UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÉLISATION GRANULAIRE DES RÉSERVES EN ASSURANCES INCENDIE, ACCIDENTS ET RISQUES DIVERS AVEC UNE STRUCTURE EN COMPOSANTES HIÉRARCHIQUES.

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN MATHÉMATIQUES

PAR

JUAN SEBASTIAN YANEZ

JUILLET 2023

REMERCIEMENTS

Je tiens à remercier tout d'abord mon directeur de recherche, Mathieu Pigeon, pour la confiance qu'il m'a accordée en acceptant de superviser mon doctorat. J'ai eu le privilège de bénéficier de ses compétences, sa rigueur et surtout sa gentillesse tout au long de mes études. En effet, sa patience et son soutien ont été primordiaux pour l'aboutissement de cette thèse et je lui en suis extrêmement reconnaissant.

Je remercie aussi la Chaire Co-operators en Analyse des Risques Actuariels, d'avoir financé mes études et de m'avoir donné accès aux bases de données qui ont été utilisées dans ce projet. Sans cet appui, il aurait été impossible de faire mes études doctorales.

Par ailleurs, je remercie tous les membres de la chaire pour leur soutien moral et leur compagnie. En particulier, je remercie Jean-Philippe Boucher, titulaire de la chaire, pour les conseils et l'expertise qu'il a apporté lors de la rédaction du deuxième article scientifique.

Je transmets aussi mes remerciements à l'ensemble des membres du jury : Montserrat Guillén, Emiliano Valdez et Jean-Philippe Boucher qui m'ont l'honneur d'évaluer mon travail.

À titre plus personnel, j'adresse mes remerciements à ma famille, particulièrement à mes parents, Juan Pablo et Amelia ainsi qu'à ma sœur Carolina, qui ont toujours été d'un grand soutien depuis le début de mes études universitaires. D'autre part, je remercie aussi mes amis et ma partenaire, qui m'ont accompagné lors des longues heures de rédaction. Gracias por siempre creer en mí y por darme su apoyo incondicional.

En mémoire de mon grand-père Germanico Pinto,
qui a fini son doctorat en sciences actuarielles en 1979.

TABLE DES MATIÈRES

REMERCIEMENTS i

LISTE DES FIGURES vii

LISTE DES TABLEAUX xi

RÉSUMÉ xv

INTRODUCTION 1

CHAPITRE I MODÉLISATION PARAMÉTRIQUE DE LA DURÉE, DE LA FRÉQUENCE ET DE LA SÉVÉRITÉ AU NIVEAU MICRO POUR LES RÉSERVES GRANULAIRES 11

1.1 Introduction 11

1.2 Statistical model 17

1.2.1 Duration component 18

1.2.2 The frequency component 23

1.2.3 The severity component 26

1.3 Loss reserves 28

1.3.1 IBNR reserve 28

1.3.2 RBNS reserve 30

1.4 Numerical Analysis 32

1.4.1 Data set 32

1.4.2 Fitting the collective models 36

1.4.3 Fitting the three-component model 36

1.4.4 Fitting an individual model based on a Poisson Process 42

1.4.5 Goodness of fit analysis 43

1.4.6 Outstanding reserve discussion 50

1.5 Conclusion 55

1.6 Appendix : Database examples 56

CHAPITRE II MODÉLISATION DE LA FRÉQUENCE DES PAIEMENTS DES RÉSERVES EN FONCTION D'UN SCORE DYNAMIQUE DE SINISTRE 59

2.1 Introduction 59

2.2 Statistical framework 63

2.2.1 Introductory notation 63

2.2.2 *A priori* distribution of the number of payments 65

2.2.3 *A posteriori* distribution of the number of payments 66

2.2.4 Payment categories and IBNR specifications for claim-score modelling 71

2.2.5 Distribution of duration of claims 72

2.2.6 Parameter estimation 73

2.3 Simulation procedure 74

2.3.1 *IBNR* simulation procedure 75

2.4 Numerical results 79

2.4.1 Data Set 79

2.4.2 Fitting the models 82

2.4.3 Goodness-of-fit analysis 83

2.4.4 Simulation analysis 87

2.5 Conclusion 92

2.6 Appendix 97

CHAPITRE III ANALYSE DE LA DÉPENDANCE DU TEMPS DE TRAITEMENT DES RÉCLAMATIONS : UNE PERSPECTIVE BASÉE SUR DES EFFETS ALÉATOIRES *FRAILTY* 107

3.1 Introduction 107

3.2 Preliminary numerical analysis 112

3.2.1 Data set description 112

3.2.2 Correlation analysis 115

3.3 Statistical Framework 118
3.3.1 The Cox proportional hazards model 121
3.3.2 The Weibull-Cox proportional hazards model 123
3.4 Frailty 125
3.4.1 Frailty for dependent coverages 125
3.4.2 Estimation of the Frailty based on claim history 126
3.4.3 Parameter estimation 131
3.5 Numerical results 136
3.5.1 Fitting the Models 136
3.5.2 Goodness-of-fit 137
3.5.3 Simulation analysis 139
3.5.4 Conclusion 142
3.6 Appendix 143
CONCLUSION 151

LISTE DES FIGURES

Figure		Page
0.1	Développement de trois réclamations	3
1.1	Development of two claims	12
1.2	Incremental loss development triangle	17
1.3	Observed delays for claims at valuation date in a loss triangle	19
1.4	Observed exposures ($E_{\ell,k}$) and observed payments ($N_{\ell,k}$), based on vectors $\mathcal{D}_\ell$ for two claims	26
1.5	Gender	34
1.6	Region	34
1.7	Vehicle age	34
1.8	Injured age	34
1.9	Reporting delay	34
1.10	Initial Reserve	34
1.11	Type of loss	34
1.12	Heat maps of ratios of observed cumulative values to fitted cumulative values for the three component model (in order, from up to down, the exposure, the number of payments and the cost)	47
1.13	Heat maps of ratios of observed values to fitted values for the gamma collective model	48
1.14	Plot of residuals of the cumulative exposure by accident, development and calendar year	48
1.15	Plot of residuals of the cumulative number of payments by accident, development and calendar year	49

1.16 Plot of residuals of the cumulative cost by accident, development and calendar year 49

1.17 Total reserve distributions until December 31, 2017 51

1.18 Total reserve distributions until December 31, 2019 51

2.1 Development of two claims 64

2.2 Dynamic claim score development (with $\psi = 2$) 75

2.3 Relativity of the dynamic risk score to the mean of medical RBNS payments 86

2.4 Relativity of the dynamic risk score to the mean of disability RBNS payments 86

2.5 Relativity of the dynamic risk score to the mean of expense RBNS payments 91

2.6 Total reserves for selected models 91

3.1 Development of three coverages 120

3.2 Distributions of the outstanding exposure. 141

3.3 1^{st} to 2^{nd} coverages 144

3.4 1^{st} to 3^{rd} coverages 144

3.5 1^{st} to 4^{th} coverages 144

3.6 1^{st} to 5^{th} coverages 144

3.7 1^{st} to 6^{th} coverages 144

3.8 2^{nd} to 3^{rd} coverages 144

3.9 2^{nd} to 4^{th} coverages 145

3.10 2^{nd} to 5^{th} coverages 145

3.11 2^{nd} to 6^{th} coverages 145

3.12 3^{rd} to 4^{th} coverages 145

3.13 3^{rd} to 5^{th} coverages 145

3.14 3^{rd} to 6^{th} coverages 145

3.15 4^{rd} to 5^{th} coverages 146

3.16 4^{rd} to 6^{th} coverages 146

3.17 5^{rd} to 6^{th} coverages 146

LISTE DES TABLEAUX

Tableau		Page
1.1	Categorical variables description	33
1.2	Full development triangle of the observed total cost	35
1.3	Estimated values	39
1.4	Fitted intensity of the event Poisson process	43
1.5	Observed intensity of the event Poisson process	44
1.6	Predicted and observed number of payments after the valuation date	44
1.7	AIC and BIC criteria for models with different covariates across all components	45
1.8	Likelihood Ratio ($L.R.$) test for models with different covariates across all components	45
1.9	Results of the total reserve predictions until December 31, 2017	50
1.10	Results of the total reserve predictions until December 31, 2019	52
1.11	Outstanding cumulative payments of four claims at the end of each development year (j)	54
1.12	Example of a duration training set	56
1.13	Example of a frequency training set (with $\boldsymbol{\delta}= \{0, 1, 2, 3, 4, 5\}$)	57
1.14	Example of a severity training set (with $\boldsymbol{\delta}= \{0, 1, 2, 3, 4, 5\}$)	57
2.1	Description of covariates	80
2.2	Claim frequency descriptive statistics for each category	81
2.3	Likelihood Information Criteria for RBNS models **M1**,**M2** and **M3**	84
2.4	Likelihood Information Criteria for IBNR models **M1**,**M2** and **M3**	85

2.5 AIC and BIC of RBNS models with and without the claim score . 87

2.6 AIC and BIC of IBNR models with and without the claim score . 88

2.7 AIC and BIC of no covariate models with and without the claim score 89

2.8 Likelihood Ratio (L. R.) test RBNS and IBNR models with and without the dynamic claim score 90

2.9 Student's t-test for parameter $\gamma^{(a)}$ for RBNS models with the dynamic claim score 92

2.10 Simulation results for RBNS outstanding payment counts from models with and without claim scores 94

2.11 Simulation results for the total outstanding payment counts from models with and without claim scores 95

2.12 Results of the total reserve predictions 96

2.13 Estimated values for the Negative Binomial (type II) Model (RBNS) 98

2.14 Estimated values for the Poisson Model (RBNS) 101

2.15 Estimated values for the Negative Binomial (type II) Model (IBNR) 104

2.16 Estimated values for the Poisson Model (IBNR) 105

3.1 Description of covariates 113

3.2 Claim cluster size in terms of coverage count 115

3.3 Spearman's ρ and Kendall's τ rank statistics 117

3.4 Wald tests for the variance of the Frailty (random effect) 138

3.5 AIC and BIC criteria for the parametric models 139

3.6 Likelihood Cross-Validation (LCV) criterion for the M-spline models 139

3.7 Likelihood Ratio tests for Cox and Weibull-Cox models 139

3.8 Results of the predictions for the total outstanding exposure $\left(E^{*(O)}\right)$ 140

3.9 Estimated values of the covariate parameters 147

3.10 p-values of the t-tests for the covariate parameters 149

RÉSUMÉ

Dans le cadre de la science actuarielle *Incendie, Accidents et Risques Divers* (I.A.R.D.), le calcul des réserves est primordial pour garantir le remboursement des engagements futurs d'un assureur envers ses assurés et pour estimer la solvabilité de la compagnie d'assurance. Traditionnellement ce montant est calculé à travers des méthodes nommées *collectives* qui agrègent les paiements futurs en fonction de la date de survenance du sinistre qui a déclenché la réclamation et la date des paiements. Or, malgré l'avantage de transformer la base données sous une forme plus simple, les méthodes collectives ont le désavantage de ne pas pouvoir incorporer de l'information plus pointue sur la réclamation dans la modélisation. Dans le but d'utiliser le plus d'information possible des bases de données qui deviennent de plus en plus riches en information, plusieurs auteurs se sont intéressé à des méthodes nommées *individuelles* (ou granulaires). En effet, les modèles individuels sont ajustés sur des données non agrégées, de sorte que le développement individuel non observé des réclamations est complété pour chaque réclamation ouverte du portefeuille.

La présente thèse est une collection de contributions à la littérature des réserves granulaires sous la forme de trois articles. Le thème principal qui relie ces trois propositions est l'incorporation de l'information individuelle. Plus précisément, on propose des modèles originaux pour les différents éléments qui composent le développement d'une réclamation : la *durée* de celle-ci, la *fréquence* des paiements et leurs coût (ou *sévérité*). On met en avant des méthodes pour que l'actuaire puisse utiliser les caractéristiques de réclamations sous forme de variables explicatives à chaque élément du développement. De suite, on ajuste les différent modèles à une base de données riche en information pour pouvoir mesurer la qualité de l'ajustement des modèles ainsi que de la qualité de l'information utilisée. De plus, on compare nos propositions avec d'autres modèles collectifs et individuels de la littérature afin de montrer sur quels aspects ils sur-performent leurs contreparties.

Mots Clés— Réserves individuelles, Modèles de survie, Modèles linéaires généralisés, Valeurs Extrêmes, Bonus Malus

INTRODUCTION

Un des rôles importants d'un actuaire dans le cadre d'une compagnie d'assurances *Incendie, Accidents et Risques Divers* (I.A.R.D.) est le calcul de la réserve. Ce montant est composé du total des paiements futurs pour les sinistres survenus avant une date donnée dans le portefeuille d'un assureur. Ainsi, il permet de garantir le remboursement total des assurés qui ont subit un sinistre pour lequel tous les paiements n'ont pas encore été faits. Précisément, la nature non observée de la réserve fait en sorte qu'il est nécessaire d'incorporer des modèles pour sa prédiction. De plus, cette prédiction doit être rigoureuse parce qu'elle permet non seulement de protéger les assurés, mais aussi d'estimer la solvabilité de la compagnie d'assurance à une date donnée.

Plus spécifiquement, lorsqu'un sinistre couvert par une police se produit, une série d'évènements est déclenchée. Tout d'abord, le sinistre, qui sera ultérieurement associé à une réclamation ℓ, survient et on peut immédiatement identifier $t_\ell^{(o)}$, le délai entre de début de l'année d'accident et la *date de survenance* de l'accident (*occurrence date*). Après, à la suite d'un délai ($t_\ell^{(r)}$) qui est souvent très court, l'assuré déclare le sinistre à l'assureur à une *date de déclaration*. Ainsi, à partir de cette date, la réclamation ℓ est présente dans le portefeuille et l'assureur a accès à des informations sur la nature du sinistre et une série de remboursements (*cash flows*) s'ensuit au fur et à mesure du développement de la réclamation. Finalement, la *date de fermeture* (*settlement date*) a lieu, mettant fin au développement de la réclamation et permettant de définir le délai entre la déclaration et la fermeture d'une réclamation par $t_\ell^{(c)}$. On suppose généralement dans la littérature qu'une ré-ouverture du dossier n'est pas possible.

À une certaine *date d'évaluation*, l'actuaire de la compagnie peut classer les réclamations présentes dans un portefeuille en fonction du stade atteint par le développement de chacune de celles-ci. On commence par distinguer les réclamations fermées pour lesquelles $t_\ell^{(e)}$, c'est-à-dire le délai entre la date de déclaration et la date d'évaluation, est supérieur au délai de fermeture, $t_\ell^{(c)}$. Par la suite, lorsque la date de fermeture ne s'est pas encore produite, $t_\ell^{(c)} > t_\ell^{(e)}$, on dit que la réclamation est déclarée mais pas encore fermée, ou *Reported But Not Settled* (RBNS). Finalement, il est possible de considérer les réclamations qui sont survenues avant la date d'évaluation mais dont la date de déclaration ne s'est pas encore produite. Les données sur ces réclamations dites encourues mais non déclarées, ou *Incurred But Not Reported* (IBNR), ne sont pas accessibles puisque l'assureur n'a pas été informé de leur existence. La Figure 0.1 représente graphiquement un exemple d'une réclamation ayant atteint chacun de ces trois statuts. Sur celle-ci, la ligne pointillée et la ligne solide représentent, respectivement, le délai de déclaration ($t_\ell^{(r)}$) et le délai de fermeture ($t_\ell^{(c)}$). De plus, les petits cercles symbolisent les paiements.

En résumé, le développement d'une réclamation est composé du,

— délai d'occurrence : $t_\ell^{(o)}$,
— délai de déclaration : $t_\ell^{(r)}$,
— délai d'évaluation : $t_\ell^{(e)}$,
— délai de fermeture : $t_\ell^{(c)}$.

Pour un portefeuille, la réserve est constituée par la somme des paiements futurs associés à chacun des sinistres dont la date de survenance est antérieure à la date d'évaluation. Le rôle de l'actuaire est de prédire cette réserve afin de permettre à la compagnie de mettre de côté ce montant. Après avoir décrit le développement des réclamations I.A.R.D., il est possible de séparer la réserve totale en fonc-

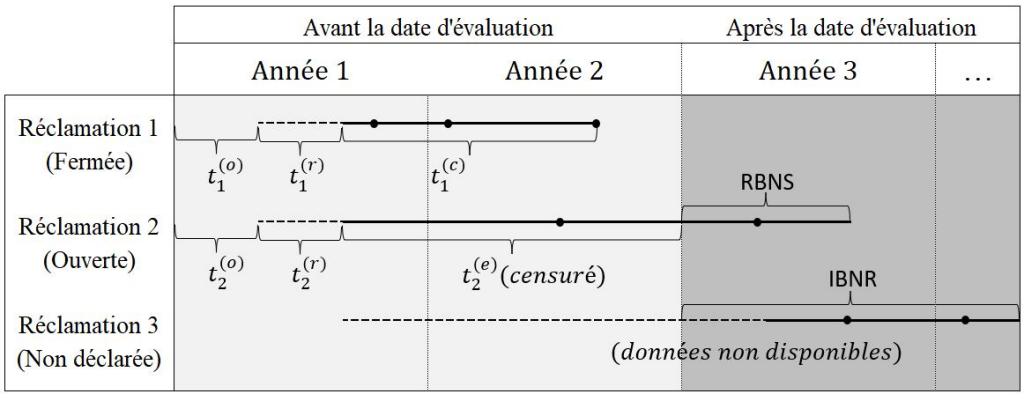

FIGURE 0.1 Développement de trois réclamations

tion du statut des réclamations considérées dans le calcul. Premièrement, on a la réserve RBNS qui englobe les paiements produits par des réclamations connues par l'assureur à la date d'évaluation. Pour cette catégorie, l'actuaire a accès à de l'information partielle constituée de la portion observée du développement, des informations sur le sinistre, des informations sur l'assuré, etc. Deuxièmement, on considère la réserve IBNR qui constitue une problématique différente puisqu'aucune information n'est disponible sur ces réclamations. De plus, le nombre de réclamations ayant ce statut est également inconnu.

À cause de leur importance, les réserves ont engendré de nombreuses contributions dans la littérature scientifique. Traditionnellement, les paiements présents dans un portefeuille sont agrégés en fonction de deux éléments : l'année de survenance de la réclamation dont ils proviennent et le nombre d'années qui se sont écoulées entre l'année de survenance et le paiement en question (connu sous le nom d'année de développement). Ceci permet de créer une structure triangulaire qui résume les remboursements faits par la compagnie d'assurance. Les modèles qui travaillent directement sur ces triangles de développement sont connus sous le nom de *méthodes collectives* ou modèles au niveau macro (*macro-level models*). Par exemple, le populaire modèle Chain Ladder et sa version stochastique (voir

Mack (1999) et Mack (1993)) sont des membres importants de cette branche de la littérature. De plus, plusieurs autres propositions ont été faites, par exemple des applications des modèles linéaires généralisés, ou *Generalized Linear Models* (GLM). Puisque dresser un portrait complet de cette branche de la littérature demanderait de nombreuses pages et éloignerait le lecteur du propos principal de cette thèse, on recommande de consulter les travaux de Wüthrich & Merz (2008) et de England & Verrall (2002) pour un survol des méthodes collectives.

La popularité de la structure triangulaire provient du fait qu'elle résume les informations sous un format intuitif. Cela conduit à des méthodes qui peuvent facilement être diffusées, telles que le modèle Chain Ladder. En effet, les triangles de développement résument l'évolution du portefeuille et peuvent facilement être annexés dans les rapports annuels ou les états financiers de la compagnie. De plus, cette structure compacte permet de mettre en place des méthodes moins exigeantes en temps et et puissance de calcul. Cependant, il n'est pas possible d'inclure des informations individuelles puisque les paiements des différentes réclamations sont agrégés en un seul montant. Ainsi, l'accès à des bases de données plus détaillées au cours des deux plus récentes décennies a déclenché un intérêt pour une autre branche de la littérature : les *modèles granulaires* ou modèles au niveau micro (*micro-level models*). Ces modèles n'agrègent pas les données et leur but est plutôt de compléter le développement de chacune des réclamations RBNS et de simuler l'entièreté du développement de chacune des réclamations IBNR. De cette façon, il est possible de profiter des informations individuelles puisque chaque réclamation ouverte est complétée séparément.

Les premières propositions de modèles granulaires ont été faites lors de la dernière décennie du siècle précédent par Arjas (1989), Norberg (1993), Haastrup & Arjas (1996) et Norberg (1999). Dans ces propositions, un processus de poisson avec marqueurs, ou *Dependent Marked-Poisson Process* (DMPP), a été utilisé pour

prédire les paiements et les fermetures des réclamations. À l'exception, notable, de ces quelques exemples, l'intérêt envers cette branche de la littérature ne s'est vraiment développé que de nombreuses années plus tard. En effet, l'augmentation phénoménale de la puissance de calcul et l'accès croissant à des bases de données plus détaillées au cours des vingt dernières années ont contribué à faire croître la popularité des approches granulaires. En particulier, en 2014, l'article Antonio & Plat (2014) reprend les premières propositions de modèles DMPP en les adaptant à une base de données. Il s'agit la, fort probablement, de la première mise en oeuvre concrète d'un modèle appartenant à cette catégorie. Ce modèle a par la suite été développé dans l'article Antonio *et al.* (2015) où une structure Markovienne avec des états interchangeables permet de prendre en compte le développement passé d'une réclamation dans la prédictions des évènements futurs.

En parallèle avec le développement des DMPP, d'autres processus semi paramétriques ont été considérés comme le processus de Cox. La première mention a eu lieu en 2009 avec l'article de Zhao *et al.* (2009) où les auteurs ont traité des réserves IBNR et, plus tard, des copules ont été incorporées pour prendre en compte la dépendance dans Zhao *et al.* (2010). Dans le cadre du processus de Cox, il est important de mentionner Badescu *et al.* (2016) et Badescu *et al.* (2019), où une application a été proposée pour compter le nombre de réclamations IBNR à la date d'évaluation. Aussi, plus tard, Avanzi *et al.* (2021) a incorporé un choc commun au modèle de Cox pour capturer la dépendance entre les réclamations. Par ailleurs, mentionnons Maciak *et al.* (2021) où le processus de Hawkes avec des intensités qui varient au cours du temps a été proposé.

Dans les années plus récentes, l'article Wüthrich (2018) a déclenché un intérêt pour des méthodes non paramétriques basées sur des techniques issues de l'apprentissage machine, ou *machine learning*. Dans cette première proposition un arbre de décision a été adapté pour prédire les flux monétaires des réclamations.

D'autres modèles basés sur la même technique ont suivit tels que Lopez *et al.* (2016), Lopez (2019) et Lopez *et al.* (2019). En parallèle d'autre suggestions ont été faites, par exemple avec l'algorithme *ExtraTrees* par Baudry & Robert (2019) ou la procédure *Gradient Boosting* par Duval & Pigeon (2019). En outre, d'autres auteurs se sont intéressés à l'application de technique de réseau de neurones pour modéliser les réserves granulaires, par exemple Kuo (2019), Gabrielli (2020) et Delong *et al.* (2022). Enfin, on doit aussi mentionner le travail fait par Blier-Wong *et al.* (2020) (section 4) qui détaille plusieurs applications d'apprentissage machine dans le contexte des réserves.

En contraste avec les autres propositions dans la littérature, certaines méthodes ont plutôt considéré des structures facilement comparables aux méthodes collectives traditionnelles. Par exemple, des articles tels que Pigeon *et al.* (2013) et Pigeon *et al.* (2014) proposent des applications qui adaptent des notions des modèles classiques, telles que les facteurs de développement du modèle Chain Ladder, au contexte des réserves granulaires. De plus, dans les articles Huang *et al.* (2015a), Huang *et al.* (2015b), Huang *et al.* (2016) et Charpentier & Pigeon (2016), la structure de temps discrète des modèles agrégés a été adaptée au contexte individuel dans le but de comparer les approches micro et macro. En effet, les auteurs ont constaté que les modèles individuels ont un avantage numérique envers ces contreparties à cause de la perte d'information des modèles collectifs. De plus, les résultats de l'article de Wang *et al.* (2021) indiquent que l'ajout d'information pointue sur les réclamations donne un avantage statistiquement significatif aux méthodes individuelles.

Cette thèse cherche à faire une contribution scientifique à la modélisation des réserves granulaires en exploitant un des avantages les plus importants de ce type de modèles : l'accès à l'information individuelle. On cherche à présenter des modèles originaux dans lesquels les actuaires pourront aisément incorporer des variables

explicatives qui caractérisent chacune des réclamations. En effet, ces variables viennent sous plusieurs formes et peuvent poser plusieurs défis à la fois théoriques et pratiques. D'un point de vue théorique, on décrit des nouveaux outils statistiques adaptés à la modélisation des réserves granulaires et, d'un point de vue pratique, on cherche à démontrer que l'information individuelle sur laquelle les modèles reposent est significative et qu'elle permet aux modèles de mieux performer par rapport à des modèles plus traditionnels.

Étant donné que le thème principal de cette thèse est l'information individuelle des réclamations, un aspect important de chaque article est la description détaillée de l'information à prendre en compte. Une fois la problématique mise en place, on présente des modèles capables d'incorporer les différents types de données considérées. Par la suite, on fait une description détaillée des méthodes d'ajustement et des algorithmes de simulation afin de permettre une application directe des modèles. De plus, grâce à l'accès à une base de données riche en information, on est capable de mesurer la qualité des modèles proposés ainsi que la qualité de l'information. Les analyses numériques sont faites de façon minutieuses en utilisant des outils statistiques pour mesurer la qualité de l'ajustement et l'importance des variables explicatives. On complète les analyses avec des comparaisons numériques avec des modèles collectifs et individuels populaires dans la littérature, en indiquant les avantages apportés par nos propositions. Concrètement, cette thèse est composée de trois projets axés sur l'inclusion des caractéristiques des réclamations dans le cadre du calcul des réserves.

Dans un premier article déjà publié Yanez & Pigeon (2021), on met en place une structure hiérarchique qui sépare le développement des réclamations en trois composantes : la durée, la fréquence et la sévérité. Cette modélisation permet à l'actuaire d'utiliser des modèles paramétriques qui peuvent facilement intégrer de l'information individuelle à chacune des étapes. La durée apparaît sous trois

formes, le délai de survenance ($t_\ell^{(o)}$), le délai de déclaration ($t_\ell^{(r)}$) et le délai de fermeture ($t_\ell^{(c)}$). Pour cette étape, des modèles classiques de survie peuvent être utilisés, par exemple la distribution Weibull ou la distribution Log-normale. La fréquence est calculée sur des intervalles entre la date de déclaration et la date de fermeture. Ainsi, la modélisation du nombre de paiement(s) par intervalle permet de prendre en compte une mesure d'exposition qui indique combien de temps la réclamation est ouverte à l'intérieur de chacun des intervalles. Ceci permet l'utilisation de modèles de comptage classiques tels que le modèle Poisson (surdispersé) ou le modèle basé sur la distribution binomiale négative. Finalement, on prend en considération le coût de chaque paiement qui demande de faire appel à la modélisation de valeurs extrêmes, voir Denuit & Trufin (2017) et Laudagé *et al.* (2019).

Les variables explicatives qui décrivent une réclamation peuvent être de nature différentes : statiques, dynamiques déterministes ou dynamiques non-déterministes. Les variables statiques ne changent pas avec le temps, contrairement aux autres types de variables. On peut distinguer les deux types de variables dynamiques par le fait que les variables dynamiques déterministes peuvent être prédites avec certitude (par exemple, l'âge d'un assuré) alors que les variables dynamiques non-déterministes ne le peuvent pas (par exemple, l'évolution d'une blessure). Le premier article est très adéquat pour incorporer l'information issue de variables statiques et de variables dynamiques déterministes mais la mise en oeuvre de l'incorporation des variables dynamiques non-déterministes a été faite dans un second projet Yanez *et al.* (2023). En effet, il existe des variables de ce type qui peuvent avoir un impact important dans la prédiction des réserves granulaires. La variable dynamique non-déterministe d'intérêt dans ce projet est le nombre de paiement(s) observé dans le passé. Ainsi, on utilise la structure en intervalle développée dans Yanez & Pigeon (2021) pour créer une nouvelle variable explicative nommée *score*

de réclamation dynamique qui résume le développement préalablement observé et se met à jour à la fin de chacun des intervalles. Cette méthode est inspirée du modèle Bonus-Malus couramment utilisé en tarification, voir Lemaire (1995), Boucher & Pigeon (2019) et Boucher (2023).

Enfin, on a décidé, dans un troisième projet Yanez & Pigeon (2023), d'étudier plus en détail la modélisation de la durée et, en particulier, du délai de fermeture. Dans ce cadre, on a remarqué que dans certaines situations, une réclamation peut affecter plusieurs couvertures qui peuvent avoir des caractéristiques différentes. Ce qui fait en sorte que l'incorporation de ces caractéristiques ne peut pas se faire directement, sauf si on modélise la durée de chaque couverture séparément. Dans ce cas, une problématique de dépendance se pose puisque les couvertures touchées partagent la même origine. Afin de considérer cette dépendance, on propose de modéliser la durée des couvertures en ajoutant un effet aléatoire commun aux membres de chaque réclamation (connu sous le nom de *frailty* dans la littérature de survie). En plus de décrire en détail ce type de modèles, on donne des outils pour que l'actuaire puisse mesurer cette dépendance dans sa base de données.

Cette thèse est composée de trois chapitres principaux. Le Chapitre 1 décrit le modèles à trois composantes et les enjeux principaux associés à celui-ci. Dans le Chapitre 2, la méthode du *score de réclamation dynamique* pour les paiements passés est développée. Par la suite, la problématique de la dépendance entre les couvertures est traitée au Chapitre 3. Finalement, on fait une conclusion sur les résultats principaux de la thèse ainsi que sur les possibles extensions pour des projets futurs.

CHAPITRE I

MODÉLISATION PARAMÉTRIQUE DE LA DURÉE, DE LA FRÉQUENCE ET DE LA SÉVÉRITÉ AU NIVEAU MICRO POUR LES RÉSERVES GRANULAIRES

1.1 Introduction

Non-life insurance companies must control their solvency in order to protect their policyholders. Therefore, a provision or, loss reserve, must be established for claims whose total amount has not been paid or fully paid. Given the importance and the complexity of this task, several models have been proposed in the actuarial literature to predict future payments and to evaluate associated risks. Traditionally, these models can be grouped into two categories, collective and individual, based on the underlying data set. Although collective models have been studied by researchers for a long time and are commonly used by practitioners, individual models have caught the eye of researchers in the more recent years, and are rarely put into practice despite their many advantages. In this paper we aim to provide a parametric framework that can use micro-level information, which in turn may shed light on the advantage of using this information.

Let us begin by looking at the typical development of two claims, as illustrated in Figure 1.1. When accident ℓ occurs, we can identify the delay between the beginning of the accident year and the exact occurrence date ($t_\ell^{(o)}$). After an

additional delay ($t_\ell^{(r)}$), claim ℓ is declared. For several situations (fire, damage to a car, etc.), this second delay may be short, but for other situations (bodily injury, civil liability), a longer period can separate the occurrence and reporting of a claim. Subsequently, one or more payments may be made (illustrated by dots in the figure) before closing the file after a final delay ($t_\ell^{(c)}$). At an valuation date, claims can be separated into several categories according to the information available. For the remainder of this paper, the main categories are as follows :

— if the valuation date is between the date of the accident and the reporting date, the loss is considered not reported (*Incurred But Not Reported* or, IBNR), meaning the actuary has no information about the claim ; and

— if the valuation date is between the reporting date and the closing date of the case, the loss is considered reported but not closed (*Reported But Not Settled* or, RBNS) meaning the actuary has only partial information about the claim.

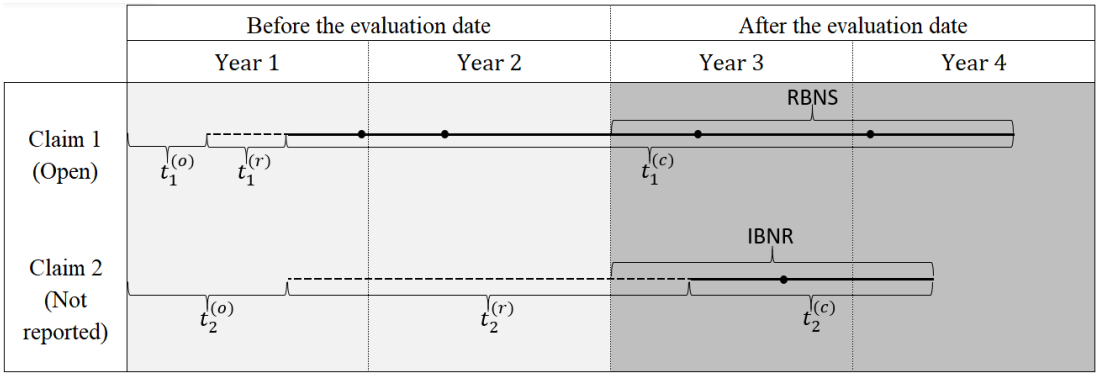

FIGURE 1.1 Development of two claims

In the literature, most stochastic collective models for loss reserving are presented in the widely used run-off triangles framework. This representation summarizes claim payments by aggregating them based on the accident and development year. In recent decades, intuitive and popular methods such as the stochastic Chain

Ladder model (Mack (1999) and Mack (1993)) have been developed. Particularly, Wüthrich & Merz (2008) and England & Verrall (2002) made a compilation of the most widely used models within this collective structure. However, one of the shortcomings of these methods is that the underlying data used to summarize payments are based on the development of many different claims in spite of their individual characteristics, making it very difficult to use micro-level information in the modeling process.

In contrast to this class of models, individual models do not require aggregating payments from different claims ; instead each claim is analysed separately. Several of the recently proposed approaches are based on techniques from the field of statistical learning. For examples, Wüthrich (2018), Lopez *et al.* (2016); Lopez (2019); Lopez *et al.* (2019); Lopez *et al.* (2019) suggested using regression trees to predict the number of outstanding claim payments. Another implementation for predicting loss reserves was suggested in Baudry & Robert (2019), using an algorithm called *ExtraTrees*. Duval & Pigeon (2019) recommend using a Gradient Boosting procedure to predict both the IBNR and RBNS reserves. The examples allow the use of micro-level information for their predictions.

In parallel, some authors have put forth parametric models. For example, Haastrup & Arjas (1996) proposed a Position Dependent Marked-Poisson Process (PDMPP) to predict claim payments continuously. In 2014, Antonio & Plat (2014) presented a similar but more elaborated model, and the authors successfully applied it to a real data set. Although Haastrup & Arjas (1996) suggested the possibility of using micro-level information in their model, Antonio & Plat (2014) only considered individual covariates for payments severity. Antonio *et al.* (2015) suggested an individual multi-state approach and applied it to a real data set. They also mention that micro-level information other that the occurrence and development period can be incorporated in their model. Furthermore, Zhao *et al.*

(2009) developed a semi-parametric model for individual claims, and Zhao *et al.* (2010) incorporated copulae to predict IBNR claims. In contrast, some authors have considered a discrete framework for claim development. For example, Pigeon *et al.* (2013) and Pigeon *et al.* (2014) considered individual development factors. Verrall *et al.* (2010) and, later Huang *et al.* (2015b), examined modeling claim counts and claim amounts separately, deriving a frequency-severity type structure.

In this paper, we propose an individual parametric model that fully takes advantage of micro-level information to predict outstanding claim payments. In retrospect, we drew inspiration from the two-component framework (frequency-severity), which is often used in ratemaking for Property & Causalty insurance. In this pricing context, general linear models or, GLM, are often used along with micro-level information (contract), to predict each component (for more information, see Ohlsson *et al.* (2010)). Moreover, an exposure measure is considered for the frequency, usually in the form of the duration of the insurance contracts. However other forms of exposure may be considered (kilometers driven Ayuso *et al.* (2019), etc.) In a loss reserving context, we considered a similar frequency-severity GLM framework for outstanding claims using micro-level information. We also establish an exposure measure based on the duration of each claim. However, unlike the contracts in ratemaking models, where the exposure measure is known beforehand, the duration of open claims is unknown at the valuation date. Therefore, an additional model needed to be fitted for this new component. We propose using parametric survival models that use micro-level information for this step, mainly to maintain a fully parametric structure similar to the one used for the aforementioned components, which in turn allows for a similar analysis of the significance of covariates in the fitting process across all steps. Note that in contrast with the studies by Verrall *et al.* (2010) and Huang *et al.* (2015b), which also contain a frequency-severity type of structure, we seek to model single

payments instead of the total cost of claims.

