

Une méthode alternative de provisionnement stochastique en Assurance Non Vie : Les Modèles Additifs Généralisés

Lheureux Elise
B&W Deloitte
185, av. Charles de Gaulle
92524 Neuilly-sur-Seine cedex
France

Direct: +33(0)1.55.61.65.32

Fax: +33 (0)1.55.61.61.31

Elheureux@bw-deloitte.com

Résumé

Cet article traite des modèles de provisionnement stochastique à travers les Modèles Additifs Généralisés. Leur méthodologie repose sur le lissage non paramétrique qui offre une grande flexibilité dans l'ajustement du modèle du fait de l'indépendance du prédicteur avec une quelconque forme rigide paramétrique, permettant ainsi au modèle de suivre la tendance présente dans les données. Cette structure est également adaptée au calcul de mesures de volatilité des estimations de réserves et d'indicateurs de valeurs centrales de la loi des réserves (quantiles,...).

Mots clés

Provisionnement stochastique, Modèles Additifs Généralisés, Modèles Linéaires Généralisés, Chain Ladder, Mack, Bootstrap, Méthode Delta, quantiles.

SOMMAIRE

I. INTRODUCTION	2
II. METHODOLOGIE DES MODELES ADDITIFS GENERALISES.....	3
III. DISCUSSION.....	8
IV. RESULTATS.....	9
V. CONCLUSIONS	21
BIBLIOGRAPHIE.....	22
ANNEXES.....	23

I. INTRODUCTION

Les nouvelles normes IFRS induisent de profonds changements en ce qui concerne les provisions techniques et en particulier la provision pour sinistre à payer. Un des principes sous-jacents à ces normes est l'évaluation au plus juste du montant des réserves. Dans les pays anglo-saxons, l'estimation en « best estimate » est désormais une pratique courante, alors que la plupart des assureurs non-vie d'Europe continentale n'ont pas de règles claires quant au niveau de provisions requis. De ce fait, certains assureurs ont pris l'habitude d'inclure une marge dans leurs réserves, dont le but recherché peut être par exemple de lisser les résultats dans le temps. Le passage en « best estimate » et la nécessité de prendre en compte le risque lié à l'incertitude du montant de réserves limiteront ce genre de pratique et obligeront la compagnie d'assurances à justifier les hypothèses sous-jacentes au niveau de son provisionnement d'une année sur l'autre.

Ainsi, le travail de l'actuaire consistera à mettre œuvre des modèles de provisionnement stochastique, lesquels deviennent incontournables pour répondre à ces problématiques.

L'objectif de cet article est de présenter une nouvelle classe de modèles pour le provisionnement stochastique en Assurance Non Vie, qui nous semble plus souple que les méthodes existantes. Il s'agit de la classe des Modèles Additifs Généralisés (Generalized Additive Models ou GAMs). Ces modèles permettent d'appliquer aux données issues de la famille exponentielle des techniques de lissage non paramétrique. Ils préservent la structure sous-jacente des Modèles Linéaires Généralisés^[9] (Generalized Linear Models ou GLMs) car la relation entre la variable réponse et les variables explicatives est modélisée. Mais la différence majeure entre ces deux modèles réside dans le fait que cette relation est paramétrique dans les GLMs alors que la réponse est liée aux variables explicatives par une procédure de lissage non paramétrique dans le cas des GAMs.

Ce type de modèle s'adapte formellement à tout triangle de liquidation mais l'estimation des réserves est d'autant plus précise que le développement des sinistres de la branche étudiée est stable. Ainsi, les études de modélisation GAMs ont jusqu'à présent été appliquées à des triangles « classiques » servant de référence pour tester de nouvelles méthodes.

Nous avons cependant choisi d'illustrer nos analyses à partir de triangles de liquidation « réels », chacun d'entre eux correspondant à une branche différente. Parmi les trois branches étudiées, l'une est à déroulement relativement court et correspond à la catégorie « Dommages aux biens Automobile ». Les deux autres branches sont à déroulement plus ou moins long et correspondent aux catégories « Responsabilité Civile Automobile » et « Responsabilité Civile Générale ». L'étude de ces différentes branches nous a ainsi permis d'illustrer la mise en

œuvre des modèles GAMs en fonction de la longueur de la cadence de développement des sinistres. Nous avons utilisé le logiciel R muni du package « mgcv » (« Minimizing Generalized Cross Validation ») qui nous fournit les méthodes et sorties informatiques nécessaires à cette étude.

Le principal avantage des modèles de provisionnement stochastique est de disposer de mesures de précision des estimations des réserves effectuées : dans ce but, nous avons calculé le « Root Mean Square Error of Prediction » (RMSEP). Nous avons également cherché à estimer la distribution des réserves à l'aide de la méthode du Bootstrap afin d'obtenir ses différents quantiles et son écart-type. Enfin, nous avons utilisé certaines méthodes alternatives au Bootstrap, basées sur les premiers moments et l'approximation de distributions ; celles-ci permettent d'estimer les quantiles du montant de réserves sans nécessiter une estimation de la distribution de ces réserves.

Cet article présente également une comparaison des résultats avec ceux obtenus par d'autres modèles déterministes et stochastiques, déjà utilisés en pratique. L'analyse de la qualité de l'ajustement de ces différents modèles (mesure analytique de variabilité, étude des résidus,...) et de leur « best estimate » nous conduit au fil de l'étude à considérer les modèles GAMs comme des modèles robustes et satisfaisants pour répondre au concept de « best estimate » et au besoin d'une mesure du risque associé.

II. METHODOLOGIE DES MODELES ADDITIFS GENERALISES

Nous proposons dans cet article d'utiliser les méthodes de lissage non paramétrique à travers les modèles GAMs^[5] comme une alternative aux méthodes de provisionnement déjà existantes.

Nous considérons sans perte de généralité un triangle de données avec n années d'origine et n années de développement. Nous désignons les montants incrémentaux relatifs à l'année d'origine i et à l'année de développement j par Y_{ij} . Les données sont alors :

$$\{Y_{ij} : i = 1, \dots, n ; j = 1, \dots, n - i + 1\}$$

Nous désignons dans la suite de cette étude les variables année d'origine i et année de développement j par $Year_i$ et Dev_j .

La composante aléatoire

Les modèles GAMs reposent sur l'hypothèse forte suivante :

Ho : Les variables aléatoires Y_{ij} sont indépendantes mais **non identiquement distribuées**.

Ces modèles supposent également l'appartenance des variables Y_{ij} à la famille exponentielle à deux paramètres dont la densité est :

$$f(y_{ij}, \theta_{ij}, \phi) = \exp\left(\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right) \quad (\text{II.1})$$

où

- θ_{ij} est un paramètre réel dit « naturel »,
- $\Phi > 0$ est un paramètre de dispersion ou d'échelle (inconnu), indépendant de i et j ,
- les fonctions b et c sont des fonctions spécifiques de la distribution exponentielle.

Sous cette hypothèse de distribution, l'espérance et la variance des Y_{ij} peuvent s'écrire :

$$E(Y_{ij}) = b'(\theta_{ij}) = \mu_{ij} \text{ et } Var(Y_{ij}) = \phi b''(\theta_{ij}) = \phi V(\mu_{ij}) \quad (\text{II.2})$$

La fonction $V(\cdot)$ est appelée la fonction de variance de la distribution et Φ le paramètre de dispersion. Ensemble, ils déterminent de manière unique la distribution.

Les différentes distributions et fonctions de lien que nous allons tester sur nos données sont résumées dans le tableau ci-dessous. Nous considérons les fonctions de lien standards Identité, Canonique, Logarithme, et Racine Carré.

Tableau II.1
Distributions et fonctions de lien testées.

Distribution	Densité de Y	Domaine des paramètres	Fonction de lien g	Fonction de variance V(.)	Paramètre de dispersion Φ inconnu et estimé
Normale	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}, \sigma^2 > 0$	$g(\mu) = \mu$	$V(\mu) = 1$	$\Phi = \sigma^2$
Poisson Surdispersé	$\frac{e^{-\mu} \mu^y}{y!}$	$\mu > 0$	$g(\mu) = \mu,$ $g(\mu) = \ln(\mu),$ $g(\mu) = \sqrt{\mu}$	$V(\mu) = \mu$	Φ
Gamma	$\frac{1}{\Gamma(r)} \left(\frac{r}{\mu}\right)^r y^{r-1} e^{-\frac{ry}{\mu}}$	$\mu, r > 0$	$g(\mu) = \mu$ $g(\mu) = \ln(\mu),$ $g(\mu) = 1/\mu$	$V(\mu) = \mu^2$	$\Phi = 1/r$
Inverse Gaussienne	$\frac{1}{\sqrt{2\pi ry^3}} \exp\left\{-\frac{1}{2r\mu^2 y} (y-\mu)^2\right\}$	$\mu, r > 0$	$g(\mu) = 1/\mu^2,$ $g(\mu) = \mu,$ $g(\mu) = \ln(\mu)$	$V(\mu) = \mu^3$	$\Phi = r$
LogNormale	$\frac{1}{\sqrt{2\pi\sigma y}} \exp\left(-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}, \sigma^2 > 0$	$g(\mu) = \mu$	$V(\mu)$	Φ

Choix des variables explicatives du modèle

Les variables intervenant dans la modélisation d'un triangle de liquidation correspondent à ses trois dimensions « naturelles » : en ligne avec les années d'origine des sinistres, en colonne avec les années de développement des paiements, et en diagonale avec les années calendaires. Cependant, nous ne prenons pas en compte dans notre modélisation cette dernière variable année calendaire qui représente l'inflation car nous travaillons sur des triangles déflatés dans le contexte de cette étude.

