to meteorological factors. In recent years, the frequency of such damage has increased. Many papers explain the relationship between climate change and drought. For instance, one report [168] highlights the fact that meteorological drought evolution differs depending on the country studied. Specifically, drought frequency, drought severity, and duration have increased in southern France ([174]). Droughts are studied using different meteorological indexes ; [167] study the standardized precipitation index (SPI), the standardized precipitation evapotranspiration index (SPEI), and the self-calibrated Palmer drought severity index scPDSI. In this paper, the indicators tested to model this risk are the SPI, the standardized soil wetness index (SSWI), and the USA reclamation droughts index (RDI). Regarding the actuarial, hydrological, and subsidence literature, Charpentier et al., 2021 [162] provides a complete overview. In short, subsidence risk is a difficult risk to model with current underwriting knowledge ; more and better data are needed. Recently, an increasing number of insurers have reduced the accessibility of new contracts due to subsidence. The recent drought of 2022 was assessed as the second-worst drought in France after the one in 2003.

Open data and geolocation According to data.europa.eu, French open data are among the leading of European countries in 2021 in terms of accessibility and transparency. All this information allows different actors to geolocate buildings using an address and to have access to a subsidence vulnerability map or information about the building (number of floors, the construction period, the surface, the vegetated surface). In this paper, a historical household insurance portfolio is used and geolocated on the basis of a data provider. The latter also provides approximately 60 variables about each building, as well as meteorological variables.

Contributions This paper makes several contributions. This article shows how to model subsidence risks for household insurance in France with the maximum data available thanks to geolocated buildings. Models of CatNat declaration at the municipality level are highly improved by adding meteorological information and aggregated information at the building level. Thanks to the latter, the results of the modeling of the frequency conditional on a CatNat declaration show that the new model is better performing and segmenting than a model using only underwriting and reserving variables. Even when the claims' development censors part of the information, our model outperforms the traditional cost models because of the variables related to the building and urbanization. Looking through the uses of these models *e.g.* reserving and prevention, this paper shows that the performance gain of these models may influence the insurability of subsidence risks. Charpentier et al., 2021 [162] focus on aggregate claims amount and frequency at the communal level, and this paper differs mainly by considering information at the building scale for mainly "pricing"/underwriting purposes at the policy level. Even if the approach of Charpentier et al., 2021 [162] is using the communal-mesh scale data and our variables are not exactly the same, their models are quite similar to the "Traditional model" developed in this paper in which few relevant information on the building are used. This paper shows the significant improvement in segmentation and reserving between the Traditional model and the model using geolocated data. For instance, the percentage of the riskiest houses of the portfolio (corresponding to a 10%- increase of combined ratio) is lowered from 3.5% to 1.4% of the portfolio, which shows that pricing is refined.

Section 7.3 lists all the data used with their particularity to fully understand all the models of the expected cost of subsidence proposed in Section 7.4. Finally, Section 7.5 discusses the problem of insurability of subsidence risks due to external data integration and presents some ideas to adapt the French CatNat program.

This paper is based on a portfolio of an insurer that has allowed us to present these results. Moreover, the same methodology has been applied to one other insurer portfolio and has led to the same results (variable selection and performance). All the numbers given are modified to anonymize the results unless otherwise mentioned.

7.3 Data and subsidence

The available information on a contract usually stems from the underwriting process (see Subsection 7.3.1) and from some external data sources (see Subsection 7.3.3). However, if the exact geolocation is known, the buildings' information can be added, as detailed in Subsection 7.3.2. Data are gathered by a data provider that has created a database for insurance purposes in France. For other damage coverage, this new information has proven its value in terms of the knowledge of risk. This data provider also provides meteorological indexes. Specifically, droughts are detected through several meteorological

indexes such as SPI, SSWI, and RDI, which are detailed in Section 7.3.4. Finally, each insurer has its own portfolio particularity, as detailed in Subsection 7.3.5.

7.3.1 Underwriting data

The portfolio considered is taken from a French insurer for MRH insurance (Multi-Peril Housing). The coverage of subsidence is compulsory only for owner-occupant or owner-non-occupant insurance contracts. This work focuses on both insurance contracts from 2015 to 2020 in the French mainland territory, excluding Corsica. Variables at disposable stem from the underwriting process. The most relevant variables are the occupant's age, the surface insured, the number of rooms, the period of construction, the personal property insured, the reconstruction value, and the type of contract (owner-occupant or owner-non-occupant). Each piece of information is taken from the last update of the contracts' database in April 2021. The quality of the information is excellent for most of the variables. The information on claims is the payment *pay*, the reserve *res*, financial recourse, and if the indemnity process is closed or open. We denote the *cost* at date t as the sum of *pay* and *res* at date t. We set aside recourse, which is negligible in number and amount.The reserve process for a claim S is as follows :

$$res(S,t) = \begin{cases} 0 & \text{if no claim is declared or closed,} \\ \hat{S} - pay(S,t) & \text{if a claim } S \text{ is evaluated and approved by an insurance expert,} \\ 20000 & \text{if a claim is declared \& the municipality declared a CatNat,} \\ 20 & \text{if claim is declared \& the municipality has yet to declare a CatNat,} \\ 0 & \text{otherwise.} \end{cases}$$

To properly model claim costs, claim triangles from 2001 to 2020 regarding the number of claims, the payment, and the cost of subsidence are provided. Different methods of reserving are used. The number of claims triangle development is better developed using GLM, as proposed by [173], where the negative binomial is the better-suited distribution (as for [162]). Then, we develop the mean cost of subsidence damage using the Mack stochastic model [169]. Several reasons explain the choice to use the mean cost and not the complete reserve or payment triangle ; the evaluation of the *res* of subsidence is quite erratic, with negative increments at some periods. Plus, in recent years starting from 2017, the payment does not provide sufficient information. The reserve is informative only once the insurance expert adjusted it. The open claims' costs are developed using the following factor Dev^{open}_{factor} for the J-the year of development :

$$Dev^{open}_{factor}(J) = \frac{CM(J)}{CM^{open}(J)} \frac{f_{Mack}(J)}{Prop_{open}(J)} \tag{7.4}$$

where $CM(J)$ is the mean cost of all claims, $CM^{open}(J)$ the mean cost of all claims still open, $f_{Mack}(t)$ the Mack factor for the J-the year of development and $Prop_{open}(J)$ is the proportion of claims still open. Table 7.1 shows the different order of magnitude.

Year	2015	2016	2017	2018	2019	2020
J	6	5	4	3	2	1
$f_{Mack}(J)$	15%	20%	17%	30%	50%	> 300 %
Nb of claims	100	1000	900	1400	300	< 10
$Prop_{open}(J)$	40%	45%	50%	75%	90%	0%
$Dev^{open}_{factor}(J)$	40%	20%	20%	25%	45%	-
$CM^{open}(J) \times Dev^{open}_{factor}(J)$	60 k	65 k	40 k	30 k	25 k	

TABLE 7.1 – Example of the development factor. All the numbers are anonymized, but the authors kept the order of magnitude.

Additional information For reserving purposes, the insurer provides the number of claims declarations reported at the end of the 1st year of development, aggregated at the municipality level. Notably, this type of declaration is independent of the CatNat declaration, *i.e.*, even if an important number of claims is declared in the municipality, it does not always lead to a CatNat declaration.

7.3.2 Geolocation of the building and data at the building scale

Historically, an insurer's portfolio is not geolocated during the underwriting process. To add new information about the building, the address is used to link the building and the contract. The data

Year	2015	2016	2017	2018	2019	2020
Proportion of claims kept	75 %	74 %	70%	90 %	90 %	0%
Geolocation rate in the area	80 %	75 %	70 %	94 %	80 %	90 %

TABLE 7.2 – Comparison between the geolocation rate in the department impacted and the proportion of claims kept. The number of claims in 2015 and 2020 is low.

provider links the address given with a geolocated address and then associates a building with it. This process is not perfect : the geolocation rate is lower in rural areas and the south of France and uncountable different settings exist. Figure 7.2 represents the two most common results. Even if these last issues are solved, only 80.3% of the addresses are linked to a building. Indeed, approximately 78 % addresses not geolocated do not have any street number, other addresses are not well reported, not updated, have the wrong spelling, and so on. Finally, among the geolocated buildings, an integrity filtering process is applied to suppress the incorrect building geolocation or demarcation thanks to different indicators such as geolocation quality indexes, the number of floors, and the footprint surface. In conclusion, 77.3 % of the addresses/contracts are kept for the modeling.

FIGURE 7.2 – Copyright OpenStreetMap (Taken the 07/03/2022) - A random street without any relation to the insurer's portfolio. Several issues can be seen. Stars correspond to address geolocation and squares to the related building. The buildings are not well reported in the building database, with no number being linked to them. Building 1 has an annex and can easily be linked to 19 Wolfloch Weg Colmar or 19 rue de WolflochWeg Colmar. Residential house 2, linked to 23 bis Wolfloch Weg Colmar, is 20 meters from the street and is composed of several buildings.

Several checks are performed to verify that this process does not bias the risk model. Different works on nonclimatic risks showed that a significant bias between 5 and 10 % appears for variables related to rural areas. Here, two conditional frequency models (See Section 7.4) using only the underwriting variables are fitted, one on the complete database and the other on the filtered database. No significant changes are observed. In our portfolio, the proportion of claims kept is the same as the addresses geolocated. Because the subsidence risk is a spatial risk, the analysis must also be performed by year and spatially. Table 7.2 shows that lost buildings are not spread uniformly. The geolocation rate is lower in the south and rural areas, where the 2015 to 2017 droughts were the most substantial. On the contrary, the north of France is better geolocated. On the severity side, the risk, by year, does not change significantly. All the areas impacted appear in the dataset used for modeling. In conclusion, the geolocation rate does not significantly bias the risk model.

The data provider has data on each building and its surroundings ; different variables, including at least the address location, were tested :

Building : Number of floors, living surface, footprint surface, annex's presence, type of roof, altitude, construction period, energy diagnostics, solar panels ...

Urbanization : Number of buildings within a radius of 50 meters, house value, distance from the closest residential building, parcel's surface, vegetated parcel's surface, number of attached houses ...

Location : Type of soil, vulnerability to subsidence risk, distance to nearest water point, altitude difference with the nearest water point ...

All variables are obtained from French Open Data Sources such as the IGN [8] pictures' database, ADEME [9], and the BRGM [10]. For each observation, quality evaluations are performed and summed up in quality indexes. For missing or incoherent information, the data provider imputes or corrects the values, *e.g.*, house value is calculated from historical data using various characteristics, the neighborhood's sales, etc. More macro-information at the municipality level was considered but was not found to be relevant for the clay shrinkage risk.

7.3.3 Gaspar, ONRN and CCR database :

Three external bases were used in this paper : the GASPAR database and information from ONRN about the number of vulnerable houses by municipality are exploited in the CatNat modeling. Also, CCR historical information helps to compare our results.

GASPAR database The GASPAR [11] database reports all the administrative procedures related to natural risks : regarding our paper, the declaration of subsidence CatNat risk and the PPRN (Municipal prevention plan for natural risks [12]). This database is of good quality [13] for most of the year. Some corrections are performed for older years, for which PPRNs were wrongly reported. If before 2000 the notion of subsidence CatNat was not sufficiently clear, the database's quality is very clean starting from 2005. The CatNat information is available through several indicators, *e.g.*, the starting date of the disaster, the duration of the episode, the end date, and the date for which the disaster is declared as a natural catastrophe. Looking through the different years, the evolution of CatNat declarations is evident. Before 2003, a subsidence event may exceed a year and the mean declaration delay was a lot higher than it currently is. The stationarity starts between 2005 and 2009 for different indicators, as exemplified by the duration (Figure 7.18).

ONRN database The ONRN [14], the French National Observatory of Natural Risk, made available for each municipality the number of buildings subject to subsidence risk. The information corresponds to the 2015 period and to the 2015 Subsidence vulnerability map. At that time, only three classes exist "*none*", "*low*" and "*high*". A total of 3 (+ 3) variables are created, counting the number of houses by risk class (resp., proportion). The downside is that the vulnerability map is not the latest. Our model also uses similar variables corresponding to the number of insured houses in the portfolio in each class of risk (from the latest vulnerability map class "*none*", "*low*", "*medium*" and "*high*").

The CCR database information The reinsurer CCR provides aggregated information by municipality from 1995 to 2018. The ratio claims amount and premium, the mean claims cost, and frequency are available. In this work, these data are used to compare our results, highlighting our models' limits. Unlike our data, the claims come from different insurers all over France, and consequently, it can be considered as market information. The three maps in Figure 7.3 sum up all the information provided by the CCR. This information will be used to test our models and not as variables.

7.3.4 Drought indexes and meteorological variables

Approximately 30 meteorological indicators were available on the municipality scale from 2010 to 2018. Each variable is available annually, and a variable referring to the mean value of 2010-2018 is created. The meteorological information was unknown for 2019 and 2020 and was replaced by the mean value calculated for the 2010-2018 historical data.

Nonexhaustive list of meteorological data used (NbD = Number of days) ;

Temperature (T°) : NbD $T° \leq 18$, NbD $T° \geq 34$, NbD $T° \geq 4$ + mean seasonal temperature, NbD $T° \geq 8$ + mean seasonal temperature, ...

8. The reference public operator for geographic and forest information in France
9. The French Agency for Ecological Transition
10. France's reference public institution for Earth Science applications
11. *fr : gestion assistée des procédures administratives relatives aux risques naturels*
12. *fr : Plan de prévention des risques naturels*
13. Few quality tests were done on subsidence risk, whereas for floods risks the database has been tested in broad terms *e.g.* [Casaux et al. 2019] [161] or for a spatial-temporal quality problem on the side of the flood by [Douvinet and Vinet 2015] [163]. Nonetheless, the latter issues are similar to drought ones.
14. *fr : Observatoire National des Risques Naturels*

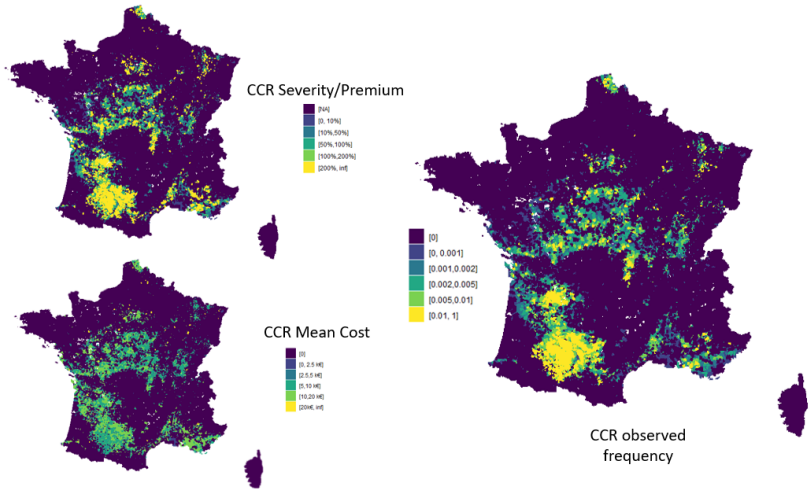


FIGURE 7.3 – Information provided by the CCR on the subsidence risk on the French market. The data correspond to aggregation from 1995 to 2018 sinistrality.

Precipitation (Prep) : NbD $Prep \geq 1$ mm, NbD $Prep \geq 1.25$ mean precipitation, NbD of heavy precipitation, prep quantity...

Frost : NbD of superficial frost, NbD of deep frost, NbD consecutive of deep frost, NbD with snow ...

Wind : NbD of heavy wind gusts, Number of wind gusts with an hourly speedy higher than 99 $km.h^{-1}$, ...

Note that not all the meteorological variables were relevant and some indicators are highly correlated.

In addition to meteorological variables, drought indexes were added by year at the municipality scale. The drought index literature has proposed several indexes to understand the connection between drought and weather. Different types of drought have been proposed by Wilhite and Glantz (1985) [176] : agricultural, meteorological, hydrological, and socio-economic. The operative event of clay shrinkage is caused mostly by hydrological and meteorological droughts. A subsidence event is triggered most often by clay shrinkage, and clay swelling is caused by higher humidity conditions, for example, rain. A quick level shift can crack the foundation or walls. Such damage is covered in a household insurance contract if the municipality has declared a natural catastrophe. Recently, a clay shrinkage-swelling natural disaster is declared if the meteorological indexes are abnormal, corresponding to a period of return higher than 25 years. Since 2010, the indicator SSWI has been used at the *inter-ministériel* meeting to classify the status of natural disasters. To evaluate the probability that a municipality declares a CatNat, three indicators are studied. These indexes were available monthly and were extrapolated to each municipality from 2010 to 2018. All this information has been summed into three variables (*severity*, *duration*, *magnitude*) by calendar year and by drought index.

SPI : Standardized precipitation index is commonly used, *e.g.* [175]. ([171]) has developed a methodology to standardize indicators. In our case, we consider the **SPI - 1 month**, the annual monthly minimum named *magnitude* (Similar ESPI variable from [162]), maximum duration of events, *duration*, when SPI is below -1 and the mean of the SPI during the event, named *severity*.

SSWI : The standardized soil wetness index (SSWI) is better adapted to agricultural drought. The SWI is used as a meteorological criterion for CatNat declaration since 2009. We consider **SSWI - 1 month**, the annual monthly minimum, maximum duration of events when SSWI is below -1, and the mean SSWI during the event.

RDI : The reclamation drought index was developed by the United States Bureau of Reclamation in 1996 to trigger drought emergency relief funds associated with public lands. This index captures drought severity and duration and can be used to predict the start and end of drought periods. RDI uses temperature and hydrological components, incorporating evaporation into the index for its calculation. We consider the **RDI** : the annual monthly minimum, the maximum duration of events when RDI is below 0 and the mean RDI during the event.

Annual but not seasonal : Contrary to Charpentier et al. 2021 [162], our goal is to predict the frequency and the claims for an insurance company that consolidates its financial statements in late

December. For reserving, the calculus is done for each year. Moreover, the different models used suppose independence of each observation, which can not be assumed seasonally. From the Gaspar database, nearly no municipality had two CatNat declarations in the same year. This is explained mainly by the delay of CatNat declaration. The period event could be increased to take into account two events, and all claims incurred are declared in the first event. Two limits of this method must be stated :

— Inclusion of droughts starting at the end of the year and ending at the start of this new year : few in number but not insignificant ;
— Meteorological criterion updated in 2018 now has defined a new seasonal threshold. Even if we used the standardized index, which partially addresses this problem, our model is based on historical events before 2018.

The idea is to use these three variables to sum the worst drought of the year according to each indicator. The downside of these three variables is their dependencies, which are not trivial and impact how the linear model should take them into account.

As explained by Charpentier et al. 2021 [162], the subsidence meteorological criteria changed over time. From 2000 to 2003, only a hydrological criterion was used. The year 2003, the worst year for subsidence, was not captured by this criteria, and since 2004, a meteorological criterion was used. Because the criteria were not easy to apprehend, the SSWI was finally used starting in 2009 and evaluated in winter, spring, and summer. In 2018, a new threshold is used for each season. Notably, the justification is quite succinct, and the presence of clay and the number of houses damaged may be better key drivers than this meteorological factor.

7.3.5 Particularity of the portfolio studied

The contract portfolio covers the entire mainland France from 2015 to 2020. With an exposure greater than 700 000 per year, the claims are spread out in all the potential territories impacted by the subsidence risks according to the vulnerability map provided by the BRGM. According to the insurers' actuaries, the claims frequency of 2016 is unusually high compared to that of other insurers for the same year. The CCR insurer has also noted this particularity. A probable underlying reason is that several buildings were damaged before 2016 but waited for the CatNat declaration in 2016 to be declared. The mean contract's seniority is approximately the same as that of the French market, approximately 10 years with a constant retention rate.

The data have undergone some temporal evolution ; the underwriting process was revised over time. For instance, the period of construction variables has new modalities, leading to lower completeness. Evolutions due to the underwriting process are not exceptional. Hence, some variables may have a limited impact, *e.g.*, the insurer's period of construction will have a lower impact than that gathered from open data, with completeness of approximately 97%. Before 2015, the number of rooms and the house's surface were considered. However, for the new underwriting process, only the number of rooms is still included. Therefore, missing values on the surface are imputed using the number of rooms and also information from the geolocation process.

Finally, the management of claims also impacts the data. For each contract, an estimate of reconstruction value is calculated using the underwriting information. If important damage occurs, such as subsidence damage, an insurance expert evaluates the damage and provides a proper evaluation of the reconstruction value. In this process, the surface is sometimes set to 0. Therefore, the authors were not able to train proper random forest and XGBoost models on the conditional frequency, as explained in Subsection 7.4.3. On the basis of external information, the surface gathered during the underwriting has been correctly imputed for use in the GLMs. The authors note that the linearity of GLM avoids learning this causal information and also helps to evaluate the bias due to our model for surface imputation.

7.4 Modeling the CatNat frequency, claim cost and legal declaration

This paper's modeling process is based on the CatNat regime process. The indemnity occurs only if the municipality has declared a subsidence natural disaster. Hereafter, within the same municipality, the frequency and the claim cost depend on the building characteristics. The first step is to model the municipalities' CatNat declaration in Subsection 7.4.1. After a filtering process to clean up the errors on the link between building and geocoding (Subsection 7.4.2), models for claims' frequency (Subsection 7.4.3) and claims' cost (Subsection 7.4.4) are applied with and without the geolocation data of the building. From a performance perspective, the improvement thanks to the data is significant, and the

aggregation of the different models leads to several insights. All the models are created to be used operationally. Machine learning models such as random forest ([160]) or XGBoost ([165]) are used with frugality for reserving and are almost forbidden in underwriting use. Therefore, the combined model uses GLMs with a limited number of splines. For models on the CatNat declaration, we used a XGBoost-based model, which can be justified due to the complexity of meteorological variables and the use of compositional variables for urbanization impact. For use in an underwriting process or other operational uses, the results are resumed in zones similar to a zoning variable added in premium models. Figure 7.4 summarizes the results.