Comparatively to the existing literature for parametric individual loss reserving, our model diverges from methods previously suggested, such as models that use development factors, as in Pigeon *et al.* (2013) and Pigeon *et al.* (2014), or models that use Poisson Processes, as in Antonio & Plat (2014) and Zhao *et al.* (2009). Indeed, instead of making use of a Poisson Process to model the delays between payments and closure of claims simultaneously, we suggest a more hierarchical structure in which we model first the duration of the claim, through survival models, and then the frequency, through discrete time parametric modeling for claim payment counts. To the best of our knowledge this is the first individual parametric loss reserving model that suggests modeling these two type of events separately, which in turn allows us to introduce the possibility of modeling payment counts discretely instead of continuously. This new structure offers flexibility in the choice of the distributions for the both the duration and the frequency, all while offering a straightforward way of introducing micro-level information from claims. Furthermore, modeling the cost of single payments have already been studied by Antonio & Plat (2014), who recommend Burr, Lognormal and generalized linear models, and Denuit & Trufin (2017) who favored a mixture of Gamma and Pareto distribution. However, we make a contribution in the modeling of this component as well by suggesting a new model based on splices, drawing inspiration from Laudagé *et al.* (2019).

In a nutshell, we propose, in this paper

— a new 3-component framework for outstanding claim payments, where the unknown exposure is based on the duration and where micro-level information can be included at all levels (duration-frequency-severity) ;

— a model structure that allows for a comparison with both collective approaches based on run-off triangles (e.g., Mack's model), and with other individual approaches proposed in the literature (e.g., individual model proposed in Antonio & Plat (2014)).

In this paper we aim to implement the use of claims covariate information in the prediction of loss reserves. However, we must consider the different kinds of covariates that are available to us. Taylor *et al.* (2008) use this information for their individual loss reserve model and, consider three types of covariates : static (such as the region), time dynamic (such as the age of the beneficiary), and, unpredictable dynamic (such as the health condition of the beneficiary). Static covariates, do not change as time passes, however dynamic covariates will. Furthermore, even though both dynamic types are affected by time, only time dynamic covariates can be predicted with certainty, which in turn makes unpredictable dynamic covariates delicate to work with. An additional model is required in order to predict these uncertain values after the valuation date. In this paper, we will consider only static and time-dependent dynamic covariates because the additional model required for unpredictable dynamic covariates could be very specific, depending on which variable we are looking at.

This paper is be structured as follows. In Section 1.2, we look at the general framework of the three-component model. In Section 1.3, we discuss the simulation procedure of the IBNR and RBNS reserves. In Section 1.4, we describe the data set used, followed by the numerical results of both our model and other comparative models. Finally, Section 1.5 contains concluding remarks and mentions further topics that could be explored based on this work.

1.2 Statistical model

In this section, we define both the individual and collective perspectives of a given portfolio, because we want our model to be interpreted from both perspectives.

On the one hand, in a micro-level structure, let $\mathcal{L} = \mathcal{L}^{(O)} \cup \mathcal{L}^{(C)}$, represent a set containing L reported claims in a portfolio, where $\mathcal{L}^{(O)}$ and $\mathcal{L}^{(C)}$ are the subsets containing open (RBNS) and closed claims, respectively. Let $\mathcal{L}^*$ be the set containing incurred but not reported (IBNR) claims, which is, obviously, unavailable at the valuation date.

On the other hand, in a macro-level structure, let i and j be, respectively, the occurrence and the development periods in a run-off triangle, or loss triangle. Also, let $Y_{i,j}$ be the total paid amount between time $i-1+j$ and $i+j$ from claims occurring during period i, where $i = 1, \ldots, I$ and $j = 0, \ldots, (I-1)$. For example, Figure 1.2 illustrates an incremental loss triangle with five occurrence and development periods.

		Development Period (j)				
		0	1	2	3	4
Occurrence Period (i)	1	$Y_{1,0}$	$Y_{1,1}$	$Y_{1,2}$	$Y_{1,3}$	$Y_{1,4}$
	2	$Y_{2,0}$	$Y_{2,1}$	$Y_{2,2}$	$Y_{2,3}$	$\hat{Y}_{2,4}$
	3	$Y_{3,0}$	$Y_{3,1}$	$Y_{3,2}$	$\hat{Y}_{3,3}$	$\hat{Y}_{3,4}$
	4	$Y_{4,0}$	$Y_{4,1}$	$\hat{Y}_{4,2}$	$\hat{Y}_{4,3}$	$\hat{Y}_{4,4}$
	5	$Y_{5,0}$	$\hat{Y}_{5,1}$	$\hat{Y}_{5,2}$	$\hat{Y}_{5,3}$	$\hat{Y}_{5,4}$

FIGURE 1.2 Incremental loss development triangle

Let us suppose that the insurance company has additional details about the accident, the insured, etc. and, wants to use them in the modeling process. Furthermore, let us suppose that all the covariates become available as soon as the claim

is reported. The information regarding one observed claim ℓ can be summarized by g categorical and/or continuous covariates,

$$\mathbf{c}_\ell = [c_{\ell,1}, c_{\ell,2}, \ldots, c_{\ell,g}], \text{ for } \ell \in \mathcal{L}.$$

Having defined these variables, we can now consider the three components of the model in detail. We present the duration component in Subsection 1.2.1, the frequency component in Subsection 1.2.2, and the severity component in Subsection 1.2.3. To better illustrate how each of these components could be obtained from a real data set, we provide examples in Appendix 1.6.

1.2.1 Duration component

The duration component can be defined as the delay between the beginning of the occurrence period and the closure of a given claim. Thus, for a claim ℓ, the component is constructed from the three following parts :

— $T_\ell^{(o)}$ the occurrence delay, i.e. the time elapsed between the beginning of the occurrence year and the exact occurrence date ;
— $T_\ell^{(r)}$ the declaration delay, i.e. the time elapsed between the exact occurrence date and the reporting date ; and
— $T_\ell^{(c)}$ the closure delay, i.e. the time elapsed between the reporting date and the closure date.

Because the claims we consider at are either open, closed or non-reported, Figure 1.3 represents how the delays are observed at the valuation date for claims having the same occurrence period at these three stages of development in a loss triangle.

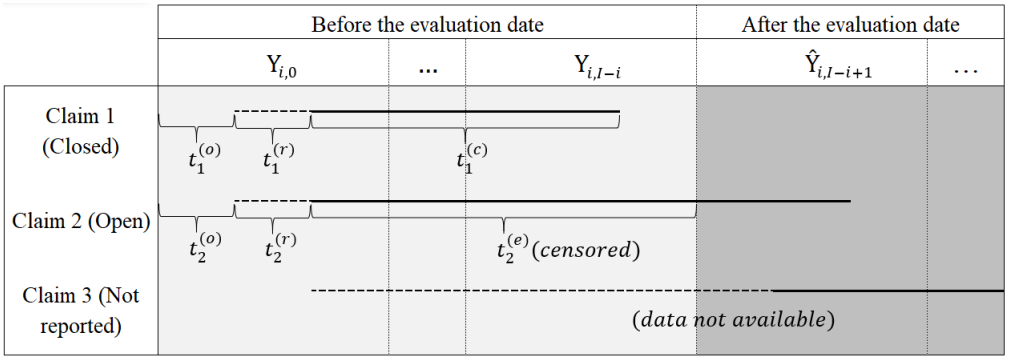

FIGURE 1.3 Observed delays for claims at valuation date in a loss triangle

Recall $t_\ell^{(o)}$ is the delay from the beginning of the occurrence period. We have covariate information about open claims, and we also know the full extent of the occurrence and declaration delays. Furthermore, for $\ell \in \mathcal{L}^{(O)}$, we only have partial information about the closure delay, in the form of a right-censored observation, $t_\ell^{(e)}$, the delay between the reporting date and the valuation date. These values coupled with the observed values of closed claims (which are similar except that the closure delay is not a censored observation) constitute the full extent of the training data set. Moreover, not reported claims are undisclosed to the insurer at the valuation date and therefore cannot be included in the training set.

In this paper, we suppose that no claim can reopen after it has been closed for the first time. However, this model can include reopening events by simply adding a variable, the delay between the closure date and a later reopening date. Given the assumption, only open and not reported claims need a reserve to be calculated. Furthermore, the predictions for the future payments of open and not reported claims constitute the full extent of the RBNS and the IBNR reserves respectively. In fact, the prediction of claims at these two stages of development bring about different challenges in the modeling process of the different delays and are explained separately.

To begin, we need to find a model for the closure delay of open claims,

$$\left(T_\ell^{(c)}|\mathbf{c}_\ell\right), \text{ for } \ell \in \mathcal{L}^{(O)}. \tag{1.1}$$

However, at the valuation date these delays have been partially observed. Therefore, we need to find the conditional distribution,

$$\left(T_\ell^{(c)}|T_\ell^{(c)} > t_\ell^{(e)}, \mathbf{c}_\ell\right), \text{ for } \ell \in \mathcal{L}^{(O)}. \tag{1.2}$$

Nevertheless, depending on the chosen distribution, and the number of covariates, making simulations from the conditional distribution (1.2) can be simplified by simulating from (1.1), and then using an acceptance-rejection method to keep only the delays greater than $t_\ell^{(e)}$. For this reason, in this subsection we focus on modeling (1.1), and in Section 1.3, we further explain the simulation procedure of the conditional delays.

We suggest using parametric survival models for this step. Chapter 1 of the book by Lawless (2011) contains some of the distributions that can be considered, among them the Weibull, Log-Logistic, Lognormal and Gamma distributions. Recall that the training set contains non-censored and right-censored observations of the closure delay. For a parametric survival distribution, the likelihood function for the closure delay is

$$\begin{aligned}\Lambda^{(RBNS)}\left(\mathbf{\Theta}_{T^{(c)}}\right) = &\prod_{\ell\in\mathcal{L}^{(C)}}\left\{f_{\left(T_\ell^{(c)}|\mathbf{c}_\ell\right)}\left(\left(t_\ell^{(c)}|\mathbf{c}_\ell\right);\mathbf{\Theta}_{T^{(c)}}\right)\right\}\times \\ &\prod_{\ell\in\mathcal{L}^{(O)}}\left\{S_{\left(T_\ell^{(c)}|\mathbf{c}_\ell\right)}\left(\left(t_\ell^{(e)}|\mathbf{c}_\ell\right);\mathbf{\Theta}_{T^{(c)}}\right)\right\},\end{aligned} \tag{1.3}$$

where, $f()$ and $S()$, are the probability density function (pdf) and the survival function of the closure delay for reported claims, respectively and, $\mathbf{\Theta}_{T^{(c)}}$ is the parameter vector.

IBNR claims are more complex to predict because the insurer does not have any information about them. They only know that some probably have occurred, and they will be reported at some point after the valuation date. Therefore, the number of IBNR claims is also unknown and, must be predicted. In Section 1.3, we suggest a procedure to predict how many not reported claims have occurred at each period i, however this particular model is better explained after defining the duration component for this type of claim. Thus, in this subsection, we suppose that i_ℓ, the occurrence period of claim ℓ, is known. Moreover, this section concerns the different time measures of the model, i.e. we focus on modeling the delays of a single not reported claim $\ell^* \in \mathcal{L}^*$, while supposing that the occurrence period is known.

Unlike open reported claims, we need to predict all three delays, instead of just completing the closure delay. Admittedly, as mentioned in the introduction of this paper, one of the advantages of micro-level models is that we can use individual information better than macro-level models. This is very much true for RBNS claims however its less so for IBNR claims. In spite of lack of micro-level information available when modeling non reported claims, individual models, unlike collective models, have the advantage of being able to predict IBNR and RBNS reserves separately, allowing the insurer to have insight about the weight of not reported claims in the portfolio.

The first delay we need to model is the occurrence delay, $T_\ell^{(o)}$, using $\ell \in \mathcal{L}$. The key aspect of this delay is that its observation period is $(0, 1]$, i.e. a year. Moreover, because we are modeling unreported claims, micro-level information is unknown

and thus, we suggest fitting models based on seasonal effects, for example. Nevertheless, this time measure varies greatly depending on the data set, and many models can be considered. Examples of distributions that can be examined are the multinomial distribution, assigning a probability of a claim occurring within a certain time-window (e.g. each month of a year) or, the empirical distribution.

The other delay that needs to be modeled is the reporting delay, $T_\ell^{(r)}$, for which many observations generally occur within only a few days after the occurrence date. In order to take into account this large number of short time observations and those that take more time, Antonio & Plat (2014) suggest using a mixture of a Weibull distribution with D degenerate components, where each component d, represents the number of days that have passed since the occurrence date, $d = 0, 1, \ldots, D-1$ days. Furthermore, contrarily to the observed closure delay, the reporting delay does not contain censored observations, but the delays are truncated by the valuation date. Thus, the likelihood of the reporting delay can be written as

$$\Lambda^{(IBNR)}\left(\boldsymbol{\Theta}^*_{T^{(r)}}\right) = \prod_{\ell \in \mathcal{L}} \frac{f^*_{T_\ell^{(r)}}\left(t_\ell^{(r)}; \boldsymbol{\Theta}^*_{T^{(r)}}\right)}{F^*_{T_\ell^{(r)}}\left(I - i_\ell + 1 - t_\ell^{(o)}; \boldsymbol{\Theta}^*_{T^{(r)}}\right)},$$

where, $f^*()$ and $F^*()$ are the pdf and the cumulative distribution function of the reporting delay for not reported claims, respectively and, $\boldsymbol{\Theta}^*_{T^{(r)}}$ is the parameter vector. Having modeled the reporting and the occurrence delay, let $U_\ell^{(r)}$ be the delay between the beginning of the occurrence period and the report date, thus,

$$U_\ell^{(r)} = T_\ell^{(o)} + T_\ell^{(r)}.$$

By definition, the report date of an IBNR claim must happen after the valuation

date. Moreover, for claim $\ell \in \mathcal{L}^*$, we can obtain the delay between the beginning of the occurrence year and the valuation date, $I - i_\ell + 1$. Thus,

$$U_\ell^{(r)} > I - i_\ell + 1, \text{ for } \ell \in \mathcal{L}^*.$$

This in turn means that we need to find the distribution of $\left(U_\ell^{(r)} | U_\ell^{(r)} > I - i + 1\right)$, for $i = 1, \ldots, I$. We suggest obtaining these distributions numerically through a simulation of both $T_\ell^{(o)}$ and $T_\ell^{(r)}$. More details are given in Section 1.3.

The final delay that needs to be addressed is the closure delay, $T_\ell^{(c)}$, where the same parametric distributions can be used in a similar manner to the the one suggested in Subsection 1.2.1, with the only difference being that most of the explanatory variables are missing, thus the likelihood to model the closure delay of IBNR in the training set is defined as :

$$\begin{aligned}\Lambda^{(IBNR)}\left(\mathbf{\Theta}^*_{T^{(c)}}\right) = &\prod_{\ell \in \mathcal{L}^{(C)}} \left\{ f^*_{\left(T_\ell^{(c)} | i_\ell\right)} \left(\left(t_\ell^{(c)} | i_\ell \right) ; \mathbf{\Theta}^*_{T^{(c)}} \right) \right\} \times \\ &\prod_{\ell \in \mathcal{L}^{(O)}} \left\{ S^*_{\left(T_\ell^{(c)} | i_\ell\right)} \left(\left(t_\ell^{(e)} | i_\ell \right) ; \mathbf{\Theta}^*_{T^{(c)}} \right) \right\}, \end{aligned} \tag{1.4}$$

where, $f^*()$ and $S^*()$, are the pdf and the survival function of the closure delay for not reported claims, respectively and, $\mathbf{\Theta}^*_{T^{(c)}}$ is the parameter vector.

1.2.2 The frequency component

Payments are time-framed by the reporting and closure dates because they can happen only between these two events (closure delay). In this subsection we aim to define a partition of this timeline in order to count the number of payments within each of the sub-intervals. We build this partition with two main goals in

mind : (1) allow the structure of a run-off triangle to be easily reconstructed from our model, and (2) capture variations in individual development as a function of the time elapsed since the reporting date of each claim.

Let $\mathcal{Q} = \{0, 1, \ldots, I\}$ be a partition of the time interval between the beginning of the period in which the claim is reported and the maximum development time in a run-off triangle. This first division allows us to easily draw parallels between our results and those obtained using a collective approach based on a triangular structure.

We define a second partition of the same time interval which, this time, will not be constrained by the regularity of the construction of the run-off triangle, i.e. the fact that a loss triangle is generally divided into development periods of one year. We aim to capture the individual development of the frequency of claims from the report date until its closure, and to identify, precisely, what stage of development each time division is in. This type of division is reminiscent of that used in the position dependent marked-Poisson process model proposed by Antonio and Plat Antonio & Plat (2014).

Thus, let us begin by noting that the observation period of claim ℓ, starting from the reporting date, is $[0, \tau_\ell]$, where $\tau_\ell = \max\left\{t_\ell^{(c)}, t_\ell^{(e)}\right\}$. Let τ be the longest possible observation period, such that $\tau = \max_{\ell \in \mathcal{L}} \{\tau_\ell\}$. Let $\boldsymbol{\delta} = \{\delta_0, \delta_1, \ldots, \delta_{K-1}, \delta_K\}$ a partition of the interval $[0, \tau]$, where $\delta_0 = 0$ and $\delta_K = \tau$. Then let $\mathcal{P}_\ell = \left\{0, \boldsymbol{\delta} + u_\ell^{(r)}, I\right\}$. We assume that $\tau < I - u_\ell^{(r)}$, $\forall \ell$, but it is easy to adapt this definition if this inequality is not satisfied. Finally, we define the common refinement of $\mathcal{Q}$ and $\mathcal{P}_\ell$, which consists of all the points of $\mathcal{Q}$ and $\mathcal{P}_\ell$:

$$\mathcal{D}_\ell = \mathcal{Q} \vee \mathcal{P}_\ell = \{D_{\ell,0}, D_{\ell,1}, \ldots, D_{\ell,M_\ell}\},$$

where $D_{\ell,0} = 0$ and M_ℓ is the number of sub-intervals of $\mathcal{D}_\ell$.

The exposure corresponding to the k^{th} sub-interval of $\mathcal{D}_\ell$ is

$$E_{\ell,k} = \begin{cases} D_{\ell,k} - D_{\ell,k-1}, & k \in \{k : k > 0, D_{\ell,k} < \tau_\ell + u_\ell^{(r)}\} \\ \tau_\ell + u_\ell^{(r)} - D_{\ell,k-1}, & k \in \{k : D_{\ell,k-1} < \tau_\ell + u_\ell^{(r)} \leq D_{\ell,k}\} \\ 0, & \text{elsewhere.} \end{cases} \tag{1.5}$$

Let $N_{\ell,k}$ be a random variable which counts the number of payments for the claim ℓ during the k^{th} sub-interval. For $\ell \in \mathcal{L}$, the distribution of each $N_{\ell,k}$, is given by

$$(N_{\ell,k}|\mathbf{c}_\ell, \mathcal{D}_\ell, E_{\ell,k}) \sim Dist^{(n)}\left(E_{\ell,k} \cdot \mu_{\ell,k}^{(n)}\left(\boldsymbol{\beta}^{(n)}\right), \cdot\right), \text{ if } E_{\ell,k} > 0, \text{ and} \tag{1.6}$$

$$(N_{\ell,k}|\mathcal{D}_\ell, E_{\ell,k}) \sim Dist^{*(n)}\left(E_{\ell,k} \cdot \mu_{\ell,k}^{*(n)}\left(\boldsymbol{\beta}^{*(n)}\right), \cdot\right), \text{ if } E_{\ell,k} > 0, \tag{1.7}$$

with $j = 0, \ldots, I-1$ and $k = 1, \ldots, M_\ell$. $Dist^{(n)}$ and $Dist^{*(n)}$ are the distributions of the number of payments from reported and unreported claims, respectively. Also, $\mu_{\ell,k}^{(n)}\left(\boldsymbol{\beta}^{(n)}\right)$ and $\mu_{\ell,k}^{*(n)}\left(\boldsymbol{\beta}^{*(n)}\right)$ are the mean parameters for one exposure unit, while, $\boldsymbol{\beta}^{(n)}$ and $\boldsymbol{\beta}^{*(n)}$ are the parameter vectors used to predict the mean. We explicitly mention j in our construction to ease comparison with collective approaches. If obtaining results in the form of a loss triangles is not needed, it is possible to simplify the model and to keep only the second partition.

Finally, having defined both the exposure and the number of payments for each claim ℓ at each k^{th} sub-interval, we can now depict these values visually for two claims in Figure 1.4. As in Figure 1.3 the reporting and closure delays are represented by a dotted and solid lines respectively, while dots represent payments.

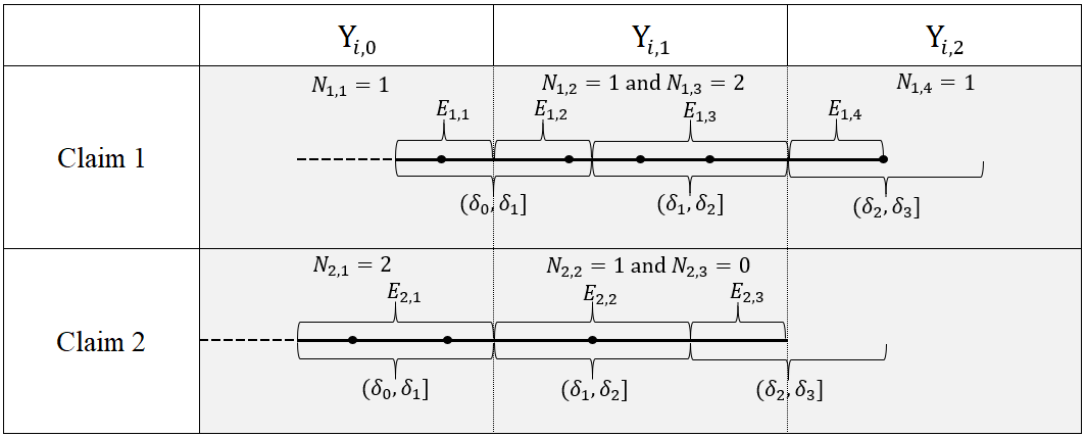

FIGURE 1.4 Observed exposures ($E_{\ell,k}$) and observed payments ($N_{\ell,k}$), based on vectors $\mathcal{D}_\ell$ for two claims

1.2.3 The severity component

With the final component we seek to predict the cost of single payment. Let $Dist^{(x)}$ and $Dist^{*(x)}$ be the distributions of the cost of single payments from reported and not reported claims respectively. Let $X_{\ell,k,m}$ be the cost of the m^{th} payment from claim ℓ occurring during the k^{th} sub-interval of $\mathcal{D}_\ell$. We have, for $\ell \in \mathcal{L}$,

$$(X_{\ell,k,m}|\mathbf{c}_\ell, \mathcal{D}_\ell) \sim Dist^{(x)}, \text{ for } m = 1, \ldots, N_{\ell,k}, \text{ and} \tag{1.8}$$

$$(X_{\ell,k,m}|\mathcal{D}_\ell) \sim Dist^{*(x)}, \text{ for } m = 1, \ldots, N_{\ell,k}. \tag{1.9}$$

Although it is possible to consider generalized linear models, where the distribution is continuous on $\mathbb{R}^+$, modeling the severity of payments may require a more complex approach. This is due to the fact that payments may be highly diverse and, a model that can accommodate both large and small payments could be preferable. Antonio & Plat (2014), for example, suggest using models such as the

Burr and the Lognormal distributions. Alternatively, one may consider a mixture of distributions, for example Denuit & Trufin (2017) suggest a discrete mixture of a Gamma and a Pareto distribution.

Finally, a splicing model can also be considered, of which some of the best known are the threshold models. Laudagé *et al.* (2019) advocate using this method in the context of claim severity for rate making and this model can be accommodated to model claim payments instead. This so-called *Threshold severity model* can better accommodate large payments by fitting the tail and the body of the distributions separately through a splicing point u, called a threshold. For a given payment, $(X_{\ell,k,m}|\mathbf{c}_\ell, \mathcal{D}_\ell)$, let $h_{\left(X_{\ell,k,m}\right)}\left(x_{\ell,k,m};\boldsymbol{\Theta}_{X^{(h)}}\right)$ and $g_{\left(X_{\ell,k,m}\right)}\left(x_{\ell,k,m};\boldsymbol{\Theta}_{X^{(g)}}\right)$ be the probability density functions of the bulk and the tail, respectively, with parameter vectors $\boldsymbol{\Theta}_{X^{(h)}}$ and $\boldsymbol{\Theta}_{X^{(g)}}$. Let $H_{\left(X_{\ell,k,m}\right)}$ and $G_{\left(X_{\ell,k,m}\right)}$ be the respective cumulative distribution functions. The probability of exceeding the splicing point is given by $q_{\left(X_{\ell,k,m}\right)}\left(\boldsymbol{\Theta}_{X^{(q)}}\right)$, where $\boldsymbol{\Theta}_{X^{(q)}}$ is the parameter vector. Thus, the pdf of $(X_{\ell,k,m}|\mathbf{c}_\ell, \mathcal{D}_\ell)$ with parameter vector $\boldsymbol{\Theta}_X = (\boldsymbol{\Theta}_{X^{(h)}}, \boldsymbol{\Theta}_{X^{(g)}}, \boldsymbol{\Theta}_{X^{(q)}})$ is given by

$$f_{\left(X_{\ell,k,m}\right)}\left(x_{\ell,k,m};\boldsymbol{\Theta}_X\right) =$$
$$\begin{cases} 0, & \text{for } x_{\ell,k,m} \leq 0 \\ \left(1 - q_{\left(X_{\ell,k,m}\right)}\left(\boldsymbol{\Theta}_{X^{(q)}}\right)\right) \dfrac{h_{\left(X_{\ell,k,m}\right)}\left(x_{\ell,k,m};\boldsymbol{\Theta}_{X^{(h)}}\right)}{H_{\left(X_{\ell,k,m}\right)}\left(u;\boldsymbol{\Theta}_{X^{(h)}}\right)}, & \text{for } 0 < x_{\ell,k,m} \leq u \\ q_{\left(X_{\ell,k,m}\right)}\left(\boldsymbol{\Theta}_{X^{(q)}}\right) g_{\left(X_{\ell,k,m}\right)}\left(x_{\ell,k,m};\boldsymbol{\Theta}_{X^{(g)}}\right), & \text{for } x_{\ell,k,m} > u, \end{cases}$$

for $m = 1, \ldots, N_{\ell,k}$.

1.3 Loss reserves

In this section we illustrate how to simulate both the IBNR and the RBNS reserves after fitting the three-component model defined in Section 1.2. For payments that happen after the valuation date, for $(i + j > I)$, we have

$$\begin{aligned} Y_{i,j} &= Y_{i,j}^{(RBNS)} + Y_{i,j}^{(IBNR)} \\ &= \sum_{\ell \in \left\{\ell | \ell \in \mathcal{L}^{(O)}, i_\ell = i\right\}} \sum_{k \in \left\{k : j < D_{\ell,k} \leq j+1\right\}} \sum_{m=1}^{N_{\ell,k}} X_{\ell,k,m} \\ &\quad + \sum_{\ell \in \{\ell | \ell \in \mathcal{L}^{*}, i_\ell = i\}} \sum_{k \in \left\{k : j < D_{\ell,k} \leq j+1\right\}} \sum_{m=1}^{N_{\ell,k}} X_{\ell,k,m}. \end{aligned}$$

We recall that i_ℓ is the occurrence period of claim ℓ. Then, we calculate the total reserve :

$$\begin{aligned} R &= R^{(RBNS)} + R^{(IBNR)} \\ &= \sum_{i+j>I} Y_{i,j}^{(RBNS)} + \sum_{i+j>I} Y_{i,j}^{(IBNR)}. \end{aligned}$$

We can now describe the simulation procedure for both parts of the total reserve.

1.3.1 IBNR reserve

Before we give the complete simulation procedure, we must predict the number of IBNR claims in the portfolio. Our approach is based on the work of Pigeon *et al.* (2013) and on the distribution of $U_\ell^{(r)}$ as defined in Subsection 1.2.1. Let L_i, the total number of claims occurring during period i, follow a Poisson distribution

with occurrence measure $\theta\omega_i$, where ω_i is the total exposure registered for period i, for $i = 1, \ldots, I$. Because we only observe reported claims, the Poisson distribution should be thinned in the following way

$$L_i \sim \text{Poisson}\left(\theta\omega_i \Pr\left[U_\ell^{(r)} \leq I - i + 1\right]\right).$$

Thus, L_i^*, the number of IBNR claim(s) from occurrence period i follows a Poisson distribution with occurrence measure given by

$$\theta\omega_i \Pr\left[U_\ell^{(r)} > I - i + 1\right]. \tag{1.10}$$

We can now proceed to the simulation procedure of an IBNR reserve.

- **Step 1 :** Obtain $\widetilde{L}^* = \sum_i \widetilde{L}_i^*$,where $\widetilde{L}_i^*$ is the simulated value of L_i^* for each occurrence period (see Equation (1.10)).
- **Step 2 :** For $\ell = 1, \ldots, \widetilde{L}^*$, go through each of the following sub-steps.
  - **Step 2a :** Obtain $\widetilde{U}_\ell^{(r)}$, the simulated value of $\left(U_\ell^{(r)} | U_\ell^{(r)} > I - i_\ell + 1\right)$, the delay between the beginning of the occurrence period and the exact reporting date, where

$$\Pr\left[U_\ell^{(r)} \leq u | U_\ell^{(r)} > I - i_\ell + 1\right] = \frac{\Pr\left[I - i_\ell + 1 < U_\ell^{(r)} \leq u\right]}{1 - \Pr\left[U_\ell^{(r)} \leq I - i_\ell + 1\right]}.$$

  - **Step 2b :** Obtain $\widetilde{T}_\ell^{(c)}$, the simulated value of $\left(T_\ell^{(c)} | i_\ell\right)$, the closure delay (see Equation (1.4)), where

$$\left(T_\ell^{(c)} | i_\ell\right) \sim Dist^{*(t^{(c)})}.$$

  - **Step 2c :** Based on $\widetilde{U}_\ell^{(r)}$, $\widetilde{T}_\ell^{(c)}$ and $\boldsymbol{\delta}$, calculate $\widetilde{\mathcal{P}}_\ell$ and $\widetilde{\mathcal{D}}_\ell = \mathcal{Q} \vee \widetilde{\mathcal{P}}_\ell = \{0, \widetilde{D}_{\ell,1}, \ldots, \widetilde{D}_{\ell,\widetilde{M}_\ell}\}$.

— **Step 2d :** Calculate

$$\widetilde{E}_{\ell,k} = \begin{cases} \widetilde{D}_{\ell,k} - \widetilde{D}_{\ell,k-1}, & k \in \{k : k > 0, \widetilde{D}_{\ell,k} < \widetilde{T}_{\ell}^{(c)} + \widetilde{U}_{\ell}^{(r)}\} \\ \widetilde{T}_{\ell}^{(c)} + \widetilde{U}_{\ell}^{(r)} - \widetilde{D}_{\ell,k-1}, & k \in \{k : \widetilde{D}_{\ell,k-1} < \widetilde{T}_{\ell}^{(c)} + \widetilde{U}_{\ell}^{(r)} \leq \widetilde{D}_{\ell,k}\} \\ 0, & \text{elsewhere,} \end{cases}$$

for $k = 1, \ldots, \widetilde{M_\ell}$.

— **Step 2e :** Obtain $\widetilde{N}_{\ell,k}$, a simulated value of $\left(N_{\ell,k}|\widetilde{\mathcal{D}}_\ell, \widetilde{E}_{\ell,k}\right)$, using Equation (1.7), for $k = 1, \ldots, \widetilde{M_\ell}$.

— **Step 2f :** Obtain $\widetilde{X}_{\ell,k,m}$, a simulated value of $\left(X_{\ell,k}|\widetilde{\mathcal{D}}_\ell\right)$, using Equation (1.9), for $m = 1, \ldots, \widetilde{N}_{\ell,k}$ and $k = 1, \ldots, \widetilde{M_\ell}$.

— **Step 3 :** Calculate the simulated IBNR reserve :

$$\widetilde{R}^{(IBNR)} = \sum_{i+j>I} \widetilde{Y}_{i,j}^{(IBNR)} = \sum_{i+j>I} \sum_{\ell \in \widetilde{\mathcal{L}}_i^*} \sum_{k \in \left\{k : j < \widetilde{D}_{\ell,k} \leq j+1\right\}} \sum_{m=1}^{\widetilde{N}_{\ell,k}} \widetilde{X}_{\ell,k,m},$$

where $\widetilde{\mathcal{L}}_i^*$ is the set containing all the simulated IBNR claims occurring at period i.

1.3.2 RBNS reserve

Let $L^{(O)}$ be the total number of open claims in the portfolio. We describe the simulation procedure for the RBNS reserve below.

— **Step 1 :** Set $\ell = 1$, the first open claim in $\mathcal{L}^{(O)}$.

— **Step 1a :** Obtain $\widetilde{T}_{\ell}^{(c)}$, the simulated value of $\left(T_{\ell}^{(c)}|\mathbf{c}_\ell\right)$, the closure delay of open claim ℓ (see Equation (1.3)), where,

$$\left(T_{\ell}^{(c)}|\mathbf{c}_\ell\right) \sim Dist^{(t^{(c)})}.$$

— **Step 1b :** If $\widetilde{T}_{\ell}^{(c)} > t_{\ell}^{(e)}$, set $\ell = \ell + 1$, the next open claim.

— **Step 1c :**

— If $\ell \leq L^{(O)}$, go to **Step 1a**.

— If $\ell = L^{(O)} + 1$, go to **Step 2**.

— **Step 2 :** Based on $u_\ell^{(r)}$, $\widetilde{T}_\ell^{(c)}$ and $\boldsymbol{\delta}$, calculate $\mathcal{P}_\ell$ and $\widetilde{\mathcal{D}}_\ell = \mathcal{Q} \vee \mathcal{P}_\ell = \{0, \widetilde{D}_{\ell,1}, \ldots, \widetilde{D}_{\ell,\widetilde{M}_\ell}\}$, for $\ell = 1, \ldots, L^{(O)}$.

— **Step 3 :** Calculate

$$\widetilde{E}_{\ell,k} = \begin{cases} \widetilde{D}_{\ell,k} - t_\ell^{(e)} - u_\ell^{(r)}, & k \in \mathcal{K}_1 \\ \widetilde{D}_{\ell,k} - \widetilde{D}_{\ell,k-1}, & k \in \mathcal{K}_2 \\ \widetilde{T}_\ell^{(c)} + u_\ell^{(r)} - \widetilde{D}_{\ell,k-1}, & k \in \mathcal{K}_3 \\ \widetilde{T}_\ell^{(c)} - t_\ell^{(e)}, & k \in \mathcal{K}_4 \\ 0, & \text{elsewhere,} \end{cases}$$

for $k = 1, \ldots, \widetilde{M}_\ell$ and $\ell = 1, \ldots, L^{(O)}$, where,

$$\mathcal{K}_1 = \{k : \widetilde{D}_{\ell,k-1} < t_\ell^{(e)} + u_\ell^{(r)} \leq \widetilde{D}_{\ell,k}, \widetilde{D}_{\ell,k} < \widetilde{T}_\ell^{(c)} + u_\ell^{(r)}\},$$
$$\mathcal{K}_2 = \{k : t_\ell^{(e)} + u_\ell^{(r)} \leq \widetilde{D}_{\ell,k-1}, \widetilde{D}_{\ell,k} < \widetilde{T}_\ell^{(c)} + u_\ell^{(r)}\},$$
$$\mathcal{K}_3 = \{k : t_\ell^{(e)} + u_\ell^{(r)} \leq \widetilde{D}_{\ell,k-1}, \widetilde{D}_{\ell,k-1} < \widetilde{T}_\ell^{(c)} + u_\ell^{(r)} \leq \widetilde{D}_{\ell,k}\},$$
$$\mathcal{K}_4 = \{k : \widetilde{D}_{\ell,k-1} < t_\ell^{(e)} + u_\ell^{(r)}, \widetilde{T}_\ell^{(c)} + u_\ell^{(r)} \leq \widetilde{D}_{\ell,k}\}.$$

— **Step 4 :** Obtain $\widetilde{N}_{\ell,k}$, a simulated value of $\left(N_{\ell,k} | \boldsymbol{c}_\ell, \widetilde{\mathcal{D}}_\ell, \widetilde{E}_{\ell,k}\right)$, using Equation (1.6), for $k = 1, \ldots, \widetilde{M}_\ell$ and $\ell = 1, \ldots, L^{(O)}$.

