UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ARBRES DE DÉCISION POUR LA FRÉQUENCE DANS LE CADRE DE LA MODÉLISATION DES RÉSERVES EN ASSURANCE NON-VIE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

ANDRA CRAINIC

OCTOBRE 2019

REMERCIEMENTS

Tout d'abord, je voudrais remercier mon directeur de recherche Mathieu Pigeon pour m'avoir soutenu tout au long de la maitrise. Je le remercie pour ses conseils, sa patience et sa disponibilité quasi-infinie qui m'ont grandement aidé dans la rédaction de ce mémoire. Je tiens à remercier également Co-operators pour m'avoir fourni les données sans lesquelles la rédaction de ce projet aurait été impossible. Je remercie mon conjoint de m'avoir encouragé à poursuivre mes études et de m'avoir soutenu tout au long de mon baccalauréat et de ma maitrise. Finalement, je tiens à remercier mes enfants Sandra et Maria pour avoir enjoué mon parcours.

TABLE DES MATIÈRES

LISTE DES TABLEAUX vii
LISTE DES FIGURES xi
RÉSUMÉ xiii
INTRODUCTION 1
CHAPITRE I MODÈLES COLLECTIFS D'ÉVALUATION DES RÉSERVES 5
1.1 Triangle de développement 6
1.2 Le modèle de Mack 8
1.3 Modèles linéaires généralisés 11
1.3.1 Famille exponentielle linéaire 11
1.3.2 Hétérogénéité et fonction de lien canonique 13
1.3.3 Inférence statistique 14
1.4 Modèles linéaires généralisés pour les modèles collectifs en réserves . . 15
CHAPITRE II ARBRE DE DÉCISION 19
2.1 Introduction 20
2.2 Divisions de l'espace des états 22
2.3 Élagage 28
2.4 Arbre de classification 29
2.5 Avantages et inconvénients 33
CHAPITRE III DEUX MODÈLES INDIVIDUELS POUR LES RÉSERVES 35
3.1 Structure générale 37
3.2 Approche paramétrique basée sur les modèles linéaires généralisés . . 38
3.2.1 Réserves RBNS 39

3.2.2 Réserves IBNR 43
3.3 Approche semi-paramétrique basée sur les arbres de régression 45
3.3.1 Réserves RBNS 48
3.3.2 Réserves IBNR 49
CHAPITRE IV ANALYSE NUMÉRIQUE 53
4.1 Données 53
4.2 Appplication des approches classiques 57
4.3 Application du modèle fréquence-sévérité individuel 59
4.3.1 Réserve RBNS 59
4.3.2 Réserve IBNR et réserve totale 62
4.4 Application des modèles non-paramétriques individuels 66
4.4.1 Réserve RBNS 66
4.4.2 Réserves IBNR 70
4.5 Comparaison des résultats 71
CONCLUSION 75
ANNEXE A TABLEAUX COMPLÉMENTAIRES AU CHAPITRE 4 . 77
RÉFÉRENCES 95

LISTE DES TABLEAUX

Tableau		Page
1.1	Triangle de développement incrémental.	6
1.2	Triangle de développement.	7
1.3	Triangle de développement cumulatif.	8
1.4	Triangle de développement complété.	10
1.5	Triangle de développement incrémental, où $J = 3$, transformé en base de données.	16
1.6	Base de donnés basée sur le tableau 1.2.	17
1.7	Coefficients du modèle GLM Poisson dans un contexte de réserve collective.	18
2.1	Base de données fictive.	22
3.1	Base de données fictive.	40
3.2	Triangle de développement.	41
3.3	Estimations des paramètres pour la fréquence.	41
3.4	Base de données fictive.	42
3.5	Estimations des paramètres pour la charge totale.	43
3.6	Base de données fictive du triangle inférieur de développement avec charge totale prédite.	44
3.7	Nombres de sinistres fictifs survenus à l'année i.	44
3.8	Fréquence prédite par l'arbre de décisions.	49
3.9	Estimations des paramètres pour la charge totale.	50
3.10	Base de données fictive du triangle inférieur de développement avec la charge totale prédite.	51

4.1 Réserve totale (RBNS + IBNR) prédite par les modèles collectifs. 59

4.2 Résultats des modèles individuels pour la réserve RBNS. 62

4.3 Proportion cumulative des sinistres fermés survenus en 2010. . . . 64

4.4 Estimation de la proportion cumulative des dossiers fermés à chaque année. 64

4.5 Résultats des modèles individuels pour la réserve IBNR. 65

4.6 La réserve totale (RBNS + IBNR) prédite par les modèles individuels. 66

4.7 Résultats des modèles individuels pour la réserve RBNS avec arbre de régression. 68

4.8 Résultats des modèles individuels pour la réserve IBNR avec arbres de régression. 70

4.9 La réserve totale (RBNS + IBNR) prédite par les modèles individuels avec arbre de régression. 71

4.10 La réserve totale (RBNS + IBNR) prédite par tous les modèles individuels. 72

A.1 Les variables et leur catégories de la base de données. 78

A.2 Les variables et leur catégories de la base de données (suite). . . . 79

A.3 Les variables et leur catégories de la base de données (suite). . . . 80

A.4 Statistiques descriptives des données en milliers de dollars. 81

A.5 Statistiques descriptives des données en milliers de dollars (suite). 82

A.6 Les coefficients et leur écart-type de la fréquence **modèle G1**. . . 83

A.7 Les coefficients et leur écart-type de la sévérité **modèle G1**. . . . 84

A.8 Les coefficients et leur écart-type de la fréquence **modèle G2**. . . 85

A.9 Les coefficients et leur écart-type de la sévérité **modèle G2**. . . . 86

A.10 Les coefficients et leur écart-type de la fréquence **modèle G3**. . . 87

A.11 Les coefficients et leur écart-type de la sévérité **modèle G3**. . . . 88

A.12 Les coefficients et leur écart-type de la fréquence **modèle G4**. . . 89

A.13 Les coefficients et leur écart-type de la sévérité **modèle G4**. . . . 90

A.14 Les coefficients et leur écart-type de la sévérité **modèle A1**. . . . 91

A.15 Les coefficients et leur écart-type de la sévérité **modèle A2**. . . . 92

A.16 Les coefficients et leur écart-type de la sévérité **modèle A3**. . . . 93

LISTE DES FIGURES

Figure		Page
0.1	*Développement habituel d'un sinistre en assurance IARD.*	1
2.1	*Espace des états divisé en trois régions.*	20
2.2	*Arbre de régression estimant le logarithme du salaire des joueurs de baseball.*	21
2.3	*Espace initial des variables explicatives.*	23
2.4	*Espace des variables explicatives.*	24
2.5	*Espace des variables explicatives à l'étape 0.*	24
2.6	*Espace des variables explicatives à l'étape 1.*	25
2.7	*Espace des variables explicatives.*	26
2.8	*Espace des variables explicatives à l'étape 2.*	26
2.9	*Espace des variables explicatives final.*	27
2.10	*Arbre de régression avec la première coupure.*	27
2.11	*Arbre de régression final.*	28
2.12	*Espace des variables explicatives.*	30
2.13	*Espace des variables explicatives avec une première coupure potentielle.*	31
2.14	*Espace des variables explicatives avec une première coupure potentielle.*	32
3.1	*Modèle saturé.*	47
3.2	*Arbre de décisions.*	48
4.1	*Le nombre de sinistres survenus entre les années 2005 et 2012.*	54

4.2 *Le nombre de sinistres déclarés entre les années 2005 et 2012.* . . 55

4.3 *Le nombre de réclamations par province.* 56

4.4 *Le nombre de réclamations ayant différents contrats.* 57

4.5 *Le nombre de réclamations se trouvant dans différentes années de développement.* 58

4.6 *La distribution prédictive du modèle de Mack et du modèle quasi-Poisson.* 60

4.7 *La distribution prédictive des quatre modèles GLM.* 63

4.8 *Graphique de l'erreur en fonction du paramètre de complexité.* . . 67

4.9 *Distribution prédictive des trois modèles.* 69

RÉSUMÉ

Les compagnies d'assurance non-vie doivent constamment s'assurer que les obligations financières sont respectées. Pour ce faire, les actuaires doivent vérifier régulièrement s'il y a suffisamment de liquidités dans la réserve. Basé sur la granularité de la base de données, presque tous les modèles peuvent être divisés en deux grandes catégories : individuel et collectif. Dans ce projet, on a décidé d'utiliser l'approche individuelle pour ses nombreux avantages malgré qu'elle soit rarement utilisée en industrie. On pose que la réserve suit une structure fréquence-sévérité où la fréquence est modélisée avec la méthode d'apprentissage statistique *arbres de décision* en ligne avec des publications récentes comme, par exemple, (Wuthrich, 2018). La sévérité est modélisée à l'aide des modèles linéaires généralisés pour réserves individuelles. On présente les résultats d'une analyse empirique utilisant un portfolio réel d'un grand assureur Canadien tout en quantifiant le risque associé à la réserve.

MOTS CLÉS : Réserves individuelles, Réserve RBNS, Réserve IBNR Modèles linéaires généralisés, Arbres de décisions, Arbres de régression, Modèle fréquence-sévérité

INTRODUCTION

Toute compagnie d'assurance doit mettre de l'argent de côté pour être en mesure de respecter ses engagements. C'est, bien sûr, le cas des assureurs non-vie aussi. Une partie de la prime collectée est mise dans un compte qui s'appelle *réserve.* Il y a plusieurs façons de calculer la réserve en assurance non-vie, mais pour le bon fonctionnement des compagnies, plusieurs lois encadrent ce calcul. À chaque année, le gouvernement vérifie la réserve de façon très rigoureuse pour protéger la société d'une éventuelle faillite.

Lorsqu'un accident de voiture arrive, il y a plusieurs étapes avant que celui-ci soit réglé. La figure 0.1 montre le développement habituel d'un sinistre.

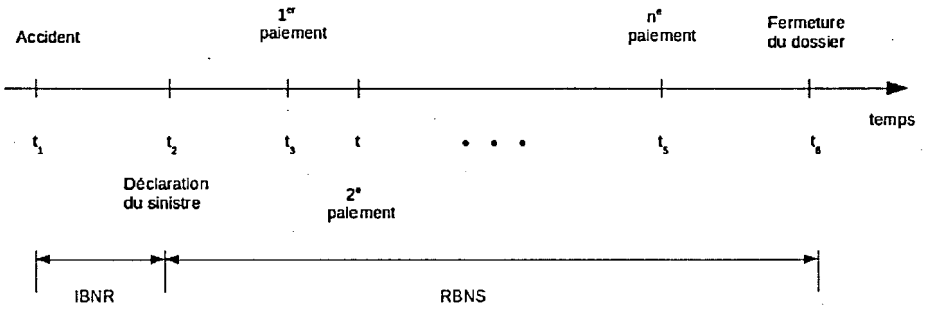

Figure 0.1: Développement habituel d'un sinistre en assurance IARD.

Au temps (t_1) il y a la survenance de l'accident. Le dossier qui se trouve entre la survenance du sinistre (t_1) et sa déclaration à l'assureur se classifie comme étant IBNR (*incurred but not reported*). Une fois le sinistre rapporté à l'assureur, on

arrive à l'étape RBNS (*reported but not settled*) et elle dure jusqu'à la fermeture du dossier.

Exemple 0.0.1. *Pour bien expliquer ces étapes, on prend un exemple fictif d'un assuré qui subit un sinistre. On suppose que Monsieur Sébastien a un accident de voiture le 9 janvier 2010 et ne le déclare à son assureur que le 23 juillet 2011. Entre les deux dates, le sinistre est nommé IBNR. Après le 23 juillet 2011, le sinistre est reclassifié dans la catégorie RBNS jusqu'au moment de sa fermeture.*

Deux grandes approches sont utilisées pour calculer les réserves : l'approche collective et l'approche individuelle. L'approche collective est utilisée depuis plus d'un siècle et elle consiste à agréger, au sein d'un portefeuille, les paiements faits pendant une période de temps. La réserve est ensuite calculée pour le portefeuille au complet. Plusieurs méthodes ont été développées au fil du temps dont l'algorithme *Chain-Ladder* (Mack, 1993) qui est fort connu en assurance non-vie. L'approche individuelle est beaucoup plus récente et est encore peu développée. Elle consiste à calculer une réserve pour chaque dossier dans le portefeuille d'assurance non-vie. Ainsi, la réserve totale est la somme de la réserve calculée pour chaque dossier.

Ce mémoire propose d'analyser des méthodes d'évaluation de réserve selon l'approche individuelle en séparant la modélisation de la fréquence et de la sévérité. En particulier, on veut utiliser des méthodes d'apprentissage statistique comme les arbres de décision pour estimer la fréquence et des méthodes d'estimations classiques pour estimer la sévérité. Mario Wuthrich a proposé, pour la première fois, la modélisation de la fréquence à l'aide des arbres de décision dans l'article (Wuthrich, 2018). Ce mémoire présente, pour la première fois, une méthodologie complète de l'application des arbres de décision lors d'évaluation des réserves non-vies.

Dans le premier chapitre, on explique les méthodes classiques d'évaluation de

réserve selon les approches collectives et individuelles. Le deuxième chapitre comporte une présentation détaillée de la méthode d'apprentissage statistique utilisée dans le modèle. Au troisième chapitre, on présente les modèles proposés pour l'évaluation des réserves. Au quatrième chapitre, on présente les résultats et, enfin, une brève conclusion termine ce document.

CHAPITRE I

MODÈLES COLLECTIFS D'ÉVALUATION DES RÉSERVES

Les réserves en assurance non-vie ont été évaluées bien avant le développement des modèles mathématiques sophistiqués utilisés depuis peu. Avant ces méthodes mathématiques, les actuaires utilisaient une méthode déterministe, une règle du pouce, pour évaluer les réserves. Cette méthode s'appelle *Chain Ladder*. Ce n'est que récemment que (Mack, 1993) a proposé une version stochastique de ce modèle. Des améliorations ont été apportées, quelques années plus tard, par (Mack, 1999).

Les modèles linéaires généralisés (GLM) ont été proposés pour la première fois par (Nelder et Wedderburn, 1972). Comme le nom l'indique, ces modèles sont une généralisation des modèles linéaires et offrent davantage de flexibilité. Il est donc possible d'utiliser les GLM pour modéliser des variables qui n'ont pas nécessairement une distribution Normale. Par contre, il arrive parfois que le modèle GLM n'offre pas assez de flexibilité. Alors, (Wedderburn, 1974) a proposé le modèle des Quasi-lois. Ce dernier modèle est capable de s'ajuster aux données davantage que le modèle GLM. Ainsi, il n'a pas besoin d'hypothèse forte pour la distribution de la variable réponse. La présentation de ce chapitre est une adaptation libre des notes de cours de Mathieu Pigeon et de Jean-Philippe Boucher.

1.1 Triangle de développement

Le modèle collectif d'évaluation des réserves utilise un outil qui aide à la visualisation des paiements des sinistres. Cet outil s'appelle le triangle de développement. Il consiste en un tableau qui contient la somme payée par la compagnie à chaque année comme dans le tableau 1.1. On définit la variable aléatoire $Y_{i,j}$ comme étant

Tableau 1.1: Triangle de développement incrémental.

Année	**1**	**2**	**3**
i	$Y_{i,1}$	$Y_{i,2}$	$Y_{i,3}$
i+1	$Y_{i+1,1}$	$Y_{i+1,2}$	
i+2	$Y_{i+2,1}$		

la somme payée à l'année j, où $j = 1, 2, \ldots, J$, pour les sinistres encourus à l'année i, où $i = 1, 2, \ldots, I$ et $I = J$. Comme il est possible que le règlement d'un sinistre s'étende sur plusieurs années, la variable $Y_{i,j}$ est le paiement fait j années après la survenance.

Le triangle de développement peut aussi contenir de l'information sur les montants cumulatifs payés pour les sinistres. On pose la variable $C_{i,j}$ comme étant le montant cumulatif payé jusqu'à l'année j pour les sinistres encourus à l'année i. Donc,

$$C_{i,j} = Y_{i,1} + Y_{i,2} + \cdots + Y_{i,j-1} + Y_{i,j}, \quad i = 1, \ldots, J, \quad j = 1, \ldots, J$$
$$C_{i,1} = Y_{i,1}.$$

Afin de simplifier la notation, on va désigner $\mathcal{D}_J$ comme étant un triangle de développement ayant comme nombre d'années de survenance et de développement maximal J. De plus, on va définir $\mathcal{D}_J^{inf}$ le triangle inférieur de développement ayant comme années de survenance et de développement maximal J.

Exemple 1.1.1. *Le tableau 1.2 est un exemple de triangle de développement avec nombres fictifs. Les années de survenance des accidents se trouvent sur les lignes,*

Tableau 1.2: Triangle de développement.

Année	1	2	3	4
1 (2010)	2650	250	300	40
2 (2011)	2800	500	100	
3 (2012)	3100	350		
4 (2013)	3900			

soient 2010, 2011, 2012 et 2013. Les années de développement se trouvent sur les colonnes 1, 2, 3 et 4. Donc, la somme totale que la compagnie a payée en 2010 pour les accidents survenus en 2010 est de 2650$. Un an plus tard, la compagnie a payé 250$ pour les sinistres survenus en 2010.

Il arrive fréquemment qu'un paiement soit fait une ou plusieurs années après la survenance de l'accident. Cette situation peut être due à des dommages causés par l'accident et observables seulement quelques années plus tard. Dans l'exemple 1.1.1, on observe que l'entreprise a payé 300$ en 2012 pour les sinistres survenus en 2010.

Pour connaitre le montant total qui a été payé dans une année calendaire, on calcule la somme des paiements sur la diagonale. Dans l'exemple 1.1.1, si on cherche le montant total payé pendant l'année 2012, on fait le calcul suivant : $3100 + 500 + 300 = 3900$. Le triangle de développement présenté dans le tableau 1.2 est appelé triangle incrémental parce que les montants inscrits dans chaque cellule sont les montants qui ont été payés à chaque année ($Y_{i,j}$). Le triangle cumulatif est un triangle de développement qui contient des montants cumulés ($C_{i,j}$) comme dans le tableau 1.3.

Tableau 1.3: Triangle de développement cumulatif.

Année	1	2	3	4
1 (2010)	2650	2900	3200	3240
2 (2011)	2800	3300	3400	
3 (2012)	3100	3450		
4 (2013)	3900			

Les espaces vides, expriment le fait qu'il n'y a pas encore d'information disponible. La tâche est alors de prédire ces montants inconnus pour ensuite être en mesure de calculer le montant de la réserve.

1.2 Le modèle de Mack

Comme il a été mentionné plus haut, la méthode d'évaluation de réserve la plus utilisée aujourd'hui est l'algorithme Chain-Ladder ou plutôt, la version stochastique de celui-ci proposée par Thomas Mack (Mack, 1993). Les valeurs prédites de la réserve sont les mêmes que celles obtenues par l'algorithme Chain-Ladder, mais une évaluation de la variabilité des paiements est proposée.

Implicitement, le modèle de Mack fait l'hypothèse qu'à la fin de la dernière année de développement, tous les dossiers sont fermés. Les hypothèses du modèle sont :

— indépendance des sommes des sinistres des années de survenance

$$\{C_{i,1}, C_{i,2}, \ldots, C_{i,J}\} \perp \{C_{l,1}, C_{l,2}, \ldots, C_{l,J}\} \quad \text{où} \quad i \neq l,$$

— $\mathrm{E}\big[C_{i,(j+1)} | C_{i,1}, C_{i,2}, \ldots, C_{i,j}\big] = C_{i,j}\lambda_j$

— $\mathrm{Var}\big[C_{i,(j+1)} | C_{i,1}, C_{i,2}, \ldots, C_{i,j}\big] = \sigma_j^2 C_{i,j}$, où $J = 1, \ldots, J-1$ et $i = 1, \ldots, J$

et les estimateurs sont

— $\hat{\lambda}_j = \frac{\sum_{i=1}^{J-j} C_{i,(j+1)}}{\sum_{i=1}^{J-j} C_{i,j}}$

— $\hat{\sigma}_j^2 = \frac{1}{J-j-1} \sum_{i=1}^{J-j} C_{i,j} \left(\frac{C_{i,(j+1)}}{C_{i,j}} - \hat{\lambda}_j \right)^2$

— $\hat{\sigma}_{J-1}^2 = \min\left[\frac{\hat{\sigma}_{J-2}^4}{\hat{\sigma}_{J-3}^2}, \min(\hat{\sigma}_{J-3}^2, \hat{\sigma}_{J-2}^2) \right]$.

