ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ADMINISTRATION ÉCONOMIQUE
3, avenue Pierre Larousse – 92245 Malakoff CEDEX, France

Mémoire d'Actuariat – Promotion 2004

MODELE DE PROVISIONNEMENT SUR DONNEES DETAILLEES EN ASSURANCE NON-VIE

RESERVING MODEL BASED ON DETAILED DATA IN NON-LIFE INSURANCE

GUILLAUME BENETEAU

MOTS-CLES : PROVISIONNEMENT, METHODES STOCHASTIQUES, GLM, IBNR, MODELES INDIVIDUELS

KEYWORDS: CLAIMS RESERVING, STOCHASTIC METHODS, GLM, IBNR, MICRO MODELS

Towers-Perrin Tillinghast, mai - octobre 2004

Directeur de mémoire : Michel Laparra
Correspondant ENSAE : Arthur Charpentier

ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ADMINISTRATION ÉCONOMIQUE
3, avenue Pierre Larousse – 92245 Malakoff CEDEX, France

Actuarial Thesis – Year 2004

RESERVING MODEL BASED ON DETAILED DATA IN NON-LIFE INSURANCE

GUILLAUME BENETEAU

- Abstract -

KEYWORDS: CLAIMS RESERVING, STOCHASTIC METHODS, GLM, IBNR, MICRO MODELS

The insurance has the distinctive nature to present an inversed economic cycle, that is, the final cost of the product is only known long after having collected the premium. However, the accounting principals of prudence and sincerity require from the insurer to recognize the ultimate cost of its liabilities in his accounts. This cost can be split into two parts: payments and outstanding claims reserve. By doing so, the actuaries have to anticipate the statistical time deviation of charges for reported losses, and late reporting of claims.

- Presentation

The most widespread methodology to assess the ultimate value of liabilities is the Chain-Ladder method. Data are usually aggregated by accident year and development year. In this model, the assumption is made that the aggregate payments' pattern is independent

from accident years. This model calculates the aggregate payments' pattern using historical data.

Yet many factors may have an influence on the payments' pattern, especially the changes in the claim handling procedures, or the occurrence of major events. For instance, in 1999, Lothar and Martin caused a lot of delay in claim management.

To account for the uncertainty of these models, stochastic methods have been developed. They assume that aggregate amounts are random variables. In particular, the Mack model defines precisely the assumptions made in the Chain-Ladder method and provides associated confidence intervals. The number of stochastic models boomed with the introduction of generalized linear models (GLM). In this framework, incremental payments are supposed to be independent and they belong to the exponential family.

However, these models suffer from drawbacks: for instance, they can hardly be used for non proportional reinsurance. On the contrary, we can study detailed data as well as individual losses. In order to do this, we will have to identify the main steps of a claim management process (occurrence, date of report, payments and settlement). Reported losses have been studied in this work.

- Reserving model based on detailed data

The data we studied comes from a professional legal liability portfolio. Such losses have a long time of development as they often need expert opinion or law decisions. This kind of data is much more difficult to evaluate with classical methods than car or house insurance. That's why studying a reserving model on detailed data can be very interesting with this portfolio.

We have constructed a model that describes the claim process. The loss development is described by a sequence of dates of payment and the associated cash-flows, and by a settlement delay.

- The dataset we studied especially includes the nature of each cash-flow (indemnity payment, direct expenses, salvage / subrogation...). Each nature of cash-flow has been modelled by a parametric distribution.

- Generalized Linear Models have been used to explain the dates of payments. In our database, we noticed that the first payment had to be considered separately from the following ones.

- Survival analysis is adapted for the study of settlement delays. A lot of claims are still opened in our database and this kind of method has the advantage to take into account the information carried by those claims.

- Prediction of outstanding liabilities

In order to understand what the effects of the model based on detailed data are, it is interesting to compare the estimates of outstanding liabilities of this model with classical methods. This has been done including all extremities but no tail factors were added.

The most widespread models are based on aggregate data. The Chain-Ladder method shows its limits with this portfolio. The predictions associated with recent years diverge. It is necessary to incorporate expert opinion. It seems that the settlement pattern has indeed changed, making it inappropriate to use the same development factor for each row.

The same phenomenon can be observed within a stochastic framework. Classical stochastic models try to extend the Chain-Ladder results. Moreover, the uncertainty given by the Mack model is not really workable.

Therefore, it seemed interesting to study new models, in particular a model of Verral which incorporates expert opinion within a stochastic framework. The results given by such a model are much more consistent. The ultimate claims associated with recent years and confidence intervals no longer diverge.

Then these various results have been compared with those given by the reserving model based on detailed data. The outstanding liabilities of the detailed model are quite similar to those of models which incorporate expert opinion. Furthermore, the uncertainty associated to this model seems to be consistent. The detailed model gives an alternative to classical reserving methods. And it is also coherent with expert opinion by modelling individual complex phenomena.

ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ADMINISTRATION ÉCONOMIQUE
3, avenue Pierre Larousse – 92245 Malakoff CEDEX, France

Mémoire d'Actuariat – Promotion 2004

MODELE DE PROVISIONNEMENT SUR DONNEES DETAILLEES EN ASSURANCE NON-VIE

GUILLAUME BENETEAU

- Note de synthèse -

MOTS-CLES : PROVISIONNEMENT, METHODES STOCHASTIQUES, GLM, IBNR, MODELES INDIVIDUELS

L'assurance a la spécificité d'avoir un cycle de production inversé : le coût total du produit n'est connu qu'après sa vente. Cependant, les principes comptables de sincérité et de prudence requièrent de l'assureur qu'il reconnaisse le coût final de ses engagements dans ses comptes. Ce coût peut se décomposer en deux parties : les règlements et la provision pour sinistres à payer (PSAP). En faisant cela, l'assureur doit anticiper deux phénomènes : la dérive dans le temps des charges de sinistres déclarés et les déclarations tardives.

- Exposé du problème

Pour calculer la charge ultime d'un portefeuille, la méthode la plus utilisée est la méthode Chain-Ladder. Les données sont généralement agrégées par année de survenance et par année de développement. Ce modèle permet de calculer une cadence de développement moyenne à partir des données historiques.

Seulement, de nombreux facteurs peuvent influer sur la cadence des paiements. On peut en particulier citer les changements de procédure dans la gestion des sinistres, la survenance d'évènements exceptionnels en assurance dommage, ou les changements de jurisprudence en responsabilité civile. Ainsi, les tempêtes Lothar et Martin du 26 et 27 décembre 1999 ont engendré de nombreux retards dans la gestion des sinistres.

Pour identifier l'incertitude liée à ces modèles, des méthodes stochastiques ont été introduites. Elles supposent que les montants agrégés cumulés ou incrémentaux sont des variables aléatoires. Le modèle de Mack permet en particulier d'expliciter les hypothèses utilisées dans la méthode Chain-Ladder et de déterminer des intervalles de confiance. Ce type d'approche s'est développé avec l'introduction des modèles linéaires généralisés (GLM – Generalized Linear Models). Dans le cadre de ces modèles, les montants incrémentaux sont supposés indépendants et distribués selon une loi de probabilité de la famille exponentielle.

Cependant, ces modèles présentent certaines limites. En particulier, ils appréhendent mal le cas où l'engagement de l'assureur est limité. Aussi, il peut être intéressant de travailler sur des données détaillées et de revenir au sinistre individuel. Pour cela, nous allons devoir déterminer les étapes clés dans la vie d'un sinistre (survenance, déclaration, paiements et clôture).

- Modèle de provisionnement individuel

Le modèle présenté ici étudie les sinistres ayant déjà été déclarés à l'assureur. Nous avons utilisé un portefeuille de responsabilité civile (RC) professionnelle pour étudier la chronique future des paiements de chacun des sinistres. Il s'agit de sinistres ayant une durée de vie longue. Ils réclament souvent des avis d'experts ou des passages devant les tribunaux. Ce type de données s'évalue beaucoup plus difficilement par les méthodes classiques que des sinistres automobiles ou MRH (multirisque habitation). C'est pourquoi utiliser un modèle de provisionnement individuel sur ce type de données est particulièrement intéressant.

Nous avons établi un modèle permettant de décrire le processus de sinistralité. Le développement d'un sinistre est décrit par un ensemble de dates de paiements, de montant associés et par une date de clôture.

- Dans les données dont nous disposons, nous connaissons la nature de chaque flux (paiement en principal, frais, honoraire ou recours). Les montants associés à chaque nature de flux ont été modélisés par des lois de probabilité paramétriques.

- En ce qui concerne les dates associées, nous les avons étudiées par des modèles linéaires généralisés. Sur notre jeu de données, on observe qu'il est important d'étudier séparément le premier paiement et les suivants.

- Enfin, l'analyse de survie s'adapte bien à l'étude des dates de clôture. En effet, beaucoup de sinistres sont encore ouverts à l'issue de l'observation et ce type de méthode permet de prendre en compte ces sinistres pour la modélisation.

- Calcul des réserves

Pour bien comprendre quels sont les effets du modèle de provisionnement individuel, il est intéressant de comparer les résultats obtenus avec des modèles classiques de provisionnement. Pour faciliter la comparaison entre les modèles, nous n'avons pas introduit de retraitement pour les sinistres graves ni de facteur de queue.

Les modèles les plus souvent utilisés se basent sur des données agrégées. Sur le portefeuille étudié, la méthode Chain-Ladder montre rapidement ses limites. Les prévisions divergent pour les années de développement les plus récentes. Il est nécessaire d'introduire des avis d'experts pour retraiter les données. Il semble en effet qu'il y ait une rupture observée dans la cadence de développement entre 1997 et les autres années.

Le même phénomène se retrouve lorsque l'on étudie des méthodes stochastiques. Les modèles stochastiques classiques visent à étendre les résultats proposés par la méthode Chain-Ladder. Aussi le modèle de Mack présente des intervalles de confiance trop importants pour être exploitables.

Il nous a donc semblé intéressant d'étudier des modèles plus récents, en particulier un modèle de Verral permettant l'introduction d'avis d'experts dans un cadre stochastique. Les résultats d'un tel modèle sont beaucoup plus cohérents. La sinistralité associée aux années récentes ne diverge plus et les intervalles de confiance associés sont acceptables.

Nous avons alors comparé ces divers résultats aux résultats fournis par le modèle individuel. Les résultats de ce modèle sont très encourageants. On retrouve en effet des

réserves similaires aux modèles permettant l'introduction d'avis d'experts. De plus, l'incertitude liée à ce modèle semble cohérente. Ce modèle propose donc une alternative aux modèles de provisionnement classiques. Ils permettent également de valider l'avis d'expert en s'attachant à la modélisation de phénomènes individuels complexes.

Je remercie tout particulièrement Fabien Faivre, Stéphane Chappellier et Michel Laparra qui m'ont conseillé et aiguillé tout au long de ce stage pour faire de ce travail de recherche une expérience enrichissante.

Je remercie également Arthur Charpentier qui a toujours été disponible pour répondre à mes nombreuses questions.

Table des Matières

INTRODUCTION

L'assurance a la spécificité d'avoir un cycle de production inversé : le coût d'un produit n'est connu qu'après sa vente. En effet, dans un premier temps, l'assureur reçoit une prime. Et dans un second temps, ce dernier devra payer à l'assuré un certain montant si un sinistre survient.

Aussi, pour assurer sa solvabilité, une compagnie d'assurance doit identifier et provisionner ses risques. A tout moment, elle doit être capable de faire face à ses engagements vis-à-vis de ses assurés.

La charge finale de tous les sinistres survenus durant la période d'observation étant inconnue, la compagnie d'assurance doit l'estimer. Les provisions techniques constituent la différence entre cette charge finale estimée et les paiements déjà effectués. L'établissement des ces provisions est très important en assurance IARD.

Il existe plusieurs phénomènes à l'origine des provisions techniques. D'une part, il y a les sinistres qui ont déjà été déclarés à l'assureur mais qui n'ont pas été totalement réglés. Pour ces derniers, une provision pour sinistre à payer (PSAP) est évaluée dossier par dossier. Cependant, cette provision peut être insuffisante et l'actuaire doit calculer une provision pour l'ensemble des sinistres déclarés mais non suffisamment provisionnés. D'autre part, il y a les sinistres survenus qui n'ont pas encore été déclarés à l'assureur. Une provision pour l'ensemble de ces sinistres doit également être calculée.

Les méthodes classiques d'estimation de la charge finale utilisent des données agrégées. Toutes les données issues par exemple de la même année de survenance et de la même année de développement sont regroupées. Ce type de méthode présente l'avantage d'être simple à utiliser et robuste à l'erreur de modèle.

Pourtant, l'utilisation de données agrégées engendre nécessairement une perte d'information. De plus, dans ce type de modèle, il est difficile, voire impossible, d'introduire des contraintes particulières (réassurance non proportionnelle,...). C'est pourquoi il peut être intéressant d'utiliser des données détaillées et de revenir au sinistre individuel. Pour cela, nous avons besoin de modéliser les diverses étapes de la vie d'un sinistre. L'objectif de cette étude est de proposer un modèle portant sur les sinistres ayant été déjà déclarés à l'assureur. Les provisions pour sinistres survenus mais non déclarés ne sont donc pas modélisées ici.

La méthode que nous avons développée ici est une méthode ad hoc. C'est pourquoi il sera indispensable de comparer les résultats de cette méthode avec les méthodes classiques de provisionnement (Chain-Ladder, bootstrap,...). Nous avons également choisi de comparer les résultats de notre modèle avec deux nouvelles méthodes. La première permet l'introduction d'avis d'experts dans un cadre stochastique, et la seconde modélise les corrélations existant dans les cadences de paiement.

Enfin, notre modèle a pour objectif non seulement de déterminer la charge ultime, mais également de déterminer la cadence de règlement des sinistres. En effet, nous cherchons une méthode qui soit compatible avec les nouvelles méthodes IAS et potentiellement applicables à des réglementations étrangères similaires à la Grande Bretagne.

1. EXPOSE DU PROBLEME

1.1 L'établissement des réserves

Généralités sur les modèles de provisionnement

Le coût ultime des sinistres

L'établissement des réserves est très important en IARD. Pour chaque sinistre déclaré à l'assureur, une provision dossier / dossier est évaluée. Des réserves doivent également être prévues pour les sinistres survenus mais non déclarés, les IBNYR (*incurred but not yet reported*) et pour les sinistres déclarés mais non suffisamment provisionnés, les IBNER (*incurred but not enough reported*).

Portefeuille de contrats

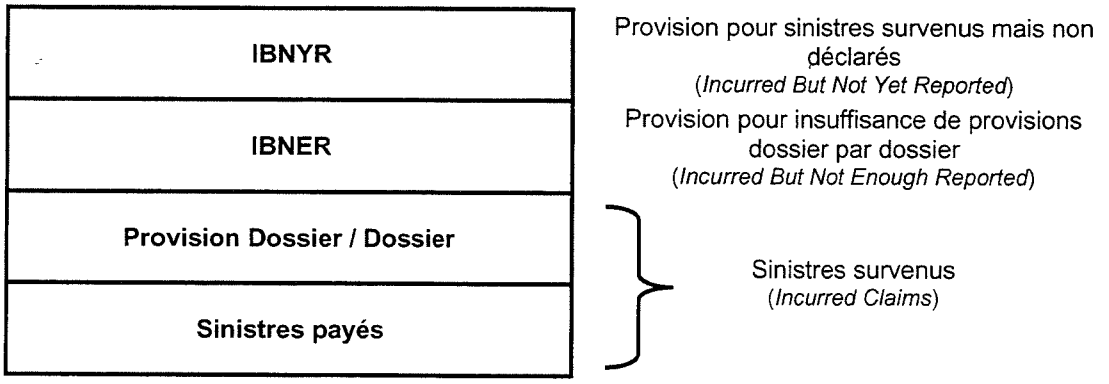


fig. 1.1 : Répartition du coût ultime des sinistres

Calcul d'une cadence de règlement

Avec la mise en place des nouvelles normes comptables IFRS04, déterminer une cadence de règlement est indispensable pour calculer les provisions d'un assureur. Cette norme suppose en effet que les provisions doivent être calculées à partir des flux futurs.

De plus, les modèles de provisionnement ne permettent pas seulement d'établir les provisions techniques. En évaluant la sinistralité ultime, ils interviennent dans la détermination d'une politique de tarification. Ils sont également utilisés pour valoriser une compagnie en run-off ou pour l'estimation du capital économique d'une entreprise (capital à détenir pour faire face à ses engagements). Or, pour la modélisation de ces divers phénomènes, il est important de connaître la cadence de règlement des sinistres. Aussi, nous ne cherchons pas à déterminer un modèle nous permettant uniquement de définir la charge ultime d'un portefeuille de sinistres, mais permettant également de calculer une cadence de paiements des sinistres.

Typologie des méthodes de provisionnement

Il existe différentes manières de calculer les réserves. On peut notamment opposer les méthodes stochastiques aux méthodes déterministes, les méthodes en temps discret aux méthodes en temps continu, les méthodes « micro » aux méthodes « macro » (les premières étudient le développement des réclamations individuelles, alors que les secondes agrègent les données).

Les méthodes le plus souvent utilisées se basent sur des données agrégées. Elles présentent l'avantage d'être simples d'utilisation. En général, elles permettent une modélisation cohérente de la sinistralité ultime et de la cadence des paiements associée. Cependant, certaines limites sont observées, et ce surtout dans les branches longues.

Pour faire face à ces limites, il peut être intéressant d'utiliser l'ensemble de l'information détenue par l'assureur au moment de l'évaluation de ses provisions.

Aussi, l'objectif de cette étude est de développer une méthode « micro » stochastique. Pour cela, nous devrons donner une description détaillée du processus de réclamation, avec en particulier les dates de survenance et de déclaration ainsi que le développement de chacun des sinistres.

De tels modèles vont inclure inévitablement un grand nombre de paramètres. Haastrup et Arjas les ont étudiés par des méthodes bayésiennes (une loi a priori est spécifiée). La distribution prédictive des réserves et la distribution a posteriori des paramètres sont déterminées à l'aide des simulations de Monte Carlo markoviennes (MCMC - Markov Chain Monte Carlo). Cependant, l'approche proposée reste très théorique.

Une méthode « micro » stochastique en temps continu présente de nombreux avantages par rapport aux méthodes dites classiques. On peut en particulier obtenir des intervalles de confiance. De plus, ce type de méthode permet de prendre en compte l'ensemble des informations disponibles. Par exemple, pour les méthodes classiques, les paiements sont typiquement évalués tous les ans. Si nous connaissons les paiements effectués quelques mois au-delà de la dernière année d'étude, ces paiements ne peuvent être pris en compte dans l'établissement des réserves. Comme le modèle est établi en temps continu, il est possible d'établir les réserves à partir de n'importe quelle date de départ.

Pour cette approche, il est nécessaire de modéliser :

- La loi des dates des cash-flows d'un sinistre
- La loi des montant associés

Pour cela, nous allons devoir déterminer quels sont les facteurs qui influent sur le processus de développement d'un sinistre.

1.2 Les méthodes classiques

Nous cherchons ici à présenter les grandes lignes des méthodes de provisionnement les plus couramment utilisées. Nous nous appuierons plus particulièrement sur leurs avantages et leurs inconvénients. Ces méthodes seront développées plus précisément dans la partie 4 afin de les comparer à notre modèle.

Ces modélisations utilisent des données agrégées et se basent sur des triangles de développement.

Pour déterminer les provisions, il est préférable que les données soient le plus homogène possible. Aussi, les provisions techniques sont évaluées par type de risque.

La méthode Chain-Ladder

Présentation du modèle

La technique d'évaluation des réserves par la méthode Chain-Ladder est l'une des plus anciennes techniques actuarielles. C'est également l'une des plus utilisées.

Les données sont regroupées par branches ou type de risque. Sur les différentes branches étudiées, les prestations à payer pour une compagnie d'assurance couvrent plusieurs années de développement. La sinistralité d'une branche est alors représentée par le triangle cumulé suivant :

	Année de développement				
Année de survenance	$C_{1,1}$	$C_{1,2}$	$\cdots$	$C_{1,n-1}$	$C_{1,n}$
	$C_{2,1}$	$C_{i,j}$	$\cdots$	$C_{2,n-1}$	
	$\vdots$	$\vdots$			
	$C_{n-1,1}$	$C_{n-1,2}$			
	$C_{n,1}$				

fig. 1.2 : Triangle de développement

où $C_{i,j}$ correspond au montant agrégé des sinistres survenus l'année i vus après j années de développement. Le triangle présenté ici est un triangle établi par année de survenance mais il est également possible d'étudier des triangles par année de déclaration.

Ce modèle cherche à établir une cadence des paiements (ou des charges) moyenne. De manière plus formelle, le modèle Chain-Ladder fait l'hypothèse que cette cadence est indépendante de l'année de survenance.

Cette technique utilise alors uniquement les données cumulées et calcule des facteurs de développement qui servent à l'établissement des réserves (cf. section 4.1).

Cette méthode présente l'avantage d'être très simple d'utilisation.

Limites de la méthode Chain-Ladder

Ce type de méthode peut manquer de précision dans la réalité. En effet, pour que cette méthode s'applique, il faut un grand nombre de contrats, avec une fréquence de sinistres qui ne soit pas trop faible. De plus, des irrégularités dans les cadences de règlement sont souvent observées. On peut citer en particulier :

- L'inflation
- L'évolution de la composition du portefeuille
- Les changements de souscripteurs ou les changements dans les modalités de souscription
- Les modifications de procédures de gestion des sinistres
- Les sinistres graves et les catastrophes
- Les tendances dans la fréquence et le coût moyen des sinistres

De plus, si l'étude porte sur des triangles de charges, l'évaluation de la sinistralité ultime dépend de la date de la dernière évaluation des provisions dossier par dossier.

Enfin, cette méthode ne permet d'obtenir qu'une estimation ponctuelle et non un intervalle de confiance.

Les méthodes stochastiques

Un grand nombre de méthodes stochastiques ont été introduites afin d'obtenir une estimation de l'erreur de prévision et d'intervalles de confiance (cf. section 4.1).

On peut en particulier citer le modèle de Mack. Il s'agit d'un modèle stochastique relatif à la méthode de Chain-Ladder. Il repose sur l'idée que les années de survenances sont indépendantes et que les facteurs de développement ne dépendent que de l'année de développement. La sinistralité ultime espérée obtenue est la même que celle calculée avec la méthode Chain-Ladder, mais le modèle de Mack permet en plus d'obtenir l'erreur de prévision.

Ce type de modèle a été généralisé par l'utilisation des modèles linéaires généralisés (GLM). De tels modèles supposent que les paiements incrémentaux $Y_{i,j}$ sont distribués selon une loi $L(A_i, B_j, C_{i+j-1})$ de telle sorte que la variable $Y_{i,j}$ dépend de trois facteurs :

- un facteur ligne A_i - par année de survenance -
- un facteur colonne B_j - par année de développement -
- un facteur diagonal C_{i+j-1} - par année calendaire -

Il y a donc un facteur dû à l'année de survenance, un facteur dû à l'année de développement, et un facteur dû à l'année calendaire.

Comme l'ont montré England et Verrall (2002) au travers des divers modèles qu'ils ont étudiés, ces modèles présentent tous en pratique une incertitude élevée.

Ceci nous a incité à travailler sur les sinistres individuels pour l'établissement des réserves.

1.3 Concepts généraux et cadre de l'étude

Le processus de sinistralité

A un moment dans le temps, un sinistre survient. Avec un certain délai, le sinistre est déclaré à la compagnie.

Le processus qui va de la déclaration à la clôture du sinistre s'appelle le développement. Il est décrit par un ensemble de dates et la nouvelle information associée à chaque date.

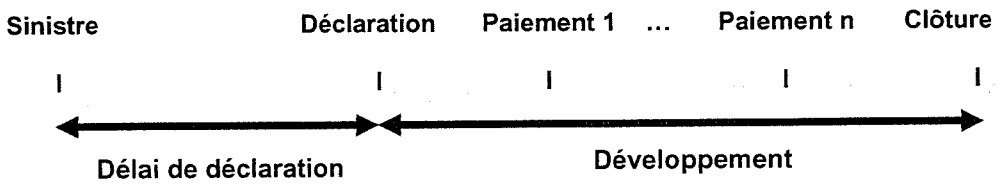


fig. 1.3 : Les étapes du processus de sinistralité

Chacune des dates est discrétisée par l'assureur. Mais si la discrétisation est suffisamment bonne (jour ou semaine), alors l'information sera traitée comme si elle était continue.