— **Step 5 :** Obtain $\widetilde{X}_{\ell,k,m}$, a simulated value of $\left(X_{\ell,k} | \boldsymbol{c}_\ell, \widetilde{\mathcal{D}}_\ell\right)$, using Equation (1.8), for $m = 1, \ldots, \widetilde{N}_{\ell,k}$, $k = 1, \ldots, \widetilde{M}_\ell$ and $\ell = 1, \ldots, L^{(O)}$.

— **Step 6 :** Calculate the simulated RBNS reserve :

$$\widetilde{R}^{(RBNS)} = \sum_{i+j>I} \widetilde{Y}_{i,j}^{(RBNS)} = \sum_{i+j>I} \sum_{\ell \in \mathcal{L}_i^{(O)}} \sum_{k \in \left\{k : j < \widetilde{D}_{\ell,k} \leq j+1\right\}} \sum_{m=1}^{\widetilde{N}_{\ell,k}} \widetilde{X}_{\ell,k,m},$$

where $\mathcal{L}_i^{(O)}$ be the set containing all the open RBNS claims occurring at period i.

1.4 Numerical Analysis

In this section we provide a detailed analysis based on a real data set from a Canadian Property & Causalty insurance company. With this example, we want to (1) illustrate the use of our new 3-component framework, (2) perform a comparison with collective approaches, and (3) perform a comparison with another individual approach. We describe our data set in Subsection 1.4.1, we adjust our model, as well as various collective models in Subsection 1.4.2, we perform a goodness of fit analysis in Subsection 1.4.5, and finally, we obtain the results for the outstanding loss reserves in Subsection 1.4.6.

1.4.1 Data set

The data set we worked on contains transactional information for $57,593$ claims occurring between January 1, 2011 and, December 31, 2015. The insurer recorded each important event (payment, case estimates, closure date, etc.), along with micro-level information until December 31, 2017. For our numerical analysis, we set the valuation date to December 31, 2015, where there were $48,855$ closed claims, $7,872$ open claims and, 866 not reported claims.

Some payments from the data set were not considered based on macro and micro level hypothesis. First, we did not consider the payments that happen after the end of the final development period in a loss triangle, i.e. $(J = I - 1)$, for each claim. In order words we did not calculate a reserve for payments that happen after time $t = i_\ell + (I - 1), \forall \ell$. This hypothesis is often used for collective approaches based on run-off triangles. Second, we also did not consider any payment after the first closure date of every claim, making the data set consistent with the no reopening hypothesis explained in Section 1.2. Moreover, these hypotheses were assumed for our individual model and the comparative collective models, in order to have a

fair comparison between the results obtained.

For the $57,593$ claims, we only consider payments for the Accident Benefits (AB) coverages, i.e. no-fault benefits for accident where the insured or a third party were hurt or killed in a car accident. Furthermore, we have micro-level information regarding each claim, which was used in the three-component model, in the form of categorical static covariates. Table 1.1 contains a summary of these variables, and Figures 1.5- 1.9 contain the percentages of each group among all claims. Among them, the initial reserve represents the first prediction of the total future cash-flows for each claim at the reporting date.

Tableau 1.1: Categorical variables description

Variable	Label	Number of levels
Gender	Gender of the injured/killed	3
Region	Geographical region	3
Type of loss	Kind of AB claim	5
Vehicle age	Age of the vehicle	6
Injured age	Age of the injured/killed	7
Reporting delay	Delay calculated in days	7
Initial Reserve	Reserve at report date	5

All the covariates are static, and some considerations must be explained. First, regarding the type of loss, in some situations a single accident may cause different kinds of losses, therefore, some claims are dependant because they originate from the same casualty, even though most of the covariates could be different. The dependence of related claims introduces an interesting, yet complex, additional problem within the framework we developed in this paper, which will be better explored in a future project. Consequently, for this analysis we assumed independence between claims. In addition, for some claims, some covariates could

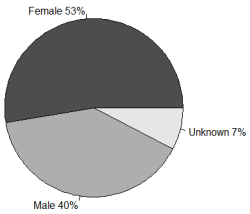

FIGURE 1.5 Gender

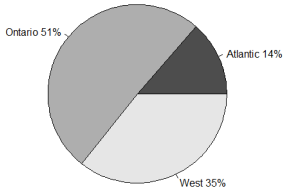

FIGURE 1.6 Region

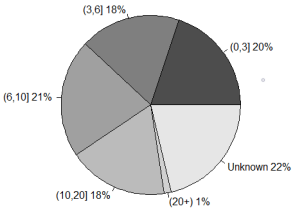

FIGURE 1.7 Vehicle age

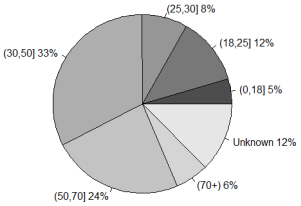

FIGURE 1.8 Injured age

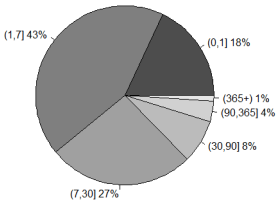

FIGURE 1.9 Reporting delay

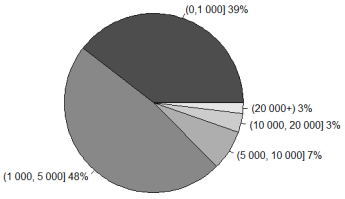

FIGURE 1.10 Initial Reserve

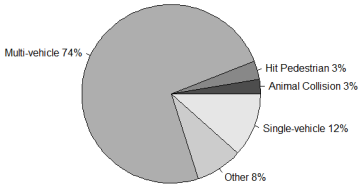

FIGURE 1.11 Type of loss

be missing (NA). We decided to keep these observations as additional categories because the number of claims with at least one unknown value is significant.

Besides the individual micro-level information for each claim, the data set also contains the exact occurrence and reporting dates, as well as the exact date and the cost of each payment up until December 31, 2017. However, at this date there are still $1,135$ open claims and thus, some data is missing (e.g. the total paid amount after this date). In order to provide a full comparison between our model and collective triangle-based approaches, we decided to complete the missing values with a Chain Ladder model that uses the latest information available. With these predicted values we can have an estimation of the total payments in 2018 and 2019, allowing us to have the full development triangle for the portfolio. Table 1.2 contains the development triangle based on the above mentioned hypothesis. The total observed reserves amount is $\$188,520,892$.

TABLEAU 1.2: Full development triangle of the observed total cost

	0	1	2	3	4
1	17,749,045	25,449,306	19,499,061	12,186,259	8,914,255
2	15,050,987	27,422,359	19,162,367	18,114,766	**8,353,106**
3	16,322,509	33,692,522	23,145,859	**18,673,999**	**12,541,295**
4	19,451,913	34,563,968	**28,254,849**	**12,360,982**	**12,180,220**
5	20,899,092	**41,120,193**	**26,399,552**	**15,939,656**	**12,697,041**

The training set consists of all the $56,727$ closed and open claims information up to the valuation date. Based on this information, we fit various collective models (see Subsection 1.4.2), our 3-component model (see Subsection 1.4.3), and the individual model based on a Poisson process suggested by Antonio & Plat (2014) (see Subsection 1.4.4).

1.4.2 Fitting the collective models

We consider four collective models based on the run-off triangle illustrated in Table 1.2. We consider two classes of approaches : stochastic Chain Ladder model, or Mack's model (see Mack (1999) and Mack (1993)), and Generalized Linear Model, or GLM, for reserves. For the first, we obtain the predictive distribution through a bootstrap procedure proposed by England & Verrall (2002) and based on the quasi-Poisson distribution (Model **Ia**) and the Gamma distribution (Model **Ib**). For the second, we use occurrence and the development periods as covariates. We considered the quasi-Poisson distribution (Model **IIa**) and the Gamma distribution (Model **IIb**) for these models as well. Since these four models are well known in the literature, we do not detail the estimation procedure further and we present the results in Subsection 1.4.6.

1.4.3 Fitting the three-component model

We adjust the 3-component model introduced in Section 1.2 (**Model III**). Regarding the closure delay, we test the Gamma, log-logistic and Weibull distributions and, based on the AIC and the BIC, we chose the Weibull distribution. Thus,

$$\left(T_{\ell}^{(c)}|\mathbf{c}_{\ell}\right) \sim \text{Weibull}\left(\lambda, \gamma_{\ell}\left(\boldsymbol{\beta}^{\left(t^{(c)}\right)}\right)\right), \text{ for } \ell \in \mathcal{L},$$

where, λ is the *shape parameter*, and $\gamma_{\ell}(\cdot)$ is the *scale parameter*. Also, $\boldsymbol{\beta}^{\left(t^{(c)}\right)}$ is the parameter vector used to predict the scale parameter.

Next, we fit the frequency component (testing the Poisson, Negative Binomial type I, and Negative Binomial type II distributions), considering the following

piece-wise development of claims :

$$\boldsymbol{\delta} = \{0, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 4, 5\},$$

where the Negative Binomial type II was chosen, again based on the AIC and the BIC. Thus,

$$(N_{\ell,k}|\mathbf{c}_\ell, \mathcal{D}_\ell, E_{\ell,k}) \sim \text{Neg Bin}\left(E_{\ell,k} \cdot \mu_{\ell,k}^{(n)}\left(\boldsymbol{\beta}^{(n)}\right), \sigma\right), \text{ if } E_{\ell,k} > 0, \text{ for } \ell \in \mathcal{L},$$

where $\mu_{\ell,k}^{(n)}(\cdot)$ and σ are such that,

$$\begin{aligned}
\mathrm{E}\left[N_{\ell,k}|\mathbf{c}_\ell, \mathcal{D}_\ell, E_{\ell,k}\right] &= E_{\ell,k} \cdot \mu_{\ell,k}^{(n)}\left(\boldsymbol{\beta}^{(n)}\right), \\
\mathrm{Var}\left[N_{\ell,k}|\mathbf{c}_\ell, \mathcal{D}_\ell, E_{\ell,k}\right] &= E_{\ell,k} \cdot \mu_{\ell,k}^{(n)}\left(\boldsymbol{\beta}^{(n)}\right) + \sigma\left(E_{\ell,k} \cdot \mu_{\ell,k}^{(n)}\left(\boldsymbol{\beta}^{(n)}\right)\right)^2.
\end{aligned}$$

Subsequently, the severity component is fitted with a splice model similar to the one suggested by Laudagé *et al.* (2019). Regarding the choice of the threshold, we used a 5-fold cross validation procedure to find the value of u that minimizes the mean square error between the predicted and the observed values of the out-of-sample sum of all payments. Regarding the choice of the distributions, we used a logit model to predict the probability of exceeding the threshold, and we chose the Gamma distribution for both the bulk and the value that exceeds the threshold. It is worth noting that in order to fit the bulk model $H_{\left(X_{\ell,k,m}\right)}$, which is a right-truncated parametric model, we used the gamlss.tr package of the statistical software R; for more information about this package we recommend the book by Stasinopoulus *et al.* (2017). Also, Laudagé *et al.* (2019) mentioned using GLM for the tail distribution $G_{\left(X_{\ell,k,m}\right)}$ is generally problematic because extreme values are rare. However in our particular problematic the data set is much larger because we are modeling payments instead of total losses, therefore we believe the

data set of payments over the threshold is large enough to circumvent the problem found by Laudagé *et al.* (2019). For reference, there are only $56,727$ claims in our training set, from which $315,527$ payments originate. Furthermore, with the optimal threshold, $u = 5,433$, the subset of the training set containing the payments exceeding the threshold contains $5,987$ observations.

Finally, regarding the IBNR claims, the occurrence delay is fitted with a multinomial distribution with 12 outcomes (one for each month of the year). Afterwards, the reporting delay is fitted with the mixture model suggested by Antonio & Plat (2014), using 8 degenerate components, i.e. $D = 7$. Then, based on $100,000$ simulations of both $T_\ell^{(o)}$ and $T_\ell^{(r)}$, the distribution of $U_\ell^{(r)}$ is obtained. Subsequently, the model of L_i^*, the number of IBNR claims for each occurrence year ($i = 1, \ldots, 5$) is fitted. It is worth noting that our data set did not contain the yearly registered exposures, thus we wrote $\omega_i = 1$, for $i = 1, \ldots, 5$. Afterwards, the closure delay, the frequency and, the severity are fitted with the same distributions used for the RBNS claims, though no information from vectors $\mathbf{c}_\ell$ was used in the fitting process.

In order to highlight the significance of micro-level information, Table 1.3 contains the estimated values of the parameters for covariates introduced in Table 1.1, and the p-value of their respective t-tests. We can observe, based on these tests, that most categories are significant in the fitting process. This in turn shows that this underlying information has an impact on the prediction of reserves. Therefore, individual models that can handle this kind of data could be attractive to insurers that have access to it.

TABLEAU 1.3: Estimated values

		Duration		Frequency		Severity (bulk)		Severity (prob)	
Variable	Category	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value
Type of loss	Single vehicle	0.17	< 0.01	0.37	< 0.01	0.12	< 0.01	0.32	0.09
	Multi vehicle	0.35	< 0.01	0.08	0.24	0.07	< 0.01	-0.04	0.85
	Hit Pedestrian	0.74	< 0.01	0.32	< 0.01	0.07	< 0.01	0.09	0.64
	Other	0.77	< 0.01	0.14	0.07	0.02	0.31	0.01	0.99
Injured Gender	Male	-0.10	< 0.01	0.15	< 0.01	0.13	< 0.01	0.20	< 0.01
	Unknown	-0.45	< 0.01	0.42	< 0.01	0.34	< 0.01	0.30	0.04
Region	Ontario	0.36	< 0.01	0.77	< 0.01	0.61	< 0.01	1.66	< 0.01
	West	-0.73	< 0.01	0.54	< 0.01	0.43	< 0.01	1.07	< 0.01
Injured Age	$(18, 25]$	0.01	0.71	0.25	< 0.01	0.12	< 0.01	0.54	< 0.01
	$(25, 30]$	0.18	< 0.01	0.23	< 0.01	0.16	< 0.01	0.56	< 0.01
	$[30, 50]$	0.25	< 0.01	0.26	< 0.01	0.14	< 0.01	0.60	< 0.01
	$(50, 70]$	0.29	< 0.01	0.26	< 0.01	0.10	< 0.01	0.63	< 0.01
	$(70, \infty)$	0.10	< 0.01	0.25	< 0.01	0.01	0.76	0.89	< 0.01
	Unknown	-0.23	< 0.01	-0.07	0.34	-0.09	< 0.01	0.25	0.10

		Duration		Frequency		Severity (bulk)		Severity (prob)	
Variable	Category	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value
Vehicle age	$(3, 6]$	0.02	0.30	0.02	0.51	-0.01	0.87	0.01	0.89
	$(6, 10]$	0.03	0.06	-0.01	0.57	-0.02	< 0.01	0.03	0.60
	$(10, 20]$	0.08	< 0.01	0.06	0.01	0.01	0.35	0.10	0.05
	$(20, \infty)$	0.14	< 0.01	0.23	< 0.01	0.02	0.40	0.32	0.01
	Unknown	0.39	< 0.01	0.05	0.05	0.06	< 0.01	0.29	< 0.01
$t_\ell^{(r)}$	$(1, 7]$	0.08	< 0.01	0.07	< 0.01	0.05	< 0.01	0.18	< 0.01
	$(7, 30]$	0.07	< 0.01	0.13	< 0.01	0.11	< 0.01	0.25	< 0.01
	$(30, 90]$	-0.12	< 0.01	0.38	< 0.01	0.16	< 0.01	0.51	< 0.01
	$(90, 180]$	-0.09	0.01	0.46	< 0.01	0.26	< 0.01	0.84	< 0.01
	$(180, 365]$	0.07	0.22	0.89	< 0.01	0.38	< 0.01	1.44	< 0.01
	$(365, \infty)$	0.25	< 0.01	0.82	< 0.01	0.37	< 0.01	1.61	< 0.01
Initial Reserve	$(1000, 5000]$	0.02	0.1	-0.24	< 0.01	0.04	< 0.01	0.08	0.03
	$(5000, 10000]$	0.63	< 0.01	-0.02	0.18	0.07	< 0.01	0.22	< 0.01
	$(10000, 20000]$	0.77	< 0.01	0.05	< 0.01	0.13	< 0.01	0.20	< 0.01
	$(20000, \infty)$	1.18	< 0.01	0.38	< 0.01	0.17	< 0.01	0.24	< 0.01

		Duration		Frequency		Severity (bulk)		Severity (prob)	
Variable	Category	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value
Time intervals	$(0.25, 0.5]$	-	-	0.59	< 0.01	0.17	< 0.01	-0.10	0.13
	$(0.5, 0.75]$	-	-	0.42	< 0.01	0.21	< 0.01	0.10	0.15
	$(0.75, 1]$	-	-	0.33	< 0.01	0.25	< 0.01	0.68	< 0.01
	$(1, 1.5]$	-	-	0.41	< 0.01	0.34	< 0.01	1.64	< 0.01
	$(1.5, 2]$	-	-	0.46	< 0.01	0.40	< 0.01	1.90	< 0.01
	$(2, 2.5]$	-	-	0.55	< 0.01	0.44	< 0.01	2.16	< 0.01
	$(2.5, 3]$	-	-	0.58	< 0.01	0.52	< 0.01	2.45	< 0.01
	$(3, 4]$	-	-	0.75	< 0.01	0.52	< 0.01	2.69	< 0.01
	$(4, 5]$	-	-	1.1	0.01	0.57	< 0.01	3.27	< 0.01
Intercept		4.50	< 0.01	5.08	< 0.01	5.34	< 0.01	-7.62	< 0.01
$\ln(1/\lambda)$		0.10	< 0.01	-	-	-	-	-	-

1.4.4 Fitting an individual model based on a Poisson Process

We have also adjusted the model suggested by Antonio & Plat (2014) (**Model IV**) in order to provide a more complete analysis and not to rely solely on collective approaches, which are by definition much simpler. We choose to keep as many similarities as possible to have more fair comparison between the models, while making minor adjustments to fit this model. Specifically, we used the the same time intervals (from $\boldsymbol{\delta}$) to delimit the intervals of the events Poisson Process which includes payments, closure and closure with payment. Finally, we used the severity component's distribution for the cost of single payments. However, in Antonio & Plat (2014) no methodology is indicated to include covariate information in the events Poisson Process, therefore micro-level information was not included for this step.

Table 1.4 contains the fitted (from events *before* the valuation date) and Table 1.5 contains the observed (from events *after* the valuation date) intensity of the Poisson process for each type of event, noted as h_p for payments, h_{sep} for closure, or settlement with payments, and h_{se} for settlement without payment. We notice that h_p is higher for the observed values compared to the fitted ones across all time intervals. Thus, the fitted h_p is does not represent accurately the intensity of the number of payments that will occur after the valuation date, and will likely result in an underestimation of the number of payments. This difference can be explained by a discrepancy between the claims considered from the training set (which contains closed and open claims) and the test set (which contains only open and unreported claims). This problem is further emphasised in Table 1.6, where the predicted number of payments after the valuation for this model was obtained by multiplying the total observed exposure by intervals to the fitted intensity of payments and closures with payments. Table 1.6 also contains the

predicted value of the chosen frequency component model (Section 1.4.3), and the predictions based on Section 1.2.2, using two simple models (Poisson and Negative Binomial) that only consider $\boldsymbol{\delta}$ as a covariate. For the latter three models, the exposure is also considered known. We notice that although all predictions are lower than the observed value considering covariate information considerably reduces the gap between the predicted and observed values, further indicating that, for this particular data set, considering covariate information is vital for better addressing the discrepancy between the training and the test data sets.

TABLEAU 1.4: Fitted intensity of the event Poisson process

	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 0.75]$	$(0.75, 1]$	$(1, 1.5]$
$\hat{h}_{se}$	2.05	1.33	0.90	0.86	0.89
$\hat{h}_{p}$	8.44	13.38	13.02	12.11	12.72
$\hat{h}_{sep}$	0.49	0.76	0.25	0.20	0.26
	$(1.5, 2]$	$(2, 2.5]$	$(2.5, 3]$	$(3, 4]$	$(4, 5]$
$\hat{h}_{se}$	0.68	0.63	0.63	0.53	0.54
$\hat{h}_{p}$	13.17	12.79	12.73	12.60	14.98
$\hat{h}_{sep}$	0.28	0.26	0.25	0.23	0.40

1.4.5 Goodness of fit analysis

The structure of our approach makes it possible to analyze the impact of micro-level covariates. We compare each of the components of our model (duration, frequency and severity) using micro-level information, with the corresponding model at the macro level, i.e., with a model that only uses the occurrence year (i) and the development year (j) as covariates. For the frequency component and for the severity component, we also compare a model using only the individual time intervals ($\boldsymbol{\delta}$) as covariates. The main objective is to weight the information provided

TABLEAU 1.5: Observed intensity of the event Poisson process

	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 0.75]$	$(0.75, 1]$	$(1, 1.5]$
h_{se}	2.04	1.26	0.80	0.80	0.78
h_p	11.12	15.38	15.50	14.08	14.08
h_{sep}	0.70	0.66	0.22	0.16	0.27
	$(1.5, 2]$	$(2, 2.5]$	$(2.5, 3]$	$(3, 4]$	$(4, 5]$
h_{se}	0.69	0.71	0.66	0.47	0.48
h_p	14.43	14.71	15.18	16.97	22.12
h_{sep}	0.26	0.27	0.26	0.35	0.38

TABLEAU 1.6: Predicted and observed number of payments after the valuation date

Model	Covariates	Exposure	Exp. number of payments
Poisson Process	$\boldsymbol{\delta}$	known	$76,236$
simple Poisson	$\boldsymbol{\delta}$	known	$77,425$
simple Neg. Bin.	$\boldsymbol{\delta}$	known	$78,455$
Neg. Bin.	$\boldsymbol{\delta}$ and $\boldsymbol{c}$	known	$84,306$
Observed value			$88,405$

by the individual time frame provided by $\boldsymbol{\delta}$ and the individual characteristics of each claim separately.

Table 1.7 contains the AIC and BIC criteria of all three kinds of models for their respective component. We can clearly observe that just by considering the individual time frame for the frequency and severity models, we obtain better results in terms of both criteria. We also notice that introducing individual claim information improves the performance of all models. Furthermore, we performed likelihood ratio tests between restricted and unrestricted models across all components. Yet

again, through these tests we want to determine whether including micro-level information, in the form of time intervals or characteristics of claims, is preferable than omitting them in the modelling process. Table 1.8 contains the results of the aforementioned Likelihood ratio tests, where we can reject the restricted models with an error of at most 0.01 %. Thus, we draw the same conclusion that was drawn from Table 1.7, i.e. that micro-level information improves the goodness of fit of models from each component.

TABLEAU 1.7: AIC and BIC criteria for models with different covariates across all components

		Macro only	Time intervals only	All micro
AIC	Duration	608,652		591,831
	Frequency	524,531	521,567	515,104
	Severity	4,756,217	4,752,714	4,737,359
BIC	Duration	608,705		592,144
	Frequency	524,630	521,716	515,581
	Severity	5,387,217	5,383,684	5,368,159

TABLEAU 1.8: Likelihood Ratio ($L.R.$) test for models with different covariates across all components

	restricted covariates	unrestricted covariates	L.R.	p-value
Duration	i	all covariates	9502	< 0.01
Frequency	i and j	i, j and $\boldsymbol{\delta}$	2974	< 0.01
	i, j and $\boldsymbol{\delta}$	all covariates	6528	< 0.01
Severity	i and j	i, j and $\boldsymbol{\delta}$	3829	< 0.01
	i, j and $\boldsymbol{\delta}$	all covariates	15,229	< 0.01

The structure of our model also allows us to perform a residual analysis based

on cells like in a run-off triangle. We drew inspiration from the residual analysis performed by Avanzi, *et al.* (2020) for their suggested collective model. Thus, we calculate residuals as ratios of observed values to the fitted values of cumulative triangles, and then we obtained heat maps based on these residuals. For our individual model, it is worth noting that the fitted cumulative exposure was obtained through simulation of the duration component while the fitted cumulative number of payments and the fitted cumulative cost were directly obtained from the expected value of each observation. Figure 1.12 contains the heat maps for all three components, while Figure 1.13 contains the heat map for the collective Gamma model. Comparatively, the severity component provides better results than the Gamma model for the first three development years, and overall seems to provide a better fit in spite of the worst results observed at the last two development years. Regarding the heat map of the duration component, accident years 2 and 5 have worst values than years 1, 3, 4 and 5 but, overall no extreme value is observed. As for the heat map of the frequency component, residuals are more variable than for the other components but overall the fitted values are close the observed ones.

Again, taking inspiration from Avanzi, *et al.* (2020), we also plot residuals in terms of accident years, development years and calendar years. This time residuals are calculated as difference between the sum of the observed values and the sum of fitted values for all cells in that year, divided by the sum of fitted values. Figures 1.14, 1.15 and 1.16 plot the residuals of each component. Residuals are close to 0 for all components and all type of years indicating that the goodness-of-fit is overall reasonable.

	0	1	2	3	4
1	1.01029	0.99154	0.99327	0.99287	0.99073
2	0.97331	0.97976	0.9817	0.9828	
3	1.00315	1.00195	1.00215		
4	0.99964	1.01364			
5	0.98262				

	0	1	2	3	4
1	1.06187	0.99546	0.98036	0.98138	0.98098
2	1.00404	0.9816	0.98551	0.98691	
3	0.97826	1.01268	1.0102		
4	1.03862	1.00827			
5	1.02336				

Year	0	1	2	3	4
1	1.00321	0.98323	0.99758	0.95059	0.95591
2	1.00685	1.05079	0.99305	1.01978	
3	1.01511	1.01913	0.98094		
4	1.00138	0.99087			
5	0.99565				

1.1
1.08
1.06
1.04
1.02
1
0.98
0.96
0.94
0.92
0.9

FIGURE 1.12 Heat maps of ratios of observed cumulative values to fitted cumulative values for the three component model (in order, from up to down, the exposure, the number of payments and the cost)

	0	1	2	3	4
1	1.17286	1.03733	1.03486	0.9967	0.99705
2	0.92822	0.95188	0.94946	0.99066	
3	0.90601	1.00885	1.01435		
4	0.9929	1.00194			
5	1				

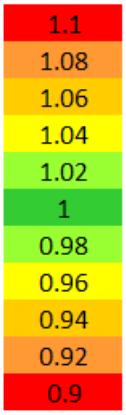

FIGURE 1.13 Heat maps of ratios of observed values to fitted values for the gamma collective model

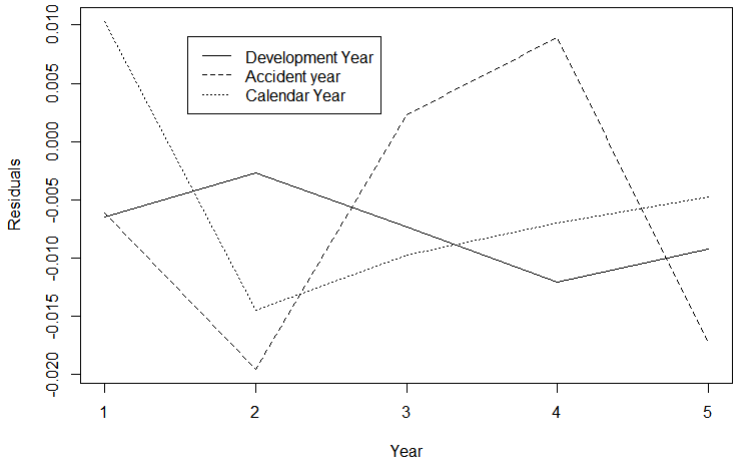

FIGURE 1.14 Plot of residuals of the cumulative exposure by accident, development and calendar year

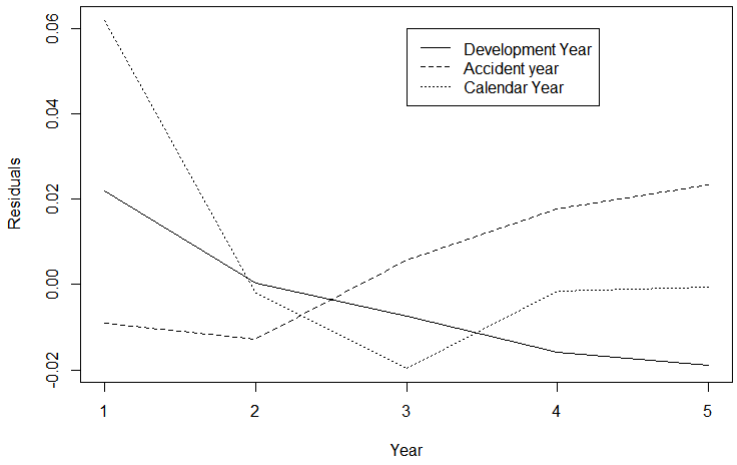

FIGURE 1.15 Plot of residuals of the cumulative number of payments by accident, development and calendar year

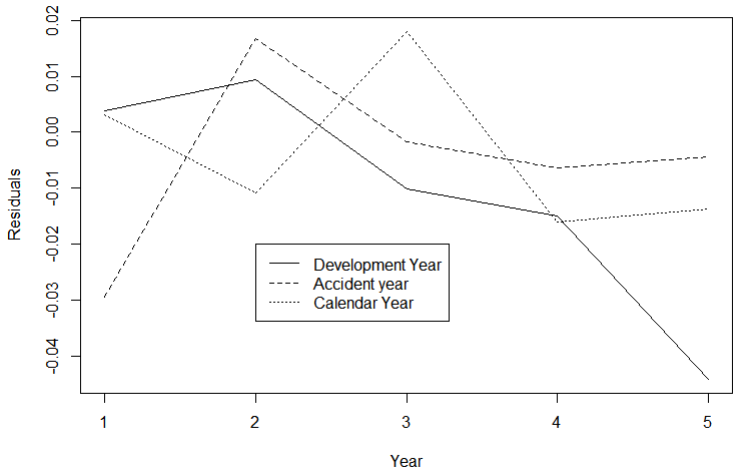

FIGURE 1.16 Plot of residuals of the cumulative cost by accident, development and calendar year

1.4.6 Outstanding reserve discussion

After the models were fitted, we proceeded to simulate the distribution of the loss reserves. For the 3-component model, this was accomplished by performing $10,000$ times the simulation procedure suggested in Section 1.3. As for the collective models and the Poisson Process model, more details about the simulation procedures are given by England & Verrall (2002), Wüthrich & Merz (2008) and Antonio & Plat (2014). For all the fitted models we obtained the distribution of the loss reserves up to the latest available date (December 31, 2017). These results are depicted in Figure 1.17 and summarized in Table 1.9. Then, we obtained these results including the missing data (from January 1, 2018 to December 31, 2019). These results are depicted in Figure 1.18 and summarized in Table 1.10.

TABLEAU 1.9: Results of the total reserve predictions until December 31, 2017

	Mean	SD	95% VaR	99% VaR
Mack ODP (**Ia**)	145,814,301	13,959,646	170,304,205	181,728,307
Mack Gamma (**Ib**)	146,025,032	13,919,184	169,926,224	180,890,690
GLM Gamma (**IIb**)	143,604,545	7,969,902	156,696,768	162,534,340
GLM ODP (**IIa**)	145,171,862	6,565,836	156,112,224	161,073,565
3-component RBNS	145,459,940	3,636,952	151,546,231	154,130,897
3-component IBNR	4,160,285	488,219	5,000,940	5,386,198
3-component (**III**)	149,620,225	3,678,054	155,830,382	158,291,786
Poisson Process RBNS	119,191,395	2,327,020	123,048,173	124,184,053
Poisson Process IBNR	3,022,166	228,903	3,416,934	3,554,742
Poisson Process (**IV**)	122,213,562	2,337,626	126,108,740	127,207,580
Observed	147,703,974			

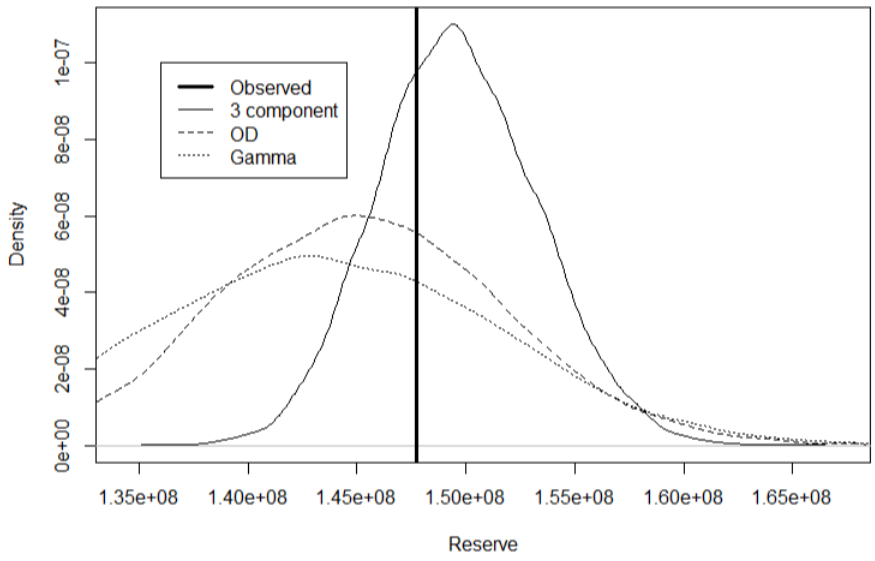

FIGURE 1.17 Total reserve distributions until December 31, 2017

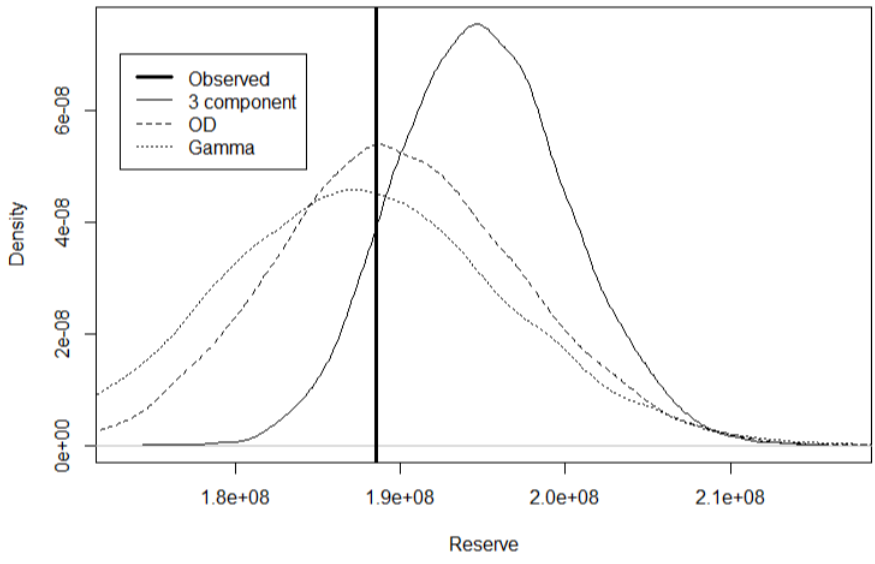

FIGURE 1.18 Total reserve distributions until December 31, 2019

TABLEAU 1.10: Results of the total reserve predictions until December 31, 2019

	Mean	SD	95% VaR	99% VaR
Mack ODP (**Ia**)	191,065,473	20,320,106	226,475,099	243,132,282
Mack Gamma (**Ib**)	190,766,649	20,350,205	226,012,897	244,050,402
GLM Gamma (**IIb**)	187,544,147	8,768,697	202,128,221	208,508,242
GLM ODP (**IIa**)	189,778,665	7,523,786	202,619,428	207,853,996
3-component RBNS	189,110,346	5,159,041	197,909,943	201,649,986
3-component IBNR	5,875,177	805,072	7,246,948	7,958,991
3-component (**III**)	194,985,523	5,233,211	203,929,434	207,493,721
Poisson Process RBNS	142,021,892	3,042,431	146,761,201	148,838,470
Poisson Process IBNR	4,949,629	599,795	5,993,530	6,507,927
Poisson Process (**IV**)	146,971,521	3,107,811	152,038,504	154,025,239
Observed*	188,520,892			

*includes predicted values

Let us analyze the obtained results. For the results until December 31, 2017, we can observe that the 95 % and the 99 % Values-at-Risk of all models is higher than the observed value, except for the Poisson Process micro-level model. The underestimation was foreseen in Section 1.4.4, where we noticed discrepancy between the observed and fitted payment intensities of the model. The rest of the models provide acceptable results, however the three-component model has the lowest standard-deviation and the lowest 95 % and 99 % Values-at-Risk across all models. This in turn suggests that it outperforms its counterparts by providing a narrower distribution all while providing the insurer with a loss reserve that covers the observed outstanding payments.