Il peut être facilement prouvé que $\hat{\lambda}_j$ et $\hat{\sigma}_j^2$ sont des estimateurs sans biais. On peut trouver les preuves dans (Mack, 1993).

Exemple 1.2.1. *On reprend l'exemple 1.1.1 dont les données sont présentées dans le tableau 1.3. On rappelle que pour cette méthode, le triangle de développement doit obligatoirement être cumulatif. Les facteurs de développement sont donnés par*

$$\hat{\lambda}_1 = \frac{2900 + 3300 + 3450}{2650 + 2800 + 3100} = 1.128655$$
$$\hat{\lambda}_2 = \frac{3200 + 3400}{2900 + 3300} = 1.064516$$
$$\hat{\lambda}_3 = \frac{3240}{3200} = 1.0125.$$

Par après, on prédit les montants manquants dans le triangle de développement comme suit :

$$\hat{C}_{2,3} = \hat{\lambda}_3 C_{2,2} = 1.0125 \times 3400 = 3442.50$$
$$\hat{C}_{3,2} = \hat{\lambda}_2 C_{3,1} = 1.06516 \times 3450 = 3672.58$$
$$\hat{C}_{4,3} = \hat{\lambda}_1 \hat{C}_{4,2} = 1.0125 \times 3674.80 = 3718.49$$

et ainsi de suite.

Le tableau 1.4 est le triangle de développement complété. Les montants en italique sont les montants cumulés calculés précédemment.

Pour trouver la réserve de chaque année, il suffit de soustraire les montants sur la diagonale, soit les plus récents montants cumulatifs connus, des montants cumulatifs à l'ultime (la dernière colonne). Pour connaitre la réserve finale, il faut

Tableau 1.4: Triangle de développement complété.

Année	1	2	3	4
1 (2010)	2650	2900	3200	3240
2 (2011)	2800	3300	3400	*3442.50*
3 (2012)	3100	3450	*3672.58*	*3718.49*
4 (2013)	3900	*4401.75*	*4685.74*	*4744.31*

additionner tous les montants des réserves annuelles :

$$\hat{R}_i = C_{i,J} - C_{J-i}, \quad i = 2, \ldots, J$$
$$\hat{R} = \sum_{i=2}^{J} \hat{R}_i.$$

Exemple 1.2.2. *Pour illustrer le concept, on poursuit l'exemple 1.2.1. On cherche la réserve avec les formules présentées plus haut :*

$$\hat{R}_2 = \hat{C}_{2011,3} - C_{2011,2} = 3442.50 - 3400 = 42.5$$
$$\hat{R}_3 = 268.49$$
$$\hat{R}_4 = 844.31$$
$$\hat{R} = 42.50 + 268.49 + 844.31 = 1155.3.$$

On calcule la variance des paiements à l'ultime, c'est-à-dire les paiements sur la dernière colonne du triangle de développement complété comme suit

$$\hat{\sigma}_1^2 = \frac{1}{4-1-1}\left[2650\left(\frac{2900}{2650} - 1.128655\right)^2 + 2800\left(\frac{3300}{2800} - 1.128655\right)^2 + 3100\left(\frac{3450}{3100} - 1.128655\right)^2\right]$$
$$= 3.622094$$

$$\hat{\sigma}_2^2 = 4.129167$$
$$\hat{\sigma}_3^2 = min\left[\frac{\hat{\sigma}_1^4}{\hat{\sigma}_2^2}, min(\hat{\sigma}_1^2, \hat{\sigma}_2^2)\right] = 3.177290.$$

L'avantage le plus important de la méthode de Mack est que le calcul de la réserve est simple et facile à implémenter. De plus, le fait qu'il s'agisse d'un modèle stochastique, permet le calcul de la variance des paiements et ouvre la porte à des méthodes de rééchantillonnage permettant d'obtenir la distribution prédictive de la réserve (Charpentier et Pigeon, 2016). Ainsi, les compagnies d'assurances sont en mesure de calculer un surplus de sécurité pour éviter l'insolvabilité. Le principal désavantage du modèle de Mack est que la méthode est heuristique, c'est-à-dire que la méthode estime les paramètres rapidement de façon non-optimale. Les compagnies d'assurances préfèrent, généralement, les modèles optimaux pour un calcul de variance précis. De plus, le modèle assume que le développement d'une année à une autre sont similaires. Enfin, l'inflation n'est pas directement prise en compte par le modèle.

1.3 Modèles linéaires généralisés

1.3.1 Famille exponentielle linéaire

Lorsque plusieurs distributions ont la même forme, on peut les classer dans une famille et ainsi dégager des propriétés générales. La famille exponentielle linéaire (FEL) regroupe les distributions dont la fonction de densité de probabilité peut s'écrire sous la forme suivante :

$$f_Y(y) = c(y, \phi)\exp\left(\frac{y\theta - a(\theta)}{\phi}\right). \tag{1.1}$$

Le paramètre θ est appelé le paramètre canonique et le paramètre ϕ est appelé le paramètre de dispersion. La fonction $c(y, \phi)$ est une mesure de base et la fonction

$a(\theta)$ est appelée la logpartition.

Il peut être démontré (Nelder et Wedderburn, 1972) que :

$$\mathrm{E}[Y] = a'(\theta) = \mu \quad \text{et}$$
$$\mathrm{Var}[Y] = \phi a''(\theta).$$

Dans la FEL, on retrouve, entre autres, la Poisson, la Bernoulli, la binomiale, la gamma, la normale, l'inverse-gaussienne, etc.

Exemple 1.3.1. *On propose en exemple de trouver les paramètres θ et ϕ et les fonction $a(\theta)$ et $c(y, \phi)$ de la distribution Poisson parce que cette dernière est fréquemment utilisée en assurance non-vie.*

Pour une variable aléatoire dont la distribution est membre de la FEL, on peut réécrire la fonction de densité de la façon suivante :

$$ln(f_Y(y)) = ln(c(y, \phi)) + \left(\frac{y\theta - a(\theta)}{\phi}\right).$$

La distribution de Poisson se réécrit de façon similaire :

$$P[Y = y] = \begin{cases} \frac{\lambda^y e^{-\lambda}}{y!}, & y = 0, 1, \ldots \\ 0, & ailleurs \end{cases}$$
$$ln(P[Y = y]) = -ln(y!) + yln(\lambda) - \lambda.$$

On voit directement que :

$$ln(c(y, \theta)) = -ln(y!) \implies c(y, \phi) = \begin{cases} \frac{1}{y!}, & y = 0, 1, \ldots \\ 0, & ailleurs \end{cases}$$
$$\theta = ln(\lambda) \implies \lambda = exp(\theta)$$
$$\phi = 1.$$

1.3.2 Hétérogénéité et fonction de lien canonique

On suppose une variable aléatoire Y dont la distribution fait partie de la FEL. Elle a alors la forme décrite par l'équation (1.1). Il est possible d'ajouter de l'hétérogénéité en posant

$$g(\mathrm{E}[Y_i]) = g(\mu_i) = g(a'(\theta_i)) = \boldsymbol{X}_i'\boldsymbol{\beta}, \tag{1.2}$$

où $\boldsymbol{X_i}$ est un vecteur regroupant les variables explicatives pour l'observation i.

La formule (1.2) a des indices i qui font référence au fait que pour chaque observation dans la base de données, le paramètre θ_i change de valeur en fonction des caractéristiques de l'assuré i. De plus, elle indique qu'une transformation g(.) de l'espérance est une fonction linéaire des variables explicatives. Cette dernière fonction $g(\mu_i)$ est appelée fonction de lien. Si $g(\mu_i) = \theta_i$ et $\theta_i = \boldsymbol{X}_i'\boldsymbol{\beta}$, $g(\mu_i)$ est appelé fonction canonique. En d'autre mots, c'est la fonction de lien qui est naturellement associée à la distribution choisie.

Exemple 1.3.2. *On considère la distribution Poisson pour illustrer le concept. Il a été expliqué plus tôt que le paramètre $\theta = ln(\lambda)$. Alors,*

$$\begin{aligned} a(\theta_i) &= \lambda_i \\ \mathrm{E}[Y] &= \mu_i = \lambda_i = exp(\theta_i) \\ g(\mu_i) &= ln(\mu_i) = \theta_i = \boldsymbol{X}_i'\boldsymbol{\beta}. \end{aligned}$$

La fonction canonique est $g(\mu_i) = ln(\mu_i)$ et est appelée lien logarithmique.

1.3.3 Inférence statistique

Pour estimer les k+1 paramètres du vecteur $\boldsymbol{\beta}$, on utilise, généralement, la méthode du maximum de vraisemblance. On a :

$$\begin{aligned} l(\boldsymbol{y};\boldsymbol{\beta},\phi) &= \sum_{i=1}^{n} \ln(f(y_i);\boldsymbol{\beta},\phi)) \\ &= \sum_{i=1}^{n} \left(\ln(c(y_i,\phi)) + \left(\frac{y_i\theta_i - a(\theta_i)}{\phi} \right) \right) \\ &= \frac{1}{\phi} \sum_{i=1}^{n} (y_i\theta_i - a(\theta_i)) + \sum_{i=1}^{n} \ln(c(y_i,\phi)). \end{aligned}$$

On maximise cette dernière fonction en prenant la dérivée première :

$$\begin{aligned} \frac{\partial l(\boldsymbol{y};\boldsymbol{\beta},\phi)}{\partial \beta_t} &= \sum_{i=1}^{n} \frac{\partial \ln(y_i;\boldsymbol{\beta},\phi)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_t} \\ &= \frac{1}{\phi}(y_i - a'(\theta_i)) \frac{\partial \theta_i}{\partial \beta_t}, \quad t = 0, 1, \ldots, k. \end{aligned}$$

Après quelques étapes, on arrive à l'expression suivante

$$\begin{aligned} \frac{\partial l(\boldsymbol{y};\boldsymbol{\beta},\phi)}{\partial \beta_t} &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{g'(\mu_i)a''(\theta_i)} x_{i,t}, \\ &\Longrightarrow \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{g'(\mu_i)a''(\theta_i)} x_{i,t} = 0, \quad t = 0, 1, \ldots, k \end{aligned}$$

qui permet d'obtenir l'estimateur des β_t, $t = 0, 1, \ldots, k$. Les estimateurs ont comme variance $\mathrm{Var}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{I}(\hat{\boldsymbol{\beta}})^{-1}$ où $\boldsymbol{I}(\hat{\boldsymbol{\beta}})$ est la matrice d'information de Fisher qui a la forme suivante :

$$\boldsymbol{I}(\hat{\boldsymbol{\beta}})_{t,j} = \sum_{i=1}^{n} \frac{1}{g(\mu_i)^2 \mathrm{Var}[Y_i]^2} x_{i,t} x_{i,j}, \quad t = 0, 1, \ldots, k, \quad j = 0, 1, \ldots, k.$$

Comme il a été mentionné plus tôt, l'espérance d'une fonction faisant partie de la FEL est $\mathrm{E}[Y] = a'(\theta)$ et la variance est $\mathrm{Var}[Y] = \phi a''(\theta)$. Les distributions membres de la FEL ont un paramètre ϕ fixe. Par exemple, $\phi = 1$ pour la loi de Poisson. En pratique, il est rarement vérifié que l'espérance et la variance sont

égales : on rencontre couramment, en assurance, des cas de surdispersion. Pour estimer le paramètre de dispersion à l'aide des données, on utilise le paramètre de Pearson

$$\hat{\phi} = \frac{1}{n-z}\sum_{i=1}^{n}\frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

où $V(.)$ est la fonction de variance. Lorsque

— $\mathrm{E}[l'(\beta)] = 0$

— $\mathrm{Var}[l'(\beta)] = \boldsymbol{I}(\beta)$

— $-\mathrm{E}[l''(\beta)] = \boldsymbol{I}(\beta)$,

en appliquant la loi faible des grands nombres, lorsque $n \to \infty$, alors,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\beta})).$$

Ceci implique que l'estimation des coefficients $\hat{\boldsymbol{\beta}}$ tend vers la vraie valeur de $\boldsymbol{\beta}$ lorsque le nombre d'observations tend vers l'infini sans avoir besoin d'hypothèse forte pour la fonction de densité de la variable aléatoire Y. Cette approche est appelée quasi-lois.

1.4 Modèles linéaires généralisés pour les modèles collectifs en réserves

Puisque les modèles linéaires généralisés ont été introduits, on explique la façon que ceux-ci s'appliquent en réserve IARD en utilisant l'approche collective. Premièrement, on change le triangle de développement cumulatif pour le triangle de développement incrémental comme au tableau 1.1. Deuxièment, on transforme le triangle de développement en une base de données habituelle comme au tableau 1.5. La variable réponse est la variable aléatoire $Y_{i,j}$ et les variables explicatives sont i et j.

Comme dans le modèle de Mack (Mack, 1993), on fait l'hypothèse qu'à la fin de la dernière année de développement, tous les dossiers sont fermés.

Tableau 1.5: Triangle de développement incrémental, où $J = 3$, transformé en base de données.

Montant	i	j
$Y_{1,1}$	1	1
$Y_{1,2}$	1	2
$Y_{1,3}$	1	3
$Y_{2,1}$	2	1
$Y_{2,2}$	2	2
$Y_{3,1}$	3	1

Les hypothèses du modèle sont :

— indépendance entre les lignes et entre les colonnes du triangle de développement $Y_{i,j} \perp Y_{i',j'}$ où $i \neq i'$ et $j \neq j'$

— la variable aléatoire $Y_{i,j}$ doit faire partie de la FEL.

Puisqu'on utilise un modèle GLM, les relations suivantes sont toujours valables :

$$\mathrm{E}[Y_{i,j}] = \mu_{i,j} = a'(\theta_{i,j})$$
$$\mathrm{Var}[Y_{i,j}] = \phi_{i,j} a''(\theta_{i,j}).$$

Pour ajouter davantage structure au modèle, on considère une relation multiplicative pour l'espérance $\mathrm{E}[Y_{i,j}] = \alpha_i \psi_j$. Alors, si on utilise une fonction de lien logarithmique, on a

$$\ln(\mathrm{E}[Y_{i,j}]) = \ln(\alpha_i) + \ln(\psi_j).$$

À ce point, il est nécessaire d'ajouter une contrainte supplémentaire pour qu'il soit possible d'estimer les paramètres α_i et ψ_j. Cette contrainte peut être $\alpha_1 = 0$ ou $\psi_1 = 0$ ou $\sum_{j=1}^{J} \psi_j = 1$.

La prédiction de l'espérance de $Y_{i,j}$ est exprimée comme suit :

$$g\Big(\mathrm{E}\Big[\hat{Y}_{i,j}\Big]\Big) = \hat{\alpha}_1 \mathbb{1}_{\{1,1\}}(i,j) + \cdots + \hat{\psi}_J \mathbb{1}_{\{J,J\}}(i,j) \quad \text{où}$$

$$\mathbb{1}_{\{1,1\}}(i,j) = \begin{cases} 1, & i = 1 \quad \text{et} \quad j = 1 \\ 0, & \text{sinon.} \end{cases}$$

La réserve totale est la somme de toutes les prédictions :

$$\hat{R} = \sum_{(i,j)\in\mathcal{D}^{inf}} \hat{Y}_{i,j}.$$

Exemple 1.4.1. *Pour illustrer le concept, on considère le tableau 1.2. Pour commencer, on transforme le triangle de développement en une base de données comme au tableau 1.6. On prend comme contrainte* $\psi_1 = 0$. *Les coefficients sont présentés*

Tableau 1.6: Base de donnés basée sur le tableau 1.2.

Montant	i	j
2650	1	1
250	1	2
300	1	3
40	1	4
2800	2	1
500	2	2
100	2	3
3100	3	1
350	3	2
3900	4	1

dans le tableau 1.7. On prédit le montant pour le triangle inférieur du triangle de

Tableau 1.7: Coefficients du modèle GLM Poisson dans un contexte de réserve collective.

Coefficient	Estimation
$\hat{\alpha}_1$	7.88736
$\hat{\alpha}_2$	7.94798
$\hat{\alpha}_3$	8.02510
$\hat{\alpha}_4$	8.26873
$\hat{\psi}_2$	-2.05062
$\hat{\psi}_3$	-2.61981
$\hat{\psi}_4$	-4.19848

développement comme suit :

$$\begin{aligned}\hat{Y}_{i,j} &= exp(\hat{\alpha}_1 \mathbb{1}_{\{1,1\}}(i,j) + \cdots + \hat{\psi}_4 \mathbb{1}_{\{4,4\}}(i,j)) \\ \hat{Y}_{1,4} &= exp(7.88736 - 4.19848) \\ &= 40\end{aligned}$$

et ainsi de suite pour toutes les cellules du triangle de développement. La réserve totale est définie par la somme des $\hat{Y}_{i,j}$ *où* $(i,j) \in \mathcal{D}^{inf}$*, soit un montant de 1155.30.*

CHAPITRE II

ARBRE DE DÉCISION

Les arbres de décision ont été proposés pour la première fois par (Morgan et Sonquist, 1963) dans le contexte de la régression. Ils ont proposé une méthode de régression qui prend en considération les interactions des données automatiquement. L'objectif était atteint, mais plusieurs problèmes se sont présentés et ils ont été fortement critiqués par la communauté scientifique, en particulier parce que la méthode répliquait les données de façon très importante. Bien sûr, les prédictions étaient médiocres. De plus, lorsqu'il y avait plusieurs variables corrélées, les arbres de régression considéraient seulement une de ces variables dans leur structure. Peu de temps après, (Messenger et Mandell, 1972) ont repris l'idée initiale dans le contexte de la classification. Seulement une dizaine d'années plus tard, (Breiman *et al.*, 1984) se sont intéressés aux arbres de régression de nouveau pour proposer des améliorations importantes. Ils ont présenté la méthode *Classification And Regression Trees* (CART), un algorithme qui n'utilise plus de critère d'arrêt, mais plutôt qui construit un grand arbre pour utiliser ensuite l'élagage afin de l'optimiser. C'est ainsi que le problème de surajustement est corrigé. De plus, cet algorithme amélioré est capable de gérer les données manquantes naturellement.

2.1 Introduction

L'idée générale derrière les arbres de régression est de diviser l'espace des variables explicatives en plusieurs groupes mutuellement exclusifs, c'est-à-dire d'en réaliser une partition. Dans le cas de deux variables explicatives, cet espace est un rectangle où la longueur est représentée par la première variable explicative et la largeur est représentée par la deuxième variable explicative. L'algorithme cherche à former plusieurs rectangles de façon à minimiser une certaine fonction de perte. La fonction de perte est choisie par l'utilisateur mais, comme il a été mentionné plus tôt, il y a des fonctions de perte préférables à d'autres. En assurance IARD, souvent, la déviance Poisson est utilisée pour que l'arbre minimise la même fonction de perte que le modèle linéaire généralisé Poisson (GLM). Typiquement, l'espace des variables explicatives est segmenté comme dans la figure 2.1 extraite de (James *et al.*, 2013). Pour construire cette figure, les auteurs ont travaillé sur la base de données *Hitters* disponible dans la librairie *ISLR* du logiciel R avec la somme des résidus au carré comme fonction de perte.

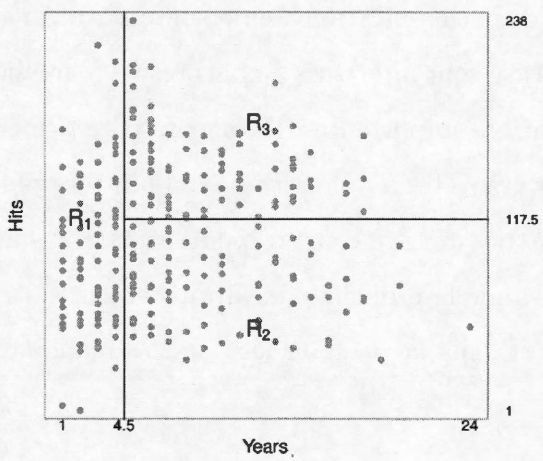

Figure 2.1: Espace des états divisé en trois régions.