La mise en œuvre d'une modélisation GAM requiert la continuité de ces variables du fait de l'utilisation de fonctions de lissage. Or la variable année de développement est une variable quantitative discrète. Nous la rendons continue en définissant une variable artificielle $Y=j$ qui sera traitée comme une variable continue. La variable année d'origine est quant à elle une variable qualitative ordinaire et est ainsi traitée comme un indice dans les modélisations classiques telles que Chain Ladder ou GLM. Mais si cette variable était vraiment un pur indice, alors les lignes pourraient être inversées sans aucun effet sur le modèle. Or ceci n'est pas supporté par la théorie sous-jacente du provisionnement et il est ainsi plus approprié de caractériser la variable année d'origine comme un *indice ordonné*. De plus, les années d'origines sont régulièrement espacées et peuvent être considérées comme une série « chronologique », suggérant ainsi de traiter cette variable, non pas comme un indice ordonné, mais comme une variable continue en définissant une variable artificielle $X=i$ qui sera elle aussi traitée comme une variable continue. Cela permet au modèle d'appliquer également une fonction de lissage à cette variable.

La composante systématique, la fonction de lien

Un modèle GAM peut être vu comme ayant la structure d'un modèle factoriel à 2 facteurs (année d'origine i et année de développement j), dont la variable réponse Y_{ij} appartient à la famille exponentielle à deux paramètres (cf. équation (II.1)). La présence supplémentaire d'un lissage des deux variables explicatives $Year_i$ et Dev_j différencie la modélisation GAM d'un modèle d'analyse de la variance classique.

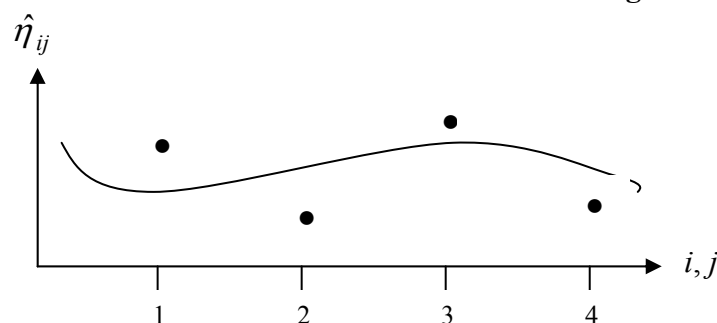
La fonction de lien notée g , strictement monotone et dérivable, relie $\mu_{ij} = E(Y_{ij})$ au prédicteur ou composante systématique η_{ij} qui fait intervenir uniquement les variables explicatives :

$$\begin{aligned} \eta_{ij} &= g(\mu_{ij}) = c + s_\lambda(Year_i) + s_\nu(Dev_j) \\ \Rightarrow \hat{\mu}_{ij} &= g^{-1}[\hat{c} + \hat{s}_\lambda(Year_i) + \hat{s}_\nu(Dev_j)] \end{aligned} \quad (II.3)$$

La fonction $s_\lambda(Year_i)$ représente un lissage de la variable explicative année d'origine i avec le paramètre de lissage λ . De façon similaire, la fonction $s_\nu(Dev_j)$ représente le lissage de la structure de développement du triangle avec le paramètre de lissage ν .

En considérant les deux variables année d'origine $Year_i$ et année de développement Dev_j comme des variables « continues », la structure du modèle GAM cherche à lisser de telle façon que les valeurs de $\hat{\eta}_{ij}$ entre i et $i+1$ et entre j et $j+1$, ne diffèrent pas trop. C'est ce qu'illustre le graphique II.1.

Graphique II.1
Illustration d'une courbe de lissage



Paramètres de la modélisation

La fonction de lissage

Un lisseur est un outil qui permet de résumer la tendance de la variable réponse Y en une fonction de une ou plusieurs variables prédictives. Cela produit une estimation de cette tendance qui est plus stable que la variable Y elle-même. Une propriété importante du lisseur est sa nature non paramétrique : il ne suppose pas une forme rigide pour la dépendance de Y en fonction des variables explicatives.

Dans cet article, nous nous concentrons uniquement sur la procédure de lissage « cubic smoothing spline » car elle seule est implémentée dans le logiciel R avec le package « mgcv ». Un spline de lissage cubic ou « cubic smoothing spline » est la solution du problème d'optimisation suivant (appelé aussi « Penalized Residual Sum Of Square ») :

Parmi toutes les fonctions $g(x)$ et $h(x)$ deux fois continuellement différentiables, trouver celles qui minimisent la somme des carrés pénalisés :

$$\underbrace{\sum_{i=1}^n \sum_{j=1}^n (y_{ij} - c - g(\text{Year}_i) - h(\text{Dev}_j))^2}_{\text{Mesure l'écart avec les données}} + \underbrace{\lambda \int_1^n (g''(t))^2 + \nu \int_1^n (h''(t))^2}_{\text{Pénalise la courbure de la fonction (Pénalité de lissage)}} \quad (\text{II.4})$$

où $1 \leq \text{Year}_i; \text{Dev}_1 \leq \dots \leq \text{Year}_n; \text{Dev}_n \leq n$
 λ, ν , constantes fixées sont les paramètres de lissage.

On peut montrer qu'il existe une solution explicite et unique à ce problème de minimisation et que le minimum est un spline cubic naturel s_λ et s_ν avec des nœuds uniquement aux valeurs Year_i et Dev_j .

Ainsi les GAMs produisent les estimations $\{\hat{s}_\lambda(\text{Year}_1), \hat{s}_\lambda(\text{Year}_2), \dots, \hat{s}_\lambda(\text{Year}_n)\}$ et $\{\hat{s}_\nu(\text{Dev}_1), \hat{s}_\nu(\text{Dev}_2), \dots, \hat{s}_\nu(\text{Dev}_n)\}$ qui correspondent respectivement à l'effet de lissage sur les années d'origine et à l'effet de lissage sur les années de développement.

Le choix du paramètre de lissage

Les deux fonctions de lissage \hat{s}_λ et \hat{s}_ν n'ont qu'un seul paramètre de lissage chacune. Chaque paramètre gouverne le niveau de lissage ainsi que le compromis entre le biais et la variance :

- Quand λ, ν tendent vers 0, il n'y a pas de pénalité de lissage et le modèle fournit un ajustement parfait : les valeurs ajustées sont les données elles-mêmes.
- Quand λ, ν sont grands (tendent vers l'infini), l'ajustement est un lissage parfait : les valeurs ajustées tombent le long d'une ligne droite forçant ainsi la liaison à être linéaire. Il en résulte une faible variance mais un biais important.

Les paramètres λ et ν doivent être compris entre ces deux cas extrêmes pour produire le niveau désiré de lissage qui permet d'avoir un bon ajustement du modèle avec un équilibre biais - variance. Le choix de ces paramètres de lissage peut se faire à l'aide de la méthode par validation croisée générale (GCV). Mais une valeur particulière pour λ et ν peut également être choisie en fixant soi-même le nombre de degrés de liberté.

- ✓ La méthode de validation croisée générale (GCV)

Wahba^[15] a proposé la fonction de validation croisée générale (GCV) pour résoudre les problèmes qui nécessitent la sélection de paramètres minimisant le risque de prédiction. On peut montrer que $\hat{\eta} = \hat{c} + A(\lambda, \nu)Y$, où $A(\lambda, \nu)$ est une matrice qui dépend de λ , ν et de la série de données ; c'est l'analogie avec une matrice de projection dans une régression linéaire. Soient a_{kk} les éléments diagonaux de $A(\lambda, \nu)$.

La fonction GCV à minimiser est alors :

$$GCV(\lambda, \nu) = \frac{\sum_{i=1}^n \sum_{j=1}^n (y_{ij} - \hat{\eta}_{ij})^2}{(n - \text{tr}(A(\lambda, \nu)))^2} \quad (\text{II.5})$$

où $\text{tr}(A(\lambda, \nu))$ est la trace de la matrice $A(\lambda, \nu)$, soit la somme des éléments diagonaux de cette matrice.

Malheureusement, la méthode GCV est très coûteuse pour les problèmes à paramètres de lissage multiples. Simon Wood^[16] a développé une approche améliorée pour résoudre ce problème dans le cas d'une classe de modèles de spline de lissage incluant les modèles GAMs avec le problème d'optimisation (II.4).

Le logiciel R utilise cette fonction GCV « améliorée » comme critère de sélection des paramètres de lissage λ et ν .

- ✓ Le nombre de Degrés de Liberté (DF)

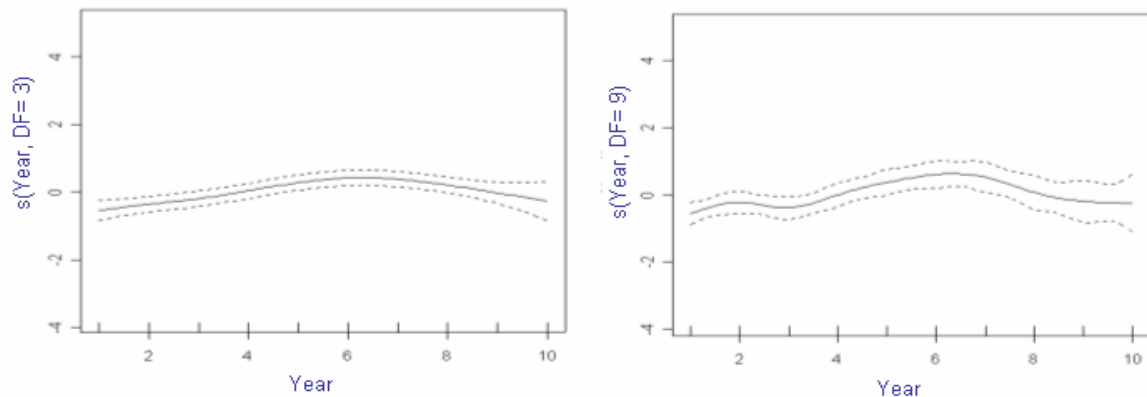
Le logiciel R permet aussi de choisir le niveau de lissage en fixant le nombre de degrés de liberté pour chaque lisseur s_λ et s_γ .