Lorsqu'un assureur étudie son portefeuille de contrats, il observe des sinistres déclarés (qui incluent les sinistres clos et les IBNER), dont le nombre est connu, et des IBNYR, dont le nombre est inconnu :

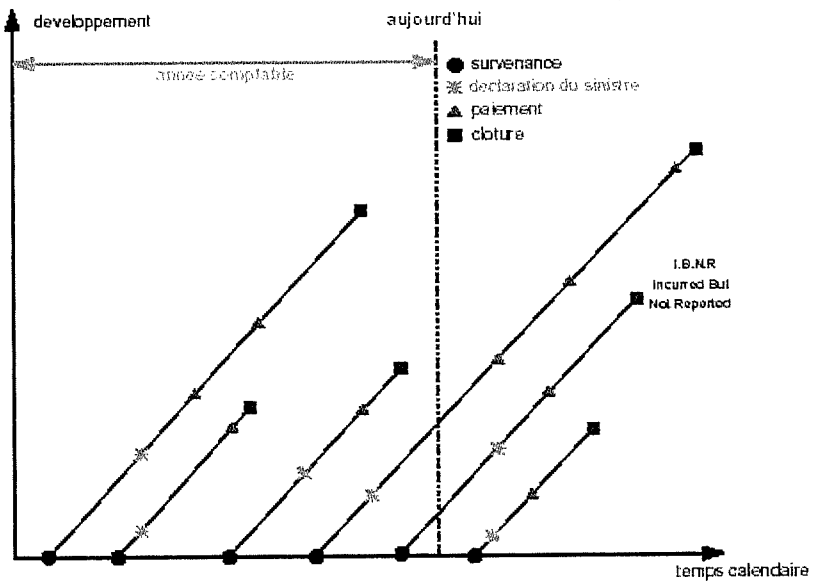


fig. 1.4 : Aspect dynamique de la gestion des sinistres
source : Arthur Charpentier

Mise en place du modèle

Nous cherchons une méthode permettant d'éviter les limites de modèles stochastiques liés aux triangles de développement. Nous allons pour cela revenir au sinistre individuel.

Avantages des modèles individuels

Ce type de méthode permet l'obtention d'intervalles de confiance.

De plus, ils permettent de séparer les provisions pour sinistres déclarés (IBNER) et les provisions pour sinistres survenus mais non déclarés (IBNYR). Dans le modèle présenté ici, nous n'étudions que les IBNER.

La réassurance non proportionnelle peut être étudiée par ce type de méthodes, ce qui n'était pas le cas avec la méthode de Chain-Ladder.

Enfin, les résultats obtenus sont additifs. Ceci signifie que l'estimation pour un portefeuille donné est la somme des estimations pour une partition de ce portefeuille.

Cependant, certaines méthodes permettent uniquement d'obtenir une estimation de l'ultime et non des cash-flows, car elles ne simulent que le montant total de chaque sinistre. Ce type de modèle ne nous intéresse pas ici. En effet, les objectifs des modèles de provisionnement sont multiples (tarification, valorisation...). Pour remplir ces critères, nous avons besoin de connaître la cadence des paiements. Aussi, nous voulons déterminer une méthode permettant de modéliser la dynamique des flux.

Les modèles classiques de provisionnement montrent quelquefois des limites. Nous cherchons ici à établir un nouveau modèle permettant de corriger ces défauts. Une modélisation des sinistres individuels nous a semblé être intéressante. Un tel modèle permet de prendre en compte l'ensemble du processus de sinistralité. Il s'agit là d'une approche très différente des méthodes de provisionnement classiques. Aussi, avant de procéder à la modélisation, il est intéressant d'étudier les données utilisées.

2. ETUDE D'UN PORTEFEUILLE RC PROFESSIONNELLE

Par soucis de confidentialité, toutes les statistiques liées au portefeuille étudié dans ce mémoire ont été changées d'échelle.

Afin de modéliser le processus de sinistralité, nous désirons savoir quel est le déroulement « type » de la vie d'un sinistre. Pour cela, nous cherchons à décrire sa manière d'évoluer, et quels pourraient être les facteurs explicatifs de cette évolution. Nous souhaitons ici pouvoir traiter les sinistres à la manière d'une chaîne de Markov, en déterminant des étapes caractéristiques dans leur vie et en trouvant les probabilités de passage associées.

Des données issues d'un portefeuille Responsabilité Civile professionnelle ont été utilisées. Ce sont des sinistres ayant une durée de vie longue. Il s'agit d'une branche complexe pour laquelle les méthodes classiques manquent de précision. C'est pourquoi il est particulièrement intéressant de revenir au sinistre individuel pour ce portefeuille.

2.1 Données utilisées

Nous disposons de données au 20 janvier 2003 sur 771 sinistres RC professionnelle survenus depuis 1997 (dont 159 ont engendré des paiements)

Pour chacun de ces sinistres, nous connaissons :

- La date de survenance
- La date de déclaration
- La date d'ouverture du dossier
- La date de clôture éventuelle du dossier

De plus, nous connaissons la date d'occurrence de chaque cash-flow (617 au total) avec des indications sur l'objet du règlement. Sont connus en particulier :

- La nature du flux (principal, frais, honoraires ou recours)
- Le type du règlement (partiel, définitif, ou complémentaire)
- Le type de bénéficiaire (expert, courtier, client, tiers ou autre)

Nous connaissons également les mouvements sur la provision dossier / dossier d'un sinistre (plus de 4000 mouvements) :

- Les dates de révision des estimations
- La variation des réserves dossier / dossier
- La variation des prévisions de recours

Un premier traitement des données a été effectué. La table contenait des flux négatifs, que ce soit du point de vue des paiements ou du point de vue des recours. Nous avons remarqué que tous les flux négatifs correspondaient à des annulations de paiements effectués quelques jours auparavant. Aussi, nous avons annulé tous les flux négatifs en supposant qu'il s'agissait d'erreurs de gestion.

2.2 Spécificité des données

Nous travaillons sur des sinistres ayant une durée de vie longue. Ils réclament souvent des avis d'experts ou des passages devant un tribunal. Il est donc beaucoup plus difficile pour ces sinistres d'en déterminer la charge définitive.

En effet, pour pouvoir étudier ce type de sinistres, nous avons besoin d'un historique suffisant. De plus, la reproductibilité du passé n'est pas évidente. Plus la durée de développement est longue, plus il y a un grand risque de changement, que ce soit au niveau de la législation ou au niveau de la politique de provisionnement de la société.

La sinistralité de ce type d'évènement est plus complexe que des sinistres automobiles ou MRH (multirisque habitation). La modélisation risque donc d'être elle aussi plus complexe. Des données comme celles de l'assurance automobile sont relativement stables et s'évaluent de manière satisfaisante par les méthodes classiques. Au contraire, le type de données que nous avons utilisé s'évalue beaucoup plus difficilement. C'est pourquoi utiliser un modèle de provisionnement individuel sur ce type de données est particulièrement intéressant.

2.3 Variables explicatives

Nous cherchons à savoir quels sont les facteurs pouvant avoir une influence sur le déroulement d'un sinistre. Pour l'instant, nous ne nous intéressons qu'aux flux des sinistres.

En dehors des informations « classiques » sur les sinistres, nous connaissons :

- La nature du flux (paiement en principal, honoraire, frais, ou recours)
- Le type du règlement (partiel ou définitif)
- Le type de bénéficiaire (expert, client, courtier, tiers ou autre)

Toutefois, parmi ces trois informations, seule la nature du flux permet de différentier les différents évènements. Par exemple, lorsque l'on observe un règlement de principal, dans 90% des cas, il s'agit d'un règlement partiel pour le type de bénéficiaire « autre ». Etant donné le nombre d'observations, les autres cas se retrouvent en nombre trop faible et risquent de ne pas donner de résultats significatifs. Aussi, dans un souci de simplification, seule la nature des flux va être exploitée.

Répartition des flux par nature

Voici la répartition des flux par nature :

REPARTITION DES FLUX PAR NATURE

	Pourcentage
Principal	42 %
Honoraire	20 %
Frais	29%
Recours	9%

Nous observons donc :

- La place prépondérante des paiements de principaux. Nous étudions ici des branches longues faisant intervenir de nombreux paiements en principal.

- Une proportion importante de frais et dans une moindre mesure, d'honoraires.
- Un nombre de recours non négligeable.

Montant des flux

Il peut être intéressant d'étudier le montant moyen et l'écart type des flux par nature :

MONTANT DES FLUX

	Moyenne	Ecart type
Principal	78 916 €	198 763 €
Honoraire	1 399 €	878 €
Frais	1 087 €	1 694 €
Recours	11 673 €	23 136 €

- Les différences de moyenne sont très marquées : les montants associés aux règlements en principal sont beaucoup plus élevés que les autres. A nouveau les frais et honoraires sont à peu près comparables. Les recours représentent des montants relativement élevés.
- Les écarts-types associés aux différents types de paiements sont élevés, sauf pour les honoraires où ils sont inférieurs à la moyenne. Aussi, si jusque-là frais et honoraires étaient relativement comparables (en nombre et en moyenne), on voit que les honoraires sont beaucoup plus stables que les frais.

De manière plus générale, en raison de la volatilité élevée, il pourra être intéressant d'utiliser d'autres facteurs, comme le délai de déclaration ou le nombre de paiements déjà effectués, pour modéliser de manière précise le sinistre.

2.4 Statistiques descriptives sur le déroulement des sinistres

Pour établir ces statistiques, nous avons travaillé sur les cash-flows des sinistres, afin de dégager une structure type pour le processus de développement d'un sinistre.

Au cours de cette étude, nous avons été amenés à regrouper les données en classes. Ce regroupement a été effectué de manière à étudier des classes de taille comparable. Une analyse plus complète de ces statistiques est donnée en annexe (partie 1).

On remarque en particulier que le nombre de paiements et l'année de développement ont une grande importance sur la nature des flux, et que ces influences semblent être relativement similaires. On peut noter par exemple que plus le nombre de paiements augmente / plus l'année de développement est grande, plus la proportion de paiement en principal diminue, alors qu'au contraire les frais augmentent. En revanche, l'influence du délai de déclaration et du délai écoulé depuis le flux précédent semble moins évidente.

Enfin, les montants des flux ne semblent pas être expliqués par des facteurs comme le délai de déclaration ou l'année de développement.

Au terme de cette étude, nous avons donc remarqué la place prépondérante de la nature des flux sur les montants associés. Seulement, nous n'avons pas été capables de déterminer une structure type dans le processus des paiements. Il ne semble pas y avoir d'ordre précis dans le développement des sinistres.

3. MODELE DE PROVISIONNEMENT INDIVIDUEL

Nous cherchons à établir un modèle de provisionnement utilisant des données détaillées. La méthode Chain-Ladder peut s'appliquer soit aux paiements, soit aux sinistres survenus (paiements et provision dossier par dossier). Nous avons choisi ici de proposer un modèle portant sur les paiements uniquement. De plus, nous étudions uniquement les sinistres déclarés. Les IBNYR ne sont donc pas analysés ici.

Nous avons trois types de processus à modéliser :

- le processus de date des flux
- les montants associés le cas échéant
- la date de clôture

Le modèle présenté ici est un modèle établi à partir de données expérimentales. Néanmoins, intuitivement, voici de quelle manière nous nous proposons d'aborder les processus à modéliser.

Pour modéliser les dates des flux, nous pouvons penser que ces flux ont lieu les uns après les autres à la manière d'un processus de Poisson. Cependant, faut-il considérer le premier paiement de la même manière que les suivants ? En effet, une fois que le premier paiement a eu lieu, il est possible que les paiements s'enchaînent de manière relativement rapide jusqu'à la clôture, un paiement en entraînant un autre.

Les montants des flux dépendent très fortement de la nature du flux. Ceci devra nécessairement être pris en compte dans la modélisation.

Les sinistres d'un portefeuille RC professionnelle ayant par nature un déroulement long, les sinistres des années de déclaration les plus anciennes ne sont pas nécessairement tous clos. C'est pourquoi nous risquons de n'observer que les sinistres ayant été clos rapidement. De plus, même si tous les sinistres de l'année de déclaration la plus ancienne sont clos, le nombre de sinistres d'une année de déclaration peut être trop faible pour permettre une modélisation correcte. L'analyse de survie permet de prendre en compte l'ensemble des sinistres, qu'ils soient clos ou non. Il est donc intéressant d'utiliser ce type de technique dans notre modèle.

3.1 Modélisation des dates de clôture

Nous ne disposons que de relativement peu de sinistres clos. Ces derniers sont des sinistres qui ont connu un traitement rapide, souvent sans recours et ne sont pas représentatifs de l'ensemble des sinistres.

Aussi, nous avons choisi d'utiliser l'analyse de survie pour notre modélisation, puisqu'elle nous permet de prendre en compte le grand nombre de sinistres qui ne sont pas encore clos.

Principe de l'analyse de la survie

L'analyse de la survie consiste à s'intéresser à la survenance au cours du temps d'un événement. L'événement considéré par les fondateurs de l'analyse de la survie est le décès, raison pour laquelle cette expression est utilisée. Il s'agit d'évaluer le délai de survenance du décès d'un individu. Cependant, le décès n'est pas le seul type d'événement analysable. Tout événement de nature binaire peut être utilisé comme critère de jugement : ici, l'évènement recherché est la clôture d'un sinistre.

Les données pour lesquelles le décès n'est pas survenu sont appelées données censurées. La spécificité de l'analyse de survie réside dans l'utilisation de ces données.

De manière générale, cela permet la description de la survie d'un groupe de sujets. Il est également possible de comparer la survie de deux ou plusieurs groupes de sujets ainsi que de rechercher des facteurs pronostiques de la survie, c'est-à-dire des facteurs susceptibles d'expliquer l'occurrence au cours du temps du décès.

Les données indispensables pour l'analyse de la survie

Quatre informations sont indispensables à cela :

- La date du début de l'observation, appelée date d'origine. Il s'agit pour nous de la date de déclaration.
- La date des dernières nouvelles. C'est soit la date de clôture du sinistre ou, pour les données censurées (sinistres encore ouverts), la date pour laquelle on dispose des dernières données relatives au sinistre sachant qu'il n'est pas clos.

- La date de fin d'observation, appelée date de point. Elle est commune à tous les sinistres inclus dans l'étude.
- La variable d'état. C'est l'état du sujet aux dernières nouvelles. Deux modalités sont possibles soit le sinistre est clos, soit il est ouvert.

On distingue deux situations :

- Le sinistre a été clos au cours du suivi, c'est-à-dire avant la date de point (ou date de fin de suivi). La durée de suivi est calculée entre la date d'origine et la date de clôture.
- La clôture n'est pas observée au cours du suivi. La durée de développement de ce sinistre est alors censurée. On sait seulement qu'elle est supérieure à la durée d'observation.

L'estimation de la survie : la méthode de Kaplan Meier

Cette méthode permet l'estimation de la probabilité de survie au cours de l'étude. Pour cela, on définit une fonction de survie $S(t)$. Le principe de l'estimation de Kaplan Meier repose sur une idée simple : être encore en vie après un instant t, c'est être en vie juste avant cet instant t et ne pas mourir à cet instant. La probabilité d'avoir survécu à un instant donné peut se calculer conditionnellement au fait d'être en vie juste avant cet instant.

On définit les variables suivantes :

- n_i est le nombre de sinistres ouverts au temps t_i^-.
- d_i le nombre de clôtures au temps t_i.

L'estimateur de Kaplan Meier s'écrit alors :

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

Un estimateur de la variance de l'estimateur de Kaplan Meier à un temps t fixé est donné par la formule de Greenwood :

$$\hat{V}\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{t_i<t} \frac{d_i}{n_i} = \left(\hat{S}(t)\right)^2 \sigma_S^2(t)$$

On peut supposer que l'estimateur de Kaplan Meier est approximativement de distribution normale pour obtenir une mesure d'incertitude. Pour un intervalle de confiance à 95%, on obtient le graphe suivant :

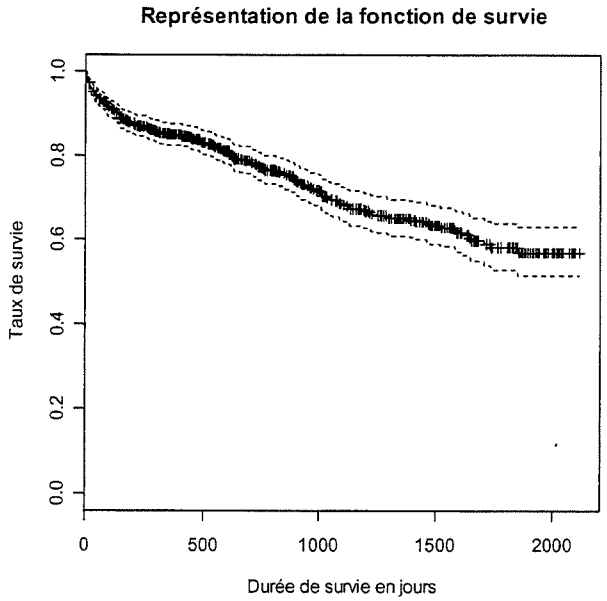


fig. 3.1 : Représentation de la fonction de survie

Modèle semi-paramétrique de Cox

Le modèle de Cox est un modèle de régression qui permet de modéliser l'effet de variables exogènes (covariables, facteurs, variables explicatives, ...) sur la distribution de la durée de vie. Contrairement à l'approche la plus classique utilisée dans la régression ordinaire, on ne modélise pas l'effet de ces variables sur la valeur de la variable aléatoire et durée de vie, mais plutôt sur la fonction de risque h.

Le modèle est donné par :

$$h(t|Z) = h_0(t) \cdot c(\beta' Z)$$

où Z est le vecteur des variables exogènes.

Nous avons utilisé diverses variables comme l'année de déclaration et le délai de déclaration. Mais aucune d'elles n'a donné de résultats significatifs.

Modèle paramétrique

Nous cherchons à ajuster la loi de manière paramétrique. Pour cela, nous allons utiliser une des lois classiques en analyse de survie : loi exponentielle, de Weibull ou loi lognormale. Ici, nous avons effectué un ajustement avec une loi lognormale. On obtient le résultat suivant :

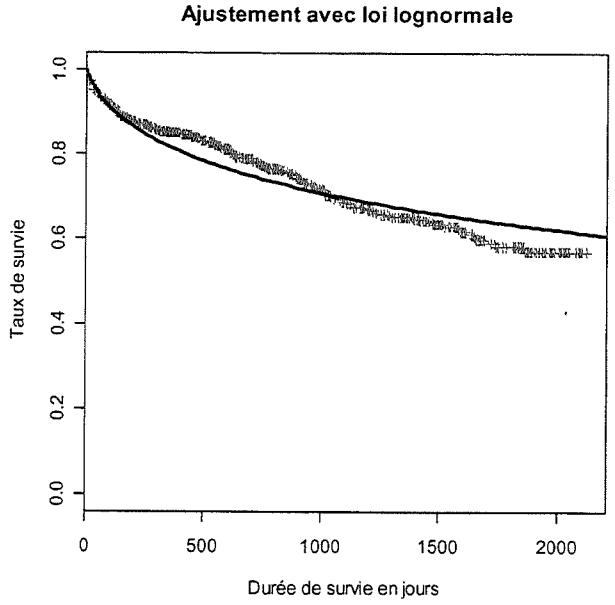


fig. 3.2 : Ajustement paramétrique de la fonction de survie par une loi lognormale

Le test du log-rank est un test qui compare la distribution de deux fonctions de survie. Ici, il s'agit de comparer la fonction de survie empirique et la fonction de survie théorique. Sous l'hypothèse nulle, ces deux fonctions sont égales en tout point. Le paramètre du test est basé sur la distance entre les deux courbes que l'on compare à un Chi deux à un degré de liberté.

Avec une modélisation par une loi normale, on ne rejette pas l'hypothèse d'égalité des distributions (tableau 2.1 en annexe). On peut en effet vérifier par le test du log-rank que le modèle est bien ajusté avec une loi lognormale de moyenne 8.5 et d'écart type 2.9.

Néanmoins, la forme générale de la courbe de survie fait penser à un mélange de lois. Lorsque la durée de survie des inférieure à 500 jours, on observe une fonction de survie « classique » (de type lognormal ou Weibull). Par la suite, il semble que la fonction de survie devienne linéaire et donc qu'il s'agisse d'une loi uniforme.

Pour les 500 premiers jours, nous avons donc ajusté la fonction de survie de manière paramétrique. Pour les délais de développement plus longs que 500 jours, nous avons choisi d'utiliser notre connaissance a priori de la branche. Nous savons en effet qu'au bout de 13 ans, environ 95% des sinistres seront clos. Ceci nous a permis de déterminer les paramètres de la loi uniforme. Graphiquement, on obtient le résultat suivant :

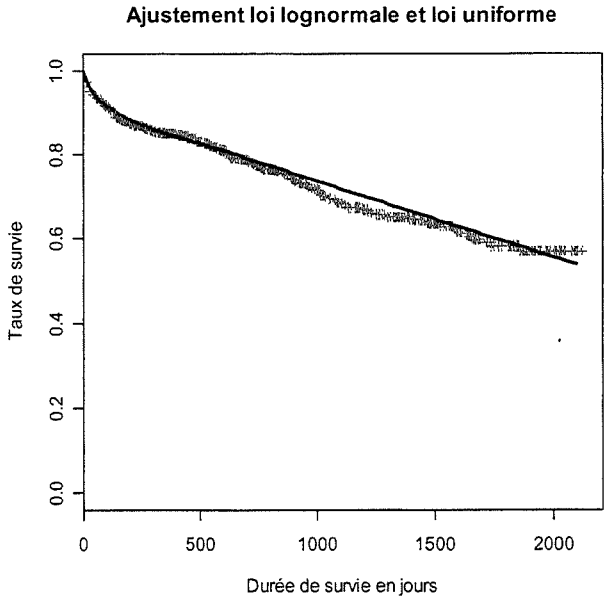


fig. 3.3 : Amélioration de l'ajustement paramétrique de la fonction de survie

Amélioration de la modélisation des dates de clôture

Le choix de la date de clôture est fortement lié à la politique de gestion du dossier. Aussi, certains gestionnaires pourront choisir de garder un dossier ouvert même si les provisions pour sinistre à payer sont nulles pour des raisons de prudence. Dans ce cas, le dossier ne sera clos que plus tard et un grand temps se sera écoulé entre le dernier paiement et la clôture. Ceci rend alors difficile la modélisation des dates de paiements.

Il y a deux approches possibles pour identifier les sinistres qui n'engendreront plus de paiements :

- Soit on regarde le temps écoulé depuis le dernier mouvement (paiement ou estimation).

Cette approche est intéressante. Seulement, nous manquons de recul pour déterminer à partir de quel seuil un sinistre qui n'a plus connu de mouvement depuis longtemps n'engendre plus de paiement. Nous étudions une branche longue et il n'est pas étonnant de ne pas avoir eu de nouvelle information depuis un an sur un sinistre. Il existe une cinquantaine de sinistres n'ayant pas connu de mouvement depuis plus d'un an. Mais certains montants de provision sont non négligeables et il est difficile d'affirmer qu'il n'y aura plus de paiement.

- Soit on regarde la provision pour sinistre à payer

Si celle-ci est trop faible, on peut supposer qu'il s'agit d'une mesure prudentielle, et qu'il n'y aura probablement plus de paiement pour ce sinistre. Mais il n'y a que trois sinistres qui laissent à penser qu'une provision a été laissée de manière prudentielle.

Il ne semble donc pas qu'il y ait des sinistres ouverts pour des raisons prudentielles dans notre portefeuille. Aussi, nous n'avons pas procédé à la clôture automatique de certains dossiers avant d'effectuer nos simulations.

3.2 Modélisation des dates des flux

Données utilisées

Nous cherchons à étudier le temps écoulé séparant la déclaration d'un sinistre et la réalisation d'un flux (paiement, honoraires,...). Tous les flux sont donc étudiés ensemble.

Pour nous faire une première idée, nous allons utiliser un modèle GLM sur toutes les données. Il s'agit d'expliquer les temps où un paiement est effectué par différents paramètres.

On utilise comme paramètres :

- le montant déjà payé
- la PSAP
- le nombre de paiements déjà effectués
- le délai de déclaration

- le temps écoulé entre la déclaration et le dernier changement

Les modèles linéaires généralisés (GLM)

Dans le modèle linéaire « classique », la valeur d'une variable expliquée est supposée suivre une loi normale dont la moyenne est une combinaison linéaire des variables explicatives :

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$

où :

$$\mu_i = x_i' \beta = \alpha + \sum_{j=1}^{p} \beta_j x_i^j$$

Le modèle linéaire généralisé permet de définir une classe de modèles plus vaste.

- Du point de vue linéaire, la moyenne est une fonction connue d'une combinaison linéaire des variables explicatives :

$$\mu_i = g(x_i' \beta)$$

- Du point de vue de la distribution de probabilité des observations, les distributions admissibles ne sont pas uniquement la loi gaussienne mais toutes les distributions de la famille exponentielle, qui comprend, outre la loi gaussienne, des distributions très utilisées comme la loi Binomiale, la loi de Poisson, la loi Gamma...

Une loi de probabilité appartient à la famille exponentielle si sa densité peut s'écrire sous la forme suivante :

$$f(y, \mu, \phi) = \exp\left[\frac{1}{\phi}(y \cdot a(\mu) - b(\mu)) + c(y, \phi)\right]$$

où les paramètres a, b et c sont connus.

D'une manière générale, la variance d'une distribution de probabilité de moyenne variable n'est pas constante mais fonction de la moyenne. Pour la famille exponentielle, on note :

$$\mathrm{Var}(\mathrm{Y_i}) = \phi \cdot V(\mathrm{Y_i})$$

où ϕ est le paramètre de dispersion et V la fonction de variance.

Aussi, par l'utilisation du modèle linéaire généralisé, on cherche à expliquer la moyenne comme une fonction simple des prédicteurs (avec le moins de prédicteurs possibles) qui restitue le mieux possible la valeur de la réponse. Pour cela, nous allons devoir faire le choix d'une distribution de probabilité de la réponse et d'une fonction de lien (définie par g^{-1}).