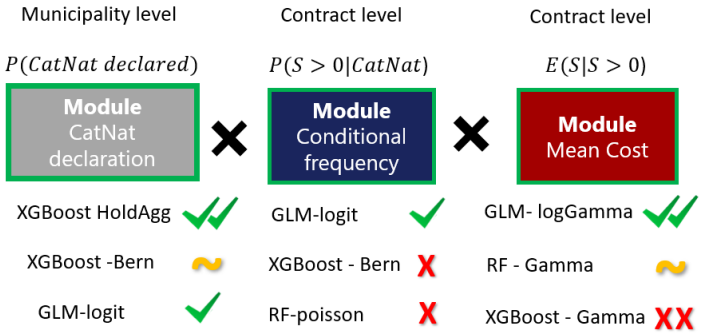


FIGURE 7.4 – S represents the cost of a claim and *Ber* is an abbreviation for the Bernoulli distribution. Due to data management, XGBoost and RF are not suitable for conditional frequency, even the GLM must be used carefully. Moreover, the mean cost modeling is impacted by the reserving process, resulting in improper convergence for the RF and even worse for XGBoost. For CatNat declaration, GLM and XGBoost have proper results. Nonetheless, the latter may learn improper causality. Thus, we propose a model with worse performance that XGBoost HoldAgg but that is more robust.

7.4.1 Natural Catastrophe (CatNat) declaration models

CatNat models focus annually on CatNat declaration at the municipality level. For reserving purposes, the probability that the municipality declared a CatNat given the annual meteorological indexes at date t and date $t-1$, $\mathbb{P}(\ CatNat(mun,t)|\ meteo(mun,t,t-1), charac(mun))$ is modeled. The same probability without any meteorological information $\mathbb{P}(CatNat(mun,t))$ is calculated based on available historical information, as follows :

$$\mathbb{P}(CatNat(mun,t)) = \sum_{t=2010}^{2018} \mathbb{P}(CatNat(mun,t)|\ meteo(mun,t,t-1), charac(mun)) \times \mathbb{P}(meteo(mun,t,t-1)), \tag{7.5}$$

where $\sum_{t=2010}^{2018} \mathbb{P}(meteo(mun,t,t-1)) = 1$, mun is the municipality, t is the year, $meteo(mun,t)$ are the weather indicators, $charac(mun)$ are the other characteristics of the municipality and $CatNat(mun,t)$ indicates if there is CatNat declaration in the municipality at date t. Equation (7.5) assumes that all the drought scenarios possible appear between 2010 and 2018 conditionally on the municipality information (urbanization, meteorological index, ...) and that there is no spatial dependency [15]. Let us assume that all probabilities are equal ; one with better knowledge could adjust the different scenario's weights.

The training method To find the hyperparameters, a spatial k-fold approach for each model is implemented, where the model is fitted on 50 % (70 % for GLM) of all regions and validated on the remaining regions [16]. The approach of time cross-validation ([158])performed by removing the future from the analysis was not relevant to our data. Starting from 2010, until 2016, only the year 2011 is a drought year. Moreover, the variables considered in this paper do not all have the same spatial properties. Indeed, urban or meteorological variables are spatially correlated. Therefore, when using a similar spatial and

15. i.e., having a municipality nearby declaring a subsidence CatNat does not increase the probability of declaring a CatNat, all other things being equal.

16. The spatial grid used is the *department* [17], other independent grids were considered (set by hierarchical cluster analysis (HCA)) with 70, 500, and 1170 groups. The changes were not significant (same hyperparameters/same likelihood).

temporal model for GLM (as in [162]), some coefficients were volatile and not interpretable, *e.g.*, the coefficient of SPI severity, magnitude, and the number of buildings in a 50 meter radius.

Reserving models used First, the CatNat declaration models were fitted using the number of claims declarations recorded the same year at the end of December and aggregated at the department scale. In this model, the annual meteorological variables are not used, only the historical mean and drought indexes. For the geolocated variables, aggregated information at the municipality scale is used; the relevant variables are the mean altitude of all the portfolio houses, the mean number of buildings in a 50 meter radius, the mean number of houses highly vulnerable to clay shrinkage (resp., medium, low, none), the mean distance to watercourses and the mean probability of being the main residence. The models' results are shown in Figure 7.5 from 2001 to 2020 using an XGBoost model.

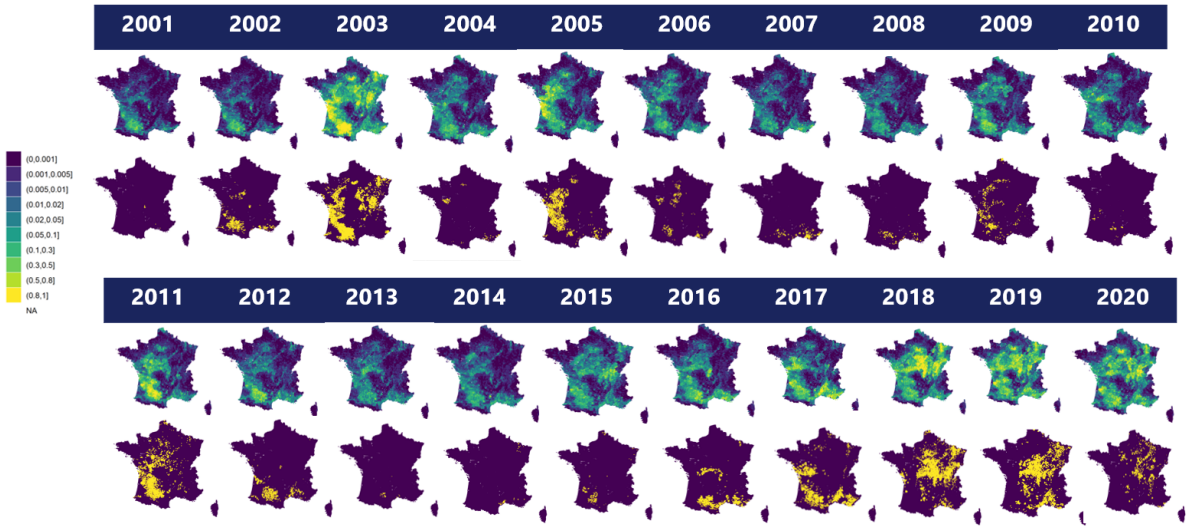


FIGURE 7.5 – CatNat declaration modeling at the municipality scale for reserving purposes. The second row corresponds to the observed CatNat declaration. This model does not use annual drought indexes nor meteorological variables but the number of subsidence damage instances declared at the end of December at the department scale.

Figure 7.5 shows that the model is correct. However, subsidence declaration does not always lead to a CatNat declaration. Furthermore, some clients wait for the CatNat declaration to declare damage to their insurer. Consequently, if the number of CatNat declarations at the end of December is informative, the structural dependency limits the model's performance. For instance, for 2004, 2010, and 2013, few CatNat declarations were enforced due to return of a drought period. Nonetheless, several claims were declared, inducing an overestimated probability. By adding all the drought indexes, the variable - the number of declarations in the first years of development - was not found to be more relevant : all the information was captured by other variables.

Meteorological and annual information is thus added to improve the model. First, two steps are performed for variable selection. To keep influential variables, a simple XGBoost and random forest are fitted. Then, a stepwise logit-GLM (Efroymson, 1960 [164]) is used to select all the nonmeteorological variables. Finally, XGBoost is fitted using the selected variables, the three indexes and two pieces of meteorological information (at most, the number of days for which the temperature is below 18 degrees and the annual precipitation quantity.).

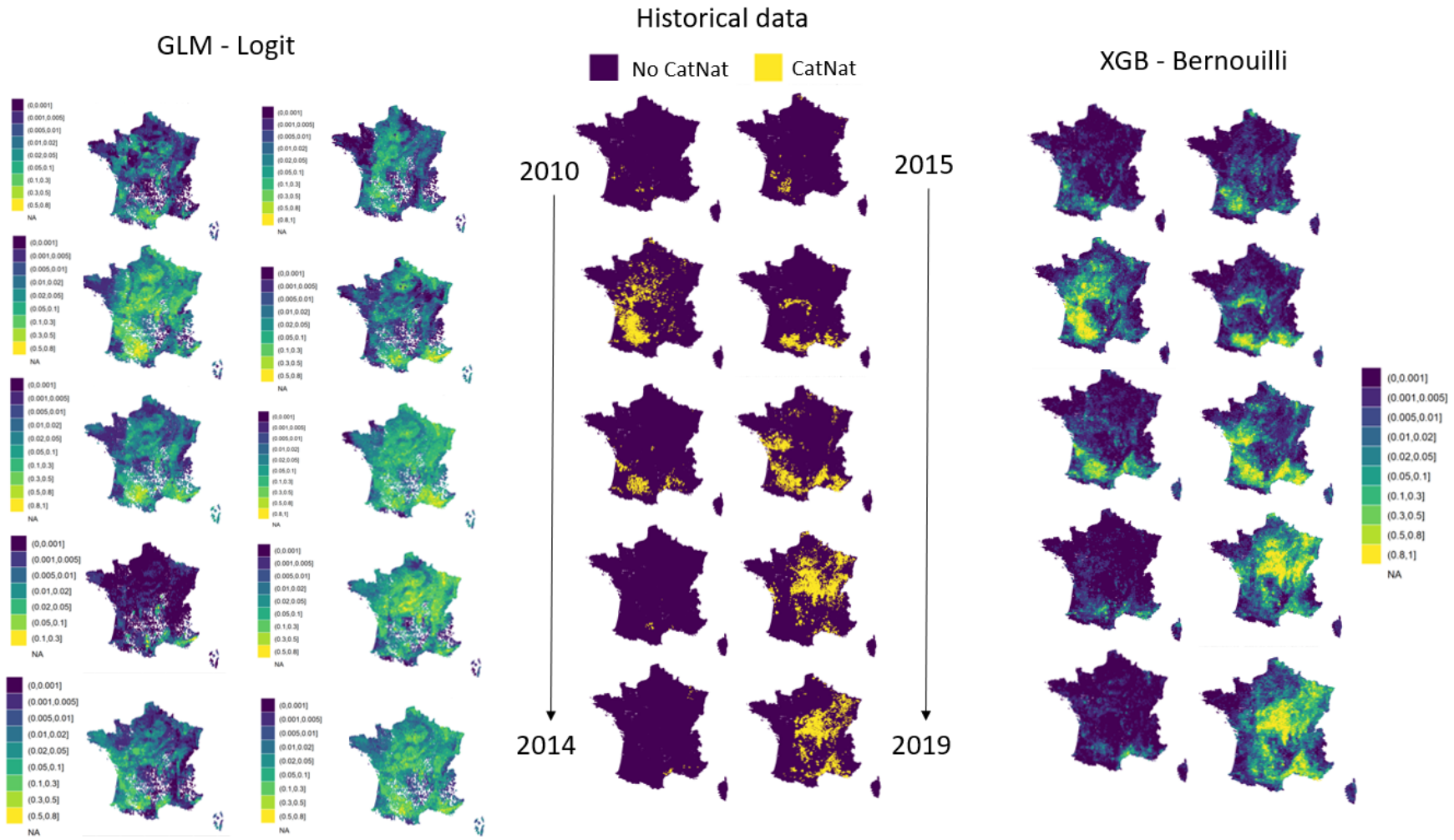


FIGURE 7.6 – Comparison between the observed CatNat declaration, the GLM logit (using RDI magnitude, SSWI severity and SPI magnitude) and the XGBoost using all the variables calculated from the three indexes.

To be used in reserving or for pricing purposes, the model must be fully interpretable, especially when using variables that change each year. Whereas the linearity of GLMs helps to fully understand quickly the learned structure, the XGBoost structure is not easy to comprehend. To partially understand the model, Shapley values [18] are used to find the cross-effect for each model for meteorological interactions and compositional/aggregated variables. Figure 7.7 shows the interactions between the 4 most important variables according to Shapley score or the traditional importance plot based on the Gini importance. The figures in 7.6.3 also show that the previous meteorological variables provide some relevant information and that the Shapley interaction values can be used to consider all the trivariable (or more) interactions learned.

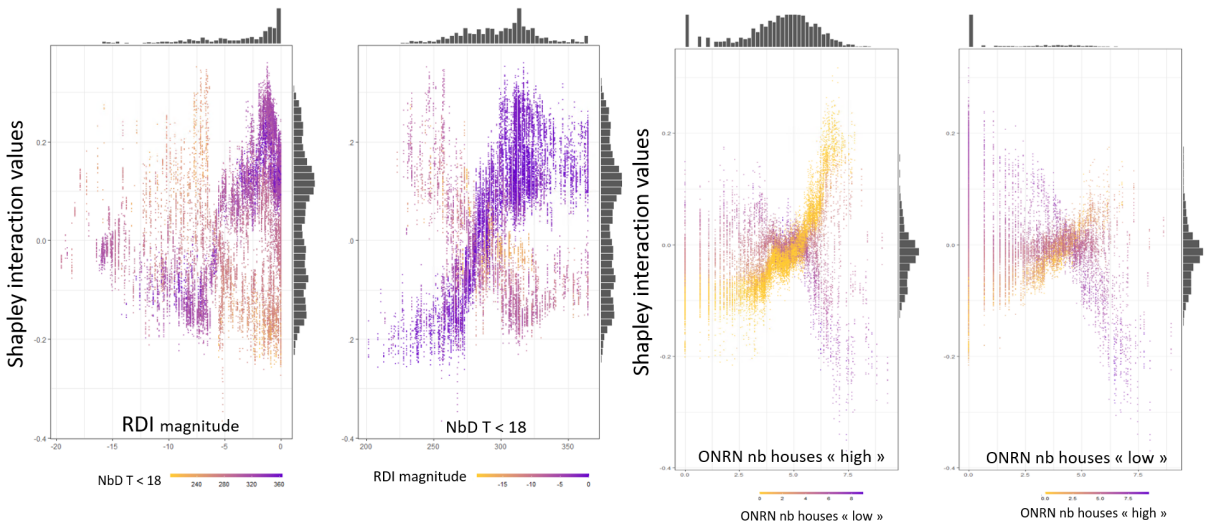


FIGURE 7.7 – Shapley interactions between meteorological variables and aggregated variables using a simple XGBoost trained on 2010-2018 historical data (20 000 rows used to calculate the Shapley values). The ONRN variables are represented on a log scale. The analysis is only bivariate; the interactions between all the meteorological variables are impossible to completely discern. A high number of interactions are present in the XGBoost learned structure and are difficult to replicate in a GLM or GAM.

Table 7.6 shows via the GLM linearity that drought indicators perform well in determining the location and the frequency on their own. Thus, the performance of the XGBoost may be explained by the use of drought indicators. In our case, the authors trained several XGBoost models, one on 2010-2018 historical data, a second on 2010-2019 historical data, and the last on 2010-2020 historical data. Notably, the meteorological information is not available for the years 2019 and 2020 and was replaced by the mean value calculated from the 2010-2018 historical data. The model would be expected to perform worse if the 2019 and 2020 years are added. Table 7.6 shows that this is not the case; this is a problematic issue and shows that the XGBoost is not predictive or has learned noncausal information. Therefore, the author added a constraint to use the model.

Operational constraint *For the years 2020 and 2019, the performance of a model should be lower or equal to that of the GLM trained from 2010 to 2020.*

Inspired by Maillard et al. 2021 [170], the next proposed structure empirically verifies our operational constraint. The method "HoldAgg" consists of cross-validation and averaging. Several XGBoosts are learned. Each XGBoost is trained on 6 years of data, and the hyperparameters used are those that maximize the cost metrics calculated on the 2 remaining years. The final result is the average value of each of the XGBoost results. The mean value is not necessarily a probability. Therefore, we train a classifier (Niculescu and Caruana, 2005 [172]) to map the average of the result to a probability. The classifier is trained on the 10 % of the training set on the mean value with a step size of 10^{-3}. Therefore, the averaging model is performs worse than the simple XGBoost model. However, when fitting this XGBoost HoldAgg model on the years between 2010 and 2020, the performances for the years 2019 and 2020 are lower than those of the GLM.

This method uses the "mrl" package [159] with a Gaussian process to find each hyperparameter more efficiently. First, 30 hyperparameter settings are fitted. Then, a Gaussian process is fitted on all 30 points

18. package *SHAPforxgboost*.

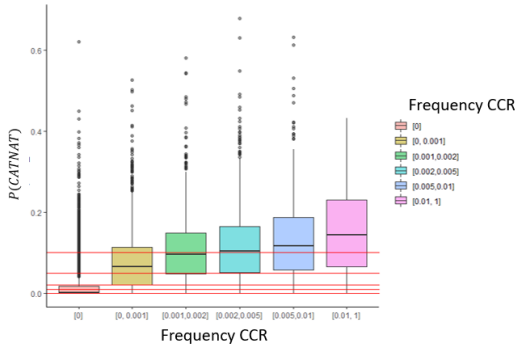
to find the next point that leads to the best expected improvement. An XGBoost is calculated with these hyperparameter settings. Then, a new Gaussian process is fitted on the 31 points and so on. We repeated this process 30 times.

The difference with the aggregated Hold out : Appearing at first in NeuroImaging ([166]) the aggregated holdout process, or more simply "CrossValidation + averaging" has been demonstrated theoretically under several assumptions by [170]. The performance of each XGBoost depends on the training and testing data, *e.g.*, for an XGBoost trained on all years except 2013-2014 (two years without any drought), the choice of hyperparameters leads to the worst performing model. Our model does not satisfy all the assumptions needed to justify the theoretical performance of the aggregated holdout process. The assumptions are as follows :

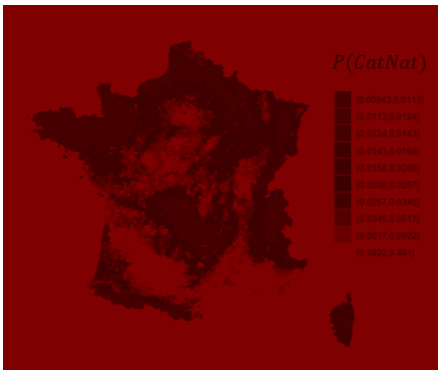
— Classification with the 0–1 risk or convex risk : Our variable is binary, but we used a regression model to calculate the probability.
— The split for the hold-out cross-validation is temporal and is correlated with the variable being modeled. Indeed, the number of CatNat declarations depends on whether the year is dry.

Model results Comparing the XGBoost results by year with the GLM results using all the nonmeteorological variables and the RDI magnitude variable, one can see all the years are well-considered. There are several limits. The year 2018 is not perfectly captured by the meteorological indicators, and for years without drought, a residual probability remains.

Given the probability, we have resimulated the number of natural disaster declarations and compared it to the real value. The low probabilities are overestimated due to the classifier's precision. For the higher probabilities, the model is well calibrated.



(a) Comparison of the aggregate value and the historical CCR frequency. Our model is based on only the 10 most recent years and the CCR is based on 23 years. The horizontal lines refer to the CCR range limit.



(b) Estimated probability from the XGBoost HoldAgg. 10% of French municipalities are likely to declare a clay shrinkage CatNat superior to 0.1.

FIGURE 7.8 – The last historical years represent all the possible drought scenarios. Our assumption is that each year/scenario has the same probability to reappear, meaning 3 of 8 years do have a drought. For 2001-2020, the probability would be approximately 7 of 20.

Moreover, the results can be compared to the CCR historical return period in Figure 7.8a. The model matches the CCR information, where the mean values increase with the period of return given by the CCR. The means are not in the interval period of return because the historical period considered is different.

7.4.2 Integrity filter

Before working at the address level, it is important to analyze the errors in the data. Indeed, the geocoding accuracy is not perfect due to the geolocation process, the address quality, and the databases used. Therefore, an integrity step is performed before modeling. Here, the integrity rule is "All the buildings geolocated are individual residential houses". Using the number of floors, the data provider's

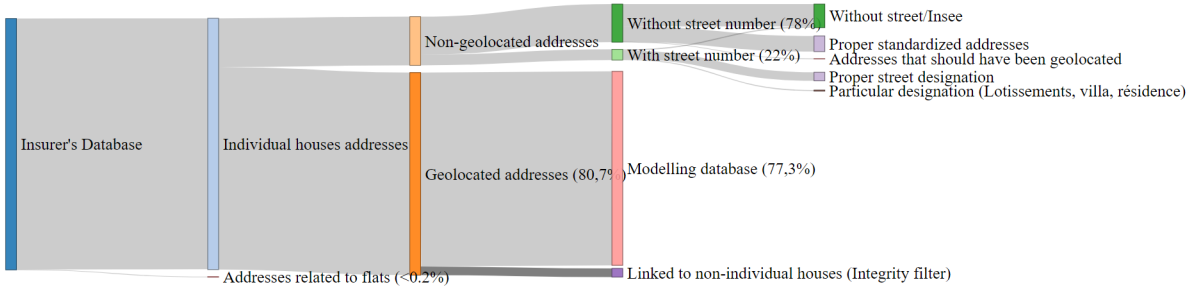


FIGURE 7.9 – Controls on the geolocation process for nongeolocated addresses. A similar process is performed for geolocated addresses; a stratified sample is chosen to compare our geolocation and that achieved with Google.

information, and the potential surface, 3.4 % of the data is set aside. Quality tests on geolocation are performed on the houses geolocated and on the nongeolocated addresses. For nongeolocated addresses, the results are summarized in Figure 7.9. Using the same process for geolocated buildings, we can state that 9 % of the addresses are potentially not well-geocoded and within these less than 4 % of the addresses are not linked to the "good" buildings in our judgment. Unfortunately, no direct rule can separate these addresses. Therefore, these addresses are kept, lowering the marginal impacts of variables using geolocation [19].

Several analyses are conducted to verify that filters do not significantly impact the sinistrality. The number of claims lost is approximately 20%, similar to the proportion of nongeocoded buildings. The claim distribution is not impacted by lost claims. The lost buildings are linked mostly to rural and southern areas, for which the standardization of addresses has not yet been completed.