Le modèle essaie de prédire le salaire des joueurs de baseball avec les deux variables explicatives *Hits* et *Years*. Sur la figure, on peut voir facilement que les joueurs de baseball ayant plus de 4.5 années d'expérience et moins de 1175 coups sûrs ont un salaire différent des joueurs ayant la même expérience mais un nombre de coups sûrs plus élevé que 1175. L'estimation est basée sur la moyenne ou sur le mode des données dans chaque région. Dans ce cas-ci, un nouveau joueur de baseball ayant plus de 4.5 années d'expérience et moins de 1175 coups sûrs aura comme prédiction du salaire la moyenne de la région R_2. Une fois l'espace des états segmenté, on peut construire l'arbre de régression tel qu'illustré à la figure 2.2.

Figure 2.2: Arbre de régression estimant le logarithme du salaire des joueurs de baseball.

Cet arbre de régression contient deux noeuds et trois feuilles. Les noeuds se trouvent à chaque division interne et les feuilles sont équivalentes aux régions R_1, R_2 et R_3. Au bout des feuilles, on trouve l'estimation du salaire soit la moyenne des salaires des joueurs de baseball dans chaque catégorie.

2.2 Divisions de l'espace des états

Pour faire les divisions, au début, il faut choisir une fonction de perte. En théorie, il serait possible d'évaluer la fonction de perte pour toutes les divisions possibles. En pratique, cette méthode n'est pas faisable à cause du très grand nombre de possibilités de divisons et l'algorithme *recursive binary splitting* proposé par (Breiman *et al.*, 1984) est alors utilisé. Cet algorithme cherche à diviser l'espace des variables explicatives à l'aide d'une question binaire de type $X_j \leq c$ où X_j est continu ou $X_j \in c_l$ où X_j est une variable catégorielle avec l catégories. On le fait pour toutes les variables explicatives X_j pour finalement choisir la variable et le point qui minimise la fonction de perte. Après chaque division, l'algorithme répète les mêmes étapes pour chaque sous-division de l'espace des variables explicatives. Évidemment, cette façon de faire fonctionne très bien sur des bases de données contenant plus de deux variables explicatives. Les divisions créent des rectangles en plusieurs dimensions, ou hyper-rectangles, permettant de regrouper les observations. Dans chacun des hyper-rectangles j ainsi formés, une prédiction est obtenue en calculant $\sum_{i:X_i \in R_j} y_i/|R_j|$.

Exemple 2.2.1. *On considère la base de données fictive présentée dans le tableau 2.1 Il faut prédire la prime des assurés sachant l'âge et la valeur de leur voiture.*

Tableau 2.1: Base de données fictive.

i	Âge	Valeur voiture	Prime ($)
1	18	4000	50.75
2	19	8000	75.30
3	20	2000	30.83
4	21	10 000	100.00

Comme mentionné plus haut, il est possible de visualiser l'espace des états par

un rectangle comme à la figure 2.3. Pour fin de simplicité, dans cet exemple, la

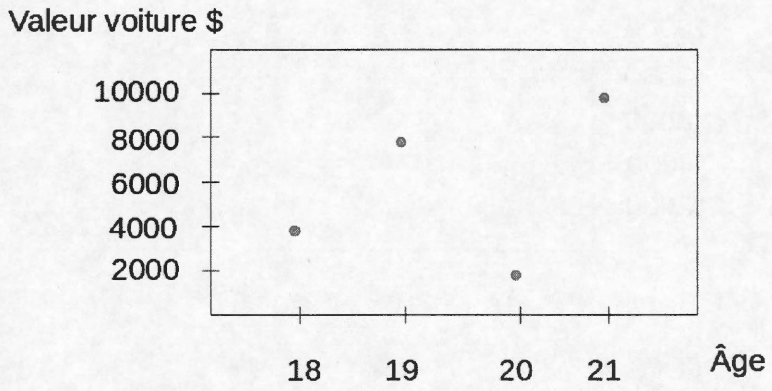

Figure 2.3: Espace initial des variables explicatives.

somme des résidus au carré est utilisée comme fonction de perte

$$RSS = \sum_{i=1}^{J} \sum_{i:x_i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

où $R_1, \ldots, R_J$ *sont les* J *régions de l'espace des états et* $\hat{y}_{R_j}$ *est la prédiction de la variable réponse* y_i *dans la région* R_j. *On commence par évaluer la fonction de perte pour la variable Âge aux points* $c_i = (18.5,\ 19.5,\ 20.5)$, *soit à distance égale entre les valeurs observées. Pour* $c_1 = 18.5$, *l'espace des états peut être vu comme dans la figure 2.4 où la linge pointillée est une division potentielle. On calcule la fonction de perte au point 18.5. La première région* (R_1) *est la région à gauche de la figure 2.4 et la deuxième région* (R_2) *est la région à droite. La prédiction au sein de* R_1 *est la moyenne des primes de l'assuré, soit* $\hat{y}_{R_1} = 50.75$. *Pour la région* R_2, *on prend la moyenne des primes des trois assurés, soit* $\hat{y}_{R_2} = 68.71$.

$$\begin{aligned} RSS &= \sum_{i=1}^{2} \sum_{i:x_i \in R_j} (y_i - \hat{y}_{R_j})^2 \\ &= (50.75 - 50.75)^2 + (75.35 - 68.71)^2 + (30.83 - 68.71)^2 + (100 - 68.71)^2 \\ &= 2457.3866. \end{aligned}$$

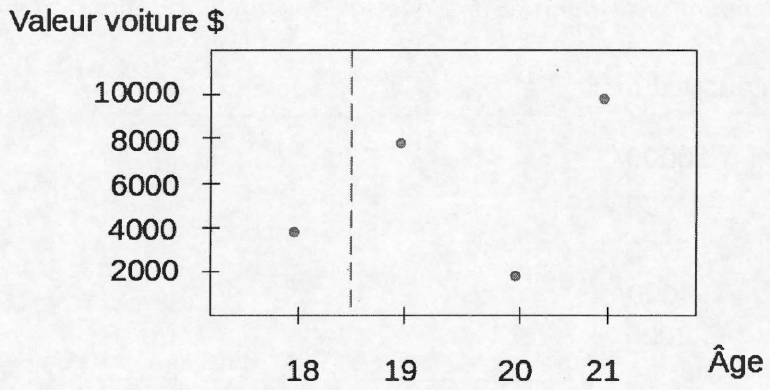

Figure 2.4: Espace des variables explicatives.

Pour $c_2 = 19.5$, *l'espace des états peut être vu comme dans la figure 2.5 où la linge pointillée est une deuxième proposition de division potentielle. On calcule la*

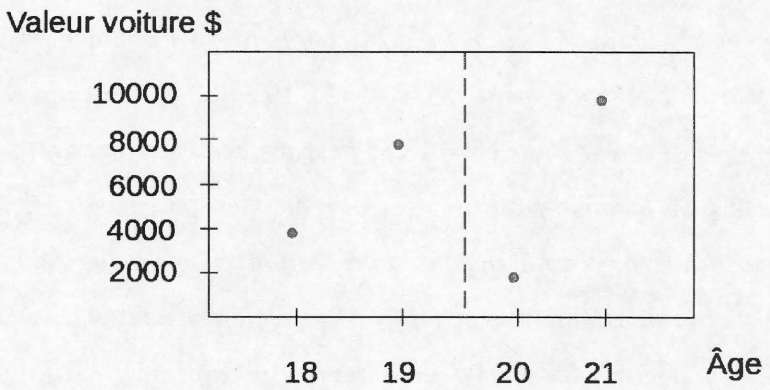

Figure 2.5: Espace des variables explicatives à l'étape 0.

fonction de perte au point 19.5. La prédiction au sein de R_1 *est* $\hat{y}_{R_1} = 63.025$ *et*

pour la région R_2 la prédiction est $\hat{y}_{R_2} = 65.415$. La fonction de perte vaut alors

$$RSS = \sum_{i=1}^{2} \sum_{i:x_i \in R_j} (y_i - \hat{y}_{R_j})^2$$
$$= (50.75 - 63.025)^2 + (75.30 - 63.025)^2 + (30.83 - 65.415)^2 + (100 - 65.415)^2$$
$$= 2690.1384.$$

On continue de la même façon pour $c_3 = 20.5$. La fonction de perte vaut alors 992.363. *On répète les mêmes étapes pour la variable explicative Valeur voiture aux points $v_i = (3000,\ 6000,\ 9000)$. Le RSS est calculé et les résultats sont respectivement* 1212.785, 503.4482 *et* 992.363. *Alors, la division finale sera sur la variable Valeur voiture au point $v_2 = 6000$, car c'est pour cette valeur que la fonction de perte est à son plus bas. Pour la deuxième itération de l'algorithme, on prend la*

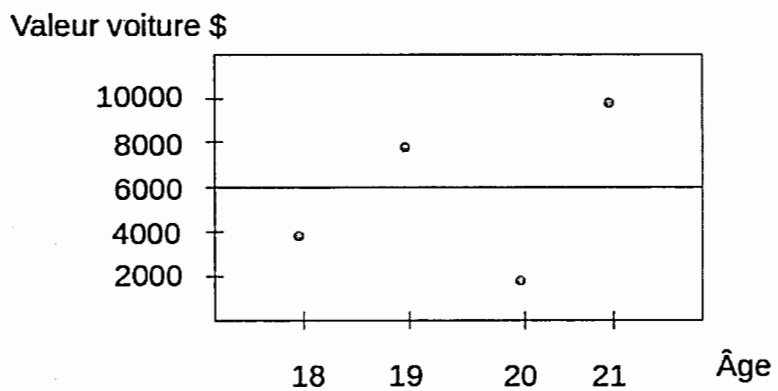

Figure 2.6: Espace des variables explicatives à l'étape 1.

région du bas de la figure 2.6 pour chercher une coupure possible. On commence avec la variable explicative Âge avec une coupure au point 19 ans comme dans la figure 2.7. On calcule le RSS et on obtient 305.045. *On calcule le RSS pour la variable Valeur voiture pour le point 3000 et on obtient* 305.045 *de nouveau. Une division amène clairement une amélioration à la fonction de perte. Puisque les deux coupures amènent la même amélioration, on va choisir de façon arbitraire une des deux, soit la variable Âge au point 19. Alors, le nouvel espace des*

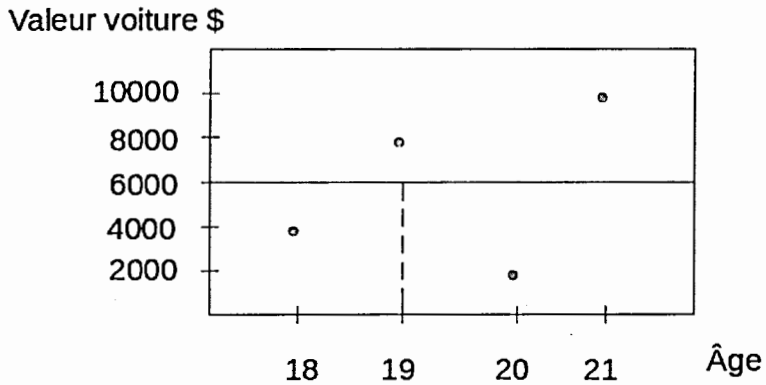

Figure 2.7: Espace des variables explicatives.

variables explicatives est montré dans la figure 2.8. L'algorithme recommence les

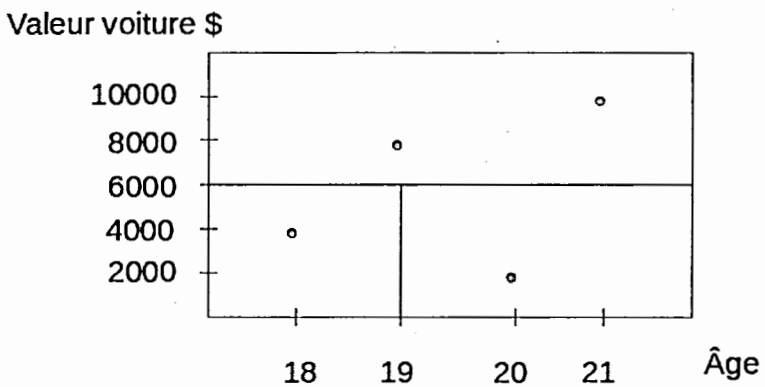

Figure 2.8: Espace des variables explicatives à l'étape 2.

mêmes étapes sur la région du haut de la figure 2.8 et calcule la fonction de perte pour la valeur 20 ans de la variable Âge et le point 9000 de la variable Valeur voiture. La situation est similaire avec la situation de la région en bas. Le RSS est 0 pour les deux dernières coupures potentielles et on choisit la coupure Âge au point 20 comme troisième division de l'espace des variables explicatives. Donc, l'espace divisé final est montré dans la figure 2.9.

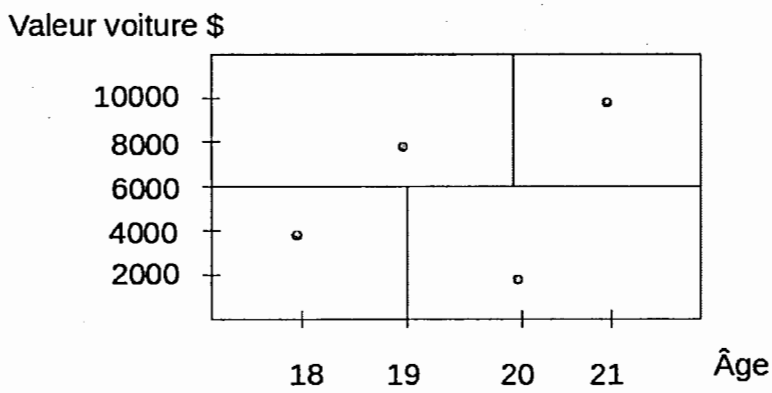

Figure 2.9: Espace des variables explicatives final.

Pour construire l'arbre de régression, il faut incorporer les coupures dans l'ordre dans lesquel elles se sont produites. Le noeud de départ représente la première coupure comme le montre la figure 3.1. Ensuite, il faut intégrer dans le graphique la deuxième et la troisième coupure comme dans la figure 2.11. À l'aide d'une simple

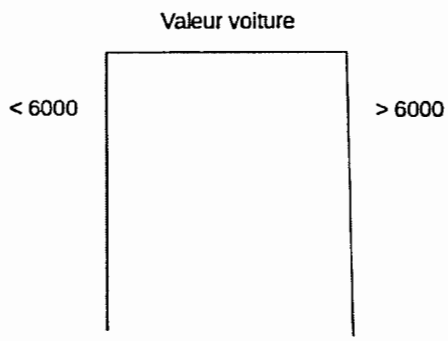

Figure 2.10: Arbre de régression avec la première coupure.

inspection visuelle il est facile d'interpréter les résultats. On voit rapidement qu'un assuré plus jeune de 19 ans possédant une voiture d'une valeur moindre de 6000$ va avoir une prime moyenne de 50.75 par mois.

Dans l'exemple présent, il n'est pas clair quand l'algorithme s'arrête, et ce, à cause

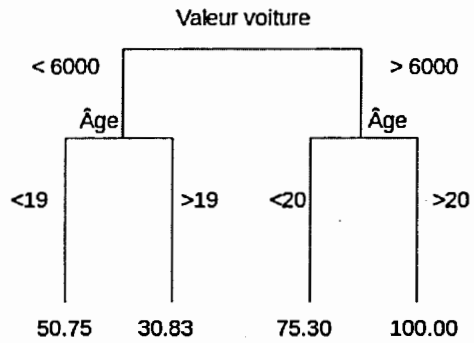

Figure 2.11: Arbre de régression final.

du nombre très petit des données. Normalement, il y a un critère d'arrêt comme le nombre minimum d'observations dans chaque rectangle R_i, le nombre de couches de noeuds, etc.

2.3 Élagage

Comme on pourrait s'y attendre, l'arbre de régression a produit du surajustement en répliquant les données parfaitement : on a quatre assurés dans la base de données et l'arbre a divisé l'espace des états en quatre parties. On se rend compte que le modèle n'a pas de biais mais il a une grande variance lors des prédictions ce qui est problématique. Cette difficulté est habituelle dans le cas des arbres de régression. Pour éviter ce problème, on utilise la méthode de (Breiman *et al.*, 1984) qui propose dans un premier temps de construire un très grand arbre où il est certain d'y avoir du surajustement. Par après, on enlève des branches qui diminuent le RSS de peu. Pour ce faire, la formule suivante est minimisée :

$$\sum_{i=1}^{n} L(y_i, f(\mathbf{x})) + c_p \times M, \ f(x_i) = \hat{y}_i$$

où $L(y_i, f(\mathbf{x}))$ est la fonction de perte, c_p est un paramètre de complexité, M est le nombre de feuilles et n est la taille de l'échantillon. Le paramètre c_p est choisi à

l'aide de la validation croisée. Sa valeur doit maximiser la performance du modèle sur une base de données de l'extérieur qui n'a pas contribué à la calibration du modèle. Lorsque $c_p = 0$, on obtient l'arbre initial.

La validation croisée est une méthode statistique qui va permettre l'estimation d'un paramètre clé d'un modèle. La méthode de validation croisée la plus utilisée s'appelle *K-Fold.* Elle consiste à diviser la base de données en K sous-bases. À chaque itération de l'algorithme, une section de la base de données est mise de côté. Le modèle est ajusté sur les sections restantes. Par après, la performance du modèle est évaluée à l'aide de la section mise de côté. Ainsi, on est capable de voir à quel point le modèle est capable de produire des estimations proches de la réalité. Une fois le modèle testé K fois (sur chaque section de la base de données), on prend la moyenne des performances calculées. Pour ajuster le paramètre c_p, on prend une multitude de valeurs possibles pour celui-ci. Pour chaque valeur de c_p, on applique la validation croisée avec un K typiquement égal à 10. Au final, on choisit la valeur du paramètre qui a permis au modèle de prédire le mieux la variable réponse. Il est vrai que cette méthode peut être couteuse en ressources informatiques si la base de données est grande. Par contre, l'avancée technologique des années récentes permet à la validation croisée d'être utilisée, car cette méthode est préférable aux méthodes classiques de sélection de modèles.

2.4 Arbre de classification

L'algorithme de l'arbre de classification est similaire à l'algorithme de l'arbre de régression. La différence la plus importante est due au fait que la variable réponse est qualitative. Il est donc impossible d'utiliser comme fonction de perte la déviance Poisson ou l'erreur quadratique moyenne parce que la variable réponse n'a pas de valeur numérique. Pour pallier ce problème, on va considérer des fonc-

tions de perte pour la classification comme le *Gini index* (Messenger et Mandell, 1972). Cette fonction de perte mesure la pureté de chaque noeud de l'arbre de classification comme suit :

$$G_r = \sum_{g \in \mathcal{G}} \hat{p}_{rg}(1 - \hat{p}_{rg}), \tag{2.1}$$

où $\hat{p}_{rg}$ est la proportion d'éléments dans la région r classés dans la catégorie $g \in \mathcal{G}$. On remarque que la fonction plus haut peut s'écrire de la façon suivante :

$$G_r = \sum_{g \in \mathcal{G}} \widehat{\mathrm{Var}}\left(\mathbb{1}_{\{\mathcal{G}=g\}}\right)$$

soit la variance d'une variable aléatoire suivant une loi de Bernoulli($\hat{p}_{rg}$). On voit aussi qu'une valeur proche de 0 ou 1 indique que la catégorie r contient un très grand nombre d'éléments classés dans une catégorie $g \in \mathcal{G}$ et très peu d'éléments classés dans les autres catégories.

