Canadian Institute of Actuaries | Institut canadien des actuaires

Document de recherche

Utilisation d'un score de sinistralité dans la modélisation dynamique du nombre de réclamations

Direction de la recherche –
Commission sur les recherches universitaires

Octobre 2019

Document 219105

This document is available in English

Utilisation d'un score de sinistralité dans la modélisation dynamique du nombre de réclamations

Jean-Philippe Boucher et Mathieu Pigeon

Département de mathématiques, UQAM. Montréal, Québec, Canada.

Courriels : *boucher.jean-philippe@uqam.ca* et *pigeon.mathieu.2@uqam.ca*

Résumé

Nous construisons un score de sinistralité en nous basant sur l'approche bonus-malus proposée par Boucher et Inoussa (2014). Ensuite, nous comparons l'adéquation et la capacité prédictive de ce nouveau modèle avec plusieurs autres modèles selon la base de panel. Plus particulièrement, nous étudions en détail un nouveau modèle dynamique basé sur l'approche de Harvey-Fernandès (HF), qui attribue différents poids aux réclamations selon leur date de survenance. Nous montrons que le modèle HF présente de graves lacunes qui limitent son application pratique. En revanche, le modèle bonus-malus ne présente pas ces défauts et possède même plusieurs caractéristiques intéressantes : intelligibilité, avantages du point de vue du calcul et facilité d'utilisation en pratique. Nous croyons que, du fait de sa souplesse, ce nouveau modèle pourrait servir dans bien d'autres contextes actuariels. En nous fondant sur une base de données réelle, nous montrons que le modèle proposé produit le meilleur ajustement et l'une des meilleures capacités prédictives parmi les autres modèles testés.

1. Introduction et motivation

Dans un article récent, Boucher et Inoussa proposait une nouvelle approche de tarification au mérite en assurance automobile qui est fonction du nombre de réclamations. Plutôt que de supposer un effet aléatoire ou une copule pour modéliser la dépendance entre tous les contrats d'un même assuré, les auteurs ont directement inclus un système de bonus-malus (SBM) dans la modélisation. En conséquence, toute l'expérience passée de sinistralité est résumée en une seule valeur numérique : le niveau actuel du SBM, ou ce que nous convenons d'appeler le *score de sinistralité* (semblable au score de solvabilité). En plus de resserrer les liens entre la théorie et ce qui est actuellement utilisé en pratique pour la tarification, l'approche proposée a montré une souplesse intéressante qui permet, par exemple, d'inclure des contraintes juridiques dans la tarification. Toutefois, malgré les avantages de l'approche, la

procédure d'estimation des paramètres proposée par Boucher et Inoussa (2014) est assez laborieuse et est couteuse en ressources informatiques.

Dans le présent document, nous proposons une modification simple du modèle de systèmes de bonus-malus qui utilise des données sous forme longitudinale (modèle SBM-panel). Il s'agit d'une modification de la méthode d'estimation des paramètres qui réduit considérablement la complexité et la durée de la procédure d'estimation. Profitant de cette nouvelle souplesse du modèle, nous procédons à une analyse exhaustive du modèle, dans la mesure où nous ne nous limitons pas dans les valeurs possibles des paramètres du modèle. En plus de montrer certaines propriétés du modèle, nous comparons le modèle SBM-panel à divers modèles de dénombrement avec données de panel, dont un nouveau modèle que nous introduisons. Nous montrons que l'ajustement du modèle SBM-panel est le meilleur parmi les distributions considérées et nous analysons sa puissance de prédiction.

Puisque le modèle SBM-panel possède une propriété markovienne qui en facilite grandement l'utilisation, nous croyons que l'approche bonus-malus pourrait être une option intéressante pour modéliser la structure complexe de l'expérience de sinistralité en actuariat. Par exemple, dans les situations où nous voulons modéliser la dépendance entre des véhicules visés par un même contrat, ou lorsque nous voulons établir un lien entre la fréquence et la gravité des sinistres, le *score de sinistralité* produite par le modèle SBM pourrait être pris en compte. Ce nouveau paradigme pourrait ensuite remplacer des approches complexes comme l'utilisation d'une série d'effets aléatoires corrélés (voir Abdallah et coll., 2016) ou d'une copule hiérarchique (voir Shi et coll., 2016).

Le document est construit de la manière suivante. À la section 2, nous présentons brièvement les principales méthodes de tarification, dans le cadre desquelles nous introduisons une nouvelle distribution dynamique du nombre de réclamations pour des données de panel. À la section 3, nous passons en revue le modèle SBM-panel, nous introduisons une version simplifiée et nous mettons en évidence certaines propriétés du modèle. À la section 4, d'après une base de données d'une société d'assurances IARD, nous estimons et comparons les modèles proposés, puis nous testons un très large éventail de paramètres structurels du modèle SBM-panel. Enfin, nous concluons et présentons des généralisations prometteuses à la section 5.

2. Techniques de tarification

Nous observons sur plusieurs années un portefeuille de M titulaires de contrats d'assurances IARD. Pour chaque contrat i, $i = 1, \dots, M$, nous définissons $N_{i,t}$, une variable aléatoire discrète indiquant le nombre de réclamations sur la période d'assurance t, et $\mathbf{X}_{i,t}$, un vecteur colonne contenant les facteurs explicatifs disponibles au début de la période t. Dans ce vecteur, nous pouvons inclure $d_{i,t}$, une grandeur scalaire mesurant l'exposition au risque. Pour ce projet, nous supposons que le but premier d'un modèle de tarification est de prédire

$$\mathrm{E}\left[N_{i,T_i+1} \middle| \underbrace{N_{i,1}, \dots, N_{i,T_i}}_{\mathbf{N}_{i,T_i}}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,T_i+1}\right],$$

où tous les titulaires sont indépendants et T_i est la dernière période observée pour le titulaire i. Notre principal objectif est d'étudier les forces et les faiblesses des catégories de modèles suivantes :

- Des modèles de données transversales, pour lesquels l'indépendance est supposée entre les contrats annuels d'un titulaire (voir en 2.1);
- Des modèles de données de panel, pour lesquels nous supposons une dépendance entre tous les contrats souscrits par un titulaire (voir en 2.2);
- Des modèles SBM-panel, pour lesquels au moins une partie de l'information passée est résumée par un SBM (voir la section 3).

2.1. Modèles de données transversales

Dans le cas des modèles transversaux, il y a indépendance entre tous les titulaires et entre tous les contrats, de sorte que nous pouvons poser

$$\Pr\left(N_{i,t+1} = n|\mathbf{N}_{i,t}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t+1}\right) = \Pr\left(N_{i,t+1} = n|\mathbf{X}_{i,t+1}\right)$$

et nous pouvons simplifier notre problème de prédiction de la façon suivante :

$$\mathrm{E}\left[N_{i,t+1}|\mathbf{N}_{i,t}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t+1}\right] = \mathrm{E}\left[N_{i,t+1}|\mathbf{X}_{i,t+1}\right] = \lambda\left(\mathbf{X}_{i,t+1}\right),$$

où $\lambda()$ est une fonction. Pour classer les risques, nous faisons traditionnellement l'hypothèse d'une relation log-linéaire entre le paramètre de moyenne et les caractéristiques du titulaire de police ou de la réclamation comme le sexe, l'âge ou l'état civil (voir Denuit et coll., 2007).

Le modèle de base est habituellement la distribution de Poisson, qui fait partie de la famille linéaire exponentielle et qui possède des propriétés statistiques utiles et bien connues (voir McCullagh et Nelder, 1989, ou Frees et coll., 2014). La fonction de masse de probabilité est définie par $\Pr\left(N_{i,t} = n|\mathbf{X}_{i,t}\right) = \left(\lambda_{i,t}\right)^n \exp\left(-\lambda_{i,t}\right)/n!,\ n = 0,1,2,\dots$ et 0 ailleurs, où $\lambda_{i,t} = d_{i,t}\exp\left(\mathbf{X}_{i,t}{}'\boldsymbol{\beta}\right)$ et $\boldsymbol{\beta}$ est un vecteur-colonne contenant les paramètres du modèle. Enfin, nous posons

$$\pi_{i,t+1}^{\mathrm{Poi}} = \mathrm{E}\left[N_{i,t+1}|\mathbf{X}_{i,t+1}\right] = \lambda_{i,t+1}.$$

Cette équation représente une prime annuelle lorsque le coût de chaque réclamation est de 1. Puisque cette prime ne dépend pas de l'expérience passée de sinistralité, nous qualifions habituellement $\pi_{i,t+1}$ de prime *a priori*.

La distribution de Poisson implique l'équidispersion, c'est-à-dire que $\mathrm{E}\left[N_{i,t}|\mathbf{X}_{i,t}\right] = \mathrm{Var}\left[N_{i,t}|\mathbf{X}_{i,t}\right]$, ce qui, habituellement, constitue une hypothèse trop forte pour la tarification en assurances IARD. Afin de surmonter ce problème, nous considérons la distribution binomiale négative (BN), qui est l'une des solutions de rechange les plus couramment utilisées au modèle de Poisson. Pour faciliter l'analyse des données, nous nous limitons aux formes les plus simples de la distribution BN, mais nous vous invitons à consulter Winkelmann (2010). La façon la plus sensée de construire une distribution BN à partir d'une distribution de Poisson consiste à introduire un terme aléatoire d'hétérogénéité dans le paramètre de moyenne. Soit Θ une

variable aléatoire obéissant à une distribution gamma qui a pour fonction de densité de probabilité $f_\Theta(\theta) = \gamma^\alpha \theta^{\alpha-1}\exp(-\theta\gamma)/\Gamma(\alpha)$ pour $\theta > 0$, avec $\alpha > 0$ et $\gamma > 0$. Nous supposons que $\alpha = 1/\tau$ et $\gamma = 1/\tau$, ce qui signifie que $\mathrm{E}[\Theta] = 1$ et $\mathrm{Var}[\Theta] = \tau$. Nous supposons ensuite que $(N_{i,t}|\Theta, \mathbf{X}_{i,t}) \sim \mathrm{Poisson}(\lambda_{i,t}\theta)$. Par conséquent, la variable aléatoire $(N_{i,t}|\mathbf{X}_{i,t})$ suit une distribution BN de type 2 (BN 2) avec paramètres τ et $\lambda_{i,t}$ et fonction de masse de probabilité donnée par

$$\Pr(N_{i,t} = n|\mathbf{X}_{i,t}) = \frac{\Gamma(n + 1/\tau)}{\Gamma(n+1)\Gamma(1/\tau)}\left(\frac{\lambda_{i,t}}{1/\tau + \lambda_{i,t}}\right)^n \left(\frac{1/\tau}{1/\tau + \lambda_{i,t}}\right)^{1/\tau}, \qquad n = 0,1,2,\ldots$$

et 0 ailleurs. Nous pouvons obtenir directement $\mathrm{E}[N_{i,t}] = \lambda_{i,t}$ et $\mathrm{Var}[N_{i,t}] = \lambda_{i,t} + \tau\lambda_{i,t}^2$.

Nous considérons également une version légèrement différente de la distribution BN (BN1) avec paramètres $\lambda_{i,t}$ et τ pour lesquels la fonction de masse de probabilité est définie par

$$\Pr(N_{i,t} = n|\mathbf{X}_{i,t}) = \frac{\Gamma\left(n + \frac{\lambda_{i,t}}{\tau}\right)}{\Gamma(n+1)\Gamma\left(\frac{\lambda_{i,t}}{\tau}\right)}(1+\tau)^{-\lambda_{i,t}/\tau}\left(1 + \frac{1}{\tau}\right)^{-n}, \qquad n = 0,1,2,\ldots$$

et 0 ailleurs. Nous constatons que $\mathrm{Var}[N_{i,t}] = \lambda_{i,t} + \tau\lambda_{i,t} = \phi\lambda_{i,t}$, ce qui correspond à la fonction de variance de la Poisson surdispersée dans le cadre du modèle linéaire généralisé. Dans tous les cas, les paramètres peuvent être estimés aisément par la procédure du maximum de vraisemblance.