Consequences of the modeling : To control the filtering process, two models using the insurer's variables are considered : one on the complete database and one on the filtered and geolocated database. No changes in coefficients are significant. In conclusion, the geocoding process does not significantly influence the subsidence risk according to the insurer's information [20]. For contracts that are not geolocated, the conditional frequency and claims are calculated using the department mean value.

7.4.3 Conditional frequency at the building scale

In this section, the goal is to model the frequency of houses conditionally on a CatNat declaration $\mathbb{E}(Card\{S > 0\}|CatNat\ declared)$. To compare the performance gained from geolocating contracts, two models are implemented, one with only underwriting variables, named *Insurer model* or *Referent model*, and the other one using all variables available, named *Performant model* [21].

Distribution used : The maximum annual number of claims in the database is equal to one. Therefore, we used a Bernoulli distribution in our XGBoost and logit-GLM. For variable selection, a random forest using a Poisson distribution from the *distRpackage* is used because of the credibility hyperparameters options [22].

Surface and reconstruction value variables of the insurer : The data analysis indicates that a null surface or a precise reconstruction value is highly linked to subsidence claims. The expert during the claim process can modify the data sent, adjusting the reconstruction value and suppressing the surface value [23]. Therefore, the null surface value has been replaced by a prediction using an insurer's variable (number of floors) and several other geolocated variables (living surface, construction period, etc.). The imputation is not perfect, especially for extreme values. For GLM, it is important to consider that it underestimates the frequency claim for buildings with a low surface and very important surface. However, for machine learning methods (XGBoost or RF), the use of the surface and reconstruction value

19. Not all variables are affected in the same way.
20. This statement could be discussed according to the data coming from the geolocation of buildings. However, by definition, the filtered buildings are wrong so nothing more can be done.
21. This model is named "performant" because, by construction, the model is performs better than the *Referent model*.
22. For low probabilities, Poisson and Logit are good approximations of each other; see [177] for a robust modified Poisson model for binary data. The credibility hyperparameters of the proposed RF Poisson lead to good results.
23. In fact, the building surface stopped being queried a few years ago. Only the number of rooms is still considered. The reconstruction value is a variable created using the surface or the number of rooms and other geographical information.

variables is problematic. The univariate and bivariate partial plots indicate that these types of models learn nonlinear relations and, worse, learn a not operationally justifiable structure, *e.g.*, reconstruction values between 155k and 165k have an important marginal impact on subsidence.

The variables used for the performant model are presented in Figure 7.10. The *Referent model* uses the same underwriting variables and also the construction period, the value of the personal insured property, and the number of rooms, which were not relevant and hard to justify.

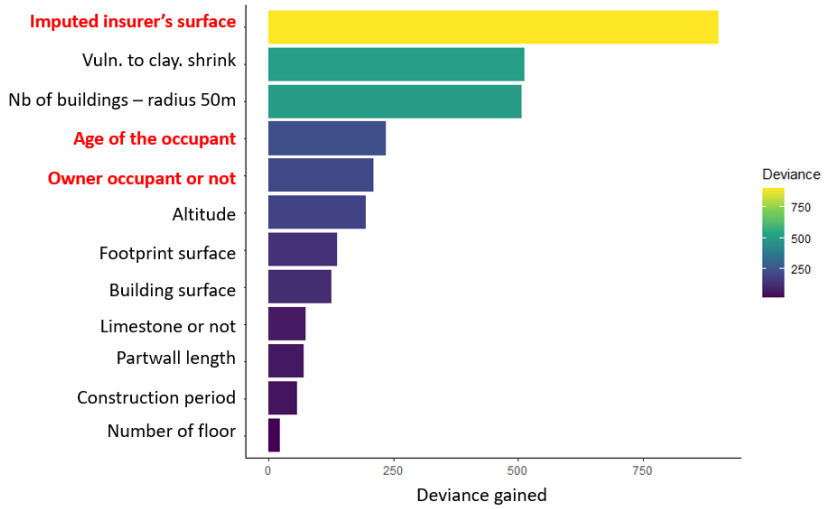


FIGURE 7.10 – Performant model : gained deviance when adding a variable to all the others in the GLM framework. The dependencies between variables should be noted. All variables are validated using a type III test and univariate graphs like the ones in the 7.6.2 and have a causal explanation.

Figure 7.10 represents the gained deviance when adding a variable to a GLM model, including all the others. Only three insurer variables remain relevant. Being an owner nonoccupant greatly reduces the subsidence declaration. Indeed, the insured person is less careful than an owner occupant. Age also impacts the declaration : older people declare less. However, the margin could also be triggered by the correlation with the type of house or period of construction. The imputed surface is not truly an insurer variable : the missing values are imputed by different variables requiring the geolocated building, especially the period of construction, surface of the building, and footprint surface, which are included in the performant model. The insurer variable cannot be entirely replaced. Indeed, it corresponds to the entire insured surface (not only the main building related to the address) with the annexes and refers also to the "used" living surface. Moreover, no geolocation errors exist.

Model	Type		EDR [a] Poisson	EDR Logistic	R^2	Gini
GLM - binomial logit	Referent	Without insurer's surface	100%	100%	100%	100%
	Referent	With the imputed surface	+ 42 %	+ 71%	+ 133%	+ 22%
	Performant	Without insurer's surface	+ 130%	+ 91 %	+ 379 %	+ 39%
	Performant	With the imputed surface	+ 164%	+133%	+ 430%	+ 48%

TABLE 7.3 – Comparison between the metrics for the conditional frequency models. Models optimize logistics deviance or equivalently the EDR logistic. The R^2, the normalized Gini and the EDR Poisson are used to control models.

a. The EDR is one minus the ratio of deviance between the model one and the satured one.

Results Table 7.3 shows adding information related to the building geolocation improves the performance for several metrics. When adding a zoning variable via the credibility spatial smoothing method, the *Insurer model* still performs worse than the *Performant model*. Moreover, no significant zoning variables can be added to the *Performant model* : all the geographical information is captured at the address level by all the other variables. The geographic information captured can be seen by looking at the risk maps in Figure 7.11. An important gap at the municipality level is observed. Moreover, the performant model segments within each municipality, whereas the Insurer model does not.

The geolocated variables used are very segmenting, *e.g.*, the greater the number of floors, the deeper

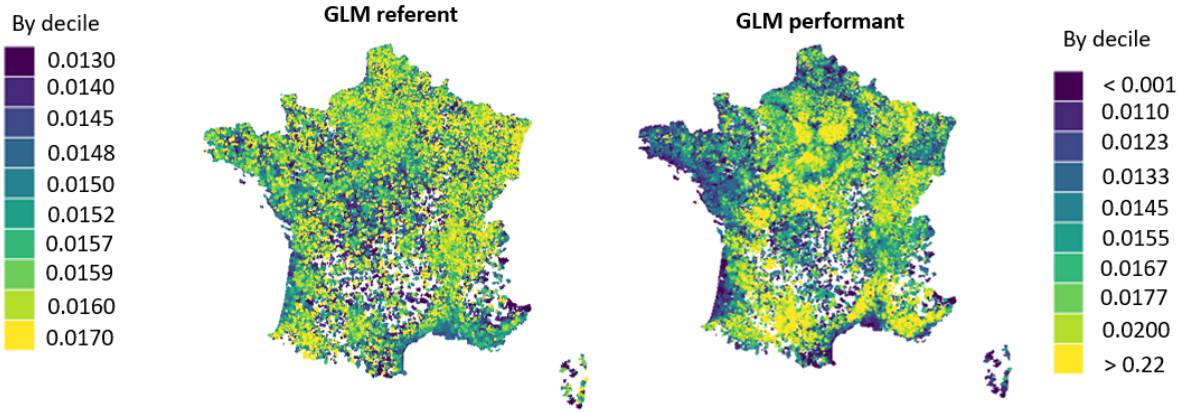


FIGURE 7.11 – Maps of the predicted conditioned frequency (if there is a CatNat, the probability of a house being damaged). The colors represent the deciles of the mean value of each house in the municipality, and the values correspond to the mean in the category. Probabilities are less relevant in zones where drought CatNat does not exist, such as Brittany.

the house's foundation is and the less subject the building is to subsidence risk. A vegetated surface and the number of buildings in a 100 m radius also capture the tarring level of the neighborhood and the probability of the presence of a canopy. The vulnerability to clay shrinkage is also very segmenting (see appendixes for univariate plots) [24].

Limits The principal limit of this conditional frequency model arises from the data. Indeed, for recent years, not all claims are declared. Therefore, the frequency is slightly underestimated. Moreover, no meteorological information is relevant for the conditional frequency ; the information may already be captured in the CatNat declaration module. Thus, the model might underestimate the frequency for drought years and overestimate the frequency for years without extreme temperatures.

7.4.4 Cost claim models and reserving

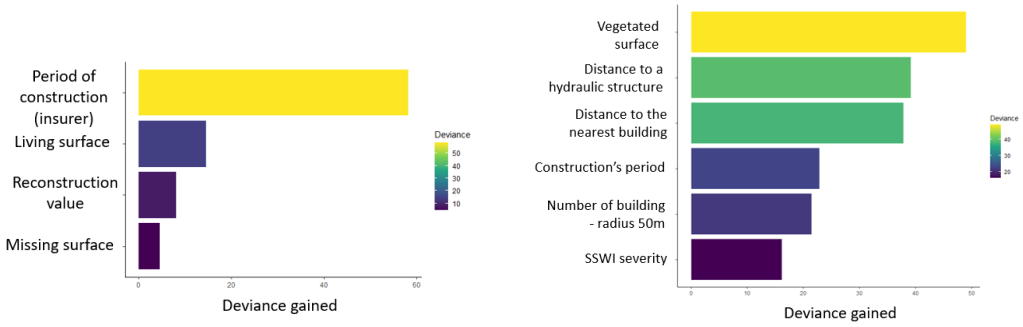
For the frequency, two models are considered : *Referent model* (Insurer's model) and *Performant model*. As explained in Subsection 7.3.1, most claim costs are yet unknown. Therefore, several databases and methods were tested (using year as a modality, looking at a ratio per year, etc.). The projection of costs still open yields the best results, though the linear impacts are similar for each method. All these controls help to validate the database's robustness. One claim was set aside due to the presence of asbestos, leading to an extreme cost.

The claims database used for modeling contains only claims that had a payment or for which an expert has evaluated the claim cost. For the others, an automatic opening claim is provisioned, waiting for an insurer expert to estimate the subsidence damage. These claims are not informative and thus are not used in the mean cost modeling.

XGBoost models do not converge to the mean value but to the median due to the claim cost particularity. Therefore, GLM and random forest gamma are used. For nearly all metrics shown in Table 7.4, the *Performant model* is better than the *Referent model*. Using one-way analysis and GLM linear structure, the marginal impacts of added variables are new and relevant, replacing the insurer variables (See Figure 7.12).

Limitations Three limitations can be stated. First, the mean cost is driven by the development factor. Thus, the average mean cost of 27 k€ is volatile, with an IC of approximately 3k€. This development factor dev_{factor} covers the fact that open claims are less informative than completely paid claims. Therefore, the marginal impacts of the variables are very likely underestimated. Finally, the dev_{factor} correction factor is obtained by year. Therefore, some meteorological information, such as the RDI severity, cannot be added. Indeed, this latter information is highly correlated with the year. Therefore, the learned marginal

24. The precision is between the street and iris levels. Indeed, the BRGM map is not sufficiently precise.



(a) Referent model GLM log-gamma

(b) Performant model GLM-Gamma : The variable SSWI severity is used only for reserving purposes. Other variables are set aside ; their marginal impacts are not relevant or justifiable.

FIGURE 7.12 – Deviance is observed when adding a variable to all the others in the GLM framework. The dependencies between the variables must be considered. All variables are validated using a type III test and univariate graphs like the ones available in the appendices and have a causal explanation.

Model	Type	EDR Gamma	R^2 (10^{-2})	Gini	Mean difference
GLM-logGamma	Referent	(0,0052)	(-0,3)	(5,98)	- 418
RF-Gamma	Referent	100 % (0,015)	100 % (1,02)	100 % (9,35)	- 451
GLM-logGamma	Performant	- 20 %	-70%	- 24,3%	-13
RF-Gamma	Performant	+ 46 %	1,59	94,4%	-8

TABLE 7.4 – Referent models do not perform well. The referent's RF might have learned improper marginal correlation based on the socio-professional information (which is not operationally justifiable). The referent's GLM has an artificially high normalized Gini linked to the poor performance. The performant model does not use the SSWI severity variable. The latter, when used, does not significantly alter the performant models' metrics.

impacts are difficult to explain properly. In summary, RDI severity and RDI magnitude could be used in cost claims if more historical data or better-developed claims were available. The latter limitation impacts the period of the construction variable. Indeed, this variable is linked with each year through spatial correlation. Depending on the year, claims occur in different areas, adding an artificial dependence to the true marginal impact of the construction period variable.

These limitations and operational constraints lead to a preference for GLM models. The nonlinear structure of random forest hinders the control of the impact of undeveloped claims. However, we use the RF for the referent model (because the GLM version has poor performance) and the GLM for the *Performant* model.

7.4.5 Combining the models

Figure 7.13 shows that the referent GLM does not perform well and that the random forest models (both the referent model and the performant model) fail to converge to the mean. Nonetheless, the latter segmentation is better than that of the GLM. The highest stable value is approximately 1 k€ for the referent model (RF + GLM) and 2.5 k€ for the performant (GLM +GLM) model.

After adding the CatNat model for reserving, the same graph 7.14 can be plotted by year. Linearity between the observation and the prediction is notable. All the predicted premiums higher than 1 000 euros are grouped together. Year 2020 is not yet completed (fewer than 5 closed claims). The number of claims in 2016 is higher than predicted, as explained in Section 7.3.5. The 2015 results are much more volatile since the year was minimally affected by drought. For the expected cost for "pricing" purposes, the CatNat model used for the *Referent model* and *Performant model* is the same : the XGBoost AggHold.

Several limitations regarding these models must be noted.

First, in our model, frequency and claims cost do not depend on previous claims. No building is our dataset declared two subsidence claims. According to the insurer's actuaries, on the basis of historical data since 2001, some contracts have declared more than one occurrence of damage : the first case often

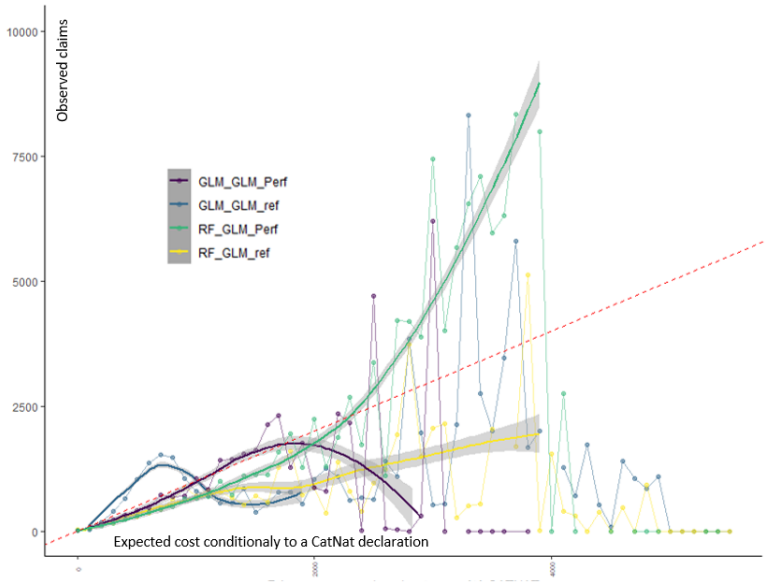


FIGURE 7.13 – Comparison between the different models combining the conditional frequency and the cost. For each type of model, a linear model using a spline and the exposition of each point is fitted to perceive the trend.

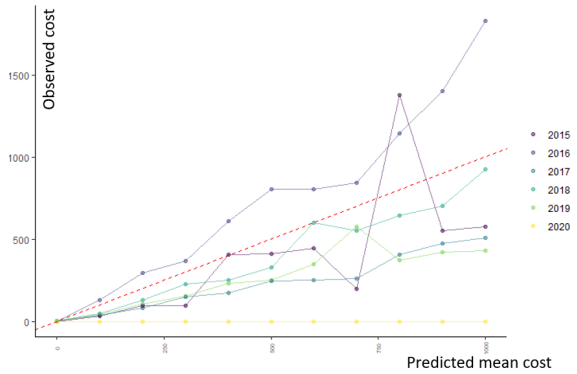


FIGURE 7.14 – Comparison between the observed sinistrality by year and the performant model (GLM + GLM + XGBoost HoldAgg).

being less expensive than the second. Two questions remain unanswered : Does the damage correspond to the same building ? Does this process correspond to poorly repaired buildings ? If both questions are answered positively, the stationarity assumed in our models may not be valid.

Second, the conditional frequency model is calculated on the basis of data from 2019 and 2020 from the perspective of April 2021. Some contracts for this time period have yet to declare their subsidence damage ; thus, our model may underestimate the conditional frequency.

Third, the cost claim model was based on a high number of incomplete claims, especially in 2019. Therefore, variables with annual variation, such as meteorological indexes, were not considered due to the lack of information.

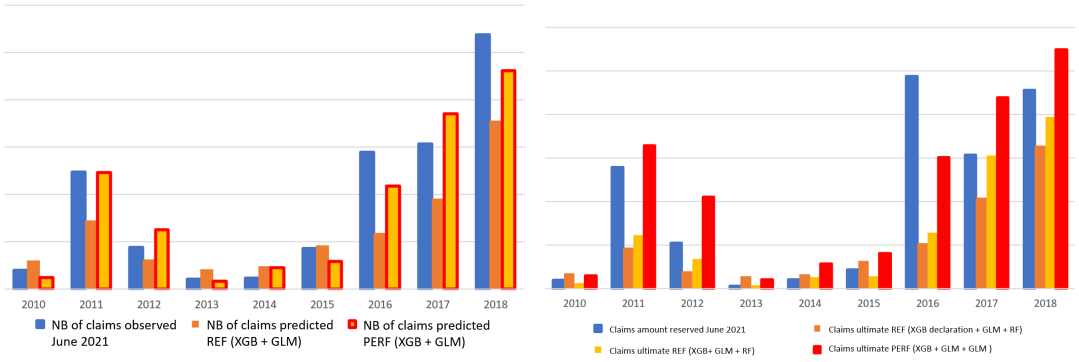
The modeling of CatNat declaration depends on the historical range considered. Given the availability of the data, 2010 - 2018 is considered. Nonetheless, one could argue that a wider scale could be used, e.g., starting from 2003 (which is the worst year for subsidence). The stationarity of the CatNat declaration and the robustness remain open questions, for which the authors have insufficient knowledge to answer properly.

Last, the legal evolution of the CatNat regime and climate change impacts the stationarity of the process. This issue will be discussed in Subsection 7.5.3. In short, the developed models have predictive power over a short period, even if different tests on other insurers' household insurance portfolios validate the robustness of the results.

7.5 Reserving, prevention and insurability statement

In France, variable pricing of CatNat coverage is not allowed. The premium is calculated as 12% of the sum of the damage coverage premiums. Nonetheless, expected cost modeling is compulsory for the reserving process (Subsection 7.5.1) and can be used to identify highly exposed segments for prevention purposes (Subsection 7.5.2). The downside of the model's performance is the insurability of this type of risk. Due to the increase in the frequency of natural disasters, urbanization, and climate change, the mutualization of the CatNat claims prevents uninsurability thanks to the CCR legal reinsurance. However, our model may lead to the identification of a segment for which the insurer has a higher negative margin. Although an important evolution of the French CatNat program was approved on 16 December 2021, it might not be sufficient to overcome the insurability issue (Subsection 7.5.3) and may even jeopardize it.

7.5.1 Reserving



(a) Prediction of the number of claims for reserving purposes. As expected, the year 2016 is abnormal. The R-squared metric is 48% and 92%, respectively, for the referent and performant models.

(b) Comparison between the predicted ultimate claims amount and the cost (payment and reserving) calculated in 2021. The R-squared metric is -21%, 43% and 69%, respectively, for the referent model XGB declaration + GLM ref + RF, the referent model XGB + GLM + RF and the performant model XGB + GLM + GLM.

The meteorological variables can be used to estimate the global claim amount. In this paper, the claim reserve of the insurer from 2001 to 2020 was disposable. However, only the address portfolio since 2015 was available. Let us assume that the 2015 portfolio is a good approximation of the 2010

to 2014 portfolio. This approximation is satisfactory in terms of exposition and portfolio stability. For nongeocoded addresses, the mean department values for the frequency and claim amount are used. For 2012, drought occurred in a zone that is well geolocated, leading to overestimation due to recalibration.

Results are shown in Figures 7.15a and 7.15b. For 2019 and 2020, the drought indexes are not available, and insufficient municipalities declared a subsidence CatNat. In recent years, notable differences arise due to the opening reserving process and the development of claims : the observed sinistrality significantly underestimates the global claims amount by approximately 10 to 20% for recent years. Finally, claims are observed for 2021 using the construction index from the FFB *fr. Fédération Française du Bâtiment* French Building Federation. The results are acceptable but are difficult to validate objectively. The results for 2016 are highly underestimated in terms of frequency and claims, whereas the results for 2017 are overestimated in terms of frequency. Comparison between the claim amounts, as shown in Figure 7.15b, is more difficult. Indeed, recent years' claims remain open, and the amount for open reserving underestimates the mean claim amount. Considering the particularity of the data, the R-squared metrics indicate that the reserving process is better evaluated when using the building geolocation information. As Charpentier et al. 2021 [162] concludes, subsidence coverage is difficult to model, even when new information is added, because proper validation is difficult.

