RESUME

L'objet de ce mémoire est de développer une méthodologie de provisionnement, pour la garantie responsabilité civile corporelle automobile, à partir de données individuelles et de l'articuler avec les méthodes de provisionnement sur données agrégées. L'idée générale développée est de considérer que sur les exercices « anciens » où les sinistres non clos sont peu nombreux et les situations sont disparates entre exercices de survenance, la prise en compte des données spécifiques au sinistre pourrait améliorer la précision des estimations.

La méthode sur données individuelles est fondée sur l'estimation de la durée de vie du sinistre et de la chronique des paiements restants pour ce sinistre. La durée de survie est confrontée à des données censurées, nous utilisons un modèle de survie paramétrique *AFT*, incorporant des variables explicatives. Les paiements sont décomposés en deux variables, la probabilité et le montant de paiement pour une année, et sont modélisés à l'aide des modèles linéaires généralisés.

Le modèle de provisionnement individuel est appliqué aux sinistres non clos appartenant aux exercices de survenance « anciens ». Le modèle est ensuite articulé avec la méthode *Chain Ladder* pour projeter les exercices de survenance récents.

Mots clés : provisionnement, modèle individuel, durée de survie, paiements, modèles linéaires généralisés, *Chain Ladder*, modèles de survie *Accelerated Failure Time.*

ABSTRACT

The aim of this document is to develop a reserving methodology, for automobile third party liability, based on individual claim data and to connect it with a reserving methodology based on aggregated data. The general concept developed is that, for longstanding years of occurrence where there are few claims and different situations amongst the years of occurrence, claim specific data can enhance the precision of the predictions.

This method is based on survival probability and claim's future payments. Survival data is confronted to censored data; we use a parametric accelerated failure time model based on explanatory variables to forecast survival probability. Payments are segmented in two variables, the probability and the amount of the payment during each development year. They are modeled using general linear models.

The individual reserving model is applied to all outstanding claims belonging to longstanding years of occurrence. Then, the model is connected with Chain Ladder in order to forecast the recent years of occurrence.

Key words: reserving, individual model, survival data, payments, general linear model, Chain Ladder, survival model Accelerated Failure Time.

REMERCIEMENTS

Je tiens tout d'abord à remercier la MAIF et plus particulièrement M. Berthoux qui m'a permis de suivre la formation du CEA, ainsi que M. Cases et M. Loizeau qui m'ont encouragé dans la réalisation de cette formation.

Je remercie mon directeur de mémoire, M. Oger, pour m'avoir encadré tout au long de la réalisation de ce mémoire et fait bénéficier de son expérience et de ses conseils.

Je souhaite témoigner ma reconnaissance à toutes les personnes qui m'ont aidé à réaliser ce mémoire dans les meilleures conditions.

Enfin, je remercie tout particulièrement ma famille pour son soutien, ses encouragements, et sa patience face à mon indisponibilité tout au long de ces travaux.

SOMMAIRE

INTRODUCTION

Le cycle de production assurantiel est inversé par rapport à une entreprise classique. Le prix de revient n'est connu qu'après le paiement de l'ensemble des sinistres de l'exercice de survenance.

Afin de rester solvable, l'assureur doit être en mesure de provisionner suffisamment pour couvrir les paiements futurs des sinistres en portefeuille et ainsi, rembourser sa dette envers les assurés.

La charge finale des sinistres survenus de l'exercice de survenance est inconnue et est estimée à l'aide de méthodologies statistiques qui se basent sur les évaluations dossier par les gestionnaires et les dépenses. Les provisions techniques sont obtenues par différence entre l'estimation de la charge finale et les paiements effectués.

Les méthodologies statistiques traditionnelles s'appuient sur l'utilisation des données agrégées. La méthode la plus répandue est basée sur les cadences de règlements. Les triangles de données agrégées utilisent l'année de survenance, l'année de développement et des coefficients de développement. Ces méthodologies sont simples d'utilisation et robustes. Néanmoins l'utilisation de données regroupées conduit à une perte d'information et engendre certaines limites telles que, par exemple, l'application des traités de réassurance non proportionnels.

De plus, sur des garanties à déroulement long, l'incertitude reste relativement importante, notamment sur les exercices de survenance anciens. L'application de méthodes agrégées sur ces exercices de survenance où le nombre de sinistre non clos est faible peut être délicate. Il est ainsi nécessaire de développer un ensemble de méthodes pour conforter les prévisions.

Les systèmes d'information retracent les principaux actes de gestion, il est ainsi possible de connaitre certains éléments de l'historique de l'évaluation et des paiements d'un sinistre. C'est à partir de cette information détaillée que peut être mise en place une méthodologie de provisionnement sur données individuelles.

Le périmètre de l'étude porte sur les sinistres survenus et déclarés à l'assureur, c'est à dire la provision *IBNER* (*Incurred But Not Enough Reported*). Les sinistres survenus mais non déclarés et plus spécifiquement la provision *IBNYR* (*Incurred But Not Yet Reported*) ne sont pas traités dans l'étude. Nous nous intéressons à la garantie automobile responsabilité civile corporelle pour les exercices de survenance 2003 à 2013.

L'objet de ce mémoire est de développer une méthodologie de provisionnement à partir de données individuelles et de l'articuler avec les méthodes de provisionnement sur données agrégées. La méthode sur données individuelles est fondée sur la connaissance de la durée de vie du sinistre et de la chronique des paiements restants pour ce sinistre. Ces éléments seront respectivement modélisés à l'aide des modèles de survie et des modèles linéaires généralisés.

L'idée générale développée est de considérer que sur les nouvelles survenances il existe un grand nombre de sinistres qui font l'objet d'actes de gestion et l'utilisation d'une méthodologie agrégée prend alors tout son sens. Par contre, sur les exercices « anciens » où les sinistres non clos sont peu nombreux et les situations sont disparates entre exercices

de survenance, l'estimation de l'évolution moyenne via le calcul de coefficients de développement peut paraitre moins justifiée. Nous étudierons si l'utilisation des données individuelles sur ces exercices fournit une estimation plus pertinente de la moyenne.

La 1ère partie du mémoire est consacrée à la description de l'étude et de la garantie automobile responsabilité civile corporelle.

La 2ème partie est consacrée aux méthodes de provisionnement traditionnelles c'est-à-dire basées sur une approche de données agrégées. Les méthodes déterministes et stochastiques méthodes serviront par la suite de base de comparaison pour le modèle de provisionnement sur données individuelles.

Dans la 3ème et 4ème partie nous aborderons respectivement la modélisation de la durée de vie et des paiements des sinistres. Les modèles seront présentés théoriquement puis appliqués aux données sinistres.

Dans la 5ème partie nous présenterons le modèle de provisionnement sur sinistres individuels et articulerons ce modèle avec les méthodes de provisions sur données agrégées. Nous effectuerons une comparaison du modèle final avec le modèle *Chain Ladder*.

Certaines données utilisées pour ce travail ont été transformées pour en assurer la confidentialité, sans pour autant porter atteinte aux résultats obtenus.

1 PARTIE 1 - PRESENTATION DE L'ETUDE

La MAIF est le 5ème assureur automobile de France et le 9ème assureur toutes assurances dommages confondues.

Les produits du groupe MAIF couvrent les univers suivants :
- l'automobile (contrat automobile ainsi que la couverture des risques corporels du conducteur),
- l'habitat (multirisque habitation, navigation de plaisance),
- la famille (accidents corporels, décès, santé),
- l'avenir (épargne, retraite),
- la vie associative (personnels morales : contrat automobile et habitation).

Fin 2012 le groupe comptait :
- 3,4 millions de sociétaires, dont 149 000 associations et collectivités,
- 3,6 millions de véhicules assurés,
- 2,3 millions de contrats habitation.

Le chiffre d'affaire consolidé atteint environ 3 Mds d'euros en 2012 et se décompose de la manière suivante :
- 2,5 Mds en assurance non-vie dont 1,4 Mds en automobile,
- 0,5 Mds en assurance vie.

Dans le cadre de ce mémoire, nous avons choisi d'étudier la garantie Responsabilité Civile Corporelle (RCC) du contrat automobile personnes physiques de la MAIF. Ce choix découle de la part importante dans les provisions de la MAIF de cette garantie, l'incertitude qui lui est associée et des difficultés de prévisions liées aux caractéristiques que nous présentons dans la partie suivante.

1.1 La garantie automobile RCC

1.1.1 Cadre juridique et conventionnel

1.1.1.1 Cadre juridique

Depuis 1985, la loi Badinter encadre le règlement des sinistres RCC[1]. Cette loi a posé le principe de la responsabilité de l'automobiliste en cas de dommages corporels consécutifs à un accident de la circulation, et a permis l'indemnisation automatique et rapide des victimes, sur la base du règlement amiable. Ses principales caractéristiques sont les suivantes :

- **Indemnisation automatique des victimes**

Piétons, cyclistes et passagers d'un véhicule sont indemnisés à 100 % des dommages corporels consécutifs à un accident de la circulation, sauf faute inexcusable de leur part qui aurait été l'unique cause de l'accident ou volonté manifeste de rechercher le dommage (suicide par exemple).

[1] La loi Lefrand (présentée par Mr Lefrand, Mme Levy, Mr Chossy et Mme Montchamp) qui a été déposée le 5 novembre 2009 réaffirme ces principes, tout en y apportant des compléments (nomenclature Dinthilac).

- **Véhicule motorisé**

La loi s'applique à tout accident dans lequel est impliqué un véhicule terrestre à moteur ainsi que ses remorques et semi-remorques.

- **Dommages indemnisés**

Les victimes sont indemnisées intégralement des conséquences du dommage corporel (blessures ou décès), qu'il s'agisse de l'atteinte à l'intégrité physique (frais médicaux, incapacité temporaire ou définitive ...), de l'atteinte morale ou économique (perte de gain pendant l'arrêt de travail ...).

Les dommages occasionnés aux fournitures et appareils délivrés sur prescription médicale (appareils auditifs ou dentaires, lunettes correctrices, etc.) sont également indemnisés.

En cas de décès de la victime, les conjoints, concubins et descendants sont indemnisés du préjudice qui leur est propre, comme le prix de la douleur, les pertes économiques causées par la disparition de la victime, avec les mêmes règles applicables. Seule la faute inexcusable de la victime leur est opposable, pouvant ainsi limiter ou exclure l'indemnisation.

- **Règlement amiable**

La loi Badinter a mis en place une procédure d'indemnisation simplifiée permettant un règlement plus rapide des sommes versées sous forme de rente ou de capital, au titre de la réparation du préjudice.

    - Transaction

L'assureur du responsable est tenu de présenter, dans un délai maximum de 8 mois à compter de l'accident, une offre d'indemnité (en cas de décès au conjoint et/ou aux héritiers). Pour être valable, l'offre doit inclure tous les éléments indemnisables du préjudice et être d'un montant manifestement suffisant sous peine de sanctions judiciaires.

Lorsque l'assureur n'a pas été informé de la consolidation de l'état de la victime dans les 3 mois de l'accident, l'offre a un caractère provisionnel. Dans ce cas, l'offre définitive doit être faite dans un délai de 5 mois suivant la date à laquelle l'assureur est informé de la consolidation.

En cas d'aggravation anormale et imprévisible de son état mais liée aux dommages consécutifs de l'accident, la victime a 10 ans, à compter de l'apparition de l'aggravation, pour présenter une demande à l'assureur qui a versé l'indemnité.

    - Règlement judiciaire

La victime est libre d'accepter l'offre, de la discuter ou de la refuser, sans aucune contrainte de délai. Elle a la possibilité de revenir sur sa décision en dénonçant la transaction dans les 15 jours de sa conclusion par lettre recommandée avec accusé réception.

De plus, elle peut, à tout moment, intenter un recours judiciaire soit parce qu'elle refuse la transaction, soit parce qu'elle estime l'offre « manifestement insuffisante ». Le juge pourra condamner l'assureur à verser des dommages et intérêts à la victime.

Enfin, lorsque sa situation personnelle le justifie, la victime peut demander au juge que la rente qui lui a été allouée soit convertie en capital, pour la totalité ou en partie seulement, pour les arrérages à venir.

1.1.1.2 Cadre conventionnel

La mise en place d'une convention a pour objectif d'accélérer l'indemnisation de préjudices corporels des victimes d'un accident de la circulation.

La convention IRCA (convention d'Indemnisation et de Recours Corporel Automobile) est applicable à tout accident survenu à compter de 1er avril 2002 entrainant un préjudice corporel inférieur à un certain seuil, soit 5% d'Atteinte à l'Intégrité Physique et Psychique (AIPP).

Cette convention désigne, dès la déclaration du sinistre, l'assureur en charge d'instruire le dossier de chaque victime et les conditions de recours entre les sociétés d'assurance concernées.

L'assureur du non responsable exerce un recours sur la base d'un barème préétabli pour des postes de dommages et de préjudice définis. A titre d'exemple, depuis l'année 2007 le montant du recours entre assureurs pour un taux d'AIPP nul et une responsabilité totale est de 1 490 €.

1.1.2 Le processus d'évaluation et règlement

La vie d'un sinistre est généralement caractérisée par le processus suivant :

Le processus d'évaluation suit les grandes étapes suivantes :

- ouverture du dossier : à la déclaration d'un événement, le gestionnaire possède généralement très peu d'information sur les victimes.

- réception d'un premier rapport d'expertise : en général, dans les 2 mois qui suivent l'événement, le gestionnaire reçoit un Certificat Médical Initial (CMI) et un

questionnaire médical par victime, qui lui permet de réaliser une première évaluation chiffrée.

- réception du rapport définitif d'expertise médicale : ce rapport correspond à la consolidation[2] de l'état de santé de la victime. Le gestionnaire dispose alors de tous les éléments pour évaluer en détail les indemnisations de la victime par poste de préjudice. Dans le cas simple, la remise de ce rapport conduit au règlement définitif du sinistre.

Ces grandes étapes sont représentatives de la gestion des sinistres. Cependant, les caractéristiques de ce processus de gestion peuvent se révéler très différentes selon la gravité des sinistres :

- la durée de consolidation de l'état de la victime est très variable. Elle est généralement d'environ 6 mois pour les sinistres de faible gravité (fracture d'un membre par exemple), tandis qu'elle atteint plus de 3 ans pour les sinistres très graves.

  A l'extrême, dans le cadre de victimes très jeunes, il est parfois nécessaire d'attendre la fin de la période de scolarisation pour estimer les incidences professionnelles.

  Ainsi, plusieurs rapports d'expertises médicales intermédiaires peuvent être produits pour les sinistres les plus graves (en pratique, environ un tous les 18 mois). Ces rapports conduisent généralement à revoir les évaluations du coût des sinistres.

- l'état de santé à un instant donné ne permet pas toujours de prévoir l'évolution future, c'est le cas notamment des traumatismes crâniens graves, pour lesquels il est très difficile d'anticiper les conséquences futures.

- la victime ou son entourage peut décider d'intenter une action en justice. Généralement, l'objectif visé est une révision de la gravité des postes de préjudices estimée par l'expert médical et en conséquence, les indemnités correspondantes.

  Dans ce cas, la durée entre le rapport de consolidation et la détermination de l'indemnité finale s'en trouve sensiblement rallongée. L'estimation à un instant donnée de la date de clôture du dossier est alors difficile.

- L'état de la victime peut quelquefois évoluer de manière subite après consolidation (évolution inattendue de l'état de la victime, évolution de l'environnement proche de la victime[3]). Cette évolution entraîne généralement la réouverture du sinistre et sa réévaluation.

L'estimation des coûts des sinistres RCC se fait en principe dossier par dossier au niveau du gestionnaire de sinistres. Les exceptions suivantes sont pratiquées à la MAIF :

- Les sinistres de faible montant : dans le cas où le gestionnaire estime que le coût ultime sera inférieur au plafond (915 €) alors l'évaluation est positionnée à un montant forfaitaire.

[2] Moment où les lésions se sont fixées et ont pris un caractère permanent tel qu'un traitement n'est plus nécessaire, si ce n'est pour éviter une aggravation, et qu'il devient possible d'apprécier un certain degré d'incapacité fonctionnelle permanente réalisant un préjudice définitif.

[3] Par exemple, dans le cas où la victime peut vivre à domicile et que certains actes essentiels de son existence sont réalisés par ses parents : lorsque ces derniers ne peuvent plus remplir ce rôle (vieillesse, décès), il peut s'avérer nécessaire de placer la victime en institution spécialisée.

- Les sinistres entrant dans le cadre du forfait de la convention IRCA (par exemple, 1490 € pour un sinistre avec un taux d'atteinte à l'intégrité physique et psychique nul, pour une responsabilité totale).

Enfin, différentes techniques et moyens ont été mis à disposition des régleurs de sinistres, pour leur permettre une meilleure estimation de la charge ultime des sinistres :

- La mise à disposition d'une base d'expérience détaillant les indemnisations correspondant à des sinistres corporels survenus depuis 2002,
- Le suivi des dossiers les plus graves centralisé dans une entité spécialisée.

1.1.3 Les principaux postes de coût

Depuis 2006, l'évaluation des dommages et préjudices conduisant à une indemnisation respecte la nomenclature dite « Dintilhac »[4]. Les informations détaillées de la nomenclature Dintilhac ne figurent pas dans les bases de données qui ont pu être utilisées dans le cadre de cette étude.

Les coûts de sinistres correspondent aux victimes directes ou par ricochet (par exemple, frais d'obsèques) et incluent les remboursements de frais engagés ou à engager par les organismes sociaux.

Les coûts des sinistres de faible gravité sont essentiellement composés de « dépenses de santé actuelles », de frais divers ainsi que d'indemnisation « de principe » pour de faibles niveaux de souffrances endurées.

Les postes de coûts les plus importants pour les sinistres graves (par exemple : traumatismes crâniens graves, para et tétraplégie, atteinte du plexus brachial…) sont constitués par l'assistance par des tierces personnes et par les pertes de gain futures. Ces paiements peuvent faire l'objet d'un paiement sous forme de rente viagère.

L'évaluation des sinistres RCC pose de grandes difficultés à l'assureur du fait :

- de la grande diversité des situations à indemniser,
- de l'aléa fort de chaque sinistre tant sur la durée du sinistre que sur le niveau du coût.

Les méthodologies statistiques sont mises en œuvre afin de compléter l'évaluation dossier-dossier. Ces méthodologies visent à acquérir des éléments de certitude sur la suffisance de la provision à constituer. Il est nécessaire que cette provision soit prudente afin qu'elle permette le paiement de tous les sinistres quelles que soient leurs évolutions.

1.2 La provision pour sinistres à payer (PSAP)

1.2.1 Le fonctionnement de PSAP

Un des critères fondamentaux à la solvabilité des organismes d'assurance est l'évaluation prudente des dettes c'est-à-dire des provisions techniques (R. 331).

[4] Nomenclature définie dans le rapport du groupe de travail chargé d'élaborer une nomenclature des préjudices corporels, dirigé par Mr Dintilhac.

La provision pour sinistres à payer (PSAP, article R.331-6) a pour objectif de permettre le règlement, en principal et en frais, des engagements envers les assurés et les bénéficiaires des contrats.

Selon l'article R331-6 4 du code des assurances :

« Valeur estimative des dépenses en principal et en frais, tant internes qu'externes, nécessaires au règlement de tous les sinistres survenus et non payés, y compris les capitaux constitutifs des rentes non encore mises à la charge de l'entreprise. »

A la date de l'inventaire, pour un exercice de survenance il existe des sinistres clos et ouverts. La PSAP correspond à l'évaluation des paiements futurs des sinistres ouverts et permet de concilier le décalage entre la survenance du sinistre et le règlement effectif du sinistre.

La PSAP englobe les sinistres survenus connus et pas encore connus :

- les sinistres connus de l'assureur à la date d'inventaire dont certains peuvent être très mal connus à cette date (peu d'information),
- les sinistres qui ne sont pas connus de l'assureur car ils n'ont pas encore fait l'objet de déclaration de la part de l'assuré (ou du tiers).

La provision pour sinistres à payer (PSAP) correspond à la charge ultime déduction faite des paiements déjà réalisés à la date d'inventaire.

Selon l'article R331-15 du code des assurances :

« La provision pour sinistres à payer est calculée exercice par exercice.

Sans préjudice de l'application des règles spécifiques à certaines branches prévues à la présente section, l'évaluation des sinistres connus est effectuée dossier par dossier, le coût d'un dossier comprenant toutes les charges externes individualisables ; elle est augmentée d'une estimation du coût des sinistres survenus mais non déclarés.

La provision pour sinistres à payer doit toujours être calculée pour son montant brut, sans tenir compte des recours à exercer ; les recours à recevoir font l'objet d'une évaluation distincte.

Par dérogation aux dispositions du deuxième alinéa du présent article, l'entreprise peut, avec l'accord de l'Autorité de contrôle des assurances et des mutuelles, utiliser des méthodes statistiques pour l'estimation des sinistres survenus au cours des deux derniers exercices. »

1.2.2 La PSAP à fin 2013

Les provisions calculées dans le cadre des arrêtés de comptes MAIF sont essentiellement basées sur les déroulés de triangles, charges et dépenses, à l'aide de méthodologies de type *Chain Ladder*. L'évaluation du niveau de prudence est effectuée à partir des techniques stochastiques.

Les triangles sont segmentés en deux triangles, un triangle de sinistres attritionnels et un triangle de sinistres importants, en fonction d'un seuil d'écrêtement. Cette segmentation permet d'appliquer des méthodologies différenciées sur chaque triangle. Ces calculs ne seront pas détaillés dans ce mémoire, pour des raisons de confidentialité.

Nous considérerons qu'ils sont effectués dans les règles de l'art, étant entendu qu'ils font l'objet d'analyses approfondies de la part des équipes en charge du provisionnement, ainsi que de revues périodiques de la part de nos Commissaires aux Comptes.

A l'arrêté des comptes 2013, les exercices de survenance de 1996 à 2013 ont fait l'objet d'une estimation de la PSAP. Les exercices antérieurs sont considérés comme totalement clos, ou alors provisionnés à leur juste montant dans les dossiers : ils ne seront donc pas étudiés ici.

Les provisions présentées par la suite sont des provisions, pour les exercices du survenance de 2003 à 2013, brutes de recours et de réassurance, conformément à la règlementation. Elles ne comprennent pas les provisions pour frais de gestion. Ces trois quantités font l'objet d'estimations par ailleurs, et ne seront pas abordées dans le cadre de ce mémoire.

1.3 Le modèle basé sur données détaillées

L'utilisation de méthodes basées sur des données regroupées conduit à une perte d'information et engendre certaines limites telles que, par exemple, l'application des traités de réassurance non proportionnels.

De plus, sur des garanties à déroulement long, l'incertitude reste relativement importante, notamment sur les exercices de survenance anciens. L'application de méthodes agrégées sur ces exercices de survenance où le nombre de sinistre non clos est faible peut être délicate. Il est ainsi nécessaire de développer un ensemble de méthodes pour conforter les prévisions.

L'approche présentée consiste à considérer les données individuelles des sinistres, pour une partie des données, et à construire des méthodes statistiques sur ces données détaillées.

Cette approche permet la prise en compte des données spécifiques au sinistre et intuitivement, avec plus d'information, la précision des estimations pourrait être améliorée.

L'idée est d'appliquer le modèle individuel sur une partie de la diagonale du triangle, l'autre partie étant traitée par l'approche agrégée.

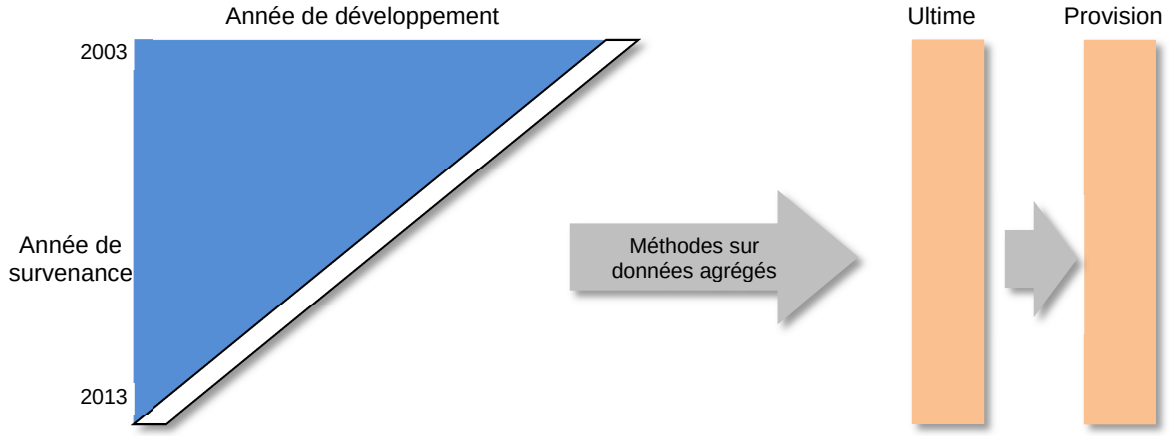
Nous avons choisi de développer le modèle individuel hors les sinistres tardifs et sans suite. Le modèle individuel fournira une estimation de la provision *IBNER*.

Compte tenu de la nature et des dynamiques d'évolution des sinistres autos RCC, nous estimons que seuls les sinistres non clos au 31/12/2013 supérieurs ou égaux à la 5ème année de développement seront traités à l'aide du modèle individuel. En effet, à partir de ce seuil de vieillissement, nous constatons les notions suivantes :

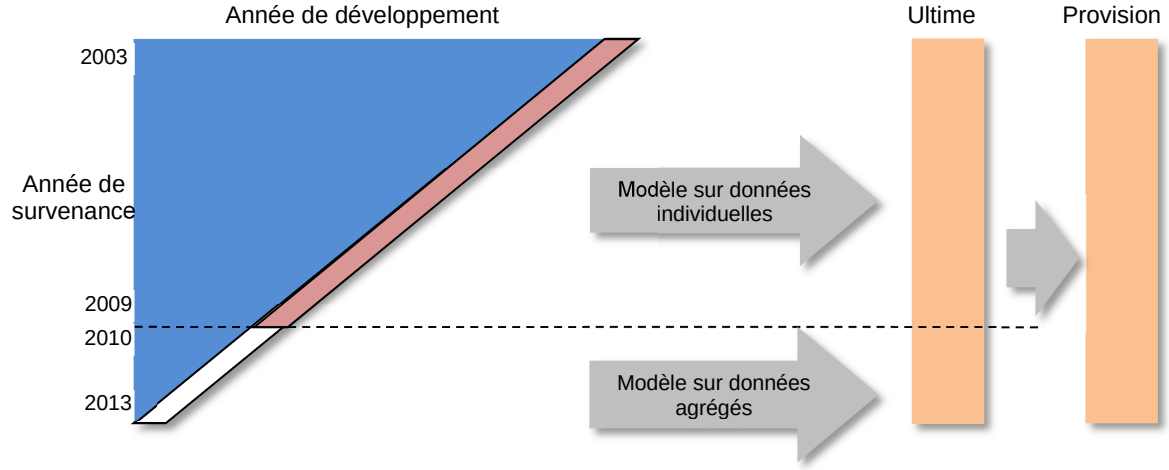
- il n'existe plus de sinistres tardifs et de sinistres évalués en forfait de gestion,
- les sinistres sans suite sont clôturés[5],
- les caractéristiques du sinistre, tel le taux d'IPP, sont stabilisés c'est-à-dire n'évolueront pas par la suite[6],
- le déroulement des paiements est spécifique aux exercices de survenance du fait du faible nombre de sinistres ouverts engendrant une volatilité plus forte.

[5] Au bout de 4 ans, plus de 99% des sinistres sans suite sont clôturés.
[6] Moins de 0,2% des sinistres a les caractéristiques (le nombre de victimes et le taux d'IPP) évoluant les années suivantes.

Provisionnement sur données agréés



Provisionnement sur données agréés et données individuelles



SCHEMA 1 – Illustration du modèle sur données individuelles

L'idée générale est de considérer que sur les survenances récentes il existe un grand nombre de sinistres qui font l'objet d'actes de gestion. L'utilisation d'une méthodologie agrégée prend alors tout son sens. Par contre, sur les exercices « anciens » où les sinistres non clos sont peu nombreux et les situations sont disparates entre exercices de survenance, l'estimation de l'évolution moyenne via le calcul de coefficients de développement parait moins justifiée. Nous étudierons si l'utilisation des données individuelles sur ces exercices fournira une estimation plus pertinente de la moyenne.

De plus, dans la méthode *Chain Ladder*, les coefficients calculés sur les exercices « anciens » jouent un rôle important car ils influent sur les projections de l'ensemble des exercices de survenances suivants. Ainsi, la projection des exercices de survenance récents s'appuie notamment sur les derniers coefficients de développement qui sont calculés sur les flux des exercices « anciens ».

Méthodes sur données individuelles

Pour chaque sinistre non clos du trapèze supérieur (trapèze rouge), nous projetons la probabilité de survie et l'estimation de l'espérance des paiements futurs sur chacune des

années de développement futures. Cette projection tient compte des caractéristiques du sinistre à la date d'inventaire. La combinaison de ces 2 éléments nous permet d'obtenir l'estimation de l'espérance des paiements probabilisés de la durée de vie.

Ensuite pour chaque exercice de survenance du trapèze supérieur, nous sommons l'ensemble des flux des sinistres non clos et obtenons ainsi la chronique des paiements jusqu'à l'ultime.

Articulation avec la méthode sur données agrégés

A partir des flux réels et projetés à l'aide du modèle individuel, nous calculons les coefficients de développements. Compte tenu des sorties du modèle sur données individuelles, les coefficients de développement sont calculés jusqu'à l'ultime. Ainsi, les flux de paiements des exercices de survenance récents (trapèze blanc) sont projetés jusqu'à l'ultime.

Provisions

Le montant de provisions pour chaque exercice est obtenu à partir de la différence entre le montant ultime et les paiements effectués à la date d'inventaire. Puis, en sommant le montant de provisions pour chaque exercice de survenance nous obtenons le montant de provision total.

1.3.1 Méthodologie de provisionnement ligne à ligne

Pour chaque sinistre, nous connaissons l'exercice de survenance, l'état du sinistre (clos ou en cours), le montant et la nature des paiements effectués. Si le dossier est en cours alors la date de clôture et les paiements restants sont inconnus et sont des variables aléatoires.

Le modèle de provisionnement est basé sur les paiements restant jusqu'à la date de clôture et sachant l'information disponible à la date t (notion de tribus). Les fondements méthodologiques sont présentés dans le mémoire de Gilles Chau et al.[7]

La provision[8] d'un sinistre individuel R_i *(Best Estimate* non actualisé) et l'estimation déterministe du montant de la provision $\hat{R}_i$, conditionnellement à l'information disponible à l'inventaire, sont obtenues à partir de la durée de vie et des paiements du sinistre :

$$R_i/F_t = \sum_{j=1}^{Ti} [X_{ij}/F_t] \quad et \ \hat{R}_i/F_t = \sum_{j=1}^{J} P(T_i > j/F_t) * E[X_{ij}/F_t]$$

avec :

- i : identifiant du sinistre dans la base de données,
- j : durée du sinistre,
- X_{ij} : le paiement du sinistre i à la date j,
- T_i : durée de vie du sinistre,
- J : durée maximale d'un sinistre,
- F_t : Information disponible à la date d'inventaire appelé tribus à la date t.

[7] Gilles Chau, Ngoc An Dinh (2012). Construction d'une méthode de provisionnement ligne à ligne pour des risques non-vie. ENSAE Paristech.

[8] C'est une provision 50/50. Par la suite, le terme provision désigne une provision 50/50.

Cette équation s'appuie sur l'hypothèse d'indépendance des durées de vie des sinistres et des paiements. Cette hypothèse peut être discutable étant donné que les gros sinistres nécessitent plus d'années pour être réglés.

Ceci étant, nous chercherons à distinguer les sinistres individuels en segments homogènes, basés sur le profil de durée, afin d'atténuer le biais lié à l'hypothèse d'indépendance et d'améliorer la qualité de l'estimation. Ainsi, en utilisant le profil de durée comme critère explicatif des paiements nous introduisons une dépendance entre la durée de vie et les paiements.

La provision globale R et l'estimation de la provision globale $\hat{R}$ sont égales à la somme des provisions des sinistres individuels :

$$R/F_t = \sum_{i=1}^{I} [R_i/F_t] \quad et \; \hat{R}/F_t = \sum_{i=1}^{I} [\hat{R}_i/F_t]$$

avec I : le nombre de sinistres.

Il est à noter qu'en général, le provisionnement classique résulte d'une approche prudente où les paiements futurs sont par conséquent surestimés en moyenne.

Le modèle de provisionnement ligne à ligne nécessite ainsi une connaissance de la date de clôture du sinistre et des paiements des sinistres à l'ultime.

A la déclaration du sinistre, ces éléments sont inconnus et sont des variables aléatoires. Au fur et à mesure du développement, nous disposons davantage d'information sur ces éléments et pouvons effectuer l'estimation avec une plus grande précision (notion de tribus F_t).

La difficulté réside dans la connaissance des dates et des montants des paiements futurs. L'implémentation de cette méthodologie requiert la modélisation de la date de clôture et la modélisation des paiements futurs.

1.3.2 Le modèle de durée de vie des sinistres

La base de données est composée de sinistres en cours et clos. La modélisation des dates de clôture uniquement sur les dossiers clos introduirait un biais et ne prendrait pas en compte l'information des sinistres en cours.

La modélisation de la durée de vie d'un sinistre est confrontée à des données censurées c'est-à-dire à des dossiers en cours où la date de clôture n'est pas encore connue.

La modélisation sera donc effectuée à l'aide des méthodes d'analyses de survies qui permettent d'inclure l'information des données censurées.