Le modèle est alors complètement défini et on peut écrire la densité de probabilité de l'observation $f(y_i, g(x_i'\beta), \phi)$. Les deux paramètres inconnus β et ϕ sont obtenus par le maximum de vraisemblance.

Les logiciels statistiques permettent l'évaluation de ces paramètres par procédure itérative.

Utilisation des GLM

Nous cherchons à étudier le temps T entre la déclaration et la survenance d'un flux. On peut supposer dans un premier temps qu'il s'agit d'un modèle de type Log-Poisson. Ce type de modèle est souvent utilisé pour les triangles lors de l'utilisation des GLM. On suppose donc que :

$$T \sim \mathrm{P}(\mathrm{E}(\mathrm{T}))$$

où :

$$\mathrm{E}(\mathrm{T}) = \exp(X'\beta) = \exp\left(\alpha + \sum_{i=1}^{p} \beta_i X_i\right)$$

Les résultats de ce GLM sont donnés en annexe (tableau 2.2). Le test de nullité d'un paramètre nous indique que chaque paramètre est significativement non nul.

Pour mesurer l'adéquation du modèle aux observations, il est intéressant d'étudier la déviance. La déviance compare les valeurs de la réponse estimées par le modèle aux valeurs effectivement observées. Elle est d'autant plus petite que les estimations sont proches des valeurs observées. La déviance est de degré de liberté n-k (où n est le nombre d'observations et k est le nombre de paramètres utilisés).

De manière plus formelle, la déviance a pour expression :

$$Dev = 2 \cdot \phi \left[L(y, y, \phi) - L(y, \hat{\beta}, \phi) \right]$$

On appelle modèle saturé le modèle pour lequel les moyennes μ_i sont estimées par les observations y_i. La déviance est donc la différence des maxima de vraisemblance du modèle GLM et du modèle saturé. Elle représente donc la perte de vraisemblance des observations en imposant aux moyennes μ_i d'être de la forme $g(x_i'\beta)$.

Dans notre modèle, la déviance est de 90 000 contre 195 000 sans l'introduction de variables explicatives.

Ces résultats ne sont pas très bons. On considère en effet que l'adéquation au modèle est bonne si la déviance normée (i.e. le rapport entre la déviance et le paramètre de dispersion) divisée par son degré de liberté est au plus de l'ordre de 1. Or ici il est de l'ordre de 100.

En choisissant une loi de Poisson, nous avons implicitement imposé un paramètre de dispersion ϕ égal à 1. Or nous pouvons travailler avec une loi de Poisson avec surdispersion. Cette dernière diffère d'une loi de Poisson classique par l'introduction d'un paramètre de dispersion non fixé à 1. Ceci permet d'avoir un rapport entre la déviance normée et le degré de liberté de l'ordre de 1. En effet :

- Il n'y avait en fait pas de raison pour que la moyenne soit égale à la variance.
- Le fait de prendre un paramètre de dispersion adapté modifie la matrice de variance – covariance des coefficients et nous permet d'obtenir les intervalles de confiance réels.

Nous avons donc ajusté un tel modèle à nos données. Nous pouvons alors observer que les paramètres ne sont plus tous significativement non nuls. Le nouveau modèle est formé du délai de déclaration, du temps écoulé entre la déclaration et le dernier flux, et du nombre de paiements déjà effectués (cf. tableau 2.3 en annexe).

Une loi Gamma peut également être utilisée pour faire la régression. Les résultats obtenus sont alors relativement similaires. Ce résultat correspond à nos attentes puisque pour simuler une loi de Poisson avec surdispersion, England utilise une loi Gamma en faisant correspondre moyenne et variance.

Vers les modèles additifs généralisés (GAM)

Dans le modèle précédent, la déviance n'est que faiblement réduite par rapport à la déviance définie sans variables explicatives. Ceci n'est pas très étonnant si l'on se souvient que les variables explicatives étaient de type très différent et qu'à la vue des premières statistiques, l'effet des variables explicatives ne semblait pas découler d'une transformation simple. Il semblait plutôt y avoir des types de comportement par classe. Par exemple, un type de comportement pour les sinistres déclarés en moins d'un mois, en moins d'un an…

Nous avons créé des indicatrices correspondant aux différentes classes utilisées pour les statistiques descriptives. Puis nous avons ensuite effectué une sélection par une analyse de type anova afin de déterminer si l'ajout ordonné de chaque variable est significatif.

Avec seulement 6 indicatrices, nous obtenons un résultat presque similaire à ce que nous avions au paragraphe précédent (déviance de 93 000). Ceci nous a amené à penser que l'effet des variables explicatives n'était pas nécessairement linéaire. Aussi, nous nous sommes intéressés aux modèles additifs généralisés (GAM – Generalized Additive Models) : en effet ces derniers permettent des transformations plus complexes des variables explicatives.

Utilisation des GAM

Il est plus intéressant dans ce contexte de travailler avec des GAM. Les LM et les GLM se basent sur des courbes de réponse paramétriques, essentiellement de premier et de deuxième ordre, limitant ainsi la forme de la réponse à une droite ou à une parabole. Les Modèles Additifs Généralisés (GAM) sont une extension non paramétrique des GLM qui introduisent des courbes de réponse lissées à partir des données d'observation.

Pratiquement, la forme linéaire $X'\beta$ est ici remplacée par une transformée :

$$\alpha + \sum_{i=1}^{p} f_i(X_i)$$

Ceci permet de lisser les facteurs utilisés.

On utilise à nouveau une distribution de Poisson avec surdispersion et un lien de type logarithmique. Trois variables sont alors significativement non nulles : le délai de déclaration, le

temps écoulé entre la déclaration et le dernier flux, et le nombre de paiement déjà effectués. On obtient avec ces derniers un R2 ajusté de 0.67 (cf. tableau 2.4 en annexe). On peut légèrement améliorer ce résultat en mettant le délai de déclaration et le temps écoulé entre la déclaration et le dernier flux dans la même fonction de lissage.

Séparer les temps en fonction de la nature du flux aurait pu nous permettre d'améliorer les résultats. En fait cela divise le nombre d'observations et les résultats ne sont pas meilleurs.

Séparation du premier flux et des suivants

La fonction de lissage du nombre de paiements nous montre qu'il y a un comportement différent selon qu'il s'agit du premier paiement ou des paiements suivants (cf. graphique 2.4 en annexe). Le premier flux de chaque sinistre a un comportement différent des autres flux. Les facteurs dont nous disposons ne nous permettent pas d'expliquer le premier flux (on obtient avec eux un R2 très petit alors qu'il est correct pour les flux suivants).

Si l'on regarde plus attentivement les données, ce n'est pas très étonnant : le premier paiement met souvent près de deux ans an à arriver, alors que les paiements suivants se renouvellent plus fréquemment, un paiement étant souvent associé à un autre.

Nous étudions des sinistres ayant une durée de vie longue. Souvent, un premier paiement a nécessité au préalable différents avis d'experts et des passages devant le tribunal. Aussi, la date du premier paiement est a priori liée aux mouvements sur les réserves.

Modélisation sans facteur explicatif

L'étude par les GLM et les GAM nous a donc montré qu'il fallait distinguer le premier flux des autres. Une fois cette distinction faite, on peut se demander s'il n'est pas intéressant de travailler sur les temps incrémentaux (i.e. le délai entre deux flux).

Nous avons choisi de modéliser le temps entre la déclaration et le premier paiement et ces temps incrémentaux par une loi Gamma.

Graphiques Quantile - Quantile

En ajustant les lois en moyenne et en variance, on en déduit les graphiques quantiles - quantiles suivants :

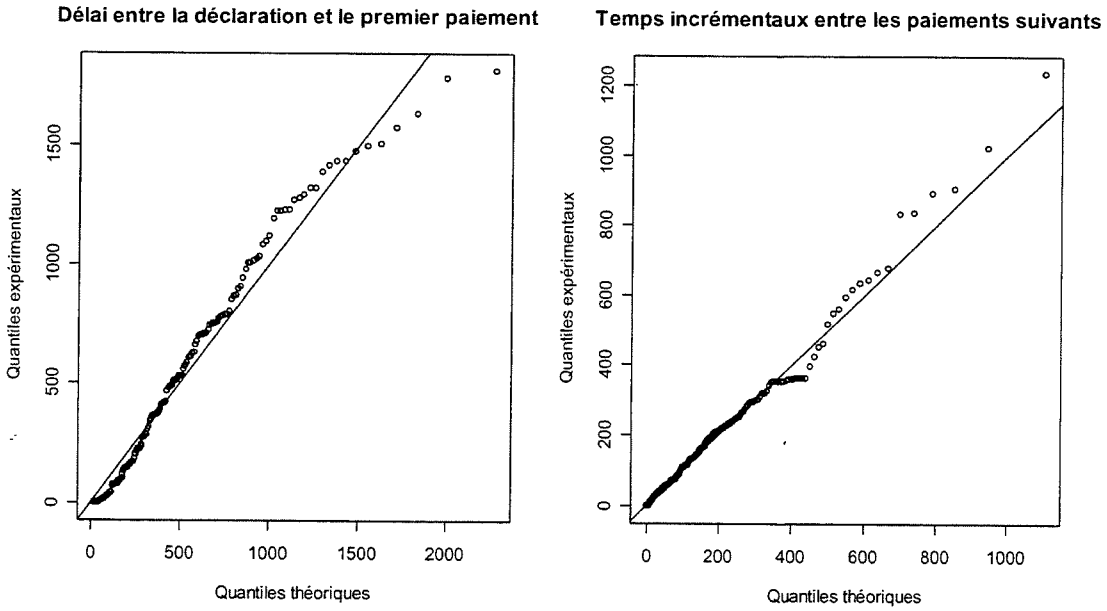


fig. 3.4 : Graphiques quantiles - quantiles

L'ajustement ne semble pas parfait pour le délai entre la déclaration et le premier paiement. Notre modélisation donne des valeurs supérieures aux valeurs expérimentales pour les faibles valeurs, et inférieures aux valeurs expérimentales pour les valeurs plus importantes.

Il est cependant meilleur pour les temps incrémentaux entre les paiements suivants. Du point de vue expérimental, on note néanmoins qu'aux alentours d'un an, on a une forte probabilité d'avoir un nouveau paiement. Puis plus le délai devient important, plus la modélisation donne des valeurs inférieures aux valeurs expérimentales.

Test d'ajustement

Pour utiles qu'elles soient, les méthodes graphiques ne constituent pas une réponse mathématique au problème de l'ajustement. Pour quantifier l'éloignement de la distribution empirique par rapport à une loi théorique, on utilise des distances entre lois de probabilités. Les deux distances les plus fréquemment utilisées sont la distance du chi-deux et la distance de Kolmogorov-Smirnov. La distance du chi-deux concerne uniquement les lois discrètes, mais on peut l'utiliser aussi pour des échantillons continus regroupés en classes.

Dans le cas présent, puisque le test du chi-deux est moins puissant que le test de Kolmogorov-Smirnov, nous préférons étudier ce dernier.

Le test de Kolmogorov-Smirnov calcule la distance de la norme uniforme entre fonctions de répartitions.

En pratique, la fonction de répartition empirique étant constante entre deux valeurs successives des statistiques d'ordre, il suffira pour calculer la distance de Kolmogorov-Smirnov, d'évaluer la différence entre les deux fonctions de répartition à chaque palier de la fonction de répartition empirique :

$$D_{KS}\left(F,\hat{F}\right)=\max\left(\left|F\left(x_{(i)}\right)-\frac{i}{n}\right|,\left|F\left(x_{(i)}\right)-\frac{i-1}{n}\right|\right)$$

où F est la fonction de répartition paramétrique, $\hat{F}$ la fonction de répartition empirique et où les $x_{(i)}$ sont les statistiques d'ordre de l'échantillon.

Le test de Kolmogorov-Smirnov ne rejette pas l'hypothèse d'égalité des distributions pour la modélisation du premier paiement, mais rejette celle faite pour les paiements suivants.

Limite de la modélisation par les temps incrémentaux.

Ce modèle montre rapidement ses limites. En effet, si on s'intéresse au temps écoulé entre le dernier flux de chaque sinistre et la date de la dernière observation (soit janvier 2003), on remarque qu'en moyenne ce délai est supérieur à la moyenne de la loi modélisant le temps entre deux paiements. En réalité, dans cette modélisation, nous sous-estimons la possibilité d'avoir des délais importants. Il s'agit là d'une caractéristique de la base étudiée. Les sinistres courts sont surreprésentés en pourcentage du fait de l'historique court.

Etude du premier paiement à l'aide de facteurs explicatifs

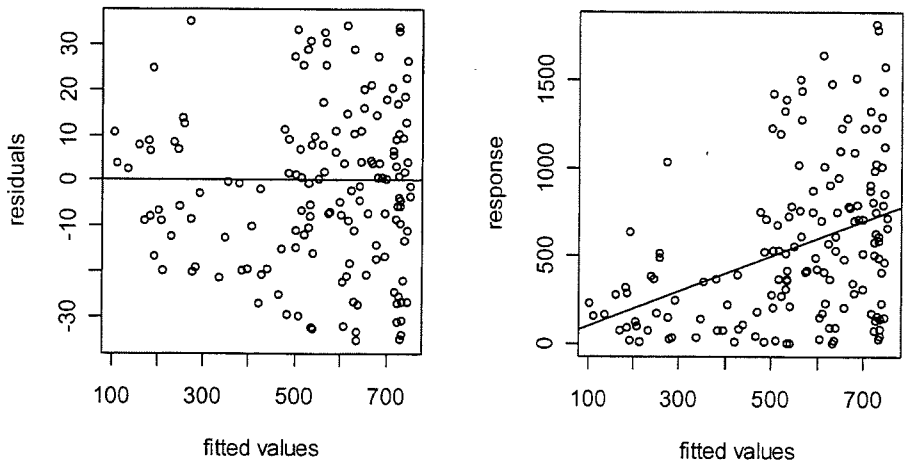
Modélisation par un GLM

Nous cherchons ici à améliorer la modélisation par une loi Gamma en introduisant des facteurs explicatifs.

Pour étudier le premier paiement, nous n'avons que peu de paramètres exploitables. Le délai de déclaration en est un. A cela, nous avons ajouté la première estimation (en montant) de la charge

du sinistre. Nous avons également essayé d'ajouter le délai d'ouverture du dossier, mais cela ne donnait pas de résultats intéressants. En effet, la plupart du temps, les sinistres qui ont été déclarés durant une même période sont ouverts le même jour. Et donc la vitesse d'ouverture du dossier ne préjuge pas du délai avant le premier paiement.

Les résultats les plus concluants sont obtenus avec la loi de Poisson (avec surdispersion) et son lien canonique (lien logarithmique). Mais seul le délai de déclaration est significativement non nul (cf. tableau 2.4 en annexe). Graphiquement, nous obtenons le résultat suivant :



Graph. 3.5 : GLM pour une loi de Poisson avec surdispersion du premier paiement

Pour essayer de voir quelle était l'influence des paramètres, nous avons travaillé sur les modèles additifs. La première estimation de la charge du sinistre devient alors un facteur significativement non nul. La fonction de lissage associée est en forme de parabole. Ce résultat est intéressant : lorsque le montant estimé du sinistre est petit, l'assureur va attendre quelque peu, certainement pour payer par la suite la totalité du sinistre d'un coup. Puis, plus le montant augmente, plus il risque d'y avoir des frais d'experts ou des petits paiements. Ces derniers peuvent arriver rapidement. Et pour les quelques sinistres exceptionnels, l'assureur devra certainement attendre une décision de justice avant de commencer à payer. On peut supposer un lien de type quadratique et revenir sur les GLM en utilisant la variable explicative et son carré. Néanmoins, ceci n'améliore pas significativement les résultats. C'est pourquoi nous préférons garder la fonction de lissage du GLM avec le délai de déclaration pour seul facteur explicatif.

Nouvel ajustement

La modélisation présentée au paragraphe précédent n'est toujours pas très bonne. En particulier, entre 300 et 500 jours, on observe graphiquement une surévaluation du temps de premier paiement. De plus, le premier paiement peut arriver très longtemps après la déclaration. C'est pourquoi, pour ne pas surreprésenter les courts délai de déclaration, nous avons choisi de n'étudier que les sinistres les plus anciens (1997), et d'ajuster une loi gamma avec la méthode des moments. La moyenne observée est alors supérieure à la moyenne prédite dans le paragraphe précédent.

Etude des paiements suivants à l'aide de facteurs explicatifs

Modélisation du temps entre la déclaration et un flux

L'étude des paiements dans leur ensemble nous a montré que nous obtenions un bon ajustement en analysant le temps séparant la date de survenance et la déclaration d'un flux avec un modèle de type GAM.

En effet, si l'on ne prend que les sinistres ayant déjà engendré un paiement, on obtient un R2 de 0.89 (cf. tableau 2.6 en annexe). Et pour obtenir un tel résultat, nous n'avons besoin que de deux paramètres :

- le délai de déclaration
- le temps écoulé entre la déclaration et le dernier flux

Le nombre de paiements déjà effectués n'importe plus. Cela nous montre que le nombre de flux ne permet de distinguer qu'entre les sinistres ayant engendré des paiements ou non.

La représentation de la fonction de lissage est la suivante :

Représentation de la fonction de lissage

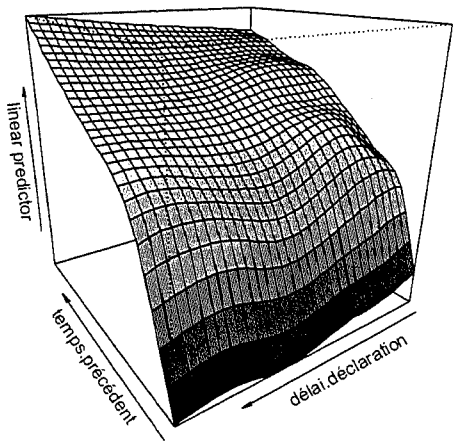


fig. 3.6 : Fonction de lissage du GAM

Si on enlève le délai de déclaration, les résultats sont à peine moins bons. C'est pourquoi nous préférons ne garder que le temps entre la déclaration et le dernier paiement. La forme de la fonction de lissage est alors logarithmique. Il peut alors être intéressant d'utiliser un GLM, non plus avec une fonction lien log mais une fonction lien identité. Cette fonction lien n'avait pas été utilisée précédemment car l'algorithme ne convergeait pas si l'on utilisait tous les sinistres. Si l'on n'étudie que les sinistres ayant engendré un paiement, l'algorithme converge, et les résultats restent bons en ne faisant intervenir qu'un seul paramètre : le temps entre la déclaration et le dernier flux.

On passe alors d'une déviance pour le modèle nul de 125 000 à une déviance de 18 000 (cf. tableau 2.7 en annexe).

En d'autres termes, le modèle nous apparaissant le plus adapté pour représenter la date du prochain flux (tps) est une loi de Poisson avec surdispersion dont la moyenne est une fonction linéaire du temps écoulé entre la déclaration et le dernier flux ($tepr$). On a donc en espérance :

$$\mathrm{E}(tps) = a \cdot \mathrm{E}(tepr) + b$$

L'intervalle de confiance à 95% du coefficient associé à la variable explicative est $[0.964, 1.051]$. La valeur 1 est donc comprise dans cet intervalle. On peut ainsi représenter le temps de paiement (au-delà du premier paiement) par une loi de Poisson avec surdispersion dont

la moyenne est le temps écoulé entre la déclaration et le dernier flux majoré d'une constante. Finalement, nous n'expliquons que ce que nous aurions pu dire avec un modèle simple : les temps incrémentaux sont constants en espérance. Cependant, la variance augmente lorsque le temps depuis la déclaration augmente.

Mais graphiquement, on s'aperçoit que ce modèle ne s'ajuste pas suffisamment bien à nos données (cf. graphique 2.7 en annexe). Le modèle théorique donne des valeurs légèrement supérieures aux valeurs expérimentales lorsque le délai entre deux paiements est faible, ce qui masque les délais importants.

De plus, lorsque l'on procède aux premières simulations, on remarque que trop de paiements s'enchaînent. Or, sur les données dont nous disposons, nous remarquons une première vague de paiements, puis une pause, et éventuellement, plus tard, une reprise des paiements. En fait, ceci s'interprète de manière assez simple. Une fois le sinistre déclaré à l'assureur, il met un certain temps à être évalué. Puis une première vague de paiements s'enchaînent. Ensuite, soit le sinistre est clos, soit il y a un jugement qui va devoir être donné. Avant qu'un jugement ne se fasse, les temps d'attente sont relativement longs. Et durant cette période, aucun paiement n'est effectué. Une fois que le jugement est rendu, une nouvelle série de paiements peut se faire.

Deux états différents

Nous avons voulu séparer les sinistres en deux états: « court » ou « long ». Nous avons remarqué que pour l'ajustement avec les temps incrémentaux, la modélisation avec une loi gamma se faisait bien pour les petits temps incrémentaux, mais qu'il était moins bon au-delà d'un an. C'est pourquoi nous avons choisi la limite d'un an pour différencier les deux états.

Pour savoir si à un moment donné, nous sommes dans l'un ou l'autre des deux états, nous avons utilisé le modèle logit.

Dans un tel modèle, on suppose que la probabilité d'être dans l'état e, conditionnellement aux variables explicatives vérifie :

$$P(e|x_i) = F(x_i b) = \frac{1}{1+\exp(-x_i b)}$$

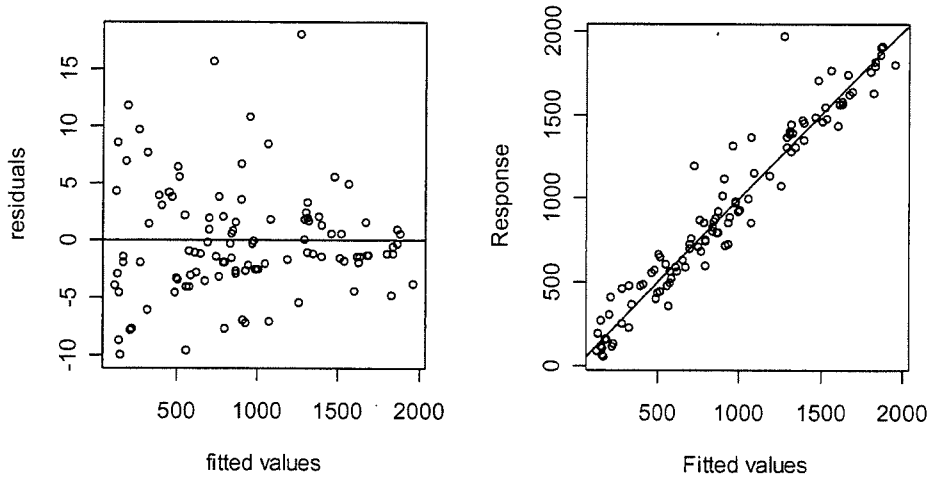
Nous avons utilisé les mêmes variables explicatives que précédemment (comme le délai de déclaration, le temps écoulé entre la déclaration et le dernier flux...). Seulement, aucune de ces

variables ne s'est avérée être significativement non nulle. Nous n'avons donc pas pris en compte les variables explicatives.

Nous modélisons donc le processus d'état par une loi de Bernoulli. L'estimateur du maximum de vraisemblance associé à la probabilité de se trouver dans un état « long » est $\hat{p} = 0.09$.

Nouvelle modélisation

Avec cette nouvelle indicatrice, nous obtenons des résultats beaucoup plus satisfaisants (cf. tableau 2.8 en annexe). Pour cela, nous avons eu besoin du temps écoulé entre la déclaration et le dernier flux, et de l'indicatrice permettant d'identifier les deux états. Graphiquement, nous obtenons le résultat suivant :



Graph. 3.7 : GLM pour une loi de Poisson avec surdispersion sur le délai entre la déclaration et un flux (hors premier paiement) avec indicatrice

Sur nos données, au moment de la dernière date d'observation, on remarque que seul un sinistre serait passé deux fois dans l'état « long » et 17 sont passés une fois par cet état Pourtant, au moment de l'étude, de nombreux sinistres n'ont pas engendré de paiement depuis plus de un an. Grâce à cette modélisation, nous prenons en compte que ces sinistres sont « en attente » et qu'ils risquent de ne pas engendrer de paiements tout de suite.

Nous avons donc établi un modèle portant sur les dates des flux. Cette étape est importante dans l'établissement de notre modèle. Ceci nous permet de déterminer une cadence de paiement.

Pour la modélisation, nous utilisons des lois Gamma. Le premier paiement est distingué des autres. En effet, le temps entre la déclaration et le premier paiement est souvent plus long que le temps entre deux paiements successifs. Les paiements suivants sont fonction de deux facteurs explicatifs. Il dépendent du temps entre la déclaration et le précédent paiement, ainsi que d'une variable d'état permettant de distinguer les sinistres qui risque d'engendrer de nouveau paiements rapidement et les autres.

3.3 Modélisation des flux

Pour la modélisation des flux, les premières statistiques effectuées en partie 2 nous ont montré qu'il était indispensable d'analyser séparément les flux par nature. Il y a alors deux phénomènes à modéliser : le type de flux et le montant du flux.