Le nombre de degrés de liberté non paramétriques (DF) pour le lisseur s_λ (respectivement pour s_γ) peut être défini par :

$$DF(\text{lisseur } s_\lambda) = \text{tr}(A(y, \lambda)) \quad (\text{II.6})$$

Si une valeur particulière est choisie pour le nombre de degrés de liberté, alors pendant toute la phase du problème d'optimisation (II.4), le paramètre λ (et respectivement ν) satisfera l'équation (II.6).

Graphique II.2
Effet du lissage sur les années d'origine pour un DF de 3 et de 9



Ainsi, comme l'illustre le graphique II.2, plus le nombre de degrés de liberté DF de s_λ augmente, plus λ diminue et moins la courbe est lissée (et inversement).

Par la suite, nous utiliserons la méthode GCV « améliorée » pour le choix du paramètre optimal et cela pour chacune des deux variables explicatives $Year_i$ et Dev_j . Nous sommes à la limite inférieure du nombre de données nécessaire ($n = 55$) pour utiliser ce critère de sélection mais ce dernier a l'avantage considérable d'être automatisé et de permettre ainsi un gain de temps non négligeable lors du processus de sélection des modèles.

III.DISCUSSION

Présentation des données

L'approche des modèles additifs généralisés est sujette au risque d'erreur de spécification du modèle, comme tout modèle stochastique. Plusieurs exigences concernant les données sont requises afin d'obtenir un modèle correctement ajusté :

- disposer d'une part d'un nombre suffisant de données (une cinquantaine de données au minimum),
- vérifier d'autre part qu'il n'y a pas de données aberrantes.

Nous disposons pour chacune des trois branches « Dommages aux biens Automobile », « Responsabilité Civile Automobile » et « Responsabilité Civile Générale », d'un jeu de 55 données correspondant à dix années de survenance de sinistres avec un développement des paiements sur un maximum de dix ans. Cependant, avant toute modélisation, nous avons supprimé dans nos données la dernière valeur de paiement connue de la première année de survenance ($Year_i=1$ et $Dev_j=10$) pour les branches « Responsabilité Civile Automobile » et « Responsabilité Civile Générale ». En effet, cette valeur était aberrante pour ces deux branches, et aurait perturbé les résultats de toute modélisation.

Pour des questions pratiques, nos trois triangles de liquidation ont été modélisés bruts de recours; la théorie sous-jacente aux modèles GAMs indique qu'il n'est pas nécessaire de

faire l'hypothèse que tous les montants incrémentaux soient positifs à condition que la somme de chaque colonne du triangle soit positive. Cependant, l'usage de certaines fonctions de lien (logarithme, racine carré,..) dans la modélisation GAM ne permet pas la présence de montants négatifs, à moins d'ignorer les cellules concernées, ou bien d'utiliser la Quasi Vraisemblance. Il nous a donc semblé plus adapté de faire cette étude sur des triangles entièrement positifs, en ne tenant pas compte des recours encaissés.

Avantages d'une structure de lissage

Un des gros avantages de l'utilisation de ces lisseurs réside dans l'absence d'une structure prédéterminée dans le développement du triangle, permettant ainsi au modèle de suivre la tendance présente dans les données, d'où une grande flexibilité dans l'ajustement du modèle.

En outre, la méthodologie nous semble intéressante car elle permet de lisser les facteurs de développement de la méthode Chain Ladder et d'obtenir directement une estimation du facteur de queue, sans nécessiter une étape supplémentaire dans le processus de provisionnement. Cet avantage non négligeable des modèles GAMs est dû à la présence de variables explicatives continues dans la modélisation offrant ainsi la possibilité d'extrapoler au-delà des rangées de données observées. C'est une alternative aux pratiques habituelles qui consistent à ajuster un modèle standard aux montants cumulés du type Chain Ladder puis d'ajuster ensuite un autre modèle aux facteurs de développement qui en découlent. L'extrapolation de la courbe de lissage est également beaucoup plus intuitive et consistante que cette dernière technique.

Notons également que les modèles GAMs, grâce à leur structure, contrôlent leur niveau d'ajustement aux données en ayant un nombre de degrés de liberté raisonnable contrairement aux modèles GLMs qui ont tendance à être sur-paramétrés par rapport à la taille des données^[14].

Enfin, il est important de signaler que les mêmes méthodes d'estimation d'une distribution du montant de réserve (Bootstrap) et de calculs de quantiles (Bootstrap, Approximations Normal Power, Gamma,..) utilisées sur les modèles GLMs s'appliquent également aux modèles GAMs.

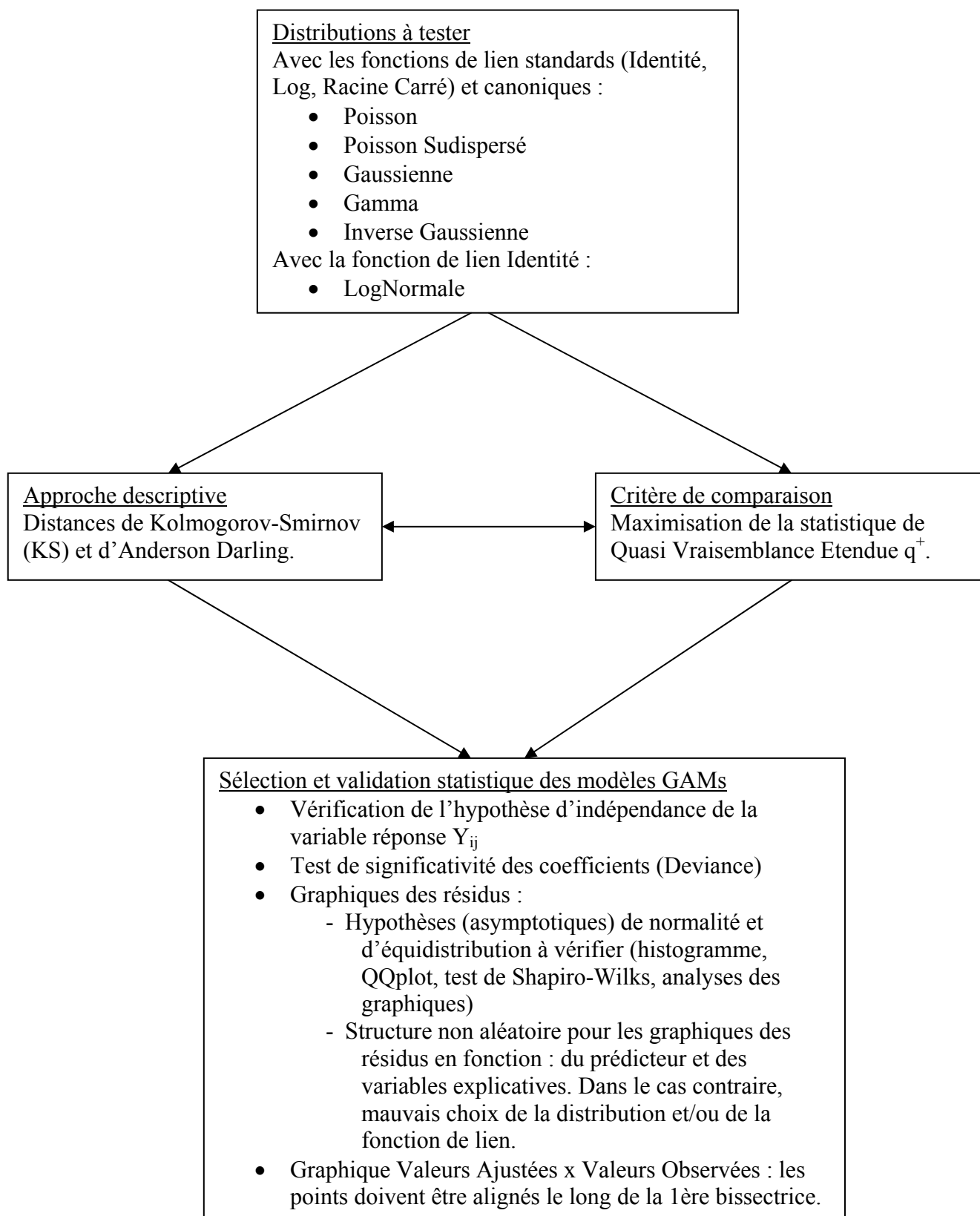
IV. RESULTATS

Processus de sélection des GAMs

Le diagramme IV.1 illustre le processus suivi au cours de notre étude pour sélectionner et valider les modèles GAMs les mieux adaptés à chacune des branches étudiées parmi la sélection décrite dans le tableau II.1. Le tableau IV.2 présente les 2 ou 3 modèles finaux retenus pour chacune des branches étudiées. Ceux-ci seront ensuite comparés avec d'autres modèles déterministes et stochastiques utilisés en pratique. Cette étude nous permettra d'illustrer l'intérêt des Modèles Additifs Généralisés par rapport à d'autres méthodes connues et utilisées jusqu'à présent.

Diagramme IV.1

Schéma récapitulatif du processus d'évaluation et de sélection des modèles GAMs.