En général, l'analyse de survie consiste à modéliser la survenance d'un décès qui dans notre étude sera la date de clôture du sinistre. Les sinistres pour lesquels il n'y a pas de date de clôture seront confrontés à une censure.

Cette approche consiste à décrire et expliquer la « survie » d'une population donnée à l'aide des approches non paramétrique et paramétrique.

L'objectif de cette partie est d'obtenir des courbes de survie différenciées en fonction de facteurs explicatifs. Un second objectif est d'obtenir une approche projective c'est-à-dire d'aller au-delà de la profondeur de l'historique. Une piste est d'utiliser des méthodes paramétriques.

La modélisation de la durée de survie permet de probabiliser la durée de survie, des sinistres non clos à la date d'inventaire, au début de chaque année de développement.

Année développement	6	7	8	…	J
$P(T_i > j / F_t)$	1.0	0.9	0.79		0

TABLEAU 1 – Exemple de vecteur de probabilité de survie

1.3.3 Le modèle des paiements

Dans cette partie, nous présenterons la modélisation du processus de paiements. Nous définirons le cadre théorique et le choix du modèle retenu.

Nous disposons d'une base de données avec les montants et les dates de paiements pour chaque sinistre. Pour chaque sinistre, nous avons un vecteur X de paiements composés d'éléments X_j.

Cependant avant la modélisation, nous modifions les paiements de la population d'étude pour obtenir une base *as-if* 2013. L'objectif de la base *as-if* est d'obtenir des paiements comparables c'est-à-dire neutralisés de l'inflation.

L'idée est de transposer l'utilisation des MLG sur les triangles agrégés sur les données individuelles. Ainsi, nous modéliserons les paiements d'un sinistre X_{ij} à l'aide de l'année de survenance α et l'année de règlement β. En plus de ces 2 variables, nous introduirons d'autres facteurs explicatifs spécifiques au sinistre tel le profil de durée γ, obtenu à la partie précédente, et le montant de la provision dossier-dossier δ évaluée par le gestionnaire.

Le profil de durée joue un rôle clé dans la modélisation des paiements car cette variable permet de relier la durée de vie du sinistre et le montant des paiements. Nous définissons l'équation suivante :

$$X_{ij} = f(\alpha, \beta, \gamma, \delta)$$

L'estimation des paiements se fait à l'aide d'un modèle basé sur 2 variables aléatoires : la probabilité de paiement et le montant du paiement.

L'objectif de cette partie est d'obtenir une chronique des paiements futurs différenciée en fonction de facteurs explicatifs et jusqu'à la date de clôture du sinistre.

Année développement	6	7	8	…	J
$E[X_{ij} > j / F_t]$	5 000 €	2 500 €	500 €		0 €

TABLEAU 2 – Exemple de vecteur des paiements

C'est à cette étape que le modèle incorpore l'inflation. Ainsi, l'inflation s'applique sur les paiements comme suit :

$$E\left[X_{ij}^{'}/F_t\right] = E\left[X_{ij}/F_t\right] * \prod_{k=6}^{j}(1 + Inflation_k)$$

Nous prenons pour hypothèse que le scénario central d'inflation sur la période de projection est de 1,50 %. Le vecteur de paiement ajusté de l'inflation est le suivant :

Année développement	6	7	8	…	J
Inflation$_k$	1,50%	1,50%	1,50%	…	1,50%
E[X'$_{ij}$ > j /F$_t$]	5 075 €	2 575 €	523 €		0 €

TABLEAU 3 – Exemple de vecteur des paiements ajusté de l'inflation

1.3.4 Le modèle individuel déterministe

Pour chaque sinistre non clos, nous projetons sur chacune des années de développement futures la probabilité de survie, au début de chaque année, et l'estimation de l'espérance des paiements futurs considérée au milieu de chaque année. La combinaison de ces 2 éléments nous permet d'obtenir l'estimation de l'espérance des paiements probabilisés de la durée de vie.

Année développement	6	7	8	…	J
P(T$_i$ > j /F$_t$)	1.0	0.9	0.79		0
E[X'$_{ij}$ > j /F$_t$]	5 075 €	2 575 €	523 €		0 €
E[R$_{ij}$ > j /F$_t$]	5 075 €	2 317€	413 €		0 €

TABLEAU 4 – Exemple de vecteur des provisions

La somme de tous les composants du vecteur nous permet d'obtenir le montant de provision du sinistre :

$$\hat{R}_i = 5\,075 + 2\,317 + 413 + \cdots$$

1.3.5 Le modèle final

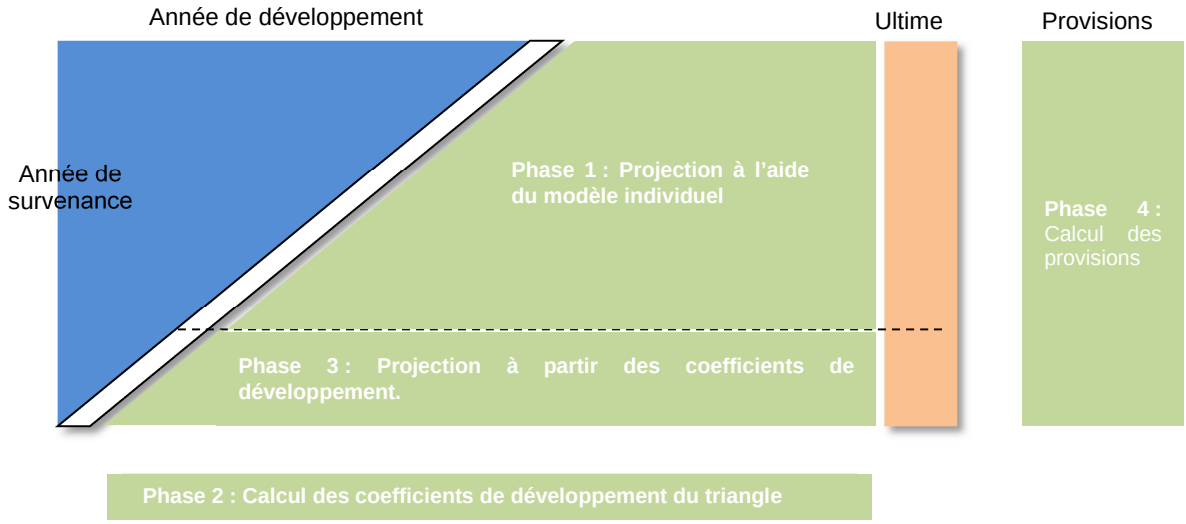
Lors de cette étape, il s'agit d'imbriquer le modèle individuel et le modèle agrégé. Le modèle final est déroulé en 4 phases :

1. Utilisation du modèle individuel pour projeter les paiements du trapèze supérieur (les sinistres en cours dans la 5ème année de développement) à l'ultime. Il est à préciser que les sinistres clos sont déjà vus à l'ultime,

2. Calcul des coefficients de développement du triangle à partir des flux réels et projetés issus du modèle individuel,

3. Projection des paiements du trapèze inférieur (les sinistres inférieurs à 4 ans de développement) à partir des coefficients de développement *Chain Ladder*,

4. Obtention des provisions à partir de l'ultime et les paiements réels.



SCHEMA 2 – Modèle final déterministe

2 PARTIE 2 - METHODES DE PROVISIONNEMENT AGREGEES

Les méthodologies traditionnelles s'appuient sur l'utilisation des données agrégées. La méthode la plus répandue est basée sur les cadences de règlements. Les triangles de données agrégées utilisent l'année de survenance, l'année de développement et des coefficients de développement. Ces méthodologies sont simples d'utilisation et robustes.

L'objet de cette partie est de présenter la méthodologie de *Chain Ladder*, le modèle de Mack et le modèle MLG. Ces deux derniers ont l'avantage de fournir une estimation de l'incertitude de prédiction. Puis, nous présentons la méthodologie *Bootstrap* et la gestion de la liquidation incomplète.

Les résultats obtenus serviront de base de comparaison au modèle de provisionnement ligne à ligne.

2.1 La méthode Chain Ladder

La méthode *Chain Ladder* est une méthode déterministe simple d'utilisation et basée sur un triangle des règlements ou un triangle de charges. Elle produit une estimation de la charge ultime qui dépend du niveau de règlements de la diagonale du triangle et d'une évolution future estimée sur les bases du développement des précédents exercices.

Les méthodes statistiques déterministes, dont *Chain Ladder*[9], reposent sur un ensemble d'hypothèses, comme le rappelle M. DENUIT, *« Les méthodes déterministes reposent sur l'hypothèse de stabilité du délai s'écoulant entre la survenance d'un sinistre et le(s) règlement(s), quel que soit l'exercice de survenance, en l'absence d'inflation, de changement de structure du portefeuille, des garanties des contrats, des franchises, et plus généralement de la gestion des sinistres. Si toutes ces hypothèses sont vérifiées sur une période suffisamment longue (au moins 5 ans), les méthodes déterministes peuvent être un premier outil intéressant pour prévoir la charge finale, en utilisant les cadences de règlement observées sur le passé »*.

Considérons les notations suivantes:

- i correspond à l'indice des années de survenance du sinistre ($i \in \{1,...,n\}$),
- j correspond à l'indice des années de développement ($j \in \{1,...,n\}$),
- $Y_{i,j}$ les dépenses correspondant aux sinistres survenus au titre de l'exercice de survenance i et payés après j années de développement,
- $C_{i,j}$ correspond aux paiements cumulés des sinistres survenus l'année i, en j années de développement :

$$C_{i,j} = \sum_{k=0}^{j} Y_{i,k}$$

Le facteur de développement individuel s'écrit $\hat{f}_{i,j} = \frac{c_{i,j+1}}{c_{i,j}}$ pour i=1, …, n et j=1, …,n.

[9] Au titre des méthodes déterministes, nous pouvons citer par ailleurs d'autres méthodes basées partiellement ou totalement sur les facteurs de développement (London-chain, Bornhutter–Ferguson …), les méthodes sur les cadences de règlement, les coûts moyens.

La méthode de *Chain Ladder* repose sur l'hypothèse d'indépendance des facteurs de développement $f_{i,j}$ de l'année de survenance.

Nous considérons alors le facteur de développement f_j pour chaque année de développement j :

$$\hat{f}_j = \frac{\sum_{i=0}^{n-j} C_{i,j+1}}{\sum_{i=0}^{n-j} C_{i,j}}$$

Il est alors possible de compléter le triangle des montants cumulés pour chaque exercice de survenance et chaque année de développement.

$$\hat{C}_{i,j} = C_{i,n-i} * \prod_{k=n-i+1}^{j-1} \hat{f}_k$$

Nous obtenons simplement :

- les charges ultimes par exercice de survenance $\hat{C}_{i,n} = C_{i,n-i} * \prod_{k=n-i+1}^{n-1} \hat{f}_k$
- les provisions par exercice de survenance $\hat{R}_i = \hat{C}_{i,n} - C_{i,n-i}$
- les provisions totales $\hat{R} = \sum_{i=1}^{n} \hat{R}_i$

Année de survenance	Année de développement 1	2	3	4	5	6	7	8	9	10	11	$\hat{C}_{i,n}$	$\hat{R}_i$
2003	10 911	50 971	71 708	81 461	89 691	96 441	104 012	107 132	108 340	108 801	109 638	109 638	0
2004	12 211	53 951	80 177	93 477	109 413	113 138	120 585	125 647	130 547	134 748	135 785	135 785	1 037
2005	12 805	52 417	74 373	89 218	99 233	106 513	113 034	117 457	121 231	123 597	124 548	124 548	3 317
2006	11 432	50 017	71 904	86 301	95 030	102 121	106 431	109 480	112 569	114 766	115 649	115 649	6 169
2007	12 124	51 389	74 043	91 068	100 852	110 624	118 925	123 117	126 591	129 061	130 054	130 054	11 130
2008	11 127	50 134	72 743	83 359	92 861	99 048	105 444	109 161	112 241	114 432	115 312	115 312	16 264
2009	10 769	46 500	67 936	78 546	90 318	96 595	102 833	106 458	109 462	111 598	112 457	112 457	22 139
2010	9 573	40 676	62 504	72 973	81 917	87 611	93 269	96 557	99 281	101 219	101 997	101 997	29 025
2011	8 233	41 966	64 121	75 378	84 618	90 499	96 343	99 740	102 554	104 555	105 360	105 360	41 239
2012	8 460	41 586	60 716	71 375	80 124	85 693	91 227	94 443	97 108	99 003	99 765	99 765	58 178
2013	9 446	42 086	61 445	72 233	81 087	86 723	92 323	95 578	98 275	100 192	100 963	100 963	91 517
												Total	**280 013**

Coefficients de passage

j	1	2	3	4	5	6	7	8	9	10
$\hat{f}_j$	4,4554	1,46	1,1756	1,1226	1,0695	1,0646	1,0353	1,0282	1,0195	1,0077

TABLEAU 5 – Application de *Chain Ladder* aux données auto rcc (en K€)

Il existe une correspondance entre les coefficients de passage et les cadences de règlements des sinistres : $Cadence_j = \frac{1}{f_j * \ldots * f_{n-1}}$.

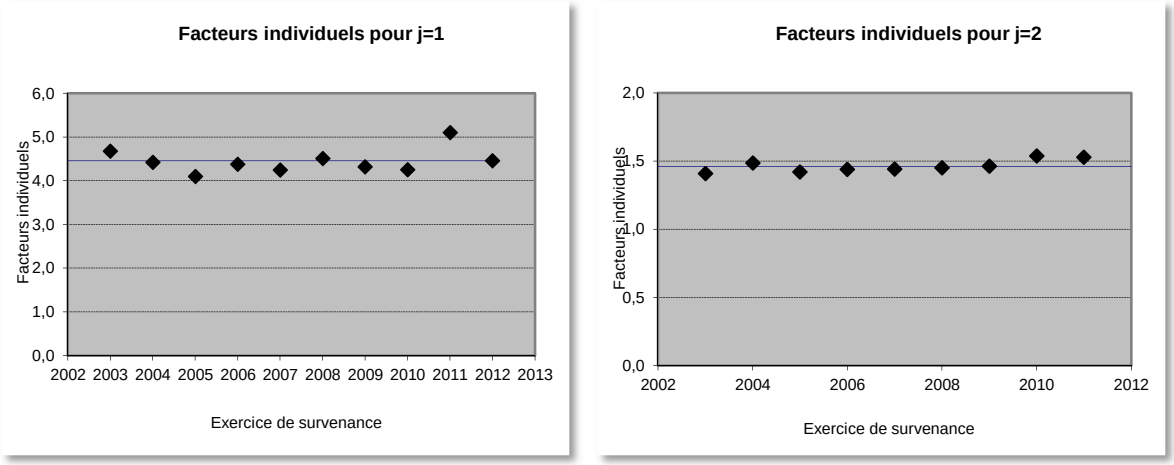
La branche automobile responsabilité civile corporelle a un développement long, c'est au bout de 8 ans que 95 % des paiements sont effectués.

j	1	2	3	4	5	6	7	8	9	10
Cadence$_j$	9,4%	41,7%	60,9%	71,5%	80,3%	85,9%	91,4%	94,7%	97,3%	99,2%

Il est important de valider l'hypothèse de constance des facteurs de développement afin de valider la méthode de *Chain Ladder*. Le triangle des facteurs de développement est défini par les facteurs de développement individuels $f_{i,j}$. Soit le triangle des facteurs de développement :

		Facteur de développement f_j									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Année de survenance	**2003**	4,672	1,407	1,136	1,101	1,075	1,079	1,030	1,011	1,004	1,008
	2004	4,418	1,486	1,166	1,170	1,034	1,066	1,042	1,039	1,032	
	2005	4,093	1,419	1,200	1,112	1,073	1,061	1,039	1,032		
	2006	4,375	1,438	1,200	1,101	1,075	1,042	1,029			
	2007	4,239	1,441	1,230	1,107	1,097	1,075				
	2008	4,505	1,451	1,146	1,114	1,067					
	2009	4,318	1,461	1,156	1,150						
	2010	4,249	1,537	1,167							
	2011	5,097	1,528								
	2012	4,455									

TABLEAU 6– Triangles des facteurs développement



GRAPHIQUE 1 – Facteurs de développement en j=1 et j=2

L'hypothèse fondamentale est vérifiée si, pour j=1,…,n-1, les facteurs individuels sont constants. Pour les premières années de développement, les facteurs individuels de développement sont dispersés autour de la droite. Dans ces conditions, Il est difficile de valider l'adéquation de la méthode *Chain Ladder*.

2.2 Le modèle de Mack

Le modèle de Mack est la version stochastique de la méthode de *Chain Ladder* et permet d'obtenir une estimation de l'incertitude de prédiction. Cette méthode repose sur trois hypothèses :

H1 : L'indépendance des années de survenance

$\{C_{i,0},...,C_{i,n}\} et \{C_{k,0},...,C_{k,n}\}$ sont indépendants pour i ≠ k

H2 : L'estimation de $C_{i,j+1}$ ne dépend que du facteur de développement f_j et $C_{i,j}$

$E\left[C_{i,j+1} / C_{i,0},...,C_{i,j}\right] = f_j * C_{i,j}$ pour j ∈ {0, 1, …, n-1} et pour i ∈ {0, 1, …, n}

Les facteurs fj sont estimés par les facteurs de développement de *Chain Ladder* et sont sans biais.

H3 : La variabilité de $C_{i,j+1}$ ne dépend que de la volatilité par période de développement σ_j et de $C_{i,j}$

$V\left[C_{i,j+1} / C_{i,0},...,C_{i,j}\right] = \sigma_j^2 * C_{i,j}$ pour j ∈ {0, 1, …, n-1} et pour i ∈ {0, 1, …, n}

Les (σ_j^2) sont estimés par : $\begin{cases} \hat{\sigma}_j^2 = \dfrac{1}{n-j-1} \sum_{i=0}^{n-j-1} C_{i,j} \left(\dfrac{C_{i,j+1}}{C_{i,j}} - \hat{f}_j \right)^2, 0 \le j \le n-2 \\ \hat{\sigma}_{n-1}^2 = \min\left(\dfrac{\hat{\sigma}_{n-2}^2}{\hat{\sigma}_{n-3}^2}, \min\left(\hat{\sigma}_{n-3}^2, \hat{\sigma}_{n-2}^2\right) \right) \end{cases}$

Ces estimateurs sont sans biais.

L'incertitude de prédiction pour un exercice de survenance est calculée à partir de l'écart quadratique moyen (MSEP : *mean squared error of prediction*) :

$$MSEP(\hat{R}_i) = E\left[(\hat{R}_i - R_i)^2 / C_{i,j/i+j\le n}\right] \Leftrightarrow MSEP(\hat{R}_i) = \hat{C}_{i,n}^2 \sum_{j=n-i}^{n-1} \frac{\hat{\sigma}_j^2}{\hat{f}_j} \left(\frac{1}{\hat{C}_{i,j}} + \frac{1}{\sum_{k=1}^{n-j} C_{i,k}} \right), i = 1,...,n$$

L'estimation du MSEP de la provision totale est alors donnée par :

$$MSEP(\hat{R}) = \sum_{i=1}^{n} \left(MSEP(\hat{R}_i) + \hat{C}_{i,n} \left(\sum_{k=i+1}^{n} \hat{C}_{k,n} \right) \sum_{j=n-i}^{n-1} \frac{2\hat{\sigma}_j^2}{\hat{f}_j^2 \sum_{z=0}^{n-j} C_{z,j}} \right),$$

L'erreur standard relative est obtenue de la manière suivante : $\dfrac{\sqrt{MSEP(\hat{R})}}{\hat{R}}$

La MSEP comprend à la fois l'erreur de processus qui mesure la variabilité interne du modèle et l'erreur d'estimation liée à l'estimation des vrais facteurs de développement f_j.

$\hat{R}$	$\sqrt{MSEP(\hat{R})}$	Erreur standard relative
280 013	29 173	10,42%

TABLEAU 7 – Résultats obtenus pour la garantie automobile RCC

2.3 Les modèles stochastiques MLG

Les modèles stochastiques *MLG* sont des modèles à facteurs qui peuvent fournir une estimation de l'incertitude de prédiction.

Les modèles linéaires généralisés ont été introduits par J. Nelder et R. Wedderburn en 1972. Ce modèle est une extension du modèle linéaire Normal et est formé de trois composantes : la composante aléatoire, la composante systématique et la fonction de lien[10].

- La composante aléatoire

Les variables aléatoires à expliquer $X_{i,j}$ sont indépendantes et suivent une loi de probabilité appartenant à la famille exponentielle. Leur densité est définie par la formule suivante :

$$f(x_{i,j};\theta_{i,j},\phi) = \exp\left(\frac{x_{i,j}\theta_{i,j} - b(\theta_{i,j})}{\phi / w_{i,j}} + c(x_{i,j},\phi)\right)$$

où :
- $\theta_{i,j}$ est le paramètre naturel,
- ϕ est le paramètre de dispersion strictement positif,
- b et c sont des fonctions spécifiques de la distribution, b étant deux fois dérivables à valeurs dans R et c à valeurs dans R^2.

- La composante déterministe

Soit M la matrice de régression et γ le vecteur des paramètres. La composante déterministe est notée η et est définie par $\eta = M\gamma$. Dans le cas du provisionnement, la composante systématique s'écrit :

$$\eta_{i,j} = \mu + \alpha_i + \beta_j, 0 \le i \le I, 0 \le j \le J$$

μ, α et β sont respectivement la constante, le premier facteur explicatif et le deuxième facteur explicatif.

- La fonction lien

La fonction g est appelée fonction lien (*link function*) et a pour rôle de linéariser l'espérance, c'est la fonction qui fait le lien entre la composante aléatoire et la composante systématique.

$$\begin{cases} g(\mu_{i,j}) = \eta_{i,j} \\ E[X_{i,j}] = \mu_{i,j} \\ V[X_{i,j}] = \phi v[\mu_{i,j}] \end{cases}$$

En utilisant une fonction lien log nous obtenons la relation suivante :

$$\ln(\mu_{i,j}) = \eta_{i,j} = \mu + \alpha_i + \beta_j \quad \Leftrightarrow \quad \mu_{i,j} = \exp\left(\mu + \alpha_i + \beta_j\right)$$

[10] une présentation plus détaillée de MLG est faite dans la partie Modélisation des paiements.

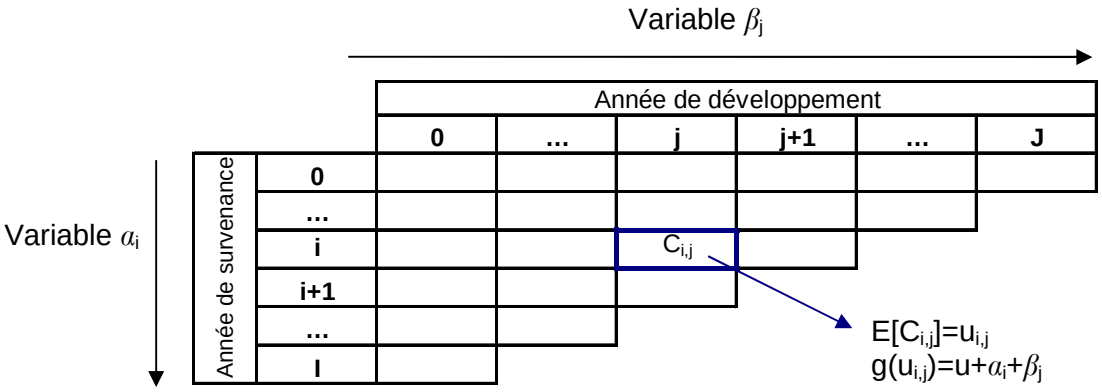
- Les moments

L'espérance et la variance s'obtiennent à partir de la fonction score ($U_{i,j} = \frac{\partial \log f(x_{i,j};\theta_{i,j},\phi)}{\partial \theta_{i,j}}$) :

$$\begin{cases} \mu_{i,j} = E[X_{i,j}] = b'(\theta_{i,j}) \\ V[X_{i,j}] = b''(\theta_{i,j})\phi = v[\mu_{i,j}]\phi \end{cases}$$

Remarque : v est la fonction de variance spécifique à chaque distribution.

La modélisation du triangle par un modèle MLG consiste à expliquer les paiements décumulés par les facteurs explicatifs que sont l'année de survenance et l'année de développement.



SCHEMA 3 – Illustration de l'utilisation des MLG

Le triangle des paiements est modélisé avec une distribution et fonction lien donnée. Dans notre cas, nous utiliserons la distribution de Poisson surdispersée et une fonction lien log.

Rappel : $X \sim P_{surd}(\lambda, \phi)$ si seulement si $\frac{X}{\phi} \sim P\left(\frac{\lambda}{\phi}\right)$ impliquant E[X]=λ et Var[X]=$\phi\lambda$.

L'estimation des incréments est obtenue par $\hat{\mu}_{i,j} = \exp\left(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j\right)$ et celle des provisions par $\hat{R} = \hat{E}[R] = \sum_{i+j>n}^{n} \hat{\mu}_{ij}$. L'erreur d'estimation des provisions $\hat{\sigma}_{\hat{R}}^2 = V\left(\hat{E}[R]\right)$ peut-être délicate, il est par contre plus simple d'utiliser des valeurs approchées basées sur la méthode Delta.

Nous avons $\hat{V}(R) = \hat{\phi} \sum_{i+j>n} V(\hat{\mu}_{ij})$ où le paramètre d'échelle est obtenu à partir des résidus de Pearson (r^p) standardisés $\hat{\phi} = \frac{1}{(N-p)} \sum_{i+j\leq n}^{n} \frac{(c_{ij} - \hat{\mu}_{ij})^2}{\sqrt{V(\hat{\mu}_{ij})}}$ *avec* $r_{ij}^P = \frac{(c_{ij} - \hat{\mu}_{ij})^2}{\sqrt{V(\hat{\mu}_{ij})}}$, N le nombre d'éléments du triangle et p le nombre de paramètres.

Le risque de prédiction est donnée par : $sep(\hat{R}) = \sqrt{\hat{V}(R) + \hat{\sigma}_{\hat{R}}^2}$

$\hat{R}$	$\sqrt{MSEP(\hat{R})}$	Erreur standard relative
280 013	25 665	9,17%

TABLEAU 8 – Résultats obtenus pour la garantie automobile RCC

2.4 Le Bootstrap

L'intérêt de cette approche est de pouvoir quantifier l'incertitude de prédiction par la construction d'intervalles de confiance. Les réserves seront considérées sous un angle probabiliste, nécessaire dans le cadre de la réglementation Solvabilité 2.

La technique du bootstrap a été introduite par Efron en 1979 et consiste, à partir d'un échantillon, à créer de nouveaux échantillons par tirage aléatoire avec remise. Cette méthode de ré-échantillonnage permet d'estimer la variabilité des provisions.

L'utilisation du bootstrap suppose que les éléments de l'échantillon de départ soient indépendants et identiquement distribués (iid). Les variables $X_{i,j}$ ne sont en général pas identiquement distribuées. Il est donc préférable d'avoir recours aux résidus du modèle, en particulier les résidus de Pearson.

Processus d'application de la méthode bootstrap :

A. Estimation des valeurs prédites $\left(\hat{u}_{i,j}\right)$ sur le triangle supérieur à l'aide d'un modèle linéaire généralisé ou de l'approche *Chain Ladder*.

B. Détermination des résidus de Pearson (r_{ij}^{p}), à partir des triangles décumulés observés $\left(x_{i,j}\right)$ et prédites $\left(\hat{u}_{i,j}\right)$, et les résidus ajustés de Pearson (r_{ij}^{ap}) :

$$r_{ij}^{p} = \frac{x_{ij} - \hat{u}_{i,j}}{\sqrt{V\left(\hat{u}_{i,j}\right)}} \text{ et } r_{ij}^{ap} = \sqrt{\frac{N}{N-p}} r_{ij}^{p}$$

où N est le nombre d'éléments de l'échantillon et p le nombre de degrés de liberté. Il est à préciser que les résidus ajustés $r_{1,J}$ et $r_{I,1}$ sont nuls.

C. Etape simulatoire, pour s= 1 à S (Nombre de simulations) :

- Ré-échantillonnage avec remise des résidus de Pearson ajustés,
- Transformation des triangles de résidus obtenus en triangles de dépenses, appelé « pseudo-triangles » :

$$(x_{ij})^{s} = \hat{u}_{i,j} + (r_{ij}^{ap})^{s} \sqrt{V\left(\hat{u}_{i,j}\right)}$$

- Ajuster un nouveau modèle MLG ou *Chain Ladder* au « pseudo-triangle »,
- Projection du triangle des paiements futurs $\hat{u}_{ij}$ (triangle inférieur), il est à préciser qu'à cette étape nous disposons uniquement de l'erreur d'estimation,
- Simuler les paiements sur le triangle inférieur à l'aide de la moyenne $\hat{u}_{ij}$ et de la variance $\phi v[\hat{u}_{ij}]$, c'est à cette étape que l'erreur de processus est intégrée à l'approche,
- Obtenir une nouvelle estimation des provisions $(R)^{s}$.

D. Détermination de la distribution empirique de la provision, nous pouvons estimer la moyenne, la variance et l'erreur de prédiction du montant de prévisions.

Nous avons « terminé » les triangles à l'aide de facteurs de queue, dont la méthodologie est exposée dans la partie suivante.

$\hat{R}$	$\sqrt{MSEP(\hat{R})}$	Erreur standard relative
279 370	25 300	9,06%

TABLEAU 9 – Résultats obtenus pour la garantie automobile RCC

2.5 La gestion de la liquidation incomplète

En général, nous ne disposons pas du développement complet de la sinistralité. C'est notamment le cas pour la garantie auto RCC où le déroulement se fait sur des dizaines d'années.

Pour ne pas sous-estimer la charge ultime, il est donc nécessaire d'estimer une queue de développement du triangle.

Les « *tail factors* » peuvent être estimés à partir des coefficients de déroulements (*Chain Ladder*).

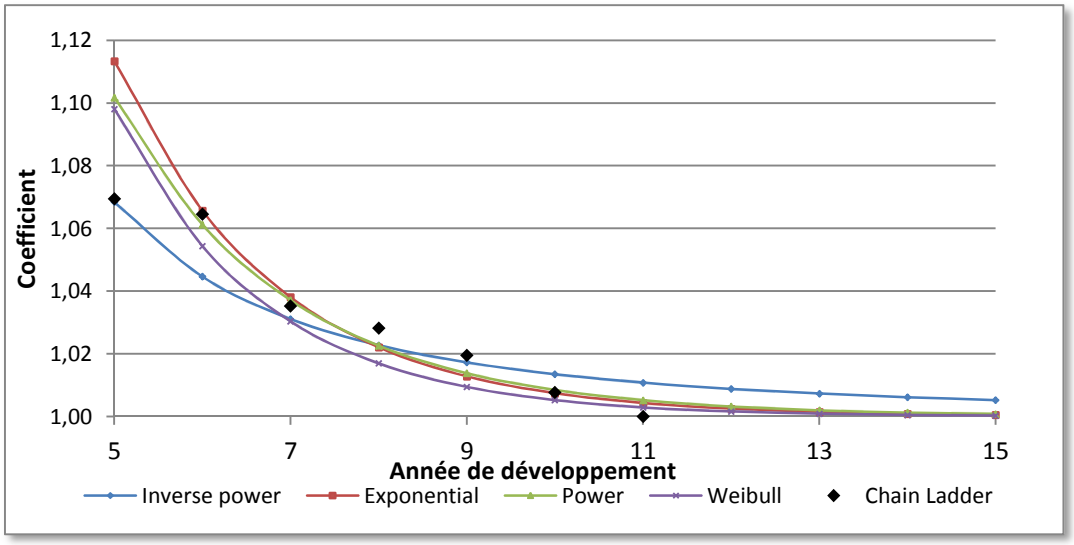
Nous présentons dans la suite deux méthodes d'estimation des « tail factors ». La première méthode, utilisée dans les modèles reposant sur les hypothèses de Chain-Ladder, est basée sur les coefficients de déroulement. La deuxième méthode, utilisable dans les modèles de type MLG, repose sur les paramètres de régression.

Extrapolation des facteurs de développement

A partir des facteurs de développement, nous cherchons à ajuster une fonction du type y=f(x) où y représente les facteurs de développement et x les années de développement. L'extrapolation des points se fera à partir de la fonction ajustée pour les années de développement supérieures à la dernière année connue. Il existe plusieurs fonctions possibles, nous citons les plus répandues :

- Inverse power : $f_{a,b}(x) = 1 + \frac{a}{x^b}$
- Exponentielle : $f_{a,b}(x) = 1 + a * e^{-bx}$
- Power : $f_{a,b}(x) = a^{b^x}$
- Weibull : $f_{a,b}(x) = \frac{1}{1 - e^{-a*b^x}}$

L'estimation des paramètres a et b se fait avec une régression linéaire à l'aide d'une transformation log ou log-log. Nous considérons que la clôture de triangle a lieu lorsque le coefficient de passage est inférieur à 1.



GRAPHIQUE 2 – Estimation des *tail factors*

Par la suite, nous choisissons de travailler avec l'ajustement power de coefficient a=3,07 et b=0,613.