Exemple 2.4.1. *On considère un espace des variables explicatives fictif présenté à la figure 2.12. On divise l'espace des variables avec l'algorithme recursive binary spliting. On voit qu'on a deux catégories : le bleu (B) et l'orange (O). La proportion*

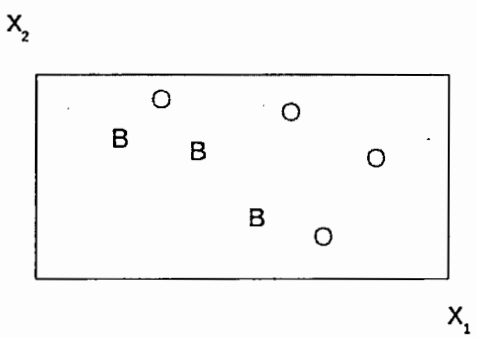

Figure 2.12: Espace des variables explicatives.

des points bleus dans la région 1 ($\hat{p}_{1B}$) est égal à 3/7. On appelle la base de données la région 1 parce qu'il n'y a pas de deuxième région pour l'instant. On garde cette

notation par soucis de cohérence. Alors, le Gini index calcule la pureté de la base de données comme suit :

$$\begin{aligned} G_1 &= \sum_{g\in\{B,O\}} \hat{p}_{1B}(1-\hat{p}_{1B}) + \hat{p}_{1O}(1-\hat{p}_{1O}) \\ &= \frac{3}{7}\times\frac{4}{7} + \frac{4}{7}\times\frac{3}{7} = \frac{24}{49}. \end{aligned}$$

Maintenant, l'algorithme prend une des deux variables explicatives, X_1 par exemple, et calcule le Gini index pour une coupure potentielle entre les deux points bleus comme à la figure 2.13. Alors, on calcule le Gini index pour la région 1, la partie

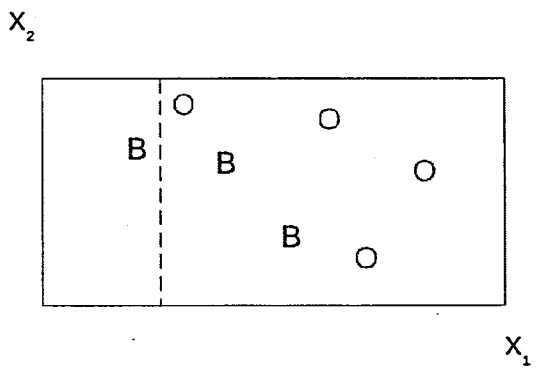

Figure 2.13: Espace des variables explicatives avec une première coupure potentielle.

gauche du rectangle, et la région 2, la partie droite du rectangle, comme suit :

$$\begin{aligned} \hat{p}_{1B} &= \frac{1}{1} = 0 \\ G_1 &= \sum_{g\in\{B,O\}} \hat{p}_{1B}(1-\hat{p}_{1B}) + \hat{p}_{1O}(1-\hat{p}_{1O}) \\ &= 1(1-1) + 0(1-0) = 0 \\ \hat{p}_{1B} &= \frac{2}{6} = \frac{1}{3} \\ G_2 &= \frac{1}{3}\times\frac{2}{3} + \frac{2}{3}\times\frac{1}{3} = \frac{4}{9}. \end{aligned}$$

On voit qu'une coupure entre ces deux premiers points est déjà avantageuse par rapport à ne pas avoir de coupure du tout parce que le Gini index a déjà diminué

pour les deux régions. On continue avec une coupure potentielle comme montré dans la figure 2.14.

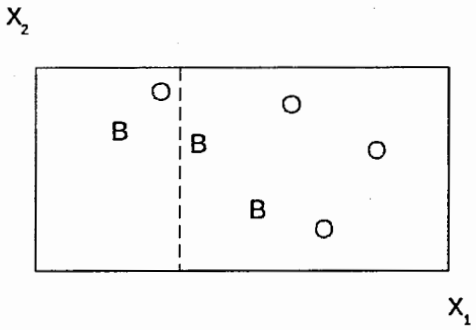

Figure 2.14: Espace des variables explicatives avec une première coupure potentielle.

Donc, le Gini index est calculé de la façon suivante :

$$\begin{aligned}
\hat{p}_{1B} &= \frac{1}{2} \\
G_1 &= \sum_{g\in\{B,O\}} \hat{p}_{1g}(1-\hat{p}_{1g}) = \frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2} \\
&= \frac{2}{4} = \frac{1}{2}, \\
\hat{p}_{2B} &= \frac{2}{5} \\
G_2 &= \hat{p}_{2B}(1-\hat{p}_{2B}) + \hat{p}_{2O}(1-\hat{p}_{2O}) \\
&= \frac{2}{5}\times\frac{3}{5}+\frac{3}{5}\times\frac{2}{5} = \frac{12}{25}.
\end{aligned}$$

On continue ainsi sur tous les points de la variable X_1 et on effectue les mêmes calculs avec la même approche sur la variable explicative X_2. On choisit la coupure qui génère le Gini index le plus petit. Une fois la première coupure choisie, l'algorithme cherche, de la même façon, les coupures subséquentes jusqu'à la satisfaction d'un critère d'arrêt comme dans les arbres de régression.

Il est possible d'utiliser d'autres fonctions de perte dans le cadre des arbres de classification comme la fonction d'entropie qui est beaucoup utilisée aussi ou la

fonction du taux d'erreur qui est moins populaire. La fonction d'entropie ressemble beaucoup au Gini index :

$$D_r = -\sum_{g \in \mathcal{G}} \hat{p}_{rg} \ln(\hat{p}_{rg}) \mathbb{1}_{(\hat{p}_{rg} \neq 0)}.$$

Pour ce qui est de la fonction du taux d'erreur, elle est moins utilisée parce que souvent les résultats sont difficiles à interpréter.

2.5 Avantages et inconvénients

Plusieurs désavantages se présentent lorsque l'arbre de décision est utilisé. Malgré qu'il soit possible de trouver un ordre d'importance parmi les variables explicatives, il est impossible de faire de l'inférence statistique, c'est-à-dire qu'il est impossible de trouver les variables explicatives significatives. Ceci empêche une interprétation du lien entre les variables explicatives et la variable réponse. De plus, l'arbre de décision manque de robustesse. Au moindre changement dans la base de données, l'arbre final pourrait être complètement différent. Pour contrer ce problème, on peut traiter les données aberrantes avant de commencer l'algorithme. Par contre, l'arbre de décision présente plusieurs avantages importants dans le domaine des réserves en assurance non-vie comme le fait que la méthode soit semi-paramétrique. Cet élément joue un rôle clé, car cette méthode ne présente pas de contrainte sur la distribution adjacente de la variable réponse. Ceci permet à la méthode de trouver facilement des liens très complexes entre les variables explicatives et la variable réponse. De plus, les bases de données ont une masse de probabilité à zéro très importante et l'arbre la gère naturellement. Aussi, cette technique d'apprentissage statistique s'occupe implicitement des interactions entre les variables explicatives et s'occupe facilement des données manquantes.

CHAPITRE III

DEUX MODÈLES INDIVIDUELS POUR LES RÉSERVES

Dans les chapitres précédents, on a introduit les méthodes et les outils mathématiques permettant la modélisation des différentes problématiques en réserves IARD. Dans ce chapitre, on propose le modèle mathématique encadrant les réserves individuelles. L'idée d'une modélisation individuelle des réserves a été introduite pour la première fois par (Arjas, 1989) et (Norberg, 1993). Ces auteurs ont suscité l'intérêt dans la communauté scientifique envers cette nouvelle méthode. Depuis, (Haastrup et Arjas, 1996), (Wüthrich, 2003), (Larsen *et al.*, 2007), (Zhao *et al.*, 2009), (Zhao et Zhou, 2010), (Antonio et Plat, 2013) et (Charpentier et Pigeon, 2016), pour n'en citer quelques uns, ont apporté leur contribution.

Les modèles qui sont présentés dans ce chapitre se basent sur une structure fréquence-sévérité, aussi appelée structure composée, qui a la forme suivante :

$$Y = \begin{cases} 0, & N = 0 \\ \sum_{i=1}^{N} X_i, & N > 0, \end{cases} \tag{3.1}$$

où Y représente le montant total d'un sinistre, N représente la fréquence de paiements et X_i représente le montant du paiement i où $i = 1, \ldots, N$. On suppose que la variable aléatoire N est indépendante de la variable Y et que les variables X_i sont indépendantes et identiquement distribuées.

Dans le cadre des réserves actuarielles, cette structure a été présentée pour la première fois dans (Norberg, 1993) et développée par la suite dans (Norberg, 1999). Dans cet article, les deux variables N et X sont modélisées à l'aide d'une distribution Poisson. Par la suite, plusieurs chercheuses et chercheurs ont repris le concept pour proposer différentes façons de modéliser N et X dont, en particulier, (Schiegl, 2002). Cet article propose un modèle où la fréquence (N) suit une loi Binomiale Négative et la sévérité (X) suit une loi Pareto. Un an plus tard, (Wüthrich, 2003) a proposé à son tour un modèle Tweedie où la fréquence suit une loi de Poisson et la sévérité suit une loi Gamma utilisant les années de survenance (i) et les années de développement (j) comme variables explicatives. En 2011, (Boucher et Davidov, 2011) utilisent un modèle Tweedie utilisant, cette fois, un double GLM (DGLM), c'est à dire qu'ils ont permis à l'espérance et à la variance de dépendre de la cellule (i,j). En 2017, (Denuit et Trufin, 2017) proposent un modèle où la fréquence est modélisée à l'aide de la distribution Poisson et la sévérité est modélisée par un mélange de distributions Gamma et Pareto. Ainsi, la distribution Gamma modélise la sévérité des paiements effectués dans les premières années de développement et la distribution Pareto modélise les paiements effectués pendant les dernières années de développement. Celles et ceux qui souhaitent compléter ce bref survol de l'utilisation de l'approche fréquence-sévérité dans la modélisation des réserves peuvent consulter (Wüthrich et Merz, 2008) et (England et Verrall, 2002).

3.1 Structure générale

Habituellement dans un modèle fréquence-sévérité, la structure du modèle est donnée par l'équation (3.1). Plus précisément, dans un cadre individuel, on aurait :

$$Y_{i,j}^{(k)} = \begin{cases} 0, & N_{i,j}^{(k)} = 0 \\ \sum_{l=1}^{N_{i,j}^{(k)}} X_{i,j,l}^{(k)}, & N_{i,j}^{(k)} > 0, \end{cases}$$

où $X_{i,j,l}^{(k)}$ représente le paiement l du sinistre k de la cellule (i,j) du triangle de développement, $N_{i,j}^{(k)}$ représente le nombre de paiements pour ce même sinistre k situé dans la cellule (i,j) et $Y_{i,j}^{(k)}$ représente la somme totale payée par la compagnie d'assurance pour le sinistre k de la cellule (i,j).

Plusieurs hypothèses sont utilisées pour construire le modèle. En premier lieu, on assume que les sinistres sont indépendants entre eux. En deuxième lieu, on suppose que les variables aléatoires $N_{i,j}^{(k)}$ et $X_{i,j,l}^{(k)}$ sont indépendantes. En troisième lieu, on suppose que $Y_{i,j}^{(k)}$ est indépendante de $Y_{m,t}^{(k)}$ pour tout $i \neq m$ et pour tout $j \neq t$. En quatrième lieu, on considère que les variables aléatoires $X_{i,j,l}$ sont indépendantes et identiquement distribuées.

Dans le cadre de ce projet, on modélise la variable aléatoire de la fréquence (N) et directement la variable aléatoire de la charge totale (Y). Malgré le fait que dans la base de données on trouvait l'information sur chaque paiement on a choisit de l'agréger et de modéliser la somme des paiements. Ce choix est motivé par le fait qu'il y a de nombreux sinistres qui ont beaucoup de paiements, par exemple, il y a un sinistre qui présente 298 paiements en une année. Il est évident que la prédiction du nombre de paiements individuels serait mauvaise. De plus, la base de données a une dimension très importante et en l'agrégeant, les modèles statistiques nécessitent moins de ressources informatiques. Même si le modèle proposé est différent du modèle fréquence-sévérité classique, les hypothèses énumérées plus

haut sont gardées.

Puisque dans ce travail la charge totale suit une distribution quasi-Poisson, l'espérance du montant total pour un sinistre k et une cellule (i,j) est donnée par (on retire la référence à k pour alléger la notation lorsqu'il n'y a pas de risque de confusion) :

$$\mathrm{E}[Y_{i,j}] = \lambda_{i,j}.$$

La variance est donnée par :

$$\mathrm{Var}[Y_{i,j}] = \psi\lambda_{i,j}$$

où ψ est le paramètre de dispersion.

3.2 Approche paramétrique basée sur les modèles linéaires généralisés

Pour modéliser la réserve, il faut diviser la modélisation en deux parties : la modélisation de la fréquence et la modélisation de la sévérité. Pour commencer, la fréquence $(N_{i,j}^{(k)})$ est modélisée par deux distributions : Poisson$(\lambda_{i,j}^{(k)})$ et Bernoulli$(p_{i,j}^{(k)})$. Dans les deux cas, la théorie des GLM permet d'inclure naturellement des variables explicatives dans la modélisation des paramètres, car les deux distributions sont membres de la FEL.

Pour la distribution de Poisson, on a

$$\lambda_{i,j} = e^{\boldsymbol{X}'_{i,j}\boldsymbol{\alpha}} \quad \text{et}$$

$$\mathrm{Var}[N_{i,j}] = \phi\lambda_{i,j} \quad \text{où} \quad \phi = 1,$$

où $\boldsymbol{\alpha}$ est le vecteur colonne contenant les paramètres du modèle et $\boldsymbol{X}_{i,j}$ est la matrice des variables explicatives. Pour la distribution Bernoulli, on a

$$\mathrm{Var}[N_{i,j}] = \phi p_{i,j}(1 - p_{i,j}) \quad \text{où}$$

$$\hat{p}_{i,j} = \frac{e^{\boldsymbol{X}'_{i,j}\hat{\boldsymbol{\alpha}}}}{1 + e^{\boldsymbol{X}'_{i,j}\hat{\boldsymbol{\alpha}}}} \quad \text{et} \quad \phi = 1.$$

Pour modéliser la charge totale par cellule ($Y_{i,j}^{(k)}$), on utilise une distribution Poisson sur-dispersée où

$$\begin{aligned} \mathrm{E}[Y_{i,j}] &= \lambda_{i,j} = \mathrm{E}[N]\, e^{\boldsymbol{X}_{i,j}'\boldsymbol{\beta}} \quad \text{et} \\ \mathrm{Var}[Y_{i,j}] &= \psi\lambda_{i,j}, \end{aligned} \tag{3.2}$$

où $\mathrm{E}[N]$ est une mesure de volume et l'estimation des paramètre est faite par le maximum de quasi-vraisemblance.

Habituellement, dans le cadre du modèle fréquence-sévérité comme décrit par l'équation (3.1), on modélise la fréquence (N) et les paiements ($X_{i,j}$) séparément. Dans le cadre de ce projet, plutôt que modéliser les paiements à venir, on modélise directement la somme des paiements $Y_{i,j}$ après que la variable fréquence ait déjà été prédite.

3.2.1 Réserves RBNS

La réserve RBNS de la cellule (i,j) est définie comme étant la somme des paiements effectués pour chaque sinistre ouvert dans la base de données :

$$R_{i,j}^{RBNS} = \sum_{k=1}^{K^{\mathrm{OBS}}} Y_{i,j}^{(k)},$$

où K^{OBS} est le nombre de sinistres observés dans la base de données.

La distribution de la variable aléatoire $Y_{i,j}^{(k)}$ a une masse de probabilité importante à zéro. Ceci est expliqué par le fait que la fréquence et la sévérité soient décomposées. Lorsque la fréquence est égale à zéro, la sévérité est automatiquement égale à zéro aussi.

Exemple 3.2.1. *On considère la base de données présentée au tableau 3.1 où on suppose que tous les dossiers sont fermés après 3 ans de développement. On suppose également que l'âge du conducteur est disponible comme variable explicative.*

Tableau 3.1: Base de données fictive.

No de sinistre	Année d'origine	Année de développement	Âge	Nombre de paiements	Montant total ($)
1	2005	1	25	1	250
1	2005	2	25	0	0
1	2005	3	25	0	0
2	2005	1	55	1	175
2	2005	2	55	1	50
2	2005	3	55	1	15
3	2006	1	85	3	1000
3	2006	2	85	2	2500
4	2006	1	43	1	500
4	2006	2	43	0	0
5	2007	1	40	1	100

Le tableau 3.2 présente le triangle de développement associé à la base de données.

La première étape consiste à calibrer le modèle GLM Poisson sur-dispersé pour modéliser la fréquence, soit la colonne «nombre de paiements». Dans cet exemple, on modélise la fréquence avec la distribution Poisson seulement pour illustrer les étapes à suivre. Dans le cas où la fréquence est modélisée à l'aide de la distribution Bernoulli, le même cheminement est suivi mais pour une variable réponse artificielle valant 0 si le nombre de paiements est égal à 0 et 1 sinon. Le tableau 3.3 présente les paramètres estimés par le modèle quasi-Poisson. Pour faire la prédiction de la fréquence moyenne, on utilise les estimations des paramètres comme

Tableau 3.2: Triangle de développement.

Année	1	2	3
2005	425	50	15
2006	1500	2500	
2007	100		

Tableau 3.3: Estimations des paramètres pour la fréquence.

Coefficients	Estimation
Ordonnée à l'origine	-1.6644
Année de survenance (2006)	-0.3476
Année de survenance (2007)	0.1524
Année de développement (2)	-0.6931
Année de développement (3)	-0.6931
Âge	0.0378
$\hat{\phi}$	0.3900

suit :

$$\hat{\lambda}_i = exp\big(\hat{\beta}_0 + \hat{\beta}_1 \times X_1 + \cdots + \hat{\beta}_6 \times X_6\big).$$

Par après, la colonne «nombre de paiements» est enlevée de la base de données pour être remplacée par une colonne contenant les valeurs moyennes prédites par le modèle de fréquence (voir le tableau 3.4).

Basé sur le tableau 3.4, on fait la modélisation de la charge totale à l'aide du modèle GLM Poisson sur-dispersé. Lors de l'application de celui-ci, la fréquence prédite est considérée comme une mesure d'exposition au risque. Cette information est utilisée comme terme «offset» dans le modèle. Le tableau 3.5 présente les

Tableau 3.4: Base de données fictive.

No de sinistre	Année d'origine	Année de développement	Âge	Montant total ($)	Fréquence prédite
1	2005	1	25	250	0.4869
1	2005	2	25	0	0.2435
1	2005	3	25	0	0.2435
2	2005	1	55	175	1.5131
2	2005	2	55	50	0.7565
2	2005	3	55	15	0.7565
3	2006	1	85	1000	3.3209
3	2006	2	85	2500	1.6605
4	2006	1	43	500	0.6791
4	2006	2	43	0	0.3395
5	2007	1	40	100	1.0000

estimations des paramètres.

Une fois la calibration des deux modèles terminée, on construit la base de données qui correspond au triangle inférieur de développement qui permet de calculer la réserve RBNS.

En premier lieu, on prédit la fréquence moyenne des paiements futurs avec le modèle GLM Poisson sur-dispersé pour la fréquence calibré plus tôt. En deuxième lieu, à l'aide de la fréquence prédite, on prédit la charge totale avec le modèle calibré plus tôt (voir le tabeau 3.6). La réserve totale RBNS est la somme de la colonne «charge totale prédite», soit 266.13$.

Tableau 3.5: Estimations des paramètres pour la charge totale.

Coefficients	Estimation
Ordonnée à l'origine	4.4581
Année de survenance (2006)	1.3312
Année de survenance (2007)	0.0066
Année de développement (2)	0.9743
Année de développement (3)	-1.9185
Âge	0.0035
$\hat{\phi}$	394.9

3.2.2 Réserves IBNR

Une réserve IBNR est nécessaire pour estimer les sinistres survenus mais non déclarés à l'assureur. Pour modéliser le nombre de sinistres IBNR, on définit une variable aléatoire K_i qui représente le nombre total de sinistres pour l'année de survenance i. On suppose que K_i suit une distribution Poisson($\delta\omega_i$) où δ est un paramètre à estimer et ω_i est un paramètre connu représentant l'exposition au risque à l'année i. De plus, on définit une variable aléatoire T_i qui représente le temps nécessaire pour qu'un assuré déclare le sinistre à la compagnie d'assurance. On suppose que $T_i \sim$ Poisson(θ). On s'intéresse à la relation $(K|T_i = t_i^*)$ où t_i^* est le nombre d'années entre la date d'évaluation et la date de survenance. L'exemple 3.2.2 permet d'illustrer cette approche.

Exemple 3.2.2. *On considère un triangle de développement de 5 ans pour lequel* $i = 1, 2, \ldots, 5$. *De plus, on suppose que l'exposition unitaire est égale à 1* ($\omega_i = 1$) *Les données fictives sont présentées au tableau 3.7.*

La variable $K_i = K_i^{OBS} + K_I^{IBNR} \sim Poisson(\delta\omega_i)$ *représente le nombre total de si-*

Tableau 3.6: Base de données fictive du triangle inférieur de développement avec charge totale prédite.