7.5.2 Prevention and uninsurability

In France, the premium associated with CatNat coverage is strictly defined. Since the global premium is unknown, let us assume the premium for damage coverage is 300 euros, the CatNat premium is calculated as 12% of 300 euros [25]. In exchange for the CatNat tax/fee/contribution, the CCR reinsurer provides a quote share treaty up to 50% and unlimited stop-loss starting at approximately 300%-400% of the CatNat premium. Figure 7.16 compares the results of the performant model (GLM + GLM + XGBoost) and the referent model (RF + GLM + XGBoost) for pricing purposes, *i.e.*, annual indicators are not used. Each line corresponds to the S/C calculated on all contracts with a mean subsidence cost prediction or premium $P_i^{Subs.}$ less than or equal to the threshold u, *i.e.*,

$$\frac{\sum_{i|P_i^{Subs.}<u} S_i}{\sum_{i|P_i^{Subs.}<u} C_i}, \tag{7.6}$$

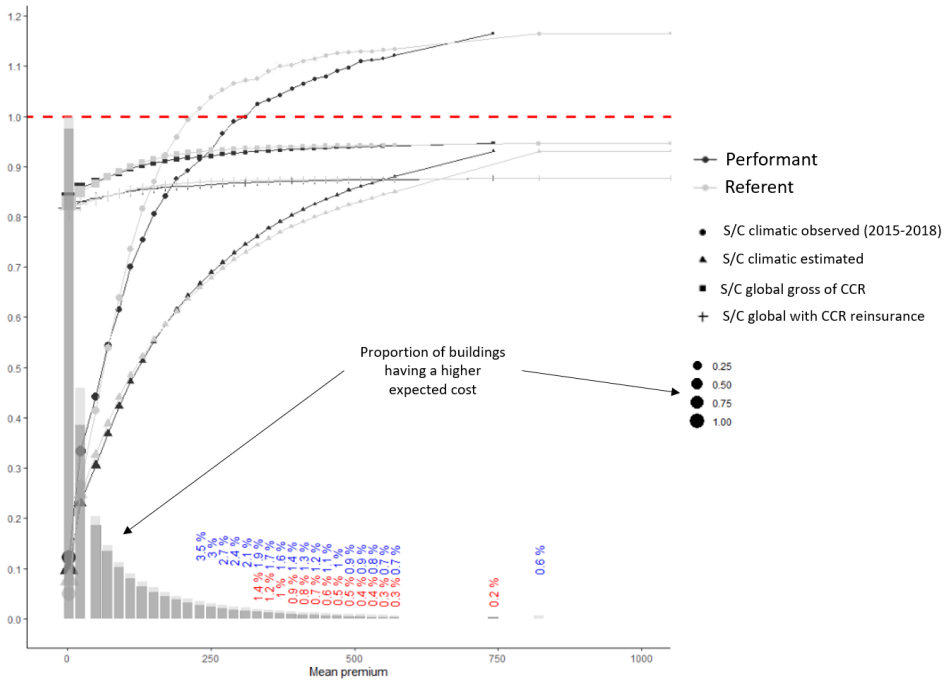
where S_i is the claim cost of the policyholder and C_i is the premium paid.

As explained in the previous section, the geolocated data greatly improve the model's performance, especially for the conditional frequency model. The differences between the referent and performant models show that insurers can identify houses for which losses greatly exceed the CatNat premium. One must also not that even if French insurers must incorporate 12% of the sum of damage insurance coverage for CatNat protection, one way around this requirement is to not insure some houses (or implement less competition or smaller discounts for these houses, higher premium increases, etc.). This approach could impact all levels, *e.g.*, the claim process via prevention during drought to very vulnerable houses for financial purposes. For instance, the CCR maintains the commission level at approximately 1% for insurers to provide proper prevention or indemnity controls to prevent fraud related to climatic risks.

The results indicate that the impact of subsidence has been important in recent years. Notably, the 12% contribution also covers floods, earthquakes, etc., which not considered in the S/C. In recent years, the cost of these climatic risks has been low, but historically, flood damage alone has had a higher cost than that of subsidence. The CCR greatly improves the S/C of reinsurers through its co-insurance. However, the increase in CatNat cost and variability increases the reinsurance cost. By means of the CCR and the mandatory CatNat program, 50% of the natural disaster losses are mitigated for all segments.

Are subsidence risks still insurable ? From the insurers' perspective, even though 4 consecutive years have a negative climatic S/C and have impacted the global S/C, insurers still have a global positive gross margin. Using the performant model, insurers could refuse to insure 1.4% of the portfolio against 3.5% (referent model with the meteorological data) to achieve the same improvement in expected climatic S/C. Refusals are possible, leading to an improvement of approximately 20 bp in the expected climatic S/C without lowering turnover substantially. Created in 1958, the French administration *Bureau Central de Tarification*(BCT) is tasked with helping people who cannot find an insurer to obtain compulsory insurance. In such cases, this administration can impose a contract with a given premium on an insurer.

25. For houses, the mean premium (houses + flat) was approximately 255 euros in 2020 according to FA (the French Insurance Federation) information. The mean premium for individual houses was around 300-400 euros.



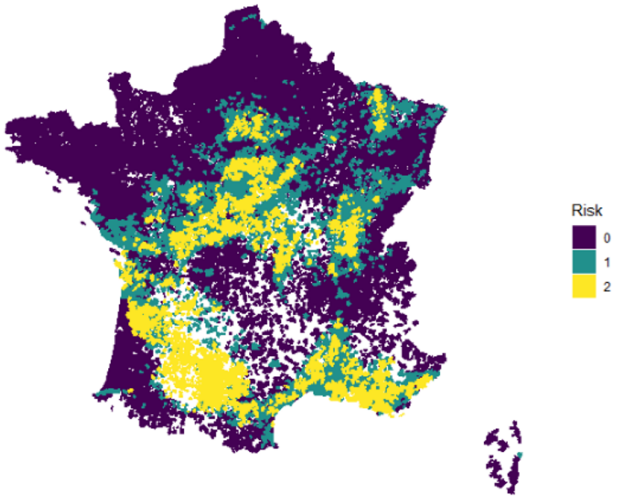
S/C	□	△	+	○
C	CatNat Premium	CatNat Premium	CatNat Premium + Damage premium	50 % CatNat Premium + Damage premium
S	Observed claims (2015-2018)	$\hat{E(S)} = P^{Subs.}$	$\hat{E(S)}$ + 80 % Damage premium	50% $\hat{E(S)}$ + 80 % Damage premium

FIGURE 7.16 – Comparison between the referent model (RF + GLM) and performant model (RF + GLM) for pricing purposes. The histogram corresponds to the proportion of contracts for which the expected subsidence cost/premium is higher. The first performant bar is not equal to 1 because the residuals contracts correspond to houses with missing information (period of construction or vegetated surface variable). Here, the S/C for a given premium represents the S/C of all contracts with an inferior of equal premium.

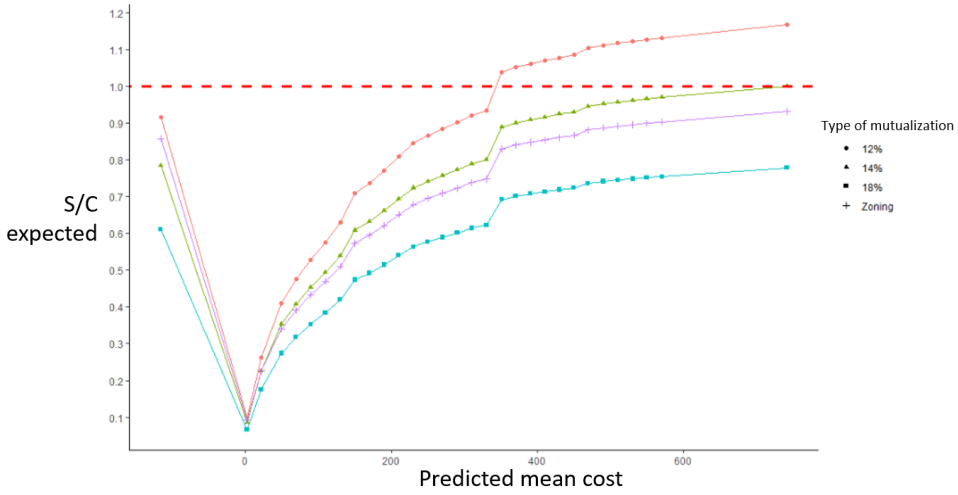
Clearly, the BCT can handle only a small volume of demand, but systemic rejection would be problematic and the existence of such an administration would be important to be aware of.

7.5.3 Evolution of the CatNat program

On 28 December 2021, a CatNat reform was enforced. The reform increased the period during which a municipality can declare a CatNat and the period during which an individual can declare subsidence damage from 2 to 5 years. This change might increase the probability of CatNat and the conditional frequency of claims compared to those considered in our model. Moreover, French insurers must pay the costs for the emergency relocation of disaster victims, as well as the costs of architects and project managers. In the reform, the different deductibles, depending on whether the municipality has a prevention plan, are unified.



(a) An example of subdividing the CatNat contribution according to drought expected cost on the municipalities available in our data. The CCR could produce an exhaustive map with the claims of the market.



(b) The combined ratio depending on the strategy with a contribution equal to 12%, 14%, 18% of damage premium or differentiated by zone. The first point refers to buildings for which the cost was not predicted due to missing values.

FIGURE 7.17 – The risk map according to drought exposure and vulnerability.

According to our results, the laws are heading in the right direction regarding insurers' contributions to climatic risk management but might have a negative outcome. The CatNat tax/fee is not sufficient in the face of the increase in drought frequency, climate change, and lack of effort toward drought prevention during building construction. The gross cost and management cost will continue to increase, especially for risky segments. To remain insurable, the expected climatic S/C should be less than one minus the management cost ratio, around 10 to 20%. Increasing the cost will increase the S/C of the riskier segments only. In the future, if the subsidence cost does not decrease, insurers for subsidence drought can focus on the most vulnerable houses to enforce a prevention plan. Nonetheless, some insurers may reluctantly insure houses vulnerable to subsidence risk. In one of the climatic ORSA scenarios of the ACPR, an increase of the CatNat fee of up to 18 % is considered. However, a better approach may be to adjust the CatNat premium based on the expected cost while maintaining the mutualization process. For instance, lower risk cases could still contribute up to 12 %, medium-risk cases up to 14% and higher risk cases up to 18%, as shown in the risk maps in Figure 7.17a. This would equilibrate the combined ratio, as shown in Figure 7.17b. The goal is to maintain the mutualization process without penalizing areas

without CatNat risks. Moreover, a municipality could change its' zone risk if sufficient anticipation and risk annihilation are implemented. This change would increase the premium but lower the financial gain of insurance refusal.

Finally, this process should also be implemented at least for floods. Some discussions on *Taxonomie1* of EIOPA (European Directives) consider including prevention measures in the design of products and asking for indicators for pricing and modeling of climate risk disasters. The brand image of these indicators is a powerful tool from our perspective if they are sufficiently refined.

In short, the French CatNat program has proven its robustness since its creation in 1989 for subsidence risk. The CatNat reform appears to be heading in the right direction, improving its ability to protect the insured individuals. Nonetheless, the increase in the CatNat premiums conceals the true problem. We recommend focusing on risk prevention and annihilation by promoting actions to address subsidence risks and homogenous risk portfolios through taxes or lower reinsurance contracts to ensure these risks are still financially attractive. Whether the building industry and not only insurers should participate in the CatNat program should also be considered.

7.6 Conclusion

This paper shows that the improvement and transparency of open data available in France enable actuaries to model subsidence claims more precisely. To access this information, insurers should geolocate each insured building. The downside of the model improvement is that some houses may become uninsurable. Even if a better prevention plan could be implemented, the authors worry that the insurance market would prefer to not accept a nonnegligible number of vulnerable houses. French and European legislators are heading in the right direction by increasing the insurers' CatNat contribution and prevention attractiveness. Nonetheless, the CatNat premium should be increased and further regulation should be implemented to address the root of the risk (during construction or during drought) to increase the insurability of this type of risk for all segments.

This work shows the discrepancies between machine learning methods (XGBoost and random forest) and GLM. The linear structure of GLM versus nonlinear ML embodies why French actuaries prefer GLM for most actuarial applications. Indeed, claims data are neither perfect nor fully developed. Therefore, a black box method often leads, without proper control, to a fully data-driven model without fully understanding the underlying particularity of the data. This paper presents an application to address the robustness of black-box methods. However, many more controls/tests must be performed for actuarial uses.

We have seen that new datasets with more granularity may change the ability of insurers to determine that some individual contracts become uninsurable. The following open research question remains : how should society address this issue and make sure that insurance remains inclusive ? This issue does not concern only household insurance. It is also relevant for biometric risks in life insurance for example. We leave this question for further reseach.

Acknowledgement

This paper stems from a mission of research of development in the context of the product "Smart Home Pricing". The authors thank the anonymous firm that has allowed us to use the portfolio, geolocate it and create this dataset. All the values in this paper have been anonymized. The views expressed in the paper do not represent the views or positions of the firms linked to this project.

Conflicts of interest

The first author declares a conflict of interest, being a full-time salaried employee of the project commercially named *Smart Home Pricing* developed by Addactis FRANCE and Nam.R. Nonetheless, he declares no financial interest. The remaining author has no conflicts of interest to declare.

7.6.1 Gaspar Database

The quality of the GASPAR database is good starting from 2005. Based on current knowledge, no proper quality evaluation has been conducted on drought and subsidence CatNat declaration and PPRN. An extensive analysis found few errors, only before 2009. The database version studied is from 19 July 2021. The subsidence PPRN is associated with the number 157, but some PPRNs -134 (drought), also refer to it. The following municipality PPRNs have been corrected "09317", "24013", "31438", "31307", "32168", "36151", "36183", "36201","68164", "68310", "81091", "81113", "81153", "81155", "89457","89275", "89213", "32022", "32081", "33063" and "89356". As shown in Figure 7.18, the CatNat declaration was not as clear before 2001. Therefore, the quality of all CatNat declarations before 2001 should ideally not be considered.

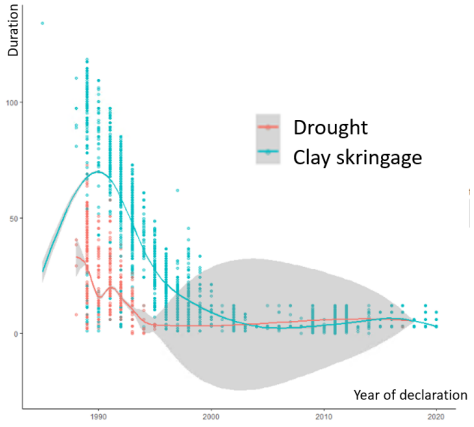


FIGURE 7.18 – The data come from the gross GASPAR database (19 July 2021). The subsidence/drought CatNat process starts in 1989 and the data become stationary starting at the end of 2003. Initially, the criteria for clay shrinkage (subsidence) and drought were not sufficiently clear. To see the trend, for each type of CatNat, a line corresponding to the marginal impact of a univariate GAM regression is plotted.

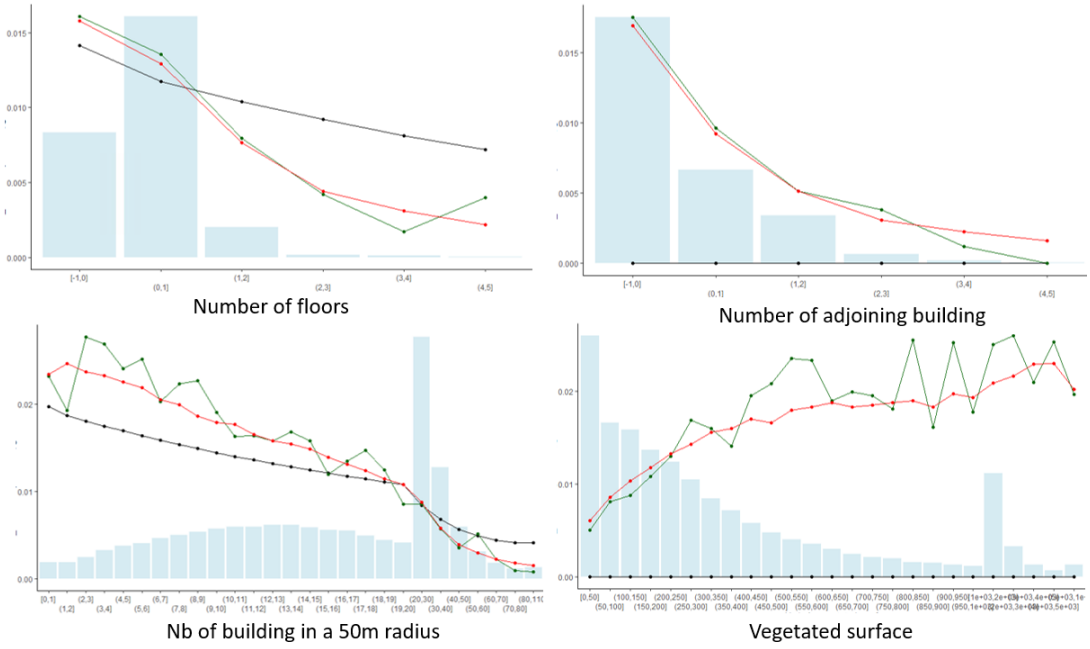
7.6.2 One-way analysis

Figures 7.19a and 7.19b present one-way analyses of variables used in GLM evaluated on the testing database. Blue histograms refer to the exposition, the green line represents the observed value, and the red line shows the predicted value. The black line refers to the marginal impact captured by the GLM. For confidentiality purposes, not all marginal impacts are plotted.

7.6.3 Some Shapley interactions

Figures 7.20 and 7.21 present Shapley values calculated for a simple XGBoost model trained on 2010-2018 historical data, including the meteorological annual indicators. The subset used to determine the Shapley values includes 20 000 random rows from the training and testing dataset. The Shapley score convergence is stable starting from 15 000 rows used.

(a) One way-analysis for GLM CatNat.



(b) One way-analysis for GLM conditional frequency.

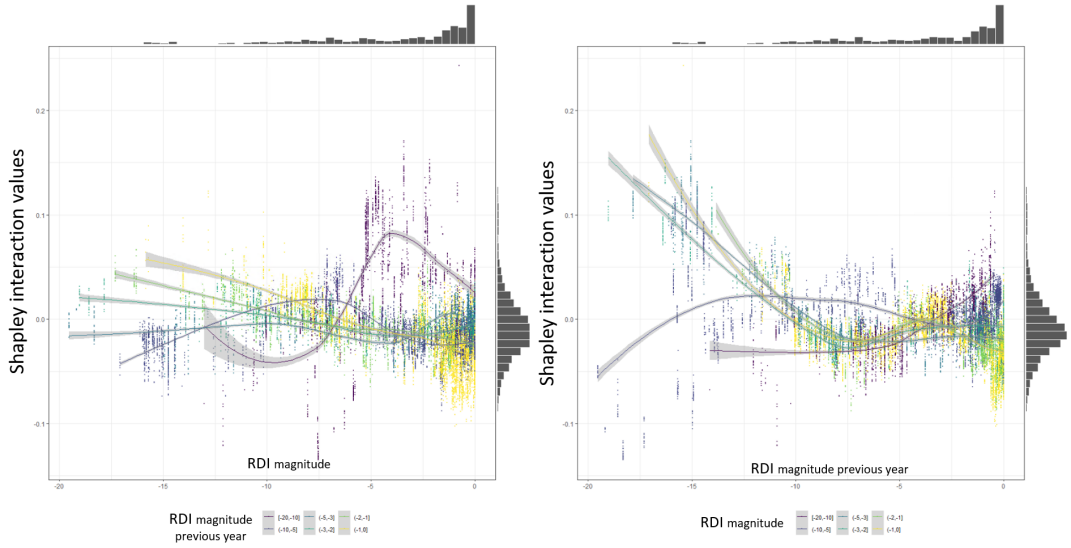


FIGURE 7.20 – Shapley interaction plot showing links between the previous and following meteorological years. Each line represents the marginal impact of a GAM model fitted by subcategories of the other variable.

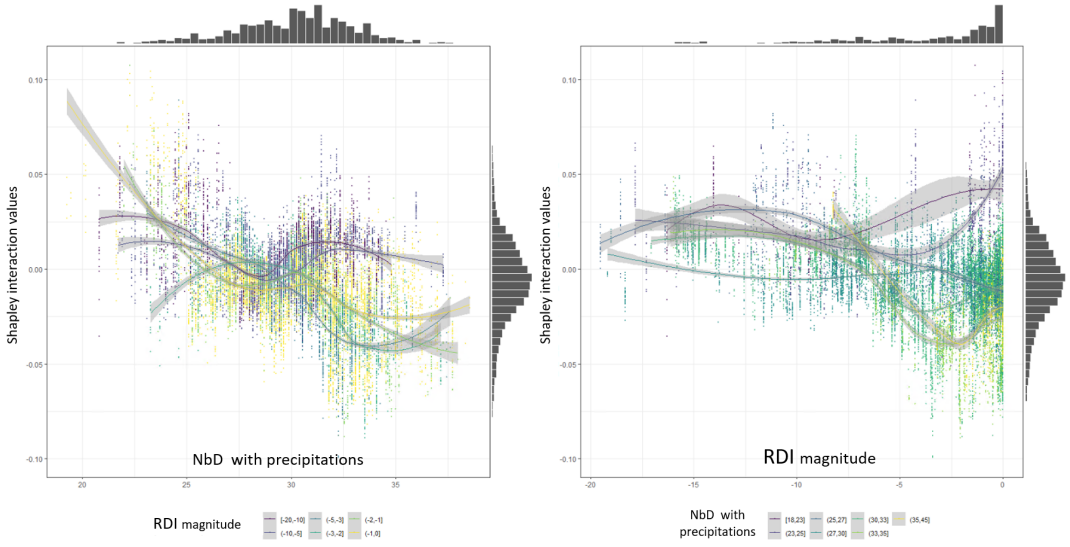


FIGURE 7.21 – Not all interactions are well captured. The variable precipitation quantity significantly improves the model's performance, but the marginal impact differs depending on the municipality's climate. Other interactions with the indicator SSWI or with the precipitation quantity of the previous year are also relevant.

Bibliography

[158] Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120 :70–83.

[159] Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M. (2016). mlr : Machine learning in r. *Journal of Machine Learning Research*, 17(170) :1–5.