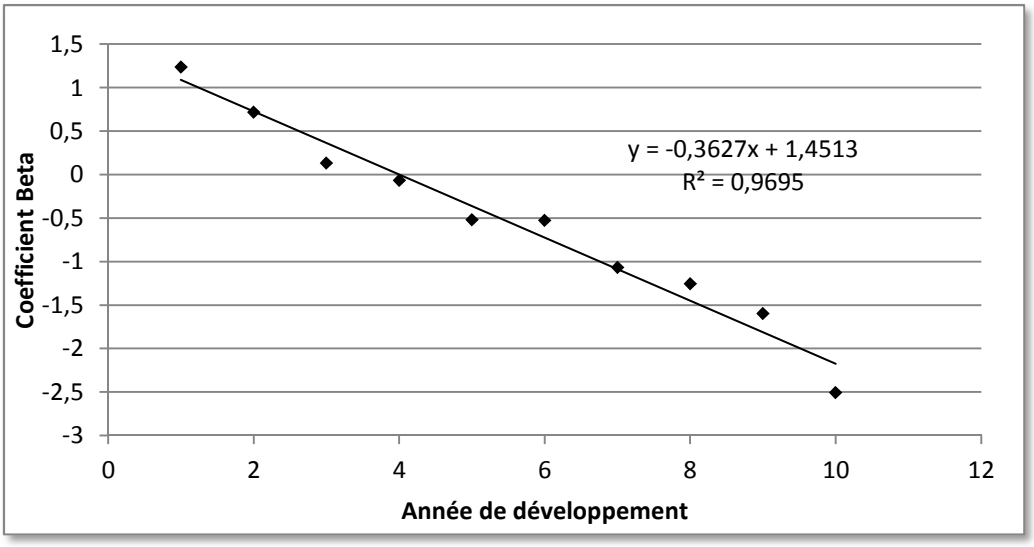
Extrapolation des paramètres de développement

Dans le cadre des modèles *MLG*, le passage d'une année de développement à la suivante est représenté par les paramètres β_j. Nous cherchons à extrapoler les coefficients de régression.

L'extrapolation se fait à partir d'une régression linéaire simple de type y=ax+b où y représente les coefficients de régression et x les années de développement.

L'extrapolation des coefficients de régression se fera à partir de la fonction ajustée pour les années de développement supérieures à la dernière année connue.

Nous appliquons l'hypothèse de continuer l'estimation des β_j tant que l'espérance de l'incrément associé est supérieur à 1.



GRAPHIQUE 3 – Estimation des paramètres

L'utilisation des méthodes d'extrapolation sur les triangles nous donne les montants de provisions suivants :

Extrapolation	$\hat{R}$
des facteurs	296 749
des paramètres	301 147

2.6 Le triangle agrégé de référence

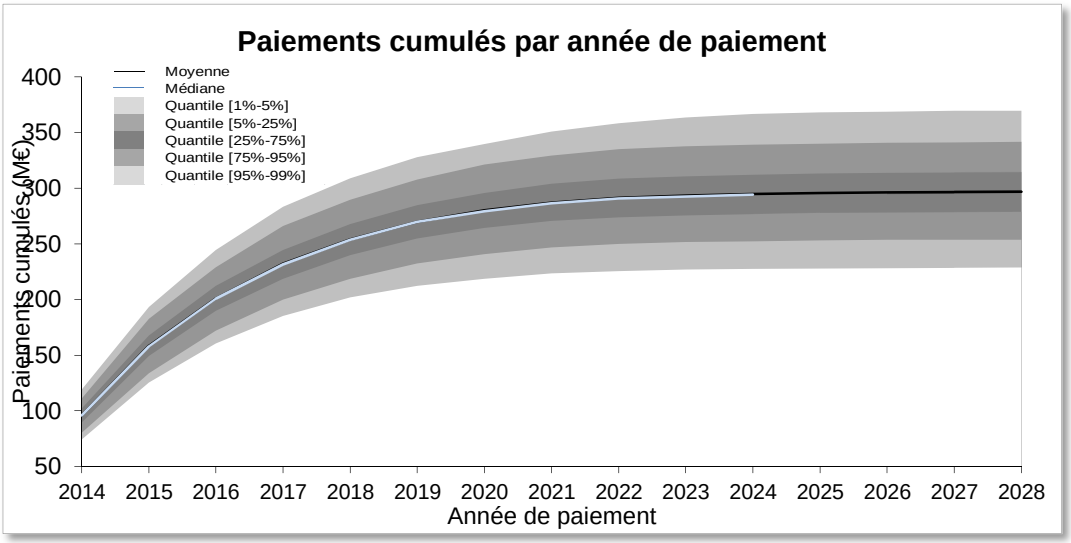
Le triangle agrégé de référence est déroulé en utilisant la méthode de *Chain Ladder* avec des *tail factors* estimés à partir de la fonction *power*.

i	f_i	i	f_i	i	f_i
1	4,45544	11	1,00515	21	1,00004
2	1,46000	12	1,00316	22	1,00002
3	1,17556	13	1,00193	23	1,00001
4	1,12258	14	1,00118	24	1,00001
5	1,06950	15	1,00073	25	1,00001
6	1,06458	16	1,00044	26	1,00000
7	1,03525	17	1,00027	27	1,00000
8	1,02822	18	1,00017	28	1,00000
9	1,01951	19	1,00010	29	1,00000
10	1,00769	20	1,00006	30	1,00000

TABLEAU 10 – Extrapolation des facteurs de développement

Les facteurs de développement f_{11} à f_{30} sont issus de l'extrapolation. A partir de la 30ème année de développement, le facteur de développement est très proche de 1 et son impact sur les provisions devient négligeable.

Nous réalisons 5 000 simulations stochastiques incorporant à la fois l'erreur d'estimation et de processus[11]. Nous obtenons les résultats suivants :



GRAPHIQUE 4 – Estimation des paiements cumulés

[11] Simulations effectuées à l'aide du logiciel IBNRS et de la méthodologie Bootstrap Mack.

La provision moyenne sur l'ensemble des exercices du triangle est de 296,8 M€ et l'écart type de 27,5 M€.

Quantile	Provision
1%	228 746 277 €
5%	253 831 419 €
25%	278 754 918 €
50%	296 201 295 €
75%	314 697 350 €
95%	341 789 027 €
99%	370 533 537 €

TABLEAU 11 – Quantiles des provisions

3 PARTIE 3 - MODELISATION DE LA DUREE DU SINISTRE

Dans cette partie, nous introduirons le concept général des modèles de durée ainsi que les données utilisées. Puis, nous présenterons de manière théorique et appliquerons les modèles de durée pour obtenir une courbe de survie de référence et des facteurs explicatifs. Enfin, nous appliquerons le modèle aux sinistres non clos de la date d'inventaire.

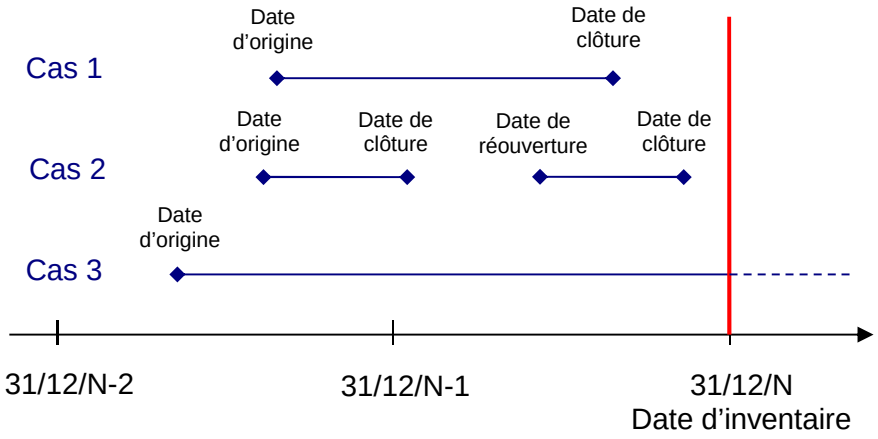
3.1 Le cadre général

La durée de survie désigne le temps écoulé entre deux événements donnés. Dans ce mémoire, la durée de survie correspond au temps écoulé entre la date de survenance d'un sinistre et la date de clôture de ce même sinistre.

L'analyse des durées de survie consiste à étudier le délai de survenance de l'événement clôture du sinistre communément appelé le temps de survie.

L'étude de survie requiert la définition de plusieurs dates :

- date d'origine : date de survenance du sinistre à la MAIF,
- date d'inventaire : date à laquelle l'étude est effectuée, au-delà de cette date aucune information n'est disponible,
- date de clôture : date de fermeture et de classement du dossier sinistre (donnée disponible uniquement si le dossier est clôturé avant la date d'inventaire).



SCHEMA 4– Illustration de la notion de censure

Cas 1 : l'événement de clôture a lieu avant la date d'inventaire. Le temps de survie correspond à l'écart entre la date d'origine et la date de clôture.

Cas 2 : l'événement de clôture a lieu avant la date d'inventaire. Le temps de survie correspond à l'écart entre la date d'origine et la dernière date de clôture.

Cas 3 : la date d'inventaire a lieu avant la date de clôture. Il y a absence d'information sur la date de clôture, la seule information est que la date de clôture se trouve après la date d'inventaire. Le temps de censure correspond à l'écart entre la date d'origine et la date d'inventaire.

3.1.1 Particularités des données

Les données de durée sont générées par des variables aléatoires positives, soit la variable aléatoire X ayant pour intervalle [0, +∞[. L'analyse de la variable aléatoire X est effectuée non plus au travers de la fonction de répartition mais au travers de la fonction de survie et de la fonction de hasard.

Les données de survies sont confrontées aux données incomplètes. Ceci peut être la conséquence des données tronquées ou censurées. Dans notre étude, nous sommes confrontés au phénomène de censure à droite où, sur certains sinistres, la date de clôture est supérieure à la date d'inventaire.

3.1.2 Distribution de la durée de survie

La répartition des données de survie peut être effectuée à partir de l'une des 5 fonctions suivantes :

- la fonction de survie S(t),
- la fonction de répartition F(t),
- la densité de probabilité f(t),
- la fonction de hasard h(t),
- la fonction de hasard cumulé H(t).

La fonction de survie S est la probabilité de survivre jusqu'à l'instant t :

$$S(t) = P(X > t);\ t > 0$$

La fonction S est décroissante telle que S(0)=1 et $\lim_{t\to\infty} S(t) = 0$.

La fonction de survie conditionnelle découle de la fonction de survie :

$$S_u(t) = P(X > t+u / X > u) = \frac{P(X > u+t)}{P(X > u)} = \frac{S(u+t)}{S(u)}$$

La fonction de répartition F représente la probabilité de mourir avant l'instant t :

$$F(t) = P(X < t) = 1 - S(t):$$

La fonction de répartition est le complément à 1 de la fonction de survie.

La densité de probabilité f est la fonction telle que pour tout t≥0 :

$$F(t) = \int_0^t f(u)du$$

Si la fonction de répartition F admet une dérivée au point t alors

$$f(t) = \frac{d}{dt}F(t) = \lim_{h\to 0}\frac{P(t \le X \le t+h)}{h} = F'(t) = -S'(t)$$

La densité de probabilité représente la probabilité de mourir dans un petit intervalle de temps après l'instant t.

La fonction de hasard h, appelée aussi risque instantané, est la probabilité de mourir dans un petit intervalle de temps après t sachant le fait d'être en vie à l'instant t ;

$$h(t) = \lim_{h \to 0} \frac{P(t \le X \le t+h / X \ge t)}{h} = \lim_{h \to 0} \frac{1}{h} \frac{P(t \le X \le t+h)}{P(X > t)} = \frac{f(t)}{S(t)} = -\ln(S(t))'$$

La fonction de hasard cumulé H est l'intégrale du risque instantané h :

$$H(t) = \int_0^t \lambda(u)du = -\ln(S(t))$$

A partir de cette relation, la fonction de survie peut être écrite en fonction du taux de hasard cumulé :

$$S(t) = \exp(-H(t)) = \exp(-\int_0^t h(u)du)$$

Cette équation illustre la relation suivante :

$$f(t) = h(t).\exp(-H(t)) = h(t).\exp(-\int_0^t h(u)du)$$

D'après la définition de la fonction de survie conditionnelle et de l'équation ci-dessus, nous obtenons :

$$S_u(t) = \exp(-\int_u^{u+t} h(u)du)$$

Les moments associés à la distribution sont :

$$E[X] = \int_0^\infty udF(u) = -\int_0^\infty udS(u) = \int_0^\infty S(u)du$$

$$V[X] = 2\int_0^\infty uS(u)du - E[X]^2$$

3.1.3 Notion de censure

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, les données sont souvent recueillies partiellement, notamment à cause des processus de censure et de troncature. Les données censurées ou tronquées découlent de l'indisponibilité de toute l'information.

Les troncatures diffèrent des censures au sens où elles concernent l'échantillonnage. S'il y a troncature, une partie des sinistres (donc des X_i) ne sont pas observables, l'étude ne porte ainsi que sur un sous-échantillon.

La durée de vie est dite censurée à droite si le sinistre n'est pas clos à la date d'inventaire. En présence de censure à droite, les durées de vie ne sont pas toutes observées. La date de clôture est ainsi supérieure à la date d'inventaire.

Dans le cadre de ce mémoire, il est à préciser les notions de troncature, de censure à gauche et de censure par intervalle ne seront pas détaillées. Les données sont confrontées à une censure à droite de type 3 c'est-à-dire que la date de censure est une variable aléatoire.

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. Pour le sinistre i, considérons :

- son temps de survie X_i,
- son temps de censure C_i,
- la durée réellement observée T_i.

Soient un premier échantillon de taille n de durées de survie (X_1, …, X_n) et un second échantillon indépendant du premier de taille n composé de temps de censures (C_1, …, C_n). Sur la période de l'étude, nous ne constatons pas de modifications des procédures et des systèmes de gestion pouvant remettre en cause l'hypothèse d'indépendance entre ces 2 variables.

Pour une censure à droite, au lieu d'observer le vecteur (X_1, …, X_n) nous observerons le couple (T_1, D_1), …, (T_n, D_n) avec :

$$T_i = X_i \wedge C_i \quad et \quad D_i = \begin{cases} 1 \; si \; X_i = T_i \\ 0 \; si \; X_i \neq T_i \end{cases}$$

L'information disponible peut être résumée par :

- la durée réellement observée T_i,
- un indicateur D_i :
  - D_i=1 : la durée du sinistre est observée (observation de la vrai durée),
  - D_i=0 : la durée du sinistre est censurée (observation de durées incomplètes).

3.1.4 Fonction de vraisemblance

Nous considérons l'indépendance des vecteurs X et C et supposons que les variables X_i et C_i ont pour densités respectives f_X et f_C et pour survies respectives S_X et S_C. La distribution de X est caractérisée par le paramètre θ.

La vraisemblance[12] de l'échantillon (T_1, D_1), …, (T_n, D_n) s'écrit :

$$L(\theta) = \prod_{i=1}^{n} L_i = \kappa \prod_{i=1}^{n} f_\theta(t_i)^{di} S_\theta(t_i)^{1-di} = \kappa \prod_{i=1}^{n} \lambda_\theta(t_i)^{di} S_\theta(t_i)$$

Le terme κ regroupe les informations en provenance de la loi de la censure qui ne dépend pas du paramètre θ. Il est à préciser que si le mécanisme de censure est indépendant de l'événement étudié alors la censure est dite non informative.

[12] Cf. Annexe pour le développement de la vraisemblance.

Lorsque le modèle comporte p variables explicatives (Z_1, …, Z_p), la vraisemblance s'écrit de la manière suivante :

$$L(\theta) = \kappa \prod_{i=1}^{n} \lambda_{\theta/Z}(t_i)^{di} S_{\theta/Z}(t_i)$$

3.1.5 Les modèles statistiques

Dans notre approche nous nous intéresserons aux modèles non paramétriques et paramétriques :

- modèles non paramétriques : aucune hypothèse a priori est faite sur la forme de la loi de survie. Nous cherchons à approximer l'une des différentes fonctions caractérisant la distribution observée (modèle de Kaplan Meier, méthode actuarielle),
- modèles paramétriques : utilisation d'une forme de distribution donnée (loi exponentielle, loi weibull…) et estimation des paramètres. Ces modèles peuvent prendre en compte l'effet de facteurs exogènes.

3.2 Les données sinistres

Afin de modéliser la durée des sinistres, il est nécessaire de bien connaître leurs principales caractéristiques. Une analyse descriptive de la durée des sinistres est présentée et permet d'identifier quelques caractéristiques principales de ces sinistres.

Nous présentons quelques définitions :

- Sinistre clos : Sinistre pour lequel le gestionnaire estime que plus aucun acte de gestion n'est nécessaire (sauf nouvel élément externe imprévisible). Par conséquent, les dépenses relatives à ce sinistre sont terminées.
- Sinistre clos sans suite : Sinistre clos et pour lequel aucun règlement n'a été effectué. Ce type de sinistre peut exister pour les raisons suivantes :
  - sinistre ouvert à tort dans la catégorie RCC (erreur de saisie),
  - sinistre dont l'instruction conduit à conclure à un refus d'indemnisation (absence de préjudice, ou préjudice n'entrant pas dans le cadre des garanties).
- Durée : écart en jours entre la date d'ouverture et la dernière date de clôture (sinistres clôturés) ou date d'inventaire (les sinistres ouverts).
- Coût d'un sinistre en fin d'inventaire :
  - sinistres clos : montant des dépenses effectuées sur ce sinistre,
  - sinistres non clos : maximum entre les dépenses effectuées sur le sinistre et l'évaluation de ce sinistre.

Les données sont extraites du système d'information décisionnel, construit à partir d'extraits du système d'information de production.

Nous disposons d'information sur la date d'ouverture, la date de clôture et les facteurs explicatifs liés aux sinistres (Cout Technique Brute[13], nombre de victimes…) et aux victimes (taux IPP) pour les exercices de survenance 2003 à 2013.

[13] Coût Technique Brut : somme des dépenses et évaluation du gestionnaire.

Les données extraites sont en adéquation avec les données provenant des états de gestions[14].

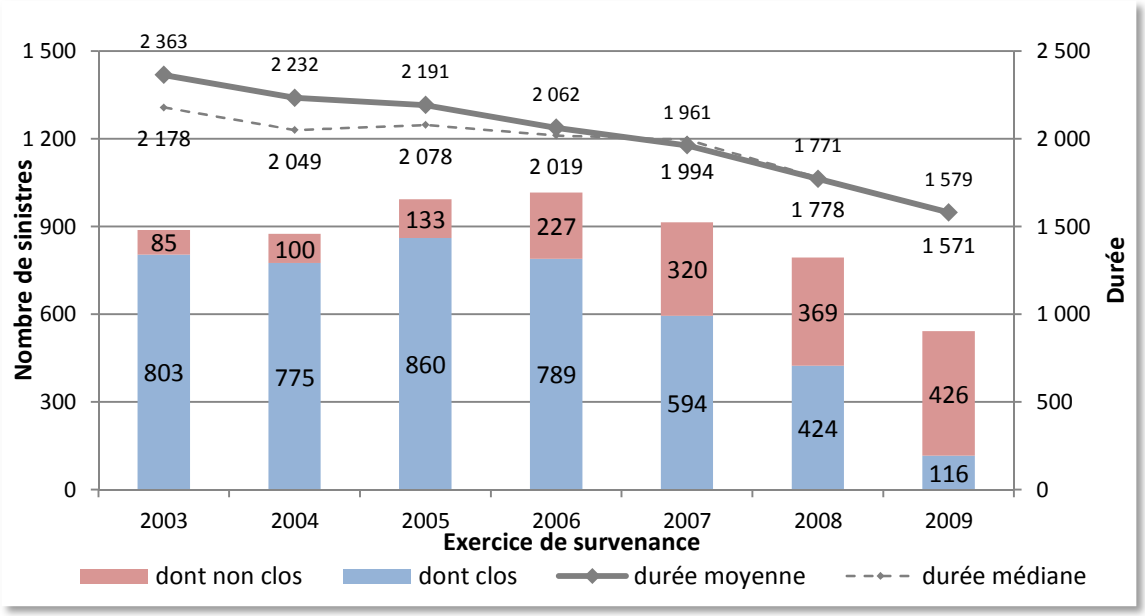
La population d'étude est restreinte aux sinistres en cours après 4 ans de développement car à partir de ce seuil, il n'existe quasi plus de sinistres tardifs et les sinistres évalués en forfait de gestion ainsi que les sinistres sans suite sont clôturés. De plus, à partir de 4 ans les caractéristiques du sinistre telles que le nombre de victimes et le taux d'IPP sont stabilisés c'est-à-dire n'évolueront pas par la suite.

Dans un premier temps, nous nous intéressons aux nombre de sinistres, ouverts et clôturés, et la durée de ces sinistres. Puis, nous ferons une analyse descriptive des facteurs explicatifs.

3.2.1 La variable d'intérêt

3.2.1.1 Le nombre et la durée des sinistres

Dans cette partie, nous nous intéressons au nombre de sinistres, ouverts et clos, et à la durée de ces sinistres de la population d'étude. La population d'étude pour la modélisation de la durée de vie des sinistres est composée de 6 021 sinistres.



GRAPHIQUE 5 – Répartition et durée moyenne des sinistres par exercice de survenance

Clé de lecture : pour l'exercice de survenance 2003 nous observons 888 sinistres en cours 4 années après la date de survenance et 85 d'entre eux, soit 9,6%, sont encore non clos au 30/09/2013.

Le nombre de sinistres en cours 4 ans après la date de survenance du sinistre varie en général entre 800 et 1000 sinistres par exercice de survenance.

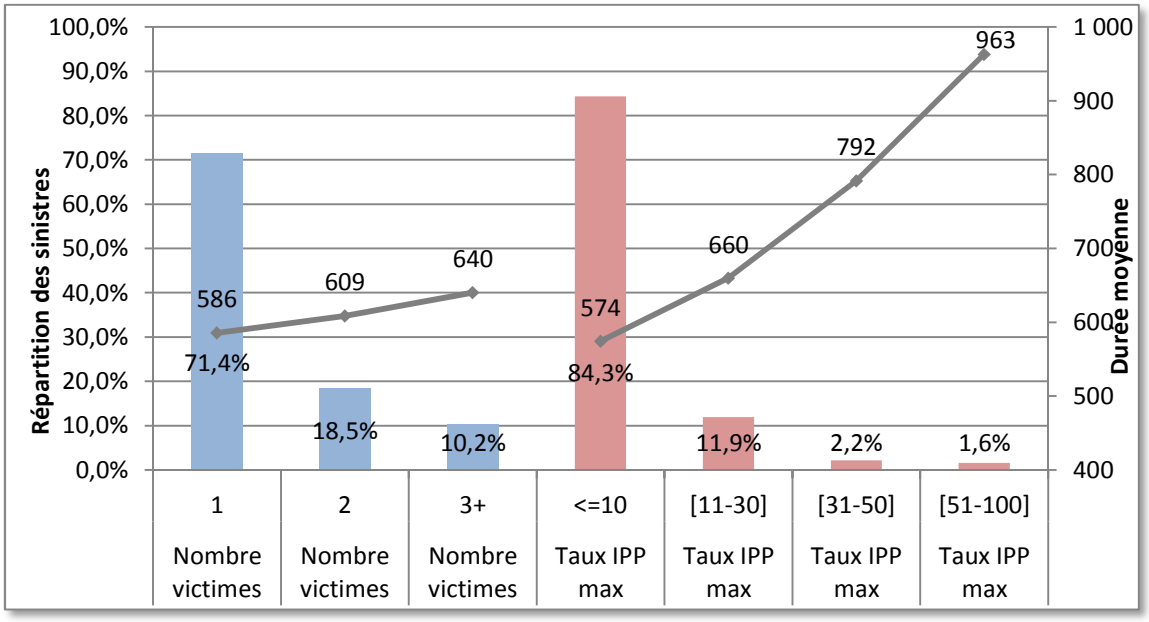
[14] Il est à noter que l'analyse descriptive est réalisée au 01/10/2013, ainsi l'année 2013 n'est que partiellement observée.

Il est difficile de comparer les durées moyennes des exercices de survenance entre elles. Néanmoins, nous constatons une moyenne supérieure à la médiane signifiant une distribution biaisée tirée par un petit nombre de sinistres avec des durées de vie élevées.

3.2.2 Les variables explicatives

Le système d'information dans sa configuration actuelle ne permet pas d'accéder à l'ensemble de l'information initialement souhaitée. Par la suite, les variables explicatives utilisées seront le nombre victimes, directe ou ayant droit, impacté par le sinistre et le taux d'incapacité physique permanente (IPP) maximal parmi la ou les victimes. A partir de la 4ème année de développement ces variables sont stables dans le temps.

Variable	Description	Modalités
Nombre de victimes	Le nombre de victimes, directe ou ayant droit, impacté par le sinistre	1 : 1 victime 2 : 2 victimes 3+ : 3 et plus victimes
Taux IPP	Le taux d'IPP maximal, parmi le ou les victimes, classé en 3 catégories	[0-10] [11-30] [31-50] [51 -100]



GRAPHIQUE 6 – Répartition et durée moyenne des sinistres

La modalité 1 victime représente 71,4% de la population et a pour durée moyenne 586 jours alors que la modalité 3 victimes et plus représente 10,2% et a pour durée moyenne 640 jours. Nous constatons que plus il y a de victimes, plus la durée est longue.

Pour la variable taux d'IPP nous observons une tendance similaire mais plus marquée. Ainsi, un sinistre ayant une victime avec un taux d'IPP supérieur à 51% durera presque que 2 fois plus longtemps qu'un sinistre ayant une victime avec un taux d'IPP inférieur à 10%.

L'analyse univariée illustre que la durée moyenne des sinistres est liée au nombre de victimes et à la gravité des victimes.

3.3 La démarche d'analyse

L'objectif du modèle de durée est d'obtenir une estimation de la probabilité de survie différenciée en fonction de facteurs explicatifs et d'apporter une approche projective afin de ne pas être limité par la profondeur de l'historique des données.

Nous nous limiterons aux sinistres ouverts 4 ans après l'occurrence du sinistre c'est-à-dire que nous modéliserons les sinistres ayant une durée de vie supérieure à 4 ans.

SCHEMA 5 – Processus d'analyse

3.4 Approche non paramétrique

3.4.1 Théorie des modèles non paramétriques

Dans cette approche aucune hypothèse a priori n'est faite sur la forme de la loi de survie. Nous cherchons à estimer directement cette fonction à partir des données. Néanmoins, les données doivent être en quantité suffisante pour obtenir des estimations fiables de la fonction de survie.

La fonction peut être caractérisée par différentes fonctions (fonctions de hasard, fonction de densité…). Les modèles présentés s'appuieront sur l'estimation empirique de la fonction de survie ou du hasard cumulé. Dans cette partie, nous nous intéresserons aux estimateurs de Kaplan-Meier et Actuarielle de la survie.

NB : d'autres méthodologies existent telles que l'estimateur de Nelson Aalen du risque cumulé mais ne seront pas présentées dans ce document.

3.4.1.1 Estimateur de Kaplan-Meier de la survie

Le principe de l'estimateur de Kaplan-Meier est fondé sur les probabilités conditionnelles et la discrétisation de l'intervalle temps en fonction des décès et des censures.

Posons $0 < t_1 < t_2 < t_3$, l'idée est d'estimer la probabilité de vivre en t_2 sachant qu'il vit en t_1, puis de calculer la probabilité de vivre en t_3 sachant qu'il vit en t_2. De manière plus mathématique, nous obtenons l'équation suivante :

$$P(X > t_3) = P(X > t_3 / X > t_2) * P(X > t_2 / X > t_1) * P(X > t_1 / X > 0)$$

La généralisation de cette équation en considérant la discrétisation en n intervalles $t_{(i)}$ (i=1, …, n) avec $t_{(0)}=0$ s'écrit comme suit :

$$P(X > t_j) = \prod_{i=1}^{j} P(X > t_i / X > t_i) = \prod_{i=1}^{j} p_i = \prod_{i=1}^{j} (1 - q_i)$$

Considérons les notations suivantes :

- p_i étant la probabilité de survivre sur l'intervalle $]T_{i-1}, T_i]$ sachant qu'on est vivant à l'instant T_{i-1},
- q_i étant la probabilité de décéder sur l'intervalle $]T_{i-1}, T_i]$ sachant qu'on est vivant à l'instant T_{i-1} et le complément à 1 de p_i ($p_i=1-q_i$),
- d_i le nombre de décès en T_i, prenant la valeur $d_i=1$ s'il y a sortie par décès en T_i et $d_i=0$ s'il y a sortie par censure en T_i,
- r_i le nombre d'individus à risque, c'est-à-dire ni décédés ni censurés, de subir l'événement juste avant le temps T_i,
- n le nombre d'individus en t=0.

Un estimateur de la probabilité q_i de décéder sur l'intervalle $]T_{i-1}, T_i]$ sachant qu'on est vivant à l'instant T_{i-1} est :

$$\hat{q}_i = \frac{d_i}{r_i} = \frac{d_i}{n - i + 1}$$

L'estimateur de Kaplan-Meier s'écrit sous la forme suivante :

$$\hat{S}(t) = \prod_{Ti \le t} \left(1 - \frac{d_i}{r_i}\right) = \prod_{Ti \le t} \left(1 - \frac{1}{r_i}\right)^{di} = \prod_{Ti \le t} \left(1 - \frac{1}{n - i + 1}\right)^{di}$$

Dans le cas d'*ex aequo* de nature différente (décès et censure), par convention les observations non censurées précèdent toujours les observations censurées. Dans le cas d'*ex aequo* de même nature, c'est-à-dire plusieurs décès en même temps, la variable d_i est strictement supérieur à 1 ($d_i>1$) et l'estimateur de Kaplan-Meier est :

$$\hat{S}(t) = \prod_{Ti \le t} \left(1 - \frac{d_i}{r_i}\right)$$

L'estimateur de Kaplan-Meier est une fonction en escalier décroissante et continue à droite.
La variance de l'estimateur de Kaplan-Meier peut être obtenue à partir de l'estimateur de Greenwood :

$$\hat{V}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{Ti \le t} \frac{d_i}{r_i (r_i - d_i)}$$

A partir de la propriété de normalité asymptotique de l'estimateur de Kaplan-Meier et de l'estimateur de variance de Greenwood, un intervalle de confiance peut être défini avec α le degré de confiance et la valeur u issue de la loi Normale centrée réduite :

$$IC(\alpha) = \left[\hat{S}(t) \pm u_{\alpha/2} \sqrt{\hat{V}(\hat{S}(t))} \right]$$

L'estimateur du risque cumulé de Breslow est obtenu à partir de l'estimateur de survie de Kaplan-Meier :

$$\hat{\Lambda}(t) = -\log(\hat{S}(t))$$

3.4.1.2 Estimateur actuarielle de la survie

L'estimateur actuarielle est fondé sur une approche similaire à l'estimateur de Kaplan-Meier à la différence de l'utilisation d'intervalles temps fixes. Ainsi, les probabilités conditionnelles sont estimées sur des intervalles temps contrairement à celle de Kaplan-Meier calculées au temps d'événements.

Nous reprenons les notations utilisées dans le cadre de Kaplan-Meier et précisons les notations suivantes :

- d_i le nombre de décès sur l'intervalle $]t_{i-1}, t_i]$,
- r_i le nombre d'individus à risque, c'est-à-dire ni décédés ni censurés, de subir l'événement sur l'intervalle $]t_{i-1}, t_i]$,
- n_{i-1} le nombre d'individus à risque t_{i-1}.

La répartition des censures est uniforme dans l'intervalle et ainsi, les individus sont exposés en moyenne un demi-intervalle. Le nombre d'individus à risque pour l'intervalle $]t_{i-1}, t_i]$ est :

$$r_i = n_{i-1} - \frac{c_i}{2}$$

Un estimateur de la probabilité q_i de décéder sur l'intervalle $]t_{i-1}, t_i]$ sachant qu'on est vivant à l'instant t_{i-1} est :

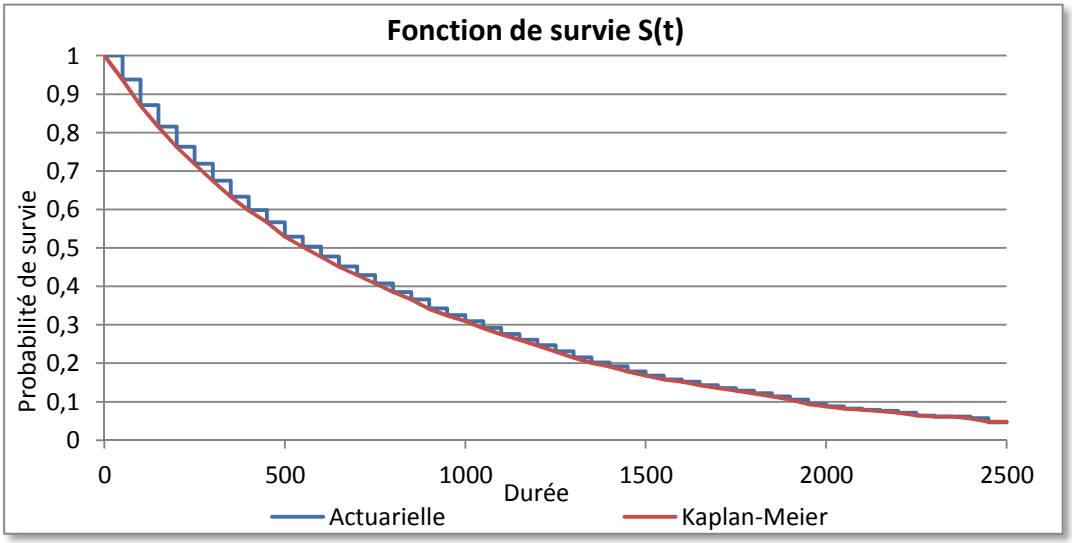
$$\hat{q}_i = \frac{d_i}{r_i}$$

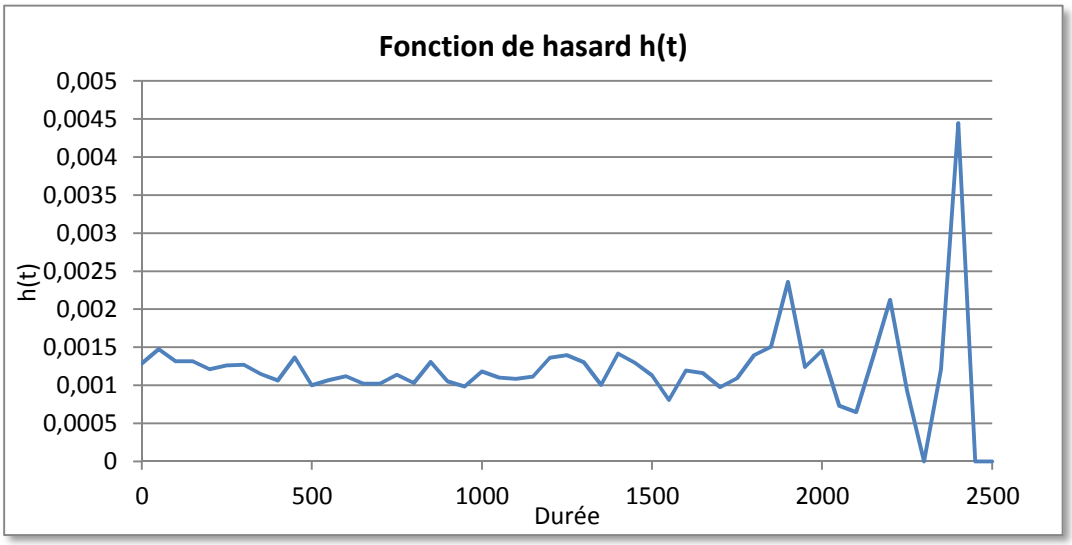
L'estimateur actuarielle de la fonction de survie s'écrit sous la forme suivante :

$$\hat{S}(t) = \prod_{ti \le t} \left(1 - \frac{d_i}{r_i}\right)$$

La variance, l'intervalle de confiance et l'estimateur du risque cumulé de Breslow s'obtiennent de la même manière que pour le modèle de Kaplan-Meier.