Nous apportons quelques remarques à prendre en compte lors de la mise en pratique de ces différents critères et tests :

- Le calcul des distances de Kolmogorov Smirnov (KS) et d'Anderson Darling consiste en une approche descriptive et non quantitative de l'adéquation du modèle aux données. En effet, il faut être prudent avec la théorie de ces tests statistiques qui repose sur l'hypothèse « forte » d'indépendance et d'équidistribution des données alors que les variables Y_{ij} ne sont pas équidistribuées. De plus, le nombre de données ($n = 55$) est trop faible et engendre une bande de confiance beaucoup trop grande nous forçant à accepter toutes les lois testées. Nous avons quand même effectué ces tests en ne conservant que les lois ayant une distance de KS et d'Anderson suffisamment petite, nous permettant ainsi d'avoir une première idée des distributions les mieux adaptées aux données.
- La loi LogNormale ne peut se comparer aux autres lois avec le critère de Quasi Vraisemblance Etendue détaillé en annexe, car cette statistique fait intervenir la déviance et ainsi intègre dans son calcul l'écart de la variable $\ln(Y_{ij})$ avec sa prédiction et non de la variable Y_{ij} avec l'exponentielle de cette prédiction. La vérification des hypothèses sur les résidus est alors une méthode alternative qui nous permet de valider ou non la cohérence du choix de cette distribution LogNormale. L'hypothèse de normalité des résidus doit être vérifiée non pas asymptotiquement mais exactement pour cette distribution.
- Il faut être prudent avec l'ensemble des tests asymptotiques effectués et leurs conclusions, car il faut garder à l'esprit que le jeu de données sur lequel nous travaillons est de taille $n=55$, ce qui est a priori insuffisant pour un cadre asymptotique.
- Il faut noter qu'une particularité des deux branches longues « Responsabilité Civile automobile » et « Responsabilité Civile Générale » est la nécessité d'extrapoler les résultats au-delà du triangle de données car la totalité des paiements n'est pas réglée au bout de 10 ans de développement. Or, certaines fonctions de lien telles que l'Identité ou la Racine Carrée ne sont pas adaptées à une extrapolation des résultats. Seule la fonction de lien logarithme est acceptable pour une modélisation GAM avec extrapolation, car son inverse g^{-1} est la fonction exponentielle. En effet, étant donné qu'à partir d'une certaine valeur de Dev_j (>10) qui diffère selon le modèle GAM, le prédicteur η_{ij} décroît négativement, l'équation (II.3) indique que pour rester cohérent avec la réalité, seule la fonction de lien logarithme permet à l'espérance des montants incrémentaux de paiements μ_{ij} de tendre vers 0 en restant positive.

A l'issue de ces tests, un à deux modèles GAMs ont été retenus selon la branche étudiée pour poursuivre cette étude, comme l'indique le tableau IV.2 ci-dessous. Nous nous intéressons à d'autres méthodes de provisionnement déterministes et stochastiques (Chain Ladder, Mack, GLM et LogNormale) et les appliquons aux mêmes trois branches d'activités étudiées pour les GAMs afin de comparer les résultats obtenus et d'analyser les avantages et particularités propres à chacune des méthodes selon des longueurs de cadence de développement des sinistres différentes. Le lecteur pourra retrouver la théorie détaillée de ce panel de modèles dans l'article d'England et Verral^[4]. Pour mener à bien cette étude comparative des GAMs avec les autres modèles, nous considérons un modèle de référence déterministe nommé « Chain Ladder Ajusté » qui est élaboré à l'aide d'un logiciel développé par la société (CROS) et qui fournit le « Best estimate » grâce à une combinaison adéquate de modèles Chain Ladder.

Tableau IV.2
Modèles GAMs sélectionnés par branche étudiée

Distribution \ Lien	Identité	Log
Poisson Surdispersé		○
Gamma		△ ○
LogNormale	x	

- x Dommages Auto
- △ RC Auto
- RC Générale

Calcul des réserves avant et après extrapolation

Après avoir ajusté chaque modèle sur le triangle de données, nous obtenons l'estimation des réserves en sommant les paiements estimés appropriés du triangle inférieur. Les montants de réserves estimés par l'ensemble des modèles étudiés sont présentés pour chaque branche dans les tableaux IV.4, IV.5 et IV.6. Nous considérons que les sinistres de la branche courte « Dommages Automobile » sont entièrement réglés au bout de 10 ans car la cadence de règlement est de 99.9%. Les deux autres branches plus longues « Responsabilité Civile Automobile » et « Responsabilité Civile Générale » ont une cadence respective de 94% et 98% au bout de 10 ans, nécessitant de prolonger les estimations de paiements pour le calcul de la charge ultime. Pour une simple comparaison des modèles entre eux et notamment avec celui de Chain Ladder, nous n'introduisons pas dès cette étape l'estimation supplémentaire liée à l'extrapolation des paiements au-delà du triangle de données.

Le calcul des réserves est effectué à la fin de la dernière année comptable connue (31/12/n) ce qui génère un montant de réserves nul la première année d'origine car tous les paiements sont connus à cette date (sous l'hypothèse de non extrapolation). Pour les deux branches longues (tableau IV.5 et IV.6), le montant de réserves de la deuxième année d'origine est également nul car le montant correspondant à ($Year=1, Dev=10$) ne rentre pas dans la modélisation (en rouge dans le tableau IV.3) ; ainsi aucun montant de paiement n'est estimé pour ($Year=i, Dev=10$), $i=2, \dots, n$, et n'intervient donc pas dans le calcul du montant de réserves correspondant s'il n'y a pas d'extrapolation. Ce point est illustré par le tableau IV.3 : les valeurs estimées par le modèle sont surlignées en vert et leur somme par année d'origine constitue le montant de réserve par année d'origine au 31/12/n qui repose sur l'hypothèse de non extrapolation, c'est-à-dire qu'il se base uniquement sur l'historique. La partie grisée correspond aux montants qui ne peuvent être estimés qu'en utilisant une méthode d'extrapolation. Cette technique fait l'objet de la section suivante et donnera lieu à un calcul de montant de réserve ultime.

Tableaux IV.3
Illustration du triangle de paiement des deux branches longues

Année d'origine	Délai de règlement									
	1	2	3	4	5	6	7	8	9	10			
1	Y ₁₁	Y ₁₂	Y ₁₃	Y ₁₄	Y ₁₅	Y ₁₆	Y ₁₇	Y ₁₈	Y ₁₉	Y _{1,10}			
2	Y ₂₁	Y ₂₂	Y ₂₃	Y ₂₄	Y ₂₅	Y ₂₆	Y ₂₇	Y ₂₈	Y ₂₉				
3	Y ₃₁	Y ₃₂	Y ₃₃	Y ₃₄	Y ₃₅	Y ₃₆	Y ₃₇	Y ₃₈	Ŷ _{3,9}				
4	Y ₄₁	Y ₄₂	Y ₄₃	Y ₄₄	Y ₄₅	Y ₄₆	Y ₄₇	Ŷ _{4,8}	Ŷ _{4,9}				
5	Y ₅₁	Y ₅₂	Y ₅₃	Y ₅₄	Y ₅₅	Y ₅₆	Ŷ _{5,7}	Ŷ _{5,8}	Ŷ _{5,9}				
6	Y ₆₁	Y ₆₂	Y ₆₃	Y ₆₄	Y ₆₅	Ŷ _{6,6}	Ŷ _{6,7}	Ŷ _{6,8}	Ŷ _{6,9}				
7	Y ₇₁	Y ₇₂	Y ₇₃	Y ₇₄	Ŷ _{7,5}	Ŷ _{7,6}	Ŷ _{7,7}	Ŷ _{7,8}	Ŷ _{7,9}				
8	Y ₈₁	Y ₈₂	Y ₈₃	Ŷ _{8,4}	Ŷ _{8,5}	Ŷ _{8,6}	Ŷ _{8,7}	Ŷ _{8,8}	Ŷ _{8,9}				
9	Y ₉₁	Y ₉₂	Ŷ _{9,3}	Ŷ _{9,4}	Ŷ _{9,5}	Ŷ _{9,6}	Ŷ _{9,7}	Ŷ _{9,8}	Ŷ _{9,9}				
10	Y _{10,1}	Ŷ _{10,2}	Ŷ _{10,3}	Ŷ _{10,4}	Ŷ _{10,5}	Ŷ _{10,6}	Ŷ _{10,7}	Ŷ _{10,8}	Ŷ _{10,9}				

Tableau IV.4
Montant de Réserves estimé pour la branche Dommages Auto

Année d'origine	Estimation des réserves sans extrapolation					
	Chain Ladder	Chain Ladder Ajusté	GLM Poisson Surdispersé - Log	GLM Gamma - Log	LogNormale LogNormale - Identité	
	1	0	0	0	0	0
	2	16	16	16	24	25
3	45	45	45	53	76	
4	107	107	107	156	176	
5	255	255	255	422	392	
6	561	611	561	1 024	840	
7	1 113	1 209	1 113	1 506	1 359	
8	1 781	2 085	1 781	1 732	2 122	
9	3 813	4 895	3 813	3 281	4 651	
10	27 873	31 478	27 873	26 468	27 662	
Total	35 563	40 700	35 563	34 666	37 303	
Ecart relatif / Chain Ladder		14%	0%	-3%	5%	
Ecart relatif / Chain Ladder Ajusté	-13%		-13%	-15%	-8%	

Quelque soit la branche étudiée, le modèle Chain Ladder Ajusté, qui est notre « Best estimate » déterministe est plus pessimiste que le Chain Ladder standard. La différence se fait sur les 2 ou 3 dernières années d'origine. En tant que modèle d'expert, le Chain Ladder Ajusté prend en compte certaines informations (après échange avec le client, les gestionnaires de sinistres,...) qui l'incitent à « gonfler » les provisions des années 8, 9 et/ou 10 selon la branche étudiée.