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

[161] Cazaux, E., Meur-Férec, C., and Peinturier, C. (2019). Le régime d'assurance des catastrophes naturelles à l'épreuve des risques côtiers. aléas versus aménités, le cas particulier des territoires littoraux. *Cyber geo, european journal of geography*.

[162] Charpentier, A., James, M. R., and Ali, H. (2021). Predicting drought and subsidence risks in france. *Natural Hazards and Earth System Sciences Discussions*, 2021 :1–27.

[163] Douvinet, J. and Vinet, F. (2015). La carte des arrêtés catnat pour les inondations : analyse spatio-temporelle. *M@ppmonde*, 1.

[164] Efroymson, M. A. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203.

[165] Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[166] Hoyos-Idrobo, A., Schwartz, Y., Varoquaux, G., and Thirion, B. (2015). Improving sparse recovery on structured images with bagged clustering. In *2015 International Workshop on Pattern Recognition in NeuroImaging*, pages 73–76. IEEE.

[167] Ionita, M. and Nagavciuc, V. (2021). Changes in drought features at the european level over the last 120 years. *Natural Hazards and Earth System Sciences*, 21(5) :1685–1701.

[168] Jonathan, S., Gustavo, N., Jürgen, V., and Barbosa, P. (2016). Meteorological droughts in europe : Events and impacts – past trends and future projections. Technical Report EUR 27748 EN, Publications Office of the European Union, Luxembourg.

[169] Mack, T. (1991). A simple parametric model for rating automobile insurance or estimating ibnr claims reserves. *ASTIN Bulletin : The Journal of the IAA*, 21(1) :93–109.

[170] Maillard, G., Arlot, S., and Lerasle, M. (2021). Aggregated hold-out. *Journal of Machine Learning Research*, 22(20) :1–55.

[171] McKee, T. B., Doesken, N. J., Kleist, J., et al. (1993). The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 17, pages 179–183. California.

[172] Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

[173] Renshaw, A. E. and Verrall, R. J. (1998). A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, 4(4) :903–923.

[174] Spinoni, J., Naumann, G., Carrao, H., Barbosa, P., and Vogt, J. (2014). World drought frequency, duration, and severity for 1951–2010. *International Journal of Climatology*, 34(8) :2792–2804.

[175] Vidal, J.-P. and Wade, S. (2009). A multimodel assessment of future climatological droughts in the united kingdom. *International Journal of Climatology : A Journal of the Royal Meteorological Society*, 29(14) :2056–2071.

[176] Wilhite, D. A. and Glantz, M. H. (1985). Understanding : the drought phenomenon : the role of definitions. *Water international*, 10(3) :111–120.

[177] Zou, G. (2004). A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, 159(7) :702–706.

Chapitre 8

Conclusion et perpectives

La tarification en non-vie est soumise à de nombreuses contraintes en particulier en MRH. Ces travaux ont permis de mieux comprendre le lien entre les données, les assurés et les risques couverts par les assurances habitations. Dès lors que les informations ne proviennent plus d'un assuré, les processus de souscriptions sont modifiés et de nouvelles problématiques apparaissent. Les contraintes de souscriptions dont l'utilisation de calculette tarifaire obligent l'utilisation de méthodes statistiques facilement interprétables, mais aussi de données endogènes à la souscription à la portée de tous.

L'ajout de données exogènes à l'adresse est tout d'abord un atout permettant d'améliorer de façon significative la connaissance des assureurs. Par nature, ces données se différencient par leur précision, leur diversité, leurs complexités et leur qualité. À ce titre, nous avons vu leur diversité, leur intégration dans un processus de souscriptions et leur apport informationnel. Dans le même département de travail où je me trouve, ces résultats techniques ont été validés par la suite sur d'autres assureurs. En performance, ces données sont suffisantes à elles-seules pour tarifer un produit d'assurance habitation. Cependant, de nouvelles problématiques légales, juridiques, opérationnelles ou techniques apparaissent et d'autres contraintes disparaissent suite à l'utilisation de nouvelles données.

L'étude du risque de sécheresse de subsidence démontre que l'utilisation de données à l'adresse permet de mieux modéliser les risques. L'amélioration de la précision vient questionner l'assurabilité de certaines habitations. La législation française en vigueur a permis aux catastrophes naturelles de rester assurables. Au vu de l'augmentation des sécheresses à venir, de l'évolution de la législation combinées avec l'amélioration de la précision des modèles et des données, cette assurabilité du risque sécheresse se questionne.

De nombreuses perspectives subsistent pour l'utilisation de données à l'adresse. La tarification n'est qu'un des sujets parmi d'autres sujets comme le provisionnement, la réassurance, la prévention, l'évaluation des émissions GES ou CO2... Le principal enseignement de ces travaux est que l'augmentation de la précision de l'information ne doit pas tendre vers un accroissement de la complexité des modèles. L'exemple de la prise en compte de la qualité de la donnée dans des modèles simples à structures linéaires en est un bon exemple. Rappelons que les actuaires doivent répondre de la qualité de données qu'ils utilisent et des modèles sous-jacents.

Finalement, il existe autant de tarifications à l'adresse que d'approches de la donnée. Les choix des informations, des référentiels, des partenaires et de leurs utilisations influent sur la transposition des méthodes présentées dans cette thèse.

Perspectives Ces travaux nous ont poussés à étudier l'intégration de la qualité de données dans des modèles de régressions simples comme les modèles linéaires et les modèles linéaires généralisés. Sans complexifier les hypothèses sur la structure de la qualité de données, les résultats démontrent que la qualité de données a une influence sur l'anti-sélection et sur la performance des modèles eux-mêmes. Cependant, les résultats démontrent aussi que même si les hypothèses étudiées sont très restrictives, l'impact de la qualité de données n'est pas immédiat à appréhender.

Cependant, l'accroissement de la complexité de la qualité se limite aussi par la connaissance que nous en avons. Opérationnellement, avoir une confiance aveugle dans des données externes est un risque. Dans le cadre linéaire, la structure sur la qualité de données pourraient être élargie : avec une dépendance entre les erreurs et les observations ou bien que d'autres hypothèses sur la distribution des erreurs suivent. À travers ses nouvelles hypothèses, des liens avec la prise en compte d'erreurs de mesures dans les observations ou la théorie des valeurs manquantes seraient davantage immédiats. De plus, pour les extensions des résultats théoriques, la généralisation de la mesure d'ordre de qualité est

un sujet indispensable pour déterminer si une variable est de meilleure qualité qu'une autre.

Une piste suivante d'étude sont les modèles linéaires pénalisés ou des modèles CARTs ou RF. Dans la cadre de modèles CARTs, une difficulté supplémentaire apparait, car la forme de la structure des modèles s'apprend sur les données. Il faudrait enquêter théoriquement sur les modifications de la structure et de sa forme à cause de la qualité de la donnée. Il va sans dire qu'il faudrait également éviter certaines hypothèses comme l'indépendance de la qualité de données et les valeurs des observations dans de telles approches.

Appendices

Annexe A

Quelques commentaires sur l'histoire de l'assurance habitation

Aux prémices, l'assurance des biens concernait principalement les biens fluviaux ou maritimes [1]. En France, si c'est Colbert qui initia avec succès des instances nationales pour les produits assurances, de nombreuses tentatives le précédaient déjà [2]. L'assurance protégeant les habitations étant le sujet de cette thèse est récente. Fin du XVII ème siècle, l'assurance habitation concerna en premier lieu le risque incendie [3]. Il fallut attendre le milieu du XXème siècle pour qu'une assurance habitation proche de l'assurance de notre époque assurant plusieurs risques, soit commercialisée à grande échelle et avec succès. J'ajouterai que c'est principalement l'avènement de l'eau courante qui a favorisé la garantie DDE. De plus, l'incendie a la particularité d'être un risque destructeur et facilement vérifiable notamment l'asymétrie d'informations du sinistre, et ainsi la fraude entre assuré et assureur est le plus faible. Si c'est aux États-Unis dans les années 1950 [194] que les assurances **MRH** ont pris leur envol [189], des essais entre l'entre deux-guerre avaient déjà été faits sur le continent européen et dans certains États américains spécifiques.

Depuis chaque pays a adapté sa couverture assurantielle en particulier climatique. En France, les produits assurances **MRH** (Multi Risques Habitations) comportent obligatoirement des garanties contre les risques climatiques. Au contraire, certains pays, comme les États-Unis, ont développé en parallèle des assurances pour différents types de risques. En particulier, les produits ont évolué différemment entre les couvertures des risques attritionnels et graves et ceux climatiques ou rares. Historiquement, les législations sur les risques climatiques ont été implémentées dans les pays développés à la fin des années 1960 - *National Flood Insurance Act* de 1968 pour les Etats-Unis, *Japanese Earthquake Reinsurance* en 1966 au Japon. Toutefois, dès 1945, la Nouvelle-Zélande comme le Pays-Bas mettent en place une législation pour les risques sismiques ou inondations [180]. L'évolution des produits d'assurances a été logiquement spécialisée en fonction des risques sous-jacents du pays. Je conseille au lecteur intéressé de lire les analyses par pays européens faites par [Donguy].

Il est amusant de remarquer qu'au départ les calculs les plus anciens pour les primes des risques climatiques se basent essentiellement sur la garantie Incendie. Sans qu'il y ait d'articles discutant de ces spécificités, plusieurs remarques sont à faire. La garantie INC contrairement aux garanties comme le VOL ou les DDE est associée à des dommages souvent importants sur le bâtiment lui-même comme pour les risques climatiques. De plus, historiquement, c'était la garantie principale qui était toujours proposée. C'est logiquement, mais aussi par défaut que les cotisations ou l'obligation d'assurer les risques climatiques sont adossées à la garantie INC. En France, on retrouve des stigmates historiques à travers les taxes sur les conventions d'assurance (TCAS). La garantie INC est l'unique garantie taxée à 30 % avec la RC auto à 33 % (Article 1001 du Code Général des Impôts). Ces taxes servent en partie à financer les services départementaux d'incendie et de secours. Si je n'ai pas réussi à trouver des commentaires sur ce choix, la taxe INC provient d'une décision pendant la seconde guerre mondiale (article 21 de la loi du 31 janvier 1944). À l'époque, le taux était réduit de 25 % pour les assurances

1. Babylonien, Chinois, puis Grec et Romain (voir [191], [198]), ces assurances furent quasiment l'unique type d'assurances non-vie durant le Moyen-âge.

2. Les échecs étaient liés aux rivalités entre grandes puissances européennes comme celles des grandes réglementations de Lyon en 1435 comme l'explique [196].

3. En Angleterre, Christopher Werns entama des réflexions sur une assurance incendie ([182, p 6]) après le grand incendie de Londres mais ce fut Nicolas Barbon qui fonda *Insurance Office for Houses* avec succès en 1680 [182, p 7].

souscrites auprès des caisses départementales (en lien avec les services de luttes contre l'incendie). Finalement, cette TACS incendie élevée servirait historiquement à financer les services de luttes contre l'incendie.

Annexe B

Comparaison du modèle de coût sur une base par sinistres individuels ou par sinistres moyens

Soit une base $\mathcal{X}^{train}$ de modélisation pour le coût moyen d'un sinistre. Notons N_{sin} le nombre de sinistres de la base, $N(c)$ le nombre de sinistres d'un contrat c et PTF l'ensemble de contrats c ayant eu au moins un sinistre. La relation suivante s'en déduit facilement, $N_{sin} = \sum_{c \in PTF} N(c)$. On suppose que les sinistres sont indépendants.

Posons $\mathcal{X}_{CM}^{train}$, associée avec la notation $_{CM}$, une base où chaque ligne correspond à un contrat c observé durant une année et $N(c) \geq 1$ le nombre de sinistres tel que :

$$Y(c) = \frac{1}{N(c)} \sum_{i=1}^{N(c)} S_i(c). \tag{B.1}$$

Posons $\mathcal{X}_{S}^{train}$, associée avec la notation $_{S}$, une base où chaque ligne correspond à un sinistre S tel que :

$$Y = S. \tag{B.2}$$

La vraisemblance d'une loi gamma de paramètres (shape, scale) (α, β) s'écrit pour des prédictions $\hat{Y}$ pour la base $\mathcal{X}_{S}^{train}$:

$$\sum_{c \in PTF} \Big(\sum_{i=1}^{N(c)} (\alpha - 1) log\Big(S_i(c)\Big) - \beta^{-1} S_i(c) - log(\beta) - \alpha log(\beta) - log(\Gamma(\alpha)) \Big). \tag{B.3}$$

Pour les GLM, on pose $\hat{Y}_i(c) = \alpha\beta$ l'espérance de notre distribution (égale à $g^{-1}(\eta_i)$) et le facteur de dispersion $\phi = \alpha^{-1}$ qui est estimé à la fin de la procédure. En actuariat, les prédictions ne dépendent pas du sinistre, mais uniquement du contrat de l'assuré donc $\hat{Y}_i(c) = \hat{Y}(c)$. Les log-vraisemblances entre les deux modèles sont égales :

$$\begin{aligned} log\mathcal{L}(S, \hat{Y}, \phi) &\propto \sum_{c \in PTF} \Big(\sum_{i=1}^{N(c)} -\frac{S_i(c)}{\hat{Y}(c)} - log(\hat{Y}(c)) \\ &\propto \sum_{c \in PTF} -\frac{\sum_{i=1}^{N(c)} S_i(c)}{\hat{Y}(c)} - N(c) log(\hat{Y}(c)) \\ &\propto \sum_{c \in PTF} -N(c) \frac{\frac{1}{N(c)} \sum_{i=1}^{N(c)} S_i(c)}{\hat{Y}(c)} - N(c) log(\hat{Y}(c)). \end{aligned} \tag{B.4}$$

Ainsi la maximisation de la vraisemblable ou de façon équivalente la minimisation de la déviance entre les deux bases sont optimales pour les mêmes valeurs de β. En conséquence, les prédictions $\hat{Y}(c)$ sont égales. Cependant en fonction de la méthode d'estimation de ϕ, les résultats peuvent différer. Dans un

cadre de simulation des sinistres, la méthode du coût moyen induit un biais. En effet, si l'estimateur des moments est utilisé (la statistique de Pearson χ^2), sur la base $\mathcal{X}_S^{train}$,

$$\hat{\phi}_S = \frac{1}{N_{sin} - p} \sum_{c \in PTF} \sum_{i=1}^{N(c)} \frac{(S_i - \hat{Y}(c))^2}{Var(\hat{Y}(c))}, \tag{B.5}$$

avec p le nombre de coefficients estimés et Var la variance. Sur la base $\mathcal{X}_{CM}^{train}$, on obtient

$$\hat{\phi}_{CM} = \frac{1}{N_{PTF} - p} \sum_{c \in PTF} N(c) \frac{(\frac{1}{N(c)} \sum_{i=1}^{N(c)} S_i - \hat{Y}(c))^2}{N(c)^{-1} Var(\hat{Y}(c))}. \tag{B.6}$$

Un rapide développement permet de montrer que $\hat{\phi}_S$ sont différents $\hat{\phi}_{CM}$ s'il existe au moins un contrat multi-sinistré. Le résultat est le même lors de l'utilisation de l'estimateur de la déviance totale ; $\hat{\phi}_S = \frac{Dev._S}{N_{sin}-p}$ et $\hat{\phi}_{CM} = \frac{Dev._{CM}}{N_{PTF}-p}$. La déviance d'une loi de gamma s'écrit pour des prédictions $\hat{Y}$ pour la base $\mathcal{X}_S^{train}$:

$$-2 \sum_{c \in PTF} \sum_{i=1}^{N(c)} log\Big(\frac{S_i(c)}{\hat{Y}(c)}\Big) + \frac{S_i(c) - \hat{Y}_i(c)}{\hat{Y}(c)}. \tag{B.7}$$

La déviance d'une loi de gamma avec des poids de N_c s'écrit pour des prédictions $\hat{Y}$ pour la base $\mathcal{X}_{CM}^{train}$:

$$-2 \sum_{c \in PTF} N(c) log\Big(\frac{\sum_{i=1}^{N} S_i(c)}{N(c)\hat{Y}(c)}\Big) + N(c) \frac{\frac{1}{N(c)} \sum_{i=1}^{N} S_i(c) - \hat{Y}(c)}{\hat{Y}(c)}. \tag{B.8}$$

La dernière partie de l'addition est égale. D'après l'inégalité de la log-somme [197] avec $a_i = 1$ et $b_i = S_i - N(c) log\Big(\frac{\sum_{i=1}^{N} S_i(c)}{N(c)}\Big) \geq -\sum_{i=1}^{N(c)} log(S_i(c))$ quand $N(c) > 1$, on a la déviance de $Dev._{CM} \geq Dev._S$.

L'estimation de $\hat{\phi}$ impacte les p-values et aussi les autres métriques comme l'EDR ou l'AIC. En d'autres mots, les deux bases sont équivalentes pour la prédiction, mais pas pour la sélection de variables. En pratique, le nombre de multisinistrés étant faible en habitation, l'impact final est très faible.

Annexe C

Détection de changement de distribution : graves

Les méthodes usuelles pour le choix d'un seuil de grave sont :
— *QQ plot* : Méthode de comparaison des quantiles des données faces aux quantiles théoriques ;
— *Mean excess plot* : Moyenne des valeurs au-dessus d'un seuil. Sous l'hypothèse d'une GPD, la courbe devient linéaire à partir d'un certain seuil. Cette méthode permet plutôt difficilement de choisir un seuil ;
— Estimateur d'Hill : Recherche un seuil en fonction de la stabilisation du paramètre de l'indice de queue ;
— *Gertensgarbe* : Recherche un changement de comportement de la série des accroissements. En pratique, je ne recommande pas cette méthode seul car elle donne de nombreux seuils.

Très volatiles, les seuils de graves sont difficiles à déterminer. Chacune des méthodes font des hypothèses préalablement d'une distribution GEV sous-jacente et d'indépendances des observations sauf Gertensgarbe. C'est conjointement que les seuils doivent être déterminés. Je conseille au lecteur intéressé de regarder le livre de référence pour l'actuariat Embrechts et al., 2013 [184].

En tarification, les actuaires sont intéressés à modéliser les sinistres attritionnels et graves. Pour cela, il est nécessaire d'obtenir un seuil tel que le nombre de graves ne soit pas trop faible pour pouvoir faire une modélisation correcte. Le fait que les sinistres graves suivent bien une loi GEV est très anecdotique.

Je propose des illustrations sur trois exemples simulés et un exemple sur une base de données du package *CASdatasets* : FRECOMFIRE. L'objectif est de montrer les méthodes quand les bases de données vérifient les hypothèses et sur un exemple réel.

Posons $(S_i)_{i=1,\dots,n}$ la suite des montants sinistres ordonnés **par ordre croissant**.

C.1 La théorie des valeurs extrêmes

La distribution (GEV) *Generalised extreme values* non standardisée a une fonction de répartition qui s'écrit :

$$F(x|\xi,\beta,\mu) = \begin{cases} \exp\Big(-\exp(-\frac{x-\mu}{\beta})\Big) & \text{for } \xi = 0 \\ \exp\Big(-(1+\xi\frac{x\mu}{\beta})^{-1/\xi}\Big) & \text{for } \xi \neq 0 \text{ and } \xi\frac{x-\mu}{\beta} > -1 \\ 0 & \text{for } \xi > 0 \text{ and } \xi\frac{x-\mu}{\beta} \leq -1 \\ 1 & \text{for } \xi < 0 \text{ and } \xi\frac{x-\mu}{\beta} \leq -1\,, \end{cases} \tag{C.1}$$

avec ξ le paramètre *shape*, μ paramètre de location et β un paramètre d'échelle positif. Selon le théorème de Fisher-Tippett, toute distribution dont la loi du max est non-dégénérée appartient à cette distribution.

3 exemples seront simulés comme suit :

1. Base 1 : 25000 montants de sinistres simulés d'une loi de Gamma de paramètres $\alpha = 0.77$ et $\beta = 2679$;

2. Base 2 : 25000 montants de sinistres simulés d'une loi de Gamma de paramètres $\alpha = 0.77$ et $\beta = 2679$ et 250 montants de sinistres simulés par une GEV($\mu = 40000$, $\xi = 0.6$, $\beta = 15000$) ;
3. Base 3 : 24000 montants de sinistres simulés d'une loi de Gamma de paramètres $\alpha = 0.77$ et $\beta = 2679$, 1000 montants de sinistres simulés par une GEV($\mu = 40000$, $\xi = 0.3$, $\beta = 15000$) et 250 montants de sinistres simulés par une GEV($\mu = 40000$, $\xi = 0.6$, $\beta = 15000$).

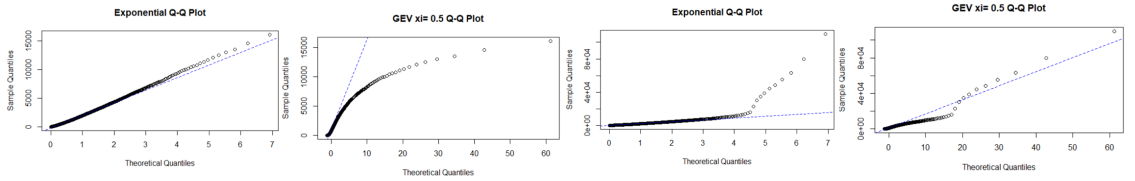
Les paramètres et les proportions ont été choisis à partir d'une base client incendie. La base 1 correspond à la base attritionnelle. La base 2 correspond à une base attritionnelle puis une base de sinistres graves. Cependant, la séparation entre les deux types est trop abrupte. Pour lisser et avoir une base plus ressemblante, des sinistres semi-graves avec une queue de distribution plus faibles sont ajoutés.