3.4.2 Application non paramétrique





GRAPHIQUES 7 – Fonction de survies et de hasard

La fonction de hasard est constante ou très légèrement décroissante sur la période. Il est à préciser que la volatilité constatée sur les durées élevées s'explique par un nombre faible de sinistres.

3.5 Choix d'une distribution avant estimation

3.5.1 Théorie des modèles paramétriques

Les modèles paramétriques s'appuient sur la connaissance de la distribution de survie qui appartient à une loi paramétrique donnée. Les modèles paramétriques sont le plus souvent caractérisés à partir de la fonction de risque instantané h(t) qui peut dépendre d'un ou

plusieurs paramètres. Cette fonction dispose de formes variées : constante, monotone, en forme de ∩, en forme de ∪, croissante ou décroissante.

L'une des difficultés lors de l'ajustement d'un modèle paramétrique est celui du choix de la distribution sous laquelle vont se faire les estimations. Une analyse a priori des données de survies offre une visibilité sur les distributions à étudier.

Nous présentons les distributions d'intérêt puis une aide graphique au choix d'une distribution.

3.5.1.1 Lois de survie

Les lois de variables aléatoires positives et continues sont appelées lois de survie. Ces lois admettent une densité de probabilité nulle sur l'intervalle]-∞, 0]. La loi de survie est, en général, définie à partir de la fonction de hasard et de la fonction de survie.

Les lois de survie paramétriques sont multiples, nous nous intéressons de manière détaillées aux lois exponentiel, weibull, log-normale et log-logistique.

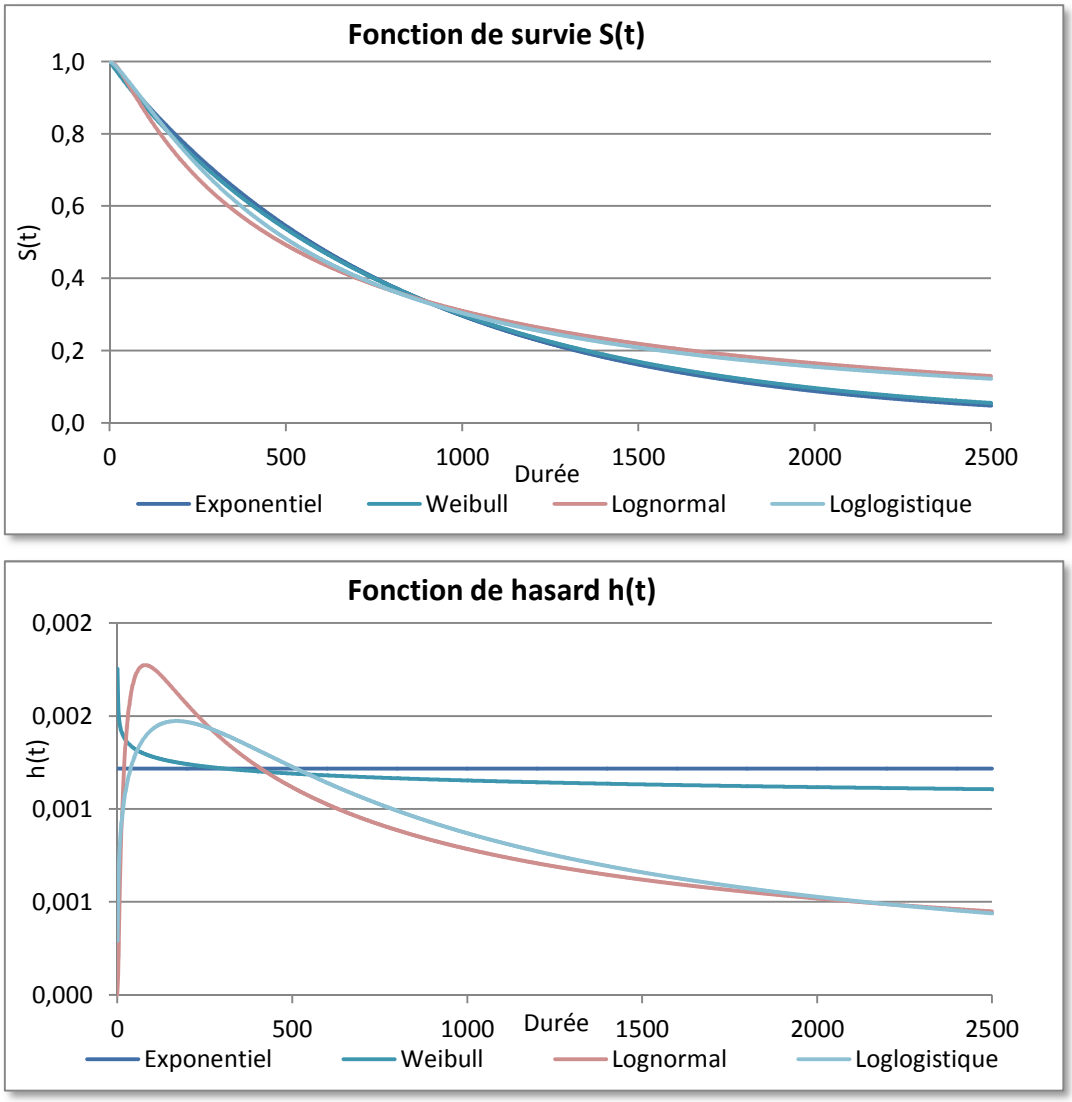
La fonction de survie et la fonction de risque instantané sont caractérisées à partir de la densité de la loi :

- la fonction de survie : $S_\theta(t) = 1 - F_\theta(t) = 1 - \int_0^t f_\theta(u)du$,
- la fonction de hasard : $\lambda_\theta(t) = \frac{f_\theta(t)}{S_\theta(t)} = \frac{f_\theta(t)}{1-F_\theta(t)} = \frac{f_\theta(t)}{1-\int_0^t f_\theta(u)du}$.

Loi	Fonction de densité	Fonction de survie	Fonction de hasard
Exponentielle	$f_\theta(t) = \theta e^{-\theta t}$	$S_\theta(t) = e^{-\theta t}$	$\lambda_\theta(t) = \theta$
Weibull	$f_{\nu,\theta}(t) = \nu\theta(\theta t)^{\nu-1} e^{-(\theta t)^\nu}$	$S_{\nu,\theta}(t) = e^{-(\theta t)^\nu}$	$\lambda_{\nu,\theta}(t) = \nu\theta(\theta t)^{\nu-1}$
Log-Normale	$f_{\nu,\theta}(t) = \frac{1}{\theta t\sqrt{2\pi}} e^{\frac{-1}{2\theta^2}(\ln(t)-\nu}$	$S_{\nu,\theta}(t) = 1 - \Phi\left(\frac{\ln(t)-\nu}{\theta}\right)$	$\lambda_{\nu,\theta}(t) = \frac{\frac{1}{\theta t}\phi(\frac{\ln(t)-\nu}{\theta})}{1-\Phi\left(\frac{\ln(t)-\nu}{\theta}\right)}$
Log-Logistique	$f_{\nu,\theta}(t) = \frac{\theta\nu t^{\nu-1}}{(1+\theta t^\nu)^2}$	$S_{\nu,\theta}(t) = \frac{1}{1+\theta t^\nu}$	$\lambda_{\nu,\theta}(t) = \frac{\theta\nu t^{\nu-1}}{\left(1+\theta t^\nu\right)^3}$

TABLEAU 12 – Les fonctions de hasard paramètriques[15]

[15] ϕ et Φ respectivement la fonction de densité et répartition de la loi Normale (0,1) (cf. annexe pour une présentation plus détaillée des lois).



GRAPHIQUE 8 – Les fonctions de hasard et de survies paramètriques

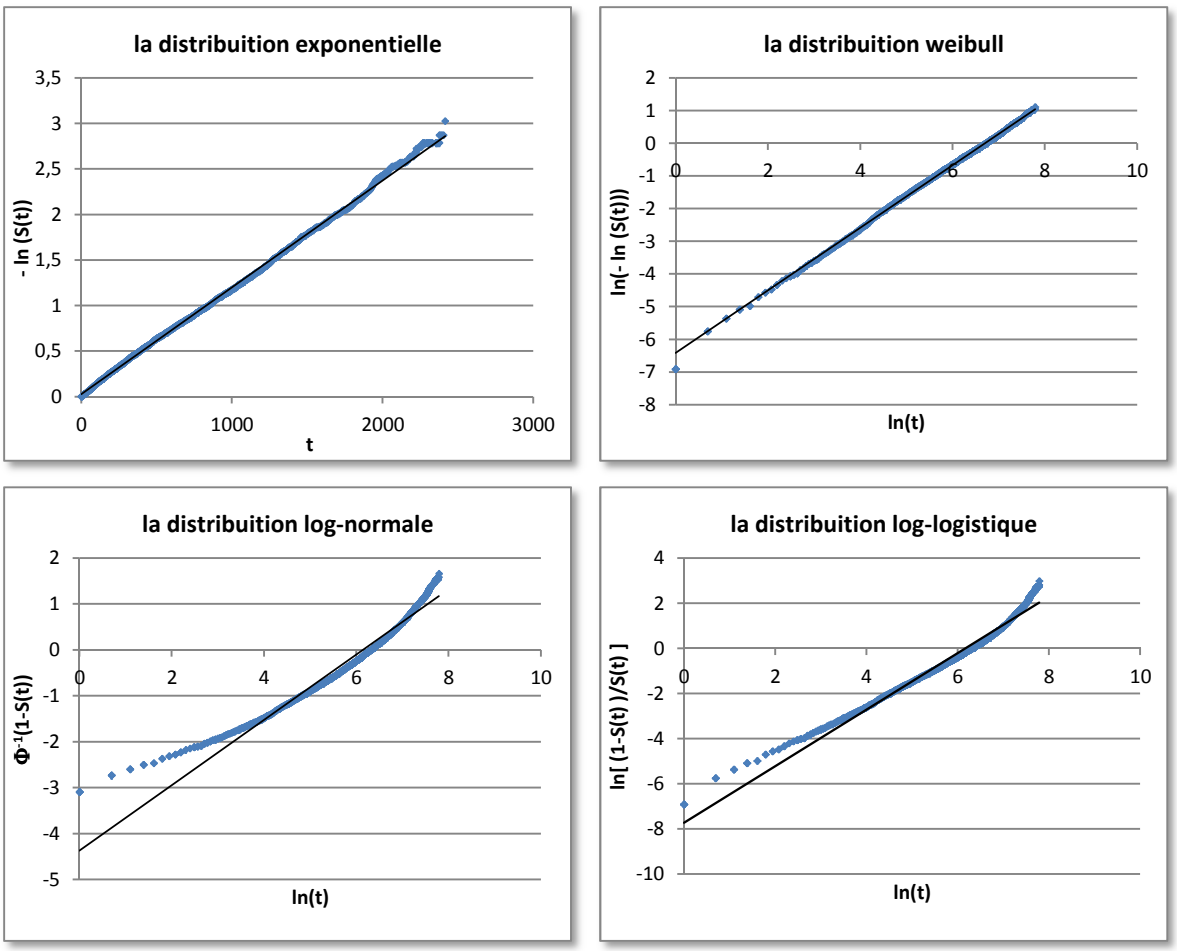
NB : la variable *duree* est la durée en jours au delà de 4 ans.

Parmi les fonctions de hasard paramétriques, la fonction de hasard de type weibull est la plus proche de la fonction de hasard obtenue à l'aide de la méthode actuarielle.

3.5.1.2 Choix d'une distribution avant estimation

L'objectif est d'appliquer une transformation à S(t) afin d'obtenir une équation linéaire en fonction de t ou ln(t). Le graphique avec t ou ln(t) en abscisse et la transformation de S(t) en ordonnée doit être représenté par une droite.

- la distribution Exponentielle : $-\ln(S(t)) = \theta t$,
- la distribution Weibull : $\ln\left[-\ln(S(t))\right] = \ln\theta + v\ln t$,
- la distribution Log-Normale : $\Phi^{-1}\left[1-S(t)\right] = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\ln t$,
- la distribution Log-logistique : $\ln\left[\frac{1-S(t)}{S(t)}\right] = \ln\theta + v\ln t$.



GRAPHIQUES 9 – Choix d'une distribution avant estimation

L'étude des graphiques met en exergue un lien linéaire pour la distribution exponentielle et la distribution de Weibull. Il est à préciser que cette approche est une aide à la décision et ne permet pas de tirer de conclusions assurées.

3.6 Identification des facteurs influents

Dans cette partie, nous chercherons à identifier les critères influents, de manière uni-variée, sur la durée de survie. Cette présélection sera ensuite utilisée lors du modèle *AFT*.

La mise en œuvre de cette présélection s'appuie sur l'étude graphique et le test de rang et de Gehan.

La comparaison de durées de vie issue de deux échantillons se fait à l'aide de tests. Les tests les plus utilisés dans le cadre des données de censures sont le test de rang et de Gehan[16].

L'objectif de cette approche est de tester l'hypothèse H_0 : l'égalité des fonctions de survie des deux échantillons contre H_1 : différence des fonctions de survie des deux échantillons.

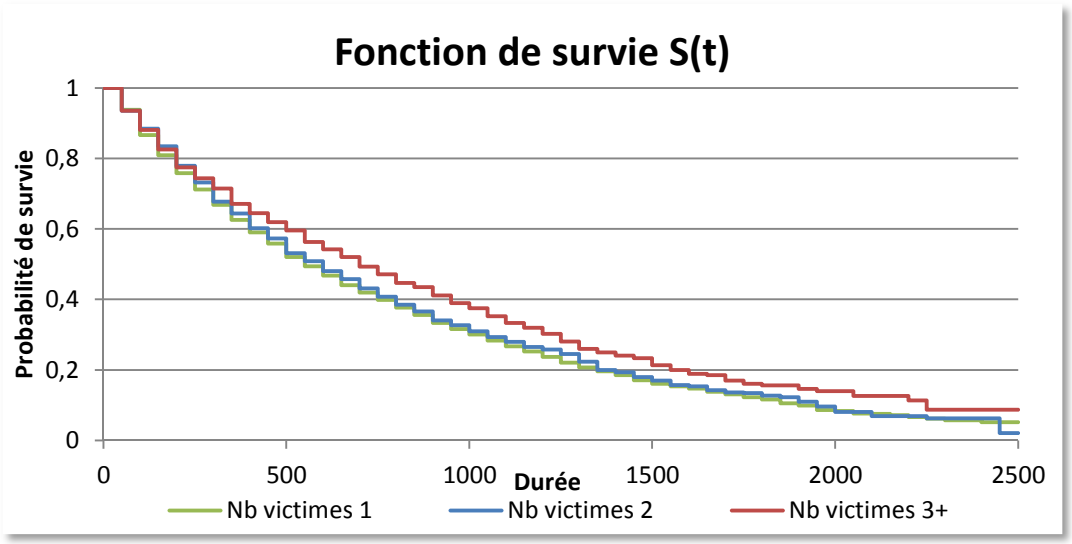
La statistique de test requiert la notion de pondération w_k. Le coefficient de pondération prend la valeur $w_k=1$ dans le cas du test de log-rank et $w_k=r_k$ dans le cas du test de Gehan. Il

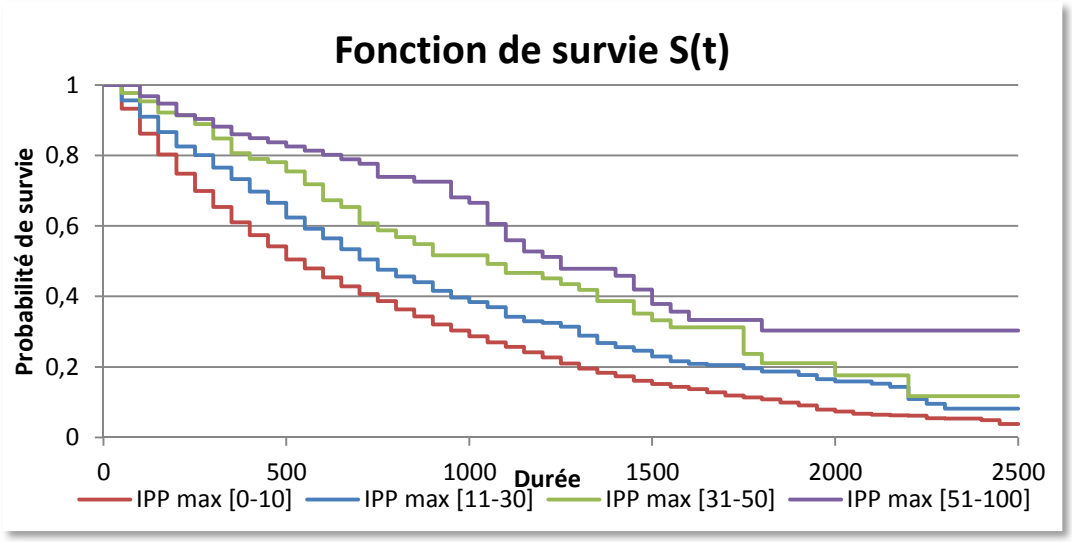
[16] Cf. Annexe pour une présentation plus détaillée.

est à préciser que la pondération appliquée ($w_k=r_k$) conduit à pondérer plus fortement les décès les plus précoces.

La statistique de test suit asymptotiquement une loi *Khi-deux* (χ^2) à 1 degré de liberté.

Les tests de rang et le test de Gehan se généralisent à la comparaison des fonctions de survie de plusieurs groupes. L'hypothèse testée est : H_0 : les probabilités de survie entre les j groupes sont identiques contre H_1 : les probabilités de survie entre les j groupes sont différentes. La statistique de test suit asymptotiquement une loi de χ^2 à j-1 degrés de liberté.





GRAPHIQUE 10 – Courbes de survies en fonction de facteurs explicatifs[17]

		Pr > Chi-square	
	ddl	*Log-Rank*	*Gehan*
Nombre de victimes	*2*	*0,0022*	*0,0139*
Taux IPP max	*3*	*<0,0001*	*<0,0001*

TABLEAU 13 – Test du *Log-rank* et *Gehan*

[17] Nota : la variable *duree* est la durée en jours au-delà de 4 ans.

Au seuil de 5%, nous rejetons l'hypothèse H_0, les probabilités de survie entre les modalités d'une même variable sont identiques, pour les 2 variables.

Nous constatons que plus le nombre de victimes est grand ou plus le taux d'IPP de la victime est élevé, plus la durée de survie sera longue.

Le nombre de victimes et le taux d'IPP sont considérés comme significativement influents sur la durée de vie du sinistre. Aussi, du fait de la proportionnalité dans le temps des modalités d'une même variable, les variables seront directement intégrées à la modélisation et n'ont pas besoin d'être retravaillées.

3.7 Modèle *Accelerated Failure Time*

3.7.1 Théorie: le modèle de vie accélérée

La durée de survie peut dépendre de variables explicatives. L'approche paramétrique incorpore les facteurs via l'utilisation d'un modèle de hasard proportionnel ou d'un modèle de vie accélérée. Dans notre étude nous utiliserons le modèle de vie accélérée.

3.7.1.1 Fondements du modèle

Les modèles de vie accélérées incorporent les facteurs explicatifs sous forme « d'accélérateurs » de temps :

$$T = T_0 \exp(-\beta' Z) \quad \text{ou} \quad \ln(T) = \ln(T_0) - \beta' Z$$

Avec T_0 et Z étant respectivement la loi de références et les facteurs exogènes.

Tout se passe comme si l'effet des facteurs explicatifs était d'allonger ou de rétrécir l'unité du temps. L'intérêt principal de ces modèles est de permettre d'interpréter l'effet des variables explicatives comme un changement d'échelle de l'axe du temps.

Ce modèle suppose que la fonction de survie S(t) conditionnée par les variables exogènes que nous désignons globalement par Z, se ramène à une fonction de survie de base $S_0(t)$, selon une relation :

$$S_{\theta/Z}(t) = S_0(t \exp(\beta' Z))$$

Car $S_{\theta/Z}(t) = P(T > t) = P(T_0 \exp(-\beta' Z) > t) = P(T_0 > t \exp(\beta' Z)) = S_0(t \exp(\beta' Z))$

La fonction de risque instantanée est déduite de la fonction de survie :

$$\lambda_{\theta/Z}(t) = \left[-\ln(S_{\theta/Z}(t))\right]' = -\frac{S_{\theta/Z}(t)'}{S_{\theta/Z}(t)} = \frac{\exp(\beta' Z) * f_0(\exp(\beta' Z))}{S_0(t \exp(\beta' Z))}$$

avec $S_{\theta/Z}(t)' = \exp(\beta' Z) * S_0'(t \exp(\beta' Z)) = -\exp(\beta' Z) * f_0(t \exp(\beta' Z))$ et $f_0(t) = -S_0'(t)$

Et comme $f_0(t) = S_0(t) * \lambda_0(t)$

Nous obtenons :

$$\lambda_{\theta/Z}(t) = \frac{\exp(\beta' Z) * \lambda_0(t\exp(\beta' Z)) * S_0(t\exp(\beta' Z))}{S_0(t\exp(\beta' Z))} = \exp(\beta' Z) * \lambda_0(t\exp(\beta' Z))$$

3.7.1.2 Estimation des paramètres

L'estimation des paramètres θ de la loi est obtenue à l'aide du maximum de vraisemblance qui s'écrit sous la forme suivante :

$$L(\theta) = \kappa \prod_{i=1}^{n} \lambda_\theta(t_i)^{di} S_\theta(t_i) \quad \text{ou} \quad L(\theta) = \kappa \prod_{i=1}^{n} f_\theta(t_i)^{di} S_\theta(t_i)^{1-di}$$

L'approche paramétrique permet la prise en compte de variables explicatives à l'aide du modèle de vie accélérée. L'estimation des paramètres de la loi θ et des paramètres explicatifs β se fait simultanément par maximisation de la vraisemblance :

$$L(\theta,\beta) = \kappa \prod_{i=1}^{n} \lambda_{\theta,\beta/Z}(t_i)^{di} S_{\theta,\beta/Z}(t_i) \quad \text{ou} \quad L(\theta,\beta) = \kappa \prod_{i=1}^{n} f_{\theta,\beta/Z}(t_i)^{di} S_{\theta,\beta/Z}(t_i)^{1-di}$$

La résolution des équations de vraisemblance est effectuée à partir de méthodes itératives et numériques telles que l'algorithme de Newton Raphson. L'estimation d'un tel modèle demande que soit spécifiée la loi de base, elle opère par la méthode classique du maximum de vraisemblance.

$$L(\theta) = \kappa \prod_{i=1}^{n} \lambda_{\theta/Z}(t_i)^{di} S_{\theta/Z}(t_i)$$

Soit le log vraisemblance : $\ln L(\theta) = \kappa \sum_{i=1}^{n} \left(d_i \ln(\lambda_{\theta/Z}(t_i)) + S_{\theta/Z}(t_i) \right)$

avec $S_{\theta/Z}(t) = S_0(t\exp(\beta' Z))$ et $\lambda_{\theta/Z}(t) = \exp(\beta' Z) * \lambda_0(t\exp(\beta' Z))$

D'où $\ln L(\theta) = \kappa \sum_{i=1}^{n} \left(d_i (\ln(\exp(\beta' Z) * \lambda_0(t\exp(\beta' Z))) + S_0(t\exp(\beta' Z)) \right)$

$$\ln L(\theta) = \kappa \sum_{i=1}^{n} \left(d_i (\exp(\beta' Z) + \ln(\lambda_0(t\exp(\beta' Z)))) + S_0(t\exp(\beta' Z)) \right)$$

L'estimation des paramètres est effectuée en maximisant la log vraisemblance par rapport au vecteur de paramètre β. Soit la fonction score : $U(\beta) = \frac{\partial Log\ L(\beta)}{\partial \beta} = 0$.

Les équations obtenues sont résolues par des méthodes numériques telles que l'algorithme de Newton-Raphson. L'estimation de la variance de β est obtenue à partir de l'inverse de la matrice d'information de Fisher I :

$Var(\hat{\beta}) = \left\{ I(\hat{\beta}) \right\}^{-1}$ où le terme (i,j) de la matrice I est $I(\beta) = \frac{\partial^2 Log L(\beta)}{\partial \beta_i \partial \beta_j}$

Les estimateurs des paramètres sont asymptotiquement gaussiens.

La significativité du paramètre se fait grâce au test de Wald et au test de vraisemblance qui ont asymptotiquement une distribution de $\chi^2(1)$. L'hypothèse de nullité des paramètres (H_0 : $\beta=\beta_0$) peut être effectuée à l'aide des trois tests classiques :

- le test du rapport de vraisemblance : $\chi_V^2 = 2\left(Log\ L(\hat{\beta}) - Log\ L(\beta_0)\right)$
- le test de Wald : $\chi_W^2 = \left(\hat{\beta} - \beta_0\right)' I(\hat{\beta})\left(\hat{\beta} - \beta_0\right)'$
- le test du Score : $\chi_S^2 = U\left(\beta_0\right)' I(\beta_0)^{-1} U\left(\beta_0\right)'$

Sous H_0 ces trois statistiques de test suivent asymptotiquement, des distributions de Khi-deux à p degrés de liberté où p est la dimension du vecteur β. Ces tests peuvent être utilisés pour tester certains termes du vecteur β. Ainsi, si nous souhaitons tester H_0 : $\beta_p=0$ nous poserons :

$$\beta = (\beta_1,...,\beta_{p-1},\beta_p) \text{ et } \beta_0 = (\beta_1,...,\beta_{p-1},0)$$

3.7.1.3 Adéquation du modèle

L'une des difficultés rencontrées lors de l'ajustement d'un modèle paramétrique est celui du choix de la distribution sous laquelle vont se faire les estimations. Nous utiliserons le test de rapport de vraisemblance (*TRV*) et les résidus de Cox-Snell.

Dès lors que le choix entre plusieurs distributions est possible, il est utile de pouvoir discriminer entre les diverses alternatives. Une solution est d'utiliser un test de type TRV. Ce test s'applique dans le cas où le modèle contraint est emboité dans le non contraint.

Soit l_{nc} et l_c respectivement les valeurs de la log-vraisemblance obtenues après estimation d'un modèle non contraint et d'un modèle contraint, la quantité TRV = $2(l_{nc} - l_c)$ est alors distribuée selon une loi du Khi-deux avec un nombre de degrés de liberté c égal au nombre de contraintes pour passer de l'un à l'autre. Si $2(l_{nc} - l_c) < \chi^2(c)$ cela signifie que le modèle contraint est aussi vraisemblable que le modèle non contraint.

Ce test est simple pour comparer un modèle exponentiel et weibull. Nous voyons que le modèle exponentiel est un cas particulier du modèle weibull. Il est obtenu lorsque le paramètre de forme v=1. Cette contrainte nous fait donc passer d'une distribution de Weibull (modèle non contraint) à une distribution exponentielle (modèle contraint).

Remarque : Il est à souligner que le test ne valide pas pour autant le modèle retenu.

Le résidu de Cox-Snell est définit par $H(T_i / Z,\theta) = -\log\left(S(T_i / Z,\theta)\right)$. Sous l'hypothèse que le modèle paramétrique est adéquat, les résidus de Cox-Snell doivent suivre une loi exponentielle de paramètre 1[18]. Pour le tester, nous représentons un graphique avec en abscisse le résidu de Cox-Snell et en ordonnée $-\log(S_{CS}(CS))$[19]. Si les points tiennent sur une droite de pente 1 passant par l'origine nous pouvons conclure sur l'adéquation du modèle.

[18] Si U est une variable uniforme sur]0,1] alors Y=–log U est une loi exponentielle de paramètre 1. $F_Y(y)=1-F_U(e^{-y})=1-e^{-y}$. En différenciant par y, nous obtenons $f_Y(y)= ye^{-y}$.

[19] Si Exponentielle(1) alors $S(t)=e^{-t}$ et –log S(t) = t.

3.7.2 Application : le modèle de vie accélérée

Distribution	l
Exponentielle	-8 730
Weibull	-8 724
Log-Normale	-8 952
Log-Logistique	-8 842

TABLEAU 14 – Maximum de la log-vraisemblance sous différentes distributions

La vraisemblance est maximisée avec les distributions exponentielles et weibull. Le test de TRV sur la distribution exponentielle et weibull s'écrit :

$$TRV = - 2 (-8\,730 + 8\,724) = 11,6.$$

Avec une valeur critique à 5% de 3,84 pour le χ^2 à un degré de liberté, nous rejetons l'hypothèse que la distribution exponentielle est équivalente à la distribution Weibull.

La modélisation permet de décorreler l'effet des critères et de mettre en exergue la significativité intrinsèque des critères.

Variables	ddl	Khi-square	Pr>Khi-sq
Nb de victimes	1	7,7155	0,0055
Taux IPP	2	54,1835	<.0001

TABLEAU 15 – Analyse des effets de type 3

L'hypothèse de nullité est rejetée pour les 2 variables et nous considérons ainsi qu'au moins une des modalités a une influence significative sur l'analyse de la durée.

Paramètre	Estimation	Ecart-type	Pr>Khi-sq
Intercept	6,6278	0,0193	<.0001
Nb victimes 3+	0,1285	0,0552	0,019
Nb victimes 2	0,0168	0,0410	0,6817
Nb victimes 1	0,0000		
Taux IPP [51-100]	0,9123	0,1515	<.0001
Taux IPP [31-50]	0,6113	0,1224	<.0001
Taux IPP [11-30]	0,3114	0,0521	<.0001
Taux IPP [0-10]	0,0000		
Scale	1,0408	0,0128	
Weibull Shape	0,9608	0,0118	

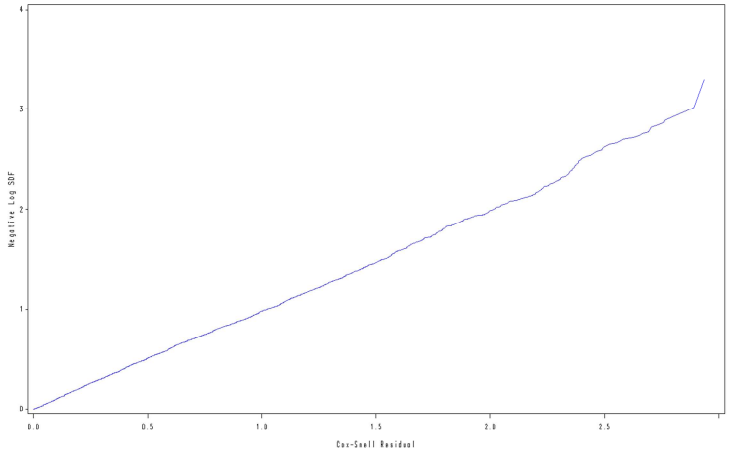
TABLEAU 16 – Estimation des paramètres

Le tableau d'estimation des paramètres met en exergue la significativité des variables et des modalités associées. Plus le coefficient est élevé et plus la durée de vie du sinistre sera longue.

Le nombre de victimes 1 et 2 ont un comportement similaire. Le nombre de victimes 3+ se distingue significativement du nombre de victime 1 et 2. Ainsi, le nombre de victimes 3+ a une durée plus longue que le nombre de victime 1 et 2.

Les modalités de la variable Taux IPP se distinguent tous significativement de la modalité de référence. Plus la modalité est élevée et plus l'estimation est grande signifiant que la durée sera plus longue.

Il est à noter que la variable Taux IPP tient un rôle clé dans le modèle et est davantage discriminant que la variable nombre de victimes. Nous observons que les estimations sont supérieures aux estimations de la variable nombre de victimes signifiant que leur impact sur la durée de vie est plus fort.