La prime prédite par les modèles BN1 et BN2 est donnée par

$$\pi_{i,t+1}^{\mathrm{BN}} = \mathrm{E}[N_{i,t+1}|\mathbf{X}_{i,t+1}] = \lambda_{i,t+1},$$

où nous posons toujours l'hypothèse d'indépendance entre les contrats d'un même assuré.

2.2. Modèles de données de panel

Les modèles de données de panel supposent une certaine dépendance entre tous les contrats annuels appartenant à un seul titulaire. Il nous faut un modèle pour le vecteur aléatoire (conditionnel) $[\mathbf{N}_{i,t}|\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t}]$, $t = 1,2,\ldots$ afin de pouvoir prédire $\mathrm{E}[N_{i,T_i+1}|\mathbf{N}_{i,T_i}, \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,T_i+1}]$. Il existe de nombreux modèles de dépendance temporelle entre les variables aléatoires, par exemple, les modèles conditionnels, les modèles marginaux et les modèles spécifiques, mais il a été démontré que les modèles à effets aléatoires étaient les mieux adaptés aux données en assurances IARD (voir Boucher et Guillén, 2009). Dans un modèle de tarification, un effet aléatoire individuel peut saisir la variabilité causée par le manque d'information sur certaines variables de classification importantes comme la rage au volant et l'usage de drogues. Soit Θ cet effet aléatoire. Conditionnellement à Θ, tous les contrats du même assuré sont supposés indépendants. La fonction de masse de probabilité conjointe est définie par

$$\Pr(N_{i,1} = n_{i,1}, \ldots, N_{i,t} = n_{i,t}|\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t})$$

$$= \int_{-\infty}^{\infty} \Pr\left(N_{i,1} = n_{i,1}, \dots, N_{i,t} = n_{i,t} | \theta_i, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\right) dG(\theta_i | \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}),$$

$$= \int_{-\infty}^{\infty} \left(\prod_{k=1}^{t} \Pr\left(N_{i,k} = n_{i,k} | \theta_i, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\right) \right) dG(\theta_i).$$

où $G(\theta_i)$ est la fonction de répartition de l'effet aléatoire. Nous supposons que la distribution de l'effet aléatoire θ_i ne dépend pas des régresseurs $\mathbf{X}$; voir Boucher et Denuit (2006) pour en savoir plus sur cette hypothèse conventionnelle en actuariat. Enfin, nous notons que la distribution conjointe peut aussi s'exprimer comme le produit de toutes les distributions prédictives de chaque contrat d'assurance de l'assuré i:

$$\Pr\left(N_{i,1} = n_{i,1}, \dots, N_{i,t} = n_{i,t} | \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\right)$$

$$= \Pr\left(N_{i,1} = n_{i,1} | \mathbf{X}_{i,1}\right) \Pr\left(N_{i,2} = n_{i,2} | N_{i,1} = n_{i,1}, \mathbf{X}_{i,1}, \mathbf{X}_{i,2}\right)$$

$$\times \cdots \times \mathbf{Pr}\left(\boldsymbol{N_{i,t} = n_{i,t} | N_{i,1} = n_{i,1}, \dots, N_{i,t-1} = n_{i,t-1}, X_{i,1}, \dots, X_{i,t}}\right)$$

2.2.1. Multinomiale négative

Ressemblant à la distribution BN2, le modèle le plus simple à effets aléatoires est donné par

$$\left(N_{i,t} | \Theta_i = \theta_i, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\right) \sim \text{Poisson}\left(\theta_i \lambda_{i,t}\right)$$

et $\Theta_i \sim \text{amma}\ \ (\alpha = \kappa, \gamma = \kappa)$, ce qui entraîne

$$\Pr\left(N_{i,1} = n | \mathbf{X}_{i,1}\right) = \frac{\Gamma(n+\kappa)}{\Gamma(n+1)\Gamma(\kappa)} \left(\frac{\lambda_{i,1}}{\kappa + \lambda_{i,1}}\right)^n \left(\frac{\kappa}{\kappa + \lambda_{i,1}}\right)^\kappa$$

$$\Pr\left(N_{i,t+1} = n | \mathbf{N}_{i,t}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t+1}\right) = \frac{\Gamma(n+\alpha)}{\Gamma(n+1)\Gamma(\alpha)} \left(\frac{\lambda_{i,t+1}}{\gamma + \lambda_{i,t+1}}\right)^n \left(\frac{\gamma}{\gamma + \lambda_{i,t+1}}\right)^\alpha,$$

avec $\alpha = \kappa + \sum_{k=1}^{t} n_{i,k}$, $\gamma = \kappa + \sum_{k=1}^{t} \lambda_{i,k}$ et $\lambda_{i,k} = d_{i,k} \exp\left(\mathbf{X}_{i,k}' \boldsymbol{\beta}\right)$. Les deux distributions de probabilité ont la forme d'une BN2 (pour tout t) et peuvent donc être désignées simplement par $\text{BN2}_{i,t}(\lambda_{i,t}, \alpha, \gamma)$.

La distribution prédictive tire son origine de la théorie de crédibilité classique (voir Bühlmann et Gisler, 2005), où la distribution *a posteriori* du terme d'hétérogénéité Θ_i est une gamma avec paramètres actualisés $\kappa + \sum_{k=1}^{t} n_{i,k}$ et $\kappa + \sum_{k=1}^{t} \lambda_{i,k}$. La distribution conjointe, appelée distribution BN à plusieurs variables (BNPV), ou multinomiale négative, est souvent utilisée en assurance de dommages. Ici encore, les paramètres peuvent être estimés par la méthode du maximum de vraisemblance. La prédiction recherchée est donc

$$\pi_{i,1}^{\text{BNPV}} = \text{E}\left[N_{i,1} | \mathbf{X}_{i,1}\right] = \lambda_{i,1} \left(\frac{\kappa}{\kappa}\right) = \lambda_{i,1}$$

$$\pi_{i,t+1}^{\mathrm{BNPV}} = \mathrm{E}\left[N_{i,t+1}|\mathbf{N}_{i,t}, \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t+1}\right] = \lambda_{i,t+1}\left(\frac{\kappa + \sum_{k=1}^{t} n_{i,k}}{\kappa + \sum_{k=1}^{t} \lambda_{i,k}}\right).$$

Étant donné que chaque contrat d'assurance d'un même titulaire a les mêmes effets aléatoires, $N_{i,j}$ et $N_{i,t+j}$ sont dépendantes :

$$\mathrm{Cov}\left[N_{i,t}, N_{i,t+j}|\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t+j}\right] = \lambda_{i,t}\lambda_{i,t+j}(1/\kappa), \qquad j > 0.$$

2.2.2. BN à effets aléatoires

Comme le souligne Boucher et coll. (2008), la distribution BN dont les effets aléatoires suivent une loi bêta, convient bien pour modéliser le nombre de réclamations, et nous l'appellerons distribution BNBêta. À partir de la fonction de masse de probabilité de la distribution BN2, nous supposons que $(1/\tau)/(\lambda + 1/\tau) \sim \mathrm{Bêta}(a, b)$. Ensuite, nous avons

$$\Pr\left(N_{i,1} = n|\mathbf{X}_{i,1}\right) = \frac{\Gamma(a+b)\Gamma(a+\lambda_{i,1})\Gamma(b+n)}{\Gamma(a)\Gamma(b)\Gamma(a+b+\lambda_{i,1}+n)}\frac{\Gamma(\lambda_{i,1}+n)}{\Gamma(\lambda_{i,1})\Gamma(n+1)}$$

$$\Pr\left(N_{i,t+1} = n|\mathbf{N}_{i,t}, \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t+1}\right) = \frac{\Gamma(\alpha+\gamma)\Gamma(\alpha+\lambda_{i,t+1})\Gamma(\gamma+n)}{\Gamma(\alpha)\Gamma(\gamma)\Gamma(\alpha+\gamma+\lambda_{i,t+1}+n)}\frac{\Gamma(\lambda_{i,t+1}+n)}{\Gamma(\lambda_{i,t+1})\Gamma(n+1)}$$

avec $\alpha = a + \sum_{k=1}^{t} \lambda_{i,k}$, $\gamma = b + \sum_{k=1}^{t} n_{i,k}$ et $\lambda_{i,k} = d_{i,k}\exp\left(\mathbf{X}_{i,k}'\boldsymbol{\beta}\right)$. Les deux distributions de probabilité ont la forme d'une BNBêta (pour tout t) et peuvent être désignées simplement par $\mathrm{BNB}_{i,t}(\lambda_{i,t}, \alpha, \gamma)$.

Nous pouvons facilement démontrer que $\mathrm{E}\left[N_{i,t}\right] = \lambda_{i,t}(b/(a-1))$ et que

$$\mathrm{Var}\left[N_{i,t}\right] = \lambda_{i,t}\frac{(a+b-1)b}{(a-1)(a-2)} + \lambda_{i,t}^2\left(\frac{(b+1)b}{(a-1)(a-2)} - \frac{b^2}{(a-1)^2}\right)$$

et obtenir une distribution *a posteriori* des effets aléatoires, qui est une distribution bêta avec paramètres actualisés $\sum_t \lambda_{i,t} + a$ et $\sum_t n_{i,t} + b$. Nous avons donc

$$\pi_{i,1}^{\mathrm{BNB}} = \mathrm{E}\left[N_{i,1}|\mathbf{X}_{i,1}\right] = \lambda_{i,1}\left(\frac{b}{a-1}\right)$$

$$\pi_{i,t+1}^{\mathrm{BNB}} = \mathrm{E}\left[N_{i,t+1}|\mathbf{N}_{i,t}, \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t+1}\right] = \lambda_{i,t+1}\left(\frac{b + \sum_{k=1}^{t} n_{i,k}}{a + \sum_{k=1}^{t} \lambda_{i,k} - 1}\right).$$

Enfin, pour ce qui est de la BNPV, il peut être démontré que la covariance entre le nombre de réclamations des contrats annuels est égale à :

$$\mathrm{Cov}\left[N_{i,t}, N_{i,t+j}|\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t+j}\right] = \lambda_{i,t}\lambda_{i,t+j}\left(\frac{b}{a-1}\right)\left(\frac{b+1}{a-2} - \frac{b}{a-1}\right), \qquad j > 0.$$

Bien entendu, en ce qui concerne les modèles d'effets aléatoires, d'autres choix que la BNPV de la BNBêta peuvent être retenus et ces modèles construits.

2.3. Modèles dynamiques de données de panel

À ce stade-ci, il y a lieu de mentionner un inconvénient majeur de l'utilisation d'un modèle

classique à effets aléatoires, comme la BNPV ou la BNBêta. En analysant les primes prédictives, on peut constater que toutes les réclamations passées ont la même importance, c'est-à-dire le même poids, pour prédire les primes futures, ce qui signifie qu'une réclamation qui date de dix ans a la même importance qu'une réclamation qui date d'un an. En pratique, il est généralement reconnu qu'il ne s'agit pas d'un scénario réaliste : le comportement des conducteurs évolue au fil du temps et l'expérience récente devrait avoir une plus grande importance que l'expérience ancienne lorsqu'il s'agit d'estimer le risque d'un conducteur.

Malheureusement, il n'est pas facile d'inclure une telle dynamique temporelle dans un modèle de données de panel. L'approche à effets aléatoires doit supposer un processus aléatoire pour $\Theta_{i,t}$. Par conséquent, les modèles dont les effets aléatoires $\Theta_{i,t}\ \ t = 1, \dots, T_i$ évoluent dans le temps ont besoin d'un entier à T dimensions pour exprimer la distribution conjointe de toutes les réclamations d'un seul titulaire. Il faut donc des procédures numériques complexes ou des méthodes d'inférence approximatives (voyez par exemple Jung et Liesenfeld, 2001). D'autres approches ont été proposées pour inclure un effet dynamique dans les modèles de dénombrement : les modèles de crédibilité évolutifs dans Albrecht (1985), les résidus de Poisson dans Pinquet et coll. (2001) ou, plus récemment, les copules avec la méthode avec gigues dans Shi et Valdez (2016).