C.2 QQ plot

Un QQ-plot (décrite par Wilk et Gnanadesikan 1968, [199]) est une représentation graphique des quantiles d'une base de données face à d'autres quantiles, ici théoriques. Dans le cadre de la détection des sinistres graves, les points d'intérêts sont les quantiles élevés. L'idée ici est de caractériser la queue de distribution (lourde, fine ou exponentielle). Pour cela, les quantiles d'une loi exponentielle (distribution à queue exponentielle) sont utilisés en représentant le graphique suivant :

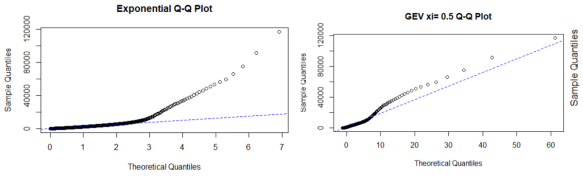
$$\Big\{ -ln(1 - \frac{i}{n+1}), -log(S_i) \Big\}_{i=1,...n}. \tag{C.2}$$

Si les points de quantiles élevés sont linéairement alignés, la queue de distribution est la même que la distribution théorique, ici normale. Si les points forment une courbe convexe, la queue de distribution est plus épaisse et inversement pour une courbe concave. Le cas d'une queue plus fine ou égale à la distribution Gamma en tarification suppose que les sinistres doivent être tous traités en tant qu'attritionnels.

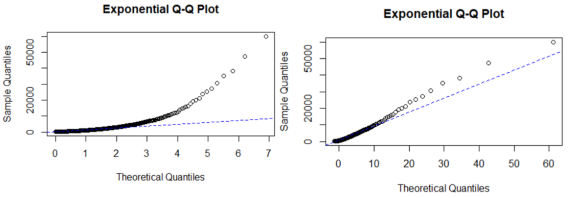


(a) QQplots sur la base 1.

(b) QQplots sur la base 2.



(c) QQplots sur la base 3.



(d) QQplots sur la base FRECOMPFIRE.

FIGURE C.1 – QQplots comparant à droite avec une loi exponentielle et à gauche une loi GEV de paramètre 0.5 .

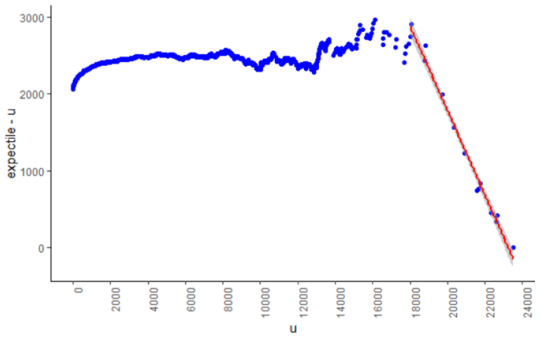
L'autre idée est de chercher les paramètres d'une loi de la théorie des valeurs extrêmes $GEV(\xi, \beta, \mu)$ sur les k plus grandes valeurs. En suite en comparant les quantiles empiriques et ceux théoriques, si les points sont linéairement alignés, le paramètre ξ permet de déterminer le type de queues.

C.3 Mean excess plot

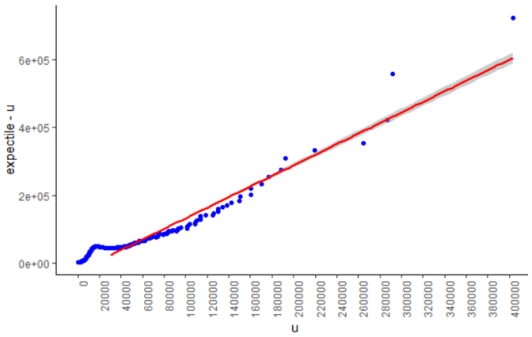
Le *Mean Excess Plot* [1] est un graphique représentant les points $\{u; \mathbb{E}(S-u|S>u)\}$ pour u positif. En pratique, la représentation suivante est utilisée :

$$\left\{S_i, \hat{M}(u) = \frac{\sum_{k=1}^{n} max(X_k - X_i, 0)}{\sum_{k=1}^{n} \mathbb{1}_{X_k > X_i}}\right\}_{i=1,...n} \tag{C.3}$$

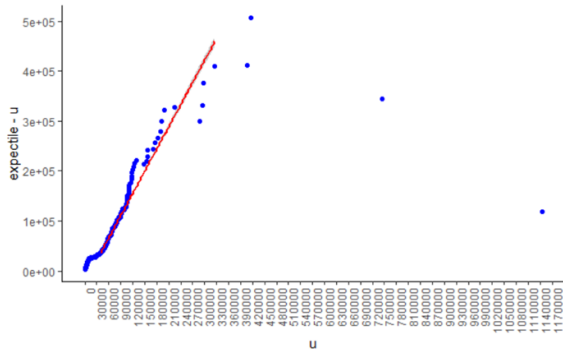
D'après la théorie des valeurs extrêmes, sous l'hypothèse d'i.i.d et des conditions sur les paramètres mentionnés après, les points de ce graphique s'alignent linéairement à partir du moment où ils suivent une $GEV(\xi, \beta, \mu)$. En effet, $M(u) = \frac{\beta}{1-\xi} + \frac{\xi}{1-\xi}u$ sous les conditions que $0 \leq u \leq \infty$ si $0 \leq \xi \leq 1$ et $0 \leq u \leq -\frac{\beta}{\xi}$ si $\xi \geq 0$. La difficulté est qu'il peut y avoir plusieurs alignements visibles. En fonction du résultat du QQ-plot, la droite peut être croissante, constante et même décroissante. Le *Mean Excess Plot* sert essentiellement à bien choisir le type de queue de la distribution. Il permet aussi de déterminer le seuil à partir duquel $\hat{M}(u)$ a un comportement adapté pour une GEV. Cette méthode est particulièrement volatile et il est nécessaire de considérer un grand nombre de points. (L'exemple d'Embrechts et al., 2013 [184] pages 297 et 298 est très parlant.)



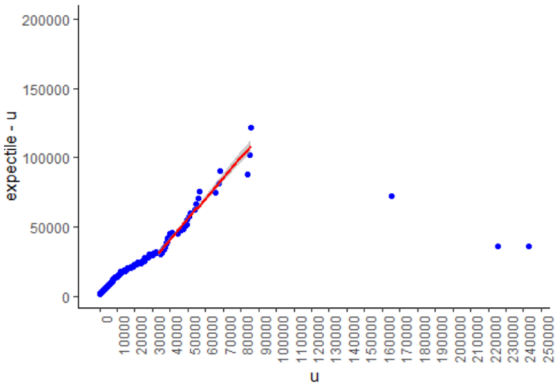
(a) *Mean Excess Plot* sur la base 1.



(b) *Mean Excess Plot* sur la base 2.



(c) *Mean Excess Plot* sur la base 3.



(d) *Mean Excess Plot* sur la base FRECOMPFIRE.

1. Proposé Davison and Smith 1990 [181] mais la métrique est étudiée par Benktander et al., 1960[178] article auquel je n'ai pas pu accéder.

C.4 Estimateur d'Hill

L'estimateur d'Hill $\xi^{Hill}(k,n)$ (Hill, 1975 [188]) est un estimateur qui converge vers l'indice de queue de GEV lorsque que le nombre d'observations $k = k(n)$ tend vers ∞ et $\xi > 0$.

$$\xi^{Hill}(k,n) = \frac{1}{n-k}\sum_{i=k}^{n} ln(\frac{S_{n-i+1}}{S_{n-k}}) \tag{C.4}$$

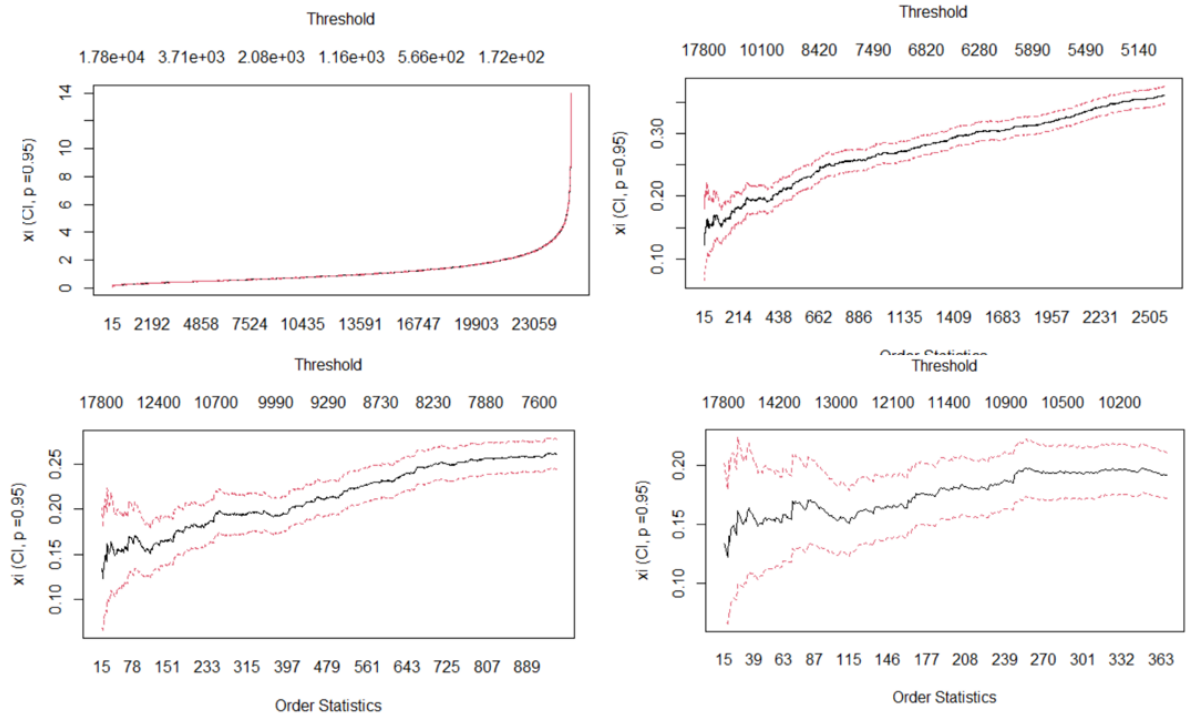
L'idée est de choisir un point à partir duquel les points sont linéairement constants sur le *Hill plot* $\left\{n - rang(S_i), \xi^{Hill}(k,n)\right\}$ pour tout $i = 1, ..., n$. L'hypothèse est qu'à partir d'un seuil, S suive une GEV d'un paramètre ξ. Dans ce cadre, $S|S > u$ suit aussi une GEV avec le même paramètre ξ d'où la stabilité du paramètre à partir d'un certain seuil. Il est important de noter que la valeur précise n'est pas intéressante pour la tarification. Ce que l'on cherche est le seuil du changement de distribution. Il existe d'autres estimateurs Pickands, Deckers Einmahl de Haan ou d'autres variantes d'Hill (voir Embrechts et al., 2013 [184]) plus robustes ou étendant les estimateurs à $\xi < 0$.

Si on reprend les hypothèses du papier (Hill, 1975[188]) et que l'on les compare à notre utilisation, deux limites apparaissent :

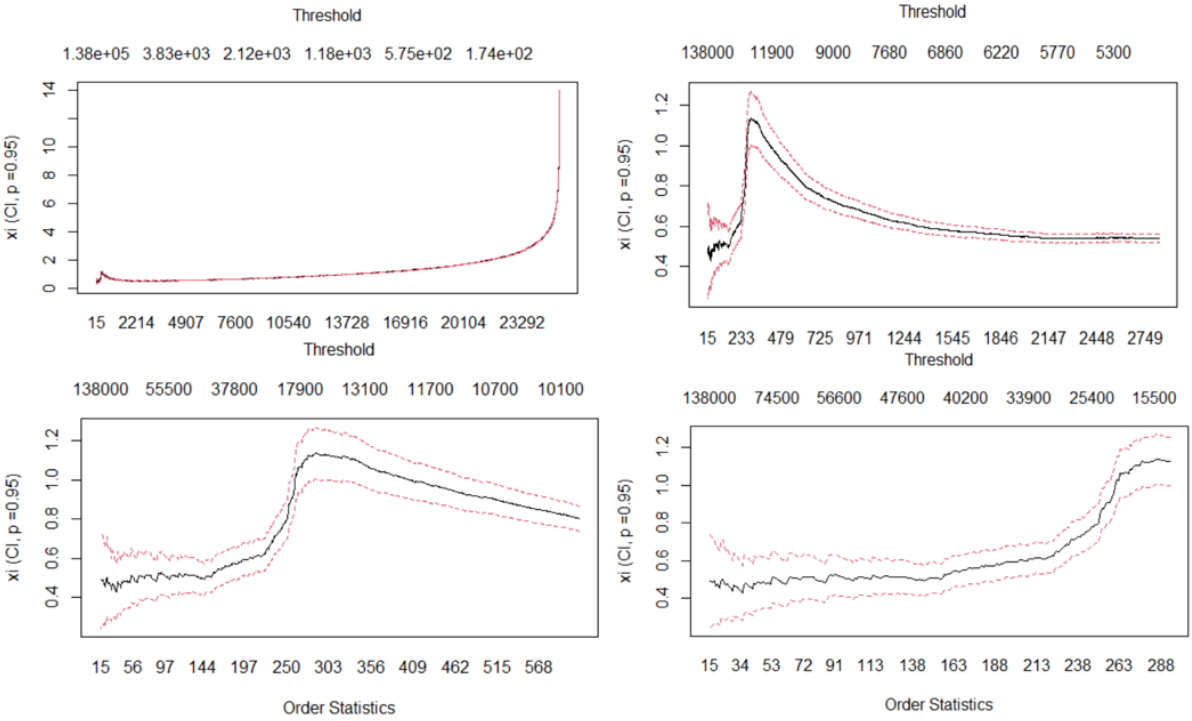
— Les $(S_i)_{i=1,...,n}$ doivent être un échantillon provenant d'une fonction strictement continue (ou Lebesgue mesurable). Dans certain cas, surtout avec les provisions d'ouvertures, cela peut ne pas être le cas ;

— Les observations doivent être indépendantes. Dans le cas des sinistres climatiques importants, les montants des sinistres sont dépendants à travers l'intensité de l'évènement ou/et spatialement et peut-être même dans le cadre du provisionnement.

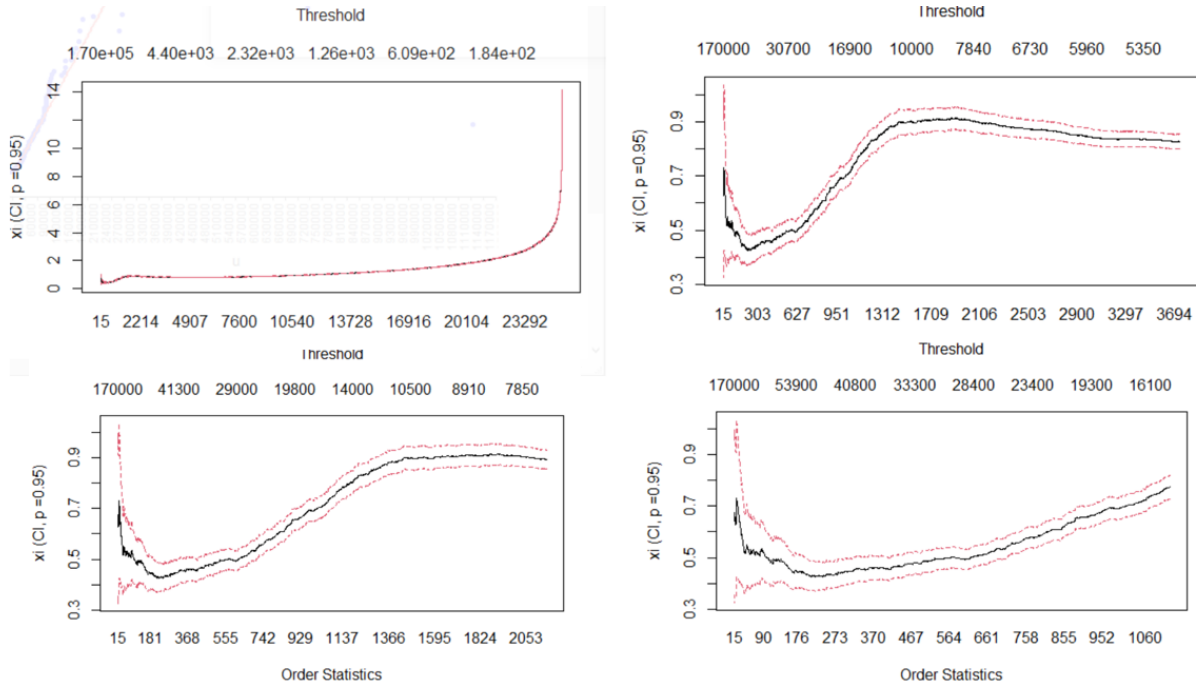
Pour le *Mean excess plot*, ces limites sont aussi valables. En pratique, à partir de l'Hill plot, le seuil correspond plutôt à la position d'une première instabilité de l'estimateur qui correspond à un changement de queue sous-jacente.



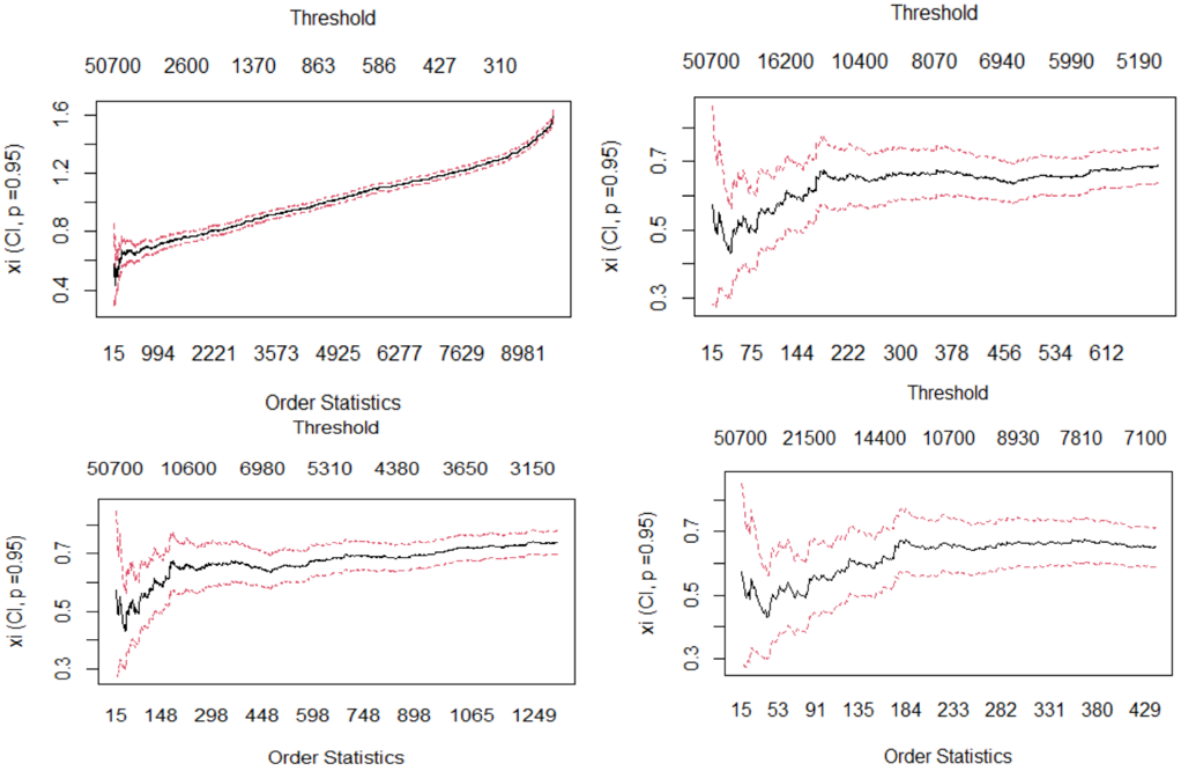
(a) *Hill Plot* sur la base 1.



(b) *Hill Plot* sur la base 2.



(a) *Hill Plot* sur la base 3.



(b) *Hill Plot* sur la base FRECOMPFIRE.

C.5 Gertensgarbe ou test Séquentiel de Mann-Kendall itératif

Popularisée en actuariat par l'article Cebrian et al. 2003 [179], cette méthode est associée à Gertensgarbe et Werner, 1989 [186]. Cependant, il a été publié en 1999 dans Climate Research [187] dont j'ai pu lire l'article [2]. Gertensgarbe est une méthode itérative et non paramétrique pour détecter un seuil de changements de la distribution entre le passage d'une distribution de queue fine à queue lourde. Dans la littérature statistique/hydrologique, cette méthode est associée à Sneyers (1990) [193] et Taubenheim (1989)[195]. Elle étudie les accroissements termes à termes en fonction des rangs et est non paramétrique. L'ajout de la méthode de Gertensgarbe et Weiner en 1999 est de réitérer cette méthode sur des sous-parties de la suite ordonnées des montants. Le *Gerstengarbe plot* est une application du test statistique séquentiel de Mann-Kendall. Ce test va détecter des changements de tendances (à la hausse ou à la baisse) dans les accroissements.