GRAPHIQUE 11 – Adéquation du modèle – les résidus de *Cox-Snell*

Le graphique des résidus de Cox-Snell vérifie l'adéquation du modèle. En effet, les points tiennent sur une droite à 45° passant par l'origine .

A partir des 3 modalités de la variable nombre de victimes et des 4 modalités de la variable taux IPP, nous obtenons 12 profils de durée classés du profil ayant la durée plus longue à la plus courte.

Profil	Nb victimes	Taux IPP	Répartition	Paramètre
1	3+	[51-100]	0,3%	1,041
2	2	[51-100]	0,3%	0,929
3	1	[51-100]	1,0%	0,912
4	3+	[31-50]	0,4%	0,740
5	2	[31-50]	0,5%	0,628
6	1	[31-50]	1,3%	0,611
7	3+	[11-30]	1,9%	0,440
8	2	[11-30]	2,6%	0,328
9	1	[11-30]	7,5%	0,311
10	3+	[0-10]	7,6%	0,128
11	2	[0-10]	15,1%	0,017
12	1	[0-10]	61,7%	0,000

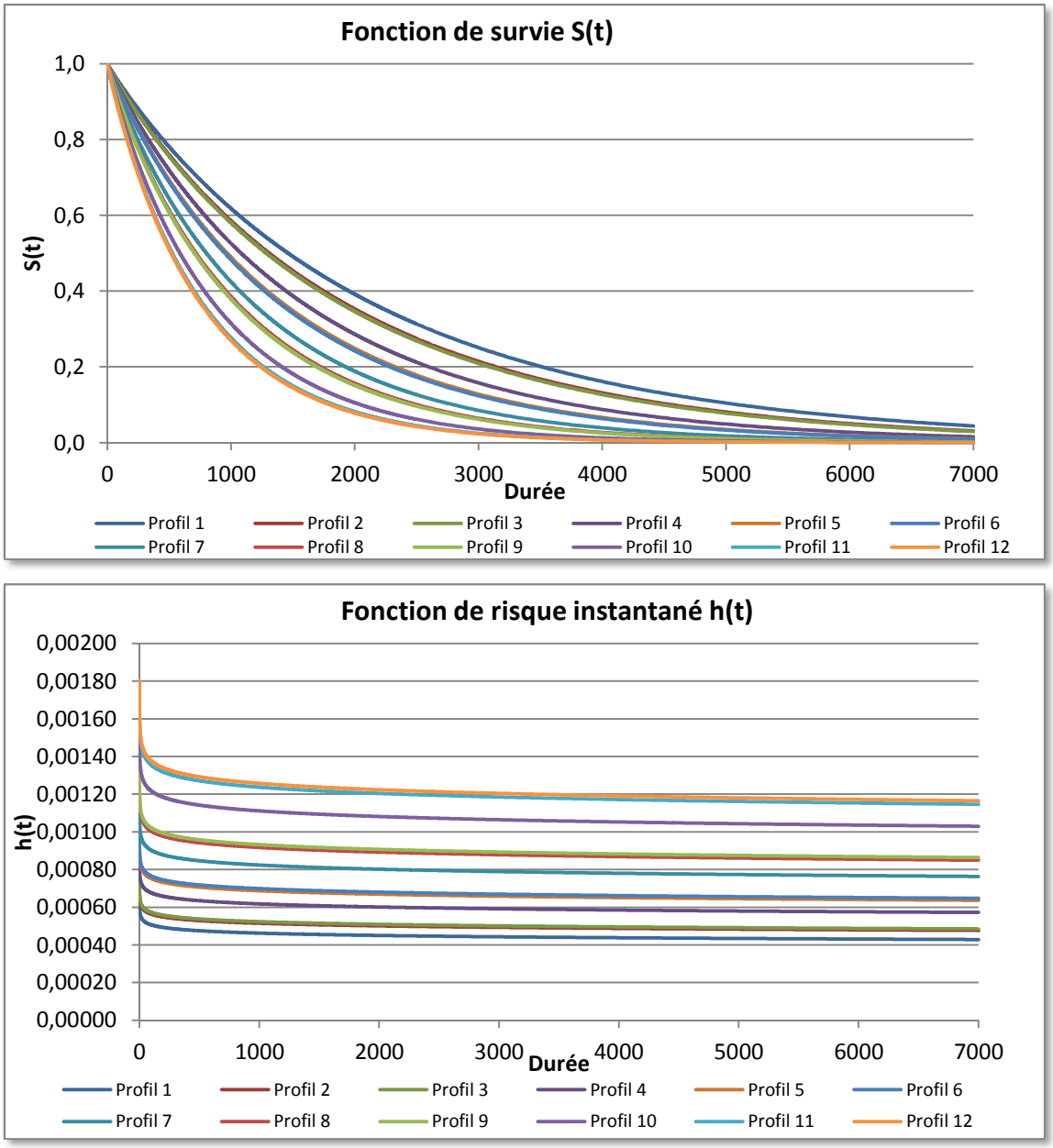
TABLEAU 17 – Répartition et paramètre des 12 profils

Le profil 12 est le profil de durée le plus représenté, soit 61,7% de la population d'étude, et constitue le profil de référence avec ainsi un paramètre[20] fixé à 0.

Le profil 12 est le profil de référence où le sinistre est constitué d'une victime avec un taux IPP inférieur à 10%. Ce profil décroit rapidement dans le temps, la probabilité de survie est de 50 % pour une durée de 516 jours.

[20] Paramètre correspond à la combinaison linéaire des estimations des modalités.

A l'opposé, nous avons le profil 1 composé d'au moins 3 victimes avec un taux IPP maximal supérieur ou égal à 51 %. La courbe de survie associée décroit plus lentement et atteint 50 % pour une durée de 1463 jours.



GRAPHIQUE 12 – Courbes différenciées de survie S(t) et de risque instantané h(t)[21]

Le profil 12 est le profil de référence où le sinistre est constitué d'une victime avec un taux IPP inférieur à 10%. Ce profil décroit rapidement dans le temps, la probabilité de survie est de 50 % pour une durée de 516 jours.

A l'opposé, nous avons le profil 1 composé d'au moins 3 victimes avec un taux IPP maximal supérieur ou égal à 51 %. La courbe de survie associée décroit plus lentement et atteint 50 % pour une durée de 1 463 jours.

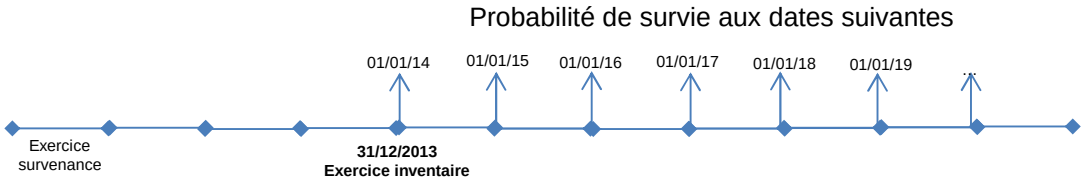
[21] Nota : la variable *duree* est la durée en jours au-delà de 4 ans.

3.8 Modèle de durée

L'objectif est d'appliquer le modèle sur les sinistres ouverts afin d'obtenir la probabilité de survie pour les années futures, c'est-à-dire les années suivantes la date d'inventaire. Les probabilités sont calculées conditionnellement à l'information disponible à la date d'inventaire.

3.8.1 L'utilisation du modèle

Nous souhaitons connaître la probabilité de survie des sinistres au 01/01 de chaque année future même si par la suite, nous faisons l'hypothèse que les règlements sont effectués en milieu d'année.



SCHEMA 6 – Probabilité de survie sur les exercices futurs

Les courbes de survie sont calibrées sur une population ayant une durée de vie supérieure à 4 ans. Ainsi, nous cherchons à déduire la probabilité conditionnellement à l'information que le sinistre est ouvert à la date d'inventaire. La date d'inventaire est fixée au 31/12/2013 et à cette date, les sinistres d'un même exercice de survenance ont des durées variables. Par exemple pour l'exercice de survenance 2009 la durée de vie du sinistre peut varier entre 4 et 5 ans.

Nous sélections les sinistres non clos d'exercice de survenance 2003 à 2009. Il est à préciser que les sinistres clos sont considérés comme étant vus à l'ultime et ne requièrent pas de projection de flux futurs.

Nous identifions la courbe de survie à utiliser à partir des caractéristiques du sinistre (Nombre de victimes, Taux IPP). Puis, nous calculons la probabilité conditionnelle qui est obtenue de la manière suivante :

$$S_u(t) = P(X > v / X > u) = \frac{P(X > v)}{P(X > u)} = \frac{S(v)}{S(u)}$$

Avec :

- u = date d'inventaire – (date de survenance + 4 ans),
- v = date d'un futur flux – (date de survenance + 4 ans),

Prenons l'exemple du profil 1 avec une date de survenance au 01/04/2009. Nous calculons la probabilité conditionnelle, $P(T>t / F_{31/12/2013})$ à différentes dates dans le futur.

Date	31/12/2013	01/01/2014	01/01/2015	01/01/2016	01/01/2017	01/01/2018	…	01/01/2048	01/01/2049
S(t)	0,872	0,872	0,735	0,622	0,528	0,449	…	0,005	0,004
$P(T>t / F_{31/12/13})$		1,000	0,843	0,713	0,605	0,514	…	0,005	0,004

TABLEAU 18 –Probabilités conditionnelles

Clé de lecture : la probabilité que le sinistre soit non clos au 01/01/2015 sachant qu'il est en cours à la date d'inventaire est 0.843 et est calculée de la manière suivante :

$$P(T > 640; F_{31\cdot12\cdot2013}) = \frac{S(640)}{S(274)} = \frac{0{,}735}{0{,}872} = 0{,}843$$

A la suite de ces 4 étapes, pour chaque sinistre non clos nous obtenons un vecteur de probabilités conditionnellement à la date d'inventaire et allant jusqu'à la date maximale possible (appelé J).

Année développement	6	7	8	…	J
$P(T_i > j / F_{31/12/2013})$	1,00	0,84	0,71		0

3.8.2 Application à la diagonale

En sommant les probabilités de survie de chaque sinistre non clos au 31/12 de chaque année de développement nous obtenons le tableau suivant :

	Année de développement															
Exercice survenance	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2 003	545	389	284	207	160	93	**74**	50	34	24	17	12	8	6	4	3
2 004	554	385	265	191	119	**95**	64	44	30	21	15	10	7	5	4	3
2 005	678	447	320	186	**117**	79	53	36	25	18	12	9	6	5	3	3
2 006	648	457	271	**203**	135	91	62	42	29	21	15	10	7	5	4	3
2 007	614	402	**290**	193	129	88	60	42	29	20	15	11	8	6	4	3
2 008	520	**354**	233	155	104	71	49	34	24	17	12	9	6	5	4	3
2 009	**548**	352	230	152	101	68	46	32	22	15	11	8	5	4	3	2

TABLEAU 19 – Projection des sinistres non clos

Pour l'exercice de survenance 2003, parmi les 74 sinistres non clos à l'inventaire 31/12/2013, 3 sinistres seraient non clos au 31/12 de la 20ème année de développement.

De même, nous voyons que parmi les 548 sinistres de l'exercice de survenance 2009, 2 sinistres seraient clos lors de la 20ème année de développement.

4 PARTIE 4 - MODELISATION DES PAIEMENTS

Dans cette partie nous présentons une approche pour modéliser les processus de paiements en utilisant un modèle composé des variables : la probabilité et le montant des paiements.

Un ensemble de sinistres est essentiellement de même structure qu'un triangle de règlements mais permet de disposer d'information détaillée sur le sinistre.

L'idée, pour la modélisation des paiements, est d'utiliser les informations génériques telles que l'année d'origine et l'année de déroulement et les informations spécifiques au sinistre telles que le profil de durée et le montant de la provision dossier évaluée par le gestionnaire.

Il est à souligner que l'utilisation du profil de durée dans la modélisation des paiements permet de corréler l'estimation de la durée au processus de paiement du sinistre.

4.1 Le cadre général

L'objectif est d'estimer la chronique des paiements futurs d'un sinistre non clos compte tenu de l'information dont nous disposons à la date d'inventaire t. Pour chaque sinistre, nous modéliserons le paiement au cours d'une année en fonction des caractéristiques constatées en début de cette année.

Nous constatons que les paiements au cours d'une année ne concernent qu'une partie des sinistres non clos. De nombreux sinistres non clos n'ont pas de paiement au cours d'une année.

Du fait de la discontinuité de la chronique des paiements, la modélisation des paiements ne peut être faite directement et nécessite d'être décomposée en 2 éléments distincts : la probabilité d'observer un paiement et le montant de ce paiement.

L'approche est fondée sur un modèle composé nécessitant l'estimation de la probabilité d'un paiement P dans l'année, le montant de ce paiement M et δ la provision dossier en début de période, pour l'ensemble des années de développement futures nommées par l'indice j.

Nous considérons une indépendance entre la probabilité d'observer un paiement et le montant de ce paiement et déduisons l'équation pour le paiement d'un sinistre i en année de développement future j :

$$\begin{cases} [X_{ij}/F_t] = [P_{ij}/F_t] * [M_{ij}/F_t] \; si \; \delta_{ij} > 0 \\ [X_{ij}/F_t] = \; 0 \; sinon \end{cases}$$

La variable X_{ij} obtenue est ensuite multipliée par l'inflation sur la période.

Nous cherchons à estimer la probabilité P_j et le montant de ce paiement M_j au cours de l'année j en fonction de l'information dont nous disposons en début de chaque année j. Pour chaque sinistre i, nous disposons de 4 variables explicatives :

- α : l'exercice de survenance,
- β : l'année de développement,
- γ : le profil de durée,
- δ : le montant de la provision dossier.

Nous obtenons les équations suivantes :

$$\begin{cases} P_j = g(\alpha, \beta, \gamma, \delta) \\ M_j = h(\alpha, \beta, \gamma, \delta) \end{cases}$$

En début de chaque année de développement j, nous devons disposer des facteurs explicatifs. Nous obtenons directement l'information concernant l'exercice de survenance, l'année de développement et le profil de durée[22].

Le montant de provision dossier est quant à lui dynamique. L'estimation en début de chaque année nécessite de poser les hypothèses suivantes :

- la provision en j+1 dépend du montant de la provision en j et du paiement en j,
- si la provision en j est nulle alors le paiement en j sera nul.

L'estimation de la provision dossier en j+1 est obtenue à partir de la provision constatée en j déduite du paiement en j. Il est à préciser que cette provision ne peut être négative et que la provision dossier à l'inventaire est une borne max du coût du sinistre. Nous définissons l'équation suivante :

$$\delta_{i,j+1} = \max(\delta_{ij} - [P_{ij}/F_t] * [M_{ij}/F_t] * Inflation; 0)$$

Dès que la provision dossier est nulle, la probabilité et le montant des paiements est fixé à 0. Par hypothèse, la provision dossier δ, évaluée à la date d'inventaire par le gestionnaire, joue le rôle d'une borne supérieure.[23]

Nous considérons que la provision évaluée par les gestionnaires intègre une vision centrale d'inflation à 1,50% sur la période de projection.

Nous ne tenons pas compte des réévaluations des sinistres par les gestionnaires compte tenu de leur nombre marginal[24].

La méthodologie proposée permet de calculer les paiements restants pour les sinistres non clos à la date d'inventaire. Les réouvertures de sinistres ne sont pas considérées. Il est à préciser que les réouvertures de sinistres clos sont en réalité marginales. A partir de la 6ème année de développement, elles concernent 1,5 % des paiements sur le déroulé entier d'un exercice de survenance.

Une fois l'estimation de P et M effectué nous obtenons une chronique des paiements pour l'ensemble des années futures.

Il est à préciser que pour des raisons de simplicité, nous regroupons les paiements d'une année en un unique paiement par année de développement qui aura lieu en milieu d'année.

[22] A partir de la 5ème année de développement le profil de durée reste identique jusqu'à la clôture du dossier.
[23] Nous observons ce phénomène sur les triangles agrégés pour la garantie Automobile RCC.
[24] A partir de la 5ème année de développement, moins de 1% des provisions dossiers des sinistres est réévalué significativement à la hausse.

La mise en place de ces modèles se fait en utilisant les modèles linéaires généralisés et plus spécifiquement un modèle logistique pour la probabilité de paiement et un modèle avec lien log pour le montant du paiement.

4.2 La population d'étude

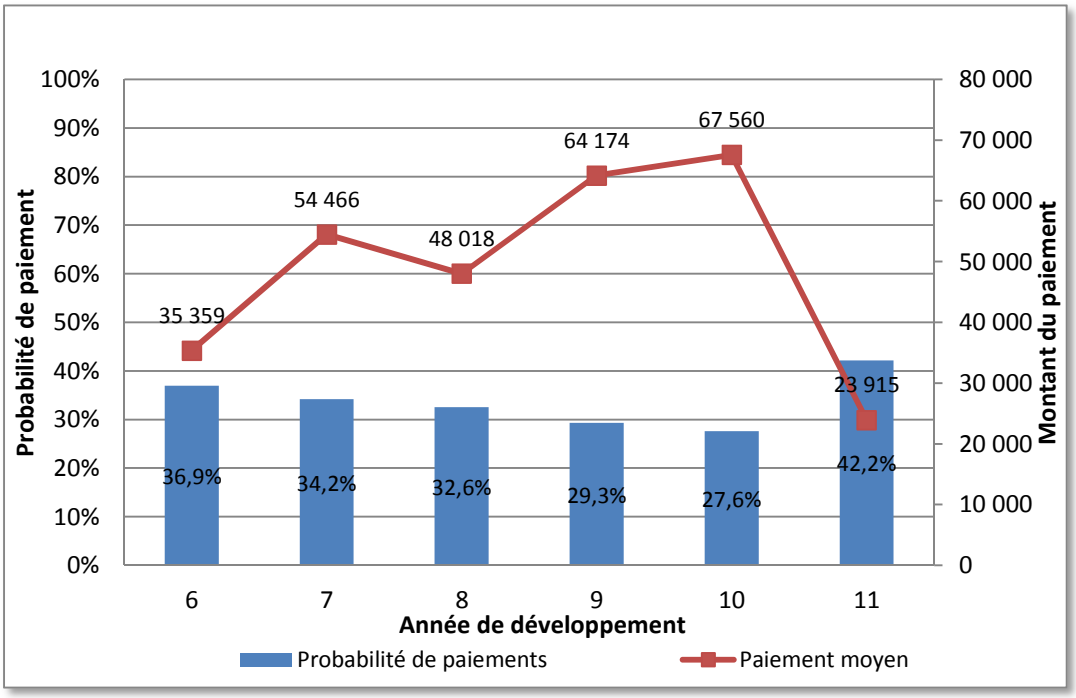
Le modèle individuel s'applique aux sinistres « anciens », où les sinistres non clos sont peu nombreux et les situations sont disparates entre exercices de survenance. Ainsi, la population d'étude est composée des sinistres non clos ayant des paiements entre la 6ème année et la 11ème année de développement.

Nous disposons d'une base d'étude composée de 6 817 sinistres répartis sur les années de développement de la manière suivante :

	Année de développement					
	6	**7**	**8**	**9**	**10**	**11**
Sinistres non clos	3 124	1 834	1 001	525	250	83

TABLEAU 20 – Nombre de sisnistres non clos en début de chaque année

Les sinistres non clos ne donnent pas tous lieu à un paiement dans l'année. Il y a 2 365 paiements répartis sur l'ensemble des exercices de survenance et l'année de développement, soit une probabilité globale d'observer un paiement au cours d'une année de 34,7 %.



GRAPHIQUE 13 – Probabilité de paiement et paiement moyen par année de développement

L'année de développement 11 se distingue des autres années de développements. Nous constatons une hausse du taux des paiements conduisant à une diminution du cout moyen. Cette dynamique s'explique par une procédure de classement automatique de certains sinistres au bout de 10 ans. Ainsi, pour ces sinistres, le gestionnaire est alerté et cherchera à revenir vers la victime afin de clôturer le sinistre.

Excepté pour la 11ème année de développement, nous constatons que la probabilité de paiements a tendance à diminuer dans le temps et qu'inversement le paiement moyen a tendance à augmenter.

Les paiements pour une année de développement sont volatils et nous constatons que l'écart type représente en général 4 à 5 fois le paiement moyen.

	Année de développement					
	6	**7**	**8**	**9**	**10**	**11**
Coefficient de variation	382%	504%	315%	395%	516%	350%

TABLEAU 21 – Coefficient de variation par année de développement

4.3 Les modèles linéaires généralisés

Les modèles linéaires généralisés ont été introduits par J. Nelder et R. Wedderburn en 1972. Ce modèle est une extension du modèle linéaire Normal et est formé de trois composantes : la composante aléatoire, la composante systématique et la fonction de lien.

L'objet de ce chapitre est d'introduire le cadre théorique global permettant de regrouper tous les modèles (linéaire gaussien, logit, log-linéaire) qui visent à exprimer l'espérance d'une variable réponse Y en fonction d'une combinaison linéaire des variables explicatives.

4.3.1 Les fondements du modèle

- La composante aléatoire

La composante aléatoire identifie la distribution de probabilités de la variable à expliquer. Nous supposons que l'échantillon statistique est constitué de n variables aléatoires $\{Y_i; i = 1, \ldots, n\}$ indépendantes admettant des distributions issues d'une structure exponentielle.

La distribution d'une variable aléatoire Y_i appartient à la famille exponentielle si la fonction de densité peut être écrite sous la forme :

$$f(y_i;\phi,\theta_i) = \exp[e(y_i)h(\theta_i) + j(\theta_i) + d(y_i)]$$

avec e, h, j et d des fonctions connues. Cependant pour les modèles linéaires généralisés il faut qu'on puisse écrire la loi avec y_i sous la forme canonique, on cherche la forme suivante :

$$f(y_i;\phi,\theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}$$

où θ_i est le paramètre naturel. Chaque Y_i dépend d'un unique paramètre θ_i. Les θ_i ne sont pas forcément identiques et dépendent des variables explicatives. S'il existe d'autres

paramètres en plus du paramètre d'intérêt ils seront considérés comme paramètre de nuisance. ϕ est la dispersion, aussi appelée paramètre de nuisance ou *scale* et est utilisée lorsqu'il y a plusieurs paramètres ce qui est le cas des lois Normale et Gamma. Les fonctions a, b et c sont des fonctions spécifiques de la distribution, b étant deux fois dérivables à valeurs dans R et c à valeurs dans R^2.

Remarque : En considérant $\theta_i=x_i^T\beta$, cette écriture nous permet de faire ressortir la fonction lien naturelle (canonique) pour chaque loi. Cette fonction est celle qui est censée linéariser le mieux l'espérance.

- La composante déterministe

Soit X la matrice de régression et γ le vecteur des paramètres. La composante déterministe, aussi appelée prédicteur linéaire, est notée η et est définie par η = Xγ. Dans le cas du provisionnement, la composante systématique s'écrit :

$$\eta_i = \mu + \alpha + \beta$$

μ, α et β sont respectivement la constante, le premier facteur explicatif et le deuxième facteur explicatif.

- La fonction lien

La fonction g est appelée fonction lien (*link function*) et a pour rôle de linéariser l'espérance, c'est la fonction qui fait le lien entre la composante aléatoire et la composante systématique.

$$\begin{cases} g(\mu_i) = \eta_i \\ E[Y_i] = \mu_i \\ V[Y_i] = \phi v[\mu_i] \end{cases}$$

La fonction g est supposée monotone et différentiable.

Pour une fonction lien de type log, nous obtenons la relation suivante :

$$\ln(\mu_i) = \eta_i = \mu + \alpha + \beta \quad \Leftrightarrow \quad \mu_i = \exp(\mu + \alpha + \beta)$$

Pour une fonction lien de type *logit*, soit ln du *ods ratio*[25], nous obtenons la relation suivante :

$$\log it(\mu_i) = \ln(\frac{\mu_i}{1-\mu_i}) = \eta_i = \mu + \alpha + \beta \quad \Leftrightarrow \quad \mu_i = \frac{\exp(\mu + \alpha + \beta)}{1+\exp(\mu + \alpha + \beta)}$$

- Expression des moments

L'espérance et la variance s'obtiennent à partir de la fonction score ($U_i = \frac{\partial \log f(x_i;\theta_i,\phi)}{\partial \theta_i}$) :

$$\begin{cases} \mu_i = E[Y_i] = b'(\theta_i) \\ V[Y_i] = b''(\theta_i)\phi = v[\mu_i]\phi \end{cases}$$

[25] $u_i/(1-u_i)$ est appelé *ods ratio.*

Remarque : v est la fonction de variance spécifique à chaque distribution.

- Estimation des paramètres

L'estimation des paramètres γ des modèles linéaires généralisés se fait grâce à la méthode du maximum de vraisemblance. En règle général l'estimation de ces paramètres est obtenue numériquement par un processus itératif. Les méthodes les plus répandues sont celle de Newton-Raphson et scoring de Fischer[26].

La famille exponentielle a pour propriété de satisfaire les conditions de régularité qui assurent que le maximum globale de la fonction de vraisemblance l(θ ;y) est donné par la solution des équations ∂l/∂β=0 (ou ∂l/∂θ=0).On calcule :

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$$

avec x_{ij} le $j^{ème}$ élément de x_i^T, $E(Y_i)=\mu_i$ et $g(\mu_i)=\eta_i$.

L'hypothèse de nullité de l'estimateur sera testée à l'aide du test de vraisemblance qui a une distribution du χ^2 asymptotique.

4.3.2 L'ajustement du modèle

- La déviance

Le test d'ajustement du modèle se fait en comparant la vraisemblance du modèle estimé et la vraisemblance du modèle maximal, aussi appelé saturé ou plein. Le modèle maximal possède autant de paramètres que d'observations et estime donc de manière exacte les données.

Cette comparaison est basée sur l'expression de la déviance D qui est le logarithme du carré du rapport des vraisemblances :

$$D = 2[\log l(b_{\max}) - \log l(b)]$$

Asymptotiquement, D suit une loi du χ^2 à n-p degrés de liberté ce qui permet de construire un test de rejet ou d'acceptation du modèle.

- Le test du χ^2 de Pearson

Le test du χ^2 de Pearson est proche du test de la déviance. Il s'intéresse à la déviation des valeurs attendues par rapport aux valeurs observées déduites du modèle considéré :

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

Asymptotiquement, le test de Pearson suit la même loi que la déviance.

[26] Ces 2 méthodes sont brièvement présentées en annexe.

- Les résidus de *Pearson* et de déviance

Le résidu de Pearson correspond au résidu brut r_i divisé par la racine de la fonction de variance. Ce résidu correspond à la contribution de chaque observation au χ^2 de Pearson.

Le résidu de *Pearson* et le résidu de *Pearson* standardisé s'écrivent respectivement de la manière suivante :

$$r_{Psi} = \frac{r_i}{\sqrt{V(\hat{\mu}_i)}\sqrt{\hat{\phi}(1-h_i)}}$$

où $\hat{\phi}$ est l'estimation du paramètre de nuisance et h_i le $i^{ième}$ élément de la diagonale de l'approximation de la matrice de projection $H = \sqrt{W}X(X^T WX)^{-1}X^T\sqrt{W}$ où W est la matrice diagonale :

$$W_{ii} = \frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

Le résidu de déviance correspond à la déviance de chaque observation. Le résidu de déviance standardisé s'écrit :

$r_{Di} = \frac{signe(r_i)\sqrt{D_i}}{\sqrt{\hat{\phi}(1-h_i)}}$ où D_i est la déviance de l'observation

L'étude des résidus permet de voir si chaque observation est bien expliquée par le modèle. Nous effectuons une comparaison des résidus standardisés à une distribution Normale et une réalisation graphique des résidus standardisés en fonction de la valeur prédite.

- Le test d'Hosmer-Lemeshow

Chaque observation est classée suivant sa valeur prédite (de la plus petite à la plus grande). Cet ensemble est divisé en plusieurs groupes, pour chaque groupe nous disposons du nombre de valeur prédite et observée. Puis nous effectuons un test de χ^2, de degrés de liberté du nombre de groupes – 2, sur les valeurs prédites et les valeurs observées permettant de construire un test de rejet ou d'acceptation du modèle.

- Le test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est un test non paramétrique mesurant l'écart maximal entre deux distributions cumulées. Nous testons l'hypothèse nulle H_0 : « $F_A=F_B$ » (égalité des distributions des 2 populations) contre l'hypothèse H_1 : « $F_A \neq F_B$ » (les distributions sont différentes).

La statistique de test est obtenue de la manière suivante : $D = \sup_{x \in R} \left| F_A(x) - F_B(x) \right|$

Cette statistique est ensuite comparée à une valeur tabulée d qui dépend du nombre d'observations et le risque de $1^{ère}$ espèce α. L'hypothèse H_0 est rejetée si la statistique D est supérieure à la valeur tabulée d.

La faiblesse du test réside dans le fait qu'il n'est appliqué qu'en un seul point et non sur l'ensemble de l'intervalle.

- La courbe de Lorenz et l'indice de Gini[27]

La courbe de Lorenz permet de représenter graphiquement une dispersion. La courbe de Lorenz est située en dessous de cette diagonale de référence. Plus la courbe est éloignée de la diagonale, plus la répartition des revenus est inégalitaire.

La courbe de Lorenz permet de calculer l'indice de Gini qui est une mesure du degré d'inégalité.

L'indice de Gini varie entre 0 et 1. Si l'indice est de 0, cela signifie que la courbe de Lorenz est la diagonale, l'égalité est parfaite. Si l'indice est de 1, cela signifie qu'une seule personne détient tout le revenu, c'est l'inégalité maximale.

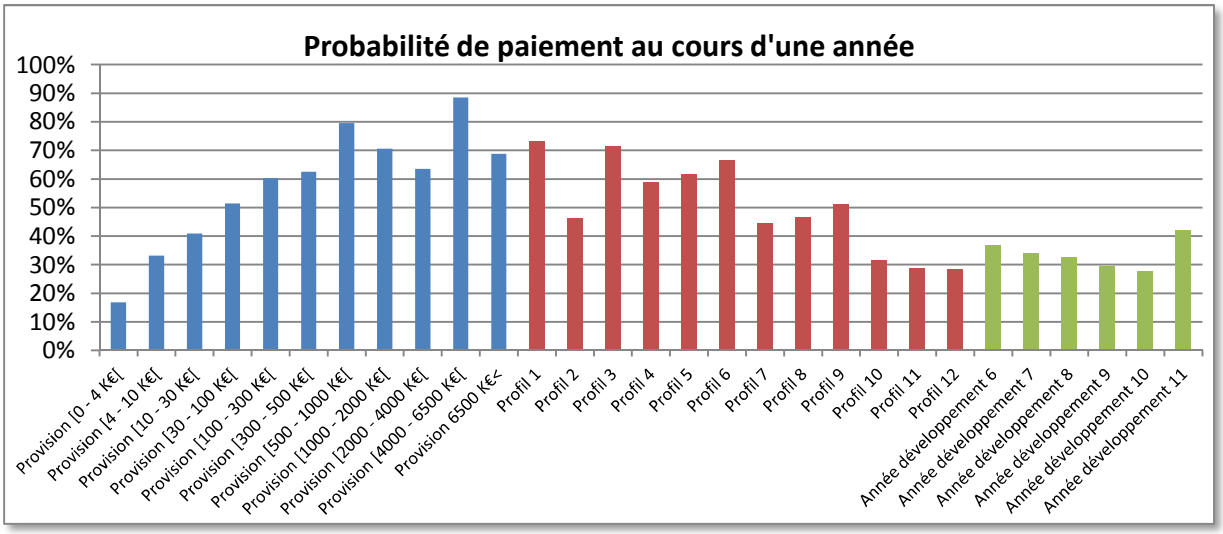
4.4 La probabilité de paiements

L'objectif est de modéliser la probabilité d'avoir un paiement au cours de l'année compte tenu des caractéristiques du sinistre en début d'année. Puis, nous chercherons à extrapoler le modèle obtenu afin de projeter ce modèle à un horizon de 40 ans[28].

4.4.1 Les données

La population d'étude est composée de 6 817 sinistres sur lesquels il y a 2 365 paiements répartis sur l'ensemble des exercices de survenance et l'année de développement, soit une probabilité globale d'observer un paiement au cours d'une année de 34,7 %.

Les variables explicatives utilisées pour ce modèle sont l'année de développement, le profil de durée et la provision dossier[29].



GRAPHIQUE 14 – Taux de paiement par critère

[27] Par la suite, le test Hosmer-Lemeshow, le test de Kolmogorov-Smirnov et l'indice de Gini seront uniquement développés dans le cadre de la régression logistique.
[28] Nous prenons pour hypothèse que la durée maximale d'un sinistre est de 40 ans.
[29] La variable exercice de survenance n'est pas présentée car non significative sur la probabilité de paiement.

Nous distinguons une tendance linéaire de chaque variable sur la probabilité de paiement au cours d'une année. Nous constatons que plus le montant de la provision dossier augmente et plus la probabilité de paiement est élevée. Ainsi, la probabilité de paiement est de 17% pour les sinistres ayant une provision [0-4K€] et atteint 70% pour les sinistres ayant des provisions supérieures à 6,5 M€.

La probabilité de paiement décroît en fonction du profil, les profils à durée longue (profils 1, 2 et 3) ont une probabilité de paiement d'environ 70% et ceux à durée plus courte (profil 10, 11 et 12) ont une probabilité de paiement de 30%.

Dans une mesure moindre et excepté pour l'année de développement 11, le taux de paiement diminue en fonction des années de développement.