Le modèle Chain Ladder standard ou sa version stochastique GLM Poisson Surdispersé donnent par construction, une importance excessive aux premiers développements connus, en particulier pour les années d'origine les plus récentes ; or il se trouve que pour les années d'origine 8, 9 et/ou 10, les montants des premiers développements disponibles sont assez faibles et se répercutent dans les estimations des paiements des développements suivants,

permettant au Chain Ladder standard ou GLM Poisson Surdispersé d'être plus optimistes qu'ils ne devraient. Cette remarque peut également être faite pour le modèle GLM Gamma qui a la même construction que le GLM Poisson Surdispersé et qui estime des montants de réserves sensiblement proches de ceux du modèle GLM Poisson Surdispersé ou du modèle Chain Ladder standard.

Nous retiendrons donc que pour chaque branche étudiée, les modèles GLMs de Poisson Surdispersé et Gamma sont très éloignés en terme de « best estimate » du modèle de référence en comparaison avec la régression LogNormale et les modèles GAMs (écarts avec Chain Ladder Ajusté surlignés en rose dans les 3 tableaux IV. 4, .5, .6).

Ce phénomène s'observe de façon importante pour la branche courte « Dommages aux biens Automobile » avec un net écart de plus de 13% entre le « best estimate » et les estimations des modèles de Chain Ladder standard et GLMs, alors qu'a priori, cette catégorie est la plus facile à modéliser et ne devrait pas engendrer de telles différences entre les modélisations.

Tableau IV.5
Montant de Réserves estimé pour la branche RC Auto

Estimation des réserves sans extrapolation						
Année d'origine	Chain Ladder	Chain Ladder Ajusté	GLM		LogNormale	GAM
			Poisson Surdispersé - Log	Gamma - Log		Gamma - Log
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	563	563	563	613	616	617
4	1 568	1 568	1 568	1 725	1 720	1 783
5	3 423	3 423	3 423	3 410	3 480	3 611
6	6 229	6 488	6 229	6 914	6 562	6 205
7	9 904	11 037	9 904	10 185	9 744	9 694
8	13 618	15 182	13 618	13 531	13 195	13 184
9	15 817	17 541	15 817	16 098	15 747	17 608
10	20 001	20 564	20 001	19 728	19 583	23 896
Total	71 122	76 367	71 122	72 203	70 645	76 598
Ecart relatif / Chain Ladder		7%	0%	2%	-1%	8%
Ecart relatif / Chain Ladder Ajusté		-7%	-7%	-5%	-7%	0%

Seuls les montants de réserve totale estimés par les modèles GAMs et parfois par le modèle LogNormal excepté pour la branche « Responsabilité Civile Automobile » sont assez proches du « Best estimate » du Chain Ladder Ajusté. Il semble que GAM s'adapte très bien aux années d'origine 8, 9 et/ou 10 selon la branche étudiée (valeurs surlignées en rose), qui sont comme nous l'avons remarqué précédemment, les années pour lesquelles Chain Ladder Ajusté diffère le plus des modèles de Chain Ladder standard, GLMs et LogNormale.

Tableau IV.6
Montant de Réserves estimé pour la branche RC Générale

Estimation des réserves sans extrapolation							
Année d'origine	Chain Ladder	Chain Ladder Ajusté	GLM		LogNormale	GAM	
			Poisson Surdispersé -	Gamma -		Poisson Surdispersé -	Gamma -
			Log	Log		Log	Log
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	373	373	373	368	518	319	337
4	667	667	667	688	883	764	798
5	1 566	1 566	1 566	1 754	2 031	1 420	1 481
6	2 234	2 315	2 234	2 347	2 343	2 348	2 433
7	3 840	4 322	3 840	3 806	3 440	3 394	3 482
8	4 591	5 162	4 591	4 574	4 537	4 416	4 478
9	4 295	5 065	4 295	4 213	3 417	5 351	5 387
10	5 555	6 954	5 555	5 537	7 996	6 797	6 657
Total	23 121	26 424	23 121	23 287	25 165	24 809	25 052
Ecart relatif / Chain Ladder		14%	0%	1%	9%	7%	8%
Ecart relatif / Chain Ladder Ajusté		-12%	-12%	-12%	-5%	-6%	-5%

Les deux branches longues « Responsabilité Civile Automobile » et « Responsabilité Civile Générale » nécessitent une extrapolation des paiements estimés au-delà du triangle de données pour chaque modèle étudié. La charge ultime prend en considération l'ensemble de ces montants estimés.

Tableau IV.7
Montant de Réserves estimé pour la branche RC Auto après extrapolation

Estimation des réserves après extrapolation					
Année d'origine	Chain Ladder Ajusté	GLM		LogNormale	GAM
		Poisson Surdispersé -	Gamma -		Gamma -
		Log	Log		Log
1	362	298	328	321	470
2	625	529	573	562	880
3	1 248	1 142	1 242	1 234	1 628
4	2 285	2 174	2 385	2 368	2 929
5	4 274	4 142	4 189	4 247	4 899
6	7 479	7 061	7 831	7 454	7 609
7	12 039	10 727	11 083	10 617	11 116
8	16 150	14 404	14 380	14 022	14 518
9	18 317	16 438	16 777	16 406	18 793
10	21 159	20 492	20 254	20 096	24 924
Total	83 938	77 407	79 042	77 326	87 766
Ecart relatif / Chain Ladder Ajusté		-8%	-6%	-8%	5%

Un des avantages des GAMs est la présence de variables explicatives continues dans le modèle qui offrent la possibilité de prolonger la courbe de lissage $s_v(Dev_j)$ au-delà des rangées de données observées. Cette technique est assez intuitive et est une alternative aux pratiques habituelles réalisées sur les modèles Chain Ladder Ajusté, LogNormale et GLMs qui consistent à ajuster une courbe paramétrique sur les facteurs de développement issus du modèle étudié et à la prolonger jusqu'à ce qu'elle devienne stable. Les facteurs de

développement sont dans le cas du modèle GLM Poisson Surdispersé les mêmes que ceux de Chain Ladder standard ; pour les autres modèles il s'agit de facteurs de développement « équivalents » obtenus^[3] en appliquant la même technique que Chain Ladder standard aux valeurs ajustées par le modèle étudié. Une fois l'ajustement de la courbe paramétrique sur les facteurs de développement effectué, les montants estimés sont calculés en appliquant la méthode de Chain Ladder standard aux facteurs de développement ajustés sur les dernières années de développement.

Tableau IV.8
Montant de Réserves estimé pour la branche RC Générale après extrapolation

Estimation des réserves après extrapolation						
Année d'origine	Chain Ladder	GLM		LogNormale	GAM	
	Ajusté	Poisson	Gamma -		Poisson	Gamma -
		Surdispersé -	Log		Surdispersé -	Log
1	335	378	395	453	615	824
2	553	592	546	694	859	1 113
3	970	1 022	966	1 289	1 189	1 487
4	1 144	1 185	1 167	1 505	1 633	1 966
5	2 160	2 206	2 355	2 813	2 279	2 647
6	2 814	2 772	2 849	2 982	3 184	3 571
7	4 914	4 473	4 388	4 156	4 194	4 565
8	5 679	5 128	5 069	5 164	5 170	5 489
9	5 464	4 689	4 572	3 821	6 054	6 319
10	7 361	5 919	5 871	8 587	7 451	7 512
Total	31 394	28 364	28 179	31 464	32 628	35 494
Ecart relatif / Chain Ladder Ajusté						
		-10%	-10%	0%	4%	13%

Une analyse des tableaux de résultats IV.7 et IV.8 nous indique que l'écart entre le « Best estimate » de Chain Ladder Ajusté et le montant de réserve des GAMs s'est creusé pour les deux branches suite à l'extrapolation. L'explication de cet écart tient essentiellement dans la différence de technique utilisée pour obtenir les estimations des paiements au-delà du triangle de données. Les résultats des GAMs restent encore très satisfaisants, notamment pour le GAM Poisson Surdispersé de la branche « Responsabilité Civile Générale » et pour le GAM Gamma de la branche « Responsabilité Civile Automobile ». L'ensemble des remarques faites avant extrapolation reste valable après extrapolation. De plus, l'extrapolation des GAMs est une technique plus consistante que celle appliquée aux autres modèles en ajustant une courbe paramétrique, renforçant ainsi nos conclusions sur les résultats des GAMs des tableaux IV.7 et IV.8.

Précision de l'estimation des réserves

L'erreur de prédiction ou RMSEP (« Root Mean Square Error of Prediction »)

Un des principaux avantages des modèles de provisionnement stochastique est l'existence de mesures de précision des estimations. Nous avons choisi d'utiliser le « Root Mean Square Error of Prediction » ou RMSEP. Supposons que nous ayons une estimation \hat{R} du montant de réserve total R . Alors le RMSEP est défini par : $\sqrt{E(R - \hat{R})^2}$.

Le RMSEP est une mesure de variabilité de l'estimation \hat{R} et peut être considéré comme la somme de deux composantes : la variabilité des données (variance du processus) et la variabilité due à l'estimation (variance de l'estimation).

Une méthode détaillée en annexe donne une approximation du calcul du RMSEP. Ce calcul nécessite la matrice de variance-covariance des coefficients paramétriques pour les GLMs et non paramétriques pour les GAMs. La méthode sera surnommée « Delta ». L'approximation du RMSEP peut être utilisée également pour construire des intervalles de confiance.