Finally, for the loss reserves until December 31, 2019, the results are narrower. Even though the three-component model still has the lowest standard-deviation among all models, the 95 % and the 99 % Values-at-Risk of the Gamma and the Over-dispersed Poisson have similar results. However, despite how close the Values-st-Risk are, the three-component model has the additional benefit of providing insurers with the loss reserve distribution of individual claims. Additionally, we can yet again notice that the Poisson Process micro-level model provides underestimated results, and thus is outclassed by our model for this particular data set.

Claims in a portfolio have different risk levels, which can be more easily identified by using their available information. Our 3-component model has the advantage of using this data in the form of explanatory variables to predict outstanding payments for each claim individually, instead of predicting the total reserve directly, as it is done for collective models. Moreover, through the simulation procedure described in Section 1.3 it is possible to estimate the distribution of each individual claim at the end of each development year. It is worth noting, however, that this type of analysis is not geared towards obtaining individual reserves with high

accuracy, but rather to identify potentially riskier claims.

As an example, we chose four claims from our data set with different characteristics and through 100,000 simulations we estimated the distribution of the cumulative payments. These values are summarized in Table 1.11.

TABLEAU 1.11: Outstanding cumulative payments of four claims at the end of each development year (j)

Claim	j	Mean	SD	75% VaR	95% VaR	99% VaR	Observed
1	2	23,108	33,863	32,512	91,900	155,409	53,676
	3	35,415	51,086	49,737	138,745	232,676	93,288
	4	42,023	62,375	57,621	168,511	288,355	n/a
2	2	78,346	96,769	112,934	272,974	432,691	34,184
	3	141,807	161,049	204,967	459,337	720,945	365,183
	4	194,478	217,788	285,711	631,308	959,678	n/a
3	1	6,949	9,411	9,320	25,817	44,102	42,580
	2	10,108	15,819	12,295	41,674	75,956	68,713
	3	11,466	20,401	12,677	48,744	101,116	n/a
	4	11,964	22,567	12,713	50,763	113,601	n/a
4	1	6,527	9,375	8,458	25,076	44,293	7,317
	2	9,668	15,556	11,457	41,169	74,348	13,181
	3	11,051	20,003	11,876	48,416	99,160	n/a
	4	11,534	21,979	11,916	50,452	111,078	n/a

We can see that all the observed values are situated under the 99 % VaR, thus the predictions provide high enough values to meet the required loss reserves under this risk measure. Further, claim 1 and 2 represent higher risks than claim 3 and 4 because the mean, standard deviations, and VaRs have larger values. Therefore,

the chosen covariates have an important impact on the prediction of outstanding payments, because the distribution of cumulative outstanding payments changes based on these values. Also, we see how an individual model that uses individual information, such as the one presented in this paper, can provide the insurer with some insight regarding the risk associated with claims. This information could be useful in the pricing process or to enable better reinsurance choices, for example.

1.5 Conclusion

Compared with macro-level models, micro-level models are capable of handling individual covariate information much easily. However, in spite of all their shortcomings, collective type models are still very popular in the industry due to their simplicity and easy to understand structure. In this paper, we suggested a model that can be interpreted in both a macro and micro level structure, while also being able to handle individual claim information. Furthermore, we derived frequency-severity structure with exposure, which is fairly similar to the one used by Property & Causalty actuaries for price-making predictions, making it even more accessible to the general public. We also put forth a fully parametric approach and proposed some models that can be considered across all components in our real data set analysis.

We managed to show that covariate information is significant in the fitting process across all components and, we even showed that utilizing this information allows insurers to make a more precise prediction of the total reserve compared with conventional models. This indicates that utilizing micro-level information can improve loss reserve predictions, thus making our model appealing to insurers that have access to precise information regarding their claims.

Moreover, this three-component structure opens the door for further research to-

pics, albeit at the cost of interpretability. One may consider the dependence between the frequency and severity as the outstanding claim develops or, even forgoing the piece-wise development triangle structure entirely to predict the total number of payments of each claim directly using more complex offset methods for the exposure, such as splines.

1.6 Appendix : Database examples

Table 1.12 contains the occurrence $\left(T_\ell^{(o)}\right)$, the reporting $\left(T_\ell^{(r)}\right)$, and the closure $\left(T_\ell^{(c)}\right)$ delays. Notice that, for open claims, the closure delay $\left(T_\ell^{(c)}\right)$ is censored by the valuation date $\left(T_\ell^{(e)}\right)$. We also have covariate information about the claims, such as the region.

TABLEAU 1.12: Example of a duration training set

ℓ	i	...	Region	$T_\ell^{(o)}$	$T_\ell^{(r)}$	$T_\ell^{(e)}$	$T_\ell^{(c)}$	Status
1	3	...	Atlantic	10/365	20/365	-	800/365	Closed
2	4	...	Ontario	200/365	35/365	-	365/365	Closed
3	5	...	West	100/365	30/365	235/365	-	Open

Table 1.13 contains the number of payments ($N_{\ell,k}$) based on the development year (j) and the time interval vector ($\boldsymbol{\delta}= \{0,1,2,3,4,5\}$). It also contains their respective exposures $E_{\ell,k}$ and the same covariate information from Table 1.12. Note that we can use the intervals from $\boldsymbol{\delta}$ as additional categorical covariates.

Table 1.14 contains the cost of single payments ($X_{\ell,k,m}$) and their respective covariate information. Here again, intervals from $\boldsymbol{\delta}$ can be used as additional covariates.

TABLEAU 1.13: Example of a frequency training set (with $\boldsymbol{\delta}= \{0, 1, 2, 3, 4, 5\}$)

ℓ	j	δ	k	i	Gender	...	$T_\ell^{(o)}$	$T_\ell^{(r)}$	$N_{\ell,k}$	$E_{\ell,k}$
1	0	$(0, 1]$	1	3	Female	...	10/365	20/365	1	335/365
1	1	$(0, 1]$	2	3	Female	...	10/365	20/365	1	35/365
1	1	$(1, 2]$	3	3	Female	...	10/365	20/365	2	330/365
1	2	$(1, 2]$	4	3	Female	...	10/365	20/365	0	35/365
1	2	$(2, 3]$	5	3	Female	...	10/365	20/365	0	65/365
2	0	$(0, 1]$	1	4	Male	...	200/365	35/365	0	130/365
2	1	$(0, 1]$	2	4	Male	...	200/365	35/365	2	235/365
3	0	$(0, 1]$	1	5	Female	...	100/365	30/365	1	235/365

TABLEAU 1.14: Example of a severity training set (with $\boldsymbol{\delta}= \{0, 1, 2, 3, 4, 5\}$)

ℓ	j	δ	k	m	i	Gender	...	$T_\ell^{(r)}$	$X_{\ell,k,m}$
1	0	$(0, 1]$	1	1	3	Female	...	20/365	\$100
1	1	$(0, 1]$	2	1	3	Female	...	20/365	\$200
1	1	$(1, 2]$	3	1	3	Female	...	20/365	\$550
1	1	$(1, 2]$	3	2	3	Female	...	20/365	\$900
2	1	$(0, 1]$	2	1	4	Male	...	35/365	\$200
2	1	$(0, 1]$	2	2	4	Male	...	35/365	\$300
3	0	$(0, 1]$	1	1	5	Female	...	30/365	\$100

CHAPITRE II

MODÉLISATION DE LA FRÉQUENCE DES PAIEMENTS DES RÉSERVES EN FONCTION D'UN SCORE DYNAMIQUE DE SINISTRE

2.1 Introduction

To accurately predict the cost of future liabilities for open claims, practitioners and researchers have suggested several models over the years. Over time, these models have changed a lot due to a significant increase in computing capacity and the quantity (and quality) of available data. While, in the past, models were always part of a collective framework, i.e., built for a data set aggregated by occurrence and development period (*run-off triangle*), today we see a wide selection of models based on the granularity of the underlying data set, ranging from raw data (micro-level) to aggregated data (macro-level). The actuarial literature on the subject has grown considerably in recent years, and we do not wish to do a detailed review here to avoid unnecessarily lengthening this paper. A review of the literature associated with some of the most essential and well-known models, such as the Chain-Ladder model (Mack (1999, 1993)), can be found in Wüthrich & Merz (2008) and England & Verrall (2002). As for individual approaches, let us mention, among others, the literature review in Blier-Wong *et al.* (2020) (section 4) and Taylor (2019). The rapid development of research in the field, partially explained by the increasing use of machine learning techniques, makes any literature review

incomplete on the day of its publication.

In this paper, we made a proposition in line with parametric and semi-parametric models. More specifically, we base our models on Position Dependent Marked-Poisson Process (PDMPP) to predict the exact time of each of the events of a claim, such as payments and settlements. One of the first papers using this type of model was Haastrup & Arjas (1996) and was expanded, in 2014, by a more practical implementation proposed by Antonio & Plat (2014), in which a more evidence-based methodology was suggested for both IBNR and RBNS reserves using a data set from an insurance company. Antonio *et al.* (2015) further developed this model type by including a multi-state approach that allowed the model to transition from one state to another as the claim evolved. Other processes that have been considered for the loss reserving literature include the Cox process (in Avanzi *et al.* (2021)) for which dependence was considered through common shock variables and the Hawkes process with time-varying intensities (see Maciak *et al.* (2021)). In contrast to these propositions, other models have been suggested. For example, let us mention Zhao *et al.* (2009), who have developed a semi-parametric model for IBNR claims and later incorporated copulae into the model. Moreover, Yanez & Pigeon (2021) introduced a more hierarchical structure, where the development of claims was divided into three components : duration of claims, payment frequency, and severity. Then in 2022, another hierarchical approach was suggested in Okine *et al.* (2022), which included the dependency between payments and settlement date. Finally, other authors have focused on the implementation of continuous Chain-Ladder methods to micro-level reserving (see Hiabu *et al.* (2016a), Hiabu *et al.* (2016b) and Hiabu (2017)).

Because of their granular structure, micro-level models can include more claim information in the modeling process than their aggregated counterparts. This information takes the form of covariates of three types (see Taylor *et al.* (2008)) :

static, time dynamic, and unpredictable time dynamic. Although time dynamic covariates change as time passes, while static covariates remain fixed, both can be predicted with certainty at any point in time. In contrast, unpredictable time dynamic covariates are, as the name suggests, unpredictable. Thus, both static and time dynamic covariates can often be included in models more straightforwardly than unpredictable time dynamic covariates. Despite the uncertainty associated with the latter type of covariates, useful claim information can be extracted from them. Specifically, when modeling RBNS claims, these covariates are abundant because a portion of the claim development has already been observed. Furthermore, a few models that can handle this information have been implemented, namely Antonio *et al.* (2015), which considered including interchangeable states based on payment counts, and Pigeon *et al.* (2014), which made use of incurred losses. This paper proposes a new method that can handle an unpredictable time dynamic covariate in a discrete-time interval framework.

For each open claim in the portfolio, we suggest using observed payments to improve the prediction of future payments. Past payments are summarized using a score system updated at the end of a given discrete time interval with the newly available information. One could implement our discrete-time scoring model into any individual model that can predict payment counts at discrete intervals and allows for including covariates. This latter element is important because the claim score will be considered a covariate. In particular, the frequency component in Yanez & Pigeon (2021) has both characteristics making it a candidate for the inclusion of this more intricate type of covariate.

Calculating a score based on previous observations is not new to the actuarial literature. The model in this paper draws inspiration from the bonus-malus scoring system (BMS) developed for claim counts. This method was developed in Boucher & Inoussa (2014), where the authors summarized previous claim counts into

a single numerical claim score. This model was further developed in Boucher & Pigeon (2019), where the claim score included linear effects. More recently, Verschuren (2021) proposed a version of the model that incorporates the claim development of different product lines into the score system. Finally, in Boucher (2023), a more compact and straightforward scoring system called a Kappa-N model was implemented. In this work, we take inspiration from all these sources to introduce a similar dynamic claim score system into the micro-level reserving literature.

The method we suggest offers a solution to including past claim information in the modeling process, fully taking advantage of a discrete interval structure. Moreover, we suggest distinguishing between different types of payments in the modeling process. This distinction is particularly relevant in loss reserving because payments occur for various reasons, such as medical bills and legal fees, and their distribution could vary. We illustrate this fact in our numerical analysis. To summarize, this paper has the following objectives :

— to implement a dynamic claim scoring system into a discrete interval payment loss reserve model and to weight the impact of such covariates in the fitting process ;
— to develop a model that considers different types of payments and analyzes their distribution ;
— to outperform models that only use static and time-dynamic covariates.

This paper is structured as follows. In Section 2.2, we look at the general framework of the model. Section 2.2.6 discusses the estimation procedure followed by Section 2.3, where we describe the simulation procedure of payment counts. Section 2.4 describes the data set used, followed by the numerical results of our model and other comparative models. Finally, Section 2.5 contains concluding remarks

and mentions other topics that could be explored based on our findings.

2.2 Statistical framework

In this section, we specify the statistical framework of our approach. We define the notation we use throughout the paper and present the construction of the dynamic claim score.

2.2.1 Introductory notation

We show the typical development of a P&C claim in Figure 2.1. First, accident i occurs, and we identify $t_i^{(o)}$, the occurrence delay, i.e., the delay between the beginning of the accident year and the exact accident date. An additional delay between the accident date and the reporting date is denoted by $t_i^{(r)}$. After the accident has been reported, several payments may be made – illustrated by dots in Figure 2.1 – before the claim is closed after a final delay $t_i^{(c)}$. At the valuation date, claims can be split into two categories depending on their development. If the claim has not yet been reported, we consider it Incurred But Not Reported, or IBNR, and if it has been reported, we consider it Reported But Not Settled, or RBNS. Furthermore, for RBNS claims, we can compute $t_i^{(e)}$, the delay between the reporting date and the valuation date.

In a loss-reserving context, we first must distinguish the status of each of the claims in the portfolio. Let $\mathcal{I} = \mathcal{I}^{(C)} \cup \mathcal{I}^{(O)}$ be the set containing the claims available at the valuation date, where $\mathcal{I}^{(C)}$ and $\mathcal{I}^{(O)}$ are the subsets containing, respectively, the closed and the open (RBNS) claims. Let $\mathcal{I}^*$ be the set containing unreported claims (IBNR), which are unknown at the valuation date.

For each claim, $i \in \mathcal{I}$, the observation period, i.e., the period between the reporting

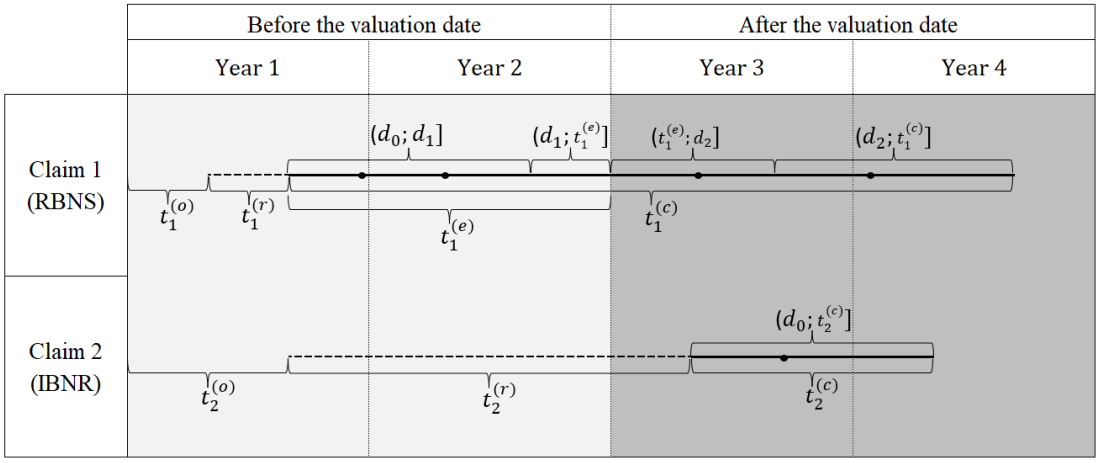

FIGURE 2.1 Development of two claims

date and the closure date (if the claim is closed) or the valuation date (if the claim is open), is denoted by $(0; \tau_i]$, where $\tau_i = \min\{t_i^{(c)}, t_i^{(e)}\}$. Afterwards, the observation period, $(0; \tau_i]$, $i \in \mathcal{I}$, can be divided into time intervals based on vector $\boldsymbol{d} = [d_0, d_1, \ldots, d_K]$, where $d_k < d_{k+1}$, $d_0 = 0$ and $d_K > \max_i\{\tau_i\}$. For the sake of simplicity, we can consider an annual framework, i.e., $\boldsymbol{d} = [0, 1, 2, \ldots]$, but one could also consider a monthly or seasonal division. We suggest basing this decision on the company's expertise or a cross-validation technique.

Furthermore, let $N_{i,k}$ be the number of payments for claim i, $i \in \mathcal{I}$, taking place over the interval $(d_k, d_{k+1}]$, and we define $\mathbf{N}_i = [N_{i,0}, N_{i,1}, \ldots, N_{i,K-1}]$. To each $N_{i,k}$, we associate an exposure measure indicating how long claim i has been open over interval $(d_k, d_{k+1}]$. Thus, let $E_{i,k}$ be the exposure measure of the claim i over the interval $(d_k, d_{k+1}]$:

$$E_{i,k} = \max\{\min\{\tau_i, d_{k+1}\} - d_k, 0\},$$

and $\mathbf{E}_i = [E_{i,0}, E_{i,1}, \ldots, E_{i,K-1}]$.

At the reporting date, micro-level information from a claim becomes available in the form of a vector $\mathbf{X}_i = [X_{i,1}, X_{i,2}, \ldots, X_{i,g}]$ of size g containing available static covariates, such as the region where the accident occurred. Note that this vector is not available for unreported claims (IBNR).

We can also identify a vector $\mathbf{Z}_{i,k} = [Z_{i,k,1}, Z_{i,k,2}, \ldots, Z_{i,k,h}]$ of size h containing time dynamic covariates available at each interval $(d_k, d_{k+1}]$. In particular, this vector contains at least one covariate indicating the interval k with which $N_{i,k}$ is associated. Thus, this vector exists for reported claimsand those that have not yet been reported (IBNR). For the latter, we define $\mathbf{Z}^*_{i,k} = [\mathbf{d}_k]$.

2.2.2 *A priori* distribution of the number of payments

RBNS claims

For open claims, $i \in \mathcal{I}^{(O)}$, we aim to predict the number of payments $N_{i,k}$, over the unobserved intervals after the valuation date $t_i^{(e)}$. We use the *a priori* information available at the reporting date (vectors $\mathbf{X}_i$ and $\mathbf{Z}_{i,k}$), as well as the exposure $\mathbf{E}_{i,k}$ before $t_i^{(e)}$. Commonly used approaches in a non-life insurance context can be considered, such as generalized linear models (GLM). The expected value of $N_{i,k}$, conditionally to $\mathbf{X}_i$, $\mathbf{Z}_{i,k}$ and $E_{i,k}$, is given by

$$\mu_{i,k} = \mathrm{E}\left[N_{i,k} | \mathbf{X}_i, \mathbf{Z}_{i,k}, E_{i,k}\right] = (E_{i,k})\, g^{-1}\left(\mathbf{X}_i \boldsymbol{\beta}' + \mathbf{Z}_{i,k} \boldsymbol{\theta}'\right),$$

where $g^{-1}()$ is the inverse of the link function, and $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are, respectively, the parameter vectors of static and time dynamic covariates.

IBNR claims

For claims that have occurred but have not been reported, $i \in \mathcal{I}^*$, we again aim to predict the number of payments $N_{i,k}$; however, given that the report date occurs after the valuation date, predictions must be made for all the intervals. Instead of having access to the information contained in the vectors $\mathbf{X}_i$ and $\mathbf{Z}_{i,k}$, we only have the information contained in $\mathbf{Z}^*_{i,k}$. Thus, the expected value of $N_{i,k}$, knowing $\mathbf{Z}^*_{i,k}$ and $E_{i,k}$, is given by,

$$\mu^*_{i,k} = \mathrm{E}\left[N_{i,k}|\mathbf{Z}^*_{i,k}, E_{i,k}\right] = (E_{i,k})\, g^{-1}\left(\mathbf{Z}^*_{i,k}(\boldsymbol{\theta}^*)'\right),$$

where $g^{-1}()$ is defined as previously, and $\boldsymbol{\theta}^*$ is the parameter vector based on time intervals $(d_k, d_{k+1}]$.

2.2.3 *A posteriori* distribution of the number of payments

We suggested a method to model frequency payments at different intervals based on information from vectors $\mathbf{X}_i$ and $\mathbf{Z}_{i,k}$, that respectively include static and time dynamic covariates. We can now focus on using information from time dynamics through various measures. Let $\epsilon_{i,k}$ and $\eta_{i,k}$ be, respectively, the cumulative number of payments and exposure of claim i over the interval $(d_0, d_k]$:

$$\epsilon_{i,k} = \sum_{j=0}^{k-1} E_{i,j}, \qquad \eta_{i,k} = \sum_{j=0}^{k-1} N_{i,j}.$$

We include the previously observed frequency in the mean parameter of claim i over interval $(d_k, d_{k+1}]$ in the following way :

$$\mu_{i,k} = \mathrm{E}\left[N_{i,k}|\mathbf{X}_i, \mathbf{Z}_{i,k}, \mathcal{H}_{i,k}\right] = (E_{i,k})\, g^{-1}\left(\mathbf{X}_i\boldsymbol{\beta}' + \mathbf{Z}_{i,k}\boldsymbol{\theta}' + \gamma_1\left(\frac{\eta_{i,k}}{\epsilon_{i,k}}\right)\right),$$

where $\mathcal{H}_{i,k}$ is the known development of claim i at time d_k, and γ_1 is the parameter associated with the new component.

Then, we want to adjust the expected value of the frequency by incorporating a covariate that identifies payment-free periods to distinguish between claims that have been open for a longer or shorter period. Thus, as a claim develops, the frequency of payment-free periods may increase or reduce the expected value. This approach is inspired by the Kappa-N structure suggested by Boucher (2023). Let $\kappa_{i,k}$ represent the total payment-free exposure observed over interval $(d_0, d_k]$, such that,

$$\kappa_{i,k} = \sum_{j=0}^{k-1} E_{i,j}\mathbb{1}\left(N_{i,j} = 0\right),$$

where $\mathbb{1}()$ is the indicator function.

We can rewrite the mean parameter by incorporating both elements into a single

claim score :

$$
\begin{aligned}
\mu_{i,k} &= \mathrm{E}\left[N_{i,k}|\mathbf{X}_i, \mathbf{Z}_{i,k}, \mathcal{H}_{i,k}\right] \\
&= (E_{i,k})\, g^{-1}\left(\mathbf{X}_i\boldsymbol{\beta}' + \mathbf{Z}_{i,k}\boldsymbol{\theta}' + \gamma_0(-\kappa_{i,k}) + \gamma_1\left(\frac{\eta_{i,k}}{\epsilon_{i,k}}\right)\right) \\
&= (E_{i,k})\, g^{-1}\left(\mathbf{X}_i\boldsymbol{\beta}' + \mathbf{Z}_{i,k}\boldsymbol{\theta}' + \gamma_0\left(-\kappa_{i,k} + \frac{\gamma_1}{\gamma_0}\left(\frac{\eta_{i,k}}{\epsilon_{i,k}}\right)\right)\right) \\
&= (E_{i,k})\, g^{-1}\left(\mathbf{X}_i\boldsymbol{\beta}' + \mathbf{Z}_{i,k}\boldsymbol{\theta}' + \gamma_0\underbrace{\left(-\kappa_{i,k} + \psi\left(\frac{\eta_{i,k}}{\epsilon_{i,k}}\right)\right)}_{\text{claim score } \ell_{i,k}}\right) \\
&= (E_{i,k})\, g^{-1}\left(\mathbf{X}_i\boldsymbol{\beta}' + \mathbf{Z}_{i,k}\boldsymbol{\theta}' + \gamma_0\ell_{i,k}\right),
\end{aligned}
$$

where $k > 0$ and ψ is defined as the *jump-parameter*.

This structure summarizes past claim experience into a single claim score that will be updated at the end of each interval. Then, the mean parameter can identify claims with a higher chance of producing payments and riskier claims. Notice that $\kappa_{i,k}$ is multiplied by -1 to accommodate better the negative impact of no-payment periods on the claim score.

Note that the mean parameter is unbounded. This situation can be an issue because upper mean parameter values can become excessively large as we include past frequency in our calculations, and outliers are not uncommon. Indeed, payment counts vary depending on how data is collected in loss reserving. Administrative reasons could cause a particular Cash flow to be divided into various payments. This situation may introduce more outliers. An actuary may regroup certain payments to consider this problem and, later, accommodate possible outliers in terms of severity through extreme value models such as Laudagé *et al.* (2019). However, outliers could still be present in the data set after regrouping payments. One may even consider situations with insufficient information to jus-

tify restructuring the data set.

A solution to this problem is the inclusion of a maximum value for the claim score (this method is consistent with the Bonus-Malus literature see Boucher (2023)). The decreasing part of the measure, based on $\kappa_{i,k}$, is bounded by the maximal duration of a claim and is less prone to excessively impacting the prediction of the mean. Thus, including a minimal value for the mean parameter is less suitable. Finally, when we look into new claims, no history has been previously observed, and we cannot include the dynamic claim score measure. Thus, by setting the initial value of the claim score to 0, all predictions of the mean parameter are based only on the other covariates available at the report date. We suggest obtaining a claim score such that :

$$\ell_{i,k} = \begin{cases} \min\left\{\left(-\kappa_{i,k} + \psi\left(\frac{\eta_{i,k}}{\epsilon_{i,k}}\right)\right), \ell_{max}\right\}, & \text{for } k = 1, 2, \dots \\ 0, & \text{for } k = 0. \end{cases} \tag{2.1}$$

One should note that the claim score for claim i is updated at the end of each interval k based on information up to the previous interval $k-1$. As such, it is possible to identify which claims are more likely to produce payments derived from past information summarized by the value of the claim score at any given time. We could also expand upon the definition of the claim score by letting $\nu_{i,k}$ be the sum of the previously observed frequencies such that :

$$\nu_{i,k} = \sum_{j=0}^{k-1} \frac{N_{i,j}}{E_{i,j}},$$

and we can then reformulate the value of the risk measure :

$$\ell_{i,k} = \begin{cases} \min\left\{\left(-\kappa_{i,k} + \psi\nu_{i,k}\right), \ell_{max}\right\}, & \text{for } k = 1, 2, \ldots \\ 0, & \text{for } k = 0. \end{cases} \tag{2.2}$$

$$= \begin{cases} \min\left\{\sum_{j=0}^{k-1}\left(-E_{i,j}\mathbb{1}\left(N_{i,j} = 0\right) + \psi\left(\frac{N_{i,j}}{E_{i,j}}\right)\right), \ell_{max}\right\}, & \text{for } k = 1, 2, \ldots \\ 0, & \text{for } k = 0. \end{cases}$$

Then we can obtain a recursive structure reminiscent of the Bonus-Malus structure used for claim count modeling :

$$\ell_{i,k} = \begin{cases} \min\left\{\left(\ell_{i,k-1} - E_{i,j}\mathbb{1}\left(N_{i,j} = 0\right) + \psi\left(\frac{N_{i,j}}{E_{i,j}}\right)\right), \ell_{max}\right\}, & \text{for } k = 1, 2, \ldots \\ 0, & \text{for } k = 0. \end{cases} \tag{2.3}$$

In particular, model (2.3) has the added advantage of being able to compute the value of any risk score $\ell_{i,k}$ just by knowing the value of the previous risk score $\ell_{i,k-1}$ and the information from the current interval $(d_{k-1}, d_k]$. Hence, unlike previous propositions (2.1 and 2.2), all information observed over the period $(d_0, d_{k-1}]$ is not mandatory to compute $\ell_{i,k}$.

For the remaining part of this paper, we label these three propositions as models **(M1)**, **(M2)** and **(M3)**, respectively, for models based on claim score definitions (2.1), (2.2) and (2.3). Further considerations will be addressed in the next section using **(M1)** as an example ; however, similar results can be obtained for models **(M2)** and **(M3)**.

2.2.4 Payment categories and IBNR specifications for claim-score modelling

Payments can be divided into several categories, e.g., payments related to medical costs or administrative costs. Suppose there are A different categories of payments. Also, we want to incorporate past payment count information in the fitting process from different payment categories as the claims develop using a claim score. For a given payment category, we propose using a dynamic claim score model with two parameters $\left(\psi^{(a)}, \ell_{max}^{(a)}\right)$ where the level of risk associated with the category a, $a = 1, \ldots, A$, at the beginning of the interval $(d_k, d_{k+1}]$ is given by

$$\ell_{i,k}^{(a)} = \begin{cases} \min\left\{\left(-\kappa_{i,k}^{(a)} + \psi^{(a)}\left(\dfrac{\eta_{i,k}^{(a)}}{\epsilon_{i,k}}\right)\right), \ell_{max}^{(a)}\right\}, & \text{for } k = 1, 2, \ldots \\ 0, & \text{for } k = 0, \end{cases}$$

where $\psi^{(a)}$ is the *jump-parameter* for category a, $\ell_{max}^{(a)}$ is the *maximum claim score* for category a, and

$$\epsilon_{i,k} = \sum_{j=0}^{k-1} E_{i,j}, \qquad \eta_{i,k}^{(a)} = \sum_{j=0}^{k-1} N_{i,j}^{(a)}, \qquad \kappa_{i,k}^{(a)} = \sum_{j=0}^{k-1} E_{i,j} \mathbb{1}\left(N_{i,j-1}^{(a)} = 0\right).$$

Information from the claim scores of each category can then be incorporated into the process. Let $\boldsymbol{\ell}_{i,k} = \left[\ell_{i,k}^{(1)}, \ell_{i,k}^{(2)}, \ldots, \ell_{i,k}^{(A)}\right]$ be the vector containing the risk levels associated with the different categories of payments. Then, for RBNS claims, we can obtain the expected value of the number of payments from category a,

$$\mu_{i,k}^{(a)} = \mathrm{E}\left[N_{i,k}^{(a)} | \mathbf{X}_i, \mathbf{Z}_{i,k}, E_{i,k}, \boldsymbol{\ell}_{i,k}\right] = (E_{i,k})\, g^{-1}\left(\mathbf{X}_i' \boldsymbol{\beta}^{(a)} + \mathbf{Z}_{i,k}' \boldsymbol{\theta}^{(a)} + \gamma^{(a)} \ell_{i,k}^{(a)}\right),$$

and we obtain the expected value of the number of payments from category a for IBNR claims :

$$\mu_{i,k}^{*(a)} = \mathrm{E}\left[N_{i,k}^{(a)} | \mathbf{Z}_{i,k}^{*}, E_{i,k}, \boldsymbol{\ell}_{i,k}^{*}\right] = (E_{i,k})\, g^{-1}\left(\mathbf{Z}_{i,k}^{*\prime}\boldsymbol{\theta}^{*(a)} + \gamma^{*(a)}\ell_{i,k}^{*(a)}\right).$$

We include the same restriction that we used in the RBNS claims by setting $\ell_{max}^{*(a)}$ as the maximal claim score and by including its respective jump-parameter $\psi^{*(a)}$. Notice that because the information from these types of claims is unknown, we can only include covariate vector $\mathbf{Z}_{i,k}^{*}$, in addition to the claim scores $\ell_{i,k}^{*(a)}$.

It is worth mentioning that unlike RBNS claims, where a portion of the development is observed, which can then be computed into the claim scores up to the valuation date, the IBNR claims are fully simulated from the occurrence date, up to the closure date. Thus, no *observed* past information can be used to compute the claim score of a given IBNR claim. In this sense, the dynamic claim score is more relevant for RBNS claims because actual observed information is included to predict future intervals.

2.2.5 Distribution of duration of claims

With pricing models, where BMS models are commonly used to predict claim counts, the duration of contracts is known beforehand. However, when we seek to predict outstanding payment counts in a loss reserve context, the entire duration of open or unreported claims is unknown. Thus an additional model is required to predict this value to obtain the exposure values after the valuation date. This problem was fully addressed in Yanez & Pigeon (2021), where, for claim i, the duration was divided into three parts modeled by three random variables :

— $T_i^{(o)}$ for the occurrence delay ;

— $T_i^{(r)}$ for the reporting delay ; and

— $T_i^{(c)}$ for the closure delay.

For RBNS claims, the report and occurrence date are known, and the information contained in the covariate vectors $\mathbf{X}_i$ and $\mathbf{Z}_{i,k}$ is also accessible. Hence, it is only necessary to model the closure delay with the added advantage of having access to micro-level information. In Yanez & Pigeon (2021), various distributions are considered from the survival literature, such as the Weibull and the Gamma distribution. It is worth noting that the training set used contains right-censored observations because of the valuation date. For more details, refer to the paper mentioned above.

For IBNR claims, however, it is necessary to model all three parts of the duration, and no individual information is available. In Yanez & Pigeon (2021), the occurrence delay is addressed with methods that consider seasonal effects. The reporting delay is based on the paper by Antonio & Plat (2014), where a mixture of a Weibull distribution with degenerate components was considered to accommodate the observations that only take a few days to complete. The closure delay was addressed similarly to the RBNS claims without considering individual information. Again, refer to Yanez & Pigeon (2021) for more details.

2.2.6 Parameter estimation

The *a priori* distribution parameters $\boldsymbol{\beta}^{(a)}$, $\boldsymbol{\theta}^{(a)}$, and $\gamma^{(a)}$ for each type of payment $a = 1, \ldots, A$ are estimated by maximizing the likelihood function given by

$$\Lambda = \prod_{i \in \mathcal{I}} \prod_{k=0}^{K-1} \prod_{a=1}^{A} p_{\left(N_{i,k}^{(a)} | \mathbf{X}_i, \mathbf{Z}_{i,k}, E_{i,k}, \boldsymbol{\ell}_{i,k}\right)} \left(n_{i,k}^{(a)} | \mathbf{x}_i, \mathbf{z}_{i,k}, e_{i,k}, \boldsymbol{\ell}_{i,k}\right),$$

where $p()$ is the probability mass function of the number of claim payments over each interval given covariates, dynamic claim score, and exposure. We suggest estimating jump-parameter $\psi^{(a)}$ and the maximal values of claim scores $\ell_{max}^{(a)}$ by looking for the values that generate the best likelihood or the best predictions, based on an out-of-sample analysis.