4.4.2 Le modèle logistique

Nous modélisons la variable d'intérêt, la présence de paiement, en fonction des variables explicatives. Nous effectuons une régression logistique qui est caractérisée par une distribution binomiale de la variable d'intérêt et une fonction de lien *logit*.

	ddl	Value	Value/ddl
Scaled Deviance	6 795	7 715	1,14
Scaled Pearson	6 795	6 901	1,02

Test Type 3	ddl	Khi-square	Pr>Khi-sq
Année développement	5	29,57	<,0001
Profil	11	34,57	0,0003
Provision	10	678,09	<,0001

TABLEAU 22 – Test basé sur la vraisemblance et de type 3

Connaissant l'aspect approximatif du test de la déviance, l'usage est souvent de comparer la statistique avec le nombre de degrés de liberté, le modèle peut être jugé satisfaisant pour un rapport *Scaled deviance/ddl* ou *Scaled Pearson/ddl* plus petit que 1. Le rapport de Pearson conduit à accepter l'hypothèse d'adéquation du modèle contrairement au rapport de déviance qui nous conduit à ne pas accepter l'adéquation du modèle.

Les test de type 3 nous conduit à rejeter l'hypothèse de nullité de chaque variable et à considérer ainsi que les variables explicatives retenues ont un effet non nul sur la probabilité de paiement.

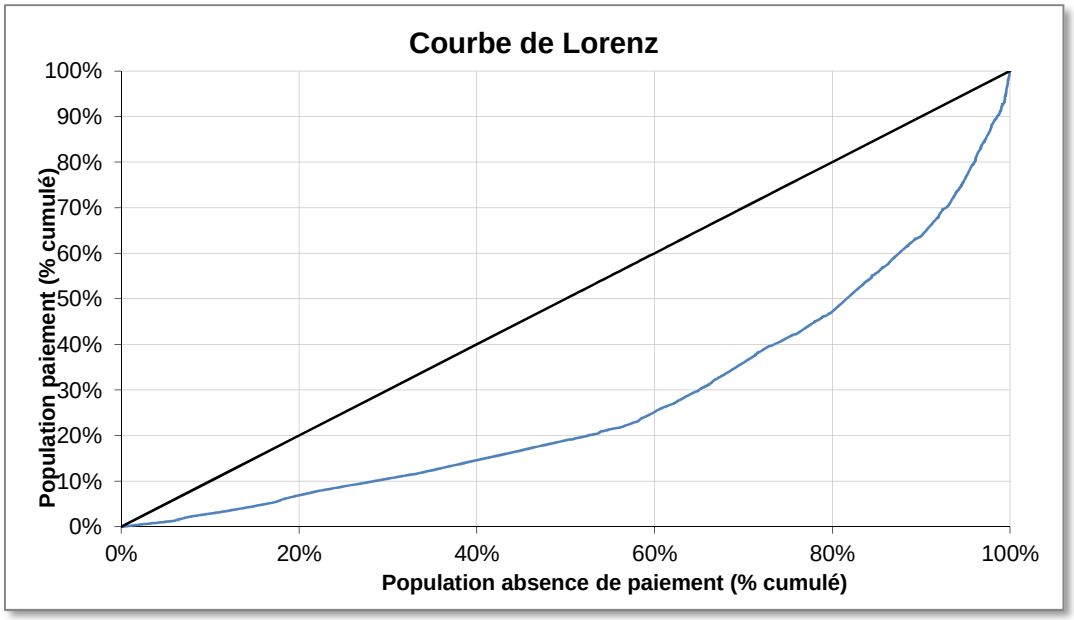
Paramètre	Estimation	Ecart-type	Pr>Khi-sq
Intercept	-1,461	0,060	<,0001
Année développement 11	0,160	0,248	0,520
Année développement 10	-0,534	0,159	0,001
Année développement 9	-0,439	0,112	<,0001
Année développement 8	-0,224	0,084	0,008
Année développement 7	-0,151	0,067	0,024
Année développement 6	0,000	0,000	,
Profil 1	0,707	0,349	0,043
Profil 2	-0,481	0,354	0,175
Profil 3	0,427	0,243	0,079
Profil 4	0,008	0,323	0,981
Profil 5	0,300	0,277	0,279
Profil 6	0,450	0,212	0,034
Profil 7	-0,086	0,171	0,618
Profil 8	0,118	0,149	0,429
Profil 9	0,279	0,098	0,005
Profil 10	-0,195	0,104	0,061
Profil 11	-0,186	0,082	0,024
Profil 12	0,000	0,000	,
Provision [0 - 4 K€[	0,000	0,000	,
Provision [4 - 10 K€[	0,917	0,085	<,0001
Provision [10 - 30 K€[	1,250	0,080	<,0001
Provision [30 - 100 K€[	1,613	0,089	<,0001
Provision [100 - 300 K€[	1,899	0,114	<,0001
Provision [300 - 500 K€[	1,969	0,178	<,0001
Provision [500 - 1000 K€[	2,761	0,207	<,0001
Provision [1000 - 2000 K€[	2,247	0,243	<,0001
Provision [2000 - 4000 K€[	1,966	0,299	<,0001
Provision [4000 - 6500 K€[	3,230	0,648	<,0001
Provision 6500 K€<	2,068	0,580	0,001

TABLEAU 23 – Estimation des paramètres

Pour chaque variable, il existe une modalité de référence, fixée à 0, correspondant à la modalité la plus représentée.

Pour les variables Provision et Année de développement, les *ods ratio* obtenus sont en phase avec l'analyse a priori. Nous constatons que les modalités de la variable Provision diffèrent significativement tous de la modalité de référence. De même pour la variable Année de développement où seule une modalité est significativement proche de la variable de référence.

Pour la variable Profil, les *ods ratio* obtenus sont moins discriminant que lors de l'analyse univariée indiquant ainsi une corrélation avec les autres variables du modèle.



GRAPHIQUE 15 – Courbe de Lorenz

L'indice de Gini est de 0,46 indiquant une capacité discriminatoire non négligeable.

ddl	Value	Pr>Khi-sq
8	12,8	0,119

TABLEAU 24 – Test de Hosmer-Lemeshow

Au seuil de 5%, la statistique de test n'appartient pas à la région de rejet, nous ne rejetons pas l'hypothèse d'égalité des valeurs prédites et observées.

L'ensemble des tests nous amène à accepter la justesse du modèle avec les données.

4.4.3 L'extrapolation des probabilités

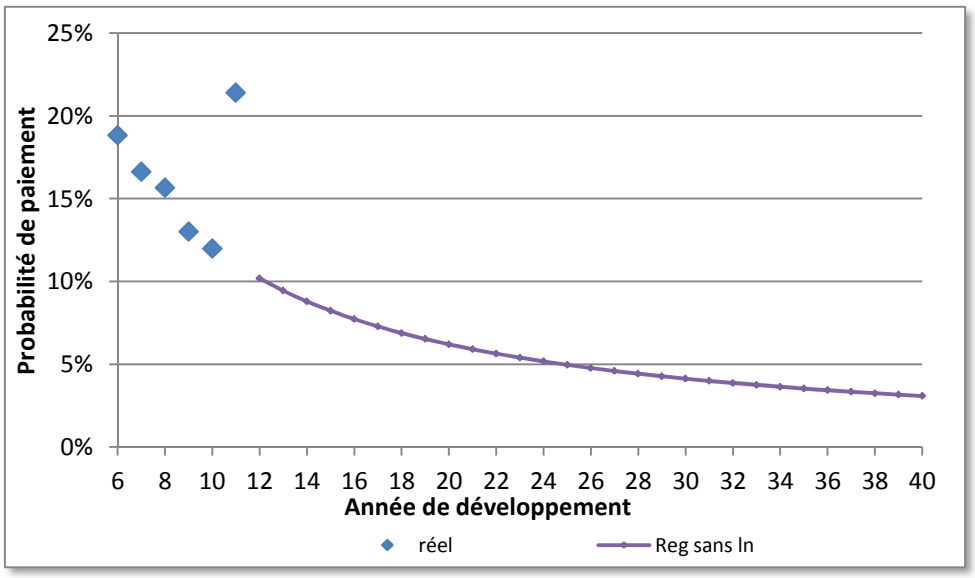
Le modèle individuel est basé sur une prédiction de paiements jusqu'à 40 ans. Le modèle en tant que tel n'offre pas une approche prospective, c'est-à-dire ne permet pas directement d'effectuer des prévisions jusqu'à l'ultime considéré dans le mémoire à 40 ans.

Comme pour la gestion de la liquidation incomplète, nous extrapolons les coefficients de régression à l'aide d'une régression linéaire de la forme y=a*ln(x)+b où y représente les coefficients de régression et x les années de développement[30].

Le modèle est développé sans le taux de paiement de la 11ème année de développement du fait de son caractère atypique. La courbe de régression retenue permet d'obtenir des coefficients pour la 12ème à la 40ème année de développement et ainsi de fournir une estimation de la probabilité de paiement jusqu'à l'ultime.

Nous illustrons l'effet du modèle sur les estimations de probabilité de paiement à l'aide de la typologie de sinistre la plus représenté {Profil 12 ; Provision [0 – 4 K€]}[31].

[30] La régression linéaire est présentée en annexe.
[31] L'estimation de probabilité de paiement pour le sinistre {Profil 1, Provision 6,5M€<} est présenté en annexe.



GRAPHIQUE 16 – Estimation de la probabilité à l'aide de la méthode d'extrapolation

Pour ce profil, la probabilité de paiement au cours d'une année est de 10% pour la 12ème année de développement et descend jusqu'à 4% la dernière année (40 ans).

4.5 Le montant des paiements

L'objectif est de modéliser le montant d'un paiement au cours de l'année compte tenu des caractéristiques du sinistre en début de l'année.

4.5.1 Les données

La population d'étude est composée de 2 365 paiements d'un montant moyen de paiement de 44 816€ et de médiane 5 523€.

	Année de développement						
Paiement	**6**	**7**	**8**	**9**	**10**	**11**	**Total**
Moyen	35 359	54 466	48 018	64 174	67 560	23 915	44 816
Médian	5 574	6 036	4 842	6 040	5 244	2 642	5 523
Coefficient de variation	382%	504%	315%	395%	516%	350%	446%

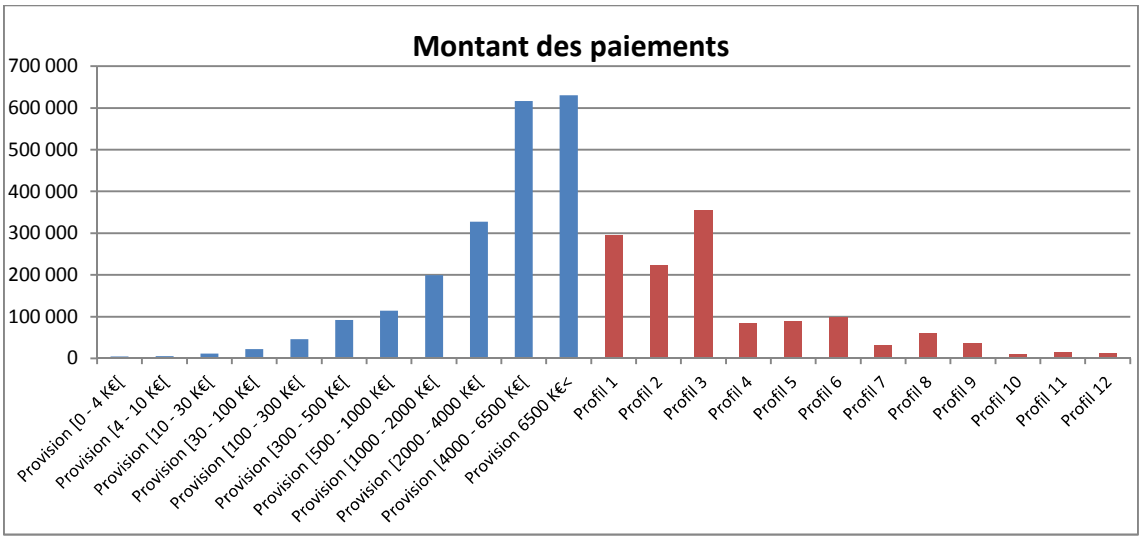
TABLEAU 25 – Paiement par année de développement

La distribution des paiements est biaisée, la moyenne est supérieure à la médiane signifiant qu'il existe des sinistres ayant des paiements élevés. Nous constatons un coefficient de variation élevé indiquant une forte volatilité au sein des paiements.

Par la suite, les variables explicatives utilisées pour ce modèle sont le profil de durée et la provision dossier. Il est à préciser qu'au seuil de 5%, les variables exercice de survenance et année de développement ne sont pas significatives et ne seront pas présentées par la suite.

Nous avons choisi de ne pas retenir la variable année de développement du fait de la *pvalue* du test de type3 estimée à 8%. La variable n'améliore pas significativement la justesse du

modèle. La non significativité de cette variable s'explique essentiellement par l'importante volatilité des paiements constatés au sein de chaque modalité[32].



GRAPHIQUE 17 – Montant des paiements par variable explicative

L'histogramme illustre la tendance entre les variables explicatives et le montant des paiements. Ainsi, nous observons que plus la provision dossier est élevée et plus le paiement est élevé. De même pour le profil de durée, les profils de durée longs ont des paiements supérieurs aux profils de durée plus courte.

4.5.2 Le modèle Lognormal

Dans cette partie, nous modélisons la variable d'intérêt, le montant du paiement, en fonction de la provision dossier et du profil de durée.

Cependant avant la modélisation, nous modifions les paiements de la population d'étude pour obtenir une base *as-if* 2013. L'objectif de la base *as-if* est d'obtenir des paiements comparables c'est-à-dire neutralisés de l'inflation.

A l'aide de l'indice des prix à la consommation[33], nous transformons les paiements de la population pour être sur une base commune de 2013.

Nous retenons une modélisation avec une distribution Lognormal de la variable d'intérêt et une fonction de lien identité.

	ddl	Value	Value/ddl
Scaled Deviance	2 343	2 365	1,009
Scaled Pearson	2 343	2 365	1,009
Test Type 3	ddl	Khi-square	Pr>Khi-sq
Profil	11	46	<0,0001
Provision	10	362	<0,0001

TABLEAU 26 – Test basé sur la vraisemblance

[32] Le choix de ne pas retenir la variable année de développement est discutable et pourra donner lieu à une extension du modèle. Les résultats du modèle sont en annexe. Il est à préciser que cette variable a une influence significative sur la probabilité de paiement.

[33] Source INSEE : Indice des prix à la consommation (série annuelle, ensemble des ménages, métropole + DOM, base 1998), Référence 000639201, Mise à jour 14 janvier 2014.

Le rapport de *Scaled* Deviance nous amène à accepter l'hypothèse d'adéquation du modèle contrairement au rapport de *Scaled* Pearson qui conduit à rejeter l'adéquation du modèle.

Les test de type 3 nous conduit à rejeter l'hypothèse de nullité de chaque variable et à considérer ainsi que les variables explicatives retenues ont un effet non nulle sur le montant de paiement.

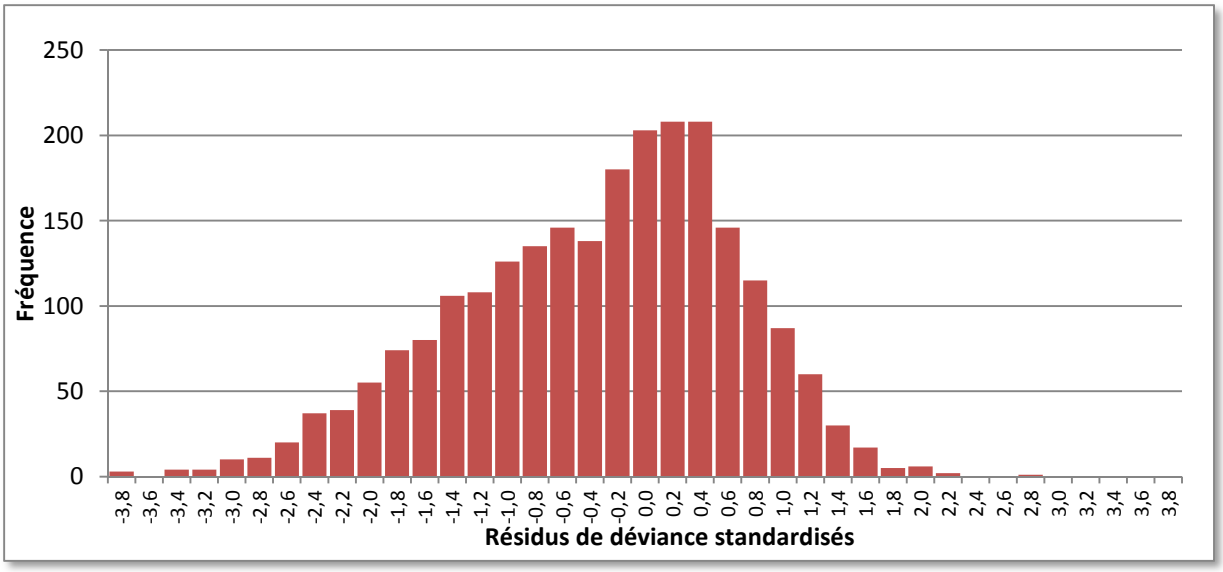
Paramètre	Estimation	Ecart-type	Pr>Khi-sq
Intercept	7,387	0,082	<.0001
Profil 1	1,068	0,310	0,001
Profil 2	0,355	0,429	0,407
Profil 3	0,907	0,238	0,000
Profil 4	-0,238	0,352	0,498
Profil 5	0,299	0,283	0,291
Profil 6	0,243	0,212	0,252
Profil 7	-0,307	0,210	0,144
Profil 8	0,025	0,180	0,889
Profil 9	0,150	0,118	0,204
Profil 10	-0,558	0,140	<.0001
Profil 11	-0,089	0,113	0,430
Profil 12	0,000	0,000	.
Provision [0 - 4 K€[	0,000	0,000	.
Provision [4 - 10 K€[	0,554	0,124	<.0001
Provision [10 - 30 K€[	1,187	0,113	<.0001
Provision [30 - 100 K€[	1,648	0,119	<.0001
Provision [100 - 300 K€[	2,018	0,140	<.0001
Provision [300 - 500 K€[	2,496	0,200	<.0001
Provision [500 - 1000 K€[	2,338	0,190	<.0001
Provision [1000 - 2000 K€[	2,635	0,245	<.0001
Provision [2000 - 4000 K€[	2,512	0,323	<.0001
Provision [4000 - 6500 K€[	3,571	0,416	<.0001
Provision 6500 K€<	3,294	0,561	<.0001
Scale	1,716	0,025	

TABLEAU 27 – Estimations des paramètres

Pour chaque variable, il existe une modalité de référence, fixée à 0, et correspondant à la modalité la plus représentée.

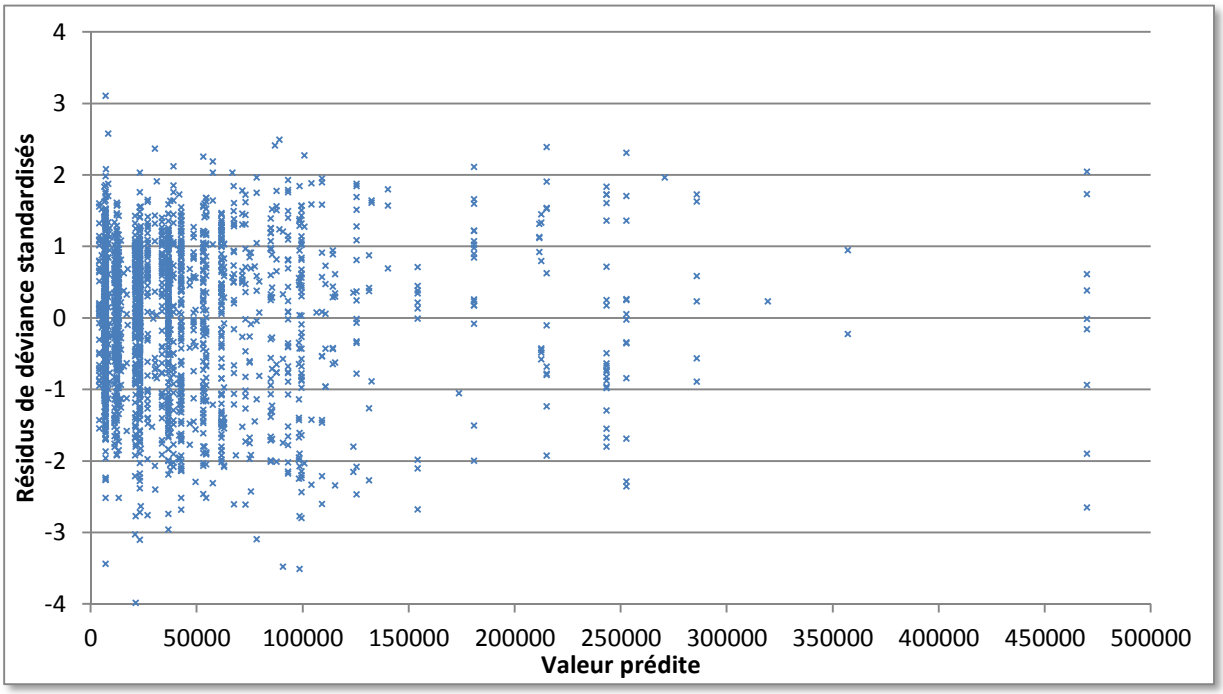
Nous constatons que toutes les modalités de la variable Provision diffèrent significativement de la modalité de référence. L'estimation des paramètres confirme l'analyse à plat mais de manière plus lissée. Ainsi, nous constatons que plus le montant de provision dossier est élevé et plus le montant de paiement est élevé. Cette variable est la plus influente des 2 variables utilisées.

Certaines modalités de la variable Profil sont significativement différentes de la modalité de référence. Nous constatons la tendance que plus le profil est de durée longue et plus le montant de paiement est élevé. La variable est moins discriminante que la provision dossier et aussi moins discriminante que lors de l'analyse à plat.



GRAPHIQUE 18 – Histogramme des résidus

La moyenne et l'écart type des résidus de déviance standardisés obtenus sont respectivement de 0 et de 1. Au seuil de 5%, nous observons que moins de 5% des valeurs n'appartiennent pas à l'intervalle [-1,96 ;1,96] et peuvent être considérées comme mal ajustées.



GRAPHIQUE 19 – Résidus en fonction des valeurs prédites

Les résidus de déviance standardisés appartiennent à un intervalle [- 1,96 ;1,96]. Les résidus ne croissent pas avec la valeur prédite et le graphique révèle une absence de structure signifiant que les variables utilisées sont pertinentes.

L'ensemble des tests nous amène à accepter la justesse du modèle avec les données.

4.6 Le modèle de paiement

Le modèle de paiement s'applique sur les sinistres ouverts, des exercices 2003 à 2009, afin d'obtenir les paiements pour les années restantes jusqu'à l'ultime. Les paiements sont estimés conditionnellement à l'information disponible à la date d'inventaire.

Nous illustrons le modèle des paiements à l'aide du sinistre non clos au 31/12/2013 ayant pour caractéristiques :

- Exercice de survenance : 2009,
- Provision dossier : 8 861 €,
- Profil de durée 12.

Nous fixons l'hypothèse que le sinistre est ouvert en début de chaque période et que l'inflation sur les années futures est nulle.

Date des paiements	Provision dossier δ	Probabilité Paiement P	Montant Paiement M	Paiements R
01/07/2014	8 861 €	37%	12 254 €	4 499 €
01/07/2015	4 362 €	33%	12 254 €	4 077 €
01/07/2016	285 €	13%	7 044 €	892 €

TABLEAU 28 – Paiement par année de développement

L'année 2014 correspond à la 6ème année de développement du sinistre. La probabilité d'avoir un paiement et le montant du paiement sont respectivement de 37 % et 12 254 €. L'espérance de paiement X est de 4 499 €.

La provision dossier 2015 est obtenue à partir de la provision dossier 2014 et le paiement X, nous déduisons une provision dossier de 4 362 € (= 8 861 € – 4 499 €). L'année 2015 correspond à la 7ème année de développement du sinistre. La probabilité de paiement est de 13 %, en diminution par rapport à l'année 2014. Le montant de paiement est identique à l'estimation de l'année 2014. Nous obtenons un paiement de 4 077 €.

Pour l'année 2016, l'application de la méthodologie permet d'obtenir une estimation de paiement de 892 €.

Pour l'année 2017, la provision dossier est nulle car les paiements cumulés sur la période sont supérieurs à la provision dossier au 31/12/2013. Ainsi, nous considérons que les paiements sont nuls à partir de l'année 2017.

5 PARTIE 5 - MODELE DE PROVISIONNEMENT FINAL

5.1 La population d'inventaire

Dans cette partie, nous nous intéressons aux exercices de survenance 2003 à 2009. Au 31/12/2013, il existe 1 681 sinistres non clos.

Exercice survenance	Effectif
2 003	74
2 004	95
2 005	117
2 006	203
2 007	290
2 008	354
2 009	548
Total	1 681

TABLEAU 29 – Sinistres non clos par exercice de survenance

Au 31/12/2013, l'exercice de survenance 2003 est dans la 11ème année de développement. Nous observons que 74 sinistres sont non clos. L'exercice de survenance 2009 atteint la 5ème année de développement et le nombre de sinistres non clos est de 548.

Le nombre de sinistres non clos est en lien avec l'année de développement. Plus l'année de développement est élevé et moins il y a de sinistres non clos.

	Année de développement						
Exercice survenance	5	6	7	8	9	10	11
2 003	545	389	284	207	160	93	**74**
2 004	554	385	265	191	119	**95**	
2 005	678	447	320	186	**117**		
2 006	648	457	271	**203**			
2 007	614	402	**290**				
2 008	520	**354**					
2 009	**548**						

TABLEAU 30 – Sinistres non clos par exercice de survenance et année de dévoppement

Le nombre de sinistres non clos et l'évolution du montant des provisions dossiers sont spécifiques à chaque exercice de survenance. Plus précisément, le montant de la provision dossier s'explique principalement par le nombre de sinistres non clos et la présence de « gros » sinistres.

Une grande partie du montant de provision dossier se concentre sur un nombre limité de sinistres non clos. Les 9 sinistres non clos évalués à plus de 4 M€ représentent 25% du montant de provision dossier total. De même, il existe 152 sinistres non clos, évalués par le gestionnaire à plus de 0,3 M€, qui représentent 85% du montant de provision dossier.

Il est à préciser que le montant de la provision dossier sur l'ensemble des exercices de survenance reste confidentiel et ne sera pas communiqué dans ce mémoire.

La répartition des sinistres non clos par exercice de survenance et classe de provision dossier évaluée par le gestionnaire illustre le caractère aléatoire des « gros » sinistres.

	Provision dossier											
Exercice de survenance	[0-4 K€[	[4-10 K€[	[10-30 K€[	[30-100 K€[	[100-300 K€[	[300-500 K€[	[0,5-1 M€[	[1-2 M€[	[2-4 M€[	[4-6,5 M€[	6,5 M€ et plus	Total
2 003	34	10	10	6	5	4	2	2	1	0	0	74
2 004	39	9	16	11	9	2	3	2	2	1	1	95
2 005	60	15	11	15	5	3	4	1	1	1	1	117
2 006	102	28	31	15	11	6	3	4	3	0	0	203
2 007	138	21	50	30	21	10	10	6	2	0	2	290
2 008	151	52	62	39	25	5	11	1	5	2	1	354
2 009	242	74	80	75	27	18	22	5	5	0	0	548

TABLEAU 31 – Nombre de sinistres non clos par provision dossier et exercice de survenance

Nous observons une absence de sinistre évalué par le gestionnaire à plus de 4 M€ pour les exercices de survenances 2003, 2006 et 2009. A contrario, l'exercice de survenance 2008 est concerné par 3 sinistres dans cette fourchette d'évaluation.

Pour les sinistres ayant une provision dossier supérieure à 4 M€, le montant cumulé des provisions dossier, par exercice de survenance, est volatil.

Exercice de survenance	Provision dossier	
	Entre 4 et 6,5 M€	6,5 M€ et plus
2 003	0	0
2 004	6 111 421	7 933 795
2 005	4 556 456	8 226 515
2 006	0	0
2 007	0	15 413 806
2 008	11 013 061	7 039 698
2 009	0	0

TABLEAU 32 – Montant cumulé de la provision dossier par exercice de survenance

Pour les exercices de survenance ayant un sinistre supérieur à 4 M€, le montant cumulé de la provision dossier varie entre 12,7 M€ (exercice de survenance 2005) et 18,1 M€ (exercice de survenance 2008).

La répartition des sinistres non clos par exercice de survenance et profil de durée met en exergue qu'en proportion, les exercices de survenance les plus anciens sont davantage composés de profils de durée longue.

	Profil			
Exercice survenance	1 à 3	4 à 6	7 à 9	10 à 12
2 003	5,4%	5,4%	17,6%	71,6%
2 004	7,4%	1,1%	15,8%	75,8%
2 005	4,3%	6,0%	14,5%	75,2%
2 006	3,5%	4,9%	15,8%	75,9%
2 007	3,8%	4,1%	15,5%	76,6%
2 008	2,5%	2,5%	18,9%	76,0%
2 009	0,9%	2,9%	14,8%	81,4%

TABLEAU 33 – Profil de durée par exercice de survenance

Pour l'exercice de survenance 2003, les profils de durée longue, c'est-à-dire les profils de 1 à 3, représentent 5,4 % des 74 sinistres non clos alors que pour l'exercice de survenance 2009, ces profils représentent 0,9 % des sinistres non clos.

Les profils 1 à 3 sont faiblement représentés dans la population mais leur répartition augmente au fur et à mesures des années de développement.

	Année de développement 5		Année de développement 11	
Profil	Effectif	Répartition	Effectif	Répartition
1 à 3	11	2,0%	4	5,4%
4 à 6	14	2,6%	4	5,4%
7 à 9	70	12,8%	13	17,6%
10 à 12	450	82,6%	53	71,6%
Total	545	100%	74	100%

TABLEAU 34 – Répartition des profils par année de développement pour l'exercice 2003

Pour l'exercice de survenance 2003, sur les 545 sinistres non clos de l'année de développement 5, il y a 11 sinistres de profil 1 à 3, soit 2,0 %. Parmi les sinistres non clos de l'année de développement 11, il y a 4 sinistres de profil 1 à 3, soit 5,4 %.

Nous nous intéressons à la dépendance entre le profil de durée et le montant de la provision dossier.

	Provision										
Profil	[0-4 K€[	[4-10 K€[	[10-30 K€[	[30-100 K€[	[100-300 K€[	[300-500 K€[	[0,5-1 M€[	[1-2 M€[	[2-4 M€[	[4-6,5 M€[	6,5 M€ et plus
1 à 3	0,8%	1,4%	0,4%	1,6%	3,9%	4,2%	10,9%	33,3%	47,4%	75,0%	80,0%
4 à 6	0,9%	1,0%	1,2%	2,1%	7,8%	29,2%	27,3%	19,0%	10,5%	0,0%	0,0%
7 à 9	7,4%	8,6%	11,9%	31,9%	50,5%	50,0%	38,2%	19,0%	10,5%	0,0%	0,0%
10 à 12	90,9%	89,0%	86,5%	64,4%	37,9%	16,7%	23,6%	28,6%	31,6%	25,0%	20,0%
Total	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

TABLEAU 35 – Répartition des provision dossier par profil de durée

Nous observons que 90,9 % des sinistres non clos avec une provision dossier [0-4 K €] appartiennent aux profils de durée courte c'est-à-dire les profils 10 à 12. A l'opposé, les sinistres à provision dossier plus forte sont caractérisés par des profils de durée longue c'est-à-dire les profils de 1 à 3. Parmi les sinistres non clos avec une provision supérieure à 2 M€, 57,1 % appartiennent aux profils 1 à 3, soit 16 sur 28 sinistres. Pour les sinistres avec une provision supérieure à 6,5 M€, 4 des 5 sinistres appartiennent aux profils 1 à 3.

La répartition des sinistres non clos par profil de durée et provision dossier met en exergue le lien entre ces critères. Plus la provision dossier est élevée et plus la durée de vie du sinistre est longue. De même, lorsque la provision dossier est faible, la durée de vie du sinistre est courte.

5.2 Fonctionnement du modèle individuel

Nous présentons le fonctionnement du modèle individuel à l'aide de 2 exemples différents de sinistres non clos à l'inventaire 31/12/2013.

	Date de survenance	Profil	Provision dossier
Exemple 1	27/12/2003	1	1 444 174 €
Exemple 2	30/01/2009	12	8 861 €

Le modèle individuel s'appuie sur la provision dossier δ, la probabilité de paiement dans l'année P, le montant de ce paiement M et la probabilité de survie en t P(T>t)[34]. Le paiement X est obtenu à partir de la probabilité et du montant de paiement.

Les exemples présentés incorporent un taux d'inflation de 1,50 % par an sur la période projective.

Dans l'exemple 1, la provision dossier est élevée et le sinistre appartient à la classe de durée longue. Durant l'année 2014, le sinistre débute sa 12ème année de développement. Compte tenu des caractéristiques, la probabilité de paiement est de 69 %, en cas de paiement le montant du paiement est 285 778 € et la probabilité que le dossier soit non clos au 01/01/2014 est de 100 %[35]. A partir de ces éléments l'estimation du paiement pour l'année 2014 est de 198 707 €.

$$\hat{R}_{12} = P(T > 12) * E[X_{12}] = 100\% * 69\% * 285\,778€ * (1 + 1{,}50\,\%) = 198\,707€$$

Le processus est séquentiel, les estimations effectuées pour l'année 2014 alimentent les estimations pour l'année 2015. La provision dossier pour l'année 2015 est obtenue à partir de la probabilité et du montant de paiement de l'année 2014 :

$$\delta_{2015} = \max(\delta_{2014} - 69\% * 285\,778€ * (1 + 1{,}50\,\%); 0) = 1\,245\,467€$$

Il est à préciser qu'une fois la provision dossier à 0, il n'y a plus de flux de paiement. La provision dossier joue le rôle de borne supérieure.