Le modèle de Mack^[6] permet également le calcul d'un RMSEP qui est un critère d'évaluation de la variation des estimations de Chain Ladder standard et de Chain Ladder Ajusté. Ce calcul diffère de celui des GAMs et GLMs car il repose sur les facteurs de développement ; il est également exposé en annexe.

Les RMSEP rapportés au montant de réserve estimé servent d'indicateur de qualité des estimations des réserves.

Le tableau IV.9 résume l'ensemble des résultats de RMSEP obtenus pour les modèles de Chain Ladder et Chain Ladder Ajusté (avec Mack), les modèles GLMs, GAMs et LogNormale ajustés sur la branche « Responsabilité Civile Générale ». Les conclusions apportées sur cette catégorie sont également valables pour les deux autres catégories.

Les résultats de l'erreur de prédiction RMSEP exprimés en un pourcentage de l'estimation du montant des réserves ne sont pas comparables pour des distributions différentes. En effet, la différence entre les erreurs de prédiction des modèles Poisson Surdispersé et Gamma est en partie due à la présence du paramètre d'échelle Φ . Quant à la distribution LogNormale, le mode de calcul du « MSEP » est différent de celui des autres modélisations. Une remarque peut cependant être faite : pour les premières années d'origine ($Year_3, Year_4, \dots$) des modèles GAMs et GLMs, le rapport $RMSEP(R_i)/R_i$ de la loi Gamma est plus faible que celui de la loi Poisson Surdispersé ; la modélisation Gamma privilégie les années les plus anciennes. Ce comportement se renverse pour les années les plus récentes ($Year_{10}, Year_9, \dots$) car le rapport $RMSEP(R_i)/R_i$ de la loi Gamma est plus élevé que celui de la loi Poisson Surdispersé. Ainsi, comme ces dernières années ont un poids très important dans le montant de réserve total, le rapport $RMSEP(R)/R$ de la loi Gamma est supérieur à celui de la Poisson Surdispersé.

En outre, la comparaison des modèles GLMs avec les modèles GAMs pour une distribution fixée peut se faire et elle indique que ces derniers donnent de meilleurs résultats en terme de variabilité du montant de réserves estimé par année d'origine et au total.

Tableau IV.9
RMSEP et Intervalle de Confiance de chaque modèle pour la branche RC Générale

Modèle de Mack			Modèle de Mack sur Chain Ladder Ajusté		
Année d'origine	Provision	RMSEP(R_i) %	Année d'origine	Provision	RMSEP(R_i) %
1	-	-	1	-	-
2	-	-	2	-	-
3	373	18%	3	373	18%
4	667	56%	4	667	56%
5	1 566	37%	5	1 566	37%
6	2 234	46%	6	2 315	44%
7	3 840	36%	7	4 322	34%
8	4 591	30%	8	5 162	29%
9	4 295	29%	9	5 065	29%
10	5 555	27%	10	6 954	29%
Total	23 121	16%	Total	26 424	16%
Intervalle de Confiance	15 797	30 445	Intervalle de Confiance	18 130	34 719

Tableau IV.9- Suite
RMSEP et Intervalle de Confiance de chaque modèle pour la branche RC Générale

Modèle GLM - Poisson Surdispersé			Modèle GLM - Gamma			LogNormale		
Année d'origine	Provision	RMSEP(Ri) %	Année d'origine	Provision	RMSEP(Ri) %	Année d'origine	Provision	RMSEP(Ri) %
1	-	-	1	-	-	1	-	-
2	-	-	2	-	-	2	-	-
3	373	75%	3	368	47%	3	518	79%
4	667	53%	4	688	34%	4	883	57%
5	1 566	36%	5	1 754	30%	5	2 031	48%
6	2 234	31%	6	2 347	28%	6	2 343	44%
7	3 840	25%	7	3 806	28%	7	3 440	44%
8	4 591	25%	8	4 574	29%	8	4 537	45%
9	4 295	30%	9	4 213	32%	9	3 417	51%
10	5 555	44%	10	5 537	41%	10	7 996	74%
Total	23 121	16%	Total	23 287	16%	Total	25 165	30%
Intervalle de Confiance	15 682	30 560	Intervalle de Confiance	16 006	30 569	Intervalle de Confiance	10 466	39 864

Modèle GAM - Poisson Surdispersé			Modèle GAM - Gamma			LogNormale		
Année d'origine	Provision	RMSEP(Ri) %	Année d'origine	Provision	RMSEP(Ri) %	Année d'origine	Provision	RMSEP(Ri) %
1	-	-	1	-	-	1	-	-
2	-	-	2	-	-	2	-	-
3	319	74%	3	337	43%	3	505	58%
4	764	48%	4	798	30%	4	1 085	41%
5	1 420	35%	5	1 481	25%	5	1 886	33%
6	2 348	27%	6	2 433	23%	6	2 790	30%
7	3 394	23%	7	3 482	21%	7	3 893	28%
8	4 416	21%	8	4 478	21%	8	4 844	27%
9	5 351	21%	9	5 387	22%	9	5 809	28%
10	6 797	22%	10	6 657	25%	10	7 859	33%
Total	24 809	14%	Total	25 052	15%	Total	28 669	19%
Intervalle de Confiance	17 839	31 778	Intervalle de Confiance	17 669	32 435	Intervalle de Confiance	17 833	39 506

La comparaison des erreurs de prédiction des différents modèles avec celles de Mack est assez difficile. Cette dernière modélisation ne repose pas sur l'hypothèse d'une distribution sous-jacente mais les résultats de variabilité du RMSEP permettent de donner une indication sur la robustesse du modèle. Dans cette optique, il est intéressant de noter que les résultats de variabilité du montant de réserve global des modèles GAMs sont plus bas que ceux de Mack sur le modèle de référence.

Pour la branche « Responsabilité Civile Générale », les modèles GAMs semblent donc plus robustes que les GLMs et ont même un indicateur de variabilité plus faible que celui du modèle de référence.

Au sein des modèles GAMs, le choix du modèle le mieux adapté peut se faire entre la modélisation de Poisson Surdispersé et celle de la loi Gamma. Il est difficile de choisir analytiquement l'un des deux car nous venons de voir que leurs résultats de RMSEP ne sont pas comparables. Mais à la vue des résultats précédents concernant l'écart relatif sur le montant de réserve avec la modélisation de Chain Ladder Ajusté avant et **après** extrapolation, on préférera le modèle GAM Poisson Surdispersé.

L'erreur de prédiction du Bootstrap

La méthodologie du Bootstrap qui n'est pas décrite dans cet article mais que le lecteur pourra lire en détail dans England P.D et Verrall R.J^[2], est une technique de ré-échantillonnage des résidus du modèle qui permet d'obtenir une estimation de la distribution des provisions. L'écart-type ou l'erreur standard de la distribution peut ainsi être calculé. Cette erreur standard issue du Bootstrap est une estimation de la racine carrée de la variance de la distribution. Cependant, elle ne peut être comparée directement avec son équivalent analytique qu'est la 2^{ème} composante du RMSEP (variance de l'estimation), car cette erreur ne tient pas compte du nombre de paramètres utilisés dans le modèle^[13]. Pour obtenir l'erreur de prédiction du Bootstrap notée PE et pouvoir ainsi la comparer avec le RMSEP, il est nécessaire d'ajouter une estimation de la variance du processus, laquelle est égale au paramètre d'échelle multiplié par l'estimation des réserves.

Tableau IV.10
Indicateurs de la volatilité des prévisions de réserves de la branche RC Générale : RMSEP (Mack et Delta) et PE (Bootstrap)

	Provision	Mack	Delta		Bootstrap	
		RMSEP	RMSEP	Intervalle Confiance	PE	Intervalle Confiance
Chain Ladder standard	23 121	16%				
Chain Ladder Ajusté	26 424	16%				
GLM Poisson Surdispersé	23 121		16%	[15 682;30 560]	20%	[17 477;30 922]
GLM Gamma	23 287		16%	[16 006;30 569]	18%	[17 855;31 743]
LogNormale	25 165		30%	[10 466;39 864]	41%	[20 467;54 980]
GAM Gamma	25 052		15%	[17 669;32 435]	16%	[19 325;33 516]
GAM Poisson Surdispersé	24 809		14%	[17 839;31 778]	16%	[19 825;32 178]

A distribution et fonction de lien fixées, les erreurs de prédiction Bootstrap (PE_{bs}) sont assez proches des erreurs de prédiction analytiques (RMSEP), à la fois au total et par année d'origine. Le tableau IV.10 illustre en partie cette constatation pour le montant total de réserve de la branche « Responsabilité Civile Générale ». Ceci est assez remarquable étant donnée la différence structurelle qui existe entre les deux méthodes utilisées.

Les résultats des trois branches étudiées ne sont pas présentés dans le détail, excepté pour la branche « Responsabilité Civile Générale », mais nous amènent aux mêmes conclusions :

- Les erreurs de prédiction Bootstrap (PE_{bs}) sont assez proches des erreurs de prédiction analytiques (RMSEP), à distribution et fonction de lien fixées, avec cependant une légère tendance pour ces dernières à être inférieures à celles du Bootstrap.
- La comparaison des modèles GLMs avec les modèles GAMs, pour une distribution fixée, indique que ces derniers donnent de meilleurs résultats en terme de variabilité du montant de réserves estimé au global et par année d'origine, et cela, quel que soit l'indicateur utilisé (la mesure analytique ou la mesure issue du Bootstrap).
- Les bandes de confiance construites avec les méthodes Delta et Bootstrap se recoupent assez bien.