No de sinistre	Année d'origine	Année de développement	Âge	Fréquence prédite	Charge totale prédite
3	2006	3	85	1.6605	107.37
4	2006	3	43	0.3395	18.95
5	2007	2	40	0.5000	132.47
5	2007	3	40	0.5000	7.34

Tableau 3.7: Nombres de sinistres fictifs survenus à l'année i.

i	K_i^{OBS}	K_i^{IBNR}	K_i
1	2089	0	2089
2	1978	1	1979
3	2130	0	2130
4	2233	3	2236
5	2057	7	2064

nistres à l'année i. La variable K_i^{OBS} *est le nombre de sinistres survenus à l'année i qui ont été déclarés avant la date d'évaluation. Donc,* $K_i^{OBS} = (K_i|T_i \leq t_i^* - 1) \sim Poisson(\delta\omega_i P(T_i \leq 5-1))$ *où* $T_i \sim Poisson(\theta)$. *La variable* K_i^{IBNR} *représente le nombre de sinistres survenus à l'année i sans avoir encore été déclarés. Donc,* $K_i^{IBNR} = (K_i|T_i > t_i^* - 1) \sim Poisson(\delta\omega_i P(T_i > 5-1))$.

En général, on dispose des réalisations de la variable aléatoire K_i^{OBS}. *À partir de celles-ci, on estime les paramètres* δ *et* θ.

La réserve IBNR de l'année i est calculée à partir du K_i^{IBNR} comme suit :

$$R_i^{IBNR} = K_i^{IBNR} \times \sum_{j=1}^{J} Y_{i,j},$$

et la réserve IBNR totale est donnée par

$$R^{IBNR} = \sum_{i=1} R_i^{IBNR}.$$

La réserve totale est définie comme étant la somme de la réserve RBNS et de la réserve IBNR

$$\begin{aligned} R_{i,j} &= R_{i,j}^{RBNS} + R_{i,j}^{IBNR} \\ R &= \sum_{(i,j)\in\mathcal{D}_J^{Inf}} R_{i,j} \end{aligned} \tag{3.3}$$

où $\mathcal{D}_J^{Inf}$ représente le triangle de développement inférieur avec J années maximales de développement.

3.3 Approche semi-paramétrique basée sur les arbres de régression

Après avoir utilisé la méthode classique GLM quasi-Poisson pour modéliser la fréquence, on utilise la méthode des arbres de décisions pour modéliser la même variable aléatoire. Cette décision est motivée par le fait que la relation entre la variable réponse et les variables explicatives est peut-être plus complexe que celle décrite par le GLM, que la masse de probabilité à zéro est très grande et que le GLM quasi-Poisson est incapable de bien la modéliser et qu'il y a fort probablement la présence des interactions entres les variables. Puisque les arbres de décisions font partie de la famille des méthodes semi-paramétriques, on s'attend à ce que ces problèmes soient mieux gérés.

Donc, la fréquence $(N_{i,j}^{(k)})$ est modélisée par un arbre de décision. Donc, le nombre

moyen de paiements par sinistre est défini par

$$\lambda_{i,j} = \sum_{m=1}^{M} c_m \mathbb{1}_{(\boldsymbol{X}_{i,j} \in R_m)},$$

où M est le nombre total de feuilles, c_m est la moyenne prédite par l'arbre si l'élément fait partie de la région R_m.

Le montant total par cellule $(Y_{i,j}^{(k)})$ est modélisé par le modèle GLM quasi-Poisson présenté à l'équation (3.2).

Pour expliquer le choix de la fonction de perte, on revient brièvement sur la théorie des GLM. On a expliqué au chapitre 2 que dans le cadre des GLM, les paramètres sont estimés en maximisant la fonction de log-vraisemblance $l(\boldsymbol{y}; \boldsymbol{\beta}, \phi)$. Pour tester si le modèle s'ajuste adéquatement aux données, on utilise souvent la mesure de déviance qui a la forme suivante :

$$\begin{aligned} D &= 2[l(\boldsymbol{y}; \boldsymbol{y}, \phi) - l(\boldsymbol{y}; \hat{\boldsymbol{\beta}}, \phi)], \quad \text{où} \\ l(\boldsymbol{y}; \boldsymbol{y}, \phi) &= \sum_{i=1}^{n} y_i \ln(y_i) - y_i - \ln(y_i!), \\ l(\boldsymbol{y}; \hat{\boldsymbol{\beta}}, \phi) &= \sum_{i=1}^{n} y_i \ln(\hat{\lambda}_i) - \hat{\lambda}_i - \ln(y_i!) \quad \text{et} \\ \hat{\lambda} &= e^{\boldsymbol{X}'_{i,j} \hat{\boldsymbol{\beta}}}. \end{aligned}$$

Donc, la déviance consiste à calculer la différence entre la fonction de log-vraisemblance du modèle saturé et la fonction de log-vraisemblance du modèle ajusté par la méthode GLM. Un modèle saturé passe par tous les points tel qu'illustré à la figure 3.1. Ainsi, on considère qu'un bon modèle a une petite déviance et c'est la motivation pour laquelle on choisit la déviance comme fonction

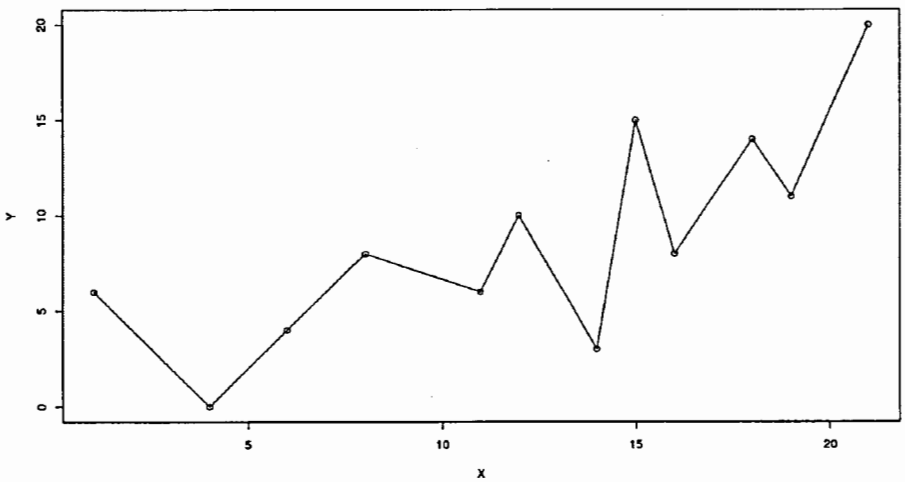

Figure 3.1: Modèle saturé.

de perte de l'arbre de régression. Lorsque $Y \sim \text{Poisson}(\hat{\lambda})$ avec $\hat{\lambda} = e^{\boldsymbol{X}'_{i,j}\hat{\boldsymbol{\beta}}}$, on a

$$\begin{aligned} D &= 2[l(\boldsymbol{y};\boldsymbol{y},\phi) - l(\boldsymbol{y};\hat{\boldsymbol{\beta}},\phi)] \\ &= 2\Big[\Big(\sum_{i=1}^{n} y_i \ln(y_i) - y_i - \ln(y_i!)\Big) - \Big(\sum_{i=1}^{n} y_i \ln(\hat{\lambda}_i) - \hat{\lambda}_i - \ln(y_i!)\Big)\Big] \\ &= 2\sum_{i=1}^{n}\Big(y_i \ln(y_i) - y_i - y_i \ln(\hat{\lambda}_i) + \hat{\lambda}_i\Big). \end{aligned}$$

Les paramètres étant estimés par maximum de quasi-vraisemblance, on a

$$\frac{\partial l(\boldsymbol{y};\hat{\boldsymbol{\beta}},\phi)}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - \hat{\lambda}_i)x_0 = 0$$

et on déduit que $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\lambda}_i$. Alors, on voit que

$$D = 2\sum_{i=1}^{n} y_i \ln\Big(\frac{y_i}{\hat{\lambda}_i}\Big).$$

Avec cette fonction de perte, l'arbre de régression estime une moyenne ($\hat{\lambda}$) différente pour chaque région.

3.3.1 Réserves RBNS

Pour calculer la réserve RBNS, on suit les mêmes étapes qu'à la section 3.2.1 avec l'exception que la fréquence est modélisée avec un arbre de régression. L'exemple 3.3.1 explique les étapes à suivre.

Exemple 3.3.1. *On considère de nouveau le tableau 3.4 et on considère les mêmes hypothèses que dans l'exemple 3.2.1.*

Pour modéliser la fréquence, on utilise les arbres de décisions. La figure 3.2 présente le graphique de l'arbre de décision.

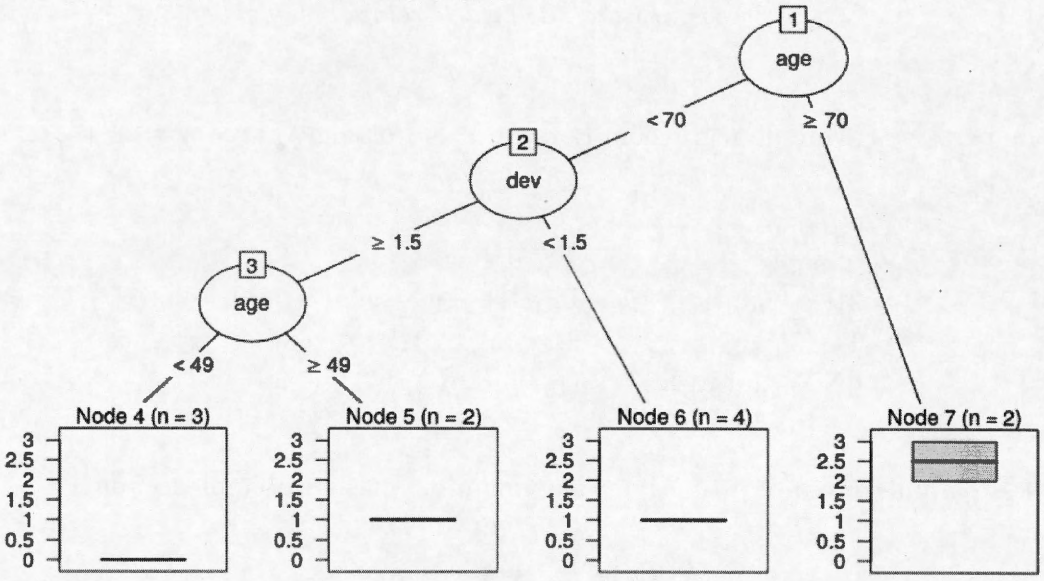

Figure 3.2: Arbre de décisions.

La fréquence prédite est présentée par le tableau 3.8.

La charge totale est prédite par un modèle GLM quasi-Poisson comme dans l'exemple 3.2.1. Les coefficients estimés sont présentés dans le tableau 3.9 Utilisant les mêmes étapes qu'à l'exemple 3.2.1, la réserve totale s'élève au 180.54$. Les ré-

Tableau 3.8: Fréquence prédite par l'arbre de décisions.

No de sinistre	Année d'origine	Année de développement	Âge	Montant total ($)	Fréquence prédite
1	2005	1	25	250	1.00
1	2005	2	25	0	0.25
1	2005	3	25	0	0.25
2	2005	1	55	175	1.00
2	2005	2	55	50	1.00
2	2005	3	55	15	1.00
3	2006	1	85	1000	2.00
3	2006	2	85	2500	2.00
4	2006	1	43	500	1.00
4	2006	2	43	0	0.25
5	2007	1	40	100	1.00

sultats sont présentés au tableau 3.10 sur la base de données représentant $\mathcal{D}_3^{Inf}$.

3.3.2 Réserves IBNR

Pour calculer la réserve IBNR dans cette section, on considère les hypothèses décrites à la section 3.2.2. La procédure générale de calcul reste inchangée mais les outils utilisés sont différents. La fréquence est modélisée par un arbre de décision, donc l'espérance de $Y_{i,j}$ est différente. L'exemple 3.3.2 illustre le concept.

Exemple 3.3.2. *On utilise les mêmes hypothèses et variables explicatives qu'à*

Tableau 3.9: Estimations des paramètres pour la charge totale.

Coefficients	**Estimation**
Ordonnée à l'origine	4.2755
Année de survenance (2006)	1.2879
Année de survenance (2007)	-0.0965
Année de développement (2)	0.5001
Année de développement (3)	-2.3203
Âge	0.0107
$\hat{\phi}$	152.24

l'exemple 3.2.2 pour prédire la valeur de K_i^{IBNR}.

$$R_i^{IBNR} = K_i^{IBNR} \times \sum_{j=1}^{J} Y_{i,j}$$

$$R^{IBNR} = \sum_{i=1} R_i^{IBNR}.$$

Tableau 3.10: Base de données fictive du triangle inférieur de développement avec la charge totale prédite.

No de sinistre	Année d'origine	Année de développement	Âge	Fréquence estimée	Charge totale prédite
3	2006	3	85	2.00	126.71
4	2006	3	43	0.25	10.12
3	2007	2	40	0.25	41.25
3	2007	3	40	0.25	2.46

CHAPITRE IV

ANALYSE NUMÉRIQUE

4.1 Données

Dans le cadre de ce projet, une base de données réelle est utilisée. Elle a été fournie par un assureur canadien. Cette base de données contient de l'information sur des sinistres survenus entre les années 1978 et les années 2017. Puisque les données récentes sont plus fiables et plus détaillées, on travaille avec les données des années 2005 à 2017 inclusivement.

Un nettoyage de la base de données a été effectué en enlevant les lignes présentant des données manquantes. Parmi un total de 9 337 420 lignes, seulement 1 255 lignes ont été enlevées, soit 0.013%. Retirer une si petite partie de la base de données n'influence probablement en rien les résultats. Aussi, les lignes contenant des paiements inférieurs à zéros ont été enlevées.

Cette dernière sous-base de données est divisée en deux parties soit les sinistres survenus entre les années 2005 à 2012 inclusivement et les sinistres survenus entre les années 2013 à 2017 inclusivement. Les modèles du chapitre 3 sont appliqués sur la première base de données et la deuxième base de données est utilisée pour obtenir le vrai montant payé pour les sinistres survenus entre les années 2005 et 2012 et valider les modèles.

La base de données contient 756 202 sinistres survenus sur l'intervalle de temps considéré. Puisque la base de données est de type transactionnel, on ne dispose que de l'information lorsqu'une transaction est effectuée. Pour faciliter les calculs, on a introduit une ligne supplémentaire par année de développement, pour chacun des sinistres, avec un nombre de paiements et une sévérité égales à zéro. Ainsi, on dispose d'un historique complet pour chaque sinistre (k). Sans surprise, on observe que la fréquence des paiements présente une grande masse de probabilité à zéro, soit de 77.2% de la base de données. Les tableaux A.4 et A.5 de l'annexe A présentent quelques statistiques descriptives supplémentaires pour chaque année de survenance et chaque année de développement.

Parmi les variables explicatives, seulement les variables qui contiennent de l'information pertinente pour la modélisation des réserves ont été gardées. Elles sont présentées aux tableaux A.1 et A.2 de l'annexe A. On observe à la figure 4.1 une croissance dans le nombre de sinistres survenus entre les années 2005 et 2012. En

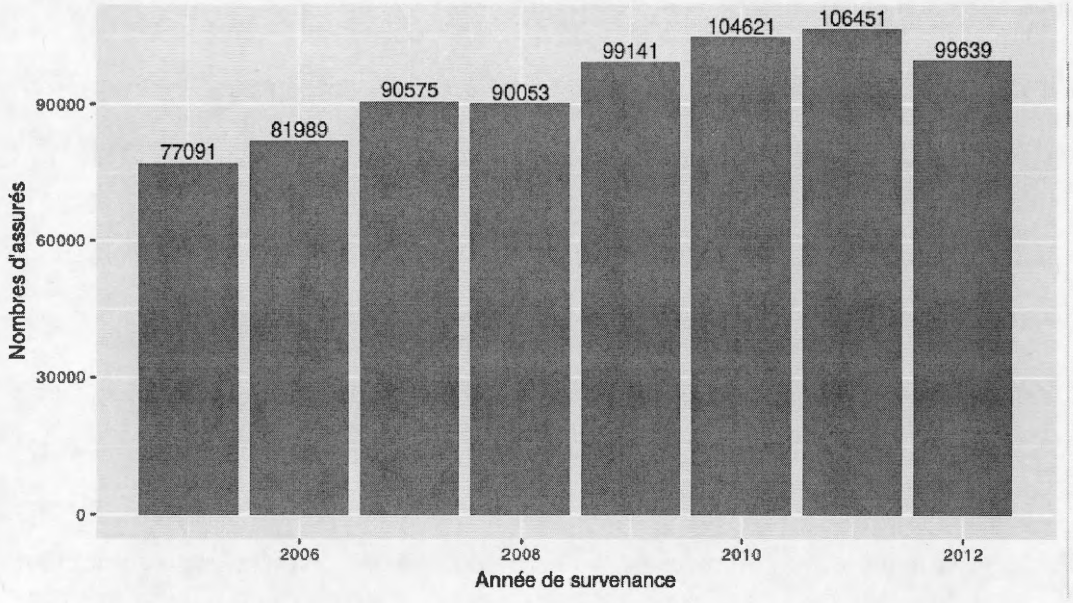

Figure 4.1: Le nombre de sinistres survenus entre les années 2005 et 2012.

absence d'information supplémentaire, on ne peut qu'émettre des hypothèses, par exemple, cette croissance pourrait être expliquée par le fait que la compagnie a réussi à acquérir une plus grande part du marché. La figure 4.2 indique le nombre de déclarations par année. On voit qu'il n'y a pas de différence importante entre

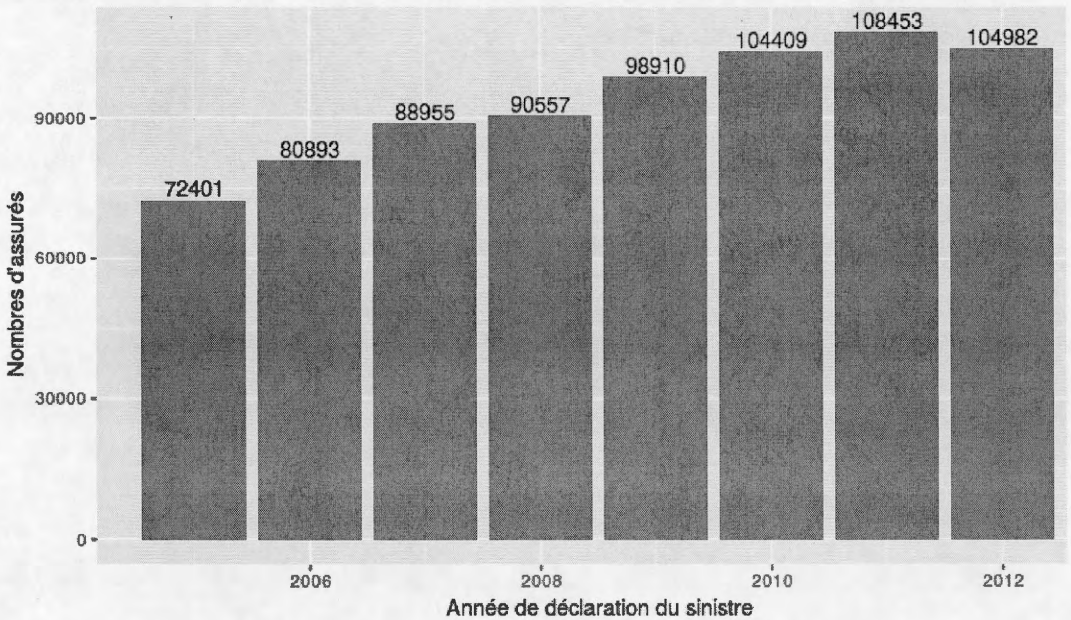

Figure 4.2: Le nombre de sinistres déclarés entre les années 2005 et 2012.

les figures 4.1 et 4.2. Cela indique qu'en très grande partie, les assurés déclarent rapidement leurs sinistres. La figure 4.3 montre la répartition des assurés dans les provinces. On voit facilement que la clientèle la plus importante se trouve en Ontario et en Alberta. La figure 4.4 donne une idée de la proportion de contrats pour véhicule personnel ou commercial parmi les réclamations. On voit que la grande majorité de réclamations implique un contrat pour leur(s) automobile(s). La figure 4.5 montre le fait que la plupart des contrats sont réglés dans la première ou deuxième année de développement. Finalement, la variable explicative *LOSS-KIND_NAME* donne de l'information concernant la cause du sinistre. Les deux catégories les plus importantes sont *multi-vehicle* constituant 35% de la base de

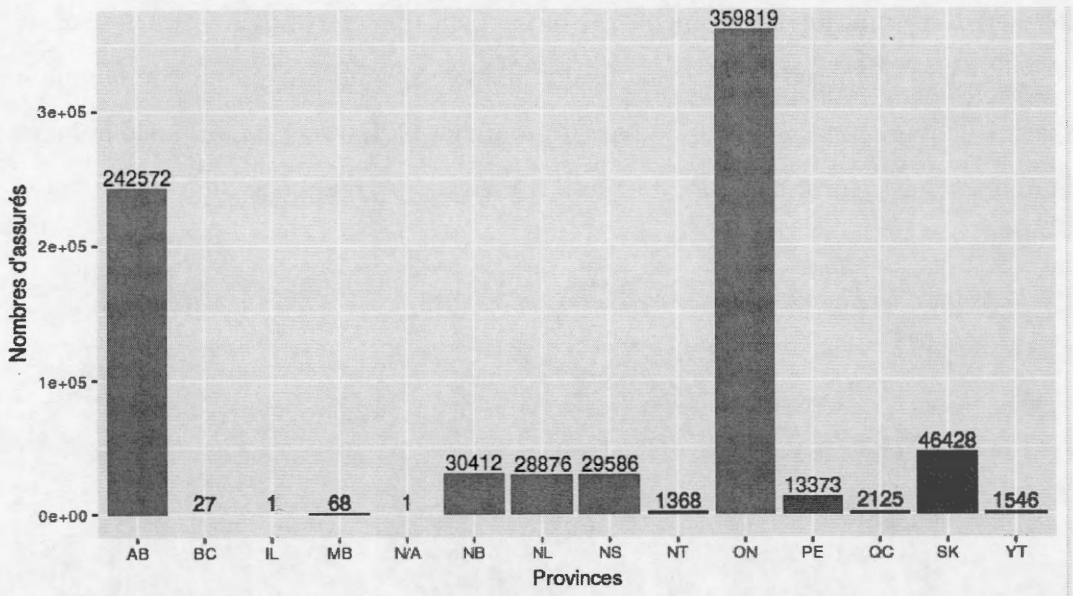

Figure 4.3: Le nombre de réclamations par province.

données et *not-available* constituant 22.7% de la base de données. Pour consulter la liste complète des variables explicatives et leurs catégories, veuillez consulter l'annexe A.