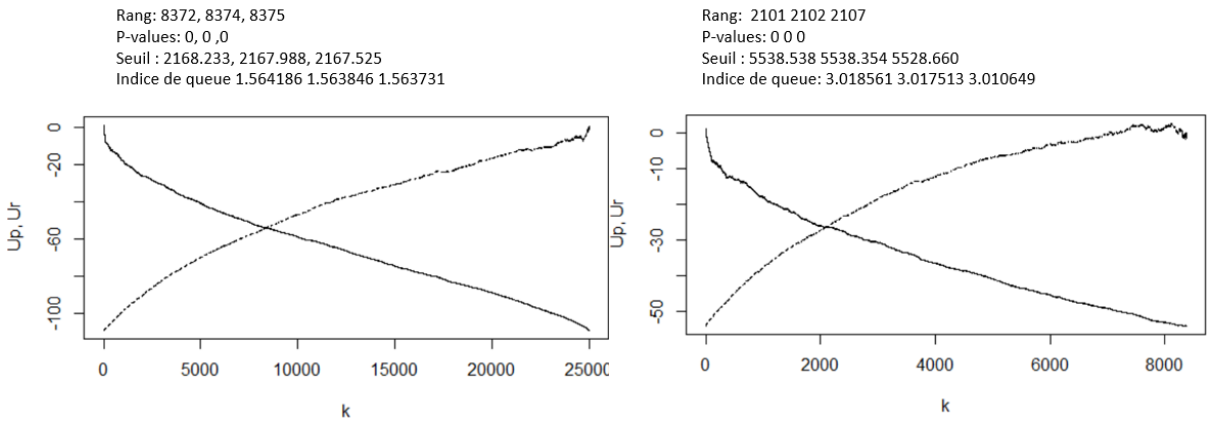
Posons $\Delta_k = x_k - x_{k-1}$, $U_i^* = \sum_{k=2}^{i} \sum_{j=2}^{k} \mathbb{1}_{\Delta_j \leq \Delta_k}$ et $\bar{U}_i^* = \sum_{k=2}^{i} \sum_{j=k}^{n} \mathbb{1}_{\Delta_j \geq \Delta_k}$. Deux grandeurs sont calculées U_i et $\bar{U}_i$:

$$\begin{aligned} U_i &= \frac{U_i^* - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i-1)(i+5)}{72}}}, \\ \bar{U}_i &= \frac{\bar{U}_i^* - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i-1)(i+5)}{72}}}. \end{aligned} \tag{C.5}$$

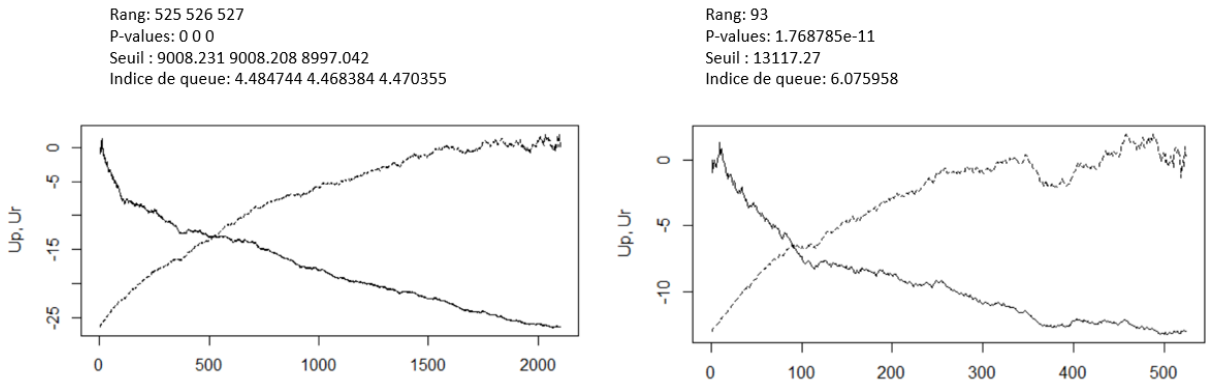
Le séquentiel de Mann-Kendall plot comporte les deux ensembles de points $\left\{i, U_i\right\}$ et $\left\{i, \bar{U}_i\right\}$. S'il existe, le point sécant représente les seuils possibles significatifs de changement de tendance. Cette procédure est itérative et réappliquée sur la sous-suite $(S_i)_{i=s,\dots,n}$ telle que s est un point sécant. Cette méthode permet d'obtenir plusieurs seuils de changements (ou aucun). Remarque : Le seuil correspond à une détection significative d'une tendance. C'est pourquoi les seuils souvent sont plus élevés que les autres méthodes. De plus, le nombre d'observations considéré influence les résultats.

Dans le cas de la détection des graves, les actuaires recherchent un unique seuil. Si on suppose que la distribution étudiée est composée uniquement de sinistres attritionnels suivant une loi de gamma et de sinistres graves suivant une loi à queue plus lourde à partir du seuil s. En commençant par la base complète, cette méthode va détecter un premier changement de tendance correspondant au changement le plus significatif : juste après la médiane des sinistres attritionnels. En effet, à partir de la médiane, les Δ_k augmentent alors qu'ils diminuaient avant la médiane. De plus, le nombre de sinistres attritionnels est prépondérant et Δ_k ne va plus que cloitre. En conclusion, le premier plot est toujours sécant qu'une seule fois. En regardant à partir de ce changement, on va obtenir un nouveau seuil correspond à l'accroissement le plus fort. On répète cette procédure jusqu'à ce que les courbes s'entrecroisent plusieurs fois, ce qui correspond dans notre cas au côté aléatoire et plus suffisamment informatif de nos sinistres. On obtient donc plusieurs seuils et il est nécessaire d'en choisir un.

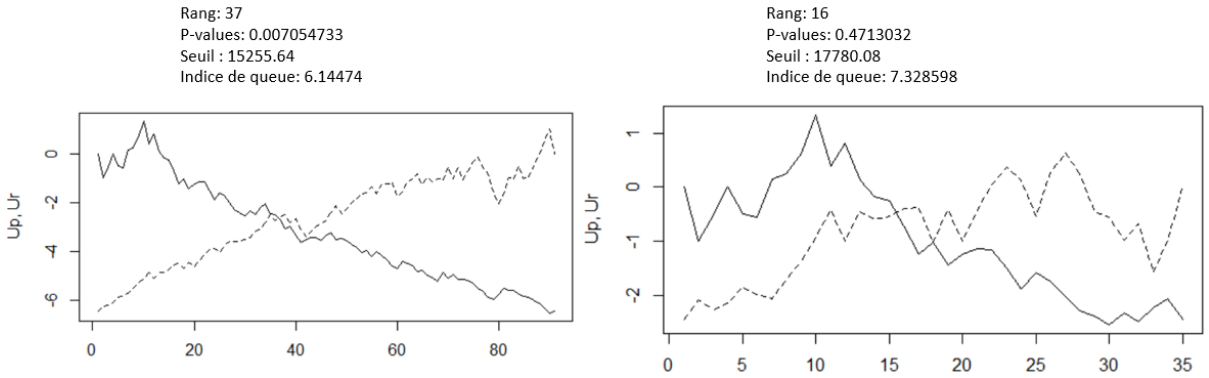
2. L'article[187] ne cite pas [186].



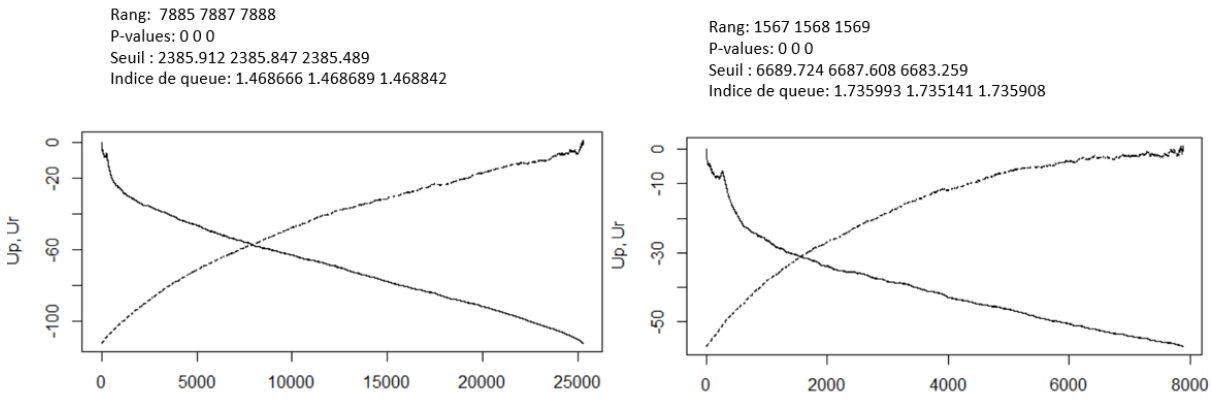
(a) *Gertensgarbe Plot* itérations 1 et 2 sur la base 1.



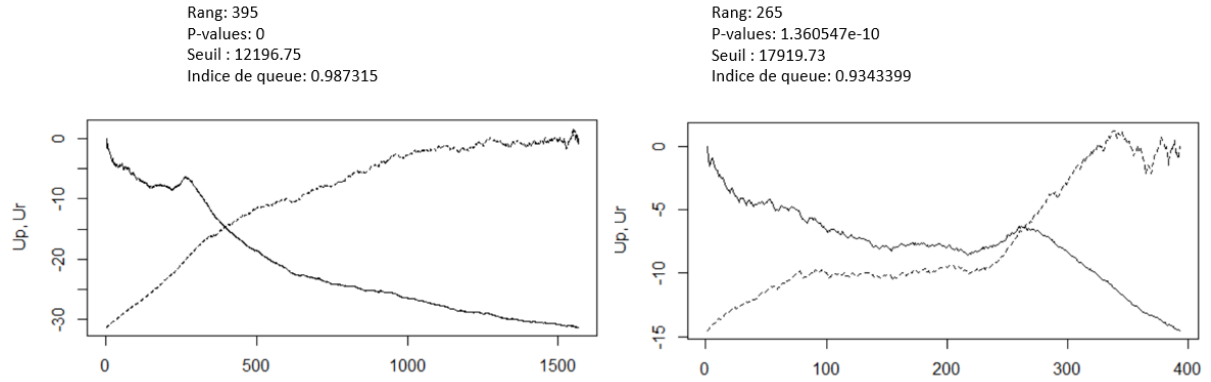
(b) *Gertensgarbe Plot* itérations 3 et 4 sur la base 1.



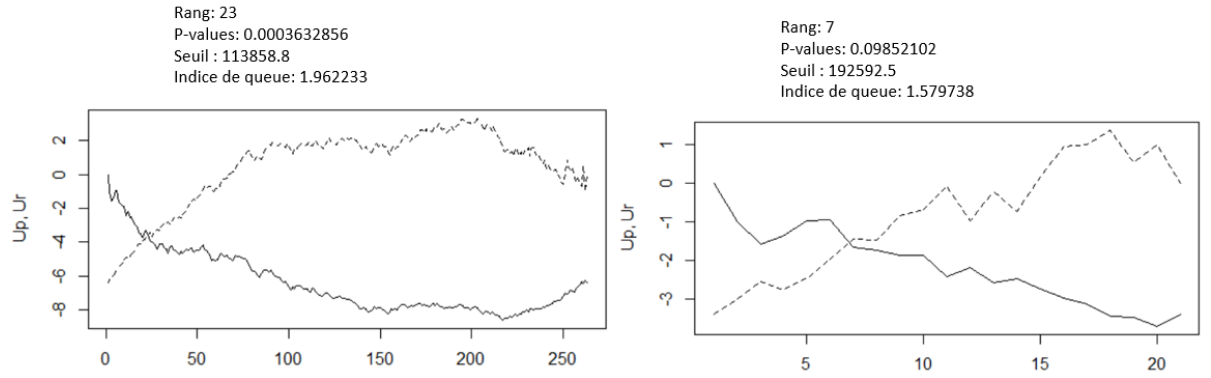
(c) *Gertensgarbe Plot* itérations 5 et 6 sur la base 1.



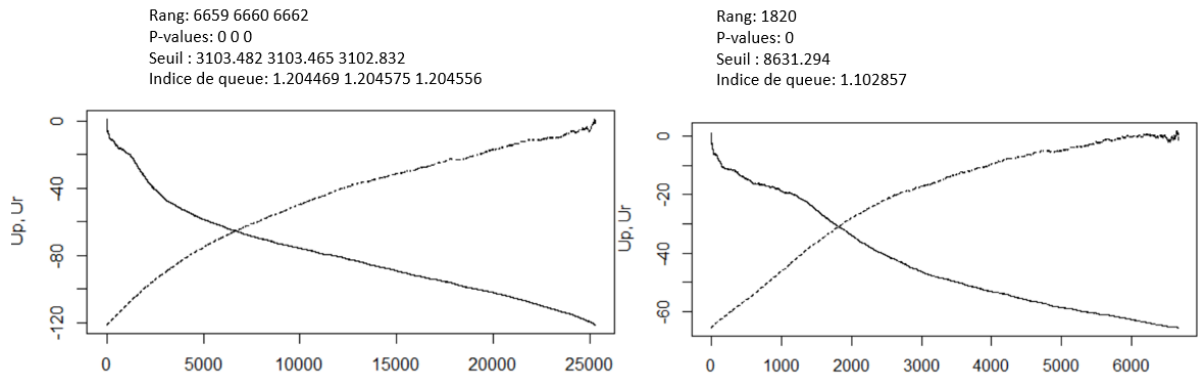
(a) *Gertensgarbe Plot* itérations 1 et 2 sur la base 2.



(b) *Gertensgarbe Plot* itérations 3 et 4 sur la base 2.

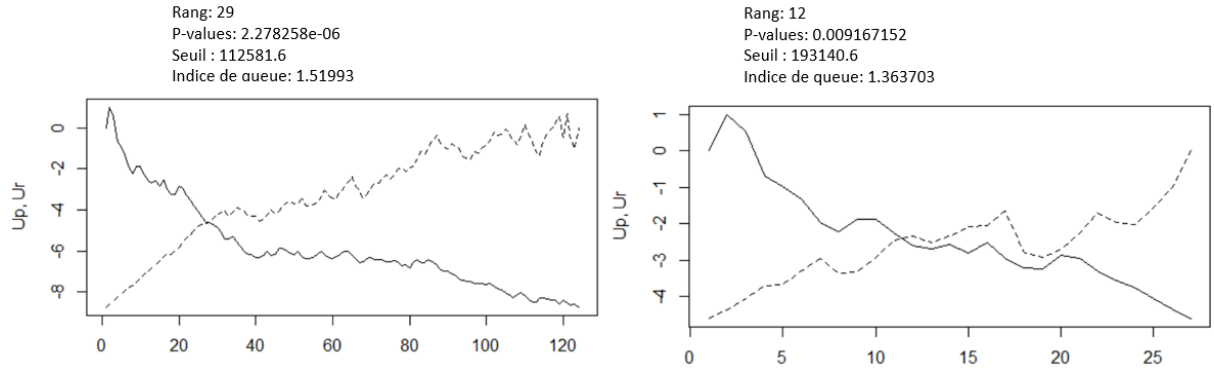


(c) *Gertensgarbe Plot* itérations 5 et 6 sur la base 2.

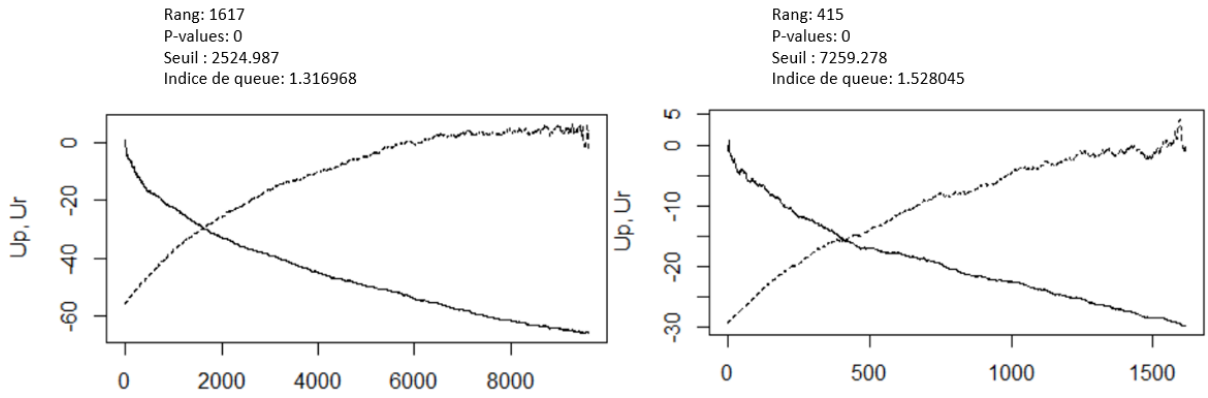


(a) *Gertensgarbe Plot* itérations 1 et 2 sur la base 3.

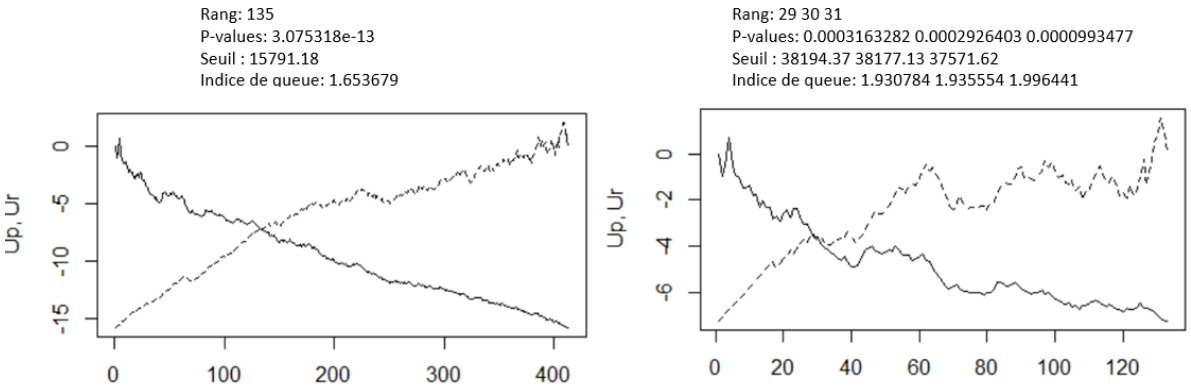
(b) *Gertensgarbe Plot* itérations 3 et 4 sur la base 3.



(c) *Gertensgarbe Plot* itérations 5 et 6 sur la base 3.



(a) *Gertensgarbe Plot* itérations 1 et 2 sur la base FRECOMPFIRE.



(b) *Gertensgarbe Plot* itérations 3 et 4 sur la base FRECOMPFIRE.

C.6 Structure attritionnelle apprenante des graves

Une autre méthode basée sur le Machine Learning a aussi été proposée par un collègue, Xuang Do-Quang, me semble aussi intéressante à mentionner, "structure attritionnelle apprenante des graves de Quang". Le principe algorithmique est détaillé dans l'algorithme 2. C'est la seule méthode avec la méthode Gertensgarbe qui cherche une base de sinistres attritionnels ne reposant pas sur un seuil pour une base de grave suivant une distribution GEV.

Algorithm 2: Graphique d'Extrapolation des graves de Quang

begin

[1] Séparation en base de train/test $\mathcal{D}_{train}, \mathcal{D}_{test}$;

[2] Choix d'un type de modèle et d'un jeu hyper-paramètres θ;

for $s_{grave} \in \mathbf{S}$ **do**

[i] $\mathcal{X}_{train} \leftarrow \{i^{eme}\ ligne\ de\ \mathcal{D}_{train} | Y_i > s_{grave}\}$;

[ii] $\mathcal{X}_{test} \leftarrow \mathcal{D}_{test}$;

[iii] Entrainement d'un modèle avec les hyper-paramètres θ sur $\mathcal{X}_{train}$;

[iv] Prédiction $\hat{\mathbf{Y}}$ d'un modèle avec les hyper-paramètres θ sur $\mathcal{X}_{test}$;

[v] $Cost(s_{grave}) \leftarrow Cost(\hat{\mathbf{Y}}, \mathbf{Y}, \mathcal{X}_{test})$;

Result: Afficher le graphique des points $(s_{grave}, Cost(s_{grave}))$;

L'astuce est d'utiliser un XGBoost avec une profondeur de 1 et avec le moins de paramétrisations possibles. Plusieurs graphiques en fonction de la métrique utilisée sont obtenues. La métrique doit être une métrique de segmentation. À la différence des autres méthodes, les critères tarifaires participent à la détermination du seuil. Le seuil correspond aux moments où les informations tarifaires ne permettent plus de détecter une augmentation de la sévérité. Cette particularité est un atout, mais aussi à des limites ; le résultat donné provient uniquement des données, mais est assujetti aux particularités de la base.

Ce point est important. Il est nécessaire d'éviter que le seuil de grave concerne une bimodalité qui peut être apprise par les données. L'exemple type est en assurance automobile où les informations sur les caractéristiques des voitures sont disponibles. Ainsi tous les sinistres associés avec voitures de luxes seraient associées à des sinistres graves alors qu'ils pourraient être inclus dans les modèles attritionnels.

Annexe D

Différentes approches pour les Ginis

En tarification, un des objectifs est de discriminer au mieux les profils risqués. Des outils graphiques et numériques permettent d'évaluer la justesse et le niveau de segmentation des modèles et les comparer entre-eux.

Le principal outil graphique utilisé est la courbe de concordance associée un Gini de concordance. Il ne faut pas la confondre avec la courbe ROC et son indice d'impureté le Gini qui est adaptée pour la classification. Une autre courbe de segmentation est aussi rencontrée : la courbe de Lorenz associée au coefficient de Gini. La différence entre l'indice de Gini et le coefficient de Gini est expliqué par dans l'article libre de droit de Schechtman et Schechtmann, 2016 [192]. Seul le coefficient de Gini de la courbe de concordance est utilisé dans cette thèse.

D.1 La courbe de Lorenz et de concordance

Développer à des fins économiques, Lorenz, 1905 [190] a proposé la courbe de Lorenz $\left\{s, LC(s) = \frac{\sum_{i=1}^{n} v_i \mathbb{1}_{rang(X_i)\leq s}}{\sum_{i=1}^{n} v_i}\right\}$ pour $s \in [0,1]$. Cette courbe de Lorenz représente la proportion de la population de la base de données en fonction d'une variable d'étude.