Dans le présent document, nous examinons le modèle de Harvey-Fernandes (HF), proposé par Harvey et Fernandes (1989) et appliqué en actuariat par Bolancé et coll. (2007). Soit $\mathcal{H}_{i,t}$ l'historique des réclamations jusqu'à l'instant t du contrat i. Cette approche inclut les effets aléatoires qui se développent au fil du temps suivant une procédure à deux étapes :

- Étape P : la distribution conditionnelle des effets aléatoires $\Theta_{i,t}|\mathcal{H}_{i,t}$ est *prédite*
- Étape U : la distribution est *actualisée* selon une certaine fonction U et $\Theta_{i,t+1}\sim U(\Theta_{i,t}|\mathcal{H}_{i,t})$.

Si nous sélectionnons une distribution conjuguée des effets aléatoires (telle que la gamma pour le modèle BNPV et la bêta pour le modèle BNBêta), $\Theta_{i,t}$ et $\Theta_{i,t}|\mathcal{H}_{i,t}$ devraient être issus de la même distribution. Ainsi, la fonction sélectionnée U peut être appliquée directement aux paramètres de la distribution de $\Theta_{i,t}|\mathcal{H}i,t$ afin d'obtenir la distribution de $\Theta_{i,t+1}$ (**étape U**). Conformément à Bolancé et coll. (2007), nous suivons cette voie et sélectionnons une fonction U pour laquelle les paramètres $\alpha_t,\ \tau_t$ de la distribution postérieure $\Theta_{i,t}|N_{i,t}$ deviendront $\alpha_t^* = \nu\alpha_t$ et $\tau_t^* = \nu\tau_t$ avec valeurs de départ α_0 et τ_0. De cette structure, nous pouvons obtenir la distribution de chaque $\Theta_{i,1}, \dots, \Theta_{i,t}$.

Étant donné que la distribution conjointe peut s'exprimer comme le produit des distributions prédictives, il peut être démontré que (nous retranchons $\mathbf{X}_i$ pour simplifier les choses) :

$$\mathbf{Pr}\big(\boldsymbol{N_{i,1}} = \boldsymbol{n_{i,1}}, \dots, \boldsymbol{N_{i,t}} = \boldsymbol{n_{i,t}}\big) =$$

$$\Pr\big(N_{i,1} = n_{i,1}|\alpha_{i,1}, \gamma_{i,1}\big)\Pr\big(N_{i,2} = n_{i,2}|\alpha_{i,2}, \gamma_{i,2}\big) \times \dots \times \Pr\big(N_{i,t} = n_{i,t}|\alpha_{i,t}, \gamma_{i,t}\big),$$

où seuls les paramètres actualisés sont explicitement mentionnés dans chaque distribution :

$$\alpha_{i,t} = (\nu)^{t-1}\alpha_0 + \sum_{k=1}^{t-1} (\nu)^k n_{i,t-k}$$

$$\gamma_{i,t} = (\nu)^{t-1}\gamma_0 + \sum_{k=1}^{t-1} (\nu)^k \lambda_{i,t-k}. \quad (1)$$

Ainsi, en accord avec Bolancé et coll. (2007), nous pouvons utiliser une BNPV dynamique (désignée par HF-BNPV) en utilisant le produit des distributions $\mathrm{BN2}_{i,t}(\lambda_{i,t}, \alpha_{i,t}, \gamma_{i,t})$ pour $t = 1, \ldots, T_i$. De même, nous pouvons construire une nouvelle distribution dynamique basée sur la distribution BNB (désignée par HF-BNB), en faisant le produit des distributions $\mathrm{BNB}_{i,t}(\lambda_{i,t}, \alpha_{i,t}, \gamma_{i,t})$. Nous pouvons constater que les deux distributions conjointes dynamiques accordent un plus grand poids aux réclamations récentes dans le calcul de la prime prédictive :

$$\begin{aligned}\pi_{i,t+1}^{\mathrm{HF-BNPV}} &= \mathrm{E}\left[N_{i,t+1} | \mathbf{N}_{i,t}, \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t+1}\right] \\ &= \lambda_{i,t+1}\left(\frac{(\nu)^t \kappa + \sum_{k=1}^{t} (\nu)^k n_{i,t-k+1}}{(\nu)^t \kappa + \sum_{k=1}^{t} (\nu)^k \lambda_{i,t-k+1}}\right)\end{aligned}$$

et

$$\begin{aligned}\pi_{i,t+1}^{\mathrm{HF-BNB}} &= \mathrm{E}\left[N_{i,t+1} | \mathbf{N}_{i,t}, \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t+1}\right] \\ &= \lambda_{i,t+1}\left(\frac{(\nu)^t b + \sum_{k=1}^{t} (\nu)^k n_{i,t-k+1}}{(\nu)^t a + \sum_{k=1}^{t} (\nu)^k \lambda_{i,t-k+1} - 1}\right).\end{aligned}$$

À la section 4, nous traitons de la covariance entre le nombre de réclamations des contrats annuels d'un même titulaire.

3. SBM et données de panel

Les SBM ont été introduits dans les procédures de tarification pour plusieurs raisons pratiques (voir Lemaire, 1995, et Denuit et coll., 2007). L'idée de base qui sous-tend les modèles SBM est de résumer l'expérience passée de sinistralité par un *score de sinistralité*, qui est une valeur discrète allant de 1 à s, où 1 représente le risque le plus faible et s, le plus élevé. En pratique, lorsqu'un nouveau titulaire est intégré au portefeuille, la société d'assurance lui attribue à l'entrée une certaine note initiale de sinistralité. Chaque année, selon l'expérience de sinistralité de l'assuré, la note du titulaire se déplacera sur l'échelle de bonus-malus, pour prendre une valeur élevée si les réclamations sont fréquentes et une valeur moindre dans le cas contraire.

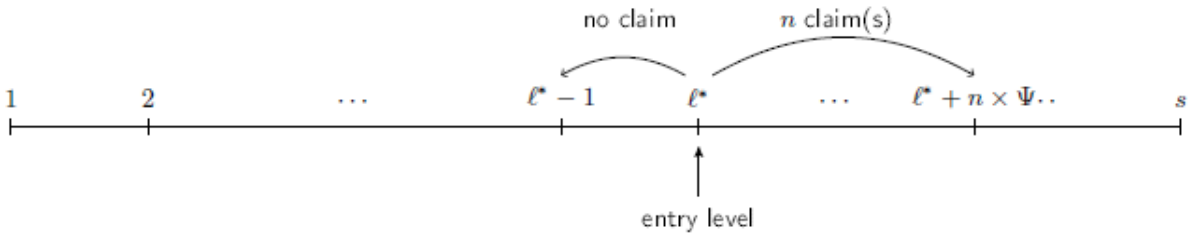
Définition 3.1 (Système de bonus-malus) : *Modèle à trois paramètres (Ψ, s, ℓ^*) pour lequel le niveau du système au début de la période $t+1$ dans la base de données est donné par*

$$L(t+1) = \min\left(\max\left(L(t) - \mathbb{I}\left(N_{i,t} = 0\right) + \Psi N_{i,t}, 1\right), s\right), \quad (2)$$

où $\mathbb{I}\left(N_{i,t} = 0\right)$ est une variable nominale indiquant une période sans réclamation, s désigne le niveau le plus élevé du système et Ψ le *paramètre de saut*. Le paramètre ℓ^* correspond au niveau d'entrée du système dans le cas d'un titulaire sans expérience. Ce type de structure du SBM est souvent désigné par $-1/+\Psi$.

Le graphique 1 présente un exemple de SBM.

Graphique 1 : Exemple de SBM. Au début de la première période, le nouveau titulaire i entre dans le système au niveau ℓ^*. S'il ne déclare aucun accident au cours du premier contrat d'assurance, il passera du niveau ℓ^* au niveau $\ell^* - 1$ au début de la deuxième période. Pour chaque sinistre, le niveau du SBM du titulaire augmente de Ψ.



*Disponible en anglais seulement

Intuitivement, en établissant la moyenne de la fréquence des réclamations pour chaque score de sinistralité, la société d'assurance pourrait ensuite établir une forme de notation au mérite, représentant les relativités pour SBM (voir, par exemple, le graphique 5 de Boucher et Inoussa, 2014). Toutefois, nous pouvons améliorer le système de notation en choisissant d'autres approches que l'établissement d'une moyenne de la fréquence de chaque score de sinistralité. Un bon nombre d'ouvrages traitent de la façon d'étalonner un SBM avec des données d'assurance transversales (voir Denuit et coll., 2007, pour une vue d'ensemble). Toutefois, lorsque nous observons plusieurs contrats annuels d'un seul assuré dans un ensemble de données (c.-à-d. une structure de données de panel), Boucher et Inoussa (2014) présente une approche plus appropriée, où les auteurs construisent ce qu'ils appellent un modèle SBM-panel.

3.1. Modèle SBM-panel

Le modèle SBM-panel a pour principal objectif d'estimer simultanément les paramètres nécessaires à la tarification *a priori* et *a posteriori*, en incluant la structure SBM directement dans la modélisation. Pour estimer les paramètres, il nous faut un modèle pour le vecteur aléatoire conditionnel $[\mathbf{N}_{i,t}|\mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}]$. Tout d'abord, soit $[N_{i,1}, N_{i,2}]$ un vecteur aléatoire dans un portefeuille avec un SBM qui répond à la définition 3.1, et supposons que nous connaissons $\ell_{i,1}$, le niveau du SBM de l'assuré i à l'instant 1. La fonction de masse de probabilité conjointe conditionnelle est donnée par (nous supprimons la condition de dépendance à $\mathbf{X}$ pour simplifier l'exposé) :

$$\Pr(N_{i,1} = n_{i,1}, N_{i,2} = n_{i,2}|L(1) = \ell_{i,1})$$

$$= \Pr(N_{i,1} = n_{i,1}|L(1) = \ell_{i,1})\Pr(N_{i,2} = n \quad i, 2|N_{i,1} = n_{i,1}, L(1) = \ell_{i,1})$$

$$= \Pr(N_{i,1} = n_{i,1}|L(1) = \ell_{i,1})$$

$$\times \left(\sum_{y=1}^{s} \Pr\big(N_{i,2} = n_{i,2}\big|N_{i,1} = n_{i,1}, L(1) = \ell_{i,1}, L(2) = y\big) \Pr\big(L(2) = y\big|N_{i,1} = n_{i,1}, L(1) = \ell_{i,1}\big) \right).$$

En raison du SBM, l'information passée est prise en compte par le dernier niveau atteint par le système et nous obtenons

$$= \Pr(N_{i,1} = n_{i,1}|L(1)$$
$$= \ell_{i,1})\left(\sum_{y=1}^{s} \Pr(N_{i,2} = n_{i,2}|L(2) = y)\Pr(L(2) = y|N_{i,1} = n_{i,1}, L(1) = \ell_{i,1})\right)$$
$$= \Pr(N_{i,1} = n_{i,1}|L(1) = \ell_{i,1})\Pr(N_{i,2} = n_{i,2}|L(2) = \ell_{i,2}),$$

où $\Pr(L(2) = y|N_{i,1} = n_{i,1}, L(1) = \ell_{i,1}) = 0$ pour tout y, sauf pour $y = \ell_{i,2}$, le niveau atteint par le SBM à l'instant 2 après que $n_{i,1}$ réclamations ont été observées durant l'année. Par conséquent,

$$\Pr\left(N_{i,1} = n_{i,1}, N_{i,2} = n_{i,2}, \dots, N_{i,t} = n_{i,t}\middle|L(1) = \ell_{i,1}\right) = \prod_{k=1}^{t} \Pr\left(N_{i,k} = n_{i,k}\middle|L(k) = \ell_{i,k}\right). \quad (3)$$

Il existe plusieurs façons de modéliser la distribution conditionnelle $\Pr\left(N_{i,t} = n_{i,t}|L(t) = \ell_{i,t}\right)$: nous pouvons considérer les distributions introduites en 2.1 ou 2.2, ainsi que les modèles à barrière ou les distributions avec surreprésentation de zéros (voir Boucher et Guillén, 2009, pour une analyse approfondie de l'utilisation des distributions de dénombrement en tarification). Pour une distribution sélectionnée, le niveau du SBM devrait être utilisé pour modéliser le paramètre de moyenne de la distribution conditionnelle. D'après la Poisson, la BN2 ou la BN1 susmentionnées, nous supposons que le paramètre de moyenne sera modélisé comme suit :

$$\pi_{i,t+1}^{\text{SBM}} = \mathrm{E}\left[N_{i,t+1}|\mathbf{N}_{i,t+1}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t+1}\right] = \lambda_{i,t+1} r_{\ell_{i,t+1}},$$

où $\lambda_{i,t}$ est la prime *a priori* basée sur les caractéristiques de l'assuré (sexe, âge, etc.) pour la période $[t, t+1)$. Plusieurs structures peuvent être choisies pour r_ℓ, mais nous limiterons à l'étude des relativités linéaires du SBM, que nous définissons ci-après.