Distribution de la prédiction et estimation des quantiles

Connaître les moments du montant de réserve R (espérance, variance) n'est pas suffisant pour obtenir certaines caractéristiques de R : probabilité d'insuffisance, quantiles,.... C'est pourquoi, il est nécessaire d'appliquer certaines méthodes comme le Bootstrap pour avoir la distribution de R ou encore d'autres méthodes alternatives^[11] pour avoir les quantiles de R. Ces dernières reposent en partie sur les premiers moments de R et consistent en :

- une approche additive qui dans le cas présent ne s'adapte qu'à la distribution de Poisson Surdispersé,
- une approximation Normal Power qui s'applique à toutes les distributions étudiées,
- une approximation Gamma qui ne s'applique qu'aux distributions ayant un coefficient d'asymétrie éloigné de 0.

Le lecteur pourra trouver toute la théorie et les champs d'application de ces méthodes dans l'article de Partrat C., Pey N. et Schilling J. ^[11].

Tableau IV.11
Estimation des quantiles de R par le Bootstrap et les méthodes alternatives sur la
branche « Responsabilité Civile Générale »

Quantiles		50	75	90	95	99
GLM Poisson Surdispersé	Bootstrap	22 870	25 231	27 688	29 222	32 665
	Additive	23 121	24 286	25 335	25 963	27 140
	NP approx	21 393	22 569	23 643	24 294	25 529
	Gamma approx					
GLM Gamma	Bootstrap	23 120	25 298	27 915	29 884	33 532
	Additive					
	NP approx	21 717	22 795	23 799	24 414	25 597
	Gamma approx	23 256	24 329	25 319	25 923	27 077
LogNormale	Bootstrap	29 088	33 767	39 845	45 943	63 197
	Additive					
	NP approx					
	Gamma approx					
GAM Poisson Surdispersé	Bootstrap	25 273	27 548	29 535	31 038	34 169
	Additive	24 808	25 991	27 056	27 693	28 888
	NP approx	23 055	24 247	25 336	25 995	27 246
	Gamma approx					
GAM Gamma	Bootstrap	25 621	28 223	30 716	32 529	35 035
	Additive					
	NP approx	23 395	24 532	25 590	26 238	27 484
	Gamma approx	25 020	26 151	27 195	27 831	29 047

Le tableau IV.11 indique que les différentes méthodes d'estimation de quantiles utilisées fournissent des résultats assez proches, notamment pour les quantiles à 50 et 75%. Toutefois, les quantiles de l'approximation Normal Power semblent légèrement en dessous des autres et les derniers quantiles du Bootstrap ont tendance à s'échapper vers le haut laissant présager quelques dérives sur la queue de distribution. Ce problème de valeurs extrêmes est d'ailleurs vérifié par une lecture graphique de l'histogramme de la distribution

des réserves. Les méthodes alternatives étudiées pourraient être une solution à ce problème. Le rôle de l'actuaire n'est cependant pas de choisir une des méthodes indiquées pour obtenir la valeur d'un quantile mais plutôt d'appliquer selon le contexte de son étude et ses objectifs, une méthode de sélection d'un quantile. Par exemple, dans le cadre des normes IFRS, une marge pour risque et incertitude (MRI) pourrait accompagner l'estimation « best estimate » des réserves. Le principe général du calcul de cette marge pourrait être le suivant :

$$\text{MRI} = \beta\% \cdot \text{Réserves « Best Estimate »}.$$

De nombreuses méthodes sont suggérées pour déterminer le facteur β de MRI, dont deux qui sont rendues possibles grâce au provisionnement stochastique :

- Approche de la variabilité dans le montant « Best estimate » des provisions :

Le facteur de MRI est égal à un facteur de l'écart-type de la distribution du montant des réserves.

- Approche « percentile » :

Le montant des réserves (« Best Estimate »+MRI) est égal à un percentile de la distribution des réserves.

Dans l'hypothèse où cette deuxième approche serait validée, la sélection du quantile pourrait être faite, si le minimum requis est par exemple un quantile de 75% comme en Australie avec les modèles MOS, en faisant la moyenne de tous les quantiles à 75% issus des différentes méthodes (Bootstrap et alternatives) et en retirant le plus faible.

V. CONCLUSIONS

La réponse à la question « Quel est le meilleur modèle d'ajustement » n'est pas évidente. En effet, les données doivent être examinées en détail afin de trouver un modèle approprié et non pas de leur appliquer la même approche de modélisation quelles que soient les circonstances.

Le sujet des modèles stochastiques appliqués au provisionnement a donné lieu ces dernières années à un certain nombre de controverses, mais il est incontournable pour répondre aux problématiques du contexte réglementaire lié aux normes IFRS et au projet Solvabilité II. Cependant, il existe une grande variété de méthodes disponibles pour le provisionnement stochastique. Nous en avons étudié quelques unes en plus des modèles GAMs et il est important que les similitudes et les différences entre les modèles soient bien comprises et leurs propriétés examinées.

Nous avons essayé d'y contribuer en présentant dans ce mémoire les particularités des modèles GAMs qui les différencient des autres modélisations. Nous avons également étudié un certain nombre de critères dont certains nous ont permis de sélectionner au sein des modélisations GAMs, les modèles les mieux adaptés à chacune des branches étudiées. Les autres critères nous ont aidé à comparer ces modèles GAMs sélectionnés avec les modélisations de Chain Ladder, Mack, GLMs, et la régression LogNormale.

De cette étude sur les trois branches de sinistres à déroulement plus ou moins long, est sorti un choix de modèle GAM différent selon la catégorie prise en compte. Nous avons pu illustrer sur ces trois branches la grande flexibilité de l'ajustement qu'offre le lissage non paramétrique du fait de l'indépendance du prédicteur avec une quelconque forme rigide paramétrique, permettant ainsi au modèle de suivre la tendance présente dans les données.

Nous avons également remarqué que ces modèles GAMs « contrôlent » leur niveau d'ajustement aux données contrairement aux modèles GLMs qui ont tendance à être sur-paramétrés : en effet, le nombre de degrés de liberté diffère selon le modèle GAM ajusté, mais il est en moyenne égal à 11 contre un nombre de paramètres égal à 19 dans le cas d'une modélisation GLM.

Cette étude a révélé que les modèles GAMs ajustés sur chaque branche semblent plus robustes que les autres modèles comparatifs grâce à une erreur de prédiction du montant de réserve global et par année d'origine nettement plus faible.

A l'analyse de la qualité de l'ajustement des modèles à travers les indicateurs de variabilité, s'ajoute l'étude de leur « best estimate » avec comme référence le modèle de Chain Ladder Ajusté, qui nous amène à considérer les modèles GAMs comme satisfaisants pour répondre au concept de « best estimate » et au besoin d'une mesure de risque associée rendus obligatoires par l'évolution probable du cadre réglementaire.

Il serait donc intéressant de poursuivre cette étude sur d'autres triangles de liquidation appartenant aux mêmes branches et à d'autres branches de longueurs différentes, afin de confronter les résultats obtenus à ceux de cette étude. D'autre part, la modélisation GAM peut être utilisée avec la structure de la Quasi-Vraisemblance qui permet de travailler sur des triangles contenant des données négatives. Il serait alors possible de modéliser, grâce à ces modèles, un triangle de charges sous la condition non évidente d'indépendance des variables.

BIBLIOGRAPHIE

- [1] **England, P. D.** Addendum to 'Analytic and bootstrap estimates of prediction errors in claims reserving', 2001. Actuarial Research Paper No. 138, Department of Actuarial Science and Statistics, City University, London, EC1V 0HB.
- [2] **England, P. D. & Verrall, R. J.** Analytic and bootstrap estimates of prediction errors in claims reserving, 1999. Insurance: Mathematics and Economics, 25, 281-293.
- [3] **England, P. D. & Verrall, R. J.** A flexible framework for stochastic claims Reserving, 2001. Proceedings of the Casualty Actuarial Society.
- [4] **England, P. D. & Verrall, R. J.** Stochastic claims reserving in general insurance, 2002. Institute of Actuaries.
- [5] **Hastie, T. J. & Tibshirani, R. J.** Generalized additive models, 1990. Chapman and Hall, London.
- [6] **Mack, T.** Distribution-free calculation of the standard error of chain-ladder reserve Estimates, 1993. ASTIN Bulletin, 23, 213-225.
- [7] **Mack, T.** Which stochastic model is underlying the chain-ladder model?, 1994a. Insurance: Mathematics and Economics, 15, 133-138.
- [8] **Mack, T.** Measuring the variability of chain-ladder reserve estimates, 1994b. Casualty Actuarial Society, Spring Forum.
- [9] **McCullagh, P. & Nelder, J.** Generalised linear models, 2nd edition, 1989. Chapman and Hall.
- [10] **Partrat C., Besson J-L.** Assurance non-vie : Modélisation, simulation, 2005. Economica.
- [11] **Partrat C., Pey N., & Schilling J.** Delta method and Reserving, 2005. Astin Colloq., Zurich.
- [12] **Plancher F., Théron P. & Jacquemin J.** Modèles financiers en assurance, Analyses de risque dynamiques, 2005. Assurance Audit Actuariat.

- [13] **Pinheiro, P., Andrade e Silva, J., & Centeno, M.** Bootstrap methodology in claim reserving, 2001. Astin Colloq., Washington.
- [14] **Verrall, R. J.** Claims reserving and generalised additive models, 1996. Insurance: Mathematics and Economics, 19, 31-43.
- [15] **Wahba, G.** Spline models for observational data, 1990. CBMS-NSF Reg. Conf. Ser. Appl. Math. :59.
- [16] **Wood, S.N.** Modelling and smoothing parameter estimation with multiple quadratic penalties, 2000. J. R. Statist. Soc. B 62, 413-428.
- [17] **Wood, S.N. & Augustin, N.H.** GAMs with integrated model selection using penalized regression splines and applications to environmental modelling, 2002. Ecol. Model. 157, 157-177.