A partir de cette estimation, nous obtenons les éléments nécessaires à l'évaluation du paiement pour l'année suivante. Nous pouvons ainsi réitérer le processus pour les années suivantes jusqu'à l'ultime.

Date des paiements	Provision dossier δ	Probabilité Paiement P	Montant Paiement M	Probabilité Survie P(T>t)	Paiements R
01/07/2014	1 444 174 €	69%	285 778 €	100%	198 707 €
01/07/2015	1 245 467 €	67%	285 778 €	85%	166 682 €
01/07/2016	1 049 235 €	65%	285 778 €	72%	140 032 €
01/07/2017	855 330 €	74%	212 388 €	61%	102 697 €
01/07/2018	688 123 €	73%	212 388 €	52%	87 165 €
01/07/2019	521 447 €	72%	212 388 €	45%	74 048 €
01/07/2020	355 258 €	52%	248 618 €	38%	54 246 €
01/07/2021	212 442 €	49%	154 118 €	32%	27 322 €
01/07/2022	128 069 €	47%	154 118 €	28%	23 006 €
01/07/2023	44 814 €	39%	106 497 €	24%	11 362 €

TABLEAU 36 – Application du modèle individuel sur l'exemple 1

[34] La probabilité de survie est calculée au 1/1 de chaque année de développement future.
[35] Nous considérons que si le sinistre est non clos au 31/12/N-1, il le sera aussi au 01/01/N.

Pour ce sinistre, nous estimons les flux de paiement jusqu'en 2023, année où la provision dossier devient nulle, pour un montant total :

$$\hat{R} = \sum_{i=12}^{40} \hat{R}_i = 885\,267€$$

Nous constatons que le montant de la provision, considéré comme un *Best Estimate* non actualisé, est inférieur à l'évaluation du gestionnaire à l'inventaire et représente 61,3 % de la provision dossier au 31/12/2013. Ceci s'explique essentiellement par le caractère prudent de l'évaluation du gestionnaire et par la méthodologie de l'estimation de la provision, qui résulte de l'espérance des flux de paiements. Pour l'ensemble des sinistres et des exercices de survenances, nous constatons que la provision dossier est supérieure à l'estimation de la provision.

Dans l'exemple 2, le sinistre est caractérisé par le profil 12 et une évaluation de provision de 8 861 €.

Date des paiements	Provision dossier δ	Probabilité Paiement P	Montant Paiement M	Probabilité Survie P(T>t)	Paiements R
01/07/2014	8 861 €	37%	12 254 €	100%	4 567 €
01/07/2015	4 295 €	33%	12 254 €	62%	2 622 €
01/07/2016	94 €	16%	7 044 €	39%	454 €

TABLEAU 37 – Application du modèle individuel sur l'exemple 2

En 2014, le sinistre débute la 6ème année de développement. La probabilité d'observer un paiement au cours de cette année est de 37% et le montant d'un paiement est évalué à 12 254 €. La probabilité que le sinistre soit non clos au 01/01/2014 est de 100 %. Nous obtenons une estimation du paiement pour l'année 2014 de 4 567 €.

$$\hat{R}_6 = P(T > 6) * E[X_6] = 100\% * 37\% * 12\,254€ * (1 + 1{,}50\,\%) = 4\,567€$$

Pour ce sinistre, nous estimons des flux de paiement jusqu'en 2016 pour un montant total :

$$\hat{R} = \sum_{i=12}^{40} \hat{R}_i = 7\,642\,€$$

Comme pour le sinistre précédent, nous observons que le montant de la provision est inférieur à l'évaluation du gestionnaire à l'inventaire et représente 86,2 % de la provision dossier au 31/12/2013.

5.3 Le modèle final

Lors de cette partie, il s'agit d'appliquer modèle individuel déterministe aux exercices de survenance 2003 à 2009 puis de l'articuler avec le modèle agrégé pour projeter les exercices de survenance 2010 à 2013.

Le modèle présenté incorpore un taux d'inflation de 1,50% par an sur la période projective.

Le modèle sera déroulé en 4 phases :

1. Utilisation du modèle individuel déterministe pour projeter les paiements des sinistres non clos des exercices de survenance 2003 à 2009 à l'ultime.

Année de survenance	Année de développement														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2003	10,911	40,060	20,737	9,753	8,230	6,750	7,572	3,120	1,208	0,461	0,837	1,029	0,755	0,544	0,414
2004	12,211	41,740	26,226	13,300	15,936	3,725	7,447	5,062	4,901	4,201	1,948	0,870	0,574	0,392	0,264
2005	12,805	39,611	21,956	14,845	10,014	7,280	6,520	4,423	3,775	1,806	1,612	0,646	0,470	0,334	0,262
2006	11,432	38,585	21,887	14,397	8,730	7,091	4,310	3,049	1,830	1,065	1,004	0,487	0,356	0,256	0,191
2007	12,124	39,265	22,654	17,025	9,784	9,772	8,301	3,731	2,176	1,441	1,301	0,724	0,519	0,372	0,272
2008	11,127	39,007	22,609	10,617	9,501	6,187	4,600	2,863	1,784	1,281	1,144	0,638	0,461	0,248	0,194
2009	10,769	35,731	21,436	10,610	11,772	6,227	3,528	2,132	1,271	0,849	0,727	0,362	0,249	0,157	0,122
2010	9,573	31,103	21,828	10,469											
2011	8,233	33,732	22,155												
2012	8,460	33,126													
2013	9,446														

TABLEAU 38 – Flux de paiements décumulés sur sinistres non clos (M€)[36]

2. Calcul des coefficients de développement du triangle à partir des flux réels et projetés issus du modèle individuel. Cette étape se fait à partir du triangle de paiements cumulés.

Année de survenance	Année de développement														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2003	10,9	51,0	71,7	81,5	89,7	96,4	104,0	107,1	108,3	108,8	109,6	110,6	111,3	111,8	112,2
2004	12,2	54,0	80,2	93,5	109,4	113,1	120,6	125,6	130,5	134,7	136,7	137,6	138,1	138,5	138,8
2005	12,8	52,4	74,4	89,2	99,2	106,5	113,0	117,5	121,2	123,0	124,6	125,3	125,8	126,1	126,4
2006	11,4	50,0	71,9	86,3	95,0	102,1	106,4	109,5	111,2	112,2	113,1	113,5	113,8	114,0	114,1
2007	12,1	51,4	74,0	91,1	100,9	110,6	118,9	122,9	125,2	126,8	128,3	129,2	129,8	130,2	130,6
2008	11,1	50,1	72,7	83,4	92,9	99,0	104,1	107,3	109,5	111,1	112,6	113,5	114,0	114,4	114,7
2009	10,8	46,5	67,9	78,5	90,3	96,5	100,1	102,2	103,5	104,3	105,1	105,4	105,7	105,8	105,9
2010	9,6	40,7	62,5	73,0											
2011	8,2	42,0	64,1												
2012	8,5	41,6													
2013	9,4														

i	f_i	i	f_i	i	f_i	i	f_i
1	4,45544	11	1,00604	21	1,00027	31	1,00002
2	1,46000	12	1,00412	22	1,00020	32	1,00002
3	1,17556	13	1,00292	23	1,00015	33	1,00001
4	1,12258	14	1,00228	24	1,00012	34	1,00001
5	1,06943	15	1,00191	25	1,00009	35	1,00001
6	1,05895	16	1,00142	26	1,00007	36	1,00001
7	1,03260	17	1,00111	27	1,00005	37	1,00000
8	1,02196	18	1,00087	28	1,00004	38	1,00000
9	1,01419	19	1,00063	29	1,00003	39	1,00000
10	1,01092	20	1,00035	30	1,00003	40	1,00000

TABLEAU 39 – Flux de paiements cumulés et facteurs de développements

Par exemple le facteur de développement f_{10} est calculé de la manière suivante :

$$f_{10} = \frac{(109{,}6 + 136{,}7 + 124{,}7 + 113{,}1 + 128{,}3 + 112{,}6 + 105{,}1)}{(108{,}8 + 134{,}7 + 123{,}0 + 112{,}2 + 126{,}8 + 111{,}1 + 104{,}3)} = 1{,}01092$$

Les facteurs de développement sont des facteurs moyens qui permettent d'obtenir un *Best Estimate* non actualisé.

[36] Il est à noter que le tableau est déroulé jusqu'à l'ultime mais pour des raisons de présentation et de confidentialité, nous affichons uniquement jusqu'à la 15ème année de développement.

Les facteurs de développement obtenus sont identiques de ceux de modèle agrégé sur la période 1 à 4 car sont calculés sur les points réels. Puis les facteurs de développement du modèle individuel f_5 à f_9 sont inférieurs à ceux du modèle agrégé et à partir du facteur f_{10} sont supérieurs à ceux du modèle agrégé.

3. Projection des paiements des exercices de survenance 2010 à 2013 à partir des coefficients de développement *Chain Ladder*,

		Année de développement														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Année de survenance	2003	10,911	40,060	20,737	9,753	8,230	6,750	7,572	3,120	1,208	0,461	0,837	0,985	0,715	0,509	0,389
	2004	12,211	41,740	26,226	13,300	15,936	3,725	7,447	5,062	4,901	4,201	1,948	0,870	0,574	0,392	0,264
	2005	12,805	39,611	21,956	14,845	10,014	7,280	6,520	4,423	3,775	1,806	1,612	0,646	0,470	0,334	0,262
	2006	11,432	38,585	21,887	14,397	8,730	7,091	4,310	3,049	1,714	0,967	0,909	0,416	0,297	0,205	0,148
	2007	12,124	39,265	22,654	17,025	9,784	9,772	8,301	3,954	2,366	1,604	1,474	0,840	0,615	0,452	0,404
	2008	11,127	39,007	22,609	10,617	9,501	6,187	5,030	3,271	2,159	1,599	1,459	0,894	0,522	0,403	0,325
	2009	10,769	35,731	21,436	10,610	11,772	6,227	3,528	2,132	1,271	0,849	0,727	0,362	0,249	0,157	0,122
	2010	9,573	31,103	21,828	10,469	8,945	5,687	5,165	3,025	2,103	1,389	1,084	0,606	0,416	0,297	0,231
	2011	8,233	33,732	22,155	11,257	9,240	5,875	5,335	3,124	2,173	1,435	1,120	0,626	0,430	0,306	0,239
	2012	8,460	33,126	19,130	10,659	8,749	5,563	5,052	2,958	2,057	1,359	1,061	0,593	0,407	0,290	0,226
	2013	9,446	32,640	19,359	10,787	8,854	5,630	5,112	2,994	2,082	1,375	1,073	0,600	0,412	0,293	0,229

TABLEAU 40 – Paiements décumulés pour les exercices de survenance 2010 à 2013 (M€)

Les paiements pour les exercices de survenances 2010 à 2013 sont projetés à l'ultime, soit 40 années de développement.

4. Obtention des provisions à partir de l'ultime et des paiements réels.

Exercice survenance	Provision
2 003	3 682 926 €
2 004	5 125 954 €
2 005	6 202 739 €
2 006	5 142 480 €
2 007	12 957 506 €
2 008	16 670 431 €
2 009	15 929 736 €
2 010	29 708 726 €
2 011	41 945 351 €
2 012	58 847 412 €
2 013	92 194 420 €
Total	288 407 680 €

TABLEAU 41 – Provision par exercice de survenance

Le montant de provision sur les exercices de survenance 2003 à 2013 est évalué à 288,4 M€, soit 65,7 M€ pour les exercices de survenances 2003 à 2009 et 222,7 M€ pour les exercices de survenance 2010 à 2013.

Nous constatons que les provisions obtenues ne sont pas toujours décroissante en fonction de l'ancienneté des exercices de survenance, pour les années 2009 et antérieurs pour lesquelles le modèle individuel est appliqué. La provision dépend ainsi de la composition des sinistres de l'exercice de survenance.

5.4 Benchmark avec les méthodologies agrégées

5.4.1 Comparaison avec *Chain Ladder*

Dans cette partie nous comparons les résultats obtenus à partir du modèle déterministe avec le modèle *Chain Ladder*[37].

Exercice survenance	Modèle individuel	*Chain Ladder*	Ecart modèle individuel et Chain Ladder
2 003	3 682 926 €	1 466 078 €	2 216 848 €
2 004	5 125 954 €	2 852 339 €	2 273 615 €
2 005	6 202 739 €	4 981 999 €	1 220 740 €
2 006	5 142 480 €	7 715 276 €	- 2 572 797 €
2 007	12 957 506 €	12 868 626 €	88 879 €
2 008	16 670 431 €	17 806 029 €	- 1 135 598 €
2 009	15 929 736 €	23 642 604 €	- 7 712 868 €
2 010	29 708 726 €	30 388 493 €	- 679 768 €
2 011	41 945 351 €	42 647 528 €	- 702 177 €
2 012	58 847 412 €	59 512 300 €	- 664 888 €
2 013	92 194 420 €	92 867 297 €	- 672 876 €
Total	288 407 680 €	296 748 569 €	- 8 340 889 €

Regroupement sur les périodes 2003 à 2009 et 2010 à 2013

Exercice survenance	Modèle individuel	*Chain Ladder*	Ecart
2003 à 2009	65 711 771 €	71 332 951 €	- 5 621 180 €
2010 à 2013	222 695 909 €	225 415 618 €	- 2 719 709 €
Total	288 407 680 €	296 748 569 €	- 8 340 889 €

TABLEAU 42 – Provisions obtenues avec le modèle individuel et *Chain Ladder*

Le montant de la provision globale obtenue avec le modèle individuel est de 228,7 M€ et celui de *Chain Ladder* de 296,7 M€, soit un écart de de 8,3 M€ (2,8 %).

Le regroupement des exercices de survenance sur le période 2003 à 2009, soit les exercices projetés avec le modèle individuel, indique un écart de provision entre les 2 modèles de - 5,6 M€. Sur la période 2010 à 2013, l'écart de provision entre les 2 modèles est de - 2,7 M€ et s'explique essentiellement par les flux projetés sur la période 2003 à 2009.

Nous observons que les écarts diffèrent pour chaque exercice de survenance. Pour les exercices 2003 à 2005, les montants de provision obtenus avec le modèle individuel sont supérieurs à ceux de *Chain Ladder*. Puis sur les exercices suivants, la tendance s'inverse et atteint -7,7 M€ sur l'exercice de survenance 2009.

Pour l'exercice de survenance 2003, l'écart de 2,2 M€, entre le modèle individuel et *Chain Ladder,* s'explique par l'effet des *tail factors* qui font diminuer les paiements plus rapidement que ceux du modèle individuel.

Le montant des provisions calculées à l'aide modèle individuel, pour les exercices de survenance 2004 et 2005, est supérieur à ceux obtenus avec la méthodologie *Chain Ladder* du fait principalement de la présence de gros sinistres non clos.

[37] Modèle *Chain Ladder* présenté à la partie 2.

L'exercice de survenance 2006 est caractérisé par une absence de sinistres non clos évalués à plus de 4 M€ contrairement aux 2 exercices précédents obtenant ainsi une estimation de la provision avec le modèle individuel inférieure à celle de *Chain Ladder*.

L'exercice de survenance 2008 est caractérisé par un nombre de sinistres ayant une provision dossier >1 M€ inférieur aux exercices précédents. Ainsi, il y a 9 sinistres non clos alors que sur les exercices précédents, au même stade de développement il y a en moyenne sur les exercices précédents 12 sinistres non clos.

Comme pour les exercices 2003 et 2006, l'exercice de survenance 2009 est caractérisé par une absence de sinistres de >4M€. L'écart entre les 2 modèles est de 7,7 M€ et s'explique essentiellement par l'absence de gros sinistres.

Les écarts de provision constatés sur les exercices de survenance résultent essentiellement de la composition en sinistres importants. La présence d'un ou plusieurs sinistres importants joue considérablement sur le montant des paiements futurs et ainsi, sur l'évaluation de la provision future.

Pour les exercices de survenance 2010 à 2013, l'écart est stable dans le temps et résulte de la méthodologie utilisée. En effet, sur ces exercices le produit des facteurs de développement calculés à partir du modèle individuel est inférieur à celui de *Chain Ladder* car les paiements projetés, des exercices de survenance 2003 à 2009, par le modèle individuel sont inférieurs à ceux projetés par le modèle *Chain Ladder*.

La duration des sinistres non clos au 31/12/2013 diffère entre les 2 méthodologies, nous constatons une duration plus élevée pour le modèle individuel notamment sur les exercices 2003 à 2009.

Exercice de survenance	Modèle individuel	*Chain Ladder*
2003-2009	3,42	2,72
2010-2013	3,28	3,17
2003-2013	3,31	3,06

TABLEAU 43 – Duration des provisions

Pour les exercices de survenance 2003-2009 l'écart de duration entre le modèle individuel et *Chain Ladder* est de 0,7 an. Sur certains exercices de survenances tel que l'exercice de survenance 2005, l'écart de duration peut être supérieur à un an.

Pour le modèle individuel, nous observons que la duration restante sur les sinistres non clos des exercices de survenances 2003 à 2009 est supérieure à ceux des exercices de survenance 2010 à 2013.

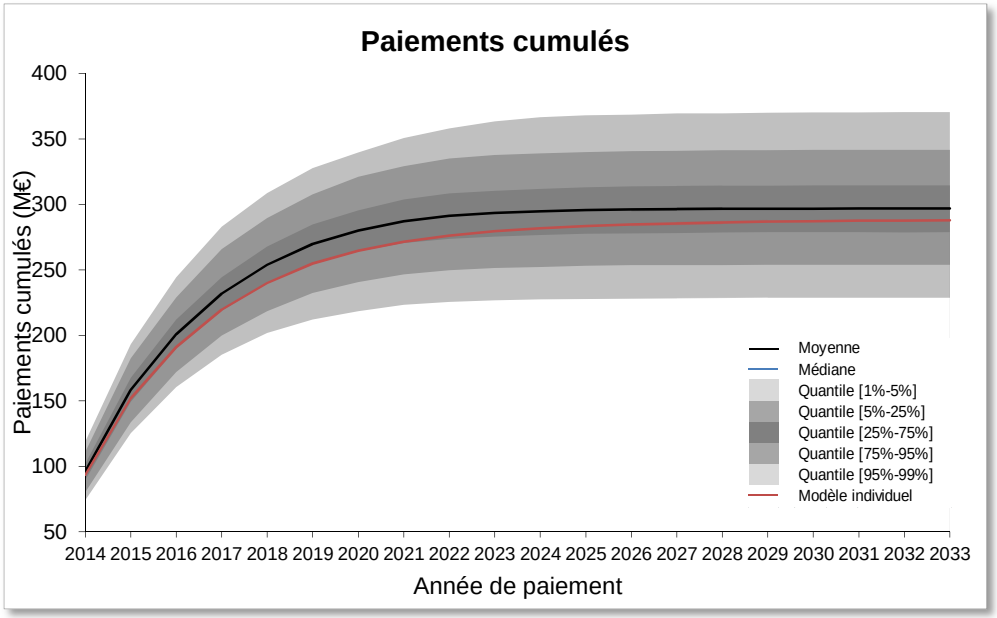
Cette dynamique s'explique essentiellement par la nature des sinistres non clos, il existe en proportion davantage de sinistres à profil de durée longue sur les exercices de survenance anciens.

Nous précisons que l'espérance de vie du sinistre est indépendant de son passé et nous remarquons que l'espérance conditionnelle sur les exercices de survenance augmente légèrement dans le temps.

5.4.2 Comparaison avec *Chain Ladder* stochastique

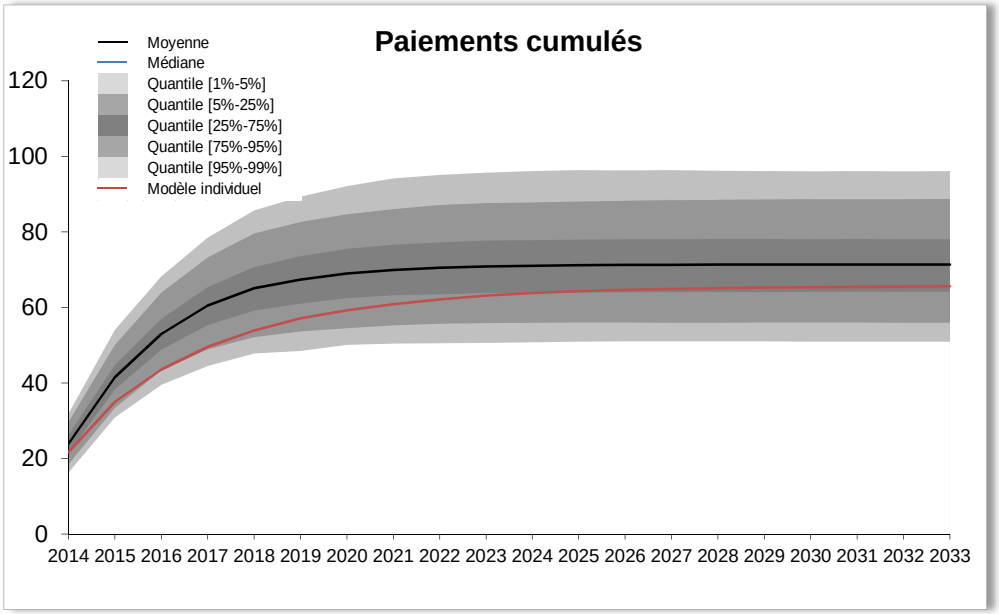
Dans cette partie nous comparons les résultats obtenus à partir du modèle individuel avec le modèle *Chain Ladder* stochastique[38].

Nous observons que les paiements cumulés issus du modèle individuel sont en phase avec le modèle *Chain Ladder* stochastique et appartiennent à un couloir situé entre les quantiles 25% et 75%.



GRAPHIQUE 20 – Exercices de survenance 2003 à 2013

Néanmoins, sur les exercices 2003 à 2009, les résultats du modèle individuel s'éloignent de ceux du modèle *Chain Ladder*.



GRAPHIQUE 21 – Exercices de survenance 2003 à 2009

[38] Modèle *Chain Ladder* présenté à la partie 2.

Les paiements cumulés issus du modèle individuel à l'ultime se situent dans le quantile 25-75 mais nous constatons une cadence plus lente.

L'écart des paiements cumulés, entre le modèle individuel et *Chain Ladder*, est de 2,5 M€ en 2014 et augmente jusqu'en 2018 pour atteindre 11,2 M€. Puis, l'écart se réduit pour atteindre 5,6 M€ à l'ultime.

L'écart sur la période 2014 à 2018 s'explique essentiellement par les exercices de survenances 2009 et 2006 qui sont caractérisés par une absence de gros sinistres. Pour l'exercice de survenance 2009, les paiements cumulés du modèle individuel sont inférieurs aux paiements *Chain Ladder* de 7,2 M€ en 2018 et atteignent 7,7 M€ à l'ultime. Pour l'exercice de survenance 2006, l'écart atteint 2,6 M€ en 2015 et reste stable jusqu'à l'ultime.

Sur la période 2019 à l'ultime, l'écart des paiements cumulés se réduit pour atteindre 5,6 M€ à l'ultime. La diminution des écarts est liée aux exercices de survenances 2003 à 2005, 2007 et 2008 où les paiements du modèle individuel sont supérieurs à ceux obtenus par les *tails factors*.

5.5 Best Estimate

Dans le cadre de Solvabilité 2, les flux de paiements restants sur sinistre sont actualisés à l'aide de la courbe EIOPA[39]. Ce qui aura pour conséquence d'obtenir un écart entre le modèle individuel et le modèle agrégé plus conséquent en solvabilité 2.

Exercice survenance	*Modèle individuel*	*Chain Ladder*	Ecart modèle individuel et *Chain Ladder*
2003 à 2009	63 064 212	69 583 509	-6 519 297
2010 à 2013	214 530 392	217 904 661	-3 374 269
Total	277 594 604	287 488 171	-9 893 566

TABLEAU 44 – *Best Estimate* hors marge de risque

Les *Best Estimate* du modèle individuel, 277,6 M€, et de *Chain Ladder*, 287,5 M€, sont inférieurs aux provisions de respectivement de 3,7 % et 3,1 %.

L'actualisation des paiements par la courbe de taux EIOPA couplé à une cadence de liquidation différente conduisent à un écart de *Best Estimate* de 9,9 M€ soit 3,5 %. Dans le cadre de solvabilité 2, nous constatons que l'écart entre les 2 modèles est plus conséquent.

5.6 Sensibilités

Dans cette partie, nous souhaitons sensibiliser le scénario d'inflation utilisé par le gestionnaire dans l'évaluation de la provision dossier.

L'évaluation de la provision par le gestionnaire incorpore implicitement un scénario d'inflation pour les paiements des années futures. Pour le scénario central nous avons pris pour hypothèse que les gestionnaires considéraient une inflation de 1,50 % sur la période de projection.

[39] Dans notre cas, nous utilisons la courbe EIOPA au 31/12/2013.

Il se peut que l'estimation de l'inflation des gestionnaires ne soit pas en phase avec le scénario central. L'inflation implicitement utilisée dans les évaluations pourrait par exemple être supérieure. Nous étudions aussi le cas où l'inflation descend en dessous du scénario central.

Nous souhaitons connaître l'impact sur le montant des provisions et sur la duration d'une inflation évaluée par le gestionnaire différente de celle du scénario central.

Inflation	Provision	Duration
0,00%	284 476 176 €	3,27
0,50%	285 648 048 €	3,28
1,00%	287 170 888 €	3,30
1,50%	**288 407 680 €**	**3,31**
2,00%	289 567 000 €	3,32
2,50%	291 090 618 €	3,34
3,00%	291 990 824 €	3,35
3,50%	293 074 329 €	3,36
4,00%	294 291 501 €	3,37
4,50%	295 618 343 €	3,39
5,00%	296 957 567 €	3,40
7,50%	302 610 340 €	3,45

TABLEAU 45 – Provision et duration en fonction de l'inflation

Le montant de provision et la duration augmentent de manière continue en fonction de l'inflation. A titre d'exemple, si l'inflation considérée par le gestionnaire sur la période de projection était de 5%, le montant de provision serait de 297,0 M€ et la duration de 3,39 soit une hausse respective par rapport au scénario central de 8,5 M€ et 0,1 an.

5.7 Backtesting

L'objectif de cette partie est d'appliquer le modèle individuel et *Chain Ladder* au triangle vu au 31/12/2011 et de comparer les résultats avec les données réelles notamment pour les exercices de survenance 2003 à 2007, exercices de survenance pour lesquels les sinistres non clos ont plus de 4 ans de durée de vie.

Il est à préciser que la méthodologie *Chain Ladder* est appliquée dans sa forme la plus simple et qu'aucun retraitement de donnée ou modification manuelle de coefficient n'ont été effectuées. Il en est de même pour la méthodologie de déroulé des sinistres individuels.

Le modèle individuel est ajusté de l'inflation réelle[40] pour correspondre aux données sinistres au 31/12/2011.

[40] Source INSEE : Indice des prix à la consommation (série annuelle, ensemble des ménages, métropole + DOM, base 1998), Référence 000639201, Mise à jour 14 janvier 2014.

	Paiements en 2012				
Exercice survenance	Réel	Modèle individuel	*Ecart Modèle ind. vs Réel*	*Chain Ladder*	*Ecart Chain Ladder vs Réel*
2 003	460 922 €	1 478 340 €	1 017 418 €	2 114 139 €	1 653 217 €
2 004	4 900 524 €	3 342 379 €	- 1 558 145 €	1 416 396 €	- 3 484 128 €
2 005	4 423 123 €	4 367 756 €	- 55 367 €	4 117 639 €	- 305 485 €
2 006	4 309 815 €	4 641 798 €	331 983 €	6 958 752 €	2 648 937 €
2 007	9 771 551 €	9 521 136 €	- 250 415 €	6 369 977 €	- 3 401 574 €
Total	23 865 935 €	23 351 409 €	- 514 526 €	20 976 902 €	- 2 889 033 €

TABLEAU 46 – Estimation des paiements durant l'année 2012

Les paiements constatés durant l'année 2012 pour les exercices de survenance 2003 à 2007 sont de 23,9 M€. L'estimation du modèle individuel, 23,4 M€, est plus proche des paiements réels durant l'année 2012 que l'estimation issue de *Chain Ladder,* 21,0 M€.

Ce résultat est intéressant à double titre : nous constatons que le modèle individuel est plus proche de la réalité et ainsi une meilleure vision *Best Estimate* alors que la méthodologie *Chain Ladder* sous-estime les paiements réels et n'offre pas une vision prudente.

De plus, l'estimation des paiements issue des deux modèles est différente par exercice de survenance. Pour chaque exercice de survenance, le modèle individuel est plus proche des données réelles que la méthodologie *Chain Ladder*. Nous observons que là où le Chain Ladder est trop imprudent, le modèle individuel est plus proche du réel et là où Chain Ladder est trop prudent, le modèle individuel est plus proche du réel. Pour le modèle individuel, l'écart maximal est de 1,5 M€ alors que pour *Chain Ladder* il existe un écart de plus de 1,5 M€ en valeur absolue pour 4 exercices de survenance.

Sur l'exercice de survenance 2004, nous observons un fort écart des deux méthodologies avec le réel. L'exercice de survenance 2004 se distingue de celui de 2003 par la présence de plusieurs paiements supérieurs à 1 M€. Le coefficient de développement utilisé par *Chain Ladder* pour estimer le paiement durant l'année 2012 conduit ainsi à sous-estimer le flux réel. Le modèle individuel est basé sur l'espérance et dès lors qu'il existe un nombre ou des montants de paiements plus élevés qu'attendus, le modèle individuel, par construction, sous-estime les paiements.

Nous observons que les estimations obtenues avec le modèle individuel sont plus en phase avec les paiements réels et moins volatiles que celles obtenues avec *Chain Ladder*.

Pour l'année 2013, les estimations de paiements issues des modèles sous-estiment les paiements réels. L'écart du modèle individuel et *Chain Ladder* avec les données réelles est respectivement de 4,8 M€ et 4,2 M€.

L'année 2013 est atypique par rapport aux autres années car est caractérisée par le montant élevé des paiements sur sinistres importants. Sur cette année, nous observons, trois gros paiements, notamment sur les exercices de survenance 2004 pour 2,8 M€, 2005 pour 2,8 M€ et 2007 pour 3,0 M€.

	Paiements en 2013				
Exercice survenance	Réel	Modèle individuel	*Ecart Modèle ind. vs Réel*	*Chain Ladder*	*Ecart Chain Ladder vs Réel*
2 003	837 009 €	1 302 574 €	465 564 €	849 728 €	12 718 €
2 004	4 200 713 €	2 593 880 €	- 1 606 833 €	2 479 507 €	- 1 721 206 €
2 005	3 774 594 €	2 892 809 €	- 881 785 €	1 320 627 €	- 2 453 967 €
2 006	3 049 149 €	2 703 653 €	- 345 496 €	3 973 599 €	924 451 €
2 007	8 301 049 €	5 912 666 €	- 2 388 383 €	7 306 355 €	- 994 695 €
Total	20 162 515 €	15 405 581 €	- 4 756 934 €	15 929 816 €	- 4 232 699 €

TABLEAU 47 – Estimation des paiements durant l'année 2013

Les estimations de la méthodologie *Chain Ladder* s'éloignent du réel, surtout pour les exercices de survenance 2004 et 2005, car ces deux exercices sont en rupture avec l'exercice 2003 sur lequel sont calculés les facteurs de développement.

Pour le modèle individuel, les forts écarts sont pour les exercices de survenance 2004 et 2007. Comme pour l'année 2012, ces écarts s'expliquent principalement par la mécanique de construction du modèle.

La présence de gros paiements est aléatoire, nous nous intéressons aux paiements de l'année en écrêtant les gros sinistres à un niveau similaire à ceux de l'année 2012 :

	Paiements en 2013				
Exercice survenance	Réel	Modèle individuel	*Ecart Modèle ind. vs Réel*	*Chain Ladder*	*Ecart Chain Ladder vs Réel*
2 003	837 009 €	1 302 574 €	465 564 €	849 728 €	12 718 €
2 004	2 756 967 €	2 593 880 €	- 163 087 €	2 479 507 €	- 277 460 €
2 005	2 308 385 €	2 892 809 €	584 424 €	1 320 627 €	- 987 758 €
2 006	3 049 149 €	2 703 653 €	- 345 496 €	3 973 599 €	924 451 €
2 007	6 626 632 €	5 912 666 €	- 713 967 €	7 306 355 €	679 722 €
Total	15 578 143 €	15 405 581 €	- 172 562 €	15 929 816 €	351 673 €

TABLEAU 48 – Estimation des paiements écrêtés durant l'année 2013

Lorsque les paiements sur gros sinistres sont ramenés à un niveau moyen constaté sur les autres années, nous observons que les paiements issus des deux modèles sont plus proches du réel par exercice de survenance.

Le *backtesting* met en exergue que le modèle individuel permet d'obtenir des résultats similaires à l'approche *Chain Ladder* tout en réduisant très légèrement l'écart des estimations avec le réel sur les différents exercices de survenance. Ceci est directement observable pour l'année 2012. La nature aléatoire du montant des paiements pour l'année 2013 rend difficile la comparaison des paiements. Néanmoins, l'écrêtement des paiements permet de retrouver une tendance similaire à l'année 2012.

Nous pensons qu'il faudrait tester sur d'autres garanties corporelles et sur un horizon plus long pour conclure plus solidement.