Modélisation du type de flux

Pour modéliser le type de flux, nous avons choisi d'utiliser les statistiques descriptives de la partie 2. Deux graphiques étaient particulièrement intéressants en ce qui concerne la fréquence des actes de gestion. Celui sur la dépendance au nombre de paiements déjà effectués et celui sur la dépendance à l'année de développement (cf. figure 1.1 et 1.2 en annexe). Nous avons ici choisi d'utiliser la dépendance à l'année de développement.

Modélisation des flux par les modèles linéaires

Pour modéliser les montants, notre première idée a été d'utiliser à nouveau des GLM et des GAM.

On peut utiliser les paramètres suivants :

- le montant déjà payé
- la PSAP
- le délai de déclaration
- le temps écoulé entre la déclaration et le dernier flux
- le temps écoulé depuis le dernier flux

- le nombre de paiements déjà effectués.

Le paiement en principal dépend de la PSAP, les frais de la PSAP et du nombre de paiements, les recours du montant déjà payé et les honoraires n'ont pas de facteur explicatif.

Cependant, graphiquement, les GLM ne donnent pas des résultats satisfaisants. De plus, l'utilisation de la PSAP comme facteur explicatif est délicate car dans ce cas, il faudrait également modéliser son évolution.

Nous avons ensuite travaillé avec des GAM. Lorsque l'on regarde la forme des fonctions de lissage, celles-ci semblent difficiles à justifier. Nous observons des ondulations difficilement explicables, surtout pour les grands montants. En réalité, l'ajustement ne se fait que pour les valeurs extrêmes. C'est pourquoi nous avons préféré ne pas utiliser de facteurs explicatifs.

Modélisation paramétrique des montants des flux

L'utilisation des modèles linéaires généralisés ne nous semble donc pas être intéressante. C'est pourquoi nous avons choisi de décrire les montants des flux par une modélisation paramétrique.

Représentation des fonctions de répartition empiriques

Nous avons séparé les paiements en fonction de leur nature. A partir des données dont nous disposons, nous obtenons les fonctions de répartition suivantes :

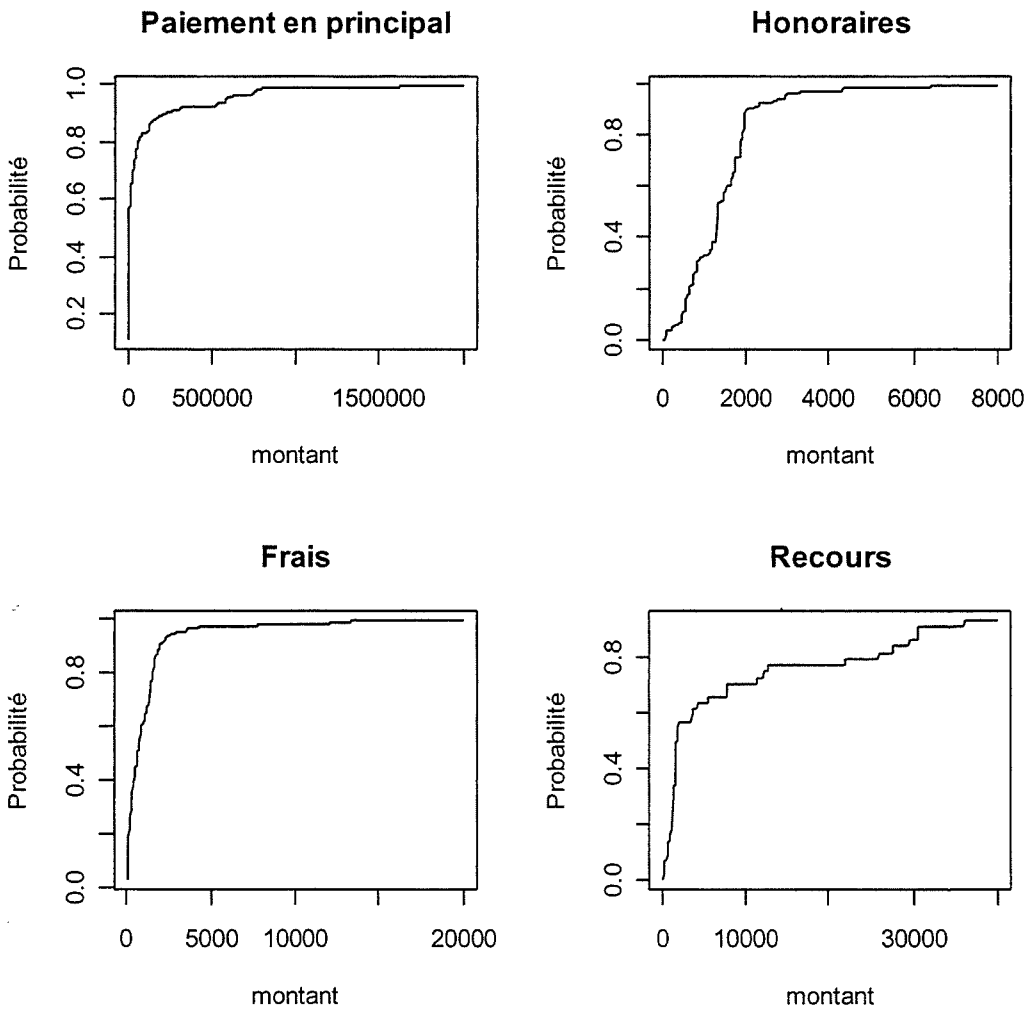


fig. 3.8 : Représentation des fonctions de répartition empiriques

Nous allons devoir ajuster nos lois aux valeurs extrêmes. Les sinistres Responsabilité civile présentent souvent une distribution à queue épaisse, c'est-à-dire avec une forte proportion de sinistres de coût élevé.

Etude des valeurs extrêmes

Le choix des seuils pour qualifier un montant d'extrême a été évalué de manière visuelle à partir des fonctions de densité empiriques des flux. Pour un paiement en principal, nous avons mis de côté les montants supérieurs à 450 000 €. Pour les honoraires et les recours, les paiements supérieurs à 2 000 € et pour les frais les paiements supérieurs à 4 000€.

Pour sélectionner notre modèle, nous avons à nouveau utilisé la technique des graphiques quantile - quantile avec les lois exponentielles, Weibull et Pareto sur les valeurs extrêmes de chaque type de flux.

On note $Q(p)$ la valeur de la fonction quantile prise au point p.

On peut alors montrer les résultats suivants :

- Pour une loi exponentielle, $(-\ln(1-p), Q(p))$ est linéaire.
- Pour une loi de Weibull, $(\ln(-\ln(1-p)), \ln(Q(p)))$ est linéaire.
- Pour une loi de Pareto, $(-\ln(1-p), \ln(Q(p)))$ est linéaire.

Les distributions que nous allons trouver sont des distributions conditionnelles (elles sont définies au-delà d'un certain montant). L'ordonnée à l'origine nous permet de déterminer (sauf pour la loi de Weibull) le paramètre a de la distribution conditionnelle $X|X \geq a$. Pour chaque type de paiement, nous avons sélectionné la loi permettant d'obtenir le meilleur coefficient de détermination.

Cependant, ce type de méthode n'est pas suffisant. Il faudra ensuite estimer les paramètres des lois par des techniques standard (maximum de vraisemblance, méthode des moments), puis tester la compatibilité du modèle avec les observations.

Pour les paiements en principal et pour les honoraires, nous avons sélectionné la loi de Pareto, alors que pour les frais et les recours, nous avons choisi une loi de Weibull. Graphiquement, voici les résultats obtenus :

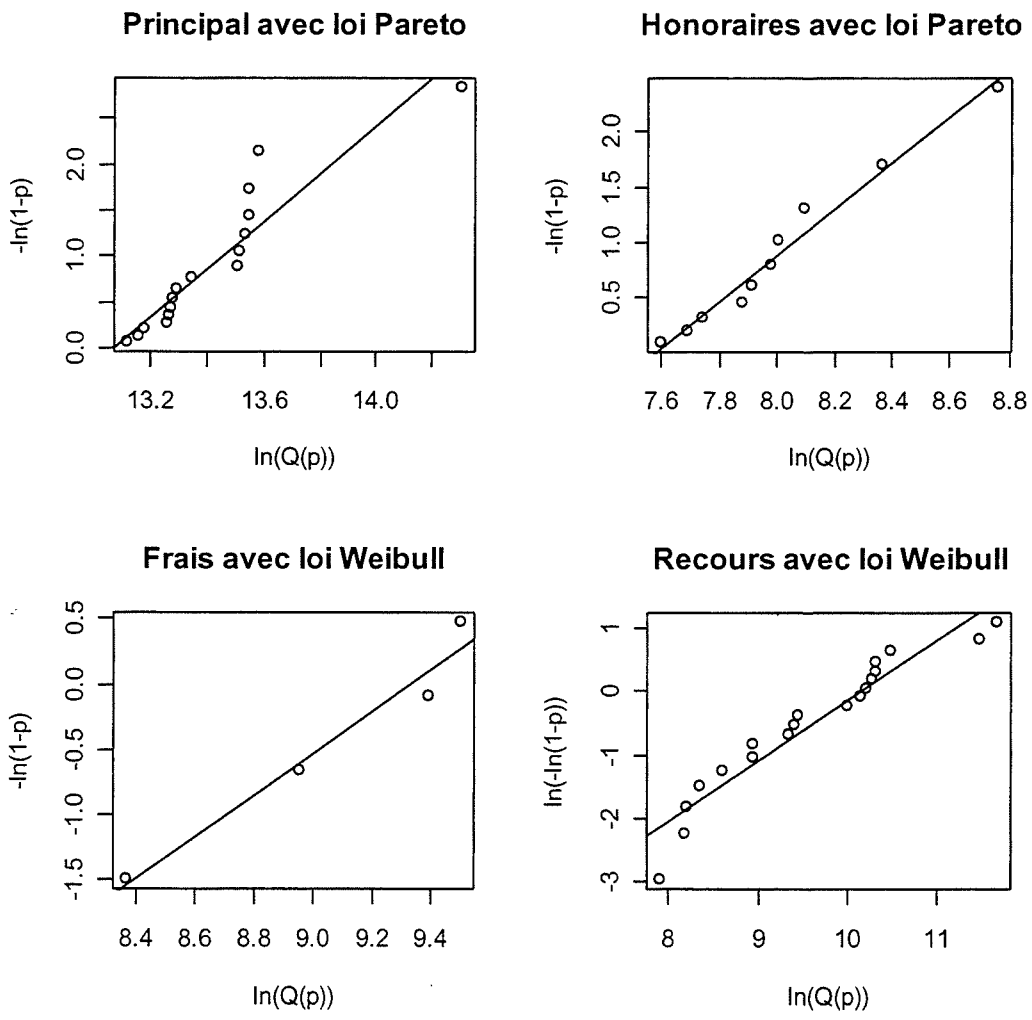


fig. 3.9 : Graphiques quantile - quantile

Il s'agit ensuite de d'estimer les paramètres associés à la loi sélectionnée. On peut utiliser les techniques standard, à savoir l'estimateur du maximum de vraisemblance ou la méthode des moments. Une fois ces estimateurs sélectionnés, nous avons utilisé le test de Kolmogorov-Smirnov afin de tester l'adéquation du modèle aux données. En particulier, les paiements en principaux ne semblent pas très bien s'ajuster avec une loi de Pareto. Il nous faut donc vérifier que le choix de cette loi est bien adapté.

Nous avons choisi d'utiliser la méthode des moments pour les paiements en principal et les honoraires, et la méthode du maximum de vraisemblance pour les frais et les recours.

Voici les ajustements obtenus :

— Pour les paiements en principal

Le meilleur ajustement se fait avec une loi de Pareto de paramètre $\alpha = 2.84$ conditionnellement à ce que l'on soit au dessus de 520 000.

— Pour les honoraires

Le meilleur ajustement se fait avec une loi de Pareto de paramètre $\alpha = 3.84$ conditionnellement à ce que l'on soit au dessus de 2 300.

— Pour les frais

Le meilleur ajustement se fait avec une loi de Weibull de paramètres $\alpha = 1.12$, $\beta \approx 5500$ conditionnellement à ce que l'on soit au dessus de 4 000.

— Pour les recours

Le meilleur ajustement se fait avec une loi de Weibull de paramètres $\alpha = 0.83$, $\beta \approx 21000$ conditionnellement à ce que l'on soit au dessus de 2 000.

Le test de Kolmogorov-Smirnov ne rejette pas l'hypothèse d'adéquation des lois sélectionnées aux données (cf. tableau 2.9 en annexe).

Ajustement global

En dehors de la queue de distribution, nous avons cherché à ajuster nos données empiriques a des lois classiques (loi lognormale ou loi gamma notamment). Cependant, frais et recours sont mieux modélisés par une loi uniforme.

Pour trouver les paramètres des lois, nous devons utiliser les résultats suivants :

- Soit deux variables aléatoire réelles X et Y. Soit Z la variable aléatoire de la loi mélangée (avec probabilité $P(Z = X) = p$). Alors Z vérifie :
  - $F_Z(t) = p \cdot F_X(t) + (1-p) \cdot F_Y(t)$
  - $\mathrm{E}(Z) = p \cdot \mathrm{E}(X) + (1-p) \cdot \mathrm{E}(Y)$
  - $V(Z) = p \cdot (1-p) \cdot (\mathrm{E}(X) - \mathrm{E}(Y))^2 + p \cdot V(X) + (1-p) \cdot V(Y)$

L'ajustement global est la variable aléatoire mélangée Z de la loi hors valeurs extrêmes et de la loi de la queue de distribution. Les paramètres de chaque loi hors valeurs extrêmes ont été ajustés à de façon à ce que la moyenne théorique de la loi mélangée soit égale à la moyenne empirique.

Voici les ajustements réalisés :

— Pour les paiements en principal

Nous avons utilisé une loi lognormale L_{pr} où $L_{pr} \sim \text{LogN}(8.68,2.1^2)$ conditionnellement à $L_{pr} < 520000$.

— Pour les honoraires

Une loi uniforme L_{ho} où $L_{ho} \sim \text{U}(100,2300)$.

— Pour les frais

Une loi Gamma L_{fr} où $L_{fr} \sim \text{G}(0.95,1000)$ conditionnellement à $L_{fr} < 4000$

— Pour les recours

Une loi uniforme L_{re} où $L_{re} \sim \text{U}(300,2000)$.

On obtient alors les ajustements suivants :

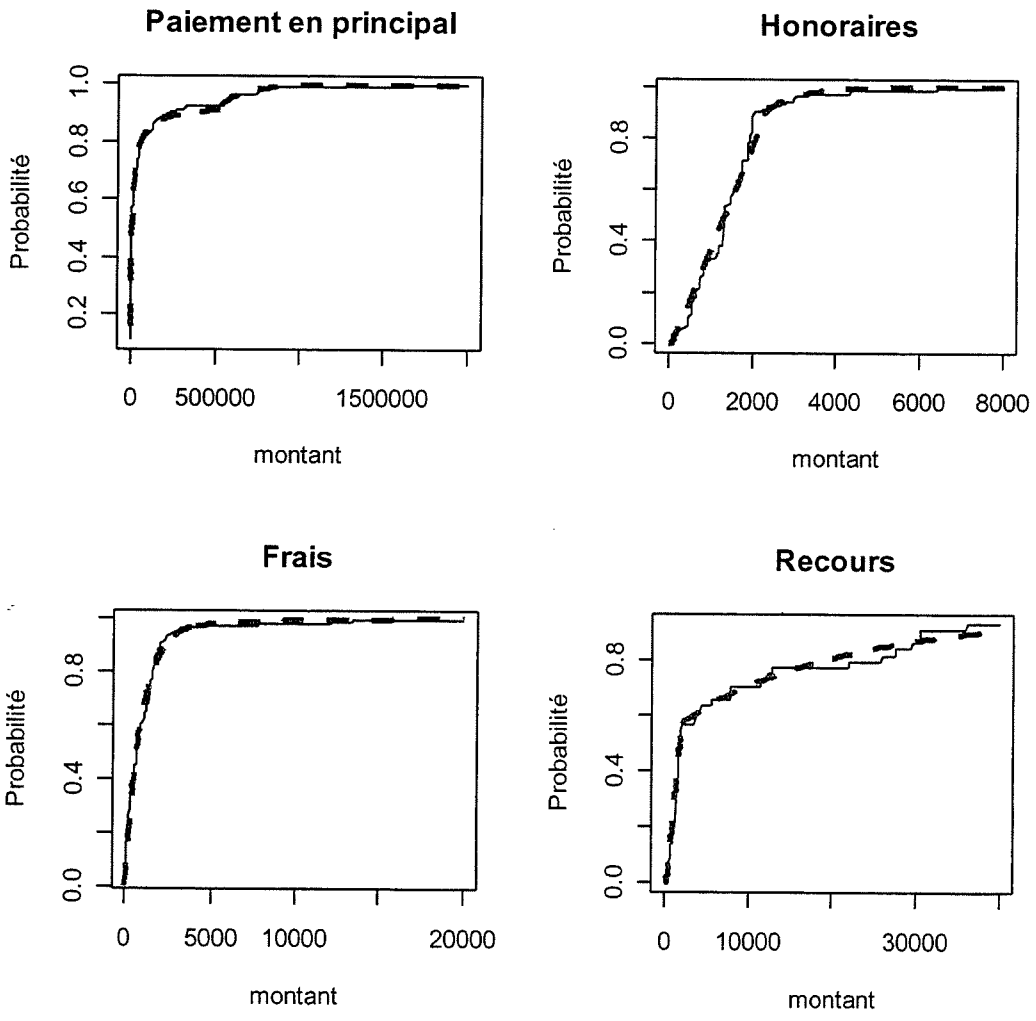


fig. 3.10 : Ajustement des fonctions de répartition

Cet ajustement permet donc d'obtenir une modélisation satisfaisante des flux des sinistres. En tirant aléatoirement 100 000 simulations de l'ensemble de ces flux, les deux premiers moments des données simulées diffèrent de moins de 1% des deux premiers moments des observations.

Nous avons donc déterminé un modèle de provisionnement individuel. Pour cela, nous utilisons des modèles linéaire généralisés pour simuler les dates des différents flux, et simulons le montant associé par des lois paramétriques. La clôture des sinistres est évaluée par des méthodes d'analyse de survie.

4. CALCUL DES RESERVES

Nous avons défini un modèle portant sur les sinistres individuels. Nous cherchons dans cette partie à pouvoir comparer les résultats de notre modèle à ceux donnés par les méthodes classiques. Etant donné que nous avons établi un modèle sur les paiements pour les sinistres déjà déclarés, nous évaluons la sinistralité ultime par année de déclaration. Les différents triangles présentés ici ne prennent pas en compte les recours, qui peuvent être étudiés de manière séparée.

Nous avons choisi de ne pas introduire de retraitement des sinistres graves afin de faciliter la comparaison entre les différents modèles. De même, nous n'avons pas introduit de facteur de queue. Nous supposons donc qu'il n'y aura plus de paiements après six années de développement, bien que ceci soit contraire à l'expérience.

Les méthodes présentées ici permettent d'obtenir des cadences de développement. En effet, cela est utile pour déterminer une politique de tarification, pour valoriser une compagnie en run-off, ou pour évaluer le capital économique de l'entreprise. Pour le calcul des réserves en elles-mêmes, il est également important de pouvoir actualiser les flux. En effet, les principes définis par la nouvelle norme comptable IFRS supposent l'actualisation des cash-flows.

Les méthodes les plus souvent utilisées se basent sur des données agrégées. Parmi les méthodes classiques, on distingue deux grandes familles, la méthode Chain-Ladder et la méthode Bornhuetter-Ferguson. La première suppose la reproductibilité du passé pour déterminer la sinistralité à venir, alors que la seconde suppose une estimation a priori de la sinistralité ultime. Les méthodes stochastiques permettent de mesurer l'incertitude liée à ces modèles. Les plus connues sont le modèle de Mack et le bootstrap.

Nous avons également choisi d'étudier des méthodes nouvelles dans notre analyse. En effet, de nombreux travaux de recherche sont effectués actuellement dans le domaine du provisionnement. La première méthode étudiée est une méthode stochastique permettant l'introduction d'avis d'experts. Pour cela, elle utilise la théorie bayésienne, où une loi a priori sur des paramètres est précisée. La seconde permet de prendre en compte des corrélations existant au sein du triangle. Dans ce modèle, des séries temporelles sont utilisées.

Après avoir étudié ces différents modèles, nous avons analysé les résultats du modèle individuel et comparé les différents résultats.

4.1 Résultats des méthodes classiques

Les méthodes déterministes

La méthode Chain-Ladder

La méthode Chain-Ladder est la méthode la plus souvent utilisée. Elle est fondée sur l'hypothèse que le déroulement des paiements est fonction de facteurs de développement λ_j, qui ne dépendent que de l'année de développement. Avec les même notations qu'en partie 1.2, le modèle sous-jacent peut s'écrire de la façon suivante : $C_{ij} = \lambda_j \cdot C_{ij-1}$.

Les coefficients λ_j peuvent être estimés à l'aide des observations par :

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}}$$

A partir de ces coefficients de passage, il est possible d'obtenir une estimation du montant des provisions, en prenant :

$$\hat{C}_{in} = \left(\prod_{k=1}^{i+j-n-1} \hat{\lambda}_{n-i+k} \right) \cdot C_{i,n+1-i}$$

Voici les résultats obtenus à partir de nos données :

Cumulative Claims Paid								
Reporting	Development year							
Year	1	2	3	4	5	6	Ult	Res
1997	4 740	210 938	1 850 617	2 608 283	3 495 492	4 607 766	4 607 766	0
1998	181 773	1 314 395	3 609 713	4 881 558	5 880 198	7 751 290	7 751 290	1 871 092
1999	112 302	1 137 729	2 286 295	2 962 406	3 708 303	4 888 293	4 888 293	1 925 887
2000	283 555	977 041	2 131 768	2 876 319	3 600 541	4 746 241	4 746 241	2 614 473
2001	474 694	1 078 972	2 928 080	3 950 756	4 945 507	6 519 178	6 519 178	5 440 206
2002	304 668	1 360 134	3 691 088	4 980 255	6 234 221	8 217 964	8 217 964	7 913 296
							36 730 731	19 764 953

Chain-Ladder						
Dev. Factor	4,464	2,714	1,349	1,252	1,318	1,000
Cumulative	26,974	6,042	2,226	1,650	1,318	1,000

tab. 4.1 : Résultats de la méthode Chain-Ladder

On obtient donc une réserve globale de 20 millions d'euros. On remarque que pour la première année de déclaration, le règlement des sinistres a été effectué plus tardivement que les autres années.

Si l'on compare la sinistralité ultime de 2002 à celle de 2000 et 2001, les chiffres de 2002 semblent surestimés. En effet, au regard des données issues du triangle, on peut se demander pourquoi la sinistralité associée à 2002 est la plus élevée. Dès la seconde année de développement, les paiements cumulés de 2002 seraient les plus importants, ce qui est plutôt étonnant.

La méthode Chain-Ladder nous montre ici ses limites. En effet, cette méthode suppose la reproductibilité du passé. Or les paiements associés aux sinistres déclarés en 1997 connaissent un développement moins rapide que les autres. Si l'on s'intéresse à l'acquisition du portefeuille, ce n'est pas très étonnant. La compagnie d'assurance a racheté le portefeuille en 1997. C'est pourquoi de nombreux sinistres ont été mis en litige afin de déterminer quel assureur devait indemniser les assurés. Aussi, le facteur de développement de la cinquième année (1.318), évalué uniquement à partir de des données de 1997, semble lui aussi surévalué. Cette opinion est renforcée par le fait que les facteurs de développement ont plutôt tendance à diminuer lorsque les années augmentent, ce qui n'est pas le cas pour ce facteur.

Il convient de noter que nous ne nous intéressons pas aux phénomènes de queues. Aussi, nous considérons que le développement des sinistres n'est étudié que pendant les six premières années de développement, et ce afin de faciliter la comparaison entre les différents modèles. Lorsque nous parlons d'ultime, il s'agit donc des paiements cumulés au bout de six années de développement, et les réserves sont les réserves associées à ces six années.

La méthode Chain-Ladder a donc l'avantage d'être très simple d'utilisation. Cependant, elle montre ici ses limites et il important de comparer le résultat de la méthode Chain-Ladder avec d'autres méthodes.

La méthode London-Chain

La méthode London-Chain est une variante de la méthode Chain-Ladder.

On suppose ici que la dynamique des données cumulées se fait selon au processus autorégressif d'ordre 1 de la forme $C_{ij} = \lambda_j \cdot C_{ij-1} + \alpha_j$.

Pour chaque année, en plus d'un facteur multiplicatif, il y a donc également un facteur additif. Pour la méthode Chain-Ladder, on imposait à ce facteur d'être nul. La méthode la plus naturelle

consiste alors à estimer ces paramètres à l'aide de la méthode des moindres carrés, c'est-à-dire que l'on cherche, pour tout k :

$$(\hat{\lambda}_k, \hat{\alpha}_k) = \arg\min\left\{\sum_{i=1}^{n-k}\left(C_{i,k+1} - \alpha_k - \lambda_k C_{i,k}\right)^2\right\}$$

Ce modèle est moins utilisé que la méthode Chain-Ladder. Cependant, étant donné que nous pensons que l'ultime associé à 2002 est trop important avec Chain-Ladder, il est intéressant de voir quelle est l'influence de l'introduction d'un facteur additif.