Because we distinguish between IBNR and RBNS reserves, it is also important to comment on the parameter estimation procedure for IBNR claims. One can follow the same procedure already described. However, instead of using micro-level covariate vectors, i.e., $\mathbf{X}_i$ and $\mathbf{Z}_i$), we only have access to the covariate vector $\mathbf{Z}_{i,k}^*$. Thus, the likelihood function is given by

$$\Lambda^* = \prod_{i\in\mathcal{I}} \prod_{k=0}^{K-1} \prod_{a=1}^{A} p^*_{\left(N_{i,k}^{(a)}|\mathbf{Z}_{i,k}^*, E_{i,k}, \boldsymbol{\ell}_{i,k}^*\right)} \left(n_{i,k}^{(a)}|\mathbf{z}_{i,k}^*, e_{i,k}, \boldsymbol{\ell}_{i,k}^*\right),$$

where $p^*()$ is the probability mass function. The procedure for estimating jump-parameters, $\psi^{*(a)}$, and the maximum values of claim scores $\ell_{max}^{*(a)}$ remains similar.

2.3 Simulation procedure

As stated previously, loss reserves are split into two types : IBNR and RBNS. We have established different modeling procedures for both reserves and in this section, we must establish the two different simulation procedures. We consider model **(M1)** for these algorithms ; however, similar algorithms can be constructed for models **(M2)** and **(M3)** by adapting the calculation of **step 5c** (2.5) for IBNR claims and **steps 3(a and b)** (2.6 and 2.7) for RBNS claims. An example when considering $\psi = 2$ is given by Figure 2.2.

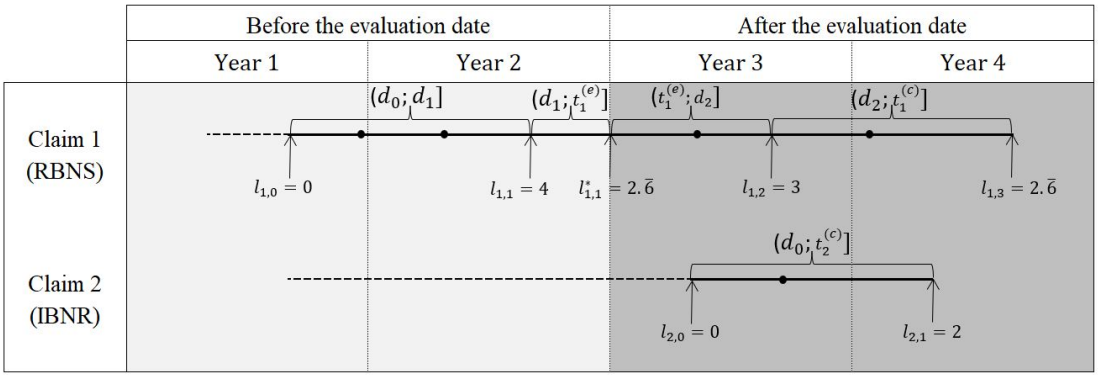

FIGURE 2.2 Dynamic claim score development (with $\psi = 2$)

2.3.1 *IBNR* simulation procedure

The exact number of IBNR claims and their information are unknown at the valuation date. Before we define the simulation procedure for the number of payments, we must perform a few steps. As indicated in Subsection 2.2.5, for these claims, all three delays must be simulated : the occurrence delay, $t_i^{(o)}$, the reporting delay, $t_i^{(r)}$, and the closure delay, $t_i^{(c)}$ (see Figure 2.1). In this particular context, we consider $u_i^{(r)} = t_i^{(o)} + t_i^{(r)}$, the delay between the beginning of the accident year and the report date of claim i. Moreover, because of the unobserved nature of IBNR claims, we must also simulate how many have occurred per accident year. Several propositions have been put forward to predict this value. For instance, in Zhao *et al.* (2009), a semi-parametric methodology was suggested, whereas, in Antonio & Plat (2014), an approach based on a Poisson process was considered. In this paper, we will accommodate the thinned-Poisson model by Pigeon *et al.* (2014) to our simulation procedure, although the aforementioned models can also be considered.

Let $m = 1, \ldots, M$ be the accident year of a given claim, where M is the total number of years considered. We select an approach inspired by the work of Pigeon

et al. (2014) and assign a distribution to I_m^*, the number of IBNR claims for each m accident year. By letting m_i be the accident year of claim i, we have :

$$I_m^* \sim \text{Poisson}\left(\theta\omega_m \Pr\left[U_i^{(r)} \leq M - m_i + 1 | m_i = m\right]\right), \tag{2.4}$$

where $\theta\omega_m$ is the occurrence measure, for which ω_m is the total exposure registered for period m. The occurrence measure is thinned by $\Pr\left[U_i^{(r)} \leq M - m_i + 1\right]$. This value represents the probability that the report date occurs before the valuation date. In order to obtain this value, we consider the distribution of the sum of the occurrence delay $T_i^{(o)}$, that is, the delay between the beginning of the accident date and the exact accident date, and the report delay $T_i^{(r)}$, the delay between the accident date and the report date. The distributions suggested for these two delays are briefly detailed in section 2.2.5. We can now define the simulation procedure for IBNR payments as follows :

— **Step 1 :** Obtain $\widetilde{I}^* = \sum_m \widetilde{I}_m^*$,where $\widetilde{I}_m^*$ is the simulated value of I_m^* for each occurrence period m (see Equation (2.4)).

— **Step 2 :** Obtain $\widetilde{U}_i^{(r)}$, the simulated value of $\left(U_i^{(r)} | U_i^{(r)} > M - m_i + 1\right)$, the delay between the beginning of the occurrence period and the exact reporting date of each simulated IBNR claim, where,

$$\Pr\left[U_i^{(r)} \leq u | U_i^{(r)} > M - m_i + 1\right] = \frac{\Pr\left[M - m_i + 1 < U_i^{(r)} \leq u\right]}{1 - \Pr\left[U_i^{(r)} \leq M - m_i + 1\right]},$$

for $i = 1, \ldots, \widetilde{I}^*$.

— **Step 3 :** Obtain $\widetilde{T}_i^{(c)}$, the simulated value of $\left(T_i^{(c)} | m_i\right)$, the closure delay of claim i, for $i = 1, \ldots, \widetilde{I}^*$.

— **Step 4 :** Calculate

$$\widetilde{E}_{i,k} = \begin{cases} d_{i,k+1} - d_{i,k}, & \text{if } d_{i,k+1} \leq \widetilde{T}_i^{(c)} \\ \widetilde{T}_i^{(c)} - d_{i,k}, & \text{if } d_{i,k+1} > \widetilde{T}_i^{(c)} \\ 0, & \text{elsewhere,} \end{cases}$$

for $k = 0, \ldots, K-1$ and $i = 1, \ldots, \widetilde{I}^*$.

— **Step 5 :** For $i = 1, \ldots, \widetilde{I}^*$, go through each of the following sub-steps.

— **Step 5a :** Set $k = 0$, the first time interval for which the exposure of claim i is positive and obtain its risk level by setting $\widetilde{\ell}_{i,0}^{*(a)} = 0$ for $a = 1, \ldots, A$.

— **Step 5b :** Obtain $\widetilde{N}_{i,k}^{(a)}$, a simulated value of $\left(N_{i,k}^{(a)} | \mathbf{Z}_{i,k}^*, \widetilde{E}_{i,k}, \widetilde{\boldsymbol{\ell}}^*{}_{i,k}\right)$, for $a = 1, \ldots, A$.

— **Step 5c :** Calculate the next risk level,

$$\widetilde{\ell}_{i,k+1}^{*(a)} = \min \left\{ -\sum_{j=1}^{k} \widetilde{E}_{i,j} \mathbb{1}\left(\widetilde{N}_{i,j}^{(a)} = 0\right) + \psi^{*(a)} \frac{\sum_{j=1}^{k} \widetilde{N}_{i,j}^{(a)}}{\sum_{m=1}^{k} \widetilde{E}_{i,j}}, \ell_{max}^{*(a)} \right\} \quad (2.5)$$

for $a = 1, \ldots, A$.

— **Step 5d :**

— If $\widetilde{E}_{i,k+1} > 0$, set $k = k + 1$, the next time interval for which the exposure of claim i is positive. Then return to **Step 5b**.

— If $\widetilde{E}_{i,k+1} = 0$ stop the simulation procedure of claim i.

RBNS simulation procedure

With RBNS claims, we have micro-level information in the form of vectors $\mathbf{X}_i$ and $\mathbf{Z}_{i,k}$. Because we are dealing with open claims, a portion of the development has already been observed. We can use the observed risk level contained in $\boldsymbol{\ell}_{i,k}$ to simulate the unobserved portion of the development. Furthermore, unlike with

IBNR claims, the exact number of open claims, $I^{(O)}$, is known beforehand. With these considerations can now describe the simulation procedure,

— **Step 1a :** Set $i = 1$, the first open claim.

— **Step 1b :** Obtain $\widetilde{T}_i^{(c)}$, the simulated value of $\left(T_i^{(c)}|\mathbf{X}_i\right)$, the closure delay of open claim i,

— **Step 1c :** If $\widetilde{T}_i^{(c)} > t_i^{(e)}$, set $i = i + 1$, the next open claim.

— **Step 1.d :**

— If $i \leq I^{(O)}$, go to **Step 1.b**.

— If $i = I^{(O)} + 1$, continue.

— **Step 2 :** Calculate the exposures after the valuation date,

$$\widetilde{E}_{i,k} = \begin{cases} d_{i,k+1} - t_i^{(e)}, & k \in \{k : d_{i,k} \leq t_i^{(e)}, d_{i,k+1} \leq \widetilde{T}_i^{(c)}\} \\ \widetilde{T}_i^{(c)} - t_i^{(e)}, & k \in \{k : d_{i,k} \leq t_i^{(e)}, d_{i,k+1} > \widetilde{T}_i^{(c)}\} \\ d_{i,k+1} - d_{i,k}, & k \in \{k : d_{i,k} > t_i^{(e)}, d_{i,k+1} \leq \widetilde{T}_i^{(c)}\} \\ \widetilde{T}_i^{(c)} - d_{i,k}, & k \in \{k : d_{i,k} > t_i^{(e)}, d_{i,k+1} > \widetilde{T}_i^{(c)}\} \\ 0, & \text{elsewhere,} \end{cases}$$

for $k = 0, \ldots, K-1$ and $i \in \mathcal{I}^{(O)}$.

— **Step 3 :** For each $i \in \mathcal{I}^{(O)}$, go through each of the following sub-steps.

— **Step 3a :** Set $k = \{k : d_{i,k} \leq t_i^{(e)} < d_{i,k+1}\}$, the first time interval that takes place after the evaluation date and obtain its risk level by calculating, if $d_{i,k} < t_i^{(e)}$,

$$\widetilde{\ell}_{i,k}^{(a)} = \min\left\{ -\sum_{j=1}^{k} E_{i,j} \mathbb{1}\left(N_{i,j}^{(a)} = 0\right) + \psi^{(a)} \frac{\sum_{j=1}^{k} N_{i,j}^{(a)}}{\sum_{j=1}^{k} E_{i,j}}, \ell_{max}^{(a)} \right\}, \quad (2.6)$$

while, setting, if $d_{i,k} = t_i^{(e)}$,

$$\widetilde{\ell}_{i,k}^{(a)} = \ell_{i,k}^{(a)},$$

for $a = 1, \ldots, A$. Note that if a portion of the interval has been observed, i.e., when $d_{i,k} < t_i^{(e)}$, we use the first portion, $(d_{i,k}, t_i^{(e)}]$, to update the risk level of the remainder of the interval. However, if no portion of the interval has been observed, i.e., when $d_{i,k} = t_i^{(e)}$, then the latest information available occurs at the previous time interval $(d_{i,k-1}, d_{i,k}]$, and the risk level is updated based on this information instead.

— **Step 3b :** Obtain $\widetilde{N}_{i,k}^{(a)}$, a simulated value of $\left(N_{i,k}^{(a)} | \mathbf{X}_i, \mathbf{Z}_{i,k}, \widetilde{E}_{i,k}, \widetilde{\boldsymbol{\ell}}_{i,k}\right)$, for $a = 1, \ldots, A$.

— **Step 3c :** Calculate the next risk level,

$$\widetilde{\ell}_{i,k+1}^{(a)} = \min \left\{ -\sum_{j=1}^{k} \widetilde{E}_{i,j} \mathbb{1}\left(\widetilde{N}_{i,j}^{(a)} = 0\right) + \psi^{(a)} \frac{\sum_{j=1}^{k} \widetilde{N}_{i,j}^{(a)}}{\sum_{j=1}^{k} \widetilde{E}_{i,j}}, \ell_{max}^{(a)} \right\} \quad (2.7)$$

for $a = 1, \ldots, A$.

— **Step 3d :**

— If $\widetilde{E}_{i,k+1} > 0$, set $k = k + 1$, the next time interval for which the exposure of claim i is positive. Then return to **Step 3b**.

— If $\widetilde{E}_{i,k+1} = 0$ stop the simulation procedure of claim i.

2.4 Numerical results

2.4.1 Data Set

We consider a data set from a Canadian insurance company for our numerical analysis. The data set contains information from 57,593 claims about Accident Benefits (AB) coverage, i.e., no-fault benefits for accidents where the driver, or

a third party, was injured or killed in a car accident. Micro-level information is incorporated in the modeling process as categorical static covariates, summarized in Table 2.1. However, some of the covariates contain missing values (NA). We can keep these observations in the process by creating a "missing value" category for each of the covariates. We decided not to remove observations with one or more missing values, as this would have deprived us of much information.

The claims considered in our analysis have occurrence dates from 2011 to 2015, and we have information regarding their development until December 31, 2017. In order to evaluate the performance of our model, we chose to set the valuation date to December 31, 2015, splitting the data set into a training and an evaluation set. Payments before the valuation date are used to fit the models, while payments until December 2017 are used for validation.At the valuation date, there were 48,855 closed claims, 7,872 open claims, and 866 unreported claims in our portfolio.

TABLEAU 2.1: Description of covariates

Covariate	Label	Number of levels
Gender	Gender of the injured/killed	3
Region	Geographical region	3
Type of loss	Kind of AB claim	5
Vehicle age	Age of the vehicle	6
Injured age	Age of the injured/killed	7
Reporting delay	Delay calculated in days	7
Initial reserve	Reserve at report date	5

Diving more deeply into the number of payments from the data set, which is the focus of this paper, we group payments into three categories :

1. **Medical** : all medical payments ;

2. **Disability** : recurrent payments such as Disability Income and Caregiver Disability Income ; and
3. **Expenses** : all other types of expenses.

We chose these groups based on the nature of the payments, as previously described, and their empirical distribution. In Table 2.2, we present some descriptive statistics of the claim frequency for each category in the training set, such as the Value-at-Risk, or VaR.

TABLEAU 2.2: Claim frequency descriptive statistics for each category

	Mean	Std. dev.	95% VaR	99% VaR
Medical	3.44	9.86	13.70	41.00
Disability	1.01	5.79	4.00	27.00
Expense	1.11	3.60	7.00	17.00
All	5.57	16.81	24.00	74.00

Finally, we make some simplifying assumptions about the possible dependency in the data set. First, in some situations, a casualty may trigger coverages from different claims, and we acknowledge that this situation can cause dependency between these claims. However, we will not address this situation in this study because the proposition made in this paper is more geared towards tackling the problem of including past information from the claims themselves rather than information from other dependent claims. Consequently, we assumed independence between those claims. Second, we do not consider the possible dependency between different types of payments from the same claim. This is a more complex issue that requires complete analysis and allows for innovative methods. We postpone this analysis to a future work where we can better deal with this point.

2.4.2 Fitting the models

This subsection describes the models we considered in our numerical analysisand the choices made regarding estimating parameters and distributions. Each step's choices and thought process are based on Section 2.2.6. As previously stated, two models are required : one for IBNR claims and one for RBNS claims. We thoroughly describe the procedure for RBNS claims and make some remarks concerning the procedure for IBNR claims.

First, we consider a time division vector with an even yearly division between each period : $\boldsymbol{d} = \{0, 1, 2, 3, 4, 5\}$. We chose this division because it is the easiest to interpret since many time divisions in the reserving literature are done year-wise, such as the development periods in a loss triangle. Although, as mentioned before, this model does allow for other time divisions. Second, we select the Poisson and Negative Binomial distributions for our frequency models.

The Negative Binomial (type II) can be described by its mean and variance :

$$\left(N_{i,k}^{(a)} | \mathbf{X}_{i,k}, \mathbf{Z}_{i,k}, E_{i,k}, \boldsymbol{\ell}_{i,k}\right) \sim \text{Neg Bin II}\left(\mu_{i,k}^{(a)}, \sigma\right), \text{ if } E_{i,k} > 0, \text{ for } i \in \mathcal{I},$$

where $\mu_{i,k}^{(a)}$ and σ are such that,

$$\begin{aligned} \text{E}\left[N_{i,k}^{(a)} | \mathbf{X}_{i,k}, \mathbf{Z}_{i,k}, E_{i,k}, \boldsymbol{\ell}_{i,k}\right] &= \mu_{i,k}^{(a)}, \\ \text{Var}\left[N_{i,k}^{(a)} | \mathbf{X}_{i,k}, \mathbf{Z}_{i,k}, E_{i,k}, \boldsymbol{\ell}_{i,k}\right] &= \mu_{i,k}^{(a)}(\sigma + 1). \end{aligned}$$

Note that there is another version of the Negative Binomial distribution (type I)

that will not be considered in this numerical analysis [1]

Finally, for our numerical analysis, we estimate parameters $\boldsymbol{\beta}^{(a)}$, $\boldsymbol{\theta}^{(a)}$, $\boldsymbol{\theta}^{*(a)}$, $\psi^{(a)}$, $\psi^{*(a)}$, $\ell_{max}^{(a)}$, and $\ell_{max}^{*(a)}$ by maximizing the likelihood function for each distribution (Poisson and Negative Binomial), each type of payment (medical, disability and, expenses) and each method to obtain a claim score (**M1**, **M2** and **M3**). A goodness of fit analysis is performed for the models considered in the next section.

2.4.3 Goodness-of-fit analysis

In order to streamline the impact of a claim score in the modeling process, we begin by selecting the best method for computing the claim score among methods **M1**, **M2**, and **M3**. This selection was achieved by comparing the Akaike information criterion (AIC) and the Bayesian (or Schwarz) information criterion (BIC) between these models. Table 2.3 and Table 2.4 contain these results, respectively, for RBNS and IBNR claims.Furthermore, in order to take into account the extra information added by using previous observations, the AIC and BIC are corrected by artificially increasing the number of parameters in the formulae. The increase is based on the maximal number of payments observed in a period, so each payment count observed in the data set is considered a covariate category. Thus, the number of parameters was increased for medical, disability, and expense payments by 208, 107, and 59. In these tables, we notice that models **M1** have consistently the lowest value for both criteria. Henceforth, since this particular data set model **M1** seems to be the most appropriate, future numerical analysis will be done only for this particular model (estimated values of the parameters are available in Appendix 2.6).

---

1. For this distribution the variance is $\text{Var}\left[N_{i,k}^{(a)}|\mathbf{X}_{i,k},\mathbf{Z}_{i,k},E_{i,k},\boldsymbol{\ell}_{i,k}\right]=\mu_{i,k}^{(a)}+\left(\mu_{i,k}^{(a)}\right)^2\sigma$.

TABLEAU 2.3: Likelihood Information Criteria for RBNS models **M1**,**M2** and **M3**

		AIC			BIC		
Dist.	Payment	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
NB	Medical	232,720	232,857	232,752	233,085	233,222	233,117
	Disability	81,099	81,379	81,233	81,464	81,745	81,597
	Expenses	122,865	122,920	122,888	123,230	123,285	123,252
POI	Medical	331,810	332,793	332,342	332,165	333,148	332,698
	Disability	203,585	205,819	204,226	203,940	206,174	204,581
	Expenses	160,369	160,608	160,505	160,725	160,963	160,861

Our main goal in this subsection is to assess the performance of including the claim score $\boldsymbol{\ell}_{i,k}$ into frequency models in terms of goodness-of-fit. Hence, we suggest comparing the AIC and BIC of two versions of our models. The first version will include $\boldsymbol{\ell}_{i,k}$ as a covariate, and the second version will not.We compare the results for RBNS, IBNR, and a covariate-free model). We present these results in Table 2.5, Table 2.6 and Table 2.7. As shown in these tables, including the claim scores provides better BIC and AIC across all models and all types of payments.

With the same goal in mind, we perform a likelihood ratio test between the models that use it and those that do not. We present the results in Table 2.8, where the claim score for the covariate-free model is given by $\boldsymbol{\ell}_{i,k}^{\otimes}$. Given low p-values, we can confidently reject all restricted models, i.e., models that do not include a claim score.

Then, we perform t-tests specifically for the parameter of the dynamic claim score, $\gamma^{(a)}$, for each RBNS model. The results are in Table 2.9. Again, with very low p-values, we can determine that the claim score is significant as a covariate.

TABLEAU 2.4: Likelihood Information Criteria for IBNR models **M1**,**M2** and **M3**

		AIC			BIC		
Dist.	Payment	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
NB	Medical	238,606	238,774	238,613	238,707	238,875	238,713
	Disability	82,481	82,951	82,698	82,581	83,056	82,798
	Expenses	130,755	131,289	130,954	130,855	131,390	131,054
POI	Medical	348,062	349,250	348,669	348,153	349,341	348,760
	Disability	217,424	221,672	219,138	217,515	221,764	219,229
	Expenses	185,596	187,235	186,242	185,687	187,326	186,333

Having assessed the increase in terms of goodness of fit, through the AIC, the BIC, the likelihood ratio test, and the Student t-test, we can also observe how changes in the dynamic claim score affect the mean of payment counts by plotting its relativity, i.e.,

$$\exp\left(\gamma^{(a)}\ell\right), \text{ for } -5 < \ell \leq \ell_{max}^{(a)},$$

for the suggested distributions. Where the claim score is bounded by its maximal value ($\ell_{max}^{(a)}$), and the value given by the maximal number of consecutive payment-free periods (i.e., $\kappa_{i,6} = 5$, leading to $\ell_{i,6} = -5$) . Figures 2.3, 2.4 and 2.5 depict these results for RBNS payments. We notice that the dynamic claim scoreimpacts the mean parameter, particularly in the extremes. For instance, the lowest increase of the mean parameter for a claim that has reached its maximum score comparatively to a claim with no history, i.e., having a score equal to zero, is 3.44 times as high (by considering the Negative Binomial for expense payments). In contrast, the highest comparative increase is 12.79 times as high (by considering the Poisson for disability payments).

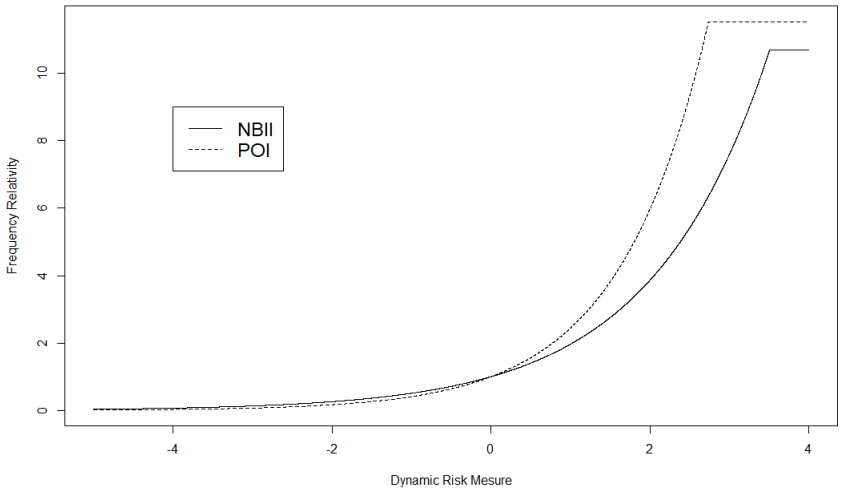

FIGURE 2.3 Relativity of the dynamic risk score to the mean of medical RBNS payments

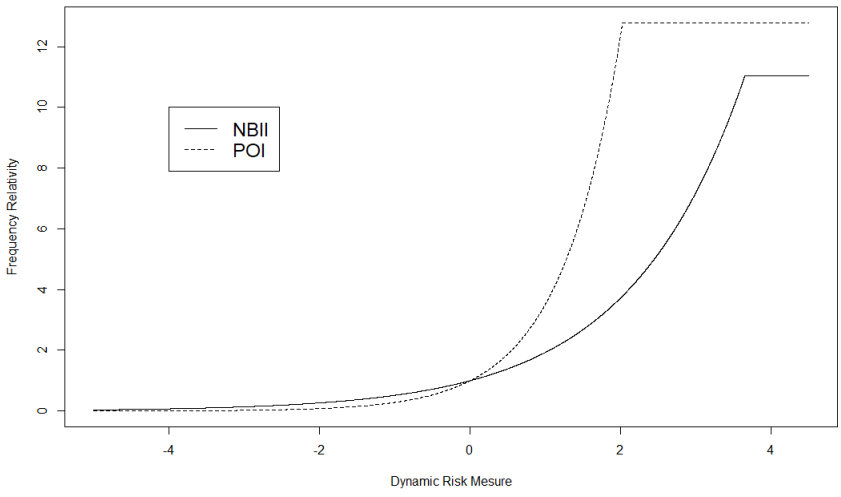

FIGURE 2.4 Relativity of the dynamic risk score to the mean of disability RBNS payments

Tableau 2.5: AIC and BIC of RBNS models with and without the claim score

		AIC		BIC	
Model	Payment type	with	without	with	without
NB	Medical	233,136	236,794	235,397	237,149
	Disability	81,313	84,576	82,653	84,931
	Expenses	122,981	123,964	123,875	124,320
POI	Medical	332,226	358,342	334,477	358,688
	Disability	203,799	240,730	205,129	241,077
	Expenses	160,485	164,481	161,370	164,828

2.4.4 Simulation analysis

We continue our numerical analysis by simulating the number of outstanding payments for each claim. By repeating algorithms described in Section 2.3 10 000 times we obtain predicted values for the frequency of payments for all our models. We summarize our IBNR, RBNS, and total reserves results in Tables 2.10 and 2.11. These tables contain results for models that use the dynamic claim score and those that do not.

Regarding the exposure, we see it is very well adjusted to the observed value of the RBNS claims : both the mean and the values-at-risk are close. Furthermore, when considering the total reserve, we include the IBNR claims, which reduces the accuracy of the exposure predictions, where 99% VaR is slightly under the observed value. We can infer that the model is less accurate when handling IBNR claims. The lack of information from IBNR claims regarding covariates and history can explain these results.

Next, we focus on frequency models. For these results, we want to compare the

Tableau 2.6: AIC and BIC of IBNR models with and without the claim score

		AIC		BIC	
Model	Payment type	with	without	with	without
NB	Medical	239,022	243,361	241,019	243,453
	Disability	82,695	86,557	83,770	86,648
	Expenses	130,871	132,828	131,500	132,920
POI	Medical	348,478	380,520	350,465	380,602
	Disability	217,638	263,932	218,704	264,014
	Expenses	185,712	193,395	186,332	193,477

results between frequency models that include the dynamic claim score to ones that do not. This analysis is done for the total number of payments and each type of payment. We will begin by looking at the results from medical payments, which represent most of the total. For these payments, including the claim score significantly brings the 95% and 99% VaR and mean values closer to the observed value, indicating a significant improvement. Next, in terms of RBNS disability payments, the inclusion of the claim score in the Negative Binomial model allows for the 95% and 99% VaR to be over the observed value.This result only occurs when the claim score is included. However, we do not see this improvement when considering the Poisson distribution. Finally, regarding the expense payments, both models without and with claim scores provide 95% and 99% VaR over the observed value ; however, the latter models tend to be more conservative with higher results mean and VaR. Overall, all types of payments are not impacted similarly, but their combined value is greatly improved when the claim score is included ; without it, the Values-at-Risk considerably fall below the observed value.

After analyzing the frequency models, we can compare the best-performing model

TABLEAU 2.7: AIC and BIC of no covariate models with and without the claim score

		AIC		BIC	
Model	Payment type	with	without	with	without
NB	Medical	240,854	244,112	242,778	244,130
	Disability	82,759	86,962	83,861	86,980
	Expenses	131,286	133,477	131,842	133,496
POI	Medical	356,122	381,063	358,037	381,073
	Disability	217,913	269,424	218,907	269,433
	Expenses	186,470	194,055	187,017	194,064

(the one that uses the Negative Binomial distribution) to other models in the literature. However, because most models directly predict the total cost of payments rather than payment counts, we choose to compare this value instead. Thus, we must add a severity model to our dynamic score frequency model. We test popular Gamma, log-Normal, and inverse-normal distributions. We find that fitting each payment type separately and including the claim score as a covariate is satisfactory, and the Gamma distribution was chosen for this numerical analysis. As for the comparative distributions, we chose two collective generalized linear models based on the quasi-Poisson distribution and the Gamma distribution (for more details, see Wüthrich & Merz (2008)). We also consider the individual model by Yanez & Pigeon (2021), which serves as a comparative baseline for including dynamic claim scores. Thus, regarding information used by the models, the GLM models only use the accident year and the development year from a loss triangle. The 3-component model incorporates covariate information from the claims (see Table 2.1), while the Dynamic Score model uses the same information as the 3-component model in addition to a dynamic claim score. Table 2.12 contains the

TABLEAU 2.8: Likelihood Ratio (L. R.) test RBNS and IBNR models with and without the dynamic claim score

Model	Payment Type	Restricted Covariates	Unrestricted Covariates	L.R.	p-value
NB	Medical	$\mathbf{X}_i, \mathbf{Z}_{i,k}$	$\mathbf{X}_i, \mathbf{Z}_{i,k}, \boldsymbol{\ell}_{i,k}$	4075	< 0.01
	Disability	$\mathbf{X}_i, \mathbf{Z}_{i,k}$	$\mathbf{X}_i, \mathbf{Z}_{i,k}, \boldsymbol{\ell}_{i,k}$	4085	< 0.01
	Expenses	$\mathbf{X}_i, \mathbf{Z}_{i,k}$	$\mathbf{X}_i, \mathbf{Z}_{i,k}, \boldsymbol{\ell}_{i,k}$	26,534	< 0.01
POI	Medical	$\mathbf{X}_i, \mathbf{Z}_{i,k}$	$\mathbf{X}_i, \mathbf{Z}_{i,k}, \boldsymbol{\ell}_{i,k}$	1100	< 0.01
	Disability	$\mathbf{X}_i, \mathbf{Z}_{i,k}$	$\mathbf{X}_i, \mathbf{Z}_{i,k}, \boldsymbol{\ell}_{i,k}$	945	< 0.01
	Expenses	$\mathbf{X}_i, \mathbf{Z}_{i,k}$	$\mathbf{X}_i, \mathbf{Z}_{i,k}, \boldsymbol{\ell}_{i,k}$	4113	< 0.01
NB	Medical	$\mathbf{Z}^*_{i,k}$	$\mathbf{Z}^*_{i,k}, \boldsymbol{\ell}^*_{i,k}$	4757	< 0.01
	Disability	$\mathbf{Z}^*_{i,k}$	$\mathbf{Z}^*_{i,k}, \boldsymbol{\ell}^*_{i,k}$	4232	< 0.01
	Expenses	$\mathbf{Z}^*_{i,k}$	$\mathbf{Z}^*_{i,k}, \boldsymbol{\ell}^*_{i,k}$	32,459	< 0.01
POI	Medical	$\mathbf{Z}^*_{i,k}$	$\mathbf{Z}^*_{i,k}, \boldsymbol{\ell}^*_{i,k}$	2075	< 0.01
	Disability	$\mathbf{Z}^*_{i,k}$	$\mathbf{Z}^*_{i,k}, \boldsymbol{\ell}^*_{i,k}$	1593	< 0.01
	Expenses	$\mathbf{Z}^*_{i,k}$	$\mathbf{Z}^*_{i,k}, \boldsymbol{\ell}^*_{i,k}$	7801	< 0.01

results of $10,000$ simulations of each described model, and Figure 2.6 displays the results.

Lets discuss the results from Table 2.12 and Figure 2.6. All the models yield satisfactory results regarding the 95 % and the 99 % VaRs as the values are higher than the observed value. The two collective models (Gamma and over-dispersed Poisson) have a mean lower than the observed value, but their standard deviation is higher than the individual models. Furthermore, the latter approaches are preferable because the 95 % and the 99 % Values-at-Risk of individual models are lower than the collective models but higher than the observed value. As for

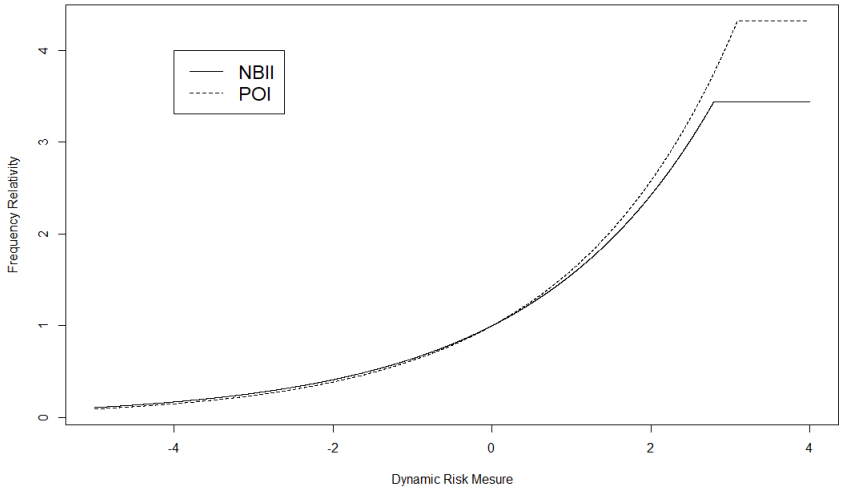

FIGURE 2.5 Relativity of the dynamic risk score to the mean of expense RBNS payments

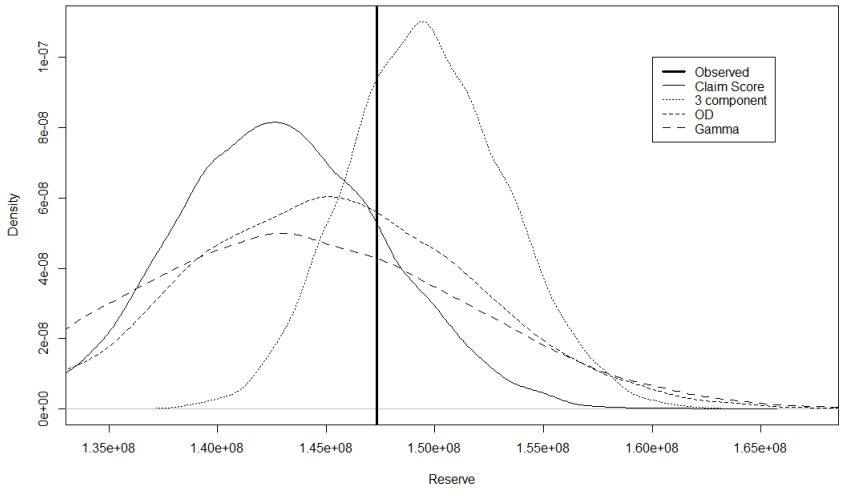

FIGURE 2.6 Total reserves for selected models

TABLEAU 2.9: Student's t-test for parameter $\gamma^{(a)}$ for RBNS models with the dynamic claim score

Dist.	Negative Binomial			Poisson		
Payment	Medical	Disability	Expenses	Medical	Disability	Expenses
t-value	77.03	62.88	33.74	167.86	156.24	63.74
p-value	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

the comparison between both individual approaches, we notice that the mean of the total reserve is lower for the dynamic score model ; however, through a higher standard deviation, the 95 % and the 99 % VaRs become lower than the model that does not make use of the claim score. This further increases the model's utility by providing values higher than the observed reserve but lower than the other predictions. Again, this shows an overall numerical preference for the model in this paper over the one suggested in Yanez & Pigeon (2021). Thus, for our data set, including more detailed information improves the reserve predictions. First, consider an individual approach (3-component model) rather than a collective approach (GLM models). Then, by also incorporating past payment counts information through the Dynamic Claim score model.

2.5 Conclusion

This paper introduced an innovative dynamic claim score to the loss reserve literature. This score allows for including past individual claim development in the fitting process of outstanding payment counts. We could feed this score information at the end of each interval through an interval-based approach and use this updated information for the next interval. We applied this new method to the model by Yanez & Pigeon (2021) because of the discrete nature of its payment count

modeling and the ease of covariate implementation it allows. However, any model that can predict payment counts at different time development states may incorporate the claim score introduced in this paper. Furthermore, we expanded the scope of payment count modeling by proposing a structure considering different payment types.