La base de données qui contient de l'information sur les années entre 2013 et 2017 permet d'obtenir le montant réellement payé qui est de 505 661 693$. Ce montant servira de validation pour les réserves RBNS obtenues avec les modèles. Puisque le modèle GLM quasi-Poisson est très répandu en industrie, on va l'utiliser comme une mesure de référence. Pour être en mesure de conclure que les modèles proposés au chapitre 3 sont supérieurs aux modèles classiques, on s'attend à calculer une réserve RBNS supérieure à la vraie réserve RBNS et inférieure aux modèle quasi-Poisson. De plus, on s'attend à ce que la variance soit moindre que la variance des modèles classiques.

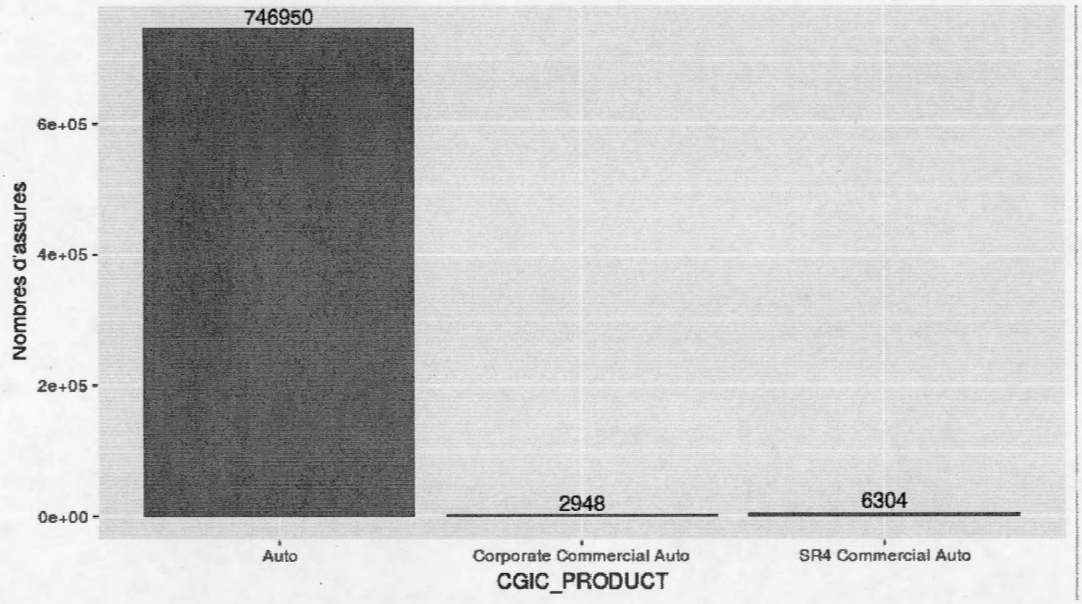

Figure 4.4: Le nombre de réclamations ayant différents contrats.

4.2 Appplication des approches classiques

Afin d'appliquer le modèle de Mack ainsi que les modèles GLM, la base de données a été modifiée. On a gardé l'année de survenance et l'année de développement comme variables explicatives et la somme des paiements comme variable réponse. On va appeler cette base de données $\mathcal{B}$. Pour commencer, on a appliqué le modèle de Mack pour évaluer la réserve totale. Le tableau 4.1 présente les résultats. Pour construire le tableau, 10 000 simulations ont été effectuées. En pratique, pour assurer leur solvabilité, les compagnies d'assurance gardent en réserve un montant correspondant à un quantile élevé de la distribution prédictive, souvent le 99^e. La figure 4.6 présente les distributions prédictives des deux modèles dans le même graphique.

On peut voir que la variance du modèle de Mack est très importante. Le montant prédit pour garder en réserve est plus de cent millions plus grand que le montant

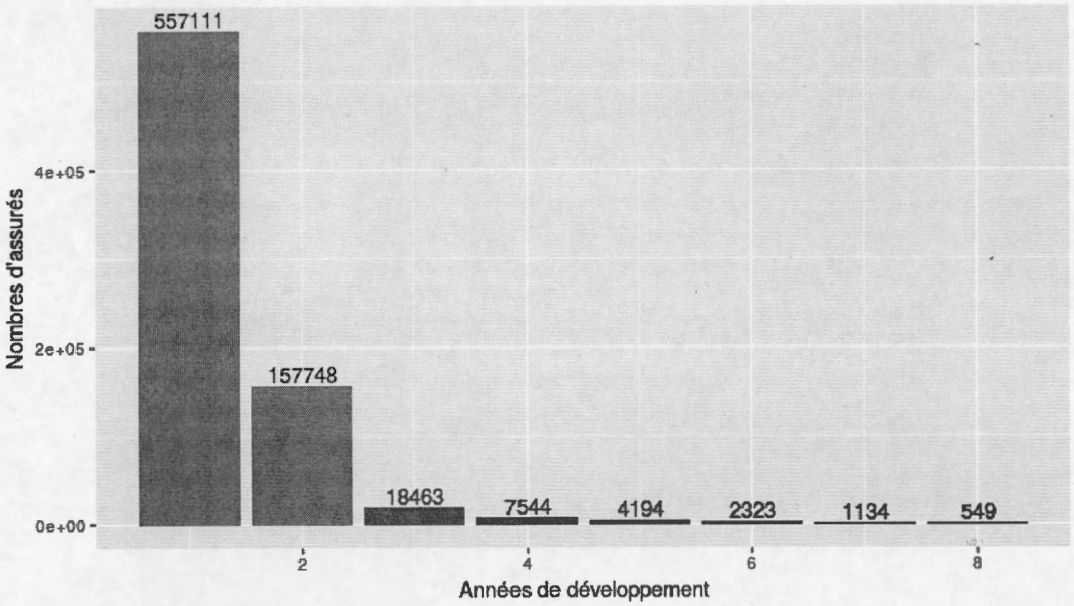

Figure 4.5: Le nombre de réclamations se trouvant dans différentes années de développement.

réellement payé par la compagnie, soit une sur-prédiction de 16.7%. Par contre, le modèle quasi-Poisson a une variance beaucoup plus petite prédisant une réserve totale de seulement 627 474 955, soit une prédiction 8.7% plus grande que le montant réel. On voit que ce dernier modèle couvre le montant réel tout en effectuant des économies par rapport au modèle le Mack. De plus, pour le modèle de Mack, la probabilité que la réserve totale soit plus grande que la réserve totale observée est de 64.3%. La même probabilité pour le modèle quasi-Poisson est 74.8%. Pour ces raisons, le modèle quasi-Poisson est préférable au modèle de Mack.

Tableau 4.1: Réserve totale (RBNS + IBNR) prédite par les modèles collectifs.

Modèle	Moyenne empirique	95^e quantile	99^e quantile
Mack	576 937 625	647 652 652	673 618 105
	(576 970 286)		
quasi-Poisson	576 765 283	612 217 438	627 474 955
	(576 970 286)		
Montant réel	562 364 975		

Note : les montants théoriques sont entre paranthèses

4.3 Application du modèle fréquence-sévérité individuel

4.3.1 Réserve RBNS

Dans cette section, on présente les résultats numériques pour les modèles individuels. Pour commencer, le **modèle G1** prédit la réserve RBNS utilisant la base de données décrite à la section 4.1 avec l'année de survenance et l'année de développement comme variables explicatives. Le **modèle G2** utilise la base de données décrite à la même section incluant toutes les variables explicatives telles quelles. Pour calibrer le **modèle G3**, on a modifié la base de données décrite à la section 4.1 en changeant la variable réponse N en une variable binaire où $N = 1$ s'il y avait au moins un paiement, sinon $N = 0$. Pour le **modèle G4**, la base de données a été changée en modifiant la variable réponse N. Lorsqu'il n'y a pas de paiemtent N est égal à zéro, lorsq'il y a un paiement, N est égal à 1, lorsqu'il y a deux paiements, N est égal à deux et ainsi de suite jusqu'à ce qu'il y a cinq ou plusieurs paiements où N prend la valeur 5. En d'autres mots, la variable N ne comprend plus que six catégories : 0, 1, 2, 3, 4 et 5+.

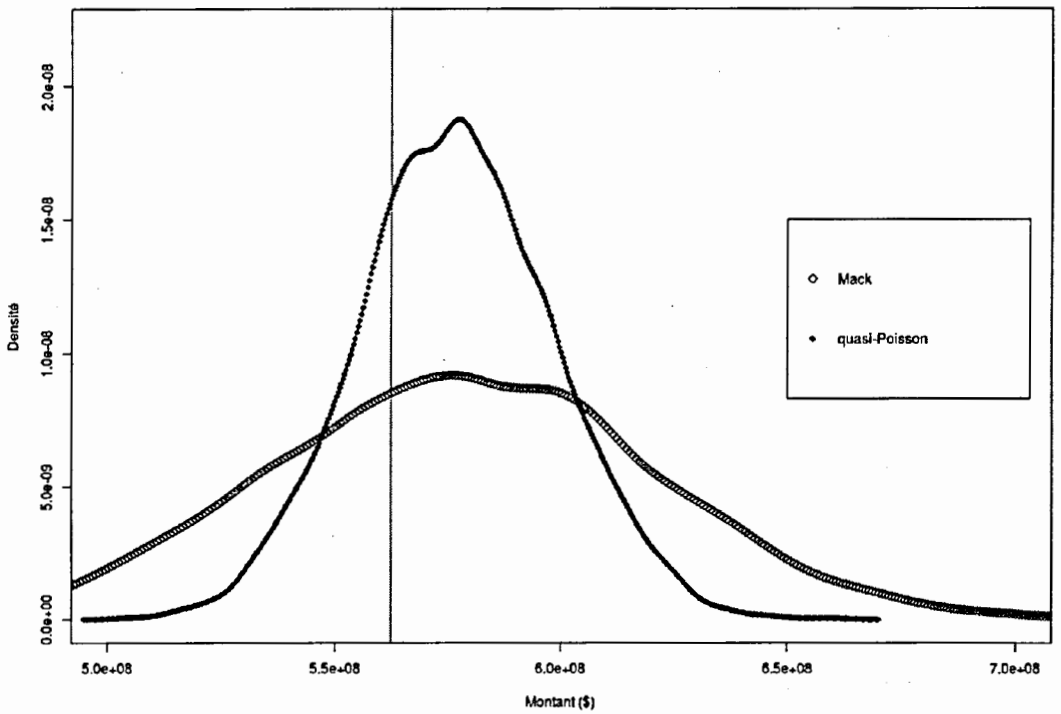

Figure 4.6: La distribution prédictive du modèle de Mack et du modèle quasi-Poisson.

Pour le modèle de la fréquence, en général, pour les modèles **G2**, **G3** et **G4**, tous les coefficients des années d'origine, des années de développement et des années de délai de déclaration sont significatifs. Les provinces significatives communes pour les trois modèles sont le Nouveau-Brunswick, la Nouvelle-Écosse, la Terre-Neuve, l'Ontario, le Québec, l'Île du Prince-Edward et le Saskatchewan. Les causes du sinistre communes significatives aux trois modèles sont les objets volants, la collision avec piéton, la collision avec plusieurs véhicules impliqués, le responsable du sinistre est parti et la réparation du pare-brise. Le paramètre de dispersion pour les modèles **G2** et **G4** sont 13.24 et 2.25 ce qui indique dans les deux cas une petite sur-dispersion. Le paramètre de dispersion pour le modèle **G3** est 0.997 ce qui indique que la fréquence suit pratiquement une distribution Bernoulli. Pour plus d'information, veuillez consulter les tableaux A.8, A.10 et A.12 présentés à

l'annexe A. Pour le modèle de la charge totale, en général, pour les modèles **G2**, **G3** et **G4**, la plupart des coefficients des années d'origine sont significatifs. Par contre, tous les coefficients des années de développement et des années de délai de déclaration des sinistres sont significatifs. Les provinces significatives communes pour les trois modèles sont le Nouveau-Brunswick, la Nouvelle-Écosse et le Saskatchewan. Les causes du sinistre communes significatives sont la collision avec des véhicules stationnées, le vol total, la réparation du pare-brise, le feu et la collision avec piéton. Le paramètre de dispersion des modèles **G2**, **G3** et **G4** est 196974.5, 196812.9 et 196974.5 respectivement ce qui indique qu'on trouve une très grande sur-dispersion dans la charge totale. Pour plus d'information, veuillez consulter les tableaux A.9, A.11 et A.13 présentés à l'annexe A.

Pour prédire la réserve RBNS, on a suivi la méthodologie de l'exemple 3.2.1. Le tableau 4.2 présente les résultats. Le **modèle G1** présente la plus grande variance des quatre modèles. Ceci pourrait être expliqué par le fait qu'il utilise seulement deux variables explicatives. Le **modèle G2** utilise toutes les variables explicatives de la base de données. Puisque ce modèle a accès à plus d'information que le modèle précédent, la variance est moins importante, ce qui lui permet de faire une prédiction plus précise que celle du **modèle G1**. Le **modèle G3** utilise la variable de la fréquence transformée en variable binaire. Le modèle Bernoulli est utilisé pour modéliser la fréquence ce qui explique une meilleure gestion de la grande masse de probabilité à zéro que le modèle quasi-Poisson. Cette explication pourrait justifier le fait que ce modèle prédit le montant de la réserve RBNS le plus proche de la réalité. Le **modèle G4** utilise toutes les variables explicatives et la variable réponse N est transformée en six catégories. On voit que la réserve RBNS est plus grande que celle du **modèle G3**. Ceci pourrait être expliqué par le fait que la fréquence est modélisée par le modèle linaire généralisé quasi-Poisson qui gère difficilement une grande masse de probabilité à zéro.

Tableau 4.2: Résultats des modèles individuels pour la réserve RBNS.

Modèle	**Moyenne empirique**	**95^e quantile**	**99^e quantile**
Modèle G1	579 724 756	601 393 552	610 190 608
	(579 724 756)		
Modèle G2	572 218 669	590 134 888	597 607 675
	(572 255 618)		
Modèle G3	564 393 790	582 252 966	589 649 740
	(564 778 596)		
Modèle G4	572 712 282	590 455 454	597 688 466
	(572 255 618)		
Montant réel	505 661 693		

Note : les montants théoriques sont entre paranthèses

La figure 4.7 présente la distribution prédictive des quatre modèles. La ligne verticale représente la réserve RBNS observée, soit 505 661 693$. Malgré que la réserve RBNS est légèrement sous-estimée, on voit que la densité prédictive des modèles se trouve loin de la droite. Ceci nous indique que pour les quatre modèles, la probabilité d'observer une réserve RBNS plus grande que la réserve RBNS observée est de 100%. De plus, on observe facilement que la variance du **modèle G1** est plus grande que la variance des trois autres modèles. La variance du **modèle G1**, **G2** et **G3** est sensiblement la même.

4.3.2 Réserve IBNR et réserve totale

Pour prédire la réserve totale, il faut prendre en compte la réserve IBNR comme il est expliqué par l'équation 3.3. Ayant en vue que la réserve IBNR observée ne

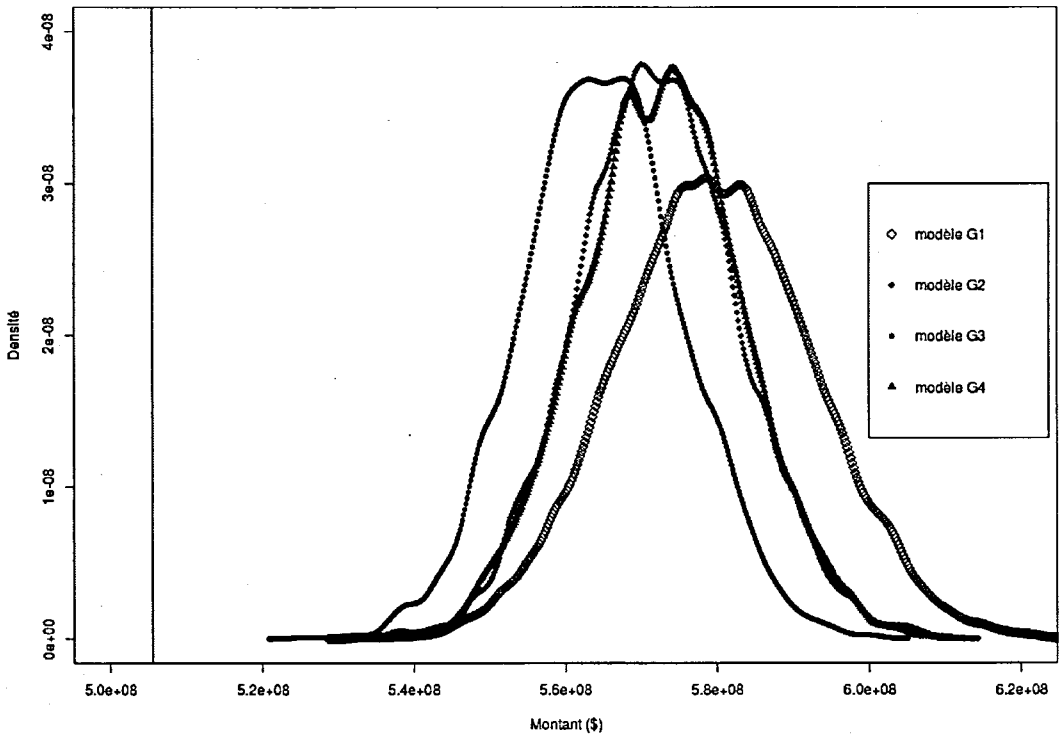

Figure 4.7: La distribution prédictive des quatre modèles GLM.

constitue que 10% de la réserve observée totale, on a choisi d'utiliser une méthode plus simple que celle décrite à la section 3.2.2 pour prédire la réserve IBNR.