Voici les résultats obtenus :

Cumulative Claims Paid

Reporting Year	Development year 1	2	3	4	5	6	Ult	Res
1997	4 740	210 938	1 850 617	2 608 283	3 495 492	4 607 766	4 607 766	0
1998	181 773	1 314 395	3 609 713	4 881 558	5 880 198	7 751 290	7 751 290	1 871 092
1999	112 302	1 137 729	2 286 295	2 962 406	3 692 203	4 867 070	4 867 070	1 904 664
2000	283 555	977 041	2 131 768	2 875 922	3 582 164	4 722 016	4 722 016	2 590 249
2001	474 694	1 078 972	2 948 098	3 978 088	4 984 519	6 570 603	6 570 603	5 491 631
2002	304 668	1 410 282	3 886 453	5 245 005	6 596 495	8 695 515	8 695 515	8 390 847
							37 214 260	20 248 482

London-Chain					
Dev. Factor	5,002	2,832	1,350	1,272	1,318
Cst	-113 689	-107 827	-2 278	-77 051	0

tab. 4.2 : Résultats de la méthode London-Chain

Les résultats sont relativement comparables avec la méthode Chain-Ladder. Le principal écart provient de l'année 2002, où la prévision des paiements futurs semble à nouveau excessive si l'on la compare avec celle des deux années précédentes. En effet, l'évaluation des paiements cumulés après deux ans de développement pour l'année 2002 (1 410 282€) est supérieure à tous les paiements cumulés après deux ans de développement des autres années de déclaration.

Cette méthode a donc tendance à augmenter le provisionnement associé à 2002. De manière plus générale, elle conduit à réduire le provisionnement des années les plus anciennes (1999 et 2000) et à augmenter celui des années les plus récentes (2001 et 2002).

Elle ne permet donc pas de prendre en compte la rupture existant au niveau de la cadence de développement entre 1997 et les années les plus récentes. Au contraire, la méthode London-Chain conduit à augmenter les réserves associées aux années récentes alors qu'elles nous semblaient déjà excessives avec Chain-Ladder.

Analyse des facteurs de développement

Les méthodes que nous venons de décrire demandent une grande stabilité dans le portefeuille étudié. Or notre portefeuille est relativement hétérogène. En particulier, tous les sinistres étudiés sont des sinistres dont la date de survenance est postérieure à 1997 inclus. Aussi tous les sinistres déclarés en 1997 sont des sinistres survenus en 1997 alors que les sinistres déclarés en 2002 ont eu lieu de 1997 à 2002. Le modèle individuel permet de distinguer les différents types de sinistres, ce qui n'est pas le cas pour les modèles agrégés. De plus, comme nous l'avons dit précédemment, le portefeuille a été revendu en 1997, ce qui a engendré des retards dans les paiements.

Afin de déterminer les facteurs de développement le plus adaptés, nous pouvons comparer différentes méthodes de calcul d'évaluation des facteurs de développement. Il est en effet possible d'introduire des pondérations lors de l'estimation des facteurs de développement. On considère alors des facteurs de développement de la forme :

$$\hat{\lambda}_k = \frac{1}{\sum_{i=1}^{n-k} w_{ik}} \sum_{i=1}^{n-k} w_{ik} \frac{C_{i,k+1}}{C_{ik}}$$

On peut noter que si $w_{ik} = C_{ik}$, on retrouve la méthode Chain-Ladder standard.

Ici, nous avons choisi d'étudier la moyenne simple, la moyenne hors extrema, la moyenne pondérée selon Chain-Ladder, et la moyenne pondérée des 3 dernières années.

Development Factors

Reporting Year	Development year 1	2	3	4	5	6
1997	44,50	8,77	1,41	1,34	1,32	
1998	7,23	2,75	1,35	1,20		
1999	10,13	2,01	1,30			
2000	3,45	2,18				
2001	2,27					
2002						

Indicateurs

Moyenne simple	13,517	3,928	1,352	1,272	1,318
Moyenne hors extrema	6,936	2,464	1,352	N/A	N/A
Moyenne pondérée	4,464	2,714	1,349	1,252	1,318
Moyenne pondérée des 3 derniers	3,669	2,341	1,349	N/A	N/A

tab. 4.3 : Indicateurs pour le choix du facteur de développement

Les facteurs de développement diffèrent énormément, et ce spécialement pour les première années de développement. L'année 1997 semble peu représentative de l'évolution des autres années. Il y

a pour cette année un retard dans le déroulement des paiements, ce qui peut être expliqué par la cession du portefeuille.

En particulier, le facteur de développement pondéré des trois dernières années de développement est sensiblement plus faible que celui obtenu par Chain-Ladder. Si l'on se base seulement sur les deux dernières années, qui sont comparables du point de vue du montant payé au bout d'un an de développement, on obtient un facteur inférieur à 3. C'est pourquoi il nous semble possible d'évaluer ce premier facteur à 3.2.

Pour ce qui est du facteur associé à la cinquième année de développement, nous n'avons comme donnée que l'année 1997. Or cette année semble en retard par rapport aux autres années. On peut en effet raisonnablement penser que le facteur de développement à tendance à diminuer au cours de dernières années et c'est pourquoi il semble possible d'évaluer ce facteur à 1.24.

Cette analyse nous permet de déterminer des facteurs de développement estimés :

Cumulative Claims Paid

Reporting Year	Development year 1	2	3	4	5	6	Ult	Res
1997	4 740	210 938	1 850 617	2 608 283	3 495 492	4 607 766	4 607 766	0
1998	181 773	1 314 395	3 609 713	4 881 558	5 880 198	7 291 445	7 291 445	1 411 248
1999	112 302	1 137 729	2 286 295	2 962 406	3 708 303	4 598 296	4 598 296	1 635 890
2000	283 555	977 041	2 131 768	2 876 319	3 600 541	4 464 670	4 464 670	2 332 903
2001	474 694	1 078 972	2 928 080	3 950 756	4 945 507	6 132 429	6 132 429	5 053 456
2002	304 668	974 936	2 645 750	3 569 818	4 468 654	5 541 131	5 541 131	5 236 463
							32 635 737	15 669 959

Analyse						
Evaluation	3,200	2,714	1,349	1,252	1,240	1,000
Cumulative	18,187	5,684	2,094	1,552	1,240	1,000

tab. 4.4 : Résultats de l'analyse

Les résultats obtenus sont alors sensiblement différents de ceux obtenus par la méthode Chain-Ladder classique. La réserve globale n'est plus que de 15.7 millions d'Euros, soit 4 millions de moins que pour la méthode Chain-Ladder classique. En particulier, la charge ultime de 2002 est plus cohérente avec la charge ultime obtenue pour les deux années précédentes. D'une manière générale, les réserves sont moins importantes, car les années de déclaration les plus anciennes sont moins prises en compte alors qu'elles avaient entraîné des paiements tardifs, et donc des facteurs multiplicatifs plus importants.

Bornhuetter-Ferguson

Cette méthode utilise les facteurs de développement trouvés par la méthode Chain-Ladder. Seulement cette fois la sinistralité ultime initialement attendue de chaque année de survenance (ou de déclaration) est supposée connue. Dans ce cas, cette sinistralité a été déterminée par des avis d'experts, par exemple la méthode du loss-ratio attendu.

Soit Ult_i l'ultime d'une année de survenance. Le dernier facteur cumulé connu $C_{i,n-i+1}$ est remplacé par :

$$Ult_i \frac{1}{\hat{\lambda}_{n-i+2}\hat{\lambda}_{n-i+3}\cdots\hat{\lambda}_n}$$

Les réserves associées à cette année sont alors :

$$Ult_i \frac{1}{\hat{\lambda}_{n-i+2}\hat{\lambda}_{n-i+3}\cdots\hat{\lambda}_n}\left(\hat{\lambda}_{n-i+2}\hat{\lambda}_{n-i+3}\cdots\hat{\lambda}_n - 1\right)$$

Aussi, si l'on compare la méthode Chain-Ladder et la méthode Bornhuetter-Ferguson, on remarque que cette dernière utilise des informations exogènes pour déterminer le « niveau » de chaque ligne dans le triangle, alors que Chain-Ladder utilise directement l'information de la ligne. La méthode de Bornhuetter-Ferguson est à rapprocher des méthodes bayésiennes, pour lesquelles des informations exogènes au modèle sont utilisées pour former des distributions a priori.

Nous avons choisi d'utiliser le même ultime pour toutes les années sauf 1998. En effet, si à partir de 2000 les paiements sont plus importants lors de la première année de développement, on remarque que les paiements cumulés après deux ans sont très similaires. Il semble donc que le résultat obtenu pour l'année 2002 n'indique pas qu'il s'agit d'une forte sinistralité. Il s'agirait plutôt d'un changement dans la cadence des paiements survenu depuis quelques années.

Voici les ultimes sélectionnés et les résultats associés :

	1997	1998	1999	2000	2001	2002	Total
Ultimes a priori	4 607 766	7 000 000	5 000 000	5 000 000	5 000 000	5 000 000	31 607 766
Réserves	0	1 689 737	1 969 897	2 754 257	4 172 463	4 814 633	15 400 987

tab. 4.5 : Résultats du Bornhuetter-Ferguson

Cette méthode conduit à réduire significativement les réserves associées à 2001 et 2002. De manière globale, les réserves calculées (15.4 millions d'Euros) sont similaires aux réserves évaluées lors de l'analyse des facteurs de développement. Cependant, étant donné que le facteur de développement de la cinquième année (1.318) n'est pas modifié par la méthode Bornhuetter-

Ferguson, les réserves associées aux années les plus anciennes (1998 et 1999) ont tendance à se rapprocher des réserves calculées par la méthode Chain-Ladder standard. Les réserves associées aux dernières années sont pour leur part plus faibles que celles calculées lors de l'analyse des facteurs de développement.

Les méthodes stochastiques

La finalité première de l'approche stochastique est de mesurer l'incertitude présente dans les triangles de liquidation et les résultats des méthodes déterministes.

De manière plus générale, les méthodes stochastiques permettent d'expliciter les hypothèses utilisées dans les modèles déterministes, et de les valider, au moins partiellement. Elles permettent également d'évaluer la variabilité de la provision « prévue » par le modèle et d'obtenir des estimations et intervalles de confiance pour des paramètres d'intérêt liés à la provision. Il est ensuite possible, par exemple à l'aide de la technique du bootstrap, d'estimer la loi de probabilité de la provision et de simuler la sinistralité des exercices futurs.

Le benchmark incontournable reste la méthode de Chain-Ladder.

Il convient de noter que ces méthodes prennent une place de plus en plus importante dans la détermination des réserves. En effet, les méthodes de calcul proposées lors de la mise en place des nouvelles normes IAS préconisent l'utilisation de ce type de méthode.

Le modèle de Mack

Mack a proposé un modèle stochastique, relatif à la méthode Chain-Ladder Ce modèle repose sur un certain nombre d'hypothèses :

- $H_1 : E\left(C_{i,k+1} \middle| C_{i,1},\ldots,C_{i,k}\right) = \lambda_k C_{ik}$
- $H_2 : \{C_{i,1},\ldots,C_{i,n}\}$ et $\{C_{j,1},\ldots,C_{j,n}\}$ sont indépendants

Aussi, comme pour Chain-Ladder, l'espérance du montant cumulé des paiements après k années de développement est entièrement déterminée par l'année précédente et un facteur de développement indépendant de l'année de survenance. De plus, deux années de survenance différentes sont indépendantes. Ces deux hypothèses permettent d'obtenir en espérance les mêmes réserves qu'avec la méthode Chain-Ladder.

A partir de ces hypothèses, il possible d'étudier l'erreur de prévision (i.e. l'écart entre la distribution l'estimateur des réserves et la vrai valeur des réserves). L'erreur quadratique moyenne (mean square error) du montant des provisions pour l'année i est estimé par :

$$\hat{mse}\left(\hat{R}_i\right) = \hat{C}_{in}^2 \sum_{k=n-i+1}^{n-1} \frac{\hat{\sigma}_k^2}{\hat{\lambda}_k^2} \left(\frac{1}{C_{ik}} + \frac{1}{\sum_{j=1}^{n-k} C_{jk}} \right)$$

où :

$$\hat{\sigma}_k^2 = \frac{1}{n-k-1} \sum_{i=1}^{n-k} C_{ik} \left(\frac{C_{ik}}{C_{ik}} - \hat{\lambda}_k \right)^2$$

Ce dernier facteur intervient dans l'hypothèse suivante, sous jacente au modèle de Mack :

- $H_3 : Var\left(C_{i,k+1} \middle| C_{i,1}, \ldots, C_{i,k}\right) = C_{i,k}\sigma_k^2$

On peut également en déduire l'erreur quadratique moyenne du montant total des provisions :

$$\hat{mse}\left(\hat{R}\right) = \sum_{i=2}^{n} \left(\hat{mse}\left(\hat{R}_i\right) + \hat{C}_{in} \left(\sum_{j=i+1}^{n} \hat{C}_{jn} \right) \sum_{k=n-i+1}^{n-1} \frac{2\hat{\sigma}_k^2}{\hat{\lambda}_k^2} \frac{1}{\sum_{j=1}^{n-k} C_{jk}} \right)$$

Dans ce modèle, il n'y a pas de distribution statistique imposée pour les données. Nous pourrons cependant faire l'hypothèse que les paiements incrémentaux sont distribués selon de lois Normales ou Lognormales.

Voici donc les estimateurs obtenus, où $se\left(R_i\right) = \sqrt{mse\left(R_i\right)}$, avec des intervalles de confiance à 95% :

	Année	1998	1999	2000	2001	2002	Ensemble
	Ri	1 871 092	1 925 887	2 614 473	5 440 206	7 913 296	19 764 953
	Se(Ri)	323 655	524 621	560 229	4 508 590	9 028 366	10 777 117
	Coef Var	17%	27%	21%	83%	114%	55%
IC Normal	Sup	2 505 455	2 954 143	3 712 522	14 277 043	25 608 893	40 888 102
	Inf	1 236 729	897 631	1 516 423	-3 396 631	-9 782 301	-1 358 196
IC LogN	Sup	2 581 367	3 139 251	3 872 499	17 281 526	31 226 891	47 169 029
	Inf	1 316 851	1 099 889	1 687 641	1 015 258	871 248	6 383 957

tab. 4.6 : Résultats du modèle de Mack

Les intervalles de confiance obtenus sont donc très grands. Le quantile à 2.5% associé à une distribution Normale prend même des valeurs négatives pour 2001, 2002 et pour l'ensemble des

réserves. Une telle distribution n'est donc pas adaptée. On lui préférera la distribution Lognormale. Cependant, l'intervalle de confiance associé aux deux dernières années reste trop important.

Pour l'année 2002, l'écart type même est supérieur à la prévision. Le modèle de Mack nous montre donc que les réserves calculées par la méthode Chain-Ladder sont soumises à une très grande incertitude. Les résultats obtenus avec la méthode Chain-Ladder standard doivent donc être pris en compte avec la plus grande précaution.

La méthode du bootstrap

Le bootstrap est une technique de plus en plus utilisée. Nous présentons ici le modèle étudié par Renshaw et Verral (1998). Ces derniers ont utilisé un modèle poissonnien, avec surdispersion de telle sorte que les données incrémentales Y_{ij} vérifient $\mathrm{E}(Y_{ij}) = \mu_{ij}$ et $Var(Y_{ij}) = \phi\mu_{ij}$, où ϕ est le paramètre de dispersion.

Sous certaines conditions (données incrémentales positives,...), England (2001) note que les réserves obtenues par cette méthode sont les mêmes que les réserves obtenues par la méthode Chain-Ladder

La méthode est la suivante. A partir des facteurs de développement de Chain-Ladder issus des données historiques, on construit un triangle ajusté tel que les paiements cumulés de la diagonale soient égaux aux paiements cumulés du triangle de départ et vérifiant exactement la condition de Chain-Ladder (à savoir $\widetilde{C}_{ij} = \lambda_j \cdot \widetilde{C}_{ij-1}$).

Il est alors possible d'obtenir le triangle des résidus ajustés de Pearson, définis par :

$$PR_{ij} = \sqrt{\frac{n}{n-p}} \frac{Y_{ij} - \widetilde{Y}_{ij}}{\sqrt{\widetilde{Y}_{ij}}}$$

où n est le nombre d'observations du triangle, et p le nombre de paramètres estimés.

La méthode du bootstrap consiste à retirer un échantillon de ces erreurs dans un nouveau triangle (en excluant la cellule en haut à droite du triangle et celle en bas à gauche, qui seront toujours nulles). Il est alors possible construire un pseudo triangle incrémental défini par $\widetilde{Y}_{ij} + rand(PR_{kl}) \cdot \sqrt{\widetilde{Y}_{ij}}$.

Les facteurs de développement de ce nouveau triangle ainsi que l'ensemble des paiements incrémentaux peuvent alors être déterminés.

Cette méthode, itérée de nombreuses fois, permet d'obtenir une mesure de l'erreur d'estimation de chaque paiement incrémental. A cela, il faut ajouter l'erreur du modèle, en prenant en compte l'hypothèse affirmant que les paiements incrémentaux sont distribués selon une loi de Poisson avec surdispersion.

Pour faire la simulation de la variabilité de chaque pseudo donnée, il nous faut connaître le paramètre de dispersion. Ce dernier est estimé comme étant la somme des résidus de Pearson ajustés divisé par le nombre de paramètres. Pour simuler la loi de la distribution, England (1999) a utilisé une loi Gamma dont la moyenne est égale aux paiements projetés incrémentaux des pseudo triangles et dont la variance est égale à la moyenne multipliée par le paramètre d'échelle. D'un point de vue pratique, lorsque cette moyenne est négative, on suppose qu'il n'y a pas de paiements. En effet, nous ne travaillons ici que sur des triangles hors recours, et il n'y a pas de paiements négatifs.

Voici les estimations des moyennes annuelles et des intervalles de confiance associés obtenues à partir d'un bootstrap réalisé avec 10 000 simulations :

	Année	1998	1999	2000	2001	2002	Ensemble
	Moyenne	1 878 463	1 929 746	2 597 262	5 444 237	7 506 241	19 355 950
IC 95%	sup	3 032 994	3 216 372	4 298 090	9 318 492	16 280 733	29 200 898
	Inf	957 144	876 173	1 076 055	1 653 816	0	9 914 484

tab. 4.7 : Résultats de la méthode du bootstrap

En moyenne, on ne retrouve pas tout à fait les réserves liées à Chain-Ladder en raison des moyennes négatives existant lors de simulation des pseudo triangles.

D'une manière générale, les intervalles de confiance semblent plus réalistes que ceux obtenus avec le modèle de Mack. L'incertitude reste tout de même importante.

Il convient de noter que l'intervalle associé à 2002 ne nous apparaît pas être cohérent. Il supposerait en effet qu'il y a plus de 2.5% de chances qu'il n'y ait pas de paiements futurs associés à l'année 2002. Or ceci ne nous semble pas être réaliste. De plus, si cette réserve était cohérente, le quantile à 2.5% associé aux autres années devrait lui aussi être nul.

Le bootstrap permet également d'obtenir une estimation de la distribution de l'ensemble des paramètres d'intérêt de l'étude. Voici une estimation de la distribution du montant total des réserves :

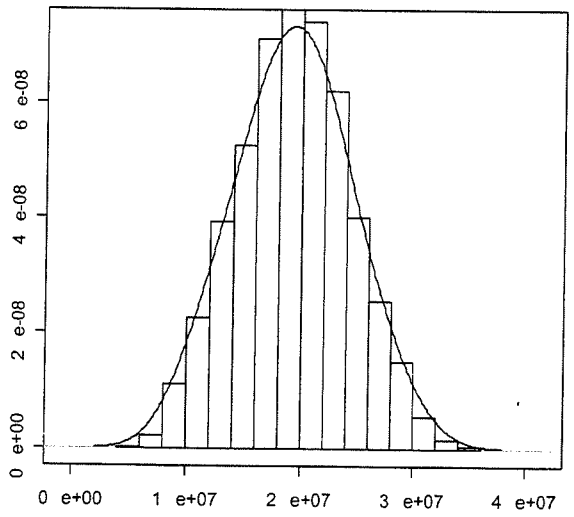


fig.. 4.1 : Estimation de la distribution du montant total des réserves

La forme de la distribution du montant total des réserves obtenue par simulation est relativement symétrique.

La dispersion de la distribution du montant total des réserves est donc à nouveau très importante. Cependant, les résultats obtenus, 2002 mis à part semblent plus cohérents que ceux obtenus par le modèle de Mack.

4.2 Méthode stochastique introduisant des avis d'experts

Présentation du modèle

Dans son article « Obtaining predictive distributions for reserves which incoporate expert opinion » (2004), Verral s'intéresse à l'introduction d'avis d'experts dans les méthodes stochastiques. Pour cela, ce dernier utilise des méthodes bayésiennes.

En effet, comme nous l'avons dit précédemment, de nombreuses méthodes de provisionnement stochastiques ont été proposées ces dernières années dans la littérature actuarielle. Seulement

jusqu'à présent, les différents modèles proposés ne permettent pas l'intervention de l'actuaire dans le processus de détermination des réserves. En effet, l'actuaire doit procéder à une analyse des résultats des méthodes classiques afin de déterminer les réserves les plus appropriées. C'est d'ailleurs ce que nous avons fait dans le paragraphe sur l'analyse des facteurs de développement. De nombreux phénomènes peuvent justifier l'intervention de l'actuaire dans le processus de détermination des réserves. On peut citer en particulier les changements de législation, ou de politique interne.

Pour l'actuaire, il est beaucoup plus difficile d'intervenir sur les méthodes stochastiques. Par exemple, il est possible de changer un ou plusieurs des résidus avant d'appliquer la méthode du bootstrap. Mais en faisant cela, les hypothèses du modèle sont modifiées. Et il n'est pas évident de savoir quel sera l'effet sur l'erreur de prédiction. Aussi, s'il est toujours possible de calculer l'erreur d'estimation de n'importe quel paramètre estimé dans un modèle stochastique, quelle estimation de l'erreur doit être utilisée pour un paramètre qui a été introduit par l'actuaire ? Pour répondre à cette question, Verral propose d'utiliser des méthodes bayésiennes.

En effet, ces méthodes permettent d'introduire des avis d'expert et de surmonter les problèmes d'erreur d'estimation. Ceci a été rendu possible grâce au développement des simulations de Monte Carlo markoviennes (MCMC – Markov Chain Monte Carlo), qui permettent d'utiliser la simulation lors de l'analyse.

Les simulations de Monte Carlo markoviennes (MCMC)

Les simulations de Monte Carlo markoviennes sont très utilisées en analyse bayésienne.

En effet, plusieurs problèmes bayésiens ont pour solution : $\mathrm{E}\left(g(\theta)\middle|y^0\right)$, où g est une fonction quelconque, y^0 le vecteur des observations et θ le vecteur des paramètre inconnus.

En particulier, la valeur $\hat{\theta}_k$ d'un paramètre minimisant l'erreur quadratique d'un paramètre $\mathrm{E}\left[\left(\theta_k-\hat{\theta}_k\right)^2\middle|y^0\right]$ est $\mathrm{E}\left(\theta_k\middle|y^0\right)$.

Il est également possible de calculer l'écart type a posteriori autour de la valeur d'un paramètre par $\mathrm{E}\left[\left(\theta_k-\hat{\theta}_k\right)^2\middle|y^0\right]^{0.5}$.

Enfin, la densité prédictive d'une variable observable est $f\left(y_{T+1}\middle|y^0\right)=\mathrm{E}\left[f\left(y_{T+1}\middle|\theta,y^0\right)\middle|y^0\right]$.

Aussi, l'analyse bayésienne amène très souvent à calculer des fonctions du type $\mathrm{E}\left(g(\theta)\middle|y^0\right)$.

Pour cela, il existe plusieurs possibilités. Le calcul analytique en est une. Seulement, si le calcul analytique permet d'obtenir des résultats exacts, sa mise en œuvre est presque toujours impossible en pratique.

C'est pourquoi les méthodes les plus souvent utilisées sont des méthodes de simulation.

La première méthode de simulation que nous allons présenter est la simulation de Monte Carlo indépendante. Il s'agit de simuler une famille (θ_m) indépendante identiquement distribuée telle que $\theta^m \sim \theta|y^0$. Dans ce cas, on a le résultat suivant :

$$\frac{1}{M}\sum_{m=1}^{M} g(\theta^m) \xrightarrow{p.s.} \mathrm{E}(g(\theta)|y^0)$$

Seulement, là encore, il est en général impossible de simuler cette famille (θ_m). C'est pourquoi, on lui préfère la simulation de Monte Carlo markovienne. Pour cette méthode, on cherche un processus markovien $f(\theta_{m+1}|\theta_m)$ pour laquelle $f(\theta|y^0)$ est la loi stationnaire unique. On peut alors montrer que :

$$\theta^m \sim \mathrm{f}(\theta|y^0) \Rightarrow \theta^{m+1} \sim \mathrm{f}(\theta|y^0)$$

Sous quelques conditions techniques, on a alors le même résultat que pour la simulation de Monte Carlo indépendante, à savoir :

$$\frac{1}{M}\sum_{m=1}^{M} g(\theta^m) \xrightarrow{p.s.} \mathrm{E}(g(\theta)|y^0)$$

Pour trouver le processus markovien $f(\theta_{m+1}|\theta_m)$ pour laquelle $f(\theta|y^0)$ est la loi stationnaire unique, il est possible d'utiliser l'échantillonnage de Gibbs.