Définition 3.2 (Système de bonus-malus linéaire) : *Une relativité linéaire est associée à chaque degré de l'échelle, selon l'équation*

$$r_{L(t)} = 1 + \delta(L(t) - 1), \qquad (4)$$

où δ est le paramètre de pénalité.

L'équation (4) implique $r_1 = 1$ et définit le risque de base. Ces relativités linéaires, proposées à l'origine par Gilde et Sundt (1989), permettent d'éviter les situations indésirables telles que $r_i > r_j$ pour $i < j$.

3.2. Niveau d'entrée

Les distributions susmentionnées ont été définies en supposant comme condition la connaissance du niveau du SBM de l'assuré i à l'instant 1. Par conséquent, la fonction de masse de probabilité conjointe du titulaire i à l'instant t est donnée par

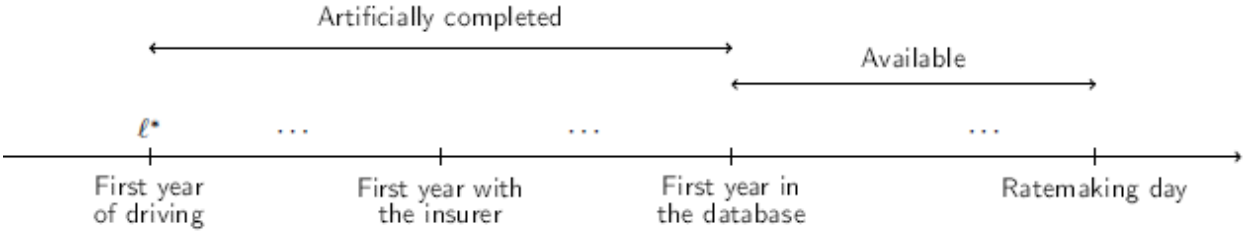
$$\Pr\left(N_{i,1} = n_{i,1}, \ldots, N_{i,t} = n_{i,t} | \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t}\right)$$

$$= \sum_{y=1}^{s} \Pr\left(N_{i,1} = n_{i,1}, \ldots, N_{i,t} = n_{i,t} | L(1) = y, \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t}\right) \Pr(L(1) = y)$$

$$= \sum_{y=1}^{s} \prod_{k=1}^{t} \Pr(N_{i,k} = n_{i,k} | L(k) = \ell_{i,k}, \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,k}) \Pr(L(1) = y),$$

où $\Pr(L(1) = y)$ est la distribution de probabilité du niveau du SBM à l'instant $t = 1$, soit la première année où un assuré figure dans la base de données. Il ne faut pas confondre la première année de conduite avec la première année où un assuré figure dans la base de données, comme le montre le graphique 2.

Graphique 2 : Exemple d'un titulaire pour lequel l'information est partiellement inconnue. Pour estimer ℓ^*, l'assureur doit compléter la trajectoire suivie par le SBM entre la première année de conduite et la première année dans la base de données.



*Disponible en anglais seulement.

Si nous voulons modéliser la distribution conjointe pour un nouveau conducteur, $\ell_{i,1}$ peut être obtenu directement : il s'agit de ℓ^*, le niveau d'entrée sélectionné pour la construction du modèle. Ainsi, $\Pr(L(1) = y) = 0$ pour tout y, sauf $y = \ell^*$. En ce qui concerne les conducteurs d'expérience, cette situation est plus difficile à résoudre. Par précaution, les assureurs ne devraient pas supposer que tous les nouveaux assurés n'ont pas enregistré de réclamations au cours des années précédentes, ni supposer automatiquement qu'ils doivent attribuer aux titulaires un niveau d'entrée ℓ^*. Dans Boucher et Inoussa (2014), pour créer la distribution de L_1, les auteurs proposent de recréer tous les événements possibles de chaque titulaire, de la première année de conduite à la première année où il figure dans la base de données. En calculant la moyenne de L_1, ils attribuent à chaque conducteur dans la base de données une valeur de ℓ_1. Bien que la méthode ait du sens, sa mise en œuvre est très complexe et nécessite beaucoup de temps et de ressources informatiques. Dans le présent document, nous proposons une méthode beaucoup plus simple qui repose sur le fait que la probabilité de ne pas déclarer un accident au cours d'une année est très élevée, souvent de l'ordre de 80 % à 95 %.

Proposition 3.3 *: Pour un titulaire i, soit $S_i(t)$ le niveau atteint par un SBM (Ψ, s, ℓ^*) après t périodes à partir de la première année d'exposition, et u_i la première année observée dans le portefeuille. Ainsi, si*

$$\Pr\left(N_{i,t} = 0 | L(t) = \ell_{i,t}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\right) > 0{,}5, \qquad \forall t, \text{ (5)}$$

alors

$$\underset{\ell \in \{1,\dots,s\}}{\operatorname{argmax}} \left(\Pr\left(S_i(1) = \ell^*, \dots, S_i(u_i+1) = \ell | \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\right)\right) = \max(\ell^* - u_i, 1).$$

Démonstration. Selon l'équation (2), les probabilités de transition sont données par (dans le cas d'un titulaire dont la fréquence des sinistres est de $\lambda_{i,t}$ pour la période t)

$$p_{k,j}(\lambda_{i,t}) = \Pr\left(L(t+1) = j | L(t) = k, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\right)$$

$$= \Pr\left(\min\left(\max\left(k - \mathbb{I}\left(N_{i,t} = 0\right) + \Psi N_{i,t}, 1\right), s\right) = j | \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\right), \qquad \text{(6)}$$

pour $j = 1, \dots, s$ $k = 1, \dots, s$, ce qui implique que $p_{k,\max(k-1,1)}(\lambda_{i,t}) > 0{,}5$ et $p_{k,j}(\lambda_{i,t}) < 0{,}5$, $j \neq \max(k-1,1)$ si $\Pr\left(N_{i,t} = 0 | L(t) = \ell_{i,t}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\right) > 0{,}5$. Même si la société d'assurance ne connaît pas les caractéristiques des titulaires avant qu'ils n'intègrent le portefeuille, il est raisonnable de supposer que celles-ci sont telles que l'équation (5) est vérifiée.

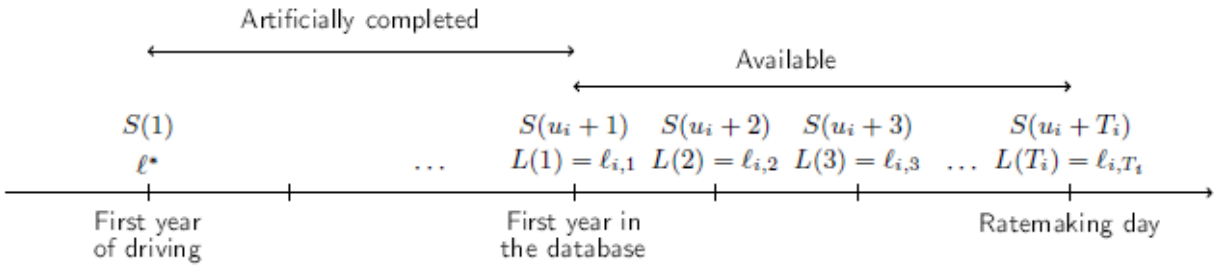
Donc,

$$p_{\ell^*,\max(\ell^*-1,1)}(\lambda_{i,1}) p_{\max(\ell^*-1,1),\max(\ell^*-2,1)}(\lambda_{i,2}) \times \dots \times p_{\max(\ell^*-u_i+1,1),\max(\ell^*-u_i,1)}(\lambda_{i,u_i})$$

est le chemin le plus probable du niveau ℓ^* au niveau $\max(\ell^* - u_i, 1)$. Enfin,

$$\underset{\ell \in \{1,\dots,s\}}{\operatorname{argmax}} \left(\Pr(S_i(1) = \ell^*, \dots, S_i(u_i+1) = \ell)\right) = \max(\ell^* - u_i, 1).$$

Graphique 3 : Exemple d'un titulaire pour lequel l'information est partiellement inconnue (u_i années sont inconnues et T_i années sont connues).



*Disponible en anglais seulement.

Le graphique 3 illustre cette situation et les nouvelles variables introduites. Au vu du résultat de la dernière proposition, nous choisissons ensuite $\max(\ell^* - u, 1)$ comme niveau d'entrée d'un

nouveau titulaire avec u années d'expérience; en d'autres termes, $\Pr(L(1) = y) = 0$ pour tout y, sauf $y = \max(\ell^* - u, 1)$. Par conséquent, chaque année d'expérience de conduite entraîne une diminution d'un niveau du SBM, ce qui est déjà l'une des façons dont les sociétés d'assurance traitent les conducteurs d'expérience dans les régimes avec tarification au mérite. Donc, cette façon de sélectionner $\ell_{i,1}$ non seulement simplifie le modèle SBM-panel, mais en plus elle justifie du point de vue théorique la procédure déjà en usage.

3.3. Paramètres

Maintenant que la distribution conjointe des N_{i,T_i} est entièrement définie pour tous les assurés dans la base de données, nous pouvons résumer les étapes nécessaires à l'estimation de tous les paramètres du modèle. Tout d'abord, l'actuaire doit sélectionner ou estimer les composantes structurelles suivantes du modèle SBM-panel :

- Le nombre de niveaux s du système;
- Le paramètre de saut Ψ pour chaque réclamation;
- Le niveau d'entrée ℓ^* pour un nouveau conducteur;
- Les caractéristiques de risque $\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,T_i}$ et les paramètres $\boldsymbol{\beta}$ associés;
- La distribution de dénombrement sous-jacente.

Ensuite, pour ce modèle SBM-panel particulier, N_{i,T_i}, les paramètres *a priori* ($\boldsymbol{\beta}$) qui relient les covariables à la valeur attendue et au *paramètre de pénalité* δ doivent être estimés par maximisation de la fonction de vraisemblance. Lorsque les composantes structurelles du modèle SBM-panel sont sélectionnées, l'estimation des paramètres s'effectue directement au moyen des logiciels statistiques standards. Les autres paramètres liés à la distribution de dénombrement sous-jacente devraient aussi faire l'objet d'une estimation. Par exemple, si on utilise une distribution BN1 ou BN2, il faudrait aussi trouver un estimateur du paramètre de surdispersion α.

À noter que, en ce qui concerne les paramètres s, Ψ et ℓ^*, la structure du modèle nous donne des informations additionnelles nous permettant de réduire l'espace des paramètres à un treillis. Il s'agit d'un problème bien connu en statistique, voyez par exemple Hammersley (1950) pour les cas simples ou Choirat et Rafaello (2012) pour les propriétés asymptotiques. Pour obtenir le meilleur modèle SBM-panel, nous ajustons toutes les combinaisons de s, Ψ et ℓ^* et nous sélectionnons le modèle qui génère le meilleur profil de vraisemblance ou la meilleure capacité prédictive (sur la base d'une analyse hors échantillon).