We applied the above-mentioned model to a data set in our numerical analysis. We showed that including a dynamic claim score improves the performance of traditional count models (such as the Poisson and Negative Binomial models) regarding goodness-of-fit. Then, we compared the predictions of outstanding payment counts between models that use this new score and models that do not, and we obtained an overall improvement of the predictions. Finally, we showed that our new approach yields better results than collective and individual models in the literature.

As mentioned before, we introduced claim scores to the micro-level loss reserving literature in this paper. Thus, given the pioneering nature of our work, it can branch out into many extensions for various contexts. In particular, we supposed that the claims of the portfolio are independent. However, a casualty may trigger different claims and thus be dependent due to their shared origin. This complex subject should be considered and will be studied in a future project. Another correlation problem that was not addressed is the dependence between the number of payments and their cost. Here claim scores could prove useful if they are computed for both the frequency and the severity of payments.

TABLEAU 2.10: Simulation results for RBNS outstanding payment counts from models with and without claim scores

Claim	Dist.	Payment	Mean	Std dev	95% VaR	99% VaR	Obs.
Score	WEI	Exposure	5893	52	5979	6015	5889
with	NB	Medical	48,941	837	50,309	50,938	51,565
		Disability	21,087	813	22,419	23,028	20,601
		Expenses	22,905	425	23,599	23,902	16,653
		Total	92,932	1452	95,299	96,303	88,819
	POI	Medical	50,749	686	51,888	52,384	51,565
		Disability	16,727	434	17,430	17,713	20,601
		Expenses	20,607	283	21,075	21,259	16,653
		Total	88,084	1126	89,945	90,669	88,819
without	NB	Medical	38,426	588	39,384	39,801	51,565
		Disability	18,519	626	19,563	20,018	20,601
		Expenses	20,820	359	21,405	21,688	16,653
		Total	77,765	1088	79,554	80,344	88,819
	POI	Medical	42,420	441	43,141	43,444	51,565
		Disability	17,464	243	17,862	18,028	20,601
		Expenses	18,277	229	18,657	18,805	16,653
		Total	78,161	797	79,487	79,970	88,819

TABLEAU 2.11: Simulation results for the total outstanding payment counts from models with and without claim scores

Claim	Dist.	Payment	Mean	Std dev	95% VaR	99% VaR	Obs.
Score	WEI	Exposure	6275	57	6369	6409	6454
with	NB	Medical	51,922	870	53,360	53,973	54,986
		Disability	21,885	825	23,248	23,843	21,620
		Expenses	24,054	440	24,780	25,085	18,080
		Total	97,861	1501	100,308	101,291	94,686
	POI	Medical	53,473	706	54,641	55,137	54,986
		Disability	17,360	437	18,066	18,354	21,620
		Expenses	21,570	291	22,051	22,236	18,080
		Total	92,403	1157	94,314	95,067	94,686
without	NB	Medical	41,480	629	42,501	42,951	54,986
		Disability	19,529	648	20,612	21,102	21,620
		Expenses	22,031	375	22,647	22,928	18,080
		Total	83,039	1157	84,940	85,735	94,686
	POI	Medical	45,353	479	46,143	46,481	54,986
		Disability	18,237	250	18,645	18,816	21,620
		Expenses	19,296	241	19,695	19,855	18,080
		Total	82,886	852	84,288	84,843	94,686

Tableau 2.12: Results of the total reserve predictions

	Mean	Std. dev.	95% VaR	99% VaR
GLM Gamma	143,604,545	7,969,902	156,696,768	162,534,340
GLM ODP	145,171,862	6,565,836	156,112,224	161,073,565
3 Component RBNS	145,459,940	3,636,952	151,546,231	154,130,897
3 Component	149,620,225	3,678,054	155,830,382	158,291,786
Claim Score RBNS	137,509,168	4,785,344	145,451,829	148,969,071
Claim Score total	142,852,107	4,842,791	150,950,931	154,342,708
Obs. RBNS	141,830,856			
Obs. total	147,308,364			

2.6 Appendix

TABLEAU 2.13: Estimated values for the Negative Binomial (type II) Model (RBNS)

Variable	Category	With the claim score			Without the claim score		
		Medical	Disability	Expenses	Medical	Disability	Expenses
Intercept		1.66	-0.56	-1.65	1.75	-0.55	-1.75
$\sigma^{(a)}$		1.46	2.70	1.08	1.59	2.93	1.13
$\gamma^{(a)}$		0.44	0.61	0.37			
$\psi^{(a)}$		0.14	0.16	0.18			
$\ell_{max}^{(a)}$		4.70	3.77	3.12			
Loss	Single vehicle	0.08	0.44	0.46	0.09	0.53	0.48
	Multi vehicle	0.22	0.07	0.33	0.23	0.06	0.34
	Hit pedestrian	0.32	0.67	0.72	0.40	0.80	0.79
	Other	0.36	0.47	0.46	0.41	0.57	0.50
Gender	Male	-0.16	0.12	0.10	-0.18	0.08	0.10
	Unknown	-0.06	0.87	0.94	-0.08	0.93	0.96
Region	Ontario	-0.12	0.30	1.88	-0.19	0.37	1.99
	West	0.50	0.44	0.48	0.45	0.51	0.57

Variable	Category	With the claim score			Without the claim score		
		Medical	Disability	Expenses	Medical	Disability	Expenses
Age	$(18, 25]$	0.06	0.54	0.30	0.07	0.63	0.32
	$(25, 30]$	0.20	0.62	0.36	0.20	0.70	0.37
	$[30, 50]$	0.23	0.59	0.38	0.25	0.67	0.40
	$(50, 70]$	0.30	0.60	0.50	0.33	0.67	0.51
	$(70, \infty)$	0.36	0.56	0.65	0.41	0.64	0.70
	Unknown	-0.11	-0.66	-0.32	-0.09	-0.61	-0.32
Vehicle age	$(3, 6]$	0.01	0.07	-0.01	0.00	0.09	-0.01
	$(6, 10]$	0.03	0.12	0.06	0.04	0.15	0.06
	$(10, 20]$	0.05	0.27	0.16	0.05	0.32	0.17
	$(20, \infty)$	-0.00	0.47	0.15	-0.01	0.59	0.15
	Unknown	0.00	0.00	-0.01	-0.03	-0.05	-0.03

Variable	Category	With the claim score			Without the claim score		
		Medical	Disability	Expenses	Medical	Disability	Expenses
$t_\ell^{(r)}$	$(1, 7]$	-0.03	0.14	0.11	-0.03	0.14	0.12
	$(7, 30]$	-0.12	-0.06	0.08	-0.14	-0.13	0.06
	$(30, 90]$	-0.30	-0.35	0.00	-0.36	-0.50	-0.04
	$(90, 180]$	-0.69	-0.57	-0.07	-0.79	-0.85	-0.15
	$(180, 365]$	-0.74	-0.57	-0.14	-0.85	-0.83	-0.28
	$(365, \infty)$	-1.10	-0.74	-0.02	-1.25	-1.07	-0.09
Initial reserve	$(1000, 5000]$	-0.10	-0.25	-0.39	-0.12	-0.27	-0.40
	$(5000, 10000]$	-0.05	0.17	-0.11	-0.08	0.18	-0.12
	$(10000, 20000]$	-0.07	0.26	0.01	-0.11	0.24	0.01
	$(20000, \infty)$	0.00	0.55	0.12	-0.03	0.74	0.13
Time intervals	$(1, 2]$	-0.68	0.17	-0.24	-0.10	0.45	0.03
	$(2, 3]$	-1.07	-0.14	-0.28	-0.18	0.43	0.16
	$(3, 4]$	-1.33	-0.16	-0.35	-0.25	0.37	0.18
	$(4, 5]$	-1.42	0.28	-0.10	-0.21	1.08	0.48

TABLEAU 2.14: Estimated values for the Poisson Model (RBNS)

Variable	Category	With the claim score			Without the claim score		
		Medical	Disability	Expenses	Medical	Disability	Expenses
Intercept		1.72	-0.66	-1.78	1.76	-0.79	-1.90
	$\gamma^{(a)}$	0.65	1.19	0.39			
	$\psi^{(a)}$	0.10	0.09	0.20			
	$\ell_{max}^{(a)}$	3.54	2.09	3.59			
Loss	Single vehicle	0.26	0.68	0.47	0.33	0.81	0.49
	Multi vehicle	0.26	0.25	0.29	0.27	0.25	0.28
	Hit pedestrian	0.61	0.92	0.80	0.77	1.10	0.87
	Other	0.44	0.62	0.49	0.50	0.75	0.52
Gender	Male	-0.14	0.09	0.09	-0.16	0.06	0.09
	Unknown	0.12	1.17	0.82	0.11	1.27	0.82
Region	Ontario	-0.17	0.76	2.22	-0.24	0.89	2.35
	West	0.23	0.36	0.60	0.16	0.48	0.70

Variable	Category	With the claim score			Without the claim score		
		Medical	Disability	Expenses	Medical	Disability	Expenses
Age	$(18, 25]$	0.07	0.17	0.29	0.08	0.18	0.32
	$(25, 30]$	0.20	0.27	0.36	0.21	0.25	0.38
	$[30, 50]$	0.23	0.29	0.37	0.25	0.27	0.40
	$(50, 70]$	0.29	0.34	0.47	0.31	0.29	0.49
	$(70, \infty)$	0.37	0.41	0.64	0.40	0.42	0.67
	Unknown	-0.22	-0.95	-0.38	-0.21	-1.04	-0.37
Vehicle age	$(3, 6]$	0.03	0.15	-0.00	0.03	0.17	-0.00
	$(6, 10]$	0.05	0.18	0.07	0.06	0.24	0.07
	$(10, 20]$	0.07	0.32	0.19	0.09	0.38	0.21
	$(20, \infty)$	0.12	0.53	0.15	0.14	0.73	0.16
	Unknown	-0.03	-0.04	-0.02	-0.07	-0.12	-0.03

Variable	Category	With the claim score			Without the claim score		
		Medical	Disability	Expenses	Medical	Disability	Expenses
$t_\ell^{(r)}$	$(1, 7]$	-0.04	0.06	0.04	-0.04	0.09	0.03
	$(7, 30]$	-0.15	-0.13	-0.04	-0.19	-0.19	-0.07
	$(30, 90]$	-0.33	-0.50	-0.17	-0.42	-0.67	-0.22
	$(90, 180]$	-0.69	-0.71	-0.30	-0.85	-1.07	-0.41
	$(180, 365]$	-0.68	-0.74	-0.35	-0.86	-1.10	-0.49
	$(365, \infty)$	-0.58	-0.74	-0.18	-0.74	-1.07	-0.27
Initial reserve	$(1000, 5000]$	-0.18	-0.24	-0.31	-0.21	-0.28	-0.32
	$(5000, 10000]$	-0.04	0.17	-0.04	-0.07	0.12	-0.06
	$(10000, 20000]$	0.00	0.30	0.10	-0.04	0.22	0.10
	$(20000, \infty)$	0.23	0.70	0.29	0.28	0.90	0.32
Time intervals	$(1, 2]$	-0.81	-0.31	-0.50	0.03	0.45	-0.07
	$(2, 3]$	-1.25	-0.69	-0.66	0.03	0.39	0.01
	$(3, 4]$	-1.57	-0.52	-0.77	-0.11	0.58	0.04
	$(4, 5]$	-1.68	-0.25	-0.70	-0.17	0.97	0.12

TABLEAU 2.15: Estimated values for the Negative Binomial (type II) Model (IBNR)

Variable	Category	With the claim score			Without the claim score		
		Medical	Disability	Expenses	Medical	Disability	Expenses
Intercept		1.93	0.54	0.68	1.95	0.69	0.72
Time intervals	$(1,2]$	-0.83	0.35	0.08	-0.23	0.51	0.32
	$(2,3]$	-0.83	0.42	0.40	-0.34	0.52	0.50
	$(3,4]$	-0.85	0.60	0.54	-0.42	0.44	0.54
	$(4,5]$	-0.71	0.97	0.94	-0.31	1.24	0.86
$\sigma^{(a)}$		1.53	2.71	1.33	1.68	2.99	1.41
$\gamma^{(a)}$		0.62	0.74	0.70			
$\psi^{(a)}$		0.12	0.13	0.12			
$\ell_{max}^{(a)}$		3.37	3.57	2.11			

TABLEAU 2.16: Estimated values for the Poisson Model (IBNR)

Variable	Category	With the claim score			Without the claim score		
		Medical	Disability	Expenses	Medical	Disability	Expenses
Intercept		1.94	0.67	0.81	1.91	0.64	0.82
Time intervals	$(1,2]$	-0.90	0.01	-0.30	-0.01	0.62	0.17
	$(2,3]$	-0.87	-0.02	-0.03	-0.03	0.60	0.30
	$(3,4]$	-0.91	0.23	0.08	-0.18	0.80	0.33
	$(4,5]$	-0.87	0.41	0.32	-0.19	1.24	0.39
$\gamma^{(a)}$		0.85	1.32	0.73			
$\psi^{(a)}$		0.09	0.07	0.15			
$\ell_{max}^{(a)}$		2.88	2.23	2.47			

CHAPITRE III

ANALYSE DE LA DÉPENDANCE DU TEMPS DE TRAITEMENT DES RÉCLAMATIONS : UNE PERSPECTIVE BASÉE SUR DES EFFETS ALÉATOIRES *FRAILTY*

3.1 Introduction

Access to more in-depth information regarding claims has recently sparked an interest in micro-level models for loss reserving. Traditionally, models were studied through the lenses of a macro-level structure, aggregating claims according to their accident and development years. This structure provides a more compact data set that one can handle with less computationally intensive methods such as the well-known Chain Ladder model (see Mack (1999) and Mack (1993)) ; however, it renders distinctions between individual claims impossible (see Wüthrich & Merz (2008) and England & Verrall (2002) for a comprehensive compendium of these models). In contrast, micro-level models forgo the traditional triangular structure associated with macro-level models by separating claims from one another rather than aggregating them. The aim is to predict the full development of each outstanding claim. This development comprises various events : the occurrence of an accident covered by a policy, the reporting of the claim by the client to the insurance company, a series of cash flows, and finally, the settlement of the claim. As such, models that use data in this granular form require various conside-

rations for the events that comprise the full development of a given claim. Despite the added complexity due to multiple elements to be considered when working on micro-level data, the structure allows for the usage of individual information in the modeling process, which, when considered, has proven to improve the quality of models that predict the overall development of claims (see Wang *et al.* (2021) and Yanez & Pigeon (2021)). In contrast, macro-level models cannot use this information due to the aggregated nature of the collected data.

Many suggestions have been put forward in the micro-level reserving literature. The very first models were introduced by Arjas (1989), Norberg (1993), Haastrup & Arjas (1996), and Norberg (1999). These propositions focused on using a dependent Marked-Poisson Process (DMPP) to predict future claim events such as closures or payments. Interest in this model type was revitalized years later by Antonio & Plat (2014), where a more practical application of a DMPP was considered. Other propositions focused on expanding the concept of development factors from the Chain Ladder model (see Mack (1999) and Mack (1993)) into individual development factors to propose an original micro-level model in Pigeon *et al.* (2013) and Pigeon *et al.* (2014). In addition, semi-parametric methods using a Cox proportional hazards model have been suggested in Zhao *et al.* (2009), Zhao *et al.* (2010), Badescu *et al.* (2016), and Badescu *et al.* (2019) for micro-level incurred but not reported claims (IBNR) claims. Several authors have recently focused on machine learning methods in various forms. For instance, regression tree methods have been proposed by Wüthrich (2018), Lopez *et al.* (2016), Lopez (2019), and Lopez *et al.* (2019). Neural networks are another example, as seen in Kuo (2019), Gabrielli (2020), and Delong *et al.* (2022). For a more in-depth review of the machine learning methods in the micro-level loss reserving literature, we recommend the work done by Blier-Wong *et al.* (2020).

An essential part of the modeling process of micro-level reserves is the prediction

of **the processing time** for outstanding claims, that is, the delay between the occurrence of a claim and its settlement. It is within this time frame that the reporting date and the cash flows that constitute the loss reserve are observed. As an essential component, the study of this time frame has been tackled with different techniques. Notably, in the context of IBNR claims, the reporting delay (the first portion of the processing time) has been studied in Verrall & Wüthrich (2016), where a three-layer approach was considered. In Antonio & Plat (2014), a Weibull distribution with degenerate components was considered. As for the settlement date predictions, Pigeon *et al.* (2013) suggested a discrete framework by using a mixture of a Geometric distribution with degenerate components to count for how many periods a claim is open. Similarly, Denuit & Lu (2021) put forward the Wishart-Gamma distribution. Other models include settlements as terminal events in the context of count processes (as in Haastrup & Arjas (1996)).

Due to the increasing popularity of micro-level modeling, suggestions with data-driven specifications have been addressed. The one of interest in the context of this paper is dependency, which can be studied between the various parts that constitute a micro-level model. For instance, some authors have used past information of a given claim to address their future development. In Antonio *et al.* (2015), their multi-state structure allowed for factors to be updated to their current state based on past information. With a similar objective in mind, in Yanez *et al.* (2023), a dynamic claim score that summarizes past information was incorporated into the modeling process. Other authors have focused on the inclusion of dependency between the settlement date and payments (see Okine *et al.* (2022) and Denuit & Lu (2021)). In contrast to these propositions, in Avanzi *et al.* (2021), claim counts from distinct lines of business are modeled through dependent Cox processes where the dependency structure is constructed through multivariate shot

noise intensity processes.

The inspiration for this paper arises from a practical problem that actuaries may encounter when dealing with detailed data sets. An accident may often affect various coverages of a client's insurance policy, which may have their development structure or covariates. For example, suppose the insurer covers the medical fees of two individuals injured in a given accident. It is tricky to incorporate specific information from these individuals (such as their age or the type of injury) to predict the development of the claim that encompasses them because two distinct data sets are available. As such, modeling coverages rather than claims to smoothly incorporate coverage-specific information may be rich in predicting power. However, in doing so, the issue of dependency arises.

In this context, we focus on one of the main variables that drive the cost of claims, the processing time. In particular, the time-to-event nature of this variable allows us to study it through the lens of the abundant literature on survival models. Among the various propositions, the Cox proportional hazards model (the Cox model, see Cox (1972)) and the models that derive from it stand out as viable options to predict processing time. Specifically, their definition through hazard rates allows for streamlined incorporation and analysis of covariate information. Indeed, a core element of our motivation relies on the quality of individual coverage information, mainly the information that diverges between coverages of the same claim. If it is not statistically significant, one could have a solid incentive to omit it and merge all coverages into a single cash flow.

Another advantage of the Cox model is the well-documented implementation of random effects to incorporate unobserved heterogeneity among observations, also known as Frailty effects. In our case, by identifying claims as clusters of coverages and linking them through a common Frailty, we can address the within-claim

coverage dependence of a portfolio. Furthermore, given the widespread and diverse uses of Cox models, several extensions, versions, and numerical applications have been added over the years (see Balan *et al.* (2020) for an extensive review). For instance, the hazard rate of the model has been redefined using M-splines (see Rondeau *et al.* (2012)). It has been restructured as a fully parametric model through the Weibull distribution (see Byar (1982)). As for the Frailty, the Gamma distribution is usually considered, as significant results can be obtained in closed forms. Other options, such as the Lognormal distribution, are also available (Balan *et al.* (2020)).

This paper provides an original solution to including coverage-specific covariates in the modeling process of micro-level reserves. Among the various elements that constitute the development of claims (such as payment counts), we direct our attention to the processing time, as it plays a crucial role in the calculation of loss reserves (see Yanez & Pigeon (2021)). Furthermore, by focusing on predicting a terminal event (such as the settlement of a claim), we can draw from the literature well-supported models such as the Cox proportional hazard model and the Weibull-Cox model. These models allow for the inclusion of Frailty random effects, which, when shared by members of the same cluster, provide a practical solution to incorporating within-claim dependence. We sustain the viability and advantages provided by the methods we cover by assessing the quality of the coverage-specific information and by measuring the correlation between members of the same cluster.

Our contribution is meticulously explained and justified through four main sections. In the first one, we justify the motivation of this paper by giving various statistical measures of intra-claim correlation. Then, in the second section, we review key definitions of the Cox propositional hazard model from the survival literature to provide a statistical framework for the processing time of coverages.

Next, in the third section, we extensively review the inference and simulation procedures developed over the years to incorporate Frailty random effects into the baseline models described in the first section. Then, in the fourth section, we provide a numerical application of our propositions and goodness-of-fit analyses. Finally, in the last section, we conclude and provide some extensions to our models.

3.2 Preliminary numerical analysis

This paper's primary motivation is to capture the dependency between the processing times of coverages originating from the same claim. Although seemingly intuitive, at least in the data set used for this work, we must show evidence of this correlation. As such, before we develop the statistical framework of the models to be considered, we dedicated this first section to a preliminary numerical analysis. It is done twofold, first by describing the data set and its variables, then by computing and analyzing correlation measures.

3.2.1 Data set description

In our data set, 43,951 Accident Benefits claims (AB) cover no-fault benefits for accidents where the driver, or a third party, was injured or killed in a car accident. Also, whenever the insurer must pay for a claim involving multiple parties, cash flows stemming from each individual are distinguished. We treat each cash flow as distinct coverages from the same claim. Thus, given that a claim may contain more than one coverage, the portfolio has a more significant number of coverages (57,593), all originating from accidents between 2011 and 2015. Furthermore, the development of each coverage is available until December 31, 2017. Then, by setting December 31, 2015, as the valuation date, we can observe that

48,855 coverages get settled, while 7,872 coverages remain open, and 866 coverages are considered unavailable because they are unreported. The development up to the valuation date of the combined 56,727 coverages that are either settled or open at the valuation date will serve as our training set. In contrast, we will use the remaining development (from December 31, 2015, to December 31, 2017) of the open claims to test our models.

Our data set benefits from individual information regarding the accident that resulted in a claim. This information becomes available in the form of categorical covariates. We only consider static covariates in our numerical analysis. Moreover, to use as much of the available data as possible, we add a "not available" category to each covariate rather than removing observations that contain missing covariate values. Table 3.1 summarises descriptive statistics for all covariates.

Tableau 3.1: Description of covariates

Covariate	Label	Number of levels
Gender	Gender of the injured/killed	3
Region	Geographical region	3
Type of loss	Kind of AB claim	5
Vehicle age	Age of the vehicle	6
Injured age	Age of the injured/killed	7
Reporting delay	Delay calculated in days	7
Initial reserve	Reserve at report date	5

In the context of our data set, coverages are divided among the parties involved in a covered accident. Indeed, recall that particular covariate information regarding each individual is considered in the fitting process (such as the age of the person), and it is because of this consideration that claims are divided into various coverages. Thus, given our particular coverage division, there is no unique way to

identify, for example, the "first" from the "second" coverage of each claim in the portfolio. Other circumstances which are not covered in this numerical example, such as clusters involving Accident Benefit (AB) and Bodily Injury (BI) coverages, would have a discernible way to identify the first (AB) from the second (BI) coverage. Nonetheless, to compute rank correlation statistics, we look for a way to label the coverages of claims through a uniform definition to designate, for example, a "first" and a "second" coverage. A seemingly natural choice to define these labels involves ordering them according to their reporting delay, given that we aim to predict processing time. However, among claims with two or more coverages, 76 % contain one or more matching reporting delays (most claim coverages are reported simultaneously), making the reporting delay unsuitable for distinguishing coverages. We looked into other possible features among Table 3.1, where we settled on the initial reserve (covariate *Initial Reserve*) as a labeling feature so that coverages are designated by their apparent severity. The coverage with the higher value is labeled as the first coverage of a claim, the second highest as the second coverage, and so on. Whenever two or more values are identical, they are randomly ordered. In addition, to add another perspective to our analysis, we also considered ordering coverages by the parties' birth dates, from youngest to oldest. Although less intuitive than the *Initial Reserve*, it is the only other covariate that does not have a considerable number of matching values.

We summarize the composition of claims in the portfolio in terms of their size (or the number of coverages) in Table 3.2. We notice that 77.54 % of claims have only one coverage in development, and among the 22.46 % other claims, their size varies (they can contain up to ten coverages). Furthermore, regarding the origin of coverages, 40.76 % are part of cluster size two or greater claims. These values indicate that, although most claims in the portfolio have only one coverage, a significant number of coverages belong to claims with multiple observations, and

a potential issue in terms of dependence should be considered.

TABLEAU 3.2: Claim cluster size in terms of coverage count

Cluster size	1	2	3	4	5	6	7-10
# of Claims	33,626	7,360	1,525	582	186	57	28
% in Portfolio	77.54%	16.97%	3.52%	1.34%	0.43%	0.13%	0.06%
Weight	59.27%	25.94%	8.06%	4.10%	1.64%	0.60%	0.39%

3.2.2 Correlation analysis

In this sub-section, we perform a non-parametric within-cluster correlation analysis. To measure this correlation, we compute two well-known rank correlation statistics : Spearman's rank correlation coefficient (also known as Spearman's ρ) and Kendall's rank correlation coefficient (also known as Kendall's τ). In the context of this paper, there are two issues when we apply these methods. First, there are claims with more than two coverages ; therefore, these tools only provide a preliminary 2-dimensional dependence analysis. The second issue is that data is censored by the valuation date. Fortunately, this is a common situation when working with survival data, and methods have been suggested to incorporate censoring into calculating correlation rank statistics. On the one hand, the proposition by Eden *et al.* (2022) uses Dabrowska's non-parametric method (Dabrowska (1988)) to compute the marginal and the joint survival probability estimates, which then can be used to calculate estimates for the Spearman's ρ statistics. On the other hand, Akritas *et al.* (1996) put forward an extension of the Theil-Sen (Thiel (1950) and Sen (1968)) non-parametric regression model, called the Akritas-Theil-Sen (ATS) regression, to accommodate censored data. The authors perform a process referred to as *inverting* the Kendall's τ statistic to obtain the slope of the Theil-Sen regression. It allows for estimating Kendall's τ

statistic and the ATS line in the context of censored data.

The **NADA** and the **survSpearman** R packages implement, respectively, the papers by Akritas *et al.* (1996) and Eden *et al.* (2022). Moreover, we applied these packages to obtain Table 3.3 results for pairs of coverages based on their position in each claim. Since there are only twenty-eight claims with at least seven coverages, we focus our correlation analysis on claims for which there are at least 30 observations, i.e., claims that have at most six observed coverages. We notice that both indicators are higher than 0, hinting at a possible positive correlation between the variables of each pair. In contrast, no value is lower than 0, indicating that a negative correlation is unlikely. Furthermore, we also perform a statistical test to verify the significativity of the ATS model, which also tests whether Kendall's τ correlation coefficient is significantly different from zero (as stipulated in Helsel (2012)). Most values fall below the 1 % error probability, with only the correlation between the first and sixth coverages having a p-value higher than 5 %. Thus, overall the indicators point towards a positive correlation between coverages. Likewise, we are able to draw similar conclusions from the rank statistics and p-values which are calculated from the version of the data set which orders coverages according to the parties' birth dates rather than the initial reserve values.

We can also interpret these numerical results through Figures 3.3- 3.17 (available in Appendix 3.6). Here, we see a clearly defined line when comparing the ranked (or ordered) values of two coverages from claims that have at most five coverages, i.e., Figures 3.3, 3.4, 3.5, 3.6, 3.8, 3.9, 3.10, 3.12, 3.13 and 3.15. Indeed, with more claims used for these illustrations, we can discern the plot patterns that point towards a positive correlation between coverages. Moreover, recall that the ATS regression significance tests in Table 3.3 also indicate the same results since p-values are lower than 1 %. As for figures for less frequently observed claims, i.e.,

TABLEAU 3.3: Spearman's ρ and Kendall's τ rank statistics

	—By Initial Reserve—			—-By Birth Date—-		
Pair	τ	p-value	ρ	τ	p-value	ρ
(1,2)	0.45	< 0.001	0.62	0.42	< 0.001	0.62
(1,3)	0.38	< 0.001	0.53	0.36	< 0.001	0.54
(1,4)	0.38	< 0.001	0.59	0.36	< 0.001	0.54
(1,5)	0.35	< 0.001	0.5	0.31	< 0.001	0.53
(1,6)	0.38	0.002	0.14	0.23	0.054	0.55
(2,3)	0.47	< 0.001	0.63	0.43	< 0.001	0.64
(2,4)	0.46	< 0.001	0.62	0.42	< 0.001	0.65
(2,5)	0.37	< 0.001	0.53	0.31	< 0.001	0.56
(2,6)	0.36	0.003	0.22	0.26	0.03	0.51
(3,4)	0.59	< 0.001	0.73	0.56	< 0.001	0.76
(3,5)	0.45	< 0.001	0.57	0.41	< 0.001	0.63
(3,6)	0.49	< 0.001	0.14	0.29	0.017	0.74
(4,5)	0.49	< 0.001	0.66	0.44	< 0.001	0.66
(4,6)	0.35	0.005	0.17	0.2	0.094	0.56
(5,6)	0.4	0.001	0.37	0.23	0.056	0.54

Figures 3.7, 3.11, 3.14, 3.16 and 3.17, it is possible to identify a positive correlation between coverages, especially for observations below the 200 day mark. However, results are less conclusive given the fewer observations for longer settlement delays. Again, these illustrations mirror the results of Table 3.3, where a positive correlation can be determined for most of these combinations, albeit with a higher error probability.

Given that we are dealing with clusters larger than two, bivariate correlation measures are not the best-suited tools to asses overall within-cluster correlation. However, they are still valuable items for preliminary data analysis. In the follo-

wing sections, we will define the models we considered, followed by a thorough goodness-of-fit analysis to assess the within-cluster coverage correlation further.

3.3 Statistical Framework

Suppose that a claim affects different coverages. For instance, in the context of an accident benefits claim, an insurer may be liable to pay for the medical bills of multiple people involved in an accident. To include detailed information about each injured person in the modeling process, cash flows issue from each person involved may be treated as different claim coverages. We could also consider a property damage claim that involves multiple cars, each with its characteristics. Again, to incorporate information related to each vehicle, the insurer may consider cash flows related to each car separately as distinct claim coverages. Let $T_{i,j}$ be the delay between the occurrence of a claim i and the settlement of its j^{th} coverage, $j = 1, \ldots, J_i$ and $i = 1, \ldots, I$. Thus, the j^{th} coverage of claim i can be identified by the pair (i, j).

Let $\mathcal{I} = \mathcal{I}^{(C)} \cup \mathcal{I}^{(O)}$ be the set containing pairs (i, j) available in the portfolio at the valuation date, where $\mathcal{I}^{(C)}$ and $\mathcal{I}^{(O)}$ are the subsets containing, respectively, closed and open coverages. Moreover, let $\mathcal{I}^{(U)}$ be the set containing the unreported coverages. Thus, the two sets that require predictions are $\mathcal{I}^{(O)}$, which includes the reported but not settled (RBNS) coverages, and $\mathcal{I}^{(U)}$ which contains the incurred but not reported (IBNR) coverages.

We can consider claim i as a cluster of J_i distinct coverages that share the same origin. In terms of observed values, for a given claim i, all its J_i coverages are framed by τ_i, the delay between the occurrence of claim i and the valuation date. Thus, τ_i denotes the censoring value of all coverages of claim i that are still open at the valuation date. Let $\mathcal{J}_i^{(O)}$ and $\mathcal{J}_i^{(C)}$ be the sets containing, respectively, open

and closed coverages for claim i. As such,

$$\mathcal{J}_i^{(O)} = \{j | t_{i,j} > \tau_i\} \qquad \mathcal{J}_i^{(C)} = \{j | t_{i,j} \leq \tau_i\}$$

for $j = 1, \ldots, J_i$, and

$$\mathcal{I}^{(O)} = \bigcup_{i=1}^{I} \left\{ (i,j) | j \in \mathcal{J}_i^{(O)} \right\} \qquad \mathcal{I}^{(C)} = \bigcup_{i=1}^{I} \left\{ (i,j) | j \in \mathcal{J}_i^{(C)} \right\}.$$

Let us briefly describe the development of coverages in the context of this paper. $T_{i,j}$ can be calculated as the sum of the delay between the occurrence date and the reporting date $T_{i,j}^{(r)}$ and the delay between the occurrence date and the closure date $T_{i,j}^{(c)}$:

$$T_{i,j} = T_{i,j}^{(r)} + T_{i,j}^{(c)}, \text{ for } (i,j) \in \mathcal{I}$$

A visual representation of the various delays is available in Figure 3.1, where three coverages are depicted (closed, open, and unreported), and the first two coverages derive from the same claim.

When a coverage is reported, information about it becomes available and can be summarized as a vector of covariates. These variables come in three different types. First, there are *static covariates* which do not change over time, e.g., the region where the event occurred. Hence, let

$$\mathbf{X}'_{i,j} = [X_{i,j,1}, \ldots, X_{i,j,k}]$$

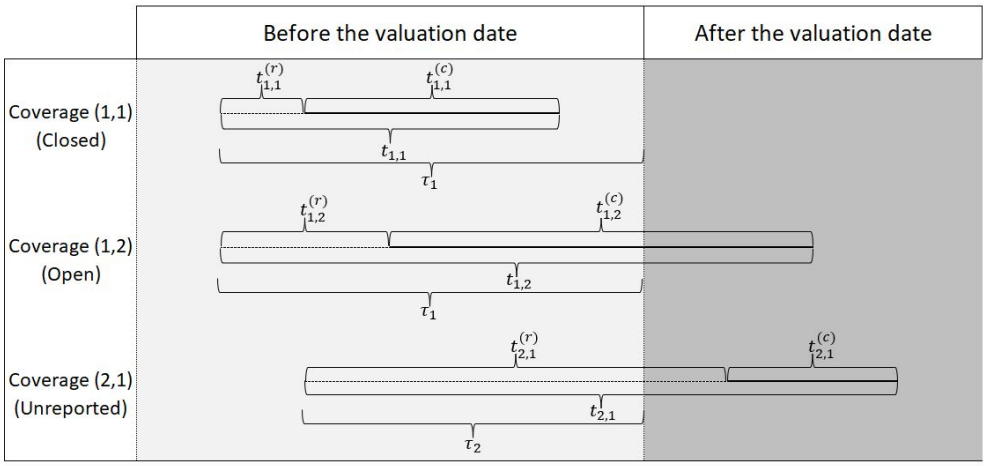

FIGURE 3.1 Development of three coverages

be the vector that contains the k static covariates from the j^{th} coverage of claim i. We can also identify *deterministic time covariates*, which change over time, although in a predictable way, e.g., the age of the person injured in an accident. Let,

$$\mathbf{Z}'_{i,j}(s) = [Z_{i,j,1}(s), \ldots, Z_{i,j,h}(s)]$$

be the vector that contains the h deterministic time covariates from the j^{th} coverage of claim i at time s. In addition, we can also consider the set of vectors that contain the values of these deterministic time covariates from time 0 to s as,

$$\mathbf{Z}^*_{i,j}(s) = \{\mathbf{Z}_{i,j}(u) | 0 < u \leq s\}.$$

Finally, we can consider *stochastic time covariates*, which evolve similarly to their deterministic counterparts. However, this evolution is not deterministic and cannot

be predicted with certainty. Thus, an additional stochastic model is required to predict their unobserved values. For this reason, although handy information can be gathered from these types of covariates, the extra considerations needed to incorporate them are outside this project's scope. They will be the subject of a dedicated project.

Their status puts forward a different challenge regarding the predictions of unreported claims. Their covariate information is unavailable at the valuation date, and how many are currently in the portfolio is unknown. In this paper, we will focus on RBNS claims, given that the main focus of this work regards the dependence of observed open coverages from the same claim. The IBNR claims are better dealt with in future separate work.

Having established the basic notation used in this paper, we can continue defining the models we will consider.