L'échantillonnage de Gibbs est une méthode permettant d'obtenir des tirages d'une loi multidimensionnelle à partir de la loi de chacune des composantes conditionnellement à toutes les autres.

Pour mettre en œuvre l'algorithme, il faut réaliser différentes itérations sur les différentes lois conditionnelles :

Itération i

1/ Tirage de $x_1^i | x_2^{i-1}, x_3^{i-1}, \cdots x_n^{i-1}$

2/ Tirage de $x_2^i | x_1^i, x_3^{i-1}, \cdots x_n^{i-1}$

...

n/ Tirage de $x_m^i \left| x_1^i, x_2^i, \cdots x_{n-1}^{i-1} \right.$

On peut montrer que cet algorithme permet d'obtenir une chaîne de Markov admettant comme loi stationnaire la loi visée.

Application à la méthode Chain-Ladder

Approche proposée

Nous considérons ici qu'une distribution a priori intervient dans des paramètres du modèle de Chain-Ladder. Dans son article, Verral considère deux cas. Dans le premier, il suppose qu'un facteur de développement particulier doit avoir une certaine valeur (fixée a priori) pour certaines années de survenance. Dans ce premier exemple, il étudie deux possibilités pour la variance. Soit elle est importante, et dans ce cas, le paramètre est estimé séparément des autres lignes. Soit il utilise une petite variance (0.1 fois la distribution a priori), et alors la moyenne a priori à une influence plus importante.

L'auteur considère également le cas où seulement l'information la plus récente est utilisée. Pour les années de développement pour lesquelles cela est possible, il sélectionne uniquement les trois années de développement les plus récentes.

Pour l'étude des réserves, Verral propose d'utiliser un modèle binomial négatif, pour lequel il est possible d'utiliser différents facteurs de développement pour chaque ligne. De cette manière, nous pouvons choisir des distributions a priori reproduisant les réserves du modèle Chain-Ladder, ou utiliser des distributions a priori fondées sur une connaissance externe du portefeuille.

Les paiements incrémentaux $Y_{ij} \left| C_{i,j-1}, \lambda_{ij}, \phi \right.$ sont distribués selon une loi binomiale négative, de moyenne et variance respectives :

$$\left(\lambda_{ij} - 1\right) \cdot C_{i,j-1} \text{ et } \phi \cdot \lambda_{ij} \cdot \left(\lambda_{ij} - 1\right) \cdot C_{i,j-1}$$

Résultats obtenus

Avant de donner les résultats de ce modèle, il convient d'ajouter que le paramètre de dispersion aurait pu être évalué de manière bayésienne. Ici, à la manière de ce qu'a fait Verral, nous avons préféré choisir celui obtenu par la méthode du bootstrap.

Changement d'un facteur de développement

Nous considérons tout d'abord l'estimateur du facteur associé à la première année de développement. L'estimateur de Chain-Ladder pour cette année est de 4,64. Ce facteur semble influencé par les premières années de déclaration. En effet, le facteur individuel de la première année est de 44, alors que la moyenne des deux derniers facteurs est plus petite que 3. C'est pourquoi, nous proposons de choisir le même facteur de développement que celui déterminé lors de l'analyse des facteurs de Chain-Ladder, à savoir 3,2.

Nous ne donnons ici que les résultats obtenus avec une variance faible, car sinon les erreurs de prévision pour l'année 2002 deviennent trop importants.

On obtient alors les résultats suivants sous le logiciel WinBugs, avec 5000 simulations, et avec 5000 simulations d'initialisation des paramètres :

	année	1998	1999	2000	2001	2002	Ensemble
	moyenne	1 933 000	1 975 000	2 686 000	5 614 000	5 721 000	17 930 000
IC 95%	sup	4 033 000	3 839 000	5 063 000	11 510 000	14 940 000	31 130 000
	inf	607 600	757 200	1 148 000	2 203 000	1 084 000	9.569 000

tab. 4.7 : Changement du premier paramètre

Bien évidemment, nous obtenons des réserves inférieures aux réserves associées à la méthode Chain-Ladder pour l'année 2002. Pour les autres années, les résultats sont quasiment identiques. Pour ce qui est de l'incertitude liée au modèle, elle est comparable aux résultats trouvés par la méthode du bootstrap. Mais pour cela, nous avons dû prendre un facteur de dispersion faible, ce qui veut dire que nous pensons que le facteur de développement proposé est un facteur de développement fiable.

Prise en compte des années les plus récentes uniquement

La deuxième méthode proposée est de considérer uniquement l'information la plus récente pour l'estimation de chaque facteur de développement. Aussi, pour calculer les facteurs de développement, nous n'utilisons que les trois années les plus récentes. Voici les résultats obtenus lors de l'analyse de cette hypothèse :

	Année	1998	1999	2000	2001	2002	Ensemble
	Moyenne	1 950 000	1 991 000	2 730 000	4 832 000	6 434 000	17 940 000
IC 95%	sup	4 109 000	3 910 000	5 100 000	10 060 000	19 340 000	33 650 000
	Inf	648 500	759 000	1 177 000	1 846 000	990 700	9 122 000

tab. 4.8 : Exclusion des années anciennes

Les résultats obtenus sont intéressants car l'on retrouve à nouveau des provisions inférieures à Chain-Ladder. Il s'agit d'une phénomène dont nous avions déjà parlé : les années les plus anciennes ont connu un développement plus tardif que les années récentes, et les prendre en compte dans l'estimation amène à surestimer les réserves des années les plus récentes. Néanmoins, les réserves associées à 2002 sont supérieures à celles obtenues par analyse des facteurs de développement.

En ce qui concerne l'intervalle de confiance associé au modèle, il est légèrement supérieur à celui obtenu par la méthode du bootstrap. Seulement cette fois, nous n'avons pas eu besoin d'introduire un paramètre de dispersion faible.

Utilisation des facteurs sélectionnés lors de l'analyse

Lors de l'analyse faite dans la section sur les méthodes déterministes, nous n'avions pas seulement changé le premier facteur de développement, mais nous avions également sélectionné un autre facteur pour la dernière année. Nous avons donc mis cela en place lors d'une nouvelle simulation. Les résultats obtenus sont les suivants :

	année	1998	1999	2000	2001	2002	Ensemble
	moyenne	1 180 000	1 510 000	2 216 000	4 992 000	5 176 000	15 070 000
IC 95%	sup	3 557 000	3 461 000	4 446 000	10 400 000	14 300 000	28 790 000
	inf	69 660	439 100	840 400	1 915 000	898 200	9 503 000

tab. 4.9 : Sélection des facteurs de l'analyse

On obtient des résultats se rapprochant sensiblement des résultats obtenus lors de l'analyse. A nouveau, l'incertitude est comparable à celle obtenue par la méthode du bootstrap, et à nouveau, pour cela nous avons dû utiliser des paramètres de dispersion faibles.

Pour ce qui est de la distribution du montant total des réserves, la simulation donne le résultat suivant :

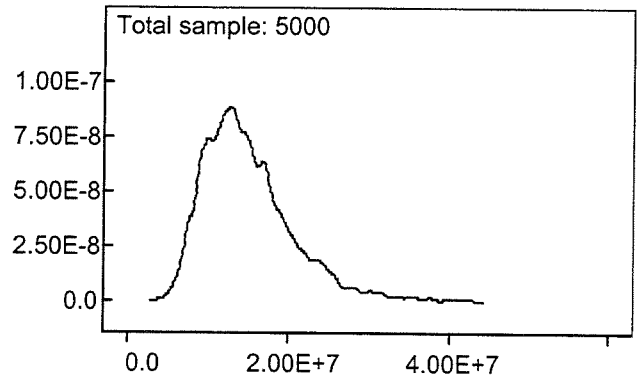


fig. 4.2 : Distribution du montant total des réserves

Cette fois, contrairement à la méthode du bootstrap, les réserves ne sont plus centrées autour de la moyenne. Le quantile à 2.5% est proche de celui obtenu par le bootstrap. Mais la moyenne et le quantile à 97.5% sont sensiblement inférieurs.

Application à la méthode Bornhuetter-Ferguson

Approche proposée

Nous considérons ici une intervention sur le niveau de chaque ligne, en utilisant la méthode de Bornhuetter-Ferguson. Pour cela, nous allons travailler sur deux exemples. Le premier propose d'utiliser des variances très petites pour les distributions a priori des paramètres de chaque ligne. Le second utilise des variances plus grandes.

Dans la suite de ce paragraphe, nous allons montrer que le premier exemple permet de retrouver des résultats similaires à la méthode Bornhuetter-Ferguson, alors qu'une variance plus grande donne un résultat entre Chain-Ladder et Bornhuetter-Ferguson.

Dans le modèle utilisé, nous avons choisi des distributions Gamma pour décrire les ultimes. On a alors :

$$Ult_i | \alpha_i, \beta_i \sim \Gamma(\alpha_i, \beta_i) \text{ indépendantes.}$$

Comme $Var(Ult_i) = \mathrm{E}(Ult_i) / \beta_i$, choisir une grande valeur de β_i signifie que nous somme très sûrs de la moyenne sélectionnée.

On peut montrer que la moyenne des paiements incrémentaux du modèle peut s'écrire de la façon suivante :

$$Z_{ij} \cdot (\lambda_j - 1) \cdot C_{i,j-1} + (1 - Z_{ij}) \cdot (\lambda_j - 1) \cdot Ult_i \frac{1}{\lambda_j \lambda_{j+1} \cdots \lambda_n}$$

où :

$$Z_{ij} = \frac{\sum_{k=1}^{j-1} y_k}{\beta_i \phi + \sum_{k=1}^{j-1} y_k}$$

les y_k étant définis comme étant les paramètres associés aux colonnes lorsque l'on considère un modèle de Poisson sur les paiements incrémentaux.

Cette formule peut être vue comme une formule de crédibilité. Il s'agit de faire un arbitrage entre Chain-Ladder (première partie de la formule) et Bornhuetter-Ferguson (deuxième partie de la formule). Aussi, nous pouvons influer sur cette relation au travers de β_i. Plus β_i est grand et plus l'on se rapproche de Bornhuetter-Ferguson.

Il est également intéressant de considérer l'estimation des paramètres des colonnes. Pour la méthode Bornhuetter-Ferguson, il s'agit simplement de reprendre les facteurs de Chain-Ladder. Cependant, étant donné que nous nous plaçons dans un contexte stochastique, Verral propose d'utiliser une loi binomiale négative avec surdispersion appliquée à :

$$C_{ij} \left| C_{1j}, C_{2j}, \ldots, C_{i-1,j}, Ult, \phi \right.$$

Résultats obtenus

Lors de la mise en pratique, nous avons choisi les mêmes ultimes que ceux choisis pour la méthode Bornhuetter-Ferguson.

Reproduction de la méthode Bornhuetter-Ferguson

Afin de faciliter la comparaison avec les résultats obtenus par la méthode Bornhuetter-Ferguson, nous avons sélectionnés le même niveau des ultimes a priori.

Comme nous l'avons vu dans le paragraphe précédent, choisir une petite variance permet de reproduire les résultats de la méthode Bornhuetter-Ferguson. Comme l'a fait Verral, nous avons choisi une variance de 1 000 pour chaque ultime.

Voici alors les résultats obtenus pour 5 000 simulations :

	année	1998	1999	2000	2001	2002	Ensemble
	moyenne	1 678 000	1 961 000	2 755 000	4 178 000	4 816 000	15 390 000
IC 95%	sup	3 265 000	3 299 000	4 357 000	5 941 000	6 738 000	21 061 000
	inf	641 000	929 800	1 529 000	2 741 000	3 244 200	11 040 000

tab. 4.10 : Reproduction de Bornhuetter-Ferguson

Nous obtenons donc des résultats très similaires à ceux obtenus avec la méthode Bornhuetter-Ferguson. En effet, l'écart des réserves observé est inférieur à 1% pour chaque année. Cette fois, l'intervalle de confiance obtenu est plus faible que pour les autres méthodes. En particulier, le quantile à 95% est nettement moins grand. Seulement, pour obtenir un tel résultat, nous avons supposé que nous sommes très confiants dans l'estimation de l'ultime. C'est pourquoi il est également intéressant d'étudier les résultats obtenus avec une confiance moindre dans nous estimateurs.

Modèle bayésien pour la méthode Borhuetter-Ferguson

Ce modèle utilise une variance plus grande, et donne des résultats entre Chain-Ladder et Bornhuetter-Ferguson. Avec une variance 1 000 fois plus grande que précédemment, nous obtenons le résultat suivant :

	année	1998	1999	2000	2001	2002	Ensemble
	moyenne	1 869 000	1 933 000	2 626 000	4 475 000	4 920 000	15 890 000
IC 95%	sup	3 915 000	3 653 000	4 747 000	7 358 000	7 885 000	23 000 000
	inf	608 200	761 100	1 177 000	2 424 000	2 724 000	10 061 000

tab. 4.11 : Modèle bayésien

Le résultat reste beaucoup plus proche du modèle de Bornhuetter-Ferguson que du modèle Chain-Ladder. Pour 2002, nous sommes en effet très loin d'obtenir les réserves de près de 8 millions obtenues avec la méthode Chain-Ladder. Cependant, comme nous pouvions le prévoir, augmenter l'incertitude tend à augmenter les réserves associées à 2002.

En ce qui concerne les intervalles de confiance associés à ces prévisions, on remarque qu'ils augmentent, mais restent tout de même inférieurs à ceux trouvés jusqu'à présent avec les autres méthodes stochastiques.

Voici la distribution du montant total des réserves associé :

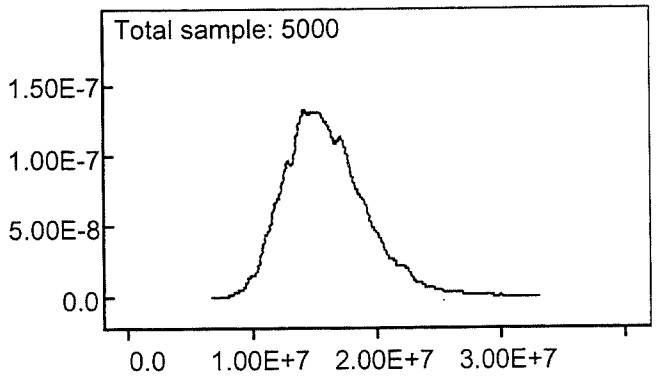


fig. 4.3 : Distribution du montant total des réserves

Comme lors de l'introduction d'avis d'expert à la méthode Chain-Ladder, la distribution n'est pas symétrique autour de la moyenne.

Il est également intéressant de noter qu'il nous a été impossible d'augmenter la variance associée aux ultimes. En effet, dans ce cas, les méthodes de simulation ne convergent plus. En particulier, dans son article, Verral retrouvait des réserves proche de Chain-Ladder avec une variance 10 fois plus grande que le modèle bayésien. Ceci a été impossible à vérifier à partir de nos données.

Néanmoins, les résultats proposés par ce modèle, que ce soit pour l'application à la méthode Chain-Ladder ou à la méthode Bornhuetter-Ferguson sont très intéressants. Ils s'adaptent relativement bien à nos données et permettent d'obtenir des résultats cohérents. De plus, l'introduction d'avis d'experts est très répandue lors de l'analyse des réserves. Il est donc important de les introduire dans un cadre stochastique.

4.3 Modélisation de la corrélation existant au sein du triangle

Présentation du modèle

Dans son modèle « Forecasting general insurance liabilities » (2004), Piet de Jong propose de prendre en compte les corrélations existant à l'intérieur du triangle de développement. Pour cela, il utilise de méthodes issues de l'étude des séries temporelles.

On s'intéresse à la réserve de chaque année de souscription

$$R_i = C_{i,n} - C_{i,n-i+1} = C_{i,n}\left(e^{g_i} - 1\right) \quad \Rightarrow \quad g_i = \ln\left(\frac{C_{i,n}}{C_{i,n-i+1}}\right)$$

où g_i est le taux continûment composé d'accroissement des réclamations pour l'année i.

L'ensemble des réserves s'écrit alors :

$$R = \sum_{i=2}^{n} R_i = \sum_{i=2}^{n} C_{i,n}\left(e^{g_i} - 1\right)$$

Dans ce modèle, on cherche à construire analytiquement la moyenne, l'écart type et les corrélations de la distribution jointe des futurs taux d'accroissement. Pour cela, des séries temporelles sont utilisées. Ce type d'approche présente l'avantage de permettre l'utilisation d'un grand nombre de modèles disponibles.

La première remarque que l'on doit faire sur ce modèle est que, bien que Chain-Ladder reste un benchmark incontournable pour l'estimation des réserves, les réserves espérées ne sont pas les mêmes que pour Chain-Ladder.

Facteurs de développement et modèle de base

Les facteurs de développement

Les différents modèles définis par Piet de Jong sont formulés en termes de facteurs de développement. On définit le taux d'accroissement de l'année de survenance i et d'année de développement j par :

$$\delta_{i,j+1} = \ln\left(\frac{C_{i,j+1}}{C_{i,j}}\right)$$

On définit également les facteurs de développement de la première colonne par $\delta_{i1} = \ln(C_{i1})$.

Ces taux peuvent être mis en parallèle avec les facteurs de développement individuels de Chain-Ladder puisque l'on a l'approximation suivante :

$$\frac{C_{i,j+1}}{C_{i,j}} \approx 1 + \delta_{i,j+1},$$

cette approximation étant valable pour les facteurs de développement individuels proches de 1.

Le taux d'accroissement pour l'année i s'écrit donc :

$$g_i = \delta_{i,n-i+2} + \cdots + \delta_{i,n}$$

Présentation du modèle de base

Ce modèle permet d'illustrer les caractéristiques d'ensemble de l'approche étudiée par l'auteur.

Il est supposé que :

$$\delta_{ij} \sim \left(\mu_j, \sigma_j^2\right)$$

Hertig (1985) avait déjà étudié un tel modèle. Seulement, il n'avait pas étudié les facteurs de développement de la première colonne et leur interaction avec les années de développement futures.

Dans ce modèle, la tendance et l'écart type des paiements agrégés fluctuent entre les années de développement mais sont les mêmes entres les années de survenance. Aussi, ce modèle implique que les paiements cumulés sont corrélés à l'intérieur des années de survenance, mais pas entre années de survenance. On suppose généralement que la loi suivie par les facteurs de développements est la loi normale.

L'estimateur du taux d'accroissement de l'année i et la variance de l'erreur de prédiction associée sont :

$$\hat{g}_i = \mu_{n-i+2} + \cdots + \mu_n \text{ et } v_i = \sigma_{n-i+2}^2 + \cdots + \sigma_n^2$$

En supposant la normalité des facteurs de développements, la valeur prédite des réserves et le coefficient de variation associé sont alors :

$$\hat{C}_{i,n-1} = C_{i,n} e^{\hat{g}_i + v_i^2/2} \text{ et } \sqrt{e^{v_i^2} - 1}$$

En pratique, le couple $\left(\mu_j, \sigma_j^2\right)$ est inconnu et à estimer. Estimer les μ_j entraîne une estimation de la corrélation entres les différences années de survenance puisque ces derniers sont utilisés dans les prévisions de plusieurs années de survenance à la fois.

Les estimateurs du couple $\left(\mu_j, \sigma_j^2\right)$ sont la moyenne et la variance empirique des données :

$$\hat{\mu}_j = \frac{1}{n-j} \sum_{i=1}^{n-j} \delta_{ij}$$

$$\hat{\sigma}_j = \sqrt{\frac{1}{n-j-1} \sum_{i=1}^{n-j} \left(\delta_{ij} - \hat{\mu}_{ij}\right)^2}$$

Aussi, la variance du processus est estimée par :

$$\hat{v}_i = \hat{\sigma}_{n-i+2}^2 + \cdots + \hat{\sigma}_n^2$$

A cette erreur, il faut ajouter l'erreur d'estimation. Comme $\hat{\mu}_j$ a pour variance $\sigma_j^2/(n-j)$, on en déduit la variance de l'erreur d'estimation:

$$\frac{\sigma_{n-i+2}^2}{i-1}+\cdots+\frac{\sigma_n^2}{1}$$

On peut vérifier que ce modèle est approprié en utilisant les résidus standardisés des facteurs de développements :

$$z_{ij}=\frac{\delta_{ij}-\hat{\mu}_j}{\hat{\sigma}_j}$$

Ces derniers sont approximativement distribués selon une loi normale standard. Pour vérifier cette hypothèse, il est possible d'effectuer des tests graphiques ou des tests formels.

Mise en pratique

Voici les facteurs de développement et les estimateurs obtenus :

Development Factors

Reporting Year	Development year 1	2	3	4	5	6
1997	8,46	3,80	2,17	0,34	0,29	0,28
1998	12,11	1,98	1,01	0,30	0,19	
1999	11,63	2,32	0,70	0,26		
2000	12,56	1,24	0,78			
2001	13,07	0,82				
2002	12,63					

Modèle de base						
mu	11,743	2,030	1,165	0,301	0,239	0,276
sigma	1,533	1,029	0,592	0,034	0,053	0,000

fig.. 4.12: Facteurs de développement et modèle de base

Le modèle de base donne alors des résultats extravagants pour les deux dernières années :

année	1998	1999	2000	2001	2002	Ensemble
moyenne	1 871 092	2 006 236	2 704 027	8 272 865	33 810 332	48 664 552

tab. 4.13 : Résultats du modèle de base

Ces résultats sont dus principalement aux premières années de développement de 1997. Etant donné que nous ne les jugeons pas comme étant révélateurs des facteurs de développement actuels, nous écartons 1997 de l'estimation des trois premiers facteurs de développement. Ce phénomène peut être détecté par l'utilisation des résidus standardisés. En effet, les facteurs des trois premières années de développement associés à 1997 sont les seuls facteurs du triangle dont

la probabilité d'occurrence devrait être inférieure à 10% si le modèle était valide (tableau 2.10 en annexe).

En ne prenant pas en compte ces trois facteurs, on obtient les résultats suivants :

année	1	2	3	4	5	6	Ensemble
mu	12,398	1,588	0,829	0,301	0,239	0,276	
sigma	0,501	0,609	0,140	0,034	0,053	0,000	
moyenne	0	1 871 092	2 006 236	2 704 027	4 586 415	9 122 711	20 290 481

tab. 4.14 : Correction des facteurs

Les réserves obtenues semblent beaucoup plus cohérentes. Seule la provision de 2002 reste « excessive ». Nous allons essayer de corriger cela par l'introduction de corrélations au sein du triangle.

Modélisations des corrélations à l'intérieur des triangles

Corrélations entre années de développement

On peut réécrire le modèle de base de la manière suivante :

$$\delta_{ij} = \mu_j + h_j \varepsilon_{ij}$$

où

$$h_j = \frac{\sigma_j}{\sigma_1} \text{ et } \varepsilon_{ij} = \frac{\delta_{ij} - \mu_j}{h_j} \sim \left(0, \sigma_1^2\right)$$

L'auteur propose d'introduire un lien entre les facteurs de développement des deux premières années. Pour cela, il modifie le modèle du second facteur de la manière suivante

$$\delta_{i2} = \mu_2 + h_2\left(\varepsilon_{i2} + \theta \cdot \varepsilon_{i1}\right)$$

Il s'agit donc d'un modèle de séries temporelles MA (1). La corrélation entre les facteurs de développements des deux premières années est alors :

$$r = \frac{\theta}{\sqrt{1+\theta^2}}$$

Cette corrélation peut être estimée à partir des estimateurs des moments, ce qui nous permet d'en déduire une valeur estimée de θ. On peut également en déduire les facteurs h.

D'autres θ peuvent être introduits. Cependant, en pratique, seule la corrélation entre les deux premiers facteurs de développement a une influence significative sur les prévisions.

Mise en pratique

Si l'on observe les facteurs de développement pris en compte dans l'évaluation des deux premières années, on remarque qu'ils sont négativement corrélés. Voici en effet le graphique des résidus standardisés.

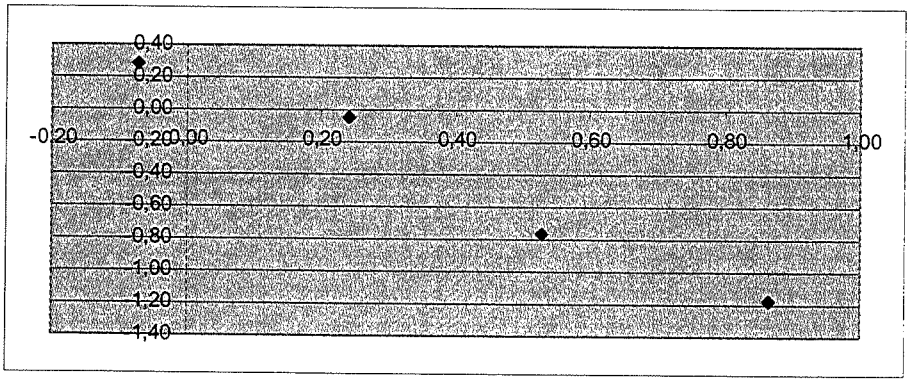


fig 4.4 : Relation entre les deux premiers facteurs de développement

Nous avons mis en pratique le modèle proposé au paragraphe précédent pour prendre en compte cette corrélation.