3.4. Propriétés du modèle SBM-panel

Entre autres propriétés du modèle SBM-panel, nous souhaitons évaluer la covariance entre les primes versées par un titulaire sur deux périodes. Pour ce faire, il est nécessaire d'étudier la covariance entre N_t et N_{t+k}, $k = 1,2, \ldots$, sous la condition que $L(t) = \ell_t$.

Pour un modèle SBM-panel, la probabilité à un an que la variable aléatoire L passe du niveau du SBM $\ell_{i,t}$ au niveau du SBM $\ell_{i,t+1}$ est désignée par $p_{\ell_{i,t},\ell_{i,t+1}}(\lambda_{i,t})$, qui est définie pour l'équation (6). Pour un titulaire, nous pouvons construire la matrice de probabilités de transition :

$$\mathbf{P}(\lambda_{i,t}) = \begin{bmatrix} p_{1,1}(\lambda_{i,t}) & p_{1,2}(\lambda_{i,t}) & \cdots & p_{1,s}(\lambda_{i,t}) \\ p_{2,1}(\lambda_{i,t}) & p_{2,2}(\lambda_{i,t}) & \cdots & p_{2,s}(\lambda_{i,t}) \\ \vdots & \vdots & \ddots & \vdots \\ p_{s,1}(\lambda_{i,t}) & p_{s,2}(\lambda_{i,t}) & \cdots & p_{s,s}(\lambda_{i,t}) \end{bmatrix}.$$

Nous pouvons démontrer que, pour tout $\boldsymbol{K} = \mathbf{1}, \mathbf{2}, \ldots$, nous avons

$$\mathbf{P}^{(K)}(\lambda_{i,t}) = \mathbf{P}^{K}(\lambda_{i,t}),$$

ce qui signifie que la matrice de probabilités de transition sur K périodes est simplement la puissance K^{e} de la matrice annuelle de probabilités de transition $\mathbf{P}(\lambda_{i,t})$.

Proposition 3.4 : *Dans un modèle SBM-panel, au début d'une période t, la covariance conditionnelle entre $N_{i,t}$ et $N_{i,t+j}$, $j = 1,2, \ldots$ est donnée par*

$$\mathbf{Cov}\left[\boldsymbol{N_{i,t}}, \boldsymbol{N_{i,t+j}} | \boldsymbol{\ell_{i,t}}, \mathbf{X_{i,1}}, \ldots, \mathbf{X_{i,t}}\right] = \boldsymbol{\lambda_{i,t+j}}$$

$$\times \sum_{m=1}^{s} r_m \left(\mathrm{E}\left[N_{i,t} p^{(j-1)}_{\min(\max(\ell_{i,t} - \mathbb{I}(N_{i,t}=0) + \Psi N_{i,t}, 1), s), m}(\lambda_{i,t}) | \ell_{i,t}, \mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,t} \right] - \lambda_{i,t} r_{\ell_{i,t}} p^{(j)}_{\ell_{i,t},m}(\lambda_{i,t}) \right),$$

où les probabilités de transition sont extraites des matrices de transition $\mathbf{P}(\lambda_{i,t})^{(j)}$ et $\mathbf{P}(\lambda_{i,t})^{(j-1)}$ en supposant que $\mathbf{X}_{i,t} = \mathbf{X}_{i,t+1} = \mathbf{X}_{i,t+2} = \cdots$.

Démonstration. Nous avons (après suppression de la mention de i et de $\mathbf{X}_{i,1}$, …, $\mathbf{X}_{i,t}$ afin d'alléger l'exposé)

$$\mathrm{E}[N_t | \ell_t] = \lambda_t r_{\ell_t}$$

$$\mathrm{E}\left[N_{t+j} | \ell_t\right] = \sum_{n=0}^{\infty} n \Pr\left(N_{t+j} = n | \ell_t\right)$$

$$= \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} n \Pr\left(N_{t+j} = n | L(t+j) = m, \ell_t\right) \Pr(L(t+j) = m | \ell_t).$$

La somme interne est

$$\sum_{n=0}^{\infty} n \Pr\left(N_{t+j} = n | L(t+j) = m, \ell_t\right) = \sum_{n=0}^{\infty} n \Pr\left(N_{t+j} = n | L(t+j) = m\right)$$

$$= \mathrm{E}\left[N_{t+j} | L(t+j) = m\right]$$

$$= \lambda_{t+j} r_m.$$

Donc,

$$\mathbf{E}\left[\boldsymbol{N_{t+j}}|\boldsymbol{\ell_t}\right] = \sum_{\boldsymbol{m=0}}^{\infty} \boldsymbol{\lambda_{t+j} r_m} \mathbf{Pr}(\boldsymbol{L(t+j)} = \boldsymbol{m}|\boldsymbol{\ell_t})$$

$$= \sum_{m=0}^{\infty} \lambda_{t+j} r_m \sum_{q_1=0}^{\infty} \Pr(L(t+j) = m|L(t+j-1) = q_1, \ell_t) \Pr(L(t+j-1) = q_1|\ell_t)$$

Et, en supposant que $\mathbf{X_{i,t}} = \mathbf{X_{i,t+1}} = \mathbf{X_{i,t+2}} = \cdots$, nous avons

$$= \lambda_{t+j} \sum_{m=0}^{\infty} r_m \sum_{q_1=0}^{\infty} p_{q_1,m}(\lambda_{i,t}) \Pr(L(t+j-1) = q_1|\ell_t).$$

Par récursivité, nous obtenons :

$$= \lambda_{t+j} \sum_{m=0}^{\infty} r_m \sum_{q_1=0}^{\infty} p_{q_1,m}(\lambda_{i,t}) \cdots \sum_{q_{j-1}=0}^{\infty} p_{q_{j-1},q_{j-2}}(\lambda_{i,t}) p_{\ell_t,q_{j-1}}(\lambda_{i,t}),$$

et

$$= \lambda_{t+j} \sum_{m=0}^{\infty} r_m p_{\ell_t,m}^{(j)}(\lambda_{i,t}).$$

$$\mathbf{E}\left[\boldsymbol{N_t N_{t+j}}|\boldsymbol{\ell_t}\right] = \sum_{\boldsymbol{n_0=0}}^{\infty} \sum_{\boldsymbol{n_j=0}}^{\infty} \boldsymbol{n_0 n_j} \mathbf{Pr}\left(\boldsymbol{N_t} = \boldsymbol{n_0}, \boldsymbol{N_{t+j}} = \boldsymbol{n_j}|\boldsymbol{\ell_t}\right)$$

$$= \sum_{n_0=0}^{\infty} \sum_{n_j=0}^{\infty} n_0 n_j \Pr(N_t = n_0|\ell_t) \Pr\left(N_{t+j} = n_j|\ell_t, N_t = n_0\right)$$

$$= \sum_{n_0=0}^{\infty} \sum_{n_j=0}^{\infty} n_0 n_j \Pr(N_t = n_0|\ell_t) \sum_{m=0}^{\infty} \Pr\left(N_{t+j} = n_j|L(t+j) = m, \ell_t, N_t = n_0\right)$$

$$\times \Pr(L(t+j) = m|\ell_t, N_t = n_0)$$

$$= \sum_{n_0=0}^{\infty} \sum_{n_j=0}^{\infty} \sum_{m=0}^{\infty} n_0 n_j \Pr(N_t = n_0|\ell_t) \Pr\left(N_{t+j} = n_j|L(t+j) = m\right)$$

$$\times \sum_{q_1=0}^{\infty} \Pr(L(t+j) = m|L(t+j-1) = q_1) \Pr(L(t+j-1) = q_1|\ell_t, N_t = n_0).$$

Par récursivité, nous obtenons :

$$= \sum_{n_0=0}^{\infty} \sum_{n_j=0}^{\infty} \sum_{m=0}^{\infty} n_0 n_j \Pr(N_t = n_0 | \ell_t) \Pr\big(N_{t+j} = n_j | L(t+j) = m\big)$$

$$\times p^{(j-1)}_{\min(\max(\ell_t - \mathbb{I}(n_0=0) + \Psi n_0, 1), s), m}(\lambda_{i,t})$$

$$= \lambda_{t+j} \sum_{n_0=0}^{\infty} \sum_{m=0}^{\infty} r_m n_0 \Pr(N_t = n_0 | \ell_t) p^{(j-1)}_{\min(\max(\ell_t - \mathbb{I}(n_0=0) + \Psi n_0, 1), s), m}(\lambda_{i,t})$$

$$= \lambda_{t+j} \sum_{m=0}^{\infty} r_m \mathrm{E}\left[N_t p^{(j-1)}_{\min(\max(\ell_t - \mathbb{I}(N_t=0) + \Psi N_t, 1), s), m}(\lambda_{i,t}) | \ell_t \right]$$

et le résultat découle directement de la définition de covariance.

Pour $j = 1$, c'est-à-dire pour deux périodes consécutives, le résultat de la proposition 3.4 se réduit à :

$$\mathrm{Cov}\big[N_{i,t}, N_{i,t+1} | \ell_{i,t}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\big] = \sum_{n=0}^{\infty} \sum_{q=0}^{\infty} nq \Pr\big(N_{i,t} = n | \ell_{i,t}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\big)$$

$$\times \Pr\big(N_{i,t+1} = q | \ell_{i,t+1}, \mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,t}\big)$$

$$-\lambda_{i,t} \lambda_{i,t+1} r_{\ell_{i,t}} \sum_{m=1}^{s} r_m p_{\ell_{i,t}, m}(\lambda_{i,t}),$$

avec $\ell_{i,t+1} = \ell_{i,t} - \mathbb{I}(n = 0) + n\Psi$.

4. Illustrations empiriques

Nous analysons une base de données d'un portefeuille de polices d'assurance de responsabilité civile générale pour particuliers d'une grande société d'assurance canadienne. Puisque les systèmes de bonus-malus sont généralement conçus pour les produits d'assurance automobile, nous ne tenons compte dans notre projet que de l'utilisation privée des véhicules. L'échantillon contient de l'information au sujet de 429 333 contrats de 140 714 titulaires de polices et recouvre les années 2012 à 2016. L'échantillon possède les propriétés suivantes :

1. Nous ne conservons que les titulaires de polices ayant au plus un véhicule assuré dans tous les contrats observés leur appartenant, afin d'éviter toute dépendance à l'intérieur des contrats;
2. Nous ne conservons que les titulaires de polices bénéficiant de couvertures complètes dans tous les contrats observés leur appartenant, afin d'éviter la censure lorsque des sinistres ne sont pas couverts ou observés.
3. Nous supprimons de la base de données les titulaires de polices qui utilisent leur voiture à des fins commerciales, car ils peuvent avoir des comportements de conduite différents.

Ces critères de sélection pourraient avoir une incidence sur les résultats. Par exemple, en ne choisissant que les assurés qui ont une seule voiture, nous pourrions avoir une plus grande proportion d'assurés seuls qu'à l'habitude. Toutefois, la présente étude n'a pas pour but d'expliquer le processus d'accident par des covariables, ni le processus d'indemnisation, mais seulement de présenter des modèles de dénombrement intéressants qui peuvent être utiles à la tarification. Pour évaluer la performance de nos modèles, nous avons divisé la base de données en deux parties : un ensemble pour l'ajustement (98 589 titulaires de polices) et un ensemble pour la validation (42 125 titulaires de polices). Le tableau 1 décrit les huit covariables sélectionnées dans la modélisation. Pour chaque contrat, nous disposons de l'information initiale au début de la période et nous souhaitons prédire le nombre de réclamations (à l'exclusion des accidents sans collision ni versement). La fréquence moyenne des réclamations est d'environ 6,5 % et nous observons un maximum de cinq réclamations par contrat. La distribution du nombre observé de contrats par assuré est indiquée au tableau 2, pour une moyenne de 3,05 contrats par titulaire de police.