Cette méthode consiste à construire une table de mortalité des dossiers, c'est-à-dire, on regarde la proportion cumulative des dossiers fermés à chaque année. On fait cette procédure pour les années 2005 et 2006 et on prend la moyenne des deux années comme proportion type. Par la suite, on fait l'hypothèse que dans le temps, la proportion des dossiers fermés reste la même. À l'aide de cette table de mortalité, on estime le nombre de dossiers IBNR pour les années 2007 à 2012.

Exemple 4.3.1. *On considère le triangle de développement qui débute à l'année 2010 et $J = 5$. On regarde la proportion des sinistres survenus en 2010 fermés en 2010. Par la suite, on regarde la proportion cumulative des sinistres survenus*

en 2010 fermés en 2011 et ainsi de suite jusqu'à l'année 2014. Posons que le tableau 4.3 présente les résultats. On pose que a proportion cumulative des dossiers

Tableau 4.3: Proportion cumulative des sinistres fermés survenus en 2010.

	2010	**2011**	**2012**	**2012**	**2013**	**2014**
Proportion	0.93	0.97	0.99	1	1	1

fermés est constante dans le temps et on applique cette table de mortalité à toutes les années de survenance tel que présenté au le tableau 4.4. Il ne reste plus que

Tableau 4.4: Estimation de la proportion cumulative des dossiers fermés à chaque année.

	2010	**2011**	**2012**	**2012**	**2013**	**2014**
2011		0.93	0.97	0.99	1	1
2012			0.93	0.97	0.99	1
2013				0.93	0.97	0.99
2014					0.93	0.97

prédire le nombre de dossiers IBNR qu'on a dans le portefeuille. Par la suite, on utilise le même cheminement que dans l'exemple 3.2.2 pour calculer la réserve IBNR.

Utilisant cet algorithme, on a prédit le nombre de dossiers IBNR dans la base de données, soit 7 412 dossiers. Pour simuler la base de données, c'est-à-dire pour obtenir les caractéristiques associées à chacun des dossiers, à chaque itération, on pige au hasard 7 412 contrats dans la base de données originale. On ajoute les lignes supplémentaire nécessaires pour appliquer les modèles présentés à la section 4.3.1 sur une base de données qui a la même forme que la base originale. Par la suite, on utilise les quatre modèles pour prédire la réserve IBNR. Le tableau 4.5

présente la réserve IBNR. Comme on peut le voir, même le 99^e quantile est plus bas

.Tableau 4.5: Résultats des modèles individuels pour la réserve IBNR.

Modèle	**Moyenne empirique**	**95^e quantile**	**99^e quantile**
Modèle G1	43 479 049	48 899 705	51 261 847
Modèle G2	43 474 291	47 947 980	51 893 233
Modèle G3	43 166 419	47 990 586	50 381 274
Modèle G4	43 474 291	47 947 980	51 893 233
Montant réel	56 703 282		

que la vraie réserve IBNR. Ceci pourrait être expliqué par plusieurs facteurs. En premier lieu, la méthode utilisée est déterministe. En deuxième lieu, les contrats pris aux hasard dans la base de données n'ont pas été adaptés pour une prédiction de modèles IBNR. En d'autres mots, si les contrats choisis sont survenus en 2008 et la date d'évaluation est en décembre 2012, il est impossible que la variable *REPORTYEAR* soit égale à 2008. Dans ce cas, la variable *REPORTYEAR* doit être égale à 2013 pour que les modèles puissent faire une bonne prédiction. Ayant en vue ces défauts, on peut améliorer la méthode d'évaluation des réserves IBNR en choisissant une méthode probabiliste comme présenté à la section 3.2.2 et de changer la variable explicative *REPORTYEAR* pour qu'elle soit cohérente avec la situation.

Le tableau 4.6 présente les résultats de la réserve totale. Malgré qu'on peut amener des amélioration importantes au modèle IBNR, on voit que la réserve observée est inférieure aux montants prédits par les modèles ce qui implique que les modèles sont valides. Par contre, on remarque que les quatre modèles présentés dans cette section ont le 99^e quantile plus grand que le 99^e quantile de la méthode classique

Tableau 4.6: La réserve totale (RBNS + IBNR) prédite par les modèles individuels.

Modèle	Moyenne empirique	95^e quantile	99^e quantile
Modèle G1	623 241 247	645 518 672	655 127 900
Modèle G2	615 755 756	634 218 322	642 052 684
Modèle G3	608 106 208	626 209 477	633 240 472
Modèle G4	616 186 572	634 488 643	642 270 642
Montant réel	562 364 975		

du quasi-Poisson.

4.4 Application des modèles non-paramétriques individuels

4.4.1 Réserve RBNS

À la section 4.3, on a présenté les résultats des modèles individuels où la fréquence est modélisée à l'aide de la méthode linéaire généralisée Poisson sur-dispersé ou par la distribution Bernoulli. Dans cette section, on présente les résultats des modèles utilisant les arbres de décision pour modéliser la fréquence.

En premier lieu, on calibre un arbre de régression tel que présenté au chapitre 2. Comme l'équation 2.3 l'indique, pour faire l'élagage, il faut choisir un paramètre de complexité. Pour ce faire, on utilise la validation croisée tel que décrite au même chapitre. Le paramètre de complexité qui présente la plus petite erreur est choisi pour le modèle final. La figure 4.8 montre le comportement de l'erreur en fonction du paramètre de complexité. Plus le paramètre de complexité se trouve à droite du graphique, plus le nombre de feuilles de l'arbre résultant est grand. En d'autres mots, il est facile de choisir un paramètre de complexité qui cause du

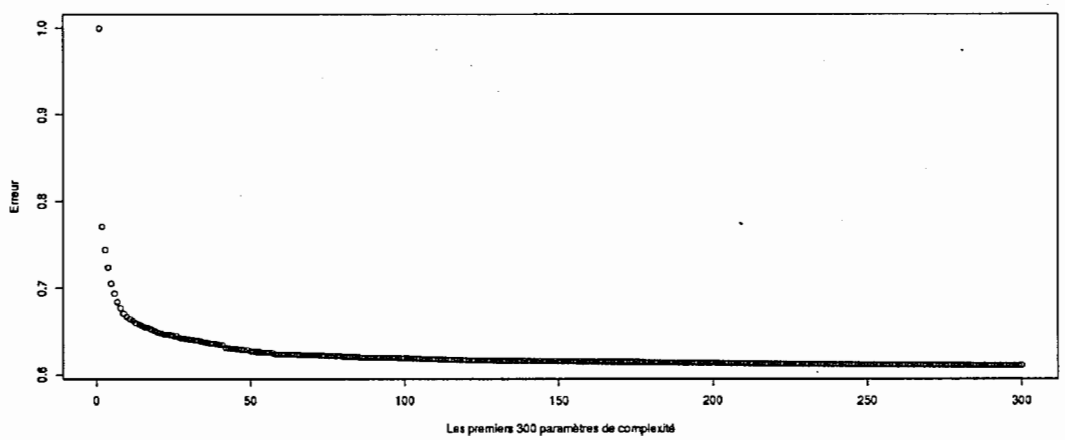

Figure 4.8: Graphique de l'erreur en fonction du paramètre de complexité.

sur-ajustement des données. Dans ce cas, le paramètre de complexité qui amène la plus petite erreur est le 908^e qui n'est pas représenté dans le graphique. Il est évident que le paramètre de complexité proposé cause du sur-ajustement. Alors, on va choisir un paramètre de complexité différent en consultant la figure 4.8. On regarde les premiers cent points sur l'axe des x qui forment la courbure importante sur le graphique. Si on choisit le point qui est à la base de la courbure, soit le 60^e, on est certains d'éviter la sur-ajustement de données sans ajouter une erreur importante aux prédictions.

Pour prédire la réserve RBNS, on a suivi la méthodologie de l'exemple 3.3.1. Le **modèle A1** utilise toutes les variables explicatives de la base de données décrite dans la section 4.1. Le **modèle A2** est appliqué sur la base de données où la variable N est transformée en variable binaire. Le **modèle A3** est calibré utilisant la base de données où la variable réponse N est divisée en six catégories : 0, 1, 2, 3, 4 et 5+ paiements. En d'autres mots, tous les sinistres qui ont reçu cinq paiments ou plus se voient attribuer comme variable réponse la catégorie 5+.

Pour le modèle de la charge totale, en général, pour les trois modèles **A1**, **A2**

et **A3**, les coefficients des années d'origine, des années de développement et des années de délai sont significatifs. Les provinces significatives communes sont le Nouveau-Brunswick, l'Ontario, l'Île du Prince-Edward et le Saskatchewan. Les causes du sinistre significatives communes aux trois modèles sont la collision avec un véhicule stationné, la réparation du pare-brise, la collision avec piéton, le responsable du sinistre est parti, le vandalisme et l'entrée dans le véhicule par infraction.

Tableau 4.7: Résultats des modèles individuels pour la réserve RBNS avec arbre de régression.

Modèle	**Moyenne empirique**	95^e **quantile**	99^e **quantile**
Modèle A1	647 878 279	664 303 034	671 086 418
	(647 877 400)		
Modèle A2	540 882 839	558 075 368	566 246 925
	(540 782 556)		
Modèle A3	537 128 645	553 248 282	557 993 407
	(537 658 228)		
Montant réel	505 661 693		

Note : les montants théoriques sont entre paranthèses

Le tableau 4.7 résume les résultats. On voit facilement que le **modèle A1** surévalue la réserve RBNS de façon importante. Ceci pourrait être expliqué par le fait que la variable N contient un nombre important de valeurs aberrantes. L'arbre de régression donne comme prédiction la moyenne des éléments qui se trouvent dans une certaine région. Comme la moyenne n'est pas une mesure robuste, elle se laisse facilement influencer par les valeurs aberrantes entrainant ainsi une sur-estimation

de la fréquence. Par contre, le **modèle A2** est moins sujet au valeurs aberrantes parce que la variable N ne prend que 1 ou 0 (paiement ou pas de paiement) comme valeurs. Il prédit un montant moyen plus petit que les modèles précédents tout en couvrant adéquatement les engagement financiers de la compagnie. Le **modèle A3** prédit un montant moyen pour la réserver encore plus petit que le **modèle A2**. La prédiction pourrait être meilleure parce qu'il utilise plus d'information sur la variable N sans être influencé par des valeurs aberrantes, car la variable N ne contient que 5 catégories.

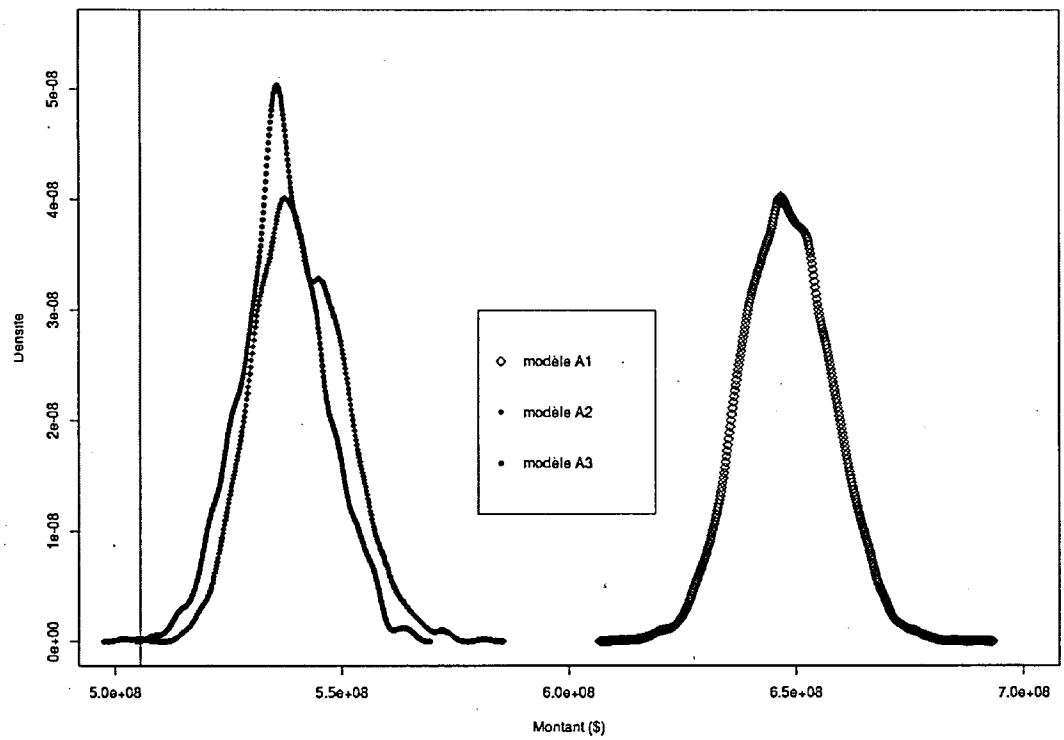

Figure 4.9: Distribution prédictive des trois modèles.

La figure 4.9 présente la distribution prédictive des trois modèles. La ligne verticale représente la réserve observée, soit le 505 661 693$. On voit facilement que la probabilité d'observer un montant plus grand que le montant observé est de pra-

tiquement 100%. De plus, sur cette figure, on peut voir rapidement la magnitude de la variance des trois modèles. On observe que la variance des trois modèles est similaire, mais plus petite que les modèles de la section 4.3.1.

4.4.2 Réserves IBNR

Après avoir calculé la réserve RBNS, on calcule la réserve IBNR pour être en mesure de déterminer la réserve totale. On applique la même méthode qu'à la section 4.3.2. Le tableau 4.8 présente la réserve IBNR calculée par les trois modèles. Comme on peut le voir, le problème d'estimation de la réserve IBNR persiste mal-

Tableau 4.8: Résultats des modèles individuels pour la réserve IBNR avec arbres de régression.

Modèle	Moyenne empirique	95^e quantile	99^e quantile
Modèle A1	44 331 128	50 942 692	52 914 184
Modèle A2	43 100 745	49 330 366	52 220 280
Modèle A3	42 391 087	46 439 991	48 957 760
Montant réel	56 703 282		

gré le changement de modèles. Comme il a été expliqué, le fait que la méthode soit déterministe et le fait que la variable REPORTYEAR soit erronée sont les causes principales des mauvaises prédictions.

Le tableau 4.9 présente la réserve totale prédite par les modèles avec arbre de régression.

Tableau 4.9: La réserve totale (RBNS + IBNR) prédite par les modèles individuels avec arbre de régression.

Modèle	**Moyenne empirique**	**95^e quantile**	**99^e quantile**
Modèle A1	692 239 561	710 030 778	716 844 130
Modèle A2	583 953 585	602 099 830	610 681 465
Modèle A3	579 519 731	595 773 631	601 744 060
Montant réel	562 364 975		

4.5 Comparaison des résultats

Les sections précédentes de ce chapitre ont été consacrées à la présentation des résultats des différents modèles présentés au chapitre 3. Le tableau 4.10 présente les résultats agrégés de tous les modèles. Puisqu'on connait le montant de la réserve totale observée, on la garde comme limite inférieure pour les modèles. En d'autres mots, si un modèle présente un 99^e quantile plus bas que la réserve totale observée est automatiquement classifié comme mauvais. Comme on peut le voir dans le tableau 4.10 aucun de modèles ne présente une telle valeur. Puisque l'approche collective utilisant la méthode des modèles linéaires généralisées Poisson sur-dispersé est très répandue en industrie, on la garde comme limite supérieure pour les modèles. En d'autres mots, si un modèle présente un 99^e quantile plus grand que celui de l'approche collective quasi-Poisson, le modèle est considéré comme mauvais, car il ne permet pas de réduire le montant nécessaire dans la réserve pour que la compagnie soit en mesure de profiter des opportunités de marché.

Comme on peut le voir, le modèle de Mack et le modèle quasi-Poisson ont une

Tableau 4.10: La réserve totale (RBNS + IBNR) prédite par tous les modèles individuels.

Modèle	Moyenne empirique	95^e quantile	99^e quantile
Mack	576 937 625	647 652 652	673 618 105
quasi-Poisson	576 765 283	612 217 438	627 474 955
Modèle G1	623 241 247	645 518 672	655 127 900
Modèle G2	615 755 756	634 218 322	642 052 684
Modèle G3	608 106 208	626 209 477	633 240 472
Modèle G4	616 186 572	634 488 643	642 270 642
Modèle A1	692 239 561	710 030 778	716 844 130
Modèle A2	583 953 585	602 099 830	610 681 465
Modèle A3	579 519 731	595 773 631	601 744 060
Montant réel	562 364 975		

moyenne de la réserve très proche, ce qui est normal. Par contre, on voit que le modèle de Mack présente une variance beaucoup plus grande que celle du modèle quasi-Poisson, ce qui empêche la compagnie de profiter d'opportunités de marché. Le **modèle G1** est le modèle individuel qui a comme variables explicatives l'année de survenance et l'année de développement et la fréquence est modélisée par GLM Poisson sur-dispersé. Tant la moyenne que le 99^e quantile sont plus grandes que le modèle quasi-Poisson. Ceci pourrait être expliqué par le fait que la distribution Poisson est incapable de gérer une grande masse de probabilité à zéro comme celle présente dans la base de données. En d'autres mots, la fréquence est sur-estimée. Donc, ce modèle n'offre pas une meilleure précision pour la prédiction que l'approche classique quasi-Poisson. Le **modèle G2** est le modèle individuel qui prend en compte toutes les variables explicatives et la fréquence est modélisée par GLM quasi-Poisson aussi. Une fois de plus, le modèle GLM quasi-Poisson a

sur-estimé la fréquence et le 99^e quantile est plus grand que la limite supérieure posée. Le **modèle G3** prend en considération toutes les variables explicatives, mais la variable N est transformée en variable binaire. La distribution Bernoulli est utilisée pour prédire la fréquence. Malgré ce changement, le modèle est incapable d'amener des économies à la compagnie d'assurance. Ceci pourrait être expliqué par le fait que la charge totale manque de l'information concernant la fréquence. Alors, le modèle utilisé pour la charge totale la sur-estime en prédisant une moyenne grossière plutôt qu'une moyenne adaptée pour chaque dossier. Le **modèle G4** prend en considération toutes les variables explicatives également et la fréquence est modélisée par GLM quasi-Poisson. Par contre, la variable N est transformée en catégories 0, 1, 2, 3, 4 et 5 ou plus paiements. Comme on peut voir, ce modèle n'amène rien de mieux que ces prédécesseurs parce que 99^e quantile est toujours plus grand que celui de l'approche collective quasi-Poisson. Ceci pourrait être expliqué toujours par l'incapacité de la distribution Poisson de traiter la grande masse de probabilité à zéro. Le **modèle A1** utilise toutes les variables explicatives de la base de données et la fréquence est modélisée par un arbre de décision. Comme on peut le voir, la fréquence est grandement sur-estimée. Ceci pourrait être expliqué par le fait que chaque région prédit la moyenne des éléments qu'elle contient. Puisque la moyenne est une mesure non-robuste, elle est fortement influencée par des valeurs aberrantes de la variables N ce qui entraine une très grande sur-estimation de la fréquence. Il est évident que ce modèle peut facilement être classifié comme mauvais. Le **modèle A2** utilise toutes les variables explicatives et la fréquence est la variable N transformée en variable binaire. Grâce à cette dernière variable N, le modèle arbre n'est clairement pas influencé par les valeurs aberrantes. On peut voir pour la première fois parmi les modèles une moyenne empirique plus petite de 600 000 000$. De plus, on observe une petite variance ce qui entraine un 99^e quantile plus bas que la limite supérieure posée ce qui nous permet de classifier ce modèle comme étant bon. Le **modèle**

A3 est appliqué sur toutes les variables explicatives et la fréquence est la variable N transformée en cinq catégories. Comme cela a été le cas pour le **modèle A2**, les valeurs aberrantes n'influencent pas la prédiction de la fréquence. Par contre, grâce au fait qu'il y a plus d'information incorporée dans la variable N, le modèle est capable de faire des prédictions plus précises et ainsi il permet à la compagnie d'assurance de faire des économies tout en se protégeant adéquatement contre l'insolvabilité.