Les résultats obtenus avec 5 000 simulations sont alors les suivants

année	1	2	3	4	5	6	Ensemble
moyenne	0	1 871 092	2 006 246	2 861 360	4 840 828	7 524 391	19 157 234

tab. 4.15 : Introduction de corrélations

Les résultats obtenus nous conduisent à une réduction de la prévision de 2002. Cependant, la réserve associée à cette année, bien que inférieure à Chain-Ladder, semble tout de même excessive.

Aussi, bien que cette approche semble intéressante, nous remarquons qu'elle ne s'adapte pas toujours très bien aux données. D'une part, le modèle de base a du être préalablement retraité pour permettre de ne pas obtenir des réserves non cohérentes. D'autre part, l'introduction de corrélations ne permet pas de réduire sensiblement les réserves.

4.4 Résultats du modèle de provisionnement individuel

Mise en œuvre du programme

Nous avons mis en œuvre la démarche proposée dans la troisième partie.

Voici les étapes nécessaires à la programmation du modèle :

Il faut dans un premier temps programmer la fonction des paiements, qui est fonction du temps écoulé depuis la déclaration. En effet, la probabilité associée à la nature de chaque flux dépend du nombre d'années écoulées depuis la déclaration. Aussi, le programme doit :

- Simuler un tirage portant sur la nature des flux.
- Puis simuler le montant associé au flux sélectionné.

Il s'agit ensuite de programmer l'algorithme principal.

Pour chaque simulation, il faut :

- Simuler la date de clôture de chacun des sinistres, à partir de la fonction paramétrique déterminée lors de l'analyse de survie.
- Simuler une date de paiement. Si aucun paiement n'a encore eu lieu, le prochain paiement est distribué selon une loi Gamma fixée. Sinon, la moyenne et la variance sont fonction du temps écoulé depuis le précédent paiement.
- Simuler un paiement associé, si la date de paiement est inférieure à la date de clôture, grâce à la fonction mise en place précédemment. Sinon, clore le sinistre.
- Renouveler le deux dernières étapes tant que le sinistre n'est pas clos.

Résultats des simulations

Nous avons effectué 5000 simulations de ce modèle. Afin de présenter nos résultats et de pouvoir les comparer avec les autres modèle déjà présentés, nous avons agrégé nos simulations dans un triangle en ne prenant pas en compte les recours.

Voici le triangle moyen obtenu :

Cumulative Claims Paid								
Reporting Year	Development year							
	1	2	3	4	5	6	Ult	Res
1997	4 740	210 938	1 850 617	2 608 283	3 495 492	4 607 766	4 607 766	0
1998	181 773	1 314 395	3 609 713	4 881 558	5 880 198	7 058 846	7 058 846	1 178 648
1999	112 302	1 137 729	2 286 295	2 962 406	3 771 592	4 555 890	4 555 890	1 593 484
2000	283 555	977 041	2 131 768	3 222 565	4 265 317	5 153 953	5 153 953	3 022 185
2001	474 694	1 078 972	2 473 683	3 952 409	5 169 987	6 157 030	6 157 030	5 078 058
2002	304 668	1 458 920	2 629 538	3 659 999	4 489 178	5 152 305	5 152 305	4 847 637
							32 685 790	15 720 012

tab. 4.7 : Résultats des simulation pour le modèle individuel

Les réserves obtenues par simulation du modèle individuel sont moindres que celles obtenues avec la méthode Chain-Ladder. La principale différence provient des réserves de l'année 2002, pour lesquelles les réserves sont beaucoup moins importantes.

Si l'on compare les résultats avec ceux obtenus par analyse des facteurs de développement, on remarque que l'on a tendance à payer un peu moins pour 2002, et un peu plus pour 2000. L'ultime obtenu pour ces deux années (5 millions et 5.3 millions) est plutôt plus satisfaisant que celui obtenu par analyse des facteurs de développement (4.5 et 5.5 millions) au regard des paiements de la première année.

Par rapport à la méthode Bornhuetter-Ferguson, les années les plus anciennes (1998 et 1999) engendrent moins de réserves, alors que 2000 en engendre plus. S'il nous apparaît normal que l'ultime associé à 1998 soit de 7 millions, celui de 1999 est peut être un peu sous évalué dans le modèle individuel.

De manière plus générale, on remarque que l'on a tendance à payer plus rapidement que la méthode Chain-Ladder au départ, et moins par la suite. En particulier, les paiements cumulés les plus importants après deux ans de développement seraient ceux de 2002. Or ceci n'a pas de raison d'être si l'on se fie uniquement aux données agrégées.

Cette méthode nous permet également d'obtenir l'incertitude liée aux réserves du modèle :

	année	1998	1999	2000	2001	2002	Ensemble
	moyenne	1 178 648	1 593 484	3 022 185	5 078 058	4 847 637	15 720 012
IC 95%	Sup	3 233 246	3 854 537	5 876 892	8 953 159	8 708 173	22 155 456
	Inf	118 470	230 556	916 583	2 176 262	1 995 133	10 132 083

tab. 4.16 : Intervalles de confiance pour le modèle individuel

Le modèle individuel nous donne un intervalle de confiance réduit par rapport aux méthodes stochastiques classiques. La borne inférieure de l'intervalle de confiance reste sensiblement la même, mais la borne supérieure est fortement réduite. Nous avions également remarqué ce phénomène lors de la mise en place des nouvelles méthodes de provisionnement, comme l'introduction d'avis d'experts ou le prise en compte de corrélations au sein du triangle. Seul le modèle permettant l'introduction des estimateurs de Bornhuetter-Ferguson dans un cadre stochastique présente une incertitude du même ordre.

Voici la distribution du montant total des réserves obtenue :

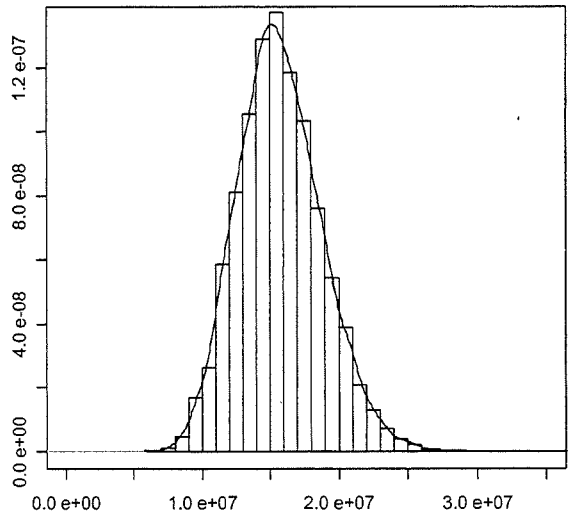


fig. 4.5 : Estimation de la distribution du montant total des réserves

Comme pour la méthode stochastique introduisant l'avis d'expert et la prise en compte de corrélations au sein du triangle, on peut noter que la distribution est légèrement asymétrique.

Les résultats obtenus par le modèle individuel nous apportent donc des résultats intéressants. Les ultimes obtenus lors des simulations nous semblent cohérents. Les intervalles de confiance associés nous apparaissent être plus adaptés que ceux obtenus par les autres modèles. Nous n'observons pas les limites des modèles classiques. On peut en particulier penser aux intervalles de confiance trop importants du modèle de Mack avec distribution normale, ou un quantile à 2.5% nul obtenu pour 2002 par la méthode du bootstrap.

Néanmoins, ce modèle reste sujet à une erreur plus importante et nécessite d'être corroboré par des méthodes traditionnelles.

4.5 Récapitulatif des résultats des différents modèles

De nombreux modèles ont été étudiés dans cette partie. Afin de mieux observer quels sont les effets de chacune des méthodes, il nous a semblé intéressant de récapituler les principaux résultats obtenus. Nous avons donc comparé les résultats moyens ainsi que les intervalles de confiance associés. Puis nous avons mis en parallèle les distributions de l'ensemble des réserves obtenues par les différents modèles stochastiques.

Comparaison des moyennes obtenues

Le principal résultat des différents modèles est l'obtention de la moyenne des réserves prévues. Voici le tableau comparatif que nous avons obtenu :

Année	1998	1999	2000	2001	2002	Ensemble
Chain-Ladder	1 871 092	1 925 887	2 614 473	5 440 206	7 913 296	19 764 953
London-Chain	1 871 092	1 904 664	2 590 249	5 491 631	8 390 847	20 248 482
Analyse	1 411 248	1 635 890	2 332 903	5 053 456	5 236 463	15 669 959
Bornhuetter-Ferguson	1 689 737	1 969 897	2 754 257	4 172 463	4 814 633	15 400 987
Mack	1 871 092	1 925 887	2 614 473	5 440 206	7 913 296	19 764 953
Bootstrap	1 878 463	1 929 746	2 597 262	5 444 237	7 506 241	19 355 950
Avis Experts CL	1 180 000	1 510 000	2 216 000	4 992 000	5 176 000	15 070 000
Avis Experts BF	1 869 000	1 933 000	2 626 000	4 475 000	4 920 000	15 890 000
Corrélations	1 871 092	2 006 246	2 861 360	4 840 828	7 524 391	19 157 234
Modèle individuel	1 178 648	1 593 484	3 022 185	5 078 058	4 847 637	15 720 012

tab. 4.17 : Tableau comparatif des réserves obtenues

On remarque qu'il y a deux catégories de modèles formés. D'une part, il y a les modèles présentant des résultats similaires à la méthode Chain-Ladder classique. Et d'autre part, il y a les modèles résultant d'une analyse plus fine de la sinistralité.

La première catégorie de modèles prévoit des réserves de l'ordre de 20 millions. On y trouve, outre le modèle Chain-Ladder, la méthode London-Chain, le modèle de Mack, le bootstrap et le modèle permettant l'introduction de corrélations. Pour le modèle de Mack, on retrouve exactement les mêmes réserves que pour la méthode Chain-Ladder. La méthode London-Chain et celle basée sur des corrélations présentent elles aussi des résultats similaires, et ce malgré qu'elles utilisent une modélisation plus complexe des facteurs de développement. Or les réserves proposées par la méthode Chain-Ladder nous ont semblé excessives, surtout en ce qui concerne

les réserves associées aux années les plus récentes. Ces modélisations plus complexes ne permettent donc pas de confirmer ces impressions. La méthode London-Chain présente même des réserves supérieures à Chain-Ladder.

Ce premier groupe de modèles ne nous permet donc pas de prendre en compte le changement de cadence de développement observé entre 1997 et les autres années.

Les réserves associées aux autres modèles sont sensiblement moindres (environ 16 millions). Ces modèles se basent sur une analyse plus fine de la cadence de développement et des particularités liées au portefeuille étudié. Les réserves calculées pour les années les plus récentes semblent plus cohérentes, surtout en ce qui concerne 2002.

Pour trouver de tels résultats, nous avons dû procéder à une analyse des facteurs de développement, ou prévoir une sinistralité ultime a priori. Seul le modèle individuel ne fait pas ce genre d'hypothèses. L'introduction d'un avis d'expert semble donc nécessaire pour modéliser des agrégations de phénomènes individuels complexes.

Les résultats de ce second groupe de modèle ne sont pas les mêmes années par année.

La méthode Bornhuetter-Ferguson prévoit en effet plus de réserves pour 1998 et 1999. Le facteur associé à la dernière année de développement, qui nous avait semblé trop important, est en effet conservé lors du calcul des réserves associées à ce modèle.

Les résultats du modèle individuels sont pour leur part relativement proches de ceux obtenus par analyse des facteurs de développement. Il convient tout de même de noter que la sinistralité associée à l'année 2000 est plus importante avec le modèle individuel qu'avec l'analyse des facteurs, alors qu'à l'inverse celle associé à 2002 est plus faible.

Un retraitement des modèles classiques doit donc être effectué pour ne pas diverger. L'introduction du modèle individuel nous permet ici de valider les avis d'experts sélectionnés.

Tableau comparatif des intervalles de confiance

Il est également intéressant de connaître l'incertitude liée aux différents modèles. Voici les résultats obtenus pour des intervalles de confiance à 95% :

Année		1998	1999	2000	2001	2002	Ensemble
CL, BF							
	Inf	N/A	N/A	N/A	N/A	N/A	N/A
	sup	N/A	N/A	N/A	N/A	N/A	N/A
Mack							
	Inf	1 316 851	1 099 889	1 687 641	1 015 258	871 248	6 383 95
	sup	2 581 367	3 139 251	3 872 499	17 281 526	31 226 891	47 169 029
Bootstrap							
	Inf	957 144	876 173	1 076 055	1 653 816	0	9 914 484
	sup	3 032 994	3 216 372	4 298 090	9 318 492	16 280 733	29 200 898
Avis Experts CL							
	Inf	69 660	439 100	840 400	1 915 000	898 20	9 503 000
	Sup	3 557 000	3 461 000	4 446 000	10 400 000	14 300 000	28 790 000
Avis Experts BF							
	Inf	608 200	761 100	1 177 000	2 424 000	2 724 000	10 061 000
	Sup	3 915 000	3 653 000	4 747 000	7 358 000	7 885 000	23 000 000
Corrélations							
	Inf	1 869 008	1 998 556	2 749 386	4 607 401	6 378 164	17 945 650
	Sup	1 869 008	2 125 609	2 976 047	5 080 121	8 819 048	20 489 936
Modèle individuel							
	Inf	118 470	230 556	916 583	2 176 262	1 995 133	10 132 083
	sup	3 233 246	3 854 537	5 876 892	8 953 159	8 708 173	22 155 456

tab. 4.17 : Tableau comparatif des intervalles de confiance obtenus

Les méthodes déterministes ne permettent pas d'obtenir de facteurs de développement. Le modèle de Mack nous montre néanmoins que l'incertitude existant autour du la méthode de Chain-Ladder classique est très élevée. Le modèle Chain-Ladder classique ne doit donc pas être utilisé en tant que tel mais des avis d'experts doivent être introduits pour ne pas diverger.

On préférera le modèle du Bootstrap au modèle de Mack. Cependant, il montre lui aussi ses limites pour l'année 2002, année pour laquelle l'intervalle de confiance obtenu est inexploitable. L'incertitude liée au modèle corrélé est moindre, mais pour obtenir de tels résultats, nous avons dû ne pas prendre en compte l'année 2002.

Le modèle permettant l'introduction d'avis d'experts sur les facteurs de développement présente un intervalle de confiance du même ordre que le bootstrap. Etant donné que la moyenne associée à ce modèle est sensiblement inférieure, la distribution est donc très asymétrique.

Pour le modèle lié à Bornhuetter-Ferguson comme pour le modèle individuel, la borne supérieure est plus petite que le modèle du Bootstrap. Ces deux distributions sont donc moins asymétriques que le modèle précédent.

Comparaison des distributions

Enfin, il est intéressant de comparer les différentes distributions obtenues. Voici un graphique comparatif du montant total des réserves :

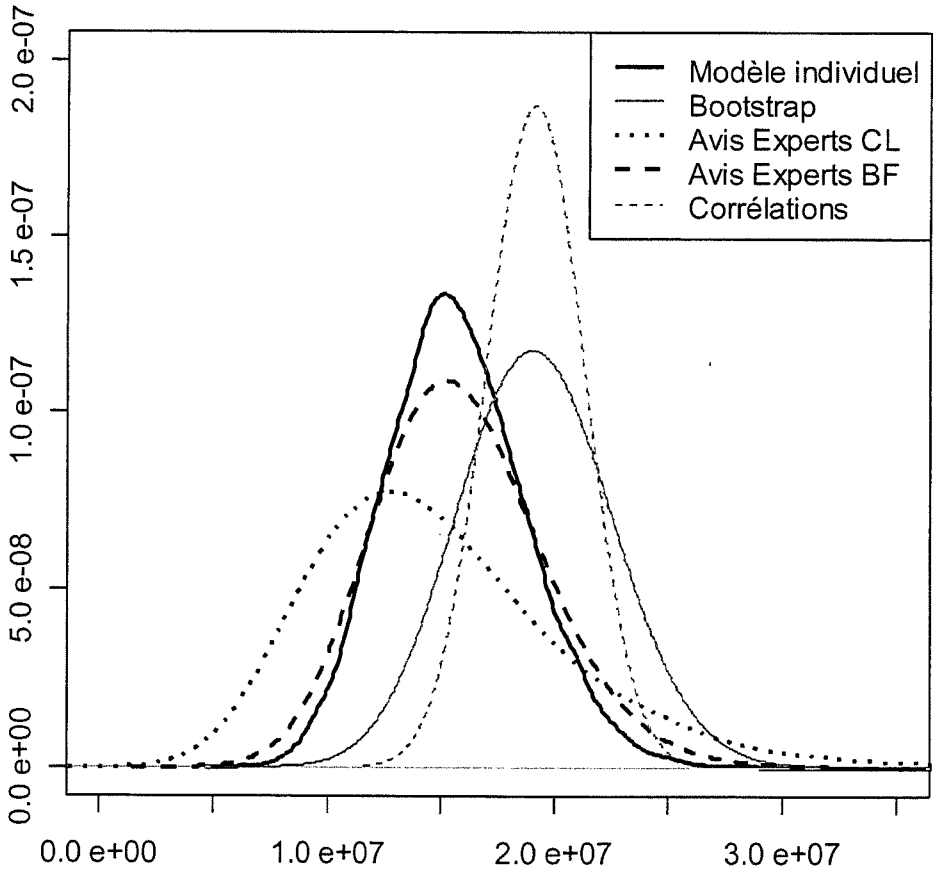


fig. 4.6 : Comparaison des distributions du montant total des réserves

Les distributions du modèle du bootstrap et du modèle permettant l'introduction de corrélations sont très différentes des autres. Elles appartiennent en effet à la catégorie de modèles présentant des réserves du même ordre que le modèle Chain-Ladder classique.

Les autres modèles sont comparables en moyenne. Cependant, si l'on compare les deux méthodes utilisant l'introduction d'avis d'expert, celle lié à la méthode Chain-Ladder présente un queue à droite beaucoup plus épaisse que celle liée à la méthode Bornhuetter-Ferguson.

Les résultats du modèle individuel sont pour leur part très proche au modèle lié à la méthode Bornhuetter-Ferguson.

Les résultats présentés par le modèle individuel sont donc très intéressants. Ils nous ont permis d'obtenir un ensemble de résultats cohérents, ce qui n'est pas le cas des toutes les autres méthodes. En effet, pour ne pas diverger, nous avons remarqué qu'il est nécessaire de modifier les modèles classiques par l'introduction d'avis d'experts. Ces opinions sont nécessaires pour prévoir des résultats provenant de l'agrégation de phénomène individuels complexes. De la même manière, si l'on veut utiliser des modèles stochastiques ne divergeant pas, l'utilisation d'un avis d'expert permet d'obtenir un ensemble de résultats plus cohérents. Le modèle individuel introduit dans ce mémoire permet de valider l'avis d'expert en s'attachant à modéliser les phénomènes individuels complexes.

CONCLUSION

Dans cette étude, nous nous sommes intéressés aux provisions techniques. Les méthodes classiques utilisent des données agrégées observées à périodicité constante. L'approche étudiée ici est différente : il s'agit d'étudier des données détaillées et de revenir au sinistre individuel. Pour cela, il est nécessaire de modéliser les différentes étapes de la vie de chaque sinistre. Ce type de modélisation présente de nombreux avantages : mesure d'incertitude, possible prise en compte de la réassurance non proportionnelle,... Pour cette approche, nous nous sommes intéressés aux sinistres ayant déjà été déclarés. Pour établir notre modèle, nous avons utilisé des données d'un portefeuille RC professionnelle ayant une durée de développement longue et de nombreux paiements.

Il y a alors plusieurs phénomènes à analyser : le processus de date des flux, les montants associés le cas échéant et la date de clôture des sinistres. En ce qui concerne la date de clôture, l'utilisation de l'analyse de survie parait être une solution intéressante. En effet, elle permet de prendre en compte les sinistres non clos dans la modélisation. Pour étudier le temps entre la déclaration et la survenance d'un flux, nous avons utilisé des modèles linéaires généralisés (GLM). Enfin, les données dont nous disposions nous permettaient de connaître la nature de chaque flux (paiement en principal, honoraire, frais ou recours). Aussi il est possible d'étudier de manière paramétrique le montant des flux selon leur nature.

Nous avons cherché à comparer les résultats du modèle individuel avec un grand nombre de modèles classiques ou plus nouveaux. De manière générale, les modèles classiques s'adaptent relativement mal à ce type de données. Les sinistres de RC professionnelle sont en effet des sinistres ayant une durée de vie longue et une gestion complexe. Des résultats incohérents peuvent apparaître avec des méthodes pourtant fréquemment utilisées.

Nous avons remarqué que les modèles classiques ne doivent pas être utilisés sans introduction d'avis d'experts. En effet, les modèles étudiés divergent pour les années les plus récentes, qu'ils soient déterministes ou stochastiques.

Il ressort de notre analyse que l'utilisation du modèle individuel peut être une alternative intéressante aux modèles les plus classiques. En effet, ces derniers nous donnent le plus souvent des réserves incohérentes, ou des incertitudes trop importantes. Même la prise en compte des corrélations existant au sein du triangle semble trop peu robuste pour étudier ce type de données.

Le seul modèle stochastique agrégé qui nous a semblé être intéressant est un modèle permettant l'introduction d'avis d'expert. Or l'utilisation du modèle individuel peut permettre de valider le retraitement effectué. Il permet en effet de valider l'avis d'expert en s'attachant à la modélisation de phénomènes individuels complexes.

En pratique, l'utilisation d'un modèle individuel est rendue difficile par le manque de données disponibles. En effet, il n'est pas toujours possible de disposer d'une base suffisamment de données détaillées. Seulement, disposer de telles données sera une nécessité avec la nouvelle norme comptable IFRS04 phase 2.

Enfin, au-delà de l'aspect statistique, des tels modèles nécessitent une connaissance des affaires sous-jacentes de manière à anticiper de la manière la plus probable l'évolution de l'environnement jurisprudentiel et les changements de procédure de gestion au sein des compagnies.

Dans le modèle présenté dans ce mémoire, nous avons choisi de ne pas introduire les sinistres survenus mais non déclarés. Néanmoins, une évolution intéressante de ce modèle serait de les prendre en compte afin de définir un modèle global permettant de calculer l'ensemble des réserves, IBNYR compris.

BIBLIOGRAPHIE

Arjas E., Haastrup S. (1996) Claims reserving in continuous time; a non parametric bayesian approach. *Astin Bulletin 26 (2), p. 139-164*

Charpentier A (2004). Ratemaking using GLM and GAM. *Lectures notes, 3rd conference in actuarial science and finance in Samos*

De Alba, E. (2002) Bayesian Estimation of Outstanding Claims Reserves, *NAAJ, Vol. 6 (4), p. 1-20*

De Alba, E. (2002) Claims reserving when there are negative values in the development triangle. *Working Paper*

De Jong, P. (2004) Forecasting general insurance liabilities. *Research paper, 2004/03, Macquarie University*

England P.D., Verrall, R.J. (1999). Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics 25, 281-293*

England P.D., Verral R.J. (2002) Stochastic claims reserving in general insurance. *Institute of actuaries and Faculty of actuaries, p.1-76*

Hayne, R.M., Estimating and Incorporating Correlation in Reserve Variability, *CAS forum, fall 2004*

Haastrup S. (1997) Some fully bayesian micro models for claims reserving. *Ph.D. thesis*

Haberman S, Renshaw A E (1996) Generalized linear models and actuarial science. *The statistician, 45 (4), p.407-436*

Hertig, J. (1985) A statistical approach to IBNR-reserves in marine reinsurance. *Astin Bulletin, 15, p.171-183*

Hesselager, O. (1994) A Markov Model for Loss Reserving. *Astin Bulletin, 24 (2), p.183-193*

Jewell, W S. (1989) Predicting IBNYR events and delays. *Astin Bulletin, 19 (1),p. 25-55*

Kerley C., Kirschner G.S., Isaacs B. (2002). Two Approaches to Calculating Correlated Reserve Indications Across Multiple Lines of Business. *CAS forum, fall 2004*

Mack, T. (1993). Distribution-free calculation of the standard error of chain-ladder

reserve estimates. *ASTIN Bulletin, 23, 213-225.*

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

Regazzoni Y., Sander J. (1997) Les provisions techniques : une approche par simulation. *Bulletin Français d'Actuariat, 1 (2)*

Verral R.J. (2004) Obtaining predictive distributions for reserves which incorporate expert opinion. *CAS forum, fall 2004*

ANNEXES

1- Statistiques descriptives

Statistiques sur la nature des flux

— Influence du nombre de paiements déjà effectués

Le nombre de paiements déjà effectués risque d'avoir une influence sur le déroulement à venir du sinistre. Ainsi, un sinistre très fractionné sera vraisemblablement la marque d'un dossier délicat, faisant peut-être l'objet d'un jugement, et requérant une multiplication d'avis d'experts et des frais de gestion plus lourds.