3.3.1 The Cox proportional hazards model

The first model we define is the well-known Cox proportional hazards model, often abbreviated simply to the Cox model (Cox (1972)). It will serve as a baseline for more complex models we will define later. This model is considered semi-parametric because it introduces a non-parametric aspect as an arbitrary baseline hazard function alongside the covariate vectors. The model's hazard rate is defined as :

$$\lambda_{i,j}\left(t|\mathbf{X}_{i,j},\mathbf{Z}^*_{i,j}(t)\right) = \lambda_0(t)\exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t)\boldsymbol{\beta}^{(z)}\right),$$

where $\lambda_0(t)$ is the baseline hazard rate, $\boldsymbol{\beta}^{(x)}$ and $\boldsymbol{\beta}^{(z)}$ are, respectively, the parameter vectors for static covariates and deterministic time covariates for (i,j).

Then, we can obtain the survival function :

$$S_{i,j}\left(t|\mathbf{X}_{i,j},\mathbf{Z}^*_{i,j}(t)\right)=\Pr\left(T_{i,j}>t|\mathbf{X}_{i,j},\mathbf{Z}^*_{i,j}(t)\right)=\exp\left\{-\Lambda_{i,j}(t)\right\},$$

where $\Lambda_{i,j}(t)=\int_0^t\lambda_0(u)\exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)}+\mathbf{Z}'_{i,j}(u)\boldsymbol{\beta}^{(z)}\right)du$.

In the specific case where no deterministic time covariates are observed, they may not be available in the data set. The survival function in this scenario becomes :

$$S_{i,j}\left(t|\mathbf{X}_{i,j}\right)=S_0(t)^{\exp(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)})},$$

where $S_0(t)=\exp\left\{-\int_0^t\lambda_0(u)du\right\}$. We can now address the estimation of $\boldsymbol{\beta}$ and the cumulative baseline hazard function defined by $\Lambda_0(t)=\int_0^t\lambda_0(u)du$. The $\boldsymbol{\beta}$ parameter can be obtained by analyzing the so-called "partial likelihood." Under independent censoring, the partial likelihood is,

$$L\left(\boldsymbol{\beta}^{(x)},\boldsymbol{\beta}^{(z)}\right)=\prod_{i=1}^{I}\prod_{j=1}^{J_i}\left(\frac{\exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)}+\mathbf{Z}'_{i,j}(t_{i,j})\boldsymbol{\beta}^{(z)}\right)}{\sum_{(i,j)\in R(t_{i,j})}\exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)}+\mathbf{Z}'_{i,j}(t_{i,j})\boldsymbol{\beta}^{(z)}\right)}\right)^{\mathbb{1}(t_{i,j}<\tau_i)},\tag{3.1}$$

where $\mathbb{1}()$ is the indicator function. Also, $R(t)$ are the coverages that have not been settled and are still open (uncensored) just before time t. That is the set of coverages that have a duration greater than t :

$$R(t)=\left\{(i,j)|t_{i,j}>t\right\}.$$

Furthermore, in additionally to the estimation of the $\boldsymbol{\beta}$ parameter, it is possible to derive an estimation of the cumulative baseline hazard function through the Breslow estimator (see Breslow (1972)) such that,

$$\widehat{\Lambda}_0(t|\mathbf{X}_{i,j}, \mathbf{Z}^*_{i,j}) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{\mathbb{1}\left(t_{i,j} < \tau_i\right) \mathbb{1}\left(t_{i,j} < t\right)}{\sum_{(i,j)\in R(t_{i,j})} \exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t_{i,j})\boldsymbol{\beta}^{(z)}\right)}, \tag{3.2}$$

where $\Lambda_{i,j}(t) = \int_0^t \lambda_0(u)du$.

3.3.2 The Weibull-Cox proportional hazards model

In parallel to the classical semi-parametric version of the Cox model, a fully parametric alternative was suggested in Byar (1982). In this version, referred to as the Weibull-Cox model, the hazard rate is also proportional but identified through a Weibull distribution. This distinction allows us to easily infer essential measures in a closed form, such as the mean or the variance. Furthermore, the fully parametric nature of the model will enable us to compare it to other parametric models through measures based on the likelihood, such as the Akaike Information Criterion. One may also perform coverage processing time simulation procedures based on the Weibull distribution, which handily provides the distribution of its future unobserved development. These features are very welcome in the context of loss reserving, given that our primary concern is the future development of claims. Let the hazard rate of the model be

$$\begin{aligned}\lambda_{i,j}\left(t|\mathbf{X}_{i,j}, \mathbf{Z}^*_{i,j}(t)\right) &= \nu\alpha t^{\alpha-1}\exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t)\boldsymbol{\beta}^{(z)}\right) \\ &= \lambda_0(t)\exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t)\boldsymbol{\beta}^{(z)}\right),\end{aligned}$$

where, $\lambda_0(t) = \nu\alpha t^{\alpha-1}$ is the baseline hazard rate. Furthermore, $\boldsymbol{\beta}^{(x)}$ and $\boldsymbol{\beta}^{(z)}$ are the vector parameters for static and time deterministic covariates, respectively. Furthermore, by letting T_0 be the baseline processing time, we can identify ν and α as the *scale parameter* and the *shape parameter* of T_0. Thus,

$$\begin{aligned} T_0 &\sim \text{Weibull}\left(\alpha, \nu\right), \\ \left(T_{i,j}|\mathbf{X}_{i,j}, \mathbf{Z}^*_{i,j}(t)\right) &\sim \text{Weibull}\left(\alpha, \nu \cdot \exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t)\boldsymbol{\beta}^{(z)}\right)\right), \end{aligned}$$

for $(i,j) \in \mathcal{I}$. It follows that if only static covariates are considered, the survival and density functions of $T_{i,j}$ become

$$\begin{aligned} S_{i,j}\left(t|\mathbf{X}_{i,j}\right) &= \exp\left\{-\nu t^{\alpha}\exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)}\right)\right\}, \text{ and} \\ f_{i,j}\left(t|\mathbf{X}_{i,j}\right) &= \nu\alpha t^{\alpha-1}\exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)}\right)\exp\left\{-\nu t^{\alpha}\exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)}\right)\right\}. \end{aligned}$$

Moreover, given that we are dealing with a fully parametric model, it is possible to determine the likelihood function directly,

$$L\left(\alpha, \nu, \boldsymbol{\beta}^{(x)}\right) = \prod_i \left(\left(\prod_{j\in\mathcal{J}_i^{(O)}} S_{i,j}\left(\tau_i|\mathbf{X}_{i,j}\right)\right)\left(\prod_{j\in\mathcal{J}_i^{(C)}} f_{i,j}\left(t_{i,j}|\mathbf{X}_{i,j}\right)\right)\right). \tag{3.3}$$

As such, one may obtain the parameter's estimates $\hat{\alpha}$, $\hat{\nu}$ and $\hat{\boldsymbol{\beta}}^{(x)}$ by finding the values that maximize the likelihood.

Having established our baseline models, we can now focus on including intra-claim dependence through random effects.

3.4 Frailty

3.4.1 Frailty for dependent coverages

The Frailty W is a non-negative random effect with a given distribution. The idea is to include a multiplicative element to the hazard rate, indicating that the coverage is more "frail" (or likely to be closed early in loss reserving). It is often assumed that $E[W] = 1$ and that W is a non-negative random variable to facilitate the interpretation of the results. From this concept, we can derive a Shared Frailty model that can be used to include dependence between clusters of coverages through common random effects. Let the hazard function, conditional on the frailty term W_i of a shared Gamma frailty model for the closure delay of the j^{th} individual involved in the i^{th} accident, be

$$\lambda_{i,j}\left(t|\mathbf{X}_{i,j}, \mathbf{Z}^*_{i,j}(t), W_i\right) = \lambda_0(t) W_i \exp(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t)\boldsymbol{\beta}^{(z)}). \tag{3.4}$$

By considering claim i as a cluster of coverages, we can identify that the coverages within it share frailty W_i. We consider that coverages from distinct claims are independent, as are coverages of cluster i conditionally to knowing W_i. Based on these assumptions, it is possible to derive the *conditional* survival function of coverage (i, j),

$$S_{i,j}\left(t|\mathbf{X}_{i,j}, \mathbf{Z}^*_{i,j}(t), W_i\right) = Pr\left(T_{i,j} > t|\mathbf{X}_{i,j}, \mathbf{Z}^*_{i,j}(t), W_i\right) = \exp\left(-W_i \Lambda_{i,j}(t)\right),$$

where $\Lambda_{i,j}(t) = \int_0^t \lambda_0(u)\exp(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t)\boldsymbol{\beta}^{(z)})du$ is the cumulative hazard rate form time 0 to time t.

Now, considering that the Frailty W_i is unobserved, it is possible to focus on the

marginal distribution, based on the mean of the survival function, that is,

$$\begin{aligned} S_{i,j}\left(t|\mathbf{X}_{i,j},\mathbf{Z}^*_{i,j}(t)\right) &= E\left[S_{i,j}\left(t|\mathbf{X}_{i,j},\mathbf{Z}^*_{i,j}(t),W_i\right)\right] \\ &= E\left[\exp\left(-W_i\Lambda_{i,j}(t)\right)\right]. \end{aligned} \tag{3.5}$$

Having defined the hazard rate, the cumulative hazard rate, and the survival function with an intra-claim Frailty random effect, we can now use the observed development of claims to predict the Frailty of each cluster of coverages.

3.4.2 Estimation of the Frailty based on claim history

We can summarize this information by defining the historical data of claim i from its occurrence up to the valuation date as $\mathcal{H}_i(\tau_i)$. Conditionally to this information, the posterior Frailty estimated value for claim i can be obtained as follows

$$E\left[W_i|\mathcal{H}_i(\tau_i)\right] = -\mathcal{L}'_{W_i|\mathcal{H}_i(\tau_i)}(0),$$

where $\mathcal{L}_{W_i|\mathcal{H}_i(\tau_i)}$ is the Laplace transform of frailty W_i, conditionally to $\mathcal{H}_i(\tau_i)$. Thus, since W_i is a non-negative random variable, we can write

$$\begin{aligned} \mathcal{L}_{W_i|\mathcal{H}_i(\tau_i)}(c) &= E\left[\exp(-cW_i)|\mathcal{H}_i(\tau_i)\right] \\ &= \int_0^\infty \exp(-cw) f_{W_i|\mathcal{H}_i(\tau_i)}(w)dw, \end{aligned}$$

where $f_{W_i|\mathcal{H}_i(\tau_i)}(w)$ is the probability density function of the frailty conditionally to $\mathcal{H}_i(\tau_i)$. Recalling that, conditionally to Frailty W_i, all the coverages' processing times $T_{i,j}$ are independent, we can use Bayes' theorem to obtain,

$$
\begin{aligned}
f_{W_i|\mathcal{H}_i}(w) &= \frac{P(\mathcal{H}_i(\tau_i)|W_i) f_{W_i}(w)}{\int_0^\infty P(\mathcal{H}_i(\tau_i)|W_i) f_{W_i}(v) dv} \\
&= \frac{f_{W_i}(w) \left(\prod_{j \in \mathcal{J}_i^{(O)}} S_{(i,j)|W_i=w}(\tau_i) \right) \left(\prod_{j \in \mathcal{J}_i^{(C)}} f_{(i,j)|W_i=w}(t_{i,j}) \right)}{\int_0^\infty f_{W_i}(v) \left(\prod_{j \in \mathcal{J}_i^{(O)}} S_{(i,j)|W_i=v}(\tau_i) \right) \left(\prod_{j \in \mathcal{J}_i^{(C)}} f_{(i,j)|W_i=v}(t_{i,j}) \right) dv} \\
&= \frac{f_{W_i}(w) \left(\prod_{j \in \mathcal{J}_i^{(O)}} e^{-w\Lambda_{i,j}(\tau_i)} \right) \left(\prod_{j \in \mathcal{J}_i^{(C)}} w\lambda_{i,j}(t_{i,j}) e^{-w\Lambda_{i,j}(t_{i,j})} \right)}{\int_0^\infty f_{W_i}(v) \left(\prod_{j \in \mathcal{J}_i^{(O)}} e^{-v\Lambda_{i,j}(\tau_i)} \right) \left(\prod_{j \in \mathcal{J}_i^{(C)}} v\lambda_{i,j}(t_{i,j}) e^{-v\Lambda_{i,j}(t_{i,j})} \right) dv} \\
&= \frac{f_{W_i}(w) \exp\left(-w \left(\sum_{j \in \mathcal{J}_i^{(O)}} \Lambda_{i,j}(\tau_i) + \sum_{j \in \mathcal{J}_i^{(C)}} \Lambda_{i,j}(t_{i,j}) \right) \right) w^{N_i^{(C)}}}{\int_0^\infty f_{W_i}(v) \exp\left(-v \left(\sum_{j \in \mathcal{J}_i^{(O)}} \Lambda_{i,j}(\tau_i) + \sum_{j \in \mathcal{J}_i^{(C)}} \Lambda_{i,j}(t_{i,j}) \right) \right) v^{N_i^{(C)}} dv} \\
&= \frac{f_{W_i}(w) \exp\left(-w \left(\sum_{j \in \mathcal{J}_i^{(O)}} \Lambda_{i,j}(\tau_i) + \sum_{j \in \mathcal{J}_i^{(C)}} \Lambda_{i,j}(t_{i,j}) \right) \right) w^{N_i^{(C)}}}{\int_0^\infty f_{W_i}(v) \exp\left(-v \left(\sum_{j \in \mathcal{J}_i^{(O)}} \Lambda_{i,j}(\tau_i) + \sum_{j \in \mathcal{J}_i^{(C)}} \Lambda_{i,j}(t_{i,j}) \right) \right) v^{N_i^{(C)}} dv} \\
&= \frac{f_{W_i}(w) \exp\left(-w\Lambda_{i,\bullet} \right) w^{N_i^{(C)}}}{\int_0^\infty f_{W_i}(v) \exp\left(-v\Lambda_{i,\bullet} \right) v^{N_i^{(C)}} dv} \\
&= \frac{f_{W_i}(w) \exp\left(-w\Lambda_{i,\bullet} \right) w^{N_i^{(C)}}}{E\left[W_i^{N_i^{(C)}} \exp\left(-\Lambda_{i,\bullet} W_i \right) \right]} \\
&= \frac{f_{W_i}(w) \exp\left(-w\Lambda_{i,\bullet} \right) w^{N_i^{(C)}}}{\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(\Lambda_{i,\bullet} \right)},
\end{aligned}
$$

where $\Lambda_{i,\bullet} = \sum_{j \in \mathcal{J}_i^{(O)}} \Lambda_{i,j}(\tau_i) + \sum_{j \in \mathcal{J}_i^{(C)}} \Lambda_{i,j}(t_{i,j})$, and the number of closed coverages from claim i at the valuation date is written as $N_i^{(C)} = \sum_{j \in \mathcal{J}_i^{(C)}} 1$. It follows that,

$$\begin{aligned}
\mathcal{L}_{W_i|\mathcal{H}_i(\tau_i)}(c) &= E[\exp(-cW_i)|\mathcal{H}_i(\tau_i)] \\
&= \int_0^\infty \exp(-cw) f_{W_i|\mathcal{H}_i(\tau_i)}(w) dw \\
&= \int_0^\infty \exp(-cw) \frac{f_{W_i}(w) \exp\left(-w\Lambda_{i,\bullet}\right) w^{N_i^{(C)}}}{\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(\Lambda_{i,\bullet}\right)} dw \\
&= \frac{\int_0^\infty f_{W_i}(w) \exp\left(-w\left(c+\Lambda_{i,\bullet}\right)\right) w^{N_i^{(C)}} dw}{\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(\Lambda_{i,\bullet}\right)} \\
&= \frac{E\left[W_i^{N_i^{(C)}} \exp\left(-\left(c+\Lambda_{i,\bullet}\right) W_i\right)\right]}{\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(\Lambda_{i,\bullet}\right)} \\
&= \frac{\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(c+\Lambda_{i,\bullet}\right)}{\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(\Lambda_{i,\bullet}\right)}.
\end{aligned}$$

Then, we can write

$$\begin{aligned}
E\left[W_i|\mathcal{H}_i(\tau_i)\right] &= -\mathcal{L}'_{W_i|\mathcal{H}_i(\tau_i)}(0) \\
&= -\frac{\frac{\partial}{\partial c}\left(\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(c+\Lambda_{i,\bullet}\right)\Big|_0\right.}{\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(\Lambda_{i,\bullet}\right)} \\
&= \frac{\mathcal{L}_{W_i}^{(N_i^{(C)}+1)}\left(\Lambda_{i,\bullet}\right)}{\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(\Lambda_{i,\bullet}\right)}, \qquad (3.6)
\end{aligned}$$

where $\mathcal{L}_{W_i}^{(k)}$ denotes the k^{th} derivative of the Laplace transform of the frailty of claim i.

In a more practical application, some distributions are more accessible regarding

tractability. In particular, the Gamma distribution has been extensively studied as a frailty distribution. By considering,

$$W_i \sim \text{Gamma}\left(\frac{1}{\theta}, \frac{1}{\theta}\right), \quad E[W_i] = 1 \quad \text{and} \quad Var[W_i] = \theta,$$

we can obtain the expected value of W_i given its history by using its Laplace transform

$$\mathcal{L}_{W_i}(c) = \left(\frac{1/\theta}{(1/\theta) + c}\right)^{1/\theta}. \tag{3.7}$$

Thus, by using the result in Equation 3.6 with this specific Laplace transform we have,

$$\begin{aligned}
E\left[W_i|\mathcal{H}_i(\tau_i)\right] &= \frac{\mathcal{L}_{W_i}^{(N_i^{(C)}+1)}\left(\Lambda_{i,\bullet}\right)}{\mathcal{L}_{W_i}^{(N_i^{(C)})}\left(\Lambda_{i,\bullet}\right)} \\
&= \frac{\frac{\partial^{N_i^{(C)}+1}}{\partial c^{N_i^{(C)}+1}}\left(\left(\frac{1/\theta}{(1/\theta)+c}\right)^{1/\theta}\Big|_{\Lambda_{i,\bullet}}\right.}{\frac{\partial^{N_i^{(C)}}}{\partial c^{N_i^{(C)}}}\left(\left(\frac{1/\theta}{(1/\theta)+c}\right)^{1/\theta}\Big|_{\Lambda_{i,\bullet}}\right.} \\
&= \frac{\frac{\partial^{N_i^{(C)}+1}}{\partial c^{N_i^{(C)}+1}}\left(\left(\frac{1}{\theta}+c\right)^{-1/\theta}\Big|_{\Lambda_{i,\bullet}}\right.}{\frac{\partial^{N_i^{(C)}}}{\partial c^{N_i^{(C)}}}\left(\left(\frac{1}{\theta}+c\right)^{-1/\theta}\Big|_{\Lambda_{i,\bullet}}\right.} \\
&= \frac{\left(\frac{1}{\theta}+\Lambda_{i,\bullet}\right)^{-1/\theta-N_i^{(C)}-1}\prod_{k=1}^{N_i^{(C)}}\left(-1/\theta-k\right)}{\left(\frac{1}{\theta}+\Lambda_{i,\bullet}\right)^{-1/\theta-N_i^{(C)}}\prod_{k=1}^{N_i^{(C)}-1}\left(-1/\theta-k\right)} \\
&= \frac{\frac{1}{\theta}+N_i^{(C)}}{\frac{1}{\theta}+\Lambda_{i,\bullet}}.
\end{aligned} \tag{3.8}$$

Note that when a Gamma Frailty is considered, the density function $f_{W_i|\mathcal{H}_i(\tau_i)}(w)$ can be obtained in a closed form :

$$\begin{aligned} f_{W_i|\mathcal{H}_i(\tau_i)}(w) &\propto f_{W_i}(w)\exp\left(-w\Lambda_{i,\bullet}\right) w^{N_i^{(C)}} \\ &\propto \frac{(1/\theta)^{(1/\theta)}}{\Gamma(1/\theta)} w_i^{(1/\theta)-1}\exp\left(-w_i/\theta\right)\exp\left(-w\Lambda_{i,\bullet}\right) w^{N_i^{(C)}} \\ &\propto w_i^{(1/\theta)+N_i^{(C)}-1}\exp\left(-w_i(1/\theta+\Lambda_{i,\bullet})\right). \end{aligned}$$

Thus, we can determine that, by letting $W_i \sim \text{Gamma}(1/\theta, 1/\theta)$, for $(i,j) \in \mathcal{I}$, the *a posteriori* distribution of each frailty given the observations its cluster also follows a Gamma distribution, that is

$$(W_i|\mathcal{H}_i(\tau_i)) \sim \text{Gamma}\left(\theta^{-1} + N_i^{(C)}, \theta^{-1} + \Lambda_{i,\bullet}\right), \tag{3.9}$$

for $(i,j) \in \mathcal{I}$. Note that we can obtain the same result from Equation 3.8 by computing the expected value of $(W_i|\mathcal{H}_i(\tau_i))$ based on its distribution. Finally, we can find an estimation of the marginal survival distribution by considering $c = \Lambda_{i,j}(t)$ in Equation 3.7, such that

$$\begin{aligned} S_{i,j}\left(t|\mathbf{X}_{i,j}, \mathbf{Z}^*_{i,j}(t)\right) &= E\left[S_{i,j}\left(t|\mathbf{X}_{i,j}, \mathbf{Z}^*_{i,j}(t), W_i\right)\right] \\ &= E\left[\exp\left(-W_i\Lambda_{i,j}(t)\right)\right] \\ &= \mathcal{L}_{W_i}(\Lambda_{i,j}(t)) \\ &= \left(\frac{1/\theta}{(1/\theta)+\Lambda_{i,j}(t))}\right)^{1/\theta}. \\ &= \frac{1}{(1+\theta\Lambda_{i,j}(t))^{1/\theta}}. \end{aligned} \tag{3.10}$$

3.4.3 Parameter estimation

So far, we have described a frailty model that is composed of three elements that require estimations :

- the baseline hazard rate, $\lambda_0(t)$,
- the covariate vectors, $\boldsymbol{\beta}^{(x)}$ and $\boldsymbol{\beta}^{(z)}$, and
- the set of parameters of the frailty distribution, Θ. Typically, Θ contains only one parameter θ and it is often set such that : $E[W_i] = 1$ and $Var[W_i] = \theta_i$.

The first element to consider is the $\lambda_0(t)$ distribution. In Section 3.3.1, we introduced the Cox model, which does not specify a particular distribution for its hazard rate, and for this reason, it is referred to as *semi-parametric*. However, it is also possible to consider parametric distributions such as a Weibull model or a spline-based estimator instead.

Assuming that the values of each claim's frailty are known, it is possible to determine the conditional likelihood function,

$$L\left(\lambda_0, \boldsymbol{\beta}^{(x)}, \boldsymbol{\beta}^{(z)} | \mathbf{W}\right) = \prod_i \left(\left(\prod_{j \in \mathcal{J}_i^{(O)}} \exp\left(-W_i \Lambda_{i,j}(\tau_i)\right) \right) \times \left(\prod_{j \in \mathcal{J}_i^{(C)}} W_i \lambda_{i,j}(t_{i,j}) \exp\left(-W_i \Lambda_{i,j}(t_{i,j})\right) \right) \right), \quad (3.11)$$

where $\mathbf{W} = [W_1, \ldots, W_I]$ the vector containing the frailties of each cluster within the claim portfolio of size I. In the fully parametric case, $\lambda_0(t)$ requires a limited number of parameters to be considered. Because of this, assuming values of vector $\mathbf{W}$ are known, a maximum likelihood approach could be considered. However,

when considering a *semi-parametric* Cox model, the baseline function can be estimated through the *partial likelihood* described in Section 3.3.1. As such, one can find an estimate of $\boldsymbol{\beta}^{(x)}$ and $\boldsymbol{\beta}^{(z)}$ by maximizing their values in,

$$\ell\left(\boldsymbol{\beta}^{(x)}, \boldsymbol{\beta}^{(z)} | \mathbf{W}\right) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \mathbb{1}\left(t_{i,j} < \tau_i\right) \times \left[-\log\left(\sum_{(i,j)\in R(t_{i,j})} W_i \exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t_{i,j})\boldsymbol{\beta}^{(z)}\right)\right) + \mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t_{i,j})\boldsymbol{\beta}^{(z)}\right] \tag{3.12}$$

along with the estimated value of the baseline hazard rate through the Breslow estimator,

$$\widehat{\Lambda}_0\left(t | \boldsymbol{\beta}^{(x)}, \boldsymbol{\beta}^{(z)}, \mathbf{W}\right) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{\mathbb{1}\left(t_{i,j} < \tau_i\right) \mathbb{1}\left(t_{i,j} < t\right)}{\sum_{(i,j)\in R(t_{i,j})} W_i \exp\left(\mathbf{X}'_{i,j}\boldsymbol{\beta}^{(x)} + \mathbf{Z}'_{i,j}(t_{i,j})\boldsymbol{\beta}^{(z)}\right)}, \tag{3.13}$$

where, $\widehat{\Lambda}_0(t) = \int_0^t \widehat{\lambda}_0(u)du$.

The previously described estimators are incomplete because vector $\mathbf{W}$ is unknown and is considered a random variable with θ as the variance parameter of its distribution. Thus, more complex approaches are required for the estimation to be made. Luckily several authors have made various suggestions to deal with this problem, along with several packages that allow for their practical applications.

The Expectation-Maximization (EM) algorithm was suggested by Nielsen *et al.* (1992) and Klein (1992) in the context of semi-parametric Gamma frailty models. The idea alternates between the "E" and the "M" steps until convergence. The

steps are :

Step 1 : Set to 1 the value of $\mathbf{W_i}$ for all frailties (i.e. $\widehat{\theta} = 1$). Then, assuming $\mathbf{W_i} = 1, \forall i = 1, \ldots, I$, the estimations of the baseline hazard rate $\widehat{\Lambda}_0(t)$ and covariate parameters as $\widehat{\boldsymbol{\beta}}^{(x)}$ and $\widehat{\boldsymbol{\beta}}^{(z)}$ are done by either finding the values that maximize the log-likelihood (see Equation 3.11), for the parametric case, or by finding the values that maximize Equation 3.12 and 3.13 in the semi-parametric case.

Step 2 : Compute $E\left[W_i|\mathcal{H}_i(\tau_i)\right]$ using the values in step 1. This value can be obtained by calculating derivatives of the Laplace transform of the frailty distribution (see 3.6). Then calculate $E\left[\log\left(L\left(\lambda_0, \boldsymbol{\beta}^{(x)}, \boldsymbol{\beta}^{(z)}|\mathbf{W}\right)\right)\right]$.

Step 3 : Find the values of $\widehat{\Lambda}_0(t)$, $\widehat{\boldsymbol{\beta}}^{(x)}$, $\widehat{\boldsymbol{\beta}}^{(z)}$ and $\widehat{\theta}$ that maximize $E\left[W_i|\mathcal{H}_i(\tau_i)\right]$.

Step 4 : Repeat steps 2 and 3 until convergence.

This algorithm has been implemented in the **frailtyEM** package from R (see Balan *et al.* (2019)). A similar application of the EM algorithm by Vaida & Xu (2000) is available in the R package **phmm**, which uses a Monte-Carlo EM algorithm. In contrast to the EM algorithm, penalized-likelihood methods have been suggested. For instance, in Therneau *et al.* (2003), a penalized likelihood function was developed for the semi-parametric case, such that,

$$\gamma\left(\boldsymbol{\beta}^{(x)}, \boldsymbol{\beta}^{(z)}, \theta|\mathbf{W}\right) = \ell\left(\boldsymbol{\beta}^{(x)}, \boldsymbol{\beta}^{(z)}|\mathbf{W}\right) - g(\mathbf{W}, \theta),$$

where $\ell\left(\boldsymbol{\beta}^{(x)}, \boldsymbol{\beta}^{(z)}|\mathbf{W}\right)$ is the cox partial likelihood and $g(\mathbf{W}, \theta)$ is a penalty function that restricts the values of θ. In particular, the authors suggested

$$g(\mathbf{W}, \theta) = -1/\theta \sum_i (W_i - \exp(W_i))$$

as the penalty function when a Gamma distribution is considered. Yet again, computational tools have been developed with penalized methods such as the one suggested by Therneau *et al.* (2003). The package **survival** in R provides an algorithm using a penalized likelihood for semi-parametric Gamma and log-normal frailty distributions. Moreover, for penalized likelihood methods, we recommend the **frailtypack** R package, which provides an even broader and more flexible algorithm for parametric and semi-parametric frailty models.

On a final note, by letting $W_i \sim \text{Gamma}(1/\theta, 1/\theta)$, for $(i, j) \in \mathcal{I}$, and by considering the Weibull-Cox model, we can obtain a fully parametric *marginal* likelihood function,

$$
\begin{aligned}
L\left(\lambda_0, \boldsymbol{\beta}^{(x)}, \boldsymbol{\beta}^{(z)}, \theta\right) &= E\left[L\left(\lambda_0, \boldsymbol{\beta}^{(x)}, \boldsymbol{\beta}^{(z)}, \theta | \mathbf{W}\right)\right] \\
&= \prod_i \left(E\left[W_i^{N_i^{(C)}} \exp\left(-W_i \Lambda_{i,\bullet}\right)\right] \prod_{j \in \mathcal{J}_i^{(C)}} \lambda_{i,j}(t_{i,j}) \right) \\
&= \prod_i \left(\int_0^\infty w_i^{N_i^{(C)}} \exp\left(-w_i \Lambda_{i,\bullet}\right) f_{W_i}(w_i) dw_i \prod_{j \in \mathcal{J}_i^{(C)}} \lambda_{i,j}(t_{i,j}) \right) \\
&= \prod_i \left(\int_0^\infty w_i^{N_i^{(C)}} \exp\left(-w_i \Lambda_{i,\bullet}\right) w_i^{\theta^{-1}-1} \exp\left(-w_i/\theta\right) dw_i \times \right. \\
&\qquad \left. \frac{\theta^{-\theta^{-1}}}{\Gamma\left(\theta^{-1}\right)} \prod_{j \in \mathcal{J}_i^{(C)}} \lambda_{i,j}(t_{i,j}) \right) \\
&= \prod_i \left(\frac{\Gamma\left(\theta^{-1} + N_i^{(C)}\right) \theta^{-\theta^{-1}}}{\Gamma\left(\theta^{-1}\right) \left(\theta^{-1} + \Lambda_{i,\bullet}\right)^{\theta^{-1} + N_i^{(C)}}} \cdot 1 \cdot \prod_{j \in \mathcal{J}_i^{(C)}} \lambda_{i,j}(t_{i,j}) \right),
\end{aligned}
$$

where,

$$
\begin{aligned}
\Lambda_{i,\bullet} &= \sum_{j \in \mathcal{J}_i^{(O)}} \nu \tau_i^\alpha \exp\left(\mathbf{X}'_{i,j} \boldsymbol{\beta}^{(x)}\right) + \sum_{j \in \mathcal{J}_i^{(C)}} \nu t_{i,j}^\alpha \exp\left(\mathbf{X}'_{i,j} \boldsymbol{\beta}^{(x)}\right), \\
\lambda_{i,j}(t_{i,j}) &= \nu \alpha t_{i,j}^{\alpha-1} \exp\left(\mathbf{X}'_{i,j} \boldsymbol{\beta}^{(x)}\right).
\end{aligned}
$$

We will now fit frailty models in the context of micro-level loss reserving using a data set from a Canadian insurance company hoping to consider the intra-claim dependence of the portfolio.

3.5 Numerical results

3.5.1 Fitting the Models

We consider various approaches for our numerical analysis based only on static covariate vectors $\mathbf{X}'_{i,j}$ (note that predictable time covariates are transformed into categorical variables in our numerical analysis). We begin by fitting simple models that do not consider dependence. The first one is the Cox model, which was described in Section 3.3.1, where we obtain estimated values for covariate parameters through the partial likelihood function 3.1, while the Brelow estimator 3.2 is used to find the cumulative hazard rate.

In addition to the classical Cox model, we consider other methods to determine the baseline hazard rate, such as cubic M-splines, which are, according to Rondeau *et al.* (2012), particularly useful because they allow for flexible shapes while reducing the number of parameters. Furthermore, we perform the estimation of the parameters through the maximization of the penalized log-likelihood suggested by Rondeau *et al.* (2003) and available through the R package **frailtypack**. In addition, we also consider the Weibull-Cox model highlighted in Section 3.3.2. We estimate parameters by maximizing the likelihood function 3.3. Then, to further evaluate the performance of our fully parametric propositions, we also fit two other well-known models in the literature, the Log-logistic and the Log-Normal. Let,

$$
\begin{aligned}
\left(T_{i,j}|\mathbf{X}'_{i,j}\right) &\sim \text{Weibull}\left(\alpha^{(Wei)}, \nu_{i,j}^{(Wei)}\left(\boldsymbol{\beta}^{(x)}\right)\right), \\
\left(T_{i,j}|\mathbf{X}'_{i,j}\right) &\sim \text{Log-Logistic}\left(\alpha^{(Loglog)}, \nu_{i,j}^{(Loglog)}\left(\boldsymbol{\beta}^{(x)}\right)\right), \text{ and} \\
\left(T_{i,j}|\mathbf{X}'_{i,j}\right) &\sim \text{Log-Normal}\left(\mu_{i,j}\left(\boldsymbol{\beta}^{(x)}\right), \sigma^2\right),
\end{aligned}
$$

for $(i,j) \in \mathcal{I}$. Where $\alpha^{(Wei)}$ and $\alpha^{(Loglog)}$ are the *shape parameters* of, respectively,

the Weibull and the Log-logistic distributions, and, similarly, $\nu^{(Wei)}$ and $\nu^{(Loglog)}$ are the *scale parameters* of these distributions. Finally, for the Log-Normal distribution, $\mu_{i,j}$ and σ^2 represent the mean and scale parameters of $\log\left(T_{i,j}|\mathbf{X}'_{i,j}\right)$ which follows a Normal distribution.

Finally, we can include dependence through Gamma Frailty random effects for the coverages of the same claim. We apply them to Cox models : the classical semi-parametric version, which considers Cubic M-splines, and the Weibull-Cox model. For the last two mentioned models, the same package (**frailtypack**) is implemented, and the penalized log-likelihood method is used for their counterparts (see Rondeau *et al.* (2003)). In contrast, the classical version is fitted with the **survival** R package.

3.5.2 Goodness-of-fit

This section describes the goodness of fit analysis for models depicted in Section 3.5.1. Let us begin by examining the significance of *static covariates* used across all models. We present estimated values for all parameters in Table 3.9 and t-tests in Table 3.6. The latter table shows that the covariates are statistically significant overall. Indeed, for the type of loss, the gender of the injured, and the region, all categories are significant with an error probability lower than 1%. The $(1,000 - 5,000]$ category of the initial reserve and the $(18 - 25]$ category of injured/vehicle age are non-significant for some distributions when considering a 5% error. It means that these intervals could be grouped with the baseline category. Nonetheless, since some distributions benefit from this division, we decide to keep the divisions as they are for all models for comparison purposes. The only other notable elements in terms of significance are the $(30-90]$ interval for the reporting delay in the Log-logistic and the log-normal model and the $(6 - 10]$ interval for

the vehicle age (in years) in the Cox model and its M-spline version.

Furthermore, given that we consider Frailty effects following a Gamma distribution, these models require an estimation for θ, which is the frailty distribution's variance. Moreover, it is possible to perform a unilateral Wald test to verify whether θ is significantly different from 0. Results are presented in Table 3.4, where we can confidently reject the null hypothesis.

TABLEAU 3.4: Wald tests for the variance of the Frailty (random effect)

	$\widehat{\theta}$	Std. err.	z	p-value
Base	0.76	-	-	-
M-Spline	0.65	0.01	46.50	< 0.01
Weibull	0.61	0.01	46.50	< 0.01

Then we perform a comparison between all estimated models. Recall that the Weibull-Cox (with and without Frailty), the Log-Logistic, and the Lognormal models are fully parametric, and this particular feature allows us to compute the log-likelihood. Thus, we can consider the Schwarz Information Criterion (BIC) and the Akaike information criterion (AIC) as comparison measures. These values are available in Table 3.5, where we can identify the Weibull-Cox model with Frailty as the best model for this criteria. For the comparison of semi-parametric M-spline models, we choose an approximation of the Likelihood Cross-Validation (LCV) criterion suggested by O'Sullivan (1988) and later incorporated by Rondeau *et al.* (2012). The results are available in Table 3.6 where the model with a Frailty random effect is picked over its counterpart. Finally, we perform likelihood ratio tests between the Weibull-Cox model (using the log-likelihood) and the classical Cox model (with the partial-log-likelihood). More specifically, a comparison in terms of likelihood is made between the versions that include a Frailty effect and the ones that do not. For both cases, we have a clear indication that the model

with Frailty is preferable (see Table 3.7).