Tableau 1 : Variables dichotomiques dans la base de données

Variable	Description
$X_1 = 1$	Le titulaire est une femme.
$X_2 = 1$	Le titulaire est marié.
$X_3 = 1$	Le titulaire est âgé de moins de 30 ans.
$X_4 = 1$	Le titulaire est âgé entre 30 et 50 ans.
$X_5 = 1$	La distance annuelle parcourue en voiture est inférieure à 10 000 km.
$X_6 = 1$	La distance annuelle parcourue en voiture se situe entre 10 000 km et 20 000 km.
$X_7 = 1$	La distance annuelle parcourue en voiture se situe entre 20 000 km et 30 000 km.
$X_8 = 1$	Le véhicule est utilisé pour se rendre au travail.

Tableau 2 : Distribution du nombre de contrats

Nombre d'années	Pourcentage
1	21,42 %
2	17,73 %
3	11,66 %
4	32,57 %
5	16,62 %

4.1. Distributions de dénombrement

Au tableau 3, nous présentons les résultats de l'ajustement de divers modèles. Nous observons que la modification, par Harvey-Fernandes, des distributions BNPV et BNBêta n'a pas amélioré l'ajustement, ce qui a donné une valeur de $\boldsymbol{\nu = 1}$ pour les deux modèles. Nous supposons que cela s'explique par l'existence d'une moyenne de 3,05 contrats par titulaire de police, où le poids des réclamations passées peut encore être supposé équivalent pour la prédiction des nouvelles réclamations.

Tableau 3 : Résultats d'ajustement

Modèle	Nombre de paramètres	Log-vraisemblance	Critère d'information d'Akaike (CIA)	Critère d'information bayésienne (CIB)
Poisson	9	$-85526{,}64$	171071,28	171280,34
BN2	10	$-85238{,}04$	170496,08	170728,37
BN1	10	$-85149{,}08$	170318,16	170550,45
BNPV	10	$-84812{,}54$	169645,08	169877,37
BNBêta	11	$-84727{,}13$	169476,26	169731,78
HF-BNPV	11	$-84812{,}54$	169647,08	169902,60
HF-BNBêta	12	$-84727{,}13$	169478,26	169757,01
BNPV *	10	$-84611{,}88$	169243,76	169476,05
BNBêta *	11	$-84426{,}63$	168875,26	169130,78
HF-BNPV *	11	$-84460{,}82$	168943,64	169199,16
HF-BNBêta *	12	$\mathbf{-84355{,}07}$	$\mathbf{168734{,}14}$	$\mathbf{169012{,}89}$

Nous pouvons observer deux résultats pour chaque distribution du dénombrement de données de panel (avec et sans *), du fait de l'ajout d'informations supplémentaires dans notre base de données. En effet, la province de l'Ontario utilise Autoplus, une base de données qui fournit un historique détaillé des polices et des réclamations en assurance automobile. Les assureurs ont donc la possibilité de retracer l'historique des réclamations d'un titulaire de police jusqu'à dix ans. Pour construire la base de données de ce projet, nous sommes donc en mesure de connaître l'expérience de sinistralité des dix dernières années de chaque assuré (avant que leur premier contrat d'assurance ne figure dans la base de données). Toutefois, seule l'expérience de sinistralité de chaque assuré est disponible, car les caractéristiques de risque (X) relatives à ces dix dernières années sont inconnues. Puisque nous utiliserons cette information dans le modèle SBM-panel, nous avons tenté d'inclure aussi cette expérience passée dans les modèles de données de panel. Nous proposons donc une version légèrement modifiée du modèle BNPV et du modèle BNBêta, que nous désignons par BNPV * et BNBêta *. Plus précisément, nous supposons que la distribution de la variable aléatoire Θ_i a déjà été adaptée pour tenir compte des réclamations passées. Après un développement similaire à celui décrit précédemment, nous pourrions démontrer que dans le cas d'un titulaire de police qui compte m années d'expérience avant son entrée dans la base de données, la distribution de Θ_i pour le premier contrat observé d'un assuré (une gamma dans le cas de la BNPV, une bêta dans celui de la BNBêta) aura les paramètres suivants :

$$\alpha^* = \alpha + \sum_{j=1}^{m_i} n^*_{i,-j} \quad \text{et} \quad \gamma^* = \gamma + \sum_{j=1}^{m_i} \lambda_{i,-j},$$

où m_i est la valeur minimale entre dix et l'expérience de conduite (en années) de l'assuré i, $n^*_{i,-t}$ est le nombre observé de réclamations t année(s) avant l'entrée du titulaire de police dans la base de données. Étant donné que nous n'observons pas les $\lambda_{i,-j}$, $j = 1, \dots, m_i$, nous les estimons par $\overline{\lambda} = 6{,}5\ \%$, la fréquence moyenne de la base de données.

Enfin, il peut être démontré que la prime prédite pour la BNPV * est égale à

$$\pi^{\text{BNPV}*}_{i,t+1} = \lambda_{i,t+1}\left(\frac{\sum_{k=1}^{t} n_{i,k} + \alpha^*}{\sum_{k=1}^{t} \lambda_{i,k} + \gamma^*}\right)$$

$$= \lambda_{i,t+1}\left(\frac{\sum_{j=1}^{m_i} n^*_{i,-j} + \sum_{k=1}^{t} n_{i,k} + \alpha}{m_i\lambda + \sum_{k=1}^{t} \lambda_{i,k} + \gamma}\right), \qquad (7)$$

où toute l'expérience de sinistralité disponible peut être utilisée pour estimer la prime future. La prime prédite pour la BNBêta * peut aussi être calculée facilement. À noter que les approches modifiées de Harvey-Fernandes (HF-BNPV * et HF-BNBêta *) peuvent aussi être construites suivant la même procédure, avec

$$\alpha^*_{i,1} = (v)^{m_i}\alpha_0 + \sum_{k=1}^{m_i} (v)^k n_{i,-k} \text{ et } \gamma^*_{i,1} = (v)^{m_i}\gamma_0 + \sum_{k=1}^{m_i} (v)^k \lambda_{i,-k}.$$

Lorsque nous incluons cette nouvelle information, la HF-BNBêta * fait mieux que toutes les autres distributions du point de vue des statistiques d'ajustement, même si nous considérons des critères pénalisés comme le CIA ou le CIB. Nous utilisons aussi les données hors échantillon pour comparer les modèles. Le tableau 4 montre les résultats, nous calculons deux mesures pour évaluer la capacité prédictive de chaque modèle : l'erreur quadratique moyenne (EQM) et − puis que nous avons affaire à des données de dénombrement et non à une distribution continue − une statistique de log-vraisemblance tirée de la distribution de Poisson. Selon les deux statistiques hors échantillon, la HF-BNPV * semble offrir la meilleure capacité prédictive.

Tableau 4 : Statistiques hors échantillon

Modèle		Log-vraisemblance (Poisson)	EQM
Poisson		$-36966{,}48$	$10609{,}54$
BN2		$-36860{,}55$	$10611{,}03$
BN1		$-36815{,}96$	$10609{,}53$
BNPV		$-36731{,}55$	$10567{,}37$
BNBêta		$-36830{,}10$	$10601{,}31$
HF-BNPV		$-36731{,}55$	$10567{,}37$
HF-BNBêta		$-36830{,}10$	$10601{,}31$
BNPV *		$-36588{,}47$	$10552{,}34$
BNBêta *		$-36628{,}14$	$10564{,}14$
HF-BNPV *		$\mathbf{-36543{,}69}$	$\mathbf{10543{,}84}$
HF-BNBêta *		$-36628{,}04$	$10570{,}82$

4.2. SBM et données de panel

Nous ajustons un modèle de données de panel à un SBM avec une Poisson, une BN1 et une BN2 pour distributions sous-jacentes. Comme nous l'avons fait pour les modèles précédents marqués d'un *, nous considérons aussi l'expérience de sinistralité passée d'*Autoplus*. Cet historique sur 10 ans des réclamations passées nous permet de trouver ℓ_1, le niveau du SBM de chaque assuré lorsqu'il est observé dans la base de données pour la première fois. Nous choisissons plusieurs combinaisons de paramètres de structure sur une grille définie par $s = 2,3,\dots,S$, $\Psi = 1,2,\dots,s$ et $\ell^* = 1,\dots,s$, où $S = 22$ pour la distribution de Poisson, $S = 16$ pour la distribution BN1 et $S = 15$ pour la distribution BN2. Ces valeurs de S recouvrent entre 1 000 et 3 000 possibilités pour chaque distribution sous-jacente. Puisque chaque étape d'estimation prend au moins de deux à cinq minutes sur un ordinateur personnel, le traitement de toutes les possibilités prend beaucoup de temps. Nous recherchons une procédure qui pourrait nous aider à rétrécir l'espace de $\{s, \Psi, \ell^*\}$ pour trouver les meilleures combinaisons d'une distribution sous-jacente particulière.

Tableau 5 : Statistiques des données SBM-panel.

Distribution	Nombre de param.	Ψ	s	ℓ^*	Log-vraisemblance	CIA	CIB
De Poisson	13	6	11	1	$-84672{,}04$	$169370{,}08$	$169508{,}07$
	13	6	11	2	$-84672{,}63$	$169371{,}26$	$169509{,}25$
	13	6	10	1	$-84673{,}27$	$169372{,}54$	$169510{,}53$
	13	6	10	2	$-84673{,}89$	$169373{,}78$	$169511{,}77$
BN1	14	6	11	1	$\mathbf{-84331{,}55}$	$\mathbf{168691{,}11}$	$\mathbf{168839{,}71}$
	14	6	10	1	$-84331{,}92$	$168691{,}84$	$168840{,}45$
	14	6	11	2	$-84332{,}13$	$168692{,}27$	$168840{,}87$
	14	6	10	2	$-84332{,}52$	$168693{,}05$	$168841{,}65$
BN2	14	6	11	1	$-84442{,}99$	$168913{,}97$	$169062{,}58$
	14	6	11	2	$-84443{,}56$	$168915{,}13$	$169063{,}73$
	14	6	10	1	$-84444{,}49$	$168916{,}98$	$169065{,}58$
	14	6	10	2	$-84445{,}09$	$168918{,}19$	$169066{,}79$

Pour chaque modèle, nous estimons $\boldsymbol{\beta}$, δ et un paramètre de surdispersion (pour les deux distributions BN). Enfin, comme nous l'avons fait pour les autres distributions, nous calculons la valeur de l'erreur quadratique moyenne ainsi que la log-vraisemblance d'une distribution de Poisson pour chaque modèle ajusté sur la base de l'échantillon de test, dans les deux cas pour prévenir un surajustement. Dans le tableau 5, nous présentons les résultats des quatre meilleurs modèles SBM-panel pour les trois distributions sous-jacentes, évalués sur l'ensemble de données d'estimation. Nous observons que le modèle $-1/+6$ avec $s = 11$ est le meilleur modèle pour la Poisson, la BN1 et la BN2. Les quatre meilleurs modèles pour chaque distribution sous-jacente ont les mêmes paramètres structurels. Seul le deuxième meilleur modèle pour la BN1 se classe au troisième rang pour la Poisson et la BN2. La valeur de δ ne semble pas dépendre de la distribution sous-jacente, mais plutôt des trois paramètres structurels s, Ψ et ℓ^*. Enfin, nous pouvons voir que la distribution BN1 sous-jacente offre de meilleures statistiques d'ajustement que la BN2, ce qui correspond à ce que nous avons déjà observé pour les distributions de données transversales. Pour la forme, des tests statistiques pourraient être effectués pour déterminer si le paramètre de surdispersion de la BN1 et la BN2 est statistiquement significatif, mais les très gros écarts entre les log-vraisemblances montrent déjà que c'est le cas. Comme nous l'avons fait précédemment, nous utilisons aussi des statistiques hors échantillon pour évaluer la capacité prédictive des modèles SBM. Le tableau 6 présente les résultats.