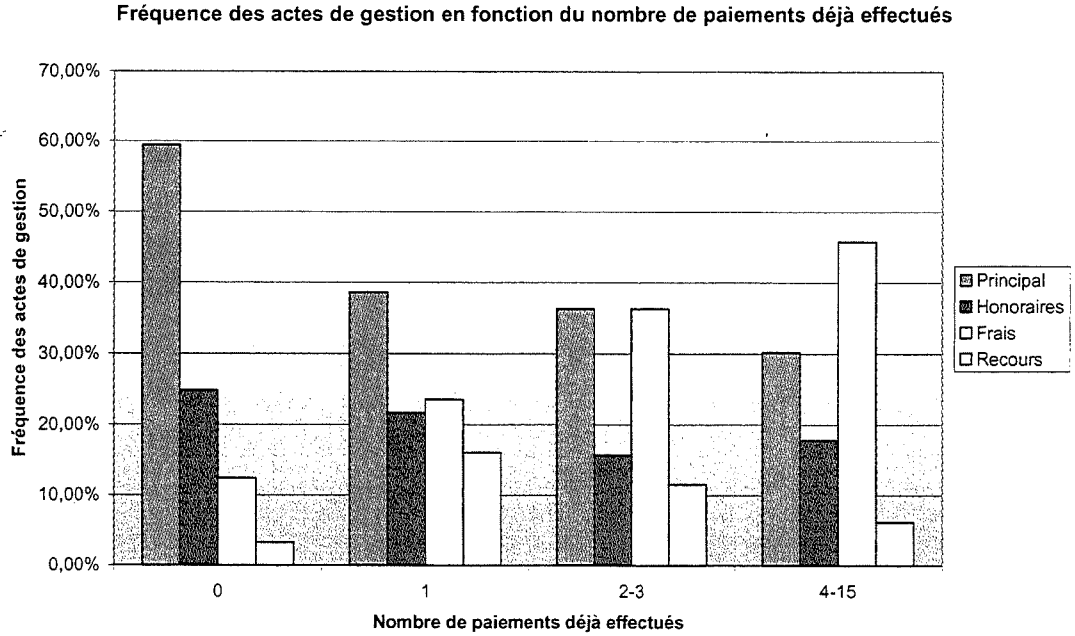


fig. 1.1 : Fréquence des actes de gestion en fonction du nombre de paiement déjà effectués

On peut ainsi remarquer que :

- Le premier paiement est très différent des autres. Il a beaucoup plus de chance d'être un règlement en principal. Cette information risque d'être déterminante pour la suite car les paiements en principaux représentent de loin les montants les plus importants.
- Plus le sinistre est fractionné, plus il engendre des frais.

— Influence de l'année de développement

On regarde ici la fréquence des actes de gestion en fonction de l'année de développement. Ici, l'année de développement est prise à partir de l'année de déclaration et non à partir de l'année de survenance. Les paiements qui adviennent plus de trois ans après la déclaration du sinistre ont été ici regroupés pour que chaque classe regroupe à peu près la même population.

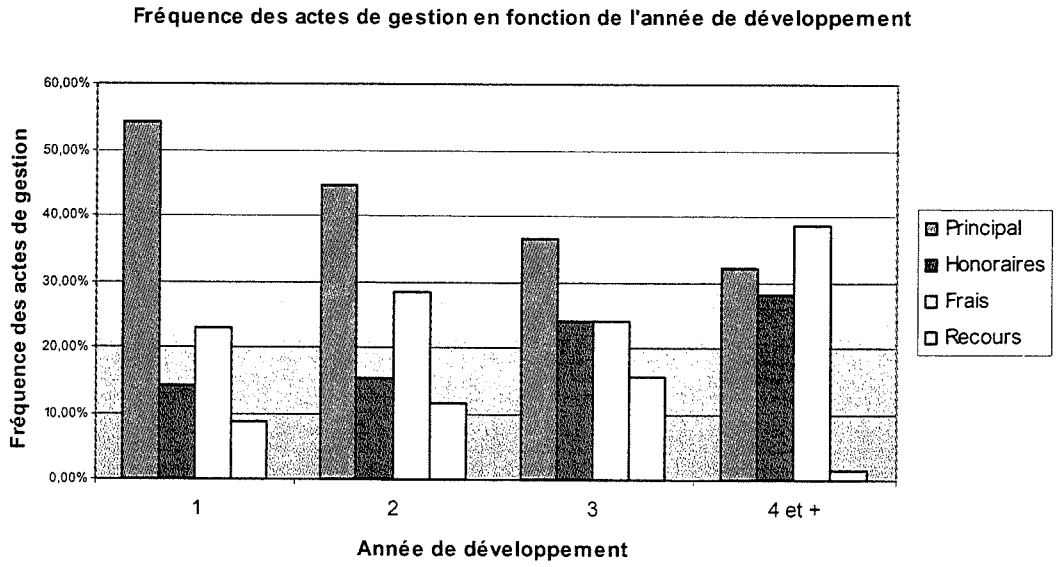


fig. 1.2 : Fréquence des actes de gestion en fonction de l'année de développement

- On remarque que plus le sinistre est déclaré depuis longtemps, plus la part relative des frais et des honoraires augmente, alors que la part relative des paiements en principal diminue.
- Les recours deviennent très rares après trois ans.

L'évolution observée ici est relativement similaire à celle observée en fonction du nombre de paiements déjà effectués.

— Influence du délai de déclaration

Le délai de déclaration peut avoir une influence sur la vie d'un sinistre. On peut penser qu'un sinistre important va être notifié rapidement.

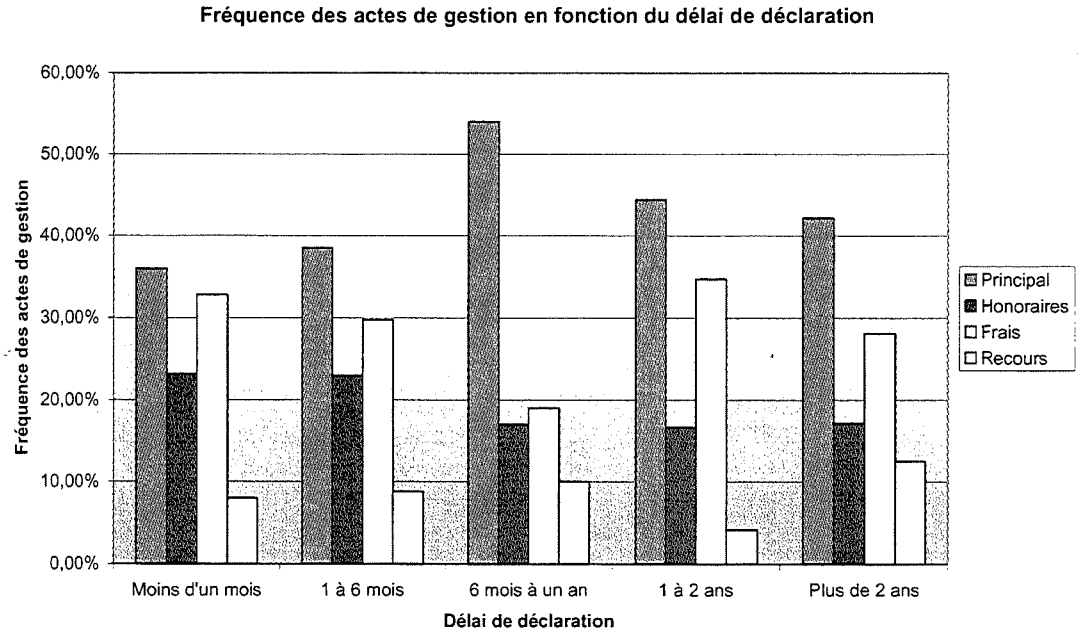


fig. 1.3 : Fréquence des actes de gestion en fonction du délai de déclaration

- Les différences ne semblent pas très marquées ici. Seuls les sinistres qui ont été déclarés entre six mois et un an après la survenance sortent du lot. Pour ces derniers, il y a plus de paiements en principal et moins de frais. Mais étant donné le faible nombre de flux observés, il n'est pas évident que ce facteur ait une véritable influence.

— Influence du temps écoulé depuis le dernier flux

S'il ne s'est rien passé depuis longtemps, cela peut également avoir de l'influence sur le développement du sinistre.

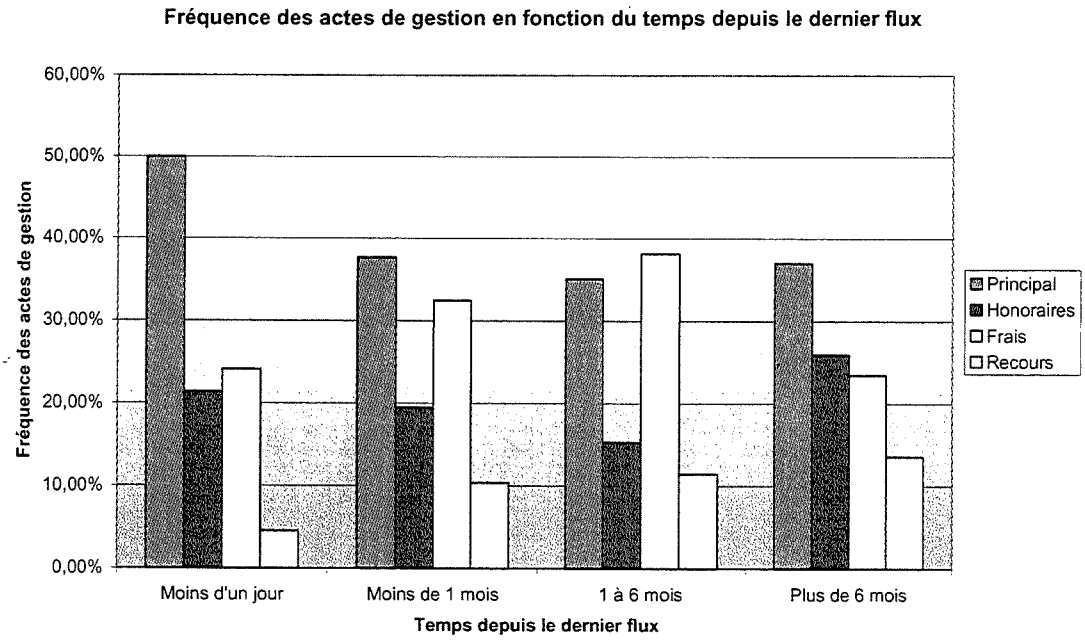


fig. 1.4 : Fréquence des actes de gestion en fonction du temps depuis le dernier flux

- La part relative des paiements en principal est relativement stable, sauf s'il s'est écoulé moins d'un jour depuis le dernier flux.
- Au contraire, si beaucoup de temps s'est écoulé, la part relative des recours augmente.

Statistiques sur le montant des flux

On peut effectuer le même genre de statistiques en travaillant non plus sur les nombres de paiements, mais sur les montants associés. Nous ne donnons ici que l'influence du nombre de paiements déjà effectués car les autres types de facteurs n'ont pas d'influence évidente au vu des graphiques.

— Influence du nombre de paiements déjà effectués

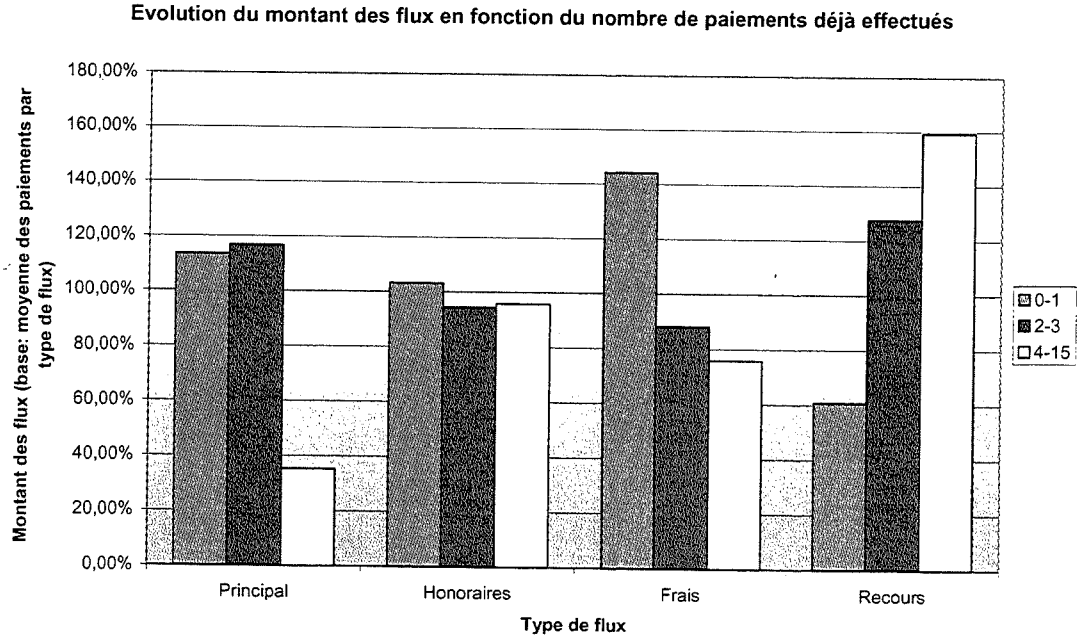


fig. 1.5 : Evolution du montant des paiements par type de flux en fonction du nombre de paiements déjà effectués

- Le montant du principal devient beaucoup plus petit après plus de trois paiements. Ce résultat n'était pas évident : on aurait pu penser que pour les sinistres longs, les paiements importants étaient réglés plus tard.

- Nous avons constaté à la partie précédente que le nombre de recours diminue lorsque le nombre d'actes de gestion devient grand. Mais après plus de 4 paiements, le montant associé est beaucoup plus important.

- Les honoraires restent relativement stables, en ligne avec ce que l'on avait déjà constaté pour les fréquences.

2- Résultats expérimentaux

Test du log-rank

— 2.1 : Test du log-rank entre la fonction de survie empirique et la fonction de survie théorique :

```
Call:
survdiff(formula = surv0 ~ offset(1 - prob0), rho = 0)
Observed Expected     Z     p
208.000  178.833  -0.158  0.874
```

Tab. 2.1 : Test du log-rank

Résultats sur l'ensemble des dates de flux

On définit les notations suivantes :

tps : « temps entre la déclaration et le flux étudié »

tdec : « delai de déclaration »

tepr : « temps entre la déclaration et le dernier flux »

num : « numéro du flux »

yest : « nomtant estimé »

mdp : « montant déjà payé »

tinc : « tps incrémental entre deux évènements »

indictinc : « indicatrice séparant les 2 états »

— 2.2 : GLM sur le temps écoulé entre la déclaration et la survenance d'un flux avec une loi de Poisson :

```
Call:
glm(formula = tps ~ tdec + tepr + num + yest + mdp, family = poisson(link = log))
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-31.626  -8.405  -2.095   4.529  43.850
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) 6.296e+00  3.649e-03 1725.183  < 2e-16 ***
tdec       -6.725e-04  7.051e-06  -95.368  < 2e-16 ***
tepr        8.154e-04  3.562e-06  228.949  < 2e-16 ***
num        -1.475e-02  6.120e-04  -24.104  < 2e-16 ***
```

```
yest        3.217e-08  2.693e-09   11.944  < 2e-16 ***
mdp         5.183e-08  6.546e-09    7.918 2.41e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 194983  on 508  degrees of freedom
Residual deviance:  90027  on 503  degrees of freedom
AIC: 94106
Number of Fisher Scoring iterations: 5
```

Tab. 2.2 : GLM global pour une loi de Poisson

— 2.3 : GLM sur le temps écoulé entre la déclaration et la survenance d'un flux avec une loi de Poisson avec surdipersion (en ne gardant que les facteurs significatifs) :

```
Call:
glm(formula = tps ~ tdec + tepr + num, family = quasipoisson(link = log))

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-31.017  -8.591  -2.423   4.670  43.728
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.303e+00  4.924e-02 128.015  < 2e-16 ***
tdec        -6.621e-04  9.493e-05  -6.974 9.68e-12 ***
tepr         8.310e-04  4.713e-05  17.631  < 2e-16 ***
num         -1.622e-02  8.235e-03  -1.970   0.0494 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for quasipoisson family taken to be 184.3999)
    Null deviance: 194983  on 508  degrees of freedom
Residual deviance:  90468  on 505  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 5
```

Tab. 2.3 : GLM pour une loi de Poisson avec surdispersion

— 2.4 : GAM sur le temps écoulé entre la déclaration et la survenance d'un flux avec une loi de Poisson avec surdipersion (en ne gardant que les facteurs significatifs) :

```
Family: quasipoisson
Link function: log
```

Formula:

tps ~ s(tdec) + s(tepr) + s(num)

Parametric coefficients:

	Estimate	std. err.	t ratio	Pr(>|t|)
constant	6.4239	0.02362	272	< 2.22e-16

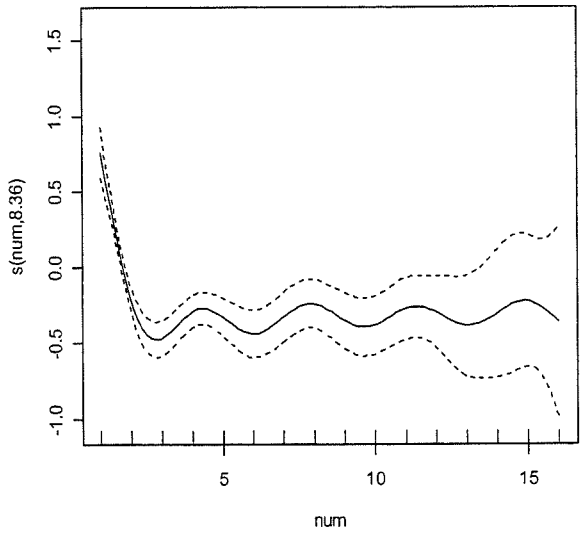
Approximate significance of smooth terms:

	edf	chi.sq	p-value
s(tdec)	4.55	45.307	1.7194e-08
s(tepr)	5.334	389.76	< 2.22e-16
s(num)	8.444	73.509	1.3297e-11

R-sq.(adj) = 0.674 Deviance explained = 62.9%

GCV score = 151.03 Scale est. = 145.29 n = 509

Tab. 2.4 : GAM pour une loi de Poisson avec surdispersion



Graph. 2.4 : Fonction de lissage du nombre de paiement associée au GAM

— 2.5 : GLM sur le premier paiement de chaque sinistre

Call:

glm(formula = tps[num == 1] ~ tdec[num == 1], family = quasipoisson)

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-35.432  -15.451  -2.393   9.048  35.017

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.6235486  0.0780359  84.878  < 2e-16 ***
tdec[num == 1] -0.0011699  0.0002309  -5.067 1.16e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 305.3225)

    Null deviance: 59470  on 152  degrees of freedom
Residual deviance: 49320  on 151  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Tab. 2.5 : GLM pour une loi de Poisson avec surdispersion du premier paiement

— 2.6 : GAM sur le délai entre la déclaration et un flux (hors premier paiement)

```
Family: quasipoisson
Link function: log
Formula:
tps[num > 1] ~ s(tdec[num > 1], tepr[num > 1])
Parametric coefficients:
          Estimate  std. err.   t ratio   Pr(>|t|)
constant    6.4609    0.01842     350.8   < 2.22e-16
Approximate significance of smooth terms:
                              edf     chi.sq    p-value
s(tdec[num > 1],tepr[num > 1])    22.88     1466.3    < 2.22e-16
R-sq.(adj) =  0.889   Deviance explained = 87.5%
GCV score = 58.797   Scale est. = 54.854    n = 356
```

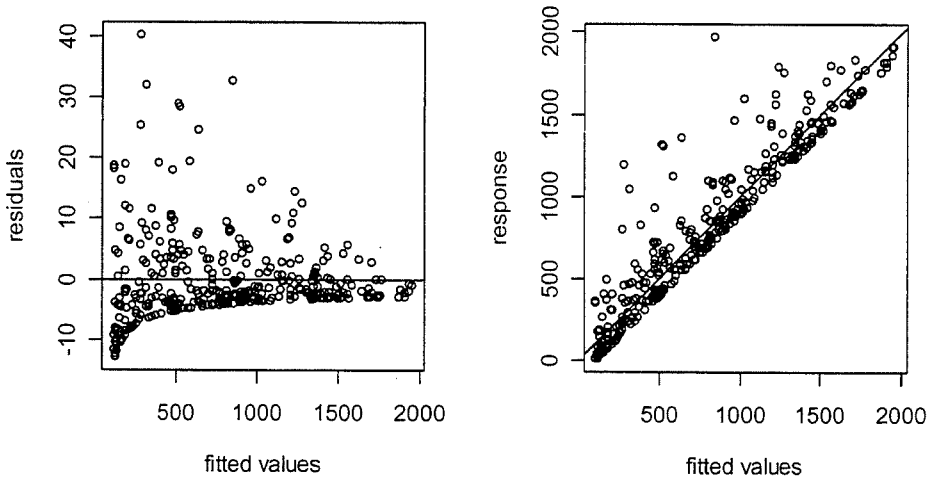
Tab. 2.6 : GAM pour une loi de Poisson avec surdispersion sur le délai entre la déclaration et un flux (hors premier paiement)

— 2.7 : GLM sur le délai entre la déclaration et un flux (hors premier paiement)

```
Call:
```

```
glm(formula = tps[num > 1] ~ (tepr[num > 1]), family = quasipoisson(link = identity))
Deviance Residuals:
   Min     1Q  Median     3Q    Max
-12.878  -3.762  -2.450   1.583  40.177
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  110.73480   11.82568   9.364   <2e-16 ***
tepr[num > 1]  1.00760    0.02229  45.201   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for quasipoisson family taken to be 62.08199)
    Null deviance: 124708  on 355  degrees of freedom
Residual deviance:  17890  on 354  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 6
```

Tab. 2.7 : GLM pour une loi de Poisson avec surdispersion sur le délai entre la déclaration et un flux (hors premier paiement)



Graph. 2.7 : GLM pour une loi de Poisson avec surdispersion sur le délai entre la déclaration et un flux (hors premier paiement)

— 2.8 : GLM sur le délai entre la déclaration et un flux (hors premier paiement) avec indicatrice pour les deux états

```
Call:
glm(formula = tps[num > 1 & dp > 0] ~ tepr[num > 1 & dp > 0] +
```

```
indictinc[num > 1 & dp > 0], family = quasipoisson(link = identity))

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-11.8638  -3.4080  -0.8107   2.5334  14.8488

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             141.53058   16.39912   8.630 8.07e-14 ***
tepr[num > 1 & dp > 0]    0.95712    0.02641  36.238  < 2e-16 ***
indictinc[num > 1 & dp > 0] 665.19199   66.77122   9.962  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 25.57479)

    Null deviance: 34342.4  on 105  degrees of freedom
Residual deviance:  2578.9  on 103  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Tab. 2.8 : GLM pour une loi de Poisson avec surdispersion sur le délai entre la déclaration et un flux (hors premier paiement) avec indicatrice

Résultats sur le montant des flux

Nous avons effectué des tests de Kolmogorov-Smirnov pour vérifier l'adéquation de nos distributions sélectionnées avec les données.

```
        Two-sample Kolmogorov-Smirnov test
data:  yprext and pareto1
D = 0.1751, p-value = 0.7115
data:  yhoext and pareto2
D = 0.289, p-value = 0.3744
data:  yfrext and weibull1
D = 0.279, p-value = 0.9147
data:  yreext and weibull2
D = 0.1167, p-value = 0.9583
alternative hypothesis: two.sided
```

Tab. 2.9 : Tests de Kolmogorov-Smirnov

Prise en compte des corrélations au sein du triangle

Voici le triangle des résidus normés obtenus avec le modèle de base. Les cellules grisées indiquent un écart significatif par rapport au modèle de base (avec un intervalle de confiance à 90%)

Reporting Year	Development Factors 0					
	1	2	3	4	5	6
1998	-2,14	1,72	1,70	1,22	1,00	0,00
1999	0,24	-0,05	-0,26	0,01	-1,00	
2000	-0,07	0,28	-0,79	-1,23		
2001	0,53	-0,77	-0,65			
2002	0,87	-1,17				
2003	0,58					

Tab. 2.10 : Résidus normés du modèle de base

3- Logiciels utilisés

R

Pour faire les simulations du modèle individuel, nous avons choisi d'utiliser le logiciel R. Il s'agit d'un logiciel libre et gratuit implémentant une panoplie variée de modèles statistiques.

R est un logiciel pour la manipulation des données, les calculs et la représentation graphique. Entre autres choses, il offre :

- La possibilité de stocker et de manipuler les données
- Une grande variété de fonctions statistiques (avec possibilité de représentation graphique)
- Une suite d'opérateurs pour les calculs sur les tableaux et les matrices
- Un langage de programmation (S) qui inclut boucles, conditions, fonctions récursives...Il s'agit en fait d'une nouvelle implémentation du langage S développé par les laboratoires AT&T et Bell laboratoires

Winbugs

Ce programme, développé par l'unité de Bio-Statistique du MRC (*medical research council*) de Cambridge, permet de mener des calculs d'inférence bayésienne avec la méthode de l'échantillonnage de Gibbs.