TABLEAU 3.5: AIC and BIC criteria for the parametric models

	Weibull	Log-Logistic	Log-Normal	Weibull (Frailty)
AIC	$552,167$	$551,355$	$553,450$	$\mathbf{547,407}$
BIC	$552,442$	$551,630$	$553,724$	$\mathbf{547,691}$

TABLEAU 3.6: Likelihood Cross-Validation (LCV) criterion for the M-spline models

	M-splines	M-splines (Frailty)
LCV	$524,069$	$\mathbf{527,597}$

TABLEAU 3.7: Likelihood Ratio tests for Cox and Weibull-Cox models

	Likelihood		Test Statistic	p-value
	Without Frailty	With Frailty		
Base Cox	$-438,554$	$-436,658$	$4,140$	< 0.01
Weibull-Cox	$-276,053$	$-273,672$	$4,762$	< 0.01

3.5.3 Simulation analysis

We begin our numerical simulation analysis by gathering the coverages from cluster size two or greater claims. As such, let $\mathcal{I}^{*(C)}$ and $\mathcal{I}^{*(O)}$ be, respectively, the sets of the closed and open claims in question, defined as

$$\mathcal{I}^{*(C)} = \left\{(i,j)|J_i \geq 2, (i,j) \in \mathcal{I}^{(O)}\right\} \text{ and } \mathcal{I}^{*(O)} = \left\{(i,j)|J_i \geq 2, (i,j) \in \mathcal{I}^{(O)}\right\}.$$

We analyzed this particular set of observations because we want to focus on coverages for which the issue of dependence is most relevant. Then, we proceed to

determine the distribution of the total outstanding exposure. Let $E^{*(O)}$ be the sum of delays between the valuation and the settlement dates for $(i,j) \in \mathcal{I}^{*(O)}$, that is,

$$E^{*(O)} = \sum_{(i,j)\in\mathcal{I}^{*(O)}} T_{i,j} - \tau_i. \tag{3.14}$$

We compare four parametric distributions : the Log-logistic, Lognormal, and Weibull (with and without Frailty) models. We obtain these distributions by performing $10,000$ simulations per model of all operational times of the coverages and then computing the total exposure. In particular, for the Weibull-Cox model with Frailty, we first simulate the Frailty of each cluster using a Gamma distribution (see 3.9). Next, for each claim, each coverage's processing time is simulated conditionally to the value of the Frailty that was simulated in the current iteration. We present the main results in Table 3.8 and Figure 3.2.

TABLEAU 3.8: Results of the predictions for the total outstanding exposure $\left(E^{*(O)}\right)$

	Mean	Std. err.	75% VaR	95% VaR	99% VaR	Obs.
Weibull	1532	24.80	1549	1568	1576	1678
Frailty Weibull	1679	27.27	1697	1725	1744	1678
Loglogistic	1814	34.62	1839	1873	1887	1678
Lognormal	1855	29.87	1875	1902	1916	1678

We notice that the log-logistic and Lognormal models overestimate the exposure value. Indeed, when we look at Figure 3.2, the two distributions in question are represented on the right side of the Figure. Here, the densities do not include the observed value (the black dotted vertical line) within their range of possible values. Then, in contrast, the Weibull model seems to underestimate the exposure.

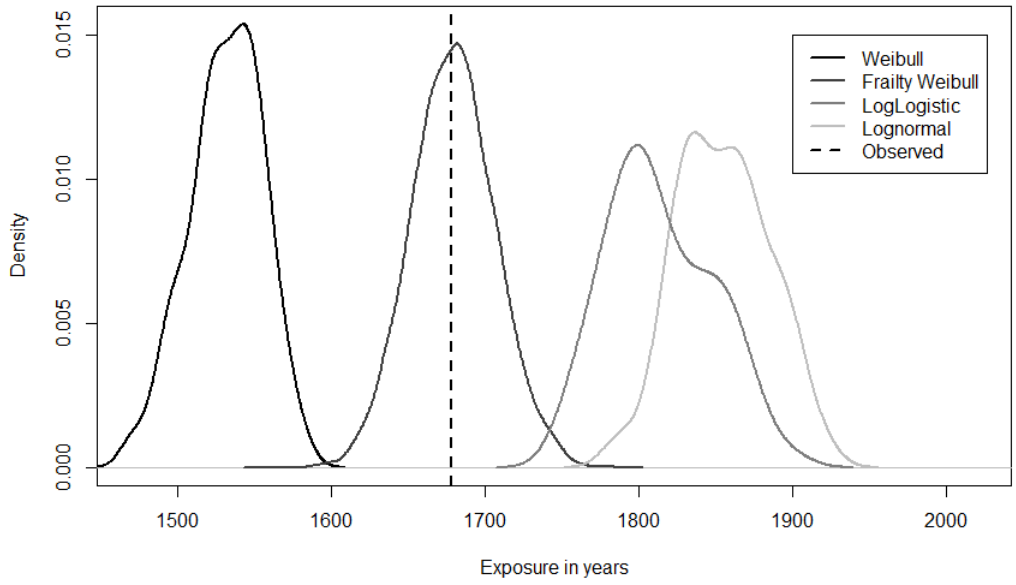

FIGURE 3.2 Distributions of the outstanding exposure.

We draw this conclusion by looking at its 99 % Value-at-Risk (VaR), which is lower than the observed exposure. Again, the density does not seem to include the observed value. It can also be observed graphically in Figure 3.2, where the dotted vertical line is clearly outside the distribution range. Finally, the Weibull model with Frailty provides the best results as its mean is close to the observed value while being lower than the 99 % and 95 % VaRs. Again, the performance of this fourth model can be seen graphically in Figure 3.2. Here, unlike the other models, the density seems to include the observed value within the range of its possible values.

3.5.4 Conclusion

In this paper, a survival analysis was performed for the processing time of claim coverages. Our work was motivated by the relevance of coverage-specific covariate information, which cannot be included in the modeling process if the processing time of the claim from which the coverages originate is directly modeled. This choice led us to consider Frailty random effects to consider that coverages of a given claim are likely to be dependent, as they share the same origin. As a baseline, we provided core definitions of classical proportional hazard models adapted to a micro-level reserve setting, such as the Weibull-Cox and Cox models. We chose these models as the inclusion of Frailty effects has been extensively studied in the survival literature, particularly when considering a Gamma distribution. We referenced and adapted relevant publications to showcase important theoretical definitions and results. In addition, we covered various methods and their numerical applications to fit the Frailty models we considered.

To complete our analysis, we performed a numerical application of the models we presented. An essential element we covered were numerical justifications for the modeling of coverages and the existence of within-cluster dependence. It was tackled on two fronts. First, we provided several tests and statistics that indicated the coverages of the same cluster are dependent, be it through non-parametric methods such as correlation rank statistics (Spearman's ρ and Kendall's τ) or through confirming the significance of the Frailty components with tests such as the likelihood ratio test. Second, we showed that coverage-specific information, such as the age of the injured person, was significant in terms of likelihood, thus motivating the division of claims into coverages. Our numerical analysis also focused on the statistical advantage of including a Frailty random effect in the benchmark models. We confirmed that models that included Frailty provided bet-

ter results regarding the goodness of fit than their counterparts. These results were provided by the AIC and BIC criteria in the parametric case and the LCV in the semi-parametric case (among other tests).

Our findings suggest that a survival perspective in micro-level loss reserving provides an efficient and well-supported solution to portfolio coverage dependence. Admittedly, the main focus of this paper was the processing time of coverages. Despite its great importance, it is not the only variable required to predict a loss reserve in a micro-level setting. In effect, a model for the number of payments and their severity are also required in conjunction with the predictions of the settlement dates. As such, a natural extension to this model would be to incorporate Frailty effects into these variables, for instance, through Cox processes for claim payment modeling. It is worth noting that such models have been extensively studied in micro-level reserving literature for unreported claims (Zhao *et al.* (2009), Zhao *et al.* (2010), Badescu *et al.* (2016)), confirming its usefulness in this context. However, coverage-based predictions have yet to be considerably researched, and thus extensions in terms of severity and frequency could be considered for both the IBNR and the RBNS coverages.

3.6 Appendix

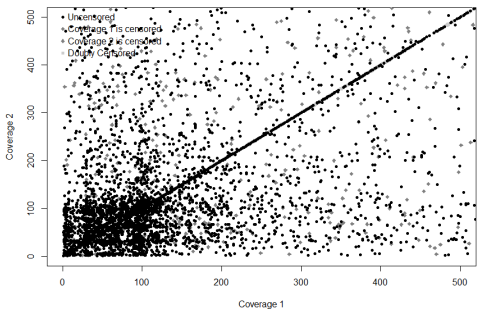

FIGURE 3.3 1^{st} to 2^{nd} coverages

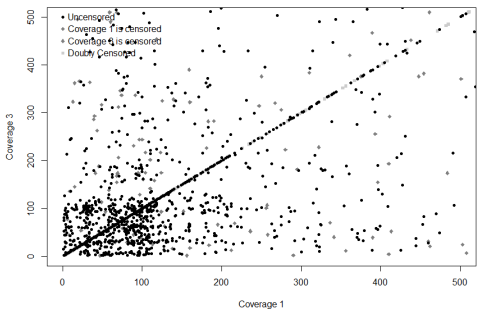

FIGURE 3.4 1^{st} to 3^{rd} coverages

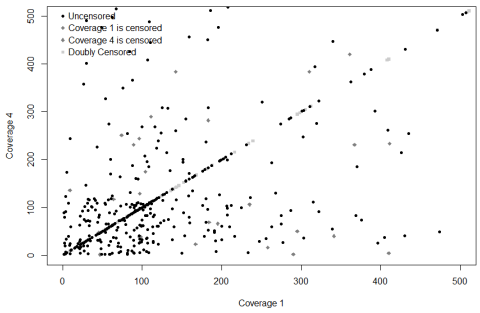

FIGURE 3.5 1^{st} to 4^{th} coverages

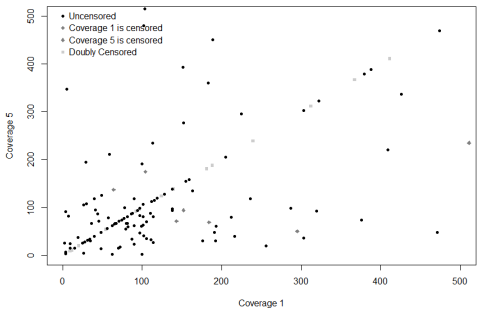

FIGURE 3.6 1^{st} to 5^{th} coverages

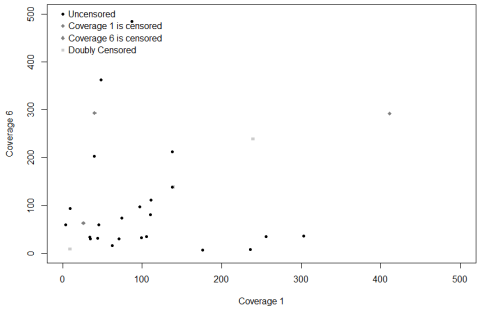

FIGURE 3.7 1^{st} to 6^{th} coverages

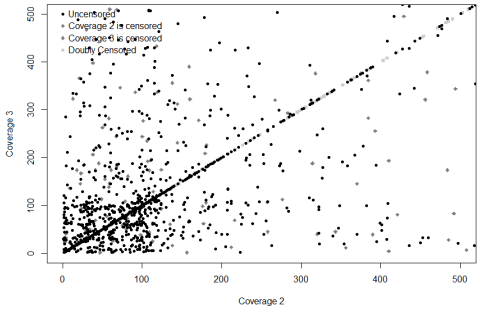

FIGURE 3.8 2^{nd} to 3^{rd} coverages

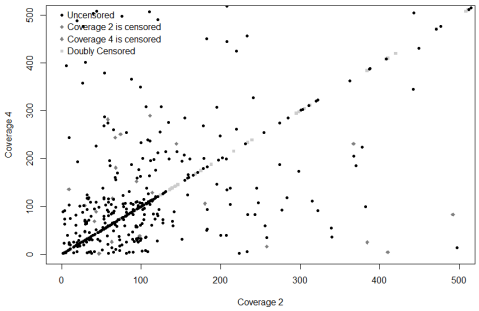

FIGURE 3.9 2^{nd} to 4^{th} coverages

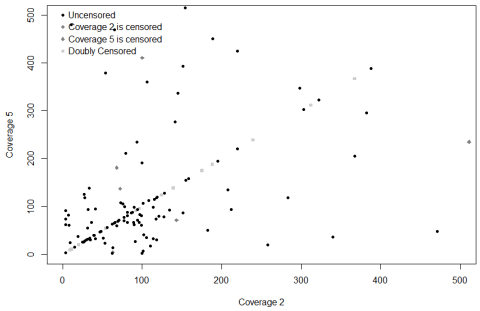

FIGURE 3.10 2^{nd} to 5^{th} coverages

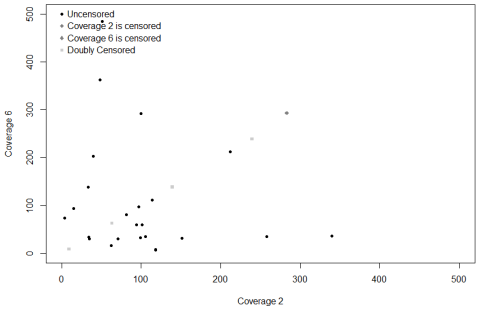

FIGURE 3.11 2^{nd} to 6^{th} coverages

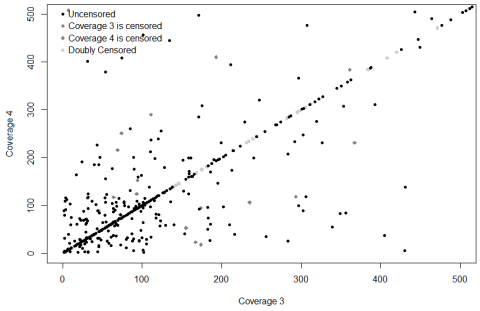

FIGURE 3.12 3^{rd} to 4^{th} coverages

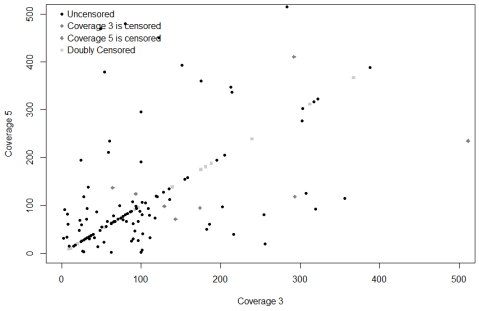

FIGURE 3.13 3^{rd} to 5^{th} coverages

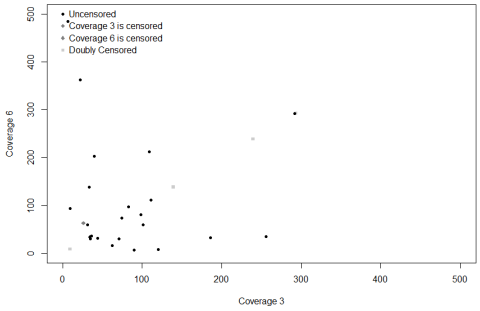

FIGURE 3.14 3^{rd} to 6^{th} coverages

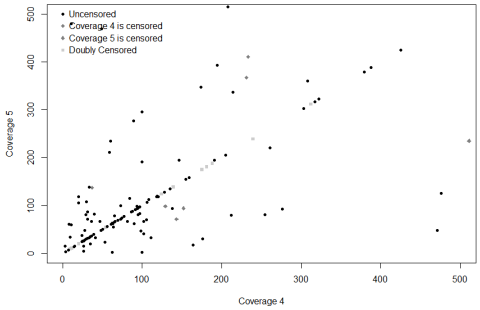

FIGURE 3.15 4^{rd} to 5^{th} coverages

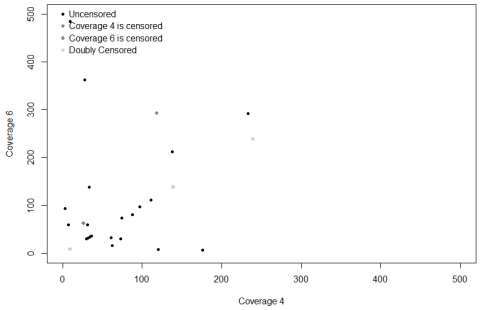

FIGURE 3.16 4^{rd} to 6^{th} coverages

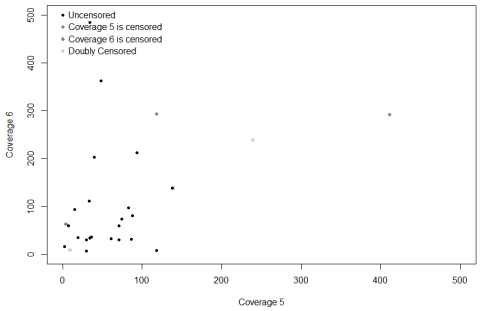

FIGURE 3.17 5^{rd} to 6^{th} coverages

TABLEAU 3.9: Estimated values of the covariate parameters

Variable		Without Frailty					With Frailty		
		Cox	Cox-spline	Weibull	Log-Log	Log-N	Cox	Cox-spline	Weibull
Type of loss	Single-vehicle	-0.18	-0.18	-0.18	0.08	0.10	-0.17	-0.16	-0.16
	Multi vehicle	-0.36	-0.36	-0.36	0.33	0.35	-0.53	-0.50	-0.49
	Hit Pedestrian	-0.69	-0.70	-0.71	0.62	0.62	-1.06	-1.00	-0.99
	Other	-0.64	-0.64	-0.64	0.46	0.45	-0.76	-0.74	-0.73
Injured gender	Male	0.10	0.10	0.09	-0.14	-0.15	0.21	0.19	0.19
	Unknown	0.37	0.36	0.34	-0.34	-0.34	0.45	0.42	0.42
Region	Ontario	-0.35	-0.35	-0.36	0.39	0.32	-0.66	-0.61	-0.60
	West	0.69	0.71	0.74	-0.48	-0.59	0.80	0.78	0.79
Injured age	$(18-25]$	-0.01	-0.01	-0.01	0.06	0.05	-0.11	-0.10	-0.10
	$(25-30]$	-0.17	-0.18	-0.17	0.19	0.20	-0.32	-0.30	-0.30
	$[30-50]$	-0.25	-0.26	-0.25	0.26	0.26	-0.43	-0.40	-0.40
	$(50-70]$	-0.30	-0.31	-0.30	0.29	0.28	-0.48	-0.46	-0.45
	$(70-\infty)$	-0.12	-0.12	-0.11	0.12	0.10	-0.22	-0.20	-0.20
	Unknown	0.18	0.19	0.20	-0.18	-0.19	0.31	0.29	0.29

Variable		Without Frailty					With Frailty		
		Cox	Cox-spline	Weibull	Log-Log	Log-N	Cox	Cox-spline	Weibull
$t_\ell^{(r)}$	$(1-7]$	-0.12	-0.12	-0.12	0.09	0.09	-0.19	-0.17	-0.17
	$(7-30]$	-0.19	-0.19	-0.19	0.13	0.16	-0.36	-0.33	-0.32
	$(30-90]$	-0.29	-0.29	-0.27	0.02	0.02	-0.58	-0.52	-0.51
	$(90-180]$	-0.67	-0.66	-0.63	0.03	0.04	-1.24	-1.12	-1.10
	$(180-365]$	-1.09	-1.12	-1.13	0.33	0.33	-1.88	-1.76	-1.75
	$(365-\infty)$	-1.73	-1.75	-1.85	0.13	0.18	-2.96	-2.77	-2.66
Vehicle age	$(3-6]$	-0.02	-0.02	-0.02	0.86	1.00	-0.02	-0.02	-0.02
	$(6-10]$	-0.03	-0.03	-0.03	1.31	1.46	-0.05	-0.05	-0.04
	$(10-20]$	-0.08	-0.08	-0.08	0.24	0.34	-0.14	-0.13	-0.13
	$(20-\infty)$	-0.13	-0.13	-0.15	0.41	0.54	-0.19	-0.19	-0.19
	Unknown	-0.34	-0.34	-0.35	1.91	2.09	-0.52	-0.49	-0.49
Initial reserve	$(1,000-5,000]$	0.02	0.02	0.04	0.02	0.04	-0.01	-0.00	0.00
	$(5,000-10,000]$	-0.50	-0.51	-0.53	0.60	0.61	-0.83	-0.79	-0.78
	$(10,000-20,000]$	-0.82	-0.83	-0.88	0.96	0.92	-1.40	-1.33	-1.29
	$(20,000-\infty)$	-1.14	-1.15	-1.22	1.23	1.24	-1.82	-1.73	-1.67

TABLEAU 3.10: p-values of the t-tests for the covariate parameters

Variable		Without Frailty					With Frailty		
		Cox	Cox-spline	Weibull	Log-Log	Log-N	Cox	Cox-spline	Weibull
Type of loss	Single-vehicle	< 0.01	< 0.01	< 0.01	0.01	< 0.01	< 0.01	< 0.01	< 0.01
	Multi vehicle	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	Hit Pedestrian	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	Other	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Injured gender	Male	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	Unknown	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Region	Ontario	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	West	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Injured age	$(18 - 25]$	0.68	0.63	0.81	0.03	0.07	< 0.01	< 0.01	< 0.01
	$(25 - 30]$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$[30 - 50]$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$(50 - 70]$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$(70 - \infty)$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	Unknown	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

Variable		Without Frailty					With Frailty		
		Cox	Cox-spline	Weibull	Log-Log	Log-N	Cox	Cox-spline	Weibull
$t_\ell^{(r)}$	$(1-7]$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$(7-30]$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$(30-90]$	< 0.01	< 0.01	< 0.01	0.16	0.33	< 0.01	< 0.01	< 0.01
	$(90-180]$	< 0.01	< 0.01	< 0.01	0.02	0.02	< 0.01	< 0.01	< 0.01
	$(180-365]$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$(365-\infty)$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Vehicle age	$(3-6]$	0.21	0.18	0.11	< 0.01	< 0.01	0.34	0.39	0.41
	$(6-10]$	0.05	0.06	0.03	< 0.01	< 0.01	0.03	0.04	0.04
	$(10-20]$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$(20-\infty)$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	Unknown	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Initial Reserve	$(1,000-5,000]$	0.13	0.08	< 0.01	0.04	< 0.01	0.75	0.97	0.79
	$(5,000-10,000]$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$(10,000-20,000]$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$(20,000-\infty)$	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

CONCLUSION

Dans cette thèse on a présenté des contributions portant sur les micro-réserves en assurance I.A.R.D. Cette branche de recherche se concentre sur la modélisation des réserves au niveau micro, c'est-à-dire sur le développement individuel de chaque réclamation. Dans ce contexte, il est nécessaire de mettre en place des modèles capables de prédire et de simuler le coût de chaque réclamation après la date d'évaluation. Pour ce faire, on a proposé une méthode hiérarchique dans laquelle le développement d'une réclamation est divisé en trois éléments principaux : la durée, la fréquence et la sévérité. Pour la durée, on a d'abord établi la problématique principale de cette variable en considérant le délai de survenance, le délai de déclaration et le délai de fermeture. Plus concrètement, dans le Chapitre 1 de cette thèse, on a présenté des méthodes paramétriques basées sur la littérature des modèles de survie, ainsi établissant une structure de base. Par la suite, cette variable a été reprise dans le Chapitre 3 au cours duquel la problématique de la dépendance entre les couvertures a été couverte par l'utilisation d'effets aléatoires associés à chacune de celles-ci au sein d'une réclamation. Pour la fréquence des paiements, on a utilisé la durée comme mesure d'exposition afin de mettre en place des modèles de comptage dépendant de cette mesure. De façon plus précise, on a considéré au Chapitre 2 des méthodes de type Bonus-Malus, souvent utilisées dans un contexte de tarification, afin d'inclure le passé d'une réclamation dans la modélisation. Finalement, au niveau de la sévérité on a proposé des méthodes basées sur la littérature des évènements extrêmes afin de mieux prendre en compte le caractère volatil des paiements individuels.

La structure hiérarchique nous a permis d'incorporer de l'information individuelle

à chaque étape de l'ajustement. En effet, la motivation principale de ce projet est justement l'inclusion de ce type de données. Dans le cadre du Chapitre 1, on a étudié les variables explicatives statiques et temporelles déterministes alors que le nombre de paiements passés, qui n'est pas déterministe, a été étudié dans le Chapitre 2. D'autre part, au Chapitre 3, une analyse par couverture, plutôt que par réclamation, a permis de plus facilement inclure des variables explicatives qui diffèrent entre les couvertures d'une même réclamation. Afin de justifier l'utilisation de modèles capables d'incorporer ces données, plusieurs analyses minutieuses ont été faites. D'une part, des mesures de la qualité de l'ajustement, telles que les critères AIC et BIC, ont démontré la significativité statistique des variables explicatives individuelles étudiées. D'autre part, lors de la comparaison entre les résultats simulés par les modèles et les valeurs observées, on a aussi démontré que les modèles qui prenaient en compte ces variables performaient mieux que les modèles qui omettaient cette information.

Par ailleurs, la structure à trois composantes mise en avant dans cette thèse ouvre la porte à des extensions afin de capturer d'autres aspects du développement des réclamations. Par exemple, on pourrait adapter les modèles proposés afin de prendre en compte des variables explicatives plus complexes telles que l'état de santé de la personne ou l'importance des moyens médicaux mis en place pour aider une victime d'accident. En effet, la nature non déterministe de ces variables implique des considérations particulières afin de pouvoir bonifier les modèles. En outre, il pourrait être possible de considérer la dépendance entres les composantes elles-mêmes, par exemple entre la sévérité et le nombre des paiements, en plus des structures de dépendances étudiées dans la cadre des Chapitre 2 et 3. Enfin, on peut mentionner que les méthodes paramétriques et semi-paramétriques présentées dans cette thèse pourraient être modifiées afin d'inclure des éléments d'apprentissage statistique.

RÉFÉRENCES

Akritas, M. G., Murphy, S. A., & Lavalley, M. P. (1995). The Theil-Sen estimator with doubly censored data and applications to astronomy. *Journal of the American Statistical Association*, 90(429), 170-177.

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

Antonio, K., Godecharle, E., & Van Oirbeek, R. (2016). A multi-state approach and flexible payment distributions for micro-level reserving in general insurance. Available at SSRN 2777467.

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

Avanzi, B., Wong, B., & Yang, X. (2016). A micro-level claim count model with overdispersion and reporting delays. *Insurance : Mathematics and Economics*, 71, 1-14.

Avanzi, B., Taylor, G., Vu, P. A., & Wong, B. (2020). A multivariate evolutionary generalised linear model framework with adaptive estimation for claims reserving. *Insurance : Mathematics and Economics*.

Avanzi, B., Taylor, G., Wong, B., & Yang, X. (2021). On the modelling of multivariate counts with Cox processes and dependent shot noise intensities. *Insurance : Mathematics and Economics*, 99, 9-24.

Ayuso, M., Guillen, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics : incorporating mileage and driver behaviour data. *Transportation*, 46(3), 735-752.

Badescu, A. L., Lin, X. S., & Tang, D. (2016). A marked Cox model for the number of IBNR claims : Theory. *Insurance : Mathematics and Economics*, 69, 29-37.

Badescu, A. L., Chen, T., Lin, X. S., & Tang, D. (2019). A marked Cox model for the number of IBNR claims : estimation and application. *ASTIN Bulletin*, 49(3), 709-739.

Balan, T. A. & Putter, H. (2019). frailtyEM : An R package for estimating semi-parametric shared frailty models. *Journal of Statistical Software*, 90, 1-29.

Balan, T. A., & Putter, H. (2020). A tutorial on frailty models. *Statistical methods in medical research*, 29(11), 3424-3454.

Baudry, M., & Robert, C. Y. (2019). A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry*, 35, 1127-1155.

Bischofberger, S. M. (2020). In-sample hazard forecasting based on survival models with operational time. *Risks*, 8(1), 3.

Blier-Wong, C., Cossette, H., Lamontagne, L. & Marceau, E. (2020). Machine learning in PC insurance : A review for pricing and reserving. *Risks*, 9(1), 4.

Boucher, J-. P., & Inoussa, R. (2014). A posteriori ratemaking with panel data. *ASTIN Bulletin*, 44(3), 587-612.

Boucher, J-. P., & Pigeon, M. (2019). A Claim Score for Dynamic Claim Counts Modelling, *Canadian Institute of Actuaries.*

Boucher, J-. P. (2023). Bonus-Malus Scale Models : Creating Artificial Past Claims History. *Annals of Actuarial Science* 17.1 (2023) : 36-62.

Byar, D. P. (1982). Cox and Weibull models with covariates. *Statistics in medical research*, 365-401.

Breslow NE (1972) Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society : Series B*, 34, 216–217.

Carroll, K. J. (2003). On the use and utility of the Weibull model in the analysis of survival data. *Controlled clinical trials*, 24(6), 682-701.

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

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society : Series B*, 34(2), 187-202.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Annals of Statistics*, 1475-1489.

Delong, Ł., Lindholm, M., & Wüthrich, M. V. (2022). Collective reserving using individual claims data. *Scandinavian Actuarial Journal*, 2022(1), 1-28.

Denuit, M., & Trufin, J. (2017). Beyond the Tweedie reserving model : The collective approach to loss development. *North American Actuarial Journal*, 21(4), 611-619.

Denuit, M., & Lu, Y. (2021). Wishart-gamma random effects models with applications to nonlife insurance. *Journal of Risk and Insurance*, 88(2), 443-481.

Duval, F., & Pigeon, M. (2019). Individual loss reserving using a gradient boosting-based approach. *Risks*, 7(3), 79.

Eden, S. K., Li, C. & Shepherd, B. E. (2022). Nonparametric estimation of Spearman's rank correlation with bivariate survival data. *Biometrics*, 78(2), 421-434.

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

Gabrielli, A. (2020). A neural network boosted double overdispersed Poisson claims reserving model. *ASTIN Bulletin*, 50(1), 25-60.

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

Helsel, D. R. (2012). *Statistics for Censored Environmental Data Using Minitab and R*, John Wily & Sons. Inc. : Hoboken, NJ.

Hiabu, M., Mammen, E., Martínez-Miranda, M. D., & Nielsen, J. P. (2016). In-sample forecasting with local linear survival densities. *Biometrika*, 103(4), 843-859.

Hiabu, M., Margraf, C., Martínez-Miranda, M. D., & Nielsen, J. P. (2016). The link between classical reserving and granular reserving through double chain ladder and its extensions. *British Actuarial Journal*, 21(1), 97-116.

Hiabu, M. (2017). On the relationship between classical chain ladder and granular reserving. *Scandinavian Actuarial Journal*, 2017(8), 708-729.

Huang, J., Qiu, C., & Wu, X. (2015a). Stochastic loss reserving in discrete time : Individual vs. aggregate data models. *Communications in Statistics - Theory and Methods*, 44(10), 2180–2206.

Huang, J., Qiu, C., Wu, X., & Zhou, X. (2015b). An individual loss reserving model with independent reporting and settlement. *Insurance : Mathematics and Economics*, 64, 232–245.

Huang, J., Wu, X., & Zhou, X. (2016). Asymptotic behaviors of stochastic reserving : Aggregate versus individual models. *European Journal of Operational Research*, 249(2), 657–666.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 795-806.

Kuo, K. (2019). Deeptriangle : A deep learning approach to loss reserving. *Risks*, 7(3), 97.

Laudagé, C., Desmettre, S., & Wenzel, J. (2019). Severity modeling of extreme insurance claims for tariffication. *Insurance : Mathematics and Economics*, 88, 77-92.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data* (Vol. 362). John Wiley & Sons.

Lemaire, J. (1995). *Bonus-malus systems in automobile insurance* (Vol. 19). Springer science & business media.

Lopez, O., Milhaud, X., & Thérond, P. E. (2016). Tree-based censored regression with applications in insurance. *Electronic journal of statistics*, 10(2), 2685-2716.

Lopez, O. (2019). A censored copula model for micro-level claim reserving. *Insurance : Mathematics and Economics*, 87, 1-14.

Lopez, O., Milhaud, X., & Thérond, P. E. (2019). A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin*, 49(3), 741-762.

Lopez, O., Milhaud, X., & Thérond, P. E. (2019). A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin*, 49(3), 741-762.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, 23(02), 213-225.

Mack, T. (1999). The standard error of chain ladder reserve estimates : Recursive calculation and inclusion of a tail factor. *ASTIN Bulletin*, 29(02), 361-366.

Maciak, M., Okhrin, O., & Pešta, M. (2021). Infinitely stochastic micro reserving. *Insurance : Mathematics and Economics*, 100, 30-58.

Nielsen, G. G., Gill, R. D., Andersen, P. K. & Sørensen, T. I. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian journal of Statistics*, 25-43.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance1. *ASTIN Bulletin*, 23(1), 95-115.

Norberg, R. (1999). Prediction of outstanding liabilities II. Model variations and extensions. *ASTIN Bulletin*, 29(1), 5-25.

Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models* (Vol. 2). Berlin : Springer.

Okine, A. N. A., Frees, E. W., & Shi, P. (2022). Joint model prediction and application to individual-level loss reserving. *ASTIN Bulletin*, 52(1), 91-116.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on scientific and statistical computing*, 9(2), 363-379.

Pigeon, M., Antonio, K., & Denuit, M. (2013). Individual loss reserving with the multivariate skew normal framework. *ASTIN Bulletin*, 43(3), 399-428.

Pigeon, M., Antonio, K., & Denuit, M. (2014). Individual loss reserving using paid–incurred data. *Insurance : Mathematics and Economics*, 58, 121-131.

Rondeau, V., Commenges, D. & Joly, P. (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime data analysis*, 9(2), 139-153.

Rondeau, V., Marzroui, Y. & Gonzalez, J. R. (2012). frailtypack : an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47, 1-28.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American statistical association*, 63(324), 1379-1389.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible regression and smoothing : using GAMLSS in R*. Chapman and Hall/CRC.

Taylor, G., McGuire, G., & Sullivan, J. (2008). Individual claim loss reserving conditioned by case estimates. *Annals of Actuarial Science*, 3(1-2), 215-256.

Taylor, G. (2019). Loss Reserving Models : Granular and Machine Learning Forms. *Risks*, 7, 82.

Therneau, T. M., Grambsch, P. M., & Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of computational and graphical statistics*, 12(1), 156-175.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes mathematicae*, 12(85), 173.

Vaida, F. & Xu, R. (2000). Proportional hazards model with random effects. *Statistics in medicine*, 19(24), 3309-3324.

Verrall, R., Nielsen, J. P., & Jessen, A. H. (2010). Prediction of RBNS and IBNR claims using claim amounts and claim counts. *ASTIN Bulletin*, 40(2), 871-887.

Verrall, R., & Wüthrich, M. (2016). Understanding reporting delay in general insurance. *Risks*, 4(3), 25.

Verschuren, R. M. (2021). Predictive claim scores for dynamic multi-product risk classification in insurance. *ASTIN Bulletin*, 51(1), 1-25.

Wang, Z., Wu, X., & Qiu, C. (2021). The impacts of individual information on loss reserving. *ASTIN Bulletin*, 51(1), 303-347.

Wei, L. J. (1992). The accelerated failure time model : a useful alternative to the Cox regression model in survival analysis. *Statistics in medicine*, 11(14-15), 1871-1879.

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

Wüthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 2018(6), 465-480.

Yanez, J. S., & Pigeon, M. (2021). Micro-level parametric duration-frequency-severity modeling for outstanding claim payments. *Insurance : Mathematics and Economics*, 98, 106-119.

Yanez, J. S., Boucher, J-.P., & Pigeon, M. (2022). Modelling payment frequency for loss reserves based on dynamic claim scores. *submitted for publication.*

Yanez, J. S., & Pigeon, M. (2023). Dependence in Claim Processing Time : A Frailty Analysis Perspective. *submitted for publication.*

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

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