Tableau 6 : Statistiques hors échantillon pour le modèle SBM-panel

Distribution	s	Ψ	ℓ^*	EQM	Log-vraisemblance (Poisson)
Poisson	11	6	1	10550,35	−36587,65
	11	6	2	10550,41	−36587,94
	10	6	1	10549,90	−36586,96
	10	6	2	10549,91	−36587,22
BN1	11	6	1	10550,41	−36587,57
	10	6	1	10549,93	−36586,85
	11	6	2	10550,42	−36587,80
	10	6	2	10549,95	−36587,11
BN2	11	6	1	10552,06	−36586,44
	11	6	2	10552,14	−36586,74
	10	6	1	10551,54	−36585,73
	10	6	2	10551,61	−36586,04

Le meilleur modèle sélectionné pour l'ensemble de données d'estimation est la BN1, $-1/+6$ avec $s = 11$ et $\ell^* = 1$. Les statistiques hors échantillon montrent que ce modèle possède une capacité prédictive intéressante et que seule la HF-BNPV * fait mieux que lui. Nous voyons aussi que le modèle de données SBM-panel donne des prédictions stables, tandis que les 15 autres modèles SBM-panel sélectionnés montrent des statistiques hors échantillon similaires (le plus gros écart est inférieur à 0,005 %).

Enfin, puisque le meilleur modèle sélectionné en fonction des données est la BN1, $-1/+6$ avec $s = 11$ et $\ell^* = 1$, nous faisons remarquer que :

- Un titulaire de police atteint le maximum du SBM après la déclaration de deux réclamations en trois ans, ce qui conduit à une prime annuelle qui est 2,2 fois supérieure à la prime d'un titulaire situé au niveau 1.
- Après la déclaration d'une réclamation, il faut six ans avant que l'assuré puisse reprendre sa position initiale.
- Les nouveaux conducteurs se voient attribuer un niveau d'entrée de 1, le meilleur du SBM. Ce résultat dépend fortement des données utilisées, mais est en parfaite contradiction avec ce que Boucher et Inoussa (2014) a obtenu. En effet, ces auteurs ont conclu que les nouveaux conducteurs devraient se voir attribuer la pire position dans l'échelle SBM.

4.3. Analyse des paramètres

Le tableau 7 présente les valeurs estimées et les erreurs-types de tous les paramètres β de certains des meilleurs modèles, notamment les modèles BN1 et HF-BNB, ainsi que du modèle SBM-BN1 avec $s = 11,\ \ell^* = 1$ et $\Psi = 6$. Pour ce dernier, nous ne considérons pas la variabilité des paramètres structurels s, ℓ^* et Ψ dans l'analyse. Comme il est dit dans Boucher et Inoussa (2014), il est intéressant de noter les écarts entre les $\hat{\beta}$, plus particulièrement lorsque l'on compare les modèles de données transversales (comme la BN1) et les deux autres modèles qui permettent une tarification au mérite. Les paramètres estimés $\widehat{\beta_5}$, $\widehat{\beta_6}$ et $\widehat{\beta_7}$ associés à la distance parcourue en voiture montrent le plus gros écart entre modèles. Les plus faibles écarts entre ces $\hat{\beta}$ sont observés entre le modèle HF-BNBeta et le modèle SBM-BN1, ce qui signifie que la forme des pénalités (surprimes) pour déclaration de réclamations a une incidence sur la tarification *a priori*.

Tableau 7 : Paramètres $\boldsymbol{\beta}$ (erreur-type) estimés pour certains modèles

	BN1		HF-BNB		SBM-BN1	
Paramètre	Estimation	Erreur-type	Estimation	Erreur-type	Estimation	Erreur-type
$\widehat{\beta_0}$	$-2{,}103$	$(0{,}036)$	$1{,}527$	$(0{,}106)$	$-2{,}356$	$(0{,}031)$
$\widehat{\beta_1}$	$0{,}037$	$(0{,}010)$	$0{,}037$	$(0{,}015)$	$0{,}031$	$(0{,}013)$
$\widehat{\beta_2}$	$-0{,}026$	$(0{,}014)$	$-0{,}036$	$(0{,}015)$	$-0{,}031$	$(0{,}014)$
$\widehat{\beta_3}$	$0{,}424$	$(0{,}018)$	$0{,}417$	$(0{,}019)$	$0{,}403$	$(0{,}018)$
$\widehat{\beta_4}$	$0{,}345$	$(0{,}015)$	$0{,}323$	$(0{,}016)$	$0{,}309$	$(0{,}016)$
$\widehat{\beta_5}$	$-0{,}579$	$(0{,}041)$	$-0{,}497$	$(0{,}071)$	$-0{,}481$	$(0{,}028)$
$\widehat{\beta_6}$	$0{,}453$	$(0{,}042)$	$-0{,}397$	$(0{,}070)$	$-0{,}385$	$(0{,}032)$
$\widehat{\beta_7}$	$-0{,}245$	$(0{,}050)$	$-0{,}222$	$(0{,}074)$	$-0{,}216$	$(0{,}039)$
$\widehat{\beta_8}$	$0{,}029$	$(0{,}013)$	$0{,}047$	$(0{,}016)$	$0{,}035$	$(0{,}016)$

4.4. Analsye prédictive et analyse des covariances

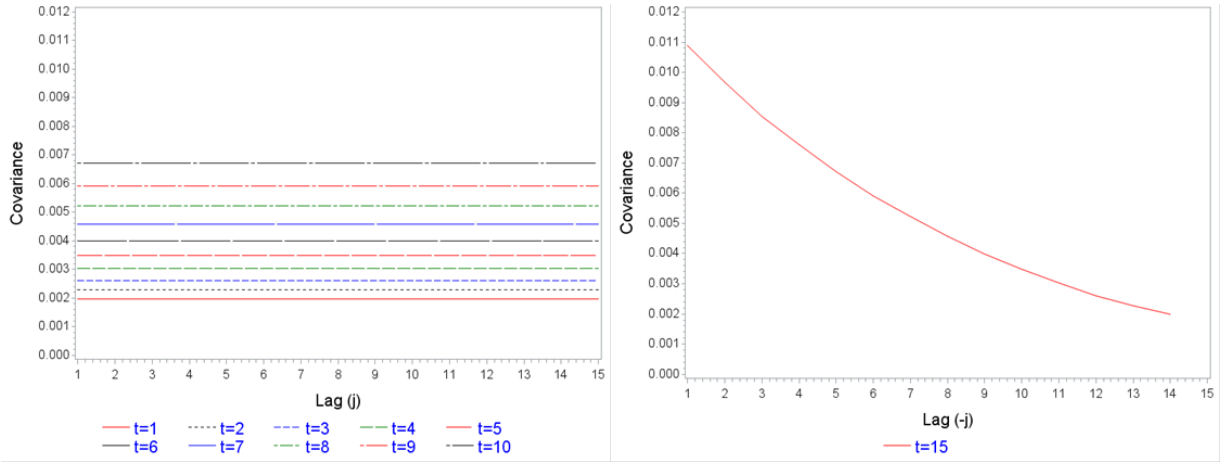
Tableau 8 : Paramètres estimés pour certains modèles

Modèle	Paramètre	Estimation	Erreur-type
BN1	τ	0,067	(0,004)
HF-BNB	a	264,818	(12,881)
	b	5,500	(0,471)
	ν	0,900	(0,008)
SBM-BN1	τ	0,062	(0,003)
	δ	0,120	(0,004)

Le tableau 8 montre les autres paramètres estimés des mêmes modèles. Pour la distribution BN1, la valeur de τ mesure principalement la surdispersion de la distribution de dénombrement et ne modélise pas la dépendance entre les nombres de réclamations. De fait, les modèles transversaux supposent l'indépendance entre les contrats et une covariance nulle entre $N_{i,t}$ et $N_{i,t+j}$, ce qui signifie qu'aucune tarification au mérite n'est possible pour cette catégorie de modèles. Par comparaison, dans le cas des modèles de données de panel plus classiques comme la BNPV ou la BNBêta, nous avons établi que la covariance entre $N_{i,t}$ et $N_{i,t+j}$ était constante et ne dépendait pas de j, ce qui a pour résultat que l'âge d'une réclamation n'est pas pris en compte dans le régime de tarification au mérite.

Les modèles HF-BNB, qui sont généralisés assez directement par une méthode similaire à celle du filtre de Kalman, ont été conçus de manière à permettre différents poids selon l'âge de la réclamation. Puisque la solution analytique qui permet de calculer la covariance est trop complexe, nous avons simplement simulé les valeurs. Le graphique 4 montre deux courbes de la covariance entre les nombres de réclamations. Celle de gauche montre la $\mathrm{Cov}\left[N_{i,t}, N_{i,t+j}\right]$ pour $t = 1, \ldots, 10$ et un retard (*lag*) $j = 1, \ldots, 15$, tandis que l'autre montre la $\mathrm{Cov}\left[N_{i,t-j}, N_{i,t}\right]$ pour $t = 15$ et un retard $j = 1, \ldots, 14$.

Graphique 4 : Covariances des modèles HF avec $\mathrm{Cov}[N_{i,t}, N_{i,t+j}]$ à gauche et $\mathrm{Cov}[N_{i,15-j}, N_{i,15}]$ à droite.



*Disponible en anglais seulement.

L'analyse des courbes indique, par exemple, que le nombre de réclamations observées pour le premier contrat de l'assuré i aura le même impact sur tous les contrats futurs. Toutefois, la covariance entre les nombres de réclamations augmentera à mesure que t grandira. Par conséquent, la relation suivante est valable :

$$\mathrm{Cov}[N_{i,t_1}, N_{i,t_1+j}] \leq \mathrm{Cov}[N_{i,t_2}, N_{i,t_2+j}], \text{pour } t_1 < t_2.$$

Le modèle HF a été construit de manière à donner des poids inégaux à la prime prédictive selon l'âge de la réclamation. Maintenant, nous constatons que le modèle donne moins de poids aux vielles réclamations, mais de plus en plus de poids aux nouvelles réclamations, ce que la proposition suivante explique plus clairement.

Proposition 4.1 : *Dans le cas d'un assuré qui n'a jamais fait de réclamation, l'augmentation de la prime qui suit une réclamation sera plus élevée si l'assuré a une longue expérience de conduite. En outre, plus l'expérience de conduite de l'assuré est élevée, plus l'incidence d'une réclamation sur la prime de l'année suivante est grande.*

Démonstration. Tout d'abord, en nous fondant sur le modèle HF-BNBêta, nous calculons la prime à l'instant $t+1$ pour un assuré qui n'a jamais fait de réclamation :

$$\pi_{i,t+1} = \lambda_{i,t+1}\left(\frac{\nu^t b}{(\nu)^t a + \sum_{k=1}^{t}(\nu)^k \lambda_{i,t-k+1} - 1}\right).$$

Ensuite, nous calculons la prime d'un assuré qui a fait une seule réclamation à l'instant $w < t$:

$$\pi_{i,t+1}^{(w)} = \lambda_{i,t+1}\left(\frac{\nu^t b + \nu^{t-w+1}}{(\nu)^t a + \sum_{k=1}^{t}(\nu)^k \lambda_{i,t-k+1} - 1}\right).$$

L'impact de la réclamation à l'instant w, ou l'augmentation de la prime, s'obtient ensuite comme suit :

$$\frac{\pi_{i,t+1}^{(w)}}{\pi_{i,t+1}} = \left(\frac{\nu^t b + \nu^{t-w+1}}{\nu^t b}\right) = 1 + \frac{\nu^{-w+1}}{b} = 1 + \left(\frac{\nu}{b}\right)\nu^{-w}$$

Puisque $\nu < 1$ pour le modèle HF, ce qui signifie que l'augmentation de la prime s'amplifiera à mesure que w grandira. Par ailleurs, nous faisons remarquer que l'augmentation ne dépend pas ni de t ni de l'âge de la réclamation ($t - w$). En d'autres termes, les modèles HF supposent que l'impact d'une réclamation à l'instant w restera le même pour tous les contrats futurs t.