CONCLUSION

Dans le cadre de ce projet, un nouveau modèle pour le calcul de réserves en assurance non-vie a été proposé. On a adopté l'approche individuelle parce qu'elle offre plusieurs avantages par rapport à l'approche collective comme l'accès à plus d'information sur les réclamations. On rappelle que le modèle quasi-Poisson dans le cadre collectif a été posé comme limite supérieure pour le modèle proposé parce qu'il est fréquemment utilisé en industrie. Le modèle proposé se base sur une structure fréquence-sévérité où la fréquence (N) est modélisée à l'aide d'un arbre de décision et la charge totale (Y) avec un modèle linéaire généralisé Poisson sur-dispersé. Malgré le manque de robustesse de la part de l'arbre de décision, ce modèle permet à la compagnie d'assurance de bien se protéger contre le risque d'insolvabilité tout en lui permettant de faire des économies pour profiter des opportunités de marché. Cette amélioration pourrait être due au fait que l'arbre de décision est une méthode semi-paramétrique et est capable de déterminer des liens complexes entre la variable réponse et les variables explicatives et de bien gérer la grande masse de probabilité à zéro et les données manquantes.

Dans une recherche future, on pourrait amener plusieurs améliorations à ce modèle. En premier lieu, on peut facilement améliorer le modèle pour la réserve IBNR en utilisant un modèle optimal comme proposé à la section 3.2.2. En deuxième lieu, on peut améliorer le modèle de fréquence, l'arbre de décision, en appliquant l'algorithme *random forest* ou *gradient boosting machine* pour des prédictions plus précises. En troisième lieu, on peut utiliser l'arbre de régression le *random forest* ou le *gradient boosting machine* pour modéliser la charge totale également.

ANNEXE A

TABLEAUX COMPLÉMENTAIRES AU CHAPITRE 4

Tableau A.1: Les variables et leur catégories de la base de données.

Nom	Description
LOSSYEAR	L'année de survenance du sinistre
2005-2012	
REPORTYEAR	L'année de la déclaration du sinistre
2005-2012	
INTC_RISK_PRIVINCE_CODE6	La province.
AB	Alberta
BC	Colombie Britanique
...	...
SK	Saskatchewan
YT	Yukon
CGIC_PRODUCT	Le type de contrat
Auto	Contrat pour particuliers
Auto commercial	Contrat pour commerce
SR4 commercial	Contrat pour commerce
DEV	L'année de développement
1 à 8	
PAID	montant payé
N	Nombre de paiements

Tableau A.2: Les variables et leur catégories de la base de données (suite).

Nom	**Description**
LOSSKIND_NAME	La cause du sinistre
Animal collision	Collision avec animaux
Collision	Collision
Collision with parked vehicle	Collision avec un véhicule stationné
Damage by water	Dommage causé par l'eau
Earthquake	Tremblement de terre
Emergency road service	Sortie de route
Explosion	Explosion
Fire	Feu
Flying / falling object	Objets volants
Glass other	Dommage aux vitres
Glass windshield	Dommage au pare-brise
Hail	Verglas
Hit and run	Le rsponsable du sinistre est parti
Hit pedestrian	Collision avec pieton
In transit	Transport en commun
Lightning	Foudre
Multi-vehicle	Plusieurs véhicules

Tableau A.3: Les variables et leur catégories de la base de données (suite).

Nom	Description
Riot	Manifestation
Rising water	Dommage causé par l'eau
Single-vehicle	Un seul véhicule
Smoke	Fumée
Total theft	Vol total de voiture
Vandalism	Vandalisme
Vehicle break-in	Entrée forcée dans le véhicule
Wind	Vent
Windshield repair	Réparation de pare-brise
Windshield replace	Remplacement de pare-brise
Other	Autre
Not available	Cause inconnue

Tableau A.4: Statistiques descriptives des données en milliers de dollars.

Année	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fréq.	55 860	48 189	12 627	4 953	2 757	1 337	1 607	835
Total	143 675	73 404	45 388	28 798	26 875	14 604	25 599	10 085
Moy.	1.622	1.523	3.594	5.814	2.748	10.923	15.930	12.077
2006								
Fréq.	89 820	53 535	13 950	5 448	2 812	1 351	681	
Total	150 629	76 898	45 450	33 768	27 009	18 764	11 829	
Moy.	1.677	1.436	3.258	6.198	9.604	13.889	17.370	
2007								
Fréq.	101 462	53 671	13 162	4 918	2 428	1 333		
Total	180 275	89 379	40 223	34 151	19 877	15 223		
Moy.	1.776	1.665	3.055	6.994	8.186	11.420		
2008								
Fréq.	99 779	54 333	14 178	5 332	3 020			
Total	179 646	95 758	51 224	37 466	29 040			
Moy.	1800	1.762	3.612	7.026	9.615			
2009								
Fréq.	90 084	48 581	11 411	4 793				
Total	180 714	79 066	37 297	36 226				
Moy.	2.006	1.627	3.268	7.558				

Tableau A.5: Statistiques descriptives des données en milliers de dollars (suite).

Année	1	2	3	4	5	6	7	8
2010								
Fréq.	79 983	41 893	10 078					
Total	156 574	75 089	33 220					
Moy.	1.957	1.792	3.296					
2011								
Fréq.	81 303	30 908						
Total	163 511	53 848						
Moy.	2.011	1.742						
2012								
Fréq.	77 460							
Total	157 885							
Moy.	2.038							

Tableau A.6: Les coefficients et leur écart-type de la fréquence ***modèle G1.***

Coefficient	**Estimé**
Ordonnée à l'origine	0.3949956 (0.0123305)
Origine 2006	0.0022828 (0.0162924)
Origine 2007	-0.0228522 (0.0160909)
...	...
Origine 2011	-0.3280643 (0.0182604)
Origine 2012	-0.1083630 (0.0207838)
Développement 2	-0.6446788 (0.0099924)
Développement 3	-1.9994781 (0.0180233)
...	...
Développement 7	-4.7472497 (0.1165122)
Développement 8	-4.6832378 (0.1616156)
ϕ	21.68294

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.7: Les coefficients et leur écart-type de la sévérité ***modèle G1.***

Coefficient	**Estimé**
Ordonnée à l'origine	7.4402332 (0.0319093)
Origine 2006	-0.0002846 (0.0405553)
Origine 2007	0.0168002 (0.0403589)
...	...
Origine 2011	0.1649105 (0.0477619)
Origine 2012	0.1796287 (0.05355730
Développement 2	-0.1091020 (0.0281147)
Développement 3	0.6181007 (0.0378636)
...	...
Développement 7	1.9535845 (0.0802863)
Développement 8	1.9588615 (0.1731291)
ϕ	292115.3

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.8: Les coefficients et leur écart-type de la fréquence ***modèle G2.***

Coefficient	Estimé
Ordonnée à l'origine	-0.315894 (0.019583)
Origine 2006	0.293686 (0.044567)
Origine 2007	0.528497 (0.059108)
...	...
Origine 2011	1.646925 (0.097764)
Origine 2012	2.054770 (0.112409)
Reportyear 2006	-0.279454 (0.044674)
Reportycar 2007	-0.514772 (0.059243)
...	...
Reportycar 2011	-1.518479 (0.097413)
Reportyear 2012	-1.726031 (0.111844)
Province BC	0.666957 (0.422453)
Province MB	0.088972 (0.318128)
...	...
Province SK	-0.067775 (0.018831)
Province YT	-0.016085 (0.081257)
Cause du sinistre Collision	0.041289 (1.377518)
Cause du sinistre Collision with anything	-0.368886 (0.435313)
...	...
Cause du sinistre Windshield repair	-0.573432 (0.028828)
Cause du sinistre Windshield replace	-0.560090 (0.027329)
Développement 2	-0.681878 (0.007831)
Développement 3	-2.048532 (0.014110)
...	...
Développement 7	-4.799829 (0.091060)
Développement 8	-4.736543 (0.1263070
ϕ	13.24348

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.9: Les coefficients et leur écart-type de la sévérité ***modèle G2.***

Coefficient	Estimé
Ordonnée à l'origine	7.972068 (0.047535)
Origine 2006	-0.288937 (0.100476)
Origine 2007	-0.691242 (0.128547)
...	...
Origine 2011	-2.398389 (0.201696)
Origine 2012	-2.750741 (0.235126)
Reportyear 2006	0.265672 (0.100741)
Reportyear 2007	0.683725 (0.128882)
...	...
Reportyear 2011	2.254678 (0.200147)
Reportyear 2012	2.647585 (0.232891)
Province BC	0.987744 (0.432239)
Province MB	0.053680 (0.553283)
...	...
Province SK	-0.447242 (0.054211)
Province YT	0.030102 (0.184830)
Cause du sinistre Collision	-0.046166 (3.854534)
Cause du sinistre Collision with anything	-0.710894 (1.461108)
...	...
Cause du sinistre Windshield repair	-0.227154 (0.084066)
Cause du sinistre Windshield replace	0.117445 (0.068974)
Développement 2	-0.087299 (0.023147)
Développement 3	0.642408 (0.031187)
...	...
Développement 7	2.262588 (0.089979)
Développement 8	1.986293 (0.142200)
ϕ	196974.5

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.10: Les coefficients et leur écart-type de la fréquence ***modèle G3.***

Coefficient	Estimé
Ordonnée à l'origine	1.4384884 (0.0116229)
Origine 2006	0.1966484 (0.0256467)
Origine 2007	0.3038794 (0.0343878)
...	...
Origine 2011	0.7465566 (0.0553593)
Origine 2012	2.1190415 (0.0632851)
Reportyear 2006	-0.2025536 (0.0257556)
Reportycar 2007	-0.3007798 (0.0345258)
...	...
Reportycar 2011	-0.8281801 (0.0552747)
Reportyear 2012	-1.1181716 (0.0619650)
Province BC	-0.0102499 (0.3500356)
Province MB	0.1656734 (0.2267030)
...	...
Province SK	-0.1457108 (0.0092612)
Province YT	-0.0667756 (0.0489729)
Cause du sinistre Collision	-1.5157253 (0.9674618)
Cause du sinistre Collision with anything	-1.0660810 (0.2167060)
...	...
Cause du sinistre Windshield repair	-0.3291637 (0.0149749)
Cause du sinistre Windshield replace	-0.3048426 (0.0141548)
Développement 2	-2.3309659 (0.0049107)
Développement 3	-4.4953485 (0.0094876)
...	...
Développement 7	-7.0301929 (0.0547164)
Développement 8	-7.6210337 (0.1047683)
ϕ	0.9973941

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.11: Les coefficients et leur écart-type de la sévérité ***modèle G3.***

Coefficient	Estimé
Ordonnée à l'origine	6.856144 (0.047504)
Origine 2006	-0.018114 (0.100440)
Origine 2007	-0.198426 (0.128489)
...	...
Origine 2011	-0.850326 (0.201575)
Origine 2012	-0.980575 (0.234957)
Reportyear 2006	0.009800 (0.100703)
Reportyear 2007	0.204148 (0.128820)
...	...
Reportyear 2011	0.845950 (0.200010)
Reportyear 2012	1.081797 (0.232701)
Province BC	1.655610 (0.432059)
Province MB	0.123266 (0.553056)
...	...
Province SK	-0.493741 (0.054189)
Province YT	0.023189 (0.184754)
Cause du sinistre Collision	0.220883 (3.853012)
Cause du sinistre Collision with anything	-0.934696 (1.460503)
...	...
Cause du sinistre Windshield repair	-0.753803 (0.084031)
Cause du sinistre Windshield replace	-0.399162 (0.068943)
Développement 2	-0.247952 (0.023137)
Développement 3	-0.675848 (0.031184)
...	...
Développement 7	-1.777979 (0.089947)
Développement 8	-1.989939 (0.142144)
ϕ	196812.9

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.12: Les coefficients et leur écart-type de la fréquence ***modèle G4.***

Coefficient	Estimé
Ordonnée à l'origine	-0.005600 (0.008450)
Origine 2006	0.170036 (0.020486)
Origine 2007	0.264756 (0.027000)
...	...
Origine 2011	0.866718 (0.044291)
Origine 2012	1.226765 (0.050321)
Reportyear 2006	-0.161966 (0.020554)
Reportyear 2007	-0.251812 (0.027088)
...	...
Reportyear 2011	-0.826576 (0.044167)
Reportyear 2012	-1.018915 (0.050128)
Province BC	0.400014 (0.193278)
Province MB	0.147044 (0.133016)
...	...
Province SK	-0.184729 (0.007886)
Province YT	-0.011988 (0.034812)
Cause du sinistre Collision	0.177725 (0.568397)
Cause du sinistre Collision with anything	-0.249535 (0.179623)
...	...
Cause du sinistre Windshield repair	-0.410132 (0.012030)
Cause du sinistre Windshield replace	-0.402998 (0.011416)
Développement 2	-1.058987 (0.004013)
Développement 3	-2.688908 (0.008635)
...	...
Développement 7	-5.066346 (0.048111)
Développement 8	-5.446187 (0.083421)
ϕ	2.254641

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.13: Les coefficients et leur écart-type de la sévérité ***modèle G4.***

Coefficient	Estimé
Ordonnée à l'origine	7.661774 (0.047535)
Origine 2006	-0.165287 (0.100476)
Origine 2007	-0.427500 (0.128547)
...	...
Origine 2011	-1.618182 (0.201696)
Origine 2012	-1.922737 (0.235126)
Reportyear 2006	0.148185 (0.100741)
Reportyear 2007	0.420765 (0.128882)
...	...
Reportyear 2011	1.562775 (0.200147)
Reportyear 2012	1.940469 (0.232891)
Province BC	1.254687 (0.432239)
Province MB	-0.004392 (0.553283)
...	...
Province SK	-0.330288 (0.054211)
Province YT	0.026005 (0.184830)
Cause du sinistre Collision	-0.182603 (3.854534)
Cause du sinistre Collision with anything	-0.830245 (1.461108)
...	...
Cause du sinistre Windshield repair	-0.390454 (0.084066)
Cause du sinistre Windshield replace	-0.039647 (0.068974)
Développement 2	0.289810 (0.023147)
Développement 3	1.282784 (0.031187)
...	...
Développement 7	2.529105 (0.089979)
Développement 8	2.695937 (0.142200)
ϕ	196974.5

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.14: Les coefficients et leur écart-type de la sévérité ***modèle A1.***

Coefficient	Estimé
Ordonnée à l'origine	8.129057 (0.041093)
Origine 2006	-0.260820 (0.089117)
Origine 2007	-0.680699 (0.112840)
...	...
Origine 2011	-2.229419 (0.172950)
Origine 2012	-2.707747 (0.202485)
Reportyear 2006	0.254728 (0.089405)
Reportyear 2007	0.674746 (0.113109)
...	...
Reportyear 2011	1.967536 (0.172155)
Reportyear 2012	2.536899 (0.200272)
Province BC	1.385582 (0.382584)
Province MB	0.099754 (0.489313)
...	...
Province SK	-0.454564 (0.048580)
Province YT	-0.081488 (0.163492)
Cause du sinistre Collision	-0.235844 (3.409149)
Cause du sinistre Collision with anything	-0.723285 (1.292288)
...	...
Cause du sinistre Windshield repair	0.003313 (0.061735)
Cause du sinistre Windshield replace	0.751185 (0.086437)
Développement 2	-0.103460 (0.020965)
Développement 3	0.759666 (0.028893)
...	...
Développement 7	2.031436 (0.080311)
Développement 8	1.835041 (0.126258)
ϕ	154078.2

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.15: Les coefficients et leur écart-type de la sévérité ***modèle A2.***

Coefficient	Estimé
Ordonnée à l'origine	8.053321 (0.046205)
Origine 2006	-0.126652 (0.098211)
Origine 2007	-0.417134 (0.125037)
...	...
Origine 2011	-1.481999 (0.195233)
Origine 2012	-1.860477 (0.229433)
Reportyear 2006	0.110697 (0.098530)
Reportycar 2007	0.414714 (0.125413)
...	...
Reportyear 2011	1.413696 (0.194179)
Reportyear 2012	1.830843 (0.226750)
Province BC	1.607998 (0.422331)
Province MB	0.101816 (0.540042)
...	...
Province SK	-0.484109 (0.053012)
Province YT	0.008385 (0.180397)
Cause du sinistre Collision	0.272959 (3.762150)
Cause du sinistre Collision with anything	-0.857328 (1.426060)
...	...
Cause du sinistre Windshield repair	-0.721333 (0.082017)
Cause du sinistre Windshield replace	-0.421975 (0.067616)
Développement 2	0.266280 (0.023103)
Développement 3	1.456014 (0.031210)
...	...
Développement 7	2.487080 (0.088140)
Développement 8	2.272062 (0.139004)
ϕ	187639.5

Note : l'écart-type des coefficient est entre parenthèses

Tableau A.16: Les coefficients et leur écart-type de la sévérité ***modèle A3.***

Coefficient	Estimé
Ordonnée à l'origine	7.90191 (0.04221)
Origine 2006	-0.26851 (0.09046)
Origine 2007	-0.56478 (0.11496)
...	...
Origine 2011	-1.81120 (0.17808)
Origine 2012	-1.83748 (0.20825)
Reportyear 2006	0.26380 (0.09072)
Reportyear 2007	0.56562 (0.11525)
...	...
Reportyear 2011	1.54529 (0.17702)
Reportyear 2012	1.74710 (0.20598)
Province BC	1.71087 (0.38868)
Province MB	0.06857 (0.49680)
...	...
Province SK	-0.27890 (0.04908)
Province YT	0.02408 (0.16596)
Cause du sinistre Collision	-0.16762 (3.46099)
Cause du sinistre Collision with anything	-0.81070 (1.31191)
...	...
Cause du sinistre Windshield repair	-0.46304 (0.07569)
Cause du sinistre Windshield replace	-0.08526 (0.06211)
Développement 2	0.24289 (0.02106)
Développement 3	1.24063 (0.02897)
...	...
Développement 7	2.19078 (0.08137)
Développement 8	1.97648 (0.12809)
ϕ	158800.6

Note : l'écart-type des coefficient est entre parenthèses

RÉFÉRENCES

Antonio, K. et Plat, R. (2013). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal, 2014*(7), 649–669.

Arjas, E. (1989). The claims reserving problem in non-life insurance : Some structural ideas. *ASTIN Bulletin, 19*(2), 139–152.

Boucher, J.-P. et Davidov, D. (2011). On the importance of dispersion modeling for claims reserving : An application with the tweedie distribution. *Variance, 5*(2), 158–172.

Breiman, L., Friedman, J., Stone, C. et Olshen, R. (1984). *Classification and Regression Trees.* The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

Charpentier, A. et Pigeon, M. (2016). Macro vs. micro methods in non-life claims reserving (an econometric perspective). *Risks, 4*(2), 12.

Denuit, M. et Trufin, J. (2017). Beyond the tweedie reserving model : The collective approach to loss development. *North American actuarial journal, 21*(4), 611–619.

England, P. D. et Verrall, R. J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal, 8*(3), 443–518.

Haastrup, S. et Arjas, E. (1996). Claims reserving in continuous time ; a nonparametric bayesian approach. *ASTIN Bulletin, 26*(2), 139–164.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Larsen, C. R. *et al.* (2007). An individual claims reserving model. *ASTIN Bulletin, 37*(01), 113–132.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin, 23*(2), 213–225.

Mack, T. (1999). The standard error of chain ladder reserve estimates : Recursive calculation and inclusion of a tail factor. *ASTIN Bulletin*, *29*(2), 361–366.

Messenger, R. et Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American statistical association*, *67*(340), 768–772.

Morgan, J. N. et Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, *58*(302), 415–434.

Nelder, J. A. et Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society : Series A (General)*, *135*(3), 370–384.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance 1. *ASTIN Bulletin*, *23*(1), 95–115.

Norberg, R. (1999). Prediction of outstanding liabilities. *ASTIN Bulletin*, *29*(1), 5–25.

Schiegl, M. (2002). On the safety loading for chain ladder estimates : A Monte Carlo simulation study. *ASTIN Bulletin*, *32*(1), 107–128.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, *61*(3), 439–447.

Wüthrich, M. V. (2003). Claims reserving using tweedie's compound Poisson model. *ASTIN Bulletin*, *33*(2), 331–346.

Wuthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 1–16.

Wüthrich, M. V. et Merz, M. (2008). *Stochastic claims reserving methods in insurance*, volume 435. John Wiley & Sons.

Zhao, X. et Zhou, X. (2010). Applying copula models to individual claim loss reserving methods. *Insurance : Mathematics and Economics*, *46*(2), 290–299.

Zhao, X. B., Zhou, X. et Wang, J. L. (2009). Semiparametric model for prediction of individual claim loss reserving. *Insurance : Mathematics and Economics*, *45*(1), 1–8.