5.8 Récapitulatif des résultats

L'étude des sinistres non clos sur les exercices de survenance « anciens » a mis en exergue qu'une grande partie du montant de la provision dossier, élément clé dans l'évaluation des paiements futurs, se concentre sur un nombre limité de sinistres non clos. Le nombre de

sinistres importants, c'est-à-dire ayant une évaluation provision dossier élevée, est spécifique à chaque exercice de survenance.

De plus, nous constatons un lien entre l'évaluation de la provision du gestionnaire et la durée du sinistre. Un sinistre à provision dossier élevée dispose d'une durée de vie plus longue.

Nous constatons que le montant de la provision issu du modèle individuel est inférieur à l'évaluation du gestionnaire à l'inventaire s'expliquant essentiellement par le caractère prudent de l'évaluation du gestionnaire.

Le modèle individuel permet de projeter les flux de paiements des sinistres non clos pour les exercices de survenance 2003 à 2009 en tenant compte des caractéristiques inhérentes à chaque sinistre.

L'articulation avec la méthodologie agrégée *Chain Ladder* se fait aisément et permet de projeter les flux pour les exercices de survenance 2010 à 2013.

La provision obtenue au global avec le modèle individuel soit 288,7 M€, est en phase avec celle obtenue à partir de *Chain Ladder,* soit 296,7 M€. Néanmoins, les provisions par exercice de survenance présentent des écarts selon le modèle utilisé. Ces écarts résultent essentiellement de la composition des exercices de survenance en sinistres importants. Nous constatons par exemple que pour l'exercice de survenance 2009 l'absence de sinistres conduit à une évaluation de la provision inférieure de plus 7 M€ à celle de *Chain Ladder*.

Le *backtesting* met en exergue que le modèle individuel permet d'obtenir des résultats similaires à l'approche *Chain Ladder* tout en réduisant légèrement l'écart des estimations avec le réel sur les différents exercices de survenance.

Des disparités de sinistres existent entre les exercices de survenances, ce qui conduit à des différences de déroulés. Ainsi, la dynamique d'un exercice de survenance peut différer significativement de l'évolution des exercices historiques.

Le modèle individuel a l'avantage d'incorporer les données propres à chaque dossier et la disparité des profils assez importante à l'intérieur des exercices de survenances.

Les résultats sont proches au global, mais permettent d'être plus précis par exercice de survenance, d'analyser plus finement les paiements et d'apporter une vision de flux différente ce qui a un impact en Solvabilité 2 via l'actualisation. C'est dans ce cadre que l'exploitation des données individuelles peut apparaître comme une alternative complémentaire aux données agrégées du triangle de dépense. Néanmoins, ces résultats restent à confirmer sur d'autres garanties corporelles et sur un horizon plus long.

CONCLUSION

Ce mémoire propose une approche de provisionnement à partir de données individuelles et l'articule avec les méthodes de provisionnement sur données agrégées. L'idée générale développée dans ce mémoire est de considérer que sur les nouvelles survenances, il existe un grand nombre de sinistres qui font l'objet d'actes de gestion et l'utilisation d'une méthodologie agrégée prend alors tout son sens. Par contre, sur les exercices « anciens » où les sinistres non clos sont peu nombreux et les situations sont disparates entre exercices de survenance, l'estimation de l'évolution moyenne via le calcul de coefficients de développement peut paraitre moins justifiée. La prise en compte des données spécifiques au sinistre permet d'affiner la précision des estimations.

La méthode sur données individuelles est fondée sur la connaissance de la durée de vie du sinistre et de la chronique des paiements restante sur ce sinistre. Par construction, l'équation s'appuie sur l'hypothèse d'indépendance des durées de vie des sinistres et des paiements. Cette hypothèse peut être discutable étant donné que les gros sinistres nécessitent plus d'années pour être réglés.

Ceci étant, nous avons distingué les sinistres individuels en segments homogènes, basés sur le profil de durée, afin d'atténuer le biais lié à l'hypothèse d'indépendance et d'améliorer la qualité de l'estimation. L'utilisation du profil de durée comme critère explicatif des paiements introduit une dépendance entre la durée de vie et les paiements.

Le modèle de survie *Accelerated Failure Time* permet d'obtenir des courbes de survie différenciées, 12 profils de durée basés sur le nombre de victimes et le taux d'IPP. C'est en toute logique que nous observons que la durée de vie d'un sinistre augmente avec le nombre de victimes et le taux d'IPP. La nature paramétrique, loi Weibull, du modèle fournit une estimation de la probabilité de survie jusqu'à l'ultime.

La chronique de paiements pour chaque année ne peut être modélisée directement, elle est décomposée en probabilité et montant de paiements qui sont modélisés à l'aide des modèles linéaires généralisés. La provision dossier et, dans un second temps, le profil de durée influent fortement sur les paiements du sinistre. Ainsi, les provisions dossier élevées et le profil de durée long conduisent à des paiements élevés.

Les sinistres non clos au 31/12/2013 pour les exercices de survenance « anciens » sont projetés à l'aide du modèle individuel. Les paiements futurs incorporent ainsi l'expertise du gestionnaire, via la provision dossier, et l'information inhérente au sinistre c'est à dire le nombre de victimes et le taux d'IPP maximal.

Les résultats obtenus à l'aide des projections sur données détaillées sont proches de ceux correspondant au provisionnement *Chain Ladder*, sur l'ensemble des exercices de survenance nous observons un écart de 2,8%. Ceci étant, il est à préciser que, dans ce mémoire, l'approche *Chain Ladder* est utilisée dans sa forme la plus simple et présente certaines limites.

Ce résultat n'est pas négligeable, car la projection de la charge de sinistres automobile RCC est un exercice délicat, compte-tenu des caractéristiques de cette garantie. Plutôt que de choisir une seule méthode, l'actuaire aura intérêt à utiliser un faisceau de méthode et à comparer les résultats obtenus. Le meilleur des cas étant celui où le faisceau de méthodes converge.

Nous constatons que les écarts sont variables sur chaque exercice de survenance. Les exercices de survenance « anciens » diffèrent entre eux par le nombre et la nature des

sinistres qui la composent. Plus spécifiquement, la présence ou l'absence de sinistres importants tient un rôle primordial dans la projection des paiements futurs.

Les résultats obtenus lors de l'étape du *backtesting* à l'inventaire 31/12/2011 ont globalement conforté les résultats initiaux, tant au niveau du montant de la provision tous exercices confondus que par exercice de survenance. Le modèle individuel permet d'obtenir des résultats similaires à l'approche *Chain Ladder* tout en réduisant très légèrement l'écart des estimations avec le réel sur les différents exercices de survenance considérés individuellement.

La méthodologie développée permet, en plus des estimations fournies, une analyse de manière plus précise des flux que Chain Ladder. Nous pouvons expliquer dans le détail le déroulé d'un exercice de survenance comparé à un autre.

La méthodologie sur données détaillées présente l'inconvénient de la lourdeur puisqu'elle nécessite la modélisation de la durée de vie du sinistre et de la chronique des paiements. Cependant, l'étude détaillée des sinistres met en exergue des résultats intéressants et une approche innovante.

Dans ce cadre, l'exploitation des données individuelles peut apparaître comme une alternative complémentaire aux données agrégées du triangle de dépense qui servent de support aux techniques usuelles de projection de charge de sinistre. Notamment dans le cadre de Solvabilité 2, où le *Best Estimate* résulte d'une actualisation des paiements.

Cette méthode peut aussi apporter un éclairage à des problématiques de réassurance. Nous pouvons disposer d'une projection individuelle des sinistres importants, notamment sur l'estimation de chacun des sinistres qui servent à calibrer les programmes de réassurance.

La méthodologie sur données individuelles peut servir au suivi du sinistre et plus spécifiquement à la détection de sinistres atypique en gestion. Par exemple, un sinistre où aucun paiement n'est constaté alors que le modèle prédit des dépenses, pourrait être identifié comme atypique.

Les résultats obtenus pour l'estimation de la provision des sinistres doivent nous encourager à persévérer dans la recherche de voies d'exploitation des informations liées à la gestion des sinistres.

Nos axes d'avancement portent sur l'utilisation du modèle sur données détaillées uniquement sur les sinistres importants et « anciens » où la dynamique du sinistre est fortement dépendante des caractéristiques du sinistre. Une amélioration du modèle serait de développer une version stochastique du modèle pour son utilisation dans le cadre de la maitrise du risque de provisionnement ou d'un modèle interne.

Certains résultats, notamment ceux portant sur les modèles de survie, pourraient être appliqués sur les sinistres de masse à d'autres fins que le provisionnement : anticipation d'évolution des effectifs ou calcul de la provision pour frais de gestion.

Les travaux développés permettent d'enrichir l'ensemble des méthodes d'estimation de la provision des sinistres. L'étude des données détaillées a mis en évidence des caractéristiques propres à certains sinistres, qui se sont révélées prédictives de leur situation finale. L'utilisation d'un ensemble de méthodes doit permettre à l'actuaire d'apporter une contribution significative à la maîtrise du risque de provisionnement, élément clé de la directive solvabilité 2.

BIBLIOGRAPHIES

Planchet F. & Thérond P. (2011). Modélisation statistique des phénomènes de durée - Applications actuarielles. Economica.

Denuit M. & Charpentier A. (2004). Mathématiques de l'assurance non-vie, Tome II. Economica.

Dobson Annette. An introduction to generalized linear models. London : Chapman & Hall, 1990, 174.

Partrat C.; Lecoeur E., Nessi J.-M., Nisiparu E., Reiz O. (2007). Provisionnement technique en assurance non-vie. Economica.

Tosetti A., Behar T., Fromenteau M., Menart S. (2002) Assurance, Comptabilité, Réglementation, Actuariat – Economica.

Ross S. (2009). Introduction to Probability and statistics for engineers and scientists, Acamedic Press.

CISIA-CERESTA (1995). Aide mémoire statistique, CISIA-CERESTA Editeur.

SAINT PIERRE. P. (Février 2013). Introduction à l'analyse des durées de survie, Support de cours Université Pierre et Marie Curie.

Gilles Chau, Ngoc An Dinh (2012). Construction d'une méthode de provisionnement ligne à ligne pour des risques non-vie. Mémoire d'actuariat, ENSAE Paristech.

Bénéteau G. (2004). Modèle de provisionnement sur données détaillées en assurance non-vie. Mémoire d'actuariat, ENSAE Paristech.

Lacoume A (2008) Mesure du risque de réserve sur un horizon de un an, Mémoire d'actuariat, ISFA.

Oger F. (2010). Provisionnement à partir de données détaillées en assurance non vie. Mémoire d'actuariat, CEA.

Habib I., Riban S. (2012). Quelle méthode de provisionnement pour des engagements non-vie dans Solvabilité 2 ? Mémoire d'actuariat, ENSAE Paristech.

ANNEXES

A – Modèle de durée

A-1 Développement de la vraisemblance

La vraisemblance de l'échantillon (T_1, D_1), …, (T_n, D_n) s'écrit :

$$L(\theta) = \prod_{i=1}^{n} L_i = \kappa \prod_{i=1}^{n} f_\theta(t_i)^{di} S_\theta(t_i)^{1-di} = \kappa \prod_{i=1}^{n} \lambda_\theta(t_i)^{di} S_\theta(t_i)$$

$$L(\theta) = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} P(T_i \in [t_i, t_i + dt], D_i = 1/\theta)^{di} P(T_i \in [t_i, t_i + dt], D_i = 0/\theta)^{1-di}$$

$$\Leftrightarrow L(\theta) = \prod_{i=1}^{n} P(X_i \in [t_i, t_i + dt], C_i \geq X_i/\theta)^{di} P(C_i \in [t_i, t_i + dt], C_i < X_i/\theta)^{1-di}$$

$$\Leftrightarrow L(\theta) = \prod_{i=1}^{n} P(X_i \in [t_i, t_i + dt]/\theta)^{di} P(C_i \geq t_i)^{di} P(C_i \in [t_i, t_i + dt])^{1-di} P(t_i < X_i/\theta)^{1-di}$$

$$\Leftrightarrow L(\theta) = \prod_{i=1}^{n} f_X(t_i/\theta)^{di} S_C(t_i)^{di} f_C(t_i)^{1-di} S_X(t_i/\theta)^{1-di}$$

$$\Leftrightarrow L(\theta) = S_C(t_i)^{di} f_C(t_i)^{1-di} \prod_{i=1}^{n} f_X(t_i/\theta)^{di} S_X(t_i/\theta)^{1-di}$$

$$\Leftrightarrow L(\theta) = \kappa \prod_{i=1}^{n} f_\theta(t_i)^{di} S_\theta(t_i)^{1-di}$$

Le terme κ regroupe les informations en provenance de la loi de la censure qui ne dépend pas du paramètre θ. Il est à préciser que si le mécanisme de censure est indépendant de l'événement étudié alors la censure est dite non informative.

En remplaçant la densité par le produit de la fonction de hasard et de la fonction du survie $f_\theta(t) = \lambda_\theta(t) S_\theta(t)$, la vraisemblance peut être écrite de la manière suivante :

$$L(\theta) = \kappa \prod_{i=1}^{n} f_\theta(t_i)^{di} S_\theta(t_i)^{1-di} = \kappa \prod_{i=1}^{n} (\lambda_\theta(t_i) S_\theta(t_i))^{di} S_\theta(t_i)^{1-di}$$

$$L(\theta) = \kappa \prod_{i=1}^{n} \lambda_\theta(t_i)^{di} S_\theta(t_i)$$

A-2 Estimateur de Nelson-Aalen

L'approche de Nelson-Aalen s'appuie sur l'estimation du taux de risque cumulé. La fonction de hasard cumulé vérifie par construction :

$$\Lambda(t + dt) - \Lambda(t) = \lambda(u)du \text{ et } \lambda(u)du = P(\textit{sortie entre } t \textit{ et } t + dt \textit{ sachant en vie en } t)$$

Considérons les notations suivantes pour chaque individu :

- le processus d'évènements non censurés : $N_i^1(t) = 1_{\{T_i \le t, D_i = 1\}}$,
- le nombre d'évènements survenus non censurés : $\overline{N}^1(t) = \sum_{i=1}^{n} N_i^1(t)$,
- l'indicateur de présence à risque comptabilisant les individus ni morts ni censurés : $R_i(t) = 1_{\{T_i \ge t,\}}$,
- l'effectif sous risque : $\overline{R}(t) = \sum_{i=1}^{n} R_i(t)$.

Un estimateur naturel de cette quantité est donc :

$$\frac{\overline{N}^1(u+du) - \overline{N}^1(u+du)}{\overline{R}(u)} = \frac{d\overline{N}^1(u)}{\overline{R}(u)} \text{ si } \overline{R}(u) > 0$$

En sommant ces quantités sur des intervalles suffisamment fin de sorte que chacun ne contienne qu'un seul événement, nous obtenons l'estimateur de Nelson-Aalen :

$$\hat{\Lambda}(t) = \int_0^t \frac{d\overline{N}^1(u)}{\overline{R}(u)}$$

Comme les processus considérés sont purement à sauts l'expression peut être écrite de la manière suivante :

$$\hat{\Lambda}(t) = \sum_{\{i / Ti \le t\}} \frac{\Delta\overline{N}^1(T_i)}{\overline{R}(T_i)}$$

Nous posons :

- le nombre de décès en t : $d(t) = \Delta\overline{N}^1(t)$
- l'effectif sous risque juste avant t : $r(t) = \overline{R}(t)$

L'équation peut être réécrite de la manière suivante :

$$\hat{\Lambda}(t) = \sum_{\{i / T_i \le t\}} \frac{d(T_i)}{r(T_i)} = \sum_{T_i \le t} \frac{d_i}{n-i+1}$$ (la seconde égalité n'est vrai que si il n'y a pas d'ex aequo).

Il est à préciser que cet estimateur est biaisé et sous-estime en moyenne la fonction de hasard cumulée.

L'estimateur de Nelson-Aalen est une fonction en escalier qui a un saut de taille di/(n-i+1) à chaque instant de décès.

La variance de l'estimateur de Nelson-Aalen est :

$$\hat{V}(\hat{\Lambda}(t)) = \sum_{Ti \le t} \frac{d_i}{(n-i+1)^2}$$

L'estimation de la fonction de survie se fait à l'aide de l'estimateur de Harrington et Fleming. Cet estimateur s'appuie sur l'estimateur du risque cumulé de Nelson-Aalen et la relation $\Lambda(t) = -\log(S(t))$.

L'estimateur de Harrington et Fleming de la survie :

$$\hat{S}(t) = \exp(-\hat{\Lambda}(t))$$

La variance de cet estimateur peut être écrite de la manière suivante :

$$\hat{V}(\hat{S}(t)) = \hat{S}(t)^2 * \hat{V}(\hat{\Lambda}(t))$$

A-3 Présentation des lois paramétriques

La loi exponentielle

La loi exponentielle est un cas particulier de la loi de gamma où ν=1. Ainsi, la loi ne dépend que d'un unique paramètre et possède la spécificité d'avoir une fonction de hasard constante. La loi exponentiel $\xi(\theta)$ (ou $\Gamma(1, \theta)$) est caractérisée par la densité suivante :

$$f_\theta(t) = \theta e^{-\theta t}$$

Soit la fonction de survie et la fonction de risque instantané :

- la fonction de survie : $S_\theta(t) = e^{-\theta t}$,
- la fonction de hasard : $\lambda_\theta(t) = \theta$.

La loi weibull

La loi weibull W(ν, θ) dépend de deux paramètres strictement positif et est caractérisée par la densité suivante :

$$f_{\nu,\theta}(t) = \nu\theta(\theta t)^{\nu-1} e^{-(\theta t)^\nu}$$

Soit la fonction de survie et la fonction de risque instantané :

- la fonction de survie : $S_{\nu,\theta}(t) = e^{-(\theta t)^\nu}$,
- la fonction de hasard : $\lambda_{\nu,\theta}(t) = \nu\theta(\theta t)^{\nu-1}$.

La forme de la fonction de hasard dépend de ν, ainsi si ν<1 alors la fonction est décroissant et si ν>1 la fonction est croissant.

Remarque :
- si ν=1 alors la loi de weibull est une loi exponentielle.
- La loi de weibull est liée à la loi gamma : si la variable aléatoire T suit une loi de weibull W(ν, θ) alors T^ν suit une loi gamma $\Gamma(\nu, \theta)$.

La loi log-normale

La loi normale N(ν, θ) dépend de deux paramètres où θ est strictement positif et sa densité est de la forme suivante :

$$f_{\nu,\theta}(t) = \frac{1}{\theta t\sqrt{2\pi}} e^{\frac{-1}{2\theta^2}(\ln(t)-\nu)^2} = \frac{1}{\theta t}\phi(\frac{\ln(t)-\nu}{\theta}) \text{ avec } \varphi \text{ la fonction densité N(0,1)}$$

La fonction de survie et la fonction de risque instantané ont respectivement les formes suivantes :

- la fonction de survie : $S_{\nu,\theta}(t) = 1-\Phi\left(\frac{\ln(t)-\nu}{\theta}\right)$ avec ϕ la fonction de répartition d'une loi normale centrée réduite,
- la fonction de hasard : $\lambda_{\nu,\theta}(t) = \frac{\frac{1}{\theta t}\phi(\frac{\ln(t)-\nu}{\theta})}{1-\Phi\left(\frac{\ln(t)-\nu}{\theta}\right)}$.

La fonction de hasard est en forme de $\cap$ avec une courbure dépendante du paramétrage.

Remarque : la loi de log-normale est liée à la loi normale : si la variable aléatoire T suit une loi de log-normale LN(ν, θ) alors ln(T) suit une loi normale N(ν, θ).

La loi log-logistique

La distribution log-logistique L(ν, θ) dépend de deux paramètres strictement positif et est caractérisée par la densité suivante :

$$f_{\nu,\theta}(t) = \frac{\theta\nu t^{\nu-1}}{(1+\theta t^{\nu})^2}$$

La fonction de survie et la fonction de risque instantané ont respectivement les formes suivantes :

- la fonction de survie : $S_{\nu,\theta}(t) = \frac{1}{1+\theta t^{\nu}}$,
- la fonction de hasard : $\lambda_{\nu,\theta}(t) = \frac{\theta\nu t^{\nu-1}}{\left(1+\theta t^{\nu}\right)^3}$.

A-4 Test de comparaison d'échantillon

La comparaison de durées de vie issue de deux échantillons se fait à l'aide de test. Les tests les plus utilisés dans le cadre des données de censures est le test de rang et de Gehan.

L'objectif de cette approche est de tester l'hypothèse H_0 : l'égalité des fonctions de survie des deux échantillons contre H_1 : différence des fonctions de survie des deux échantillons.

Si il n'y avait pas de données censurées, nous pourrions utiliser les test de Kolmogorov-Smirnov.

Le test de rang a été adapté aux données censurées. Le principe est de trier l'échantillon en fonction des dates de décès $t_1 < t_2 < \ldots < t_n$. A chaque temps t_k, nous définissons d_{kj} le nombre de décès et r_{kj} l'effectif sous risque dans le groupe j. L'effectif sous risque est calculé avant t_k et le nombre de vivant après t_k est r_{kj}-d_{kj}.

	Groupe 1		Groupe 2		Total	
Durées	Décès en t_k	Survivants après t_k	Décès en t_k	Survivants après t_k	Décès en t_k	Survivants après t_k
t_k	d_{k1}	r_{k1}-d_{k1}	d_{k2}	r_{k2}-d_{k2}	d_k	r_k-d_k

Sous H_0, à chaque t_k les proportions de décès dans les deux groupes doivent être égales. La variable aléatoire d_{kj} est distribué selon une loi hypergéométrique $H(r_k, d_k, r_{kj}/r_k)$. Ainsi, l'espérance et la variance de d_{kj} sont :

$$E[d_{kj}] = d_k \frac{r_{kj}}{r_k} \quad \text{et} \quad V[d_{kj}] = d_k \frac{r_k - d_k}{r_k - 1} \frac{r_{k1} r_{k2}}{r_k^2}$$

La statistique de test requiert la notion de pondération w_k. Le coefficient de pondération prend la valeur $w_k=1$ dans le cas du test de log-rank et $w_k=r_k$ dans le cas du test de Gehan. Il est à préciser que dans le cas du test de Gehan la pondération appliquée ($w_k=r_k$) conduit à pondérer plus fortement les décès les plus précoces.

La statistique de test ϕ_j suit asymptotiquement une loi Khi-deux de degré 1 et est calculée de la manière suivante :

$$\phi_j = \frac{\left[\sum_{k=1}^{n} w_k \left(d_{kj} - d_k \frac{r_{kj}}{r_k} \right)\right]^2}{\sum_{k=1}^{n} w_k^2 d_k \frac{r_k - d_k}{r_k - 1} \frac{r_{k1} r_{k2}}{r_k^2}} = \frac{\left[\sum_{k=1}^{n} w_k \left(d_{kj} - d_k \frac{r_{kj}}{r_k} \right)\right]^2}{\sigma^2} \text{ avec } \sigma^2 = \sum_{k=1}^{n} w_k^2 d_k \frac{r_k - d_k}{r_k - 1} \frac{r_{k1} r_{k2}}{r_k^2}$$

Les tests de rang et le test de Gehan se généralisent à la comparaison des fonctions de survie de plusieurs groupes. L'hypothèse testée est : H_0 : les probabilités de survie entre les j groupes sont identiques contre H_1 : les probabilités de survie entre les j groupes sont différentes. Pour cela il est nécessaire de calculer la matrice de variance-covariance V de seulement j−1 termes arbitrairement pris parmi les j statistiques et un vecteur de j-1 statistiques.

$$V = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \ldots & \sigma_{1,j-1} \\ \sigma_{2,1} & \sigma_1^2 & \ldots & \sigma_{2,j-1} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{j-1,1} & \sigma_{j-1,2} & \ldots & \sigma_{j-1}^2 \end{pmatrix} \quad s = \begin{pmatrix} \left[\sum_{k=1}^{n} w_k \left(d_{k1} - d_k \frac{r_{k1}}{r_k} \right)\right]^2 \\ \ldots \\ \left[\sum_{k=1}^{n} w_k \left(d_{kj-1} - d_k \frac{r_{kj-1}}{r_k} \right)\right]^2 \end{pmatrix}$$

La statistique de test suit asymptotiquement une loi de χ^2 à j-1 degrés de liberté et est obtenue par : $\phi = s' V^{-1} s$.

B – Modèle Paiement

B-1 Méthode d'estimation des paramètres

La méthode de Newton-Raphson

Si la méthode de Newton-Raphson est utilisée alors à la mième itération est donnée par :

$$b^{(m)} = b^{(m-1)} - \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]^{-1}_{\beta=b^{(m-1)}} U^{(m-1)}$$

où b est un vecteur de taille p, $\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]_{\beta=b^{(m-1)}}$ est la matrice des dérivées secondes de taille p*p évalué pour β=b$^{(m-1)}$ et $U^{(m-1)} = \begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix} = \begin{pmatrix} \partial l / \partial \beta_1 \\ \vdots \\ \partial l / \partial \beta_p \end{pmatrix}$ est le vecteur des dérivées premières évalué pour β=b$^{(m-1)}$. L'étape 0 consiste à initialiser les paramètres β à n'importe quelle valeur.

La méthode de scoring de Fisher

Une méthode alternative qui est souvent plus simple que la méthode de Newton-Raphson est la méthode dite du scoring de Fisher. Elle consiste à remplacer la matrice des dérivées secondes dans la méthode de Newton-Raphson par son espérance au lieu de $\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]_{\beta=b^{(m-1)}}$ on a $E\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]_{\beta=b^{(m-1)}} = E(U_j U_k) = J_{jk}$. On appelle J=E(UU^T) la matrice information. On obtient l'équation suivante à la m^{ème} étape :

$$b^{(m)} = b^{(m-1)} + \left[J^{(m-1)} \right]^{-1} U^{(m-1)}$$

A l'étape 0 on prend b$^{(0)}$ solution de $(X^T X) b^{(0)} = X^T y$.

B-2 Méthode d'estimation des paramètres

Dans le cas des modèles linéaires classiques où la variable réponse est distribuée suivant une loi normale, la distribution d'échantillonnage peut être déterminée exactement. Cependant dans le cas où la variable ne suit pas une loi Normale, nous nous appuierons sur des résultats asymptotiques afin de trouver les distributions d'échantillonnage, c'est à dire que pour un effectif suffisamment grand, une loi donnée peut être assimilée à une loi Normale.

Soit $\hat{\theta}$ un estimateur du paramètre θ et $Var(\hat{\theta})$ est la variance de l'estimateur. Pour des échantillons de grandes tailles on a les résultats suivants :

- $\hat{\theta}$ est un estimateur non biaisé de θ.

- $\frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}} \sim N(0,1)$

Nous nous intéressons à la distribution d'échantillonnage pour les estimations du maximum de vraisemblance. On suppose que la fonction de log-vraisemblance a un maximum unique en b et que cet estimateur b est proche de la vraie valeur du paramètre β. Pour les modèles linéaires généralisés il existe deux tests : le test de Wald et le test de vraisemblance.

Test de Wald

Pour des échantillons de grandes tailles on obtient :

$$(b-\beta)^T J(b-\beta) \overset{n\to+\infty}{\sim} \chi_p^2 \quad \text{et} \quad (b-\beta) \overset{n\to+\infty}{\sim} N(0, J^{-1})$$

avec J=E(UU^T) la matrice d'information, U le vecteur des dérivées premières et b et β des vecteurs de taille p.

Test de vraisemblance

Cette statistique correspond à deux fois la différence entre le maximum de log-vraisemblance sans contrainte et le maximum de log-vraisemblance avec contrainte. On a :

$$2[l(b;y)-l(\beta;y)] \sim \chi_p^2$$

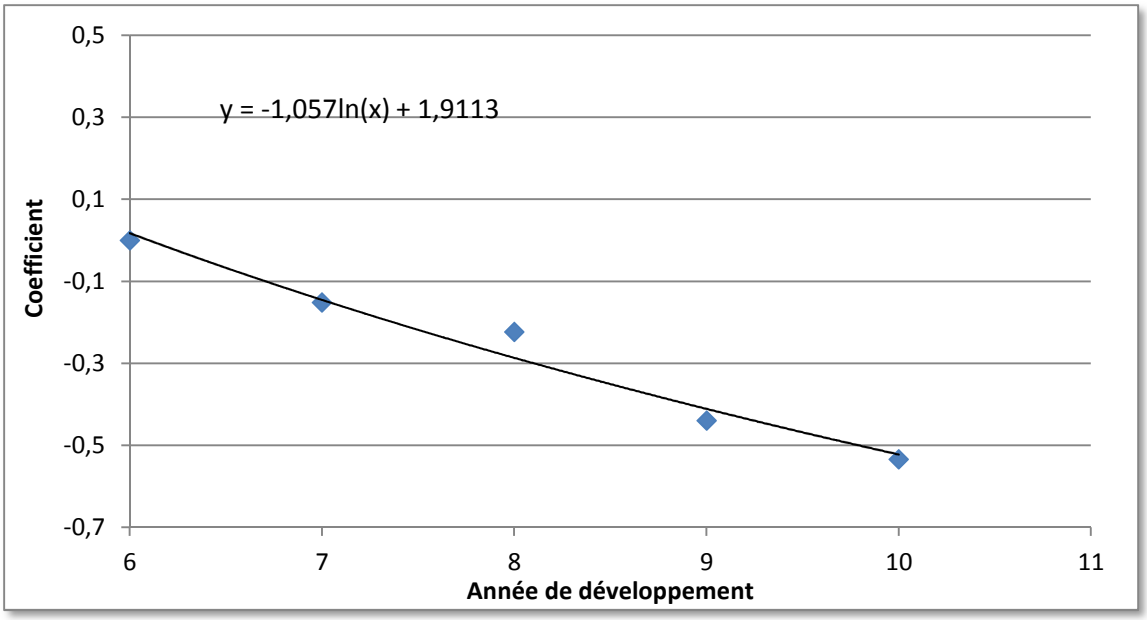
Remarque :Si on teste la nullité de b on aura pour contrainte β=0.

Le test de Wald et le test de vraisemblance ont une distribution du χ^2 asymptotique. On considère ainsi que l'estimateur du maximum de vraisemblance b suit une loi Normale(β,J^{-1}). Pour un seuil α=5% on a l'intervalle de confiance $b_j \pm 1{,}96\sqrt{v_{jj}}$ où vjj est le $j^{ième}$ élément de la matrice diagonale J^{-1}.

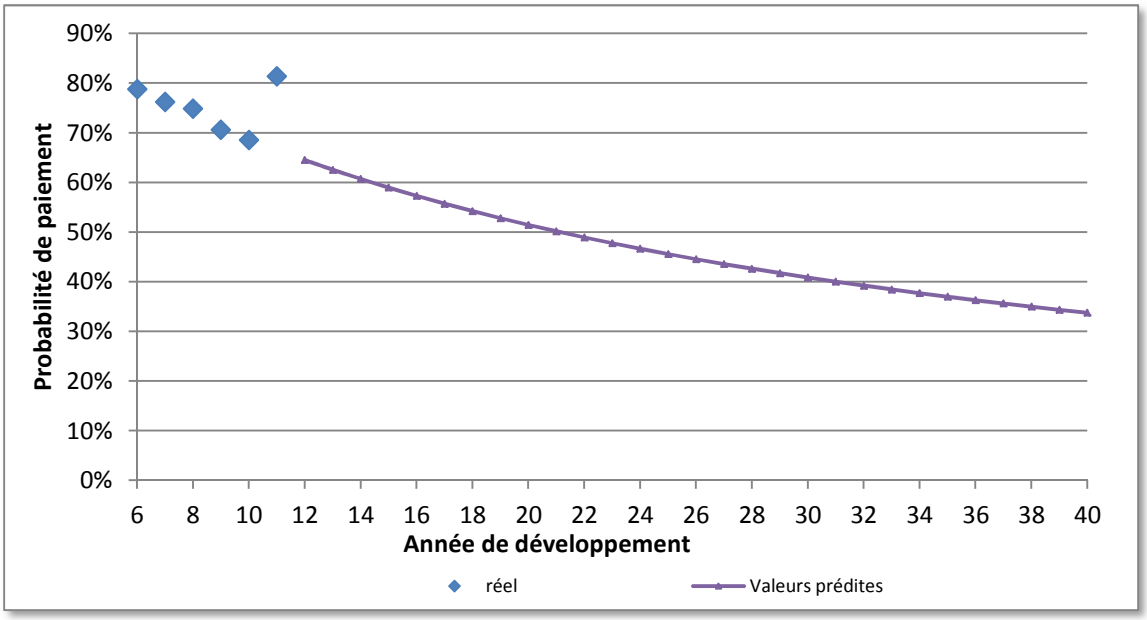
B-3 Extrapolation des probabilités de paiements

Régression linéaire de la forme y=a*ln(x)+b

Il est à souligner que la régression linéaire présente certaines limites du fait du faible nombre de points utilisés (5 années).



Exemple des estimations de probabilité de paiement pour la typologie de sinistre {Profil 1 ; Provision 6,5 M€<}.



B-4 Modèle de montant des paiements

Les tests basés sur la vraisemblance du modèle du montant des paiements avec les 3 variables sont présentés ci-dessous :

	ddl	Value	Value/ddl
Scaled Deviance	2 343	2 365	1,0115
Scaled Pearson	2 343	2 365	1,0115

Test Type 3	ddl	Khi-square	Pr>Khi-sq
Année développement	5	9,7	0,0843
Profil	11	47,5	<0,0001
Provision	10	361,0	<0,0001

La variable année de développement n'améliore pas le modèle et l'hypothèse de nullité de la variable année de développement n'est pas rejetée.