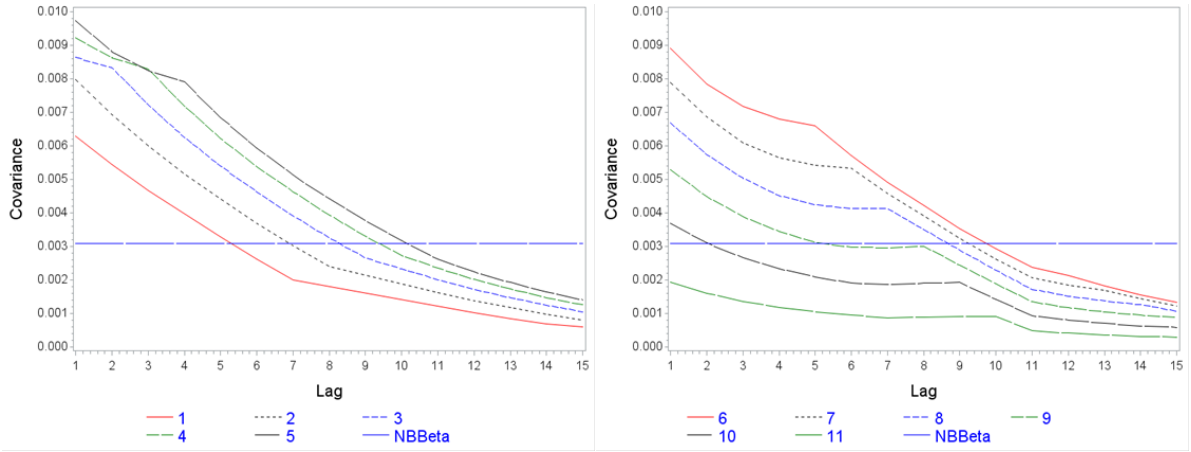
Dans les modèles classiques de données de panel comme la BNPV ou la BNBêta, il est impossible pour un assuré qui a fait au moins une réclamation de se voir demander la même prime qu'un assuré qui n'a jamais fait de réclamation, même lorsque plusieurs années se sont écoulées depuis la première et unique réclamation. Avec la généralisation de HF et l'introduction du paramètre de poids ν, nous nous étions attendus à ce que l'impact des vieilles réclamations sur les primes futures devienne graduellement négligeable. Or, nous venons de voir que ce n'est pas le cas. Pour le HF-BNBêta utilisé avec nos données, nous avons obtenu $\hat{b} = 5{,}5$ et $\hat{\nu} = 0{,}9$, ce qui signifie qu'une réclamation à l'instant $w = 1$ entraînera une hausse de 18,2 % de la prime des contrats futurs, tandis qu'une réclamation à l'instant $w = 15$ augmentera toujours la prime de 78,9 % par comparaison à un assuré qui n'a jamais déclaré d'accident. Même si l'ajustement des modèles HF est intéressant, cette propriété fait en sorte que ces modèles ne peuvent être utilisés dans la pratique : plus grande est l'expérience de conduite de l'assuré, plus grande est la pénalité qu'il se verra imposer pour avoir déclaré un accident. Le modèle SBM-panel ne possède pas cette propriété : par exemple, lorsque le nombre maximal de niveaux $s = 11$, l'assuré qui n'a fait aucune réclamation pendant 11 années consécutives se verra demander la même prime qu'un assuré semblable qui n'a jamais fait de réclamation du tout. Nous croyons que cette propriété est plus réaliste et plus souhaitable pour les assureurs.

Le modèle SBM-BN1 avec $s = 11, \ell^* = 1$ et $\Psi = 6$ suppose une relativité linéaire pour la pénalité pour réclamation, avec $\hat{\delta} = 0{,}12$. Selon le niveau du SBM de chaque assuré, cela signifie que la prime sera égale à $1{,}12; 1{,}24; \ldots; 2{,}08; 2{,}20$ et multipliée par la prime de base (niveau 1). Par exemple, dans le cas d'un assuré au niveau 2, la réclamation durant l'année entraîne un saut de $\Psi = 6$ niveaux, ce qui représente une augmentation de la prime d'environ 64 %. Inversement, une année sans réclamation entraîne une réduction de prime de 11 %.

Afin d'évaluer avec plus de précision l'impact du niveau actuel et la dépendance entre les nombres de réclamations sur les différents contrats, et en nous fondant sur la proposition 3.4, nous calculons la covariance du modèle SBM-panel, comme le montre le graphique 5. Pour comparer les covariances, nous avons aussi inclus la covariance qu'implique le modèle BNBêta, qui reste constante dans le temps. À la différence du modèle HF, la covariance dépend ici de ℓ_t, le niveau du SBM à l'instant t. La propriété dynamique du modèle SBM-panel peut être observée : la dépendance entre les contrats annuels diminue à mesure que le retard augmente. Nous pensons qu'il s'agit là d'une des plus importantes propriétés d'un régime de tarification au mérite. L'impact de ℓ_1 (le niveau du SBM à l'instant 1) sur la covariance est net. Par exemple, puisqu'un assuré situé à $\ell_1 = 1$ ne peut atteindre un niveau plus bas à l'instant 2, la covariance entre N_1 et N_2 est limitée. À mesure que ℓ_1 augmente, la covariance augmente

aussi, mais puisque l'impact d'une réclamation sur le prochain niveau du SBM est de $\Psi = 6$, nous observons que la covariance commence à diminuer pour $\ell_1 > 5$. Enfin, les assurés situés au niveau $\ell_1 = s = 11$ sont aussi limités, cette fois du fait qu'ils ne peuvent atteindre un plus haut niveau du SBM. À mesure que le retard entre deux nombres de réclamations augmente, la covariance diminue et l'impact de ℓ_1, le niveau du SBM à l'instant 1, s'estompe.

Graphique 5 : Covariance du SBM-panel selon le niveau $\boldsymbol{\ell}$



5. Conclusion

L'approche SBM proposée ici produit un score de sinistralité qui est simple à expliquer. Le score de sinistralité est aussi facile à utiliser puisqu'elle résume toute l'expérience de sinistralité d'un assuré, en tenant compte également de l'âge de chaque réclamations. Même si, au départ, nous voulions montrer que le modèle SBM-panel était pratique et souple, nous avons été surpris de constater qu'il produisait également un bien meilleur ajustement statistique que certaines des distributions de dénombrement les plus connues de la littérature actuarielle. La covariance décroissante du modèle, qui attribue un poids moindre aux réclamations âgées, semble expliquer le mieux le pouvoir prédictif intéressant du modèle. En effet, la grande majorité des modèles existants ne font pas preuve d'une telle souplesse. Le modèle HF proposé ici, qui semblait être une simple généralisation des modèles classiques de données de panel, semblait être de prime abord une solution intéressante. Toutefois, en analysant la prime prédictive de ces modèles HF, nous avons démontré qu'ils devenaient tout à fait impraticables lorsque l'historique d'assurance du réclamant s'allongeait. En effet, l'impact d'une seule réclamation par un assuré ayant une longue expérience de conduite devient si important qu'aucun assureur ne serait intéressé à mettre en œuvre une telle approche.

De nombreuses études et généralisations des modèles SBM-panel sont aujourd'hui possibles :

- Nous pouvons étudier une généralisation bidimensionnelle de l'approche SBM, dans laquelle plusieurs types de réclamations (comme les accidents avec ou sans égard à la responsabilité) pourraient être modélisés.
- Nous pouvons chercher à mieux comprendre la dynamique à plusieurs variables en assurance.
- Nous pouvons appliquer, au modèle SBM-panel, de nombreuses propriétés du SBM qui ont déjà été développées et comprises, par exemple, les propriétés asymptotiques du SBM, des outils pour comparer des régimes de tarification au mérite, la soif du bonus, etc.

Remerciements

Jean-Philippe Boucher et Mathieu Pigeon tiennent à remercier l'Institut canadien des actuaires pour son aide financière sous forme d'une subvention de recherche (n^{o} CS000168).

Bibliographie

Abdallah, A., J.-P. Boucher, H. Cossette et J. Trufin. « Sarmanov family of bivariate distributions for multivariate loss reserving analysis », *North American Actuarial Journal*, vol. 20, n^{o} 2, (2016), p. 184-200.

Albrecht, P. « An evolutionary credibility model for claim numbers », *ASTIN Bulletin*, vol. 15, n^{o} 1, (1985), p. 1-17.

Bolancé, C., M. Denuit, M. Guillén et P. Lambert. « Greatest accuracy credibility with dynamic heterogeneity: the Harvey-Fernandes model », *Belgian Actuarial Bulletin*, vol. 7, n^{o} 1, (2007), p. 14-18.

Boucher, J.-P. et M. Denuit. « Fixed versus random effects in Poisson regression models for claim counts: case study with motor insurance », *ASTIN Bulletin*, vol. 36, n^{o} 1, (2006), p. 285-301.

Boucher, J.-P., M. Denuit et M. Guillén. « Models of insurance claim counts with time dependence based on generalisation of Poisson and negative binomial distributions », *Variance*, vol. 2, n^{o} 1, (2008), p. 135-162.

Boucher, J.-P. et M. Guillén. « A survey on models for panel count data with applications to insurance », *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales*, vol. 103, n^{o} 2, (2009), p. 277-295.

Boucher, J.-P. et R. Inoussa. « A posteriori ratemaking with panel data », *ASTIN Bulletin*, vol. 44, n^{o} 3, (2014), p. 587-612.

Bühlmann, H. et A. Gisler. *A Course in Credibility Theory and its Applications*, Berlin, Heidelberg et New York: Springer, 2005.

Choirat, C. et S. Raffaello. « Estimation in discrete parameter models », *Statistical Science*, vol. 27, n^{o} 2, (2012), p. 278–293.

Denuit, M., X. Maréchal, S. Pitrebois et J.-F. Walhin. *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Scales*, Wiley, New York, 2007.

Frees, E.W., R.A. Derrig et G. Meyers. *Predictive Modeling Applications in ActuarialSscience*, Cambridge University Press, Cambridge, 2014.

Gilde, V. et B. Sundt. « On bonus systems with credibility scales », *Scandinavian Actuarial Journal*, vol. 1989, n^{o} 1, p. 13-22.

Gourieroux, C. et J. Jasiak. « Heterogeneous INAR(1) model with application to car insurance », *Insurance: Mathematics and Economics*, vol. 34, n^{o} 2, 2004, p. 177-192.

Harvey, A.C. et C. Fernandes. « Time series models for count or qualitative observations », *Journal of Business & Economics Statistics*, vol. 7, n^{o} 4, 1989, p. 407-422.

Hammersley, J. M. « On estimating restricted parameters » (avec discussion), *Journal of the Royal Statistical Society, Series B*, vol. 12, n^{o} *2*, 1950, p. 192-240.

Jung, R. et R. Liesenfeld. « Estimating time series models for count data using efficient importance sampling », *AStA Advances in Statistical Analysis*, vol. 4, n^{o} 85, 2001, p. 387-407.

Lemaire, J. *Bonus-Malus Systems in Automobile Insurance*, Kluwer, Boston 1995.

McCullagh, P. et J.A. Nelder. *Generalized Linear Models*, Londres, Chapman and Hall, 2^{e} éd., 1989.

Pinquet, J., M. Guillén et C. Bolancé. « Allowance for the age of claims in bonus-malus systems », *ASTIN Bulletin*, vol. 31, n^{o} 2, 2001, p. 337-348.

Shi, P. et E. Valdez. « Longitudinal modeling of insurance claim counts using jitters », *Scandinavian Actuarial Journal*, vol. 2014, n^{o} 2, 2016, p. 159-179.

Shi, P., X. Feng et J.-P. Boucher. « Multilevel modeling of insurance claims using copulas », *Annals of Applied Statistics*, vol. 10, n^{o} 2, 2016, p. 834-863.

Winkelmann, R. *Econometric Analysis of Count Data*. Berlin et Heidelberg : Springer-Verlag, 5^{e} éd., 2010.