ANNEXES

Quasi Vraisemblance Etendue (ou Extended Quasi-Likelihood)

Ce critère repose en partie sur la déviance : la déviance est une mesure de bon ajustement au sein du modèle et permet de comparer des modèles avec des fonctions de lien g différentes. Cette notion de déviance est détaillée ci-dessous.

Cependant, cette mesure de comparaison des modèles n'est plus valable lorsque la fonction de variance $V(\cdot)$, soit la distribution des Y_{ij} , est changée : si l'on considère V' , telle que $V'(s) \geq V(s) \forall s$, alors la déviance diminue de façon mécanique sans rendre compte de la qualité des modèles. Nous avons donc cherché un autre indicateur qui corrige cet effet, et qui permet de comparer les modèles avec des fonctions de variance et des fonctions de liens différentes. Il s'agit de l'Extended Quasi-Likelihood q^+ développée par Nelder et Pregibon puis Nelder et Lee :

$$-2q^+(y, \hat{\mu}) = \frac{1}{\phi} \sum_{i+j \leq n} d_{ij} + \sum_{i+j \leq n} \log[2\pi\phi V(y_{ij})]$$

où ϕ est estimé par $\hat{\phi} = \frac{X^2}{m} = \frac{1}{m} \sum_{i+j \leq n} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{V(\hat{\mu}_{ij})}$, m étant le nombre de degrés de

liberté des résidus du modèle (nombre de données - nombre de degrés de liberté du modèle), et X^2 étant la statistique de Pearson;

• $\sum_{i+j \leq n} d_{ij}$ est la quasi-déviance avec $d_{ij} = 2 \int_{\hat{\mu}_{ij}}^{y_{ij}} \frac{y_{ij} - u}{V(u)} du$

Maximiser la statistique q^+ permet de comparer les modèles qui ont une structure (V, ϕ, g) différente. Cela revient à minimiser la statistique $-2q^+$.

Notion de déviance

La déviance joue le rôle de la somme au carré des résidus pour les modèles généralisés et peut être utilisée pour évaluer la qualité de l'ajustement et pour comparer les modèles entre eux :

$$D = 2 \sum_{i+j \leq n} d_{ij}$$

où $d_{ij} = 2 \left\{ y_{ij} (\tilde{\theta}_{ij} - \hat{\theta}_{ij}) - [b(\tilde{\theta}_{ij}) - b(\hat{\theta}_{ij})] \right\}$ est le (i,j) ème terme de déviance avec les notations : $\tilde{\theta}_{ij} = b^{-1}(y_{ij})$ et $\hat{\theta}_{ij} = b^{-1}(\hat{\mu}_{ij})$

Test de significativité des coefficients de lissage

Pour $H_0 : s_\lambda = 0$, la statistique de test est la suivante :

$$X_\lambda = p'_\lambda V_\lambda^{k-} p_\lambda$$

Où p_λ est le vecteur de paramètres du terme de lissage s_λ ,

V_λ est la matrice de covariance,

V_λ^{k-} est la pseudo-inverse de rang k de V_λ , avec $k = \text{rang estimé de } V_\lambda$

et elle vérifie :

$$X_\lambda = p'_\lambda V_\lambda^{k-} p_\lambda \approx \chi^2(k)$$

Cette statistique asymptotique est utilisable pour étudier la significativité de s_λ . Le paramètre ϕ intervient dans l'expression de X_λ . Lors des tests statistiques, nous disposons de $\hat{\phi}$ et non de ϕ (ce dernier étant inconnu). Il faut donc prendre en compte la distribution asymptotique de ϕ dans notre statistique. Le calcul de ϕ est effectué à partir de la somme des résidus qui convergent vers une loi du chi deux :

$$m \frac{\hat{\phi}}{\phi} \approx \chi^2(m)$$

où m est le nombre de degré de liberté des résidus du modèle
(nombre de données - nombre de degrés de liberté du modèle).

En combinant ces deux statistiques qui sont supposées être indépendantes, il est alors possible d'étudier la significativité des fonctions de lissage s_λ et s_ν .

$$F_\lambda = X_\lambda \frac{\hat{\phi}}{k\phi} \approx F(k, m)$$

Ainsi l'hypothèse contrainte sera rejetée au seuil α si $F_\lambda \geq F_{1-\alpha}(k, m)$.

Il est possible d'associer la région critique à l'hypothèse H_0 du modèle contraint :

$$W = F_\lambda \geq F_{1-\alpha}(k, m)$$

La frontière de cette région critique fournit l'intervalle de confiance associée au terme de lissage estimé.

Résidus

Les résidus de Pearson (avec paramètre d'échelle Φ) sont définis par :

$$r_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\phi V(\hat{\mu}_{ij})}}$$

Les résidus de déviance (avec paramètre d'échelle Φ) sont définis par :

$$r_{ij} = \text{sign}(y_{ij} - \hat{\mu}_{ij}) \sqrt{\frac{d_{ij}}{\phi}}$$

où d_{ij} a été défini précédemment.

Calcul du RMSEP de la méthode Delta

Tous les détails de ce calcul se trouvent dans l'article^[4] d'England et Verrall. Un certain nombre de résultats sont présentés ici.

La décomposition du RMSEP en somme de deux variances, peut s'écrire dans le cas des GLMs et des GAMs:

$$MSEP(\hat{Y}_{ij}) = E\left[\left(Y_{ij} - \hat{Y}_{ij}\right)^2\right] \approx \text{Var}(Y_{ij}) + \text{Var}(\hat{Y}_{ij}) \quad \text{où} \quad \hat{Y}_{ij} = \hat{\mu}_{ij} = \exp(\hat{\eta}_{ij})$$

et $\forall n+1 < i+j$

•Distributions Poisson Surdispersé et Gamma

On peut montrer que le calcul du RMSEP pour l'estimation $\hat{R}_i = \sum_{i+j>n+1} \hat{Y}_{ij}$ est :

$$MSEP(\hat{R}_i) \approx \sum_{j>n+1-i}^n \phi \hat{m}_{ij}^\rho + \sum_{j>n+1-i}^n \hat{m}_{ij}^2 \text{Var}(\hat{\eta}_{ij}) + 2 \sum_{\substack{j,k>n+1-i \\ j<k}}^n \hat{m}_{ij} \hat{m}_{ik} \text{Cov}(\hat{\eta}_{ij}, \hat{\eta}_{ik}), \quad \forall 1 \leq i \leq n$$

Et pour le total $\hat{R} = \sum_{i=2}^n \hat{R}_i$, on a :

$$MSEP(\hat{R}) \approx \underbrace{\sum_{i+j>n+1} \text{Var}(Y_{ij})}_{1^{\text{er}} \text{ terme}} + \underbrace{\sum_{i+j>n+1} \hat{m}_{ij}^2 \text{Var}(\hat{\eta}_{ij}) + 2 \sum_{(i,j) \neq (i_2,k)} \hat{m}_{i_1j} \hat{m}_{i_2k} \text{Cov}(\hat{\eta}_{i_1j}, \hat{\eta}_{i_2k})}_{2^{\text{ème}} \text{ terme}}$$

Le 1^{er} terme est la somme des variances du processus alors que le 2^{ème} terme relate la variance de l'estimation.

• **Distribution LogNormale**

Pour la loi LogNormale avec la fonction de lien identité, le calcul du RMSEP est un peu différent :

$$MSEP(\hat{R}_i) \approx \sum_{j>n+1-i}^n MSEP(\hat{Y}_{ij}) + 2 \sum_{\substack{j,k>n+1-i \\ j<k}}^n \hat{Y}_{ij} \hat{Y}_{ik} (\exp(\text{Cov}(\hat{\eta}_{ij}, \hat{\eta}_{ik})) - 1), \quad \forall 1 \leq i \leq n$$

$$MSEP(\hat{R}) \approx \sum_{i+j>n+1} MSEP(\hat{Y}_{ij}) + 2 \sum_{(i,j) \neq (i_2,k)} \hat{Y}_{ij} \hat{Y}_{i_2k} (\exp(\text{Cov}(\hat{\eta}_{ij}, \hat{\eta}_{i_2k})) - 1)$$

Calcul du RMSEP de Mack

Toute la théorie de Mack et ses hypothèses de travail se trouvent en annexe^{[7] et [8]}. Sous certaines hypothèses détaillées dans cet article, l'incertitude présente dans la prédiction de R_i par \hat{R}_i , classiquement mesurée par l'erreur quadratique moyenne (Mean Square Error Of Prediction ou MSEP), s'écrit:

$$\hat{MSEP}(\hat{R}_i) = \hat{C}_{in}^2 \sum_{j=n-i+1}^{n-1} \frac{\hat{\sigma}_j^2}{\hat{f}_j^2} \left[\frac{1}{\hat{C}_{ij}} + \frac{1}{\sum_{k=1}^{n-j} C_{kj}} \right] \quad (\hat{C}_{i,n-i+1} = C_{i,n-i+1})$$

et

$$\hat{MSEP}(\hat{R}) = \sum_{i=2}^n \left[\hat{MSEP}(\hat{R}_i) + \hat{C}_{in} \left(\sum_{k=i+1}^n \hat{C}_{kn} \right) \sum_{j=n-i+1}^{n-1} \frac{\frac{2\hat{\sigma}_j^2}{\hat{f}_j^2}}{\sum_{h=1}^{n-j} C_{hj}} \right]$$

où C_{in} représente les paiements cumulés jusqu'à l'année de développement n relatifs aux

sinistres survenus en i et $\hat{f}_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{ij}}$ est le facteur de Chain Ladder.