Dans le cas de la tarification, cette courbe appelée la courbe de concordance compare les prédictions $\hat{\mu}$ à l'observé $\hat{Y}$ (prime avec exposition Π_i et la sinistralité observée S_i). Dans la littérature, on parle d'indice de concordance des rangs. La courbe associée est la suivante, pour s allant de 1 à n :

$$\left\{\frac{\sum_{i=1}^{n} \mathbb{1}_{rang(\hat{\mu}_i)\geq s}}{n}, \frac{\sum_{i=1}^{n} \hat{Y}\mathbb{1}_{rang(\hat{\mu}_i)\geq s}}{\sum_{i=1}^{n} \hat{Y}}\right\}. \tag{D.1}$$

La courbe de concordance est égale à la courbe de Lorenz à une symétrie près. Il est facile de remarquer que toutes transformations des valeurs$\hat{\mu}_i$ qui ne changent pas le rang des $\hat{\mu}_i$ ne modifient pas la courbe. Par exemple, la courbe sera la même entre les prédictions $\hat{\mu}$ et $\hat{\mu}$ + 100. C'est pourquoi cette courbe ne mesure pas la performance d'un modèle. De plus, même si on suppose que les prédictions sont en moyenne égales, les courbes sont exactement les mêmes ; par exemple, pour $\hat{\mu}$ et $\hat{\mu}^*$ tel que pour $i = 1, ..., n$:

$$\hat{\mu}_i^* = \hat{\mu}_i \begin{cases} +100, \text{ si } rang(\hat{\mu}_i) < \left\lfloor \frac{n}{2} \right\rfloor \\ -100, \text{ si } rang(\hat{\mu}_i) > \left\lfloor \frac{n}{2} \right\rfloor \end{cases} \tag{D.2}$$

alors $\frac{\sum_{i=1}^{n} \hat{\mu}}{n} = \frac{\sum_{i=1}^{n} \hat{\mu}^*}{n}$ et les courbes sont les mêmes s'il n'y a pas de valeurs égales dans les prédictions.

Lors de l'implémentation des courbes, nous avons souvent remarqué que les algorithmes ordonnent les prédictions puis calculent ces courbes. Cependant, les cas, où les prédictions sont égales, ne sont pas traitées et entrainent une variabilité de la courbe. Celle-ci est souvent négligeable pour les modèles à beaucoup de variables, mais pour des modèles comme les CARTs ou GLM avec deux ou trois critères, il existe un nombre non négligeable de valeurs égales.

Il existe un moyen très simple pour supprimer cette volatilité. Il suffit de calculer la moyenne entre la courbe de Lorenz sur $(\hat{\mu}_i)_{i)1,..,n}$ et sur la courbe de Lorenz $(\hat{\mu}_{n-i})_{i=1,..,n}$ avec une méthode de tri qui ne change pas l'ordre des valeurs égales. Ainsi l'AUC calculé est celui voulu [1].

1. La preuve se fait bien mais nécéssite l'introduction de notation en dehors du scope de la thèse.

Finalement dans certaines entreprises, un graphique similaire est regardé :

$$\Big\{\frac{\sum_{i=1}^{n} v_i \mathbb{1}_{rang(\hat{\mu}_i)\geq s}}{\sum_{i=1}^{n} v_i}, \frac{\sum_{i=1}^{n} \frac{\hat{Y}}{v_i} \mathbb{1}_{rang(\hat{\mu}_i)\geq s}}{\sum_{i=1}^{n} \frac{\hat{Y}}{v_i}}\Big\}. \tag{D.3}$$

avec $\sum_{i=1}^{n} v_i$ l'exposition. Il permet de lire graphiquement un niveau d'exposition avec un niveau de charge. Une autre variante est d'ordonner en fonction de $rang(\frac{\hat{\mu}_i}{v_i})$ au lieu de $rang(\hat{\mu}_i)$ et regarder la segmentation sur les primes annuelles.

Dans la littérature actuarielle, Frees et al., 2014 [185] définissent aussi une courbe nommée *Ordered Lorenz Curve* :

$$\Big\{F_L(s) = \frac{\sum_{i=1}^{n} S_i \mathbb{1}_{Score_i/\Pi_i\leq s}}{\sum_{i=1}^{n} \Pi_i}, F_P(s) = \frac{\sum_{i=1}^{n} \Pi_i \mathbb{1}_{Score_i/\Pi_i\leq s}}{\sum_{i=1}^{n} \Pi_i}\Big\}. \tag{D.4}$$

pour un individu i S_i la sinistralité observée durant une période t, Π_i la prime ou prédiction associée à cette période et $Score_i$ un score de risque associé. Les points sont placés en $F_L(s), F_P(s)$ ou L sous-entend Loss and P Premium.

D.2 L'AUC et les Ginis

Une métrique usuellement utilisée pour la transposition numérique de la courbe est l'AUC (Area Under the Curve). L'AUC correspond à l'aire sous la courbe de la courbe de concordance. L'estimateur le plus simple est :

$$\frac{1}{n} \frac{\sum_{s=1}^{n} \sum_{i=1}^{n} \hat{Y} \mathbb{1}_{rang(\hat{\mu}_i)\geq s\}}}{\sum_{i=1}^{n} \hat{Y}}.$$

Si l'AUC est égale à 0.5, le modèle segmentant au global comme s'il n'y avait aucune segmentation. Si l'AUC > 0.5, le modèle segmente au global les observations. Si l'AUC < 0.5, de manière générale, cela veut dire que le modèle a sur-appris, a appris de mauvais liens de causalités ou que l'échantillon est trop faible. Tout comme le ROC, l'AUC est borné supérieurement par l'AUC du modèle dit saturé. En actuariat - tarification, cette borne est largement inatteignable à cause de la variabilité des sinistres.

Historiquement, l'indice de ("d'impureté") Gini (ROC) est souvent utilisé. Il se calcule comme :

$$Gini = 2(AUC - 1), \tag{D.5}$$

qui est deux fois l'aire entre la courbe et la bissectrice représentant une segmentation aléatoire.

Pour calculer la proportion de la segmentation existante apprise, le Gini normalisé $Gini^{norm}$ est défini et correspond à :

$$Gini^{norm} = \frac{Gini}{Gini_{sat}}. \tag{D.6}$$

Le $Gini^{norm}$ n'a d'interprétation que pour les courbes de concordances. Ces trois métriques sont équivalentes, cependant $Gini^{norm}$ permet plus facilement de comparer des gains entre des modélisations de variables différentes et bases différentes de données.

Annexe E

Exemple de calculette tarifaire

Les images suivantes (E.1, E.2 et E.3) sont une représentation de la calculette tarifaire. L'ensemble des coefficients se présente par garantie sous forme de tabulaire. De manière générale, un grand nombre d'options/garanties optionnelles y est ajouté. L'ensemble des chiffres sont arbitraires et sont à titre illustratif.

Calculette tarifaire MRH - maison exemple - prime nette

		Garanties	INC	ATT	TGN	ELEC	DDE	BDG	VOL	RC
		Etape								
Etape A : Intercepte		A								
A	=	Prime Agence	100,000	15,000	40,000	25,000	70,000	5,000	60,000	40,000
		Prime Courtier	90,000	13,500	36,000	22,500	63,000	4,500	54,000	36,000
Etape B :Nb pièces x Capital mobiliers		B								
B	= A x	Nb pièces x Capital mobiliers								
		[0;10000[	0,76	0,66	0,36	0,51	0,41	0,57	0,73	1,00
		[10000;20000[	0,77	0,68	0,36	0,67	0,47	0,58	0,82	1,00
		[20000;30000[	0,76	0,69	0,40	0,65	0,47	0,53	0,99	1,00
	1	[30000;40000[	0,79	0,82	0,42	0,68	0,53	0,59	1,10	1,00
		[40000;60000[	0,83	0,85	0,49	0,80	0,67	0,63	1,16	1,00
		[60000;80000[	1,02	0,87	0,55	0,79	0,71	0,76	1,20	1,00
		[80000;+[	1,09	1,07	0,55	0,99	0,68	0,79	1,37	1,00
		[0;10000[	0,82	0,91	0,59	0,60	0,59	0,71	0,78	1,00
		[10000;20000[	0,83	0,80	0,61	0,64	0,68	0,76	0,83	1,00
		[20000;30000[	0,88	0,92	0,60	0,73	0,72	0,72	0,98	1,00
	2	[30000;40000[	0,92	0,96	0,64	0,79	0,84	0,78	1,00	1,00
		[40000;60000[	1,12	1,11	0,71	0,88	0,93	0,87	1,17	1,00
		[60000;80000[	1,15	1,21	0,88	0,93	1,08	0,86	1,20	1,00
		[80000;+[	1,35	1,29	0,98	1,03	1,03	1,02	1,40	1,00
		[0;10000[	0,83	0,87	0,77	0,78	0,84	0,90	0,93	1,00
		[10000;20000[	0,92	0,96	0,84	0,89	0,96	0,92	1,00	1,00
		[20000;30000[	0,95	0,97	0,86	0,99	1,00	0,92	1,16	1,00
	3	[30000;40000[	1,01	1,04	0,96	1,12	1,03	1,05	1,17	1,00
		[40000;60000[	1,18	1,22	1,07	1,14	1,06	1,05	1,26	1,00
		[60000;80000[	1,19	1,16	1,14	1,16	1,36	1,34	1,36	1,00
		[80000;+[	1,20	1,37	1,43	1,32	1,41	1,40	1,54	1,00
		[0;10000[	1,04	0,87	1,08	0,90	1,11	1,20	0,94	1,00
		[10000;20000[	1,13	0,89	1,20	1,05	1,21	1,26	1,11	1,00
		[20000;30000[	1,14	0,99	1,24	1,11	1,31	1,23	1,33	1,00
	4	[30000;40000[	1,18	1,14	1,45	1,27	1,48	1,28	1,46	1,00
		[40000;60000[	1,19	1,12	1,40	1,46	1,59	1,40	1,65	1,00
		[60000;80000[	1,34	1,37	1,79	1,57	1,70	1,58	1,70	1,00
		[80000;+[	1,64	1,40	1,70	1,63	1,96	1,67	1,73	1,00
		[0;10000[	1,10	1,00	1,45	1,17	1,16	1,27	1,25	1,00
		[10000;20000[	1,15	1,07	1,76	1,30	1,49	1,54	1,59	1,00
		[20000;30000[	1,23	1,21	1,52	1,51	1,59	1,56	1,63	1,00
	5	[30000;40000[	1,32	1,22	1,78	1,61	1,61	1,57	1,67	1,00
		[40000;60000[	1,41	1,39	2,05	1,64	1,95	1,59	1,73	1,00
		[60000;80000[	1,59	1,52	2,23	1,80	2,00	2,01	1,96	1,00
		[80000;+[	1,60	1,58	2,33	2,03	2,10	1,96	2,29	1,00

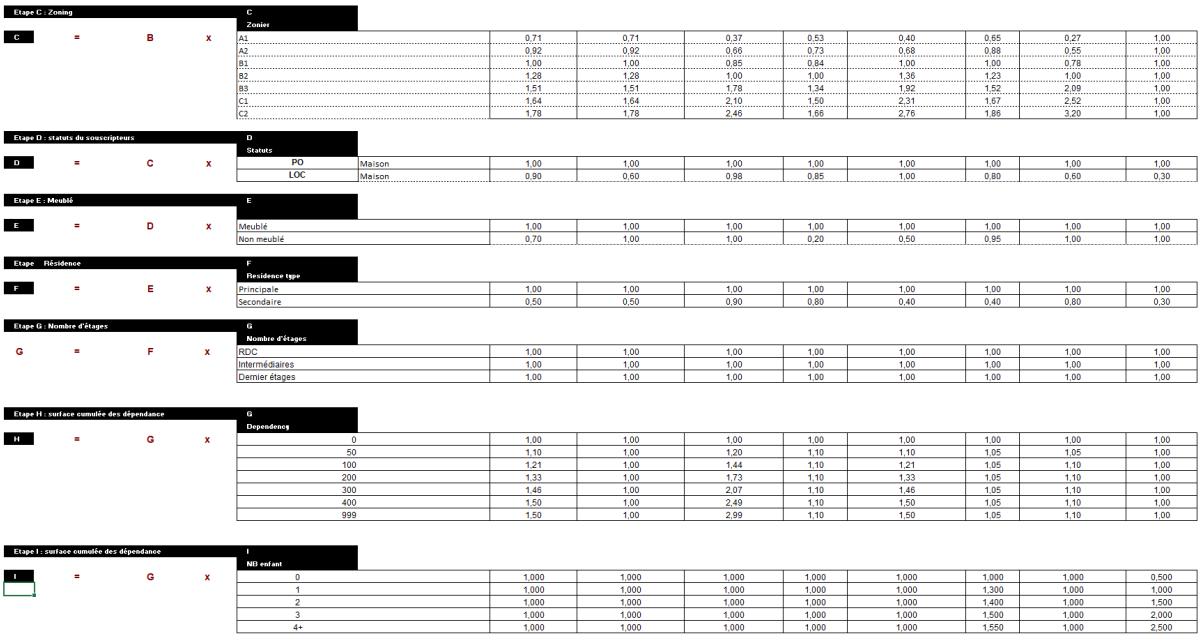
FIGURE E.1 – Exemple de calculette tarifaire pour le calcul de la prime de base. Dans le cadre d'interaction un grand nombre de coefficients peuvent apparaitre. Ici, j'ai choisi de prendre le nombre de pièces croisées avec le capital mobilier.

Mathématiquement et rigoureusement, une grille peut être définie de la façon suivante :

$$\Pi_{garantie}(c) = M_0 \prod_{\forall I_a(X_h) \& \forall h \in 1,...p | X_h(c) \in I_a(X_h)} M_{I_a(X)} \tag{E.1}$$

avec

- M_0 la constante ou le montant de référence de la garantie. Certain acteur module directement cette constante par les chargements et aussi en fonction du réseau de souscription ;
- $X_h(c)$ est la valeur du critère tarifaire X_h pour le contrat c ;
- $I_a(X)$ est un sous ensemble de X tel que $\forall a \& b, I_a(X) \cap I_b(X) = \emptyset$ et qu'il existe un intervalle pour toutes les valeurs possibles de X ;
- $M_{I_a(X)}$ le multiplieur du critère tarifaire $I_a(X)$ associé.



Etape C : Zoning

C	=	B	x

C	Zonier								
A1		0,71	0,71	0,37	0,53	0,40	0,65	0,27	1,00
A2		0,92	0,92	0,66	0,73	0,68	0,88	0,55	1,00
B1		1,00	1,00	0,85	0,84	1,00	1,00	0,78	1,00
B2		1,28	1,28	1,00	1,00	1,36	1,23	1,00	1,00
B3		1,51	1,51	1,78	1,34	1,92	1,52	2,09	1,00
C1		1,64	1,64	2,10	1,50	2,31	1,67	2,52	1,00
C2		1,78	1,78	2,46	1,66	2,76	1,86	3,20	1,00

Etape D : statuts du souscripteurs

D	=	C	x

D	Statuts								
PO	Maison	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
LOC	Maison	0,90	0,60	0,98	0,85	1,00	0,80	0,60	0,30

Etape E : Meublé

E	=	D	x

E								
Meublé	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Non meublé	0,70	1,00	1,00	0,20	0,50	0,95	1,00	1,00

Etape Résidence

F	=	E	x

F	Residence type								
Principale		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Secondaire		0,50	0,50	0,90	0,80	0,40	0,40	0,80	0,30

Etape G : Nombre d'étages

G	=	F	x

G	Nombre d'étages								
RDC		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Intermédiaires		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Dernier étages		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Etape H : surface cumulée des dépendance

H	=	G	x

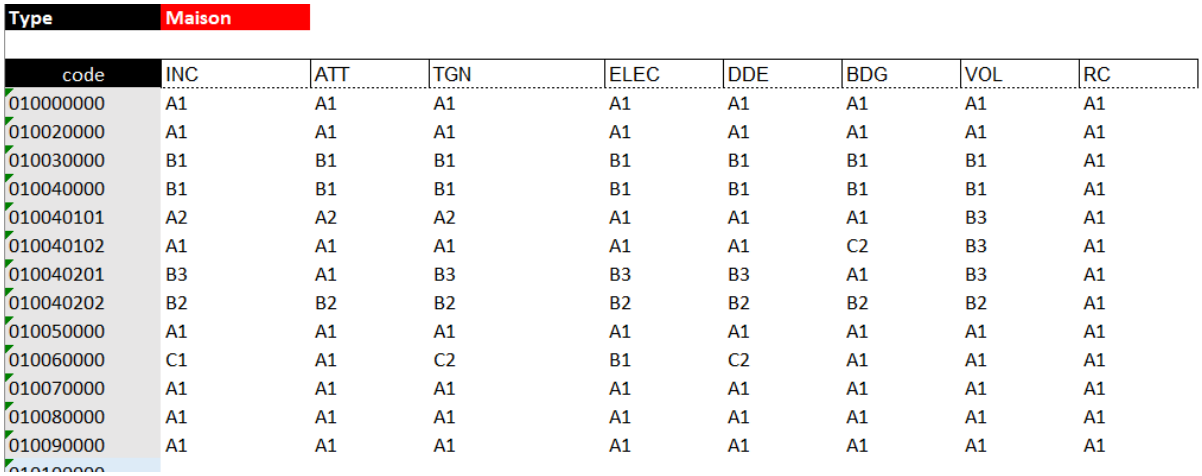
G	Dependence								
0		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
50		1,10	1,00	1,20	1,10	1,10	1,05	1,05	1,00
100		1,21	1,00	1,44	1,10	1,21	1,05	1,10	1,00
200		1,33	1,00	1,73	1,10	1,33	1,05	1,10	1,00
300		1,46	1,00	2,07	1,10	1,46	1,05	1,10	1,00
400		1,50	1,00	2,49	1,10	1,50	1,05	1,10	1,00
999		1,50	1,00	2,99	1,10	1,50	1,05	1,10	1,00

Etape I : surface cumulée des dépendance

I	=	G	x

I	NB enfant								
0		1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,500
1		1,000	1,000	1,000	1,000	1,000	1,300	1,000	1,000
2		1,000	1,000	1,000	1,000	1,000	1,400	1,000	1,500
3		1,000	1,000	1,000	1,000	1,000	1,500	1,000	2,000
4+		1,000	1,000	1,000	1,000	1,000	1,550	1,000	2,500

FIGURE E.2 – En fonction des garanties, les critères tarifaires ont un impact ou non. L'avantage de cette disposition est que l'impact des variables sont directement lisibles. En pratique, il y a un nombre plus conséquent de variables.



Type	Maison							
code	INC	ATT	TGN	ELEC	DDE	BDG	VOL	RC
010000000	A1	A1	A1	A1	A1	A1	A1	A1
010020000	A1	A1	A1	A1	A1	A1	A1	A1
010030000	B1	B1	B1	B1	B1	B1	B1	A1
010040000	B1	B1	B1	B1	B1	B1	B1	A1
010040101	A2	A2	A2	A1	A1	A1	B3	A1
010040102	A1	A1	A1	A1	A1	C2	B3	A1
010040201	B3	A1	B3	B3	B3	A1	B3	A1
010040202	B2	B2	B2	B2	B2	B2	B2	A1
010050000	A1	A1	A1	A1	A1	A1	A1	A1
010060000	C1	A1	C2	B1	C2	A1	A1	A1
010070000	A1	A1	A1	A1	A1	A1	A1	A1
010080000	A1	A1	A1	A1	A1	A1	A1	A1
010090000	A1	A1	A1	A1	A1	A1	A1	A1

FIGURE E.3 – Pour le zonier, un tableau exhaustif entre les communes et leur classe de risque est fait par garantie. Toutes les communes doivent être renseignées.

Références

[178] Benktander, G. and Segerdahl, C.-O. (1960). On the analytical representation of claim distributions with special reference to excess of loss reinsurance. In *Transactions of the international Congress of Actuaries*.

[179] Cebrian, A. C., Denuit, M., and Lambert, P. (2003). Generalized pareto fit to the society of actuaries' large claims database. *North American Actuarial Journal*, 7(3) :18–36.

[180] Commission, E. (2008). Briefing for the minister in charge of the earthquake commission. Briefing, Parliament New Zealand.

[181] Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society : Series B (Methodological)*, 52(3) :393–425.

[182] Dickson, P. G. M. (1960). *The Sun Insurance Office, 1710-1960 : The history of two and a half centuries of British insurance*. London, Oxford U. P.

[Donguy] Donguy, A. *Contribution of geographical information to insurance industry for the management of extreme events*. PhD thesis.

[184] Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events : for insurance and finance*, volume 33. Springer Science & Business Media.

[185] Frees, E. W., Meyers, G., and Cummings, A. D. (2014). Insurance ratemaking and a gini index. *Journal of Risk and Insurance*, 81(2) :335–366.

[186] Gerstengarbe, F. and Werner, P. (1989). A method for the statistical definition of extreme-value regions and their application to meteorological time series. *Zeitschrift fuer Meteorologie ;(German Democratic Republic)*, 39(4).

[187] Gerstengarbe, F.-W. and Werner, P. C. (1999). Estimation of the beginning and end of recurrent events within a climate regime. *Climate Research*, 11(2) :97–107.

[188] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174.

[189] Hunt Jr, F. J. (1962). Homeowners–the first decade. *Proceedings of the Casualty Actuarial Society*, 49(91) :12–40.

[190] Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70) :209–219.

[191] Rebuffat, F. (2003). Pret à la grosse aventure - source contre lakritos.

[192] Schechtman, E. and Schechtman, G. (2016). The relationship between gini methodology and the roc curve. *Available at SSRN 2739245*.

[193] Sneyers, R. (1990). On the statistical analysis of series of observations. wmo publ. 415, tech. Technical report, Note 143, 192 pp.

[194] Swift., T. P. (1950). Single insurance offered on homes : Fire, theft, liability and other hazards are covered in inclusive transaction new policy offers in a 'package.' coverage of most home risks. *New York Times*, page 131.

[195] Taubenheim, J. (1989). An easy procedure for detecting a discontinuity in a digital time series. *Zeitschrift für Meteorologie*, 39(6) :344–347.

[196] Thiveaud, J.-M. (1988). La naissance des assurances maritimes et colbert. *Revue d'économie financière*, 4 :151–156.

[197] Thomas, M. and Joy, A. T. (2006). *Elements of information theory*. Wiley-Interscience.

[198] Westbrook, R. (2003). *A History of Ancient Near Eastern Law (2 vols) : Volumes 1 and 2*. Brill.

[199] Wilk, M. B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1) :1–17.