Assurances et gestion des risques
Insurance and Risk Management

L'impact de la sinistralité passée sur la sinistralité future : approche empirique en assurance automobile

Olga A. Vasechko et Michel Grun-Réhomme

Volume 78, numéro 1-2, 2010

URI : https://id.erudit.org/iderudit/1106241ar
DOI : https://doi.org/10.7202/1106241ar

Aller au sommaire du numéro

Éditeur(s)

Faculté des sciences de l'administration, Université Laval

ISSN

1705-7299 (imprimé)
2371-4913 (numérique)

Découvrir la revue

Citer cet article

Vasechko, O. & Grun-Réhomme, M. (2010). L'impact de la sinistralité passée sur la sinistralité future : approche empirique en assurance automobile. *Assurances et gestion des risques / Insurance and Risk Management*, *78*(1-2), 71–91. https://doi.org/10.7202/1106241ar

Résumé de l'article

Le risque individuel de chaque assuré automobile n'est pas prévisible et n'est connu qu'a posteriori, à l'inverse du risque collectif qui est prévisible dans la mesure où l'on dispose de l'expérience du passé le plus récent observé sur une population assez grande comparable à celle du portefeuille actuel. Dans cet article, on souhaite examiner de façon empirique, si la sinistralité passée (avant l'année de référence) et la sinistralité actuelle constituent un bon indicateur prévisionnel de la sinistralité future, conditionnellement aux caractéristiques de la classe de risque (ou case tarifaire) de l'assuré. On suppose, en fonction de la sinistralité passée, que chaque classe de risques est constituée de deux catégories de conducteurs : les assurés à bas risques et ceux à hauts risques. À l'aide d'une loi binomiale négative et d'une approche bayésienne, on montre que la probabilité d'être un conducteur à bas risques est plus importante en l'absence de sinistres (ou avec un seul sinistre) et qu'à l'inverse la probabilité d'être un assuré à hauts risques augmente fortement dès que l'assuré a 2 ou 3 sinistres au cours de l'année de référence. Bien sûr le niveau de probabilité varie selon les classes de risque. Dans presque tous les cas, la sinistralité passée est un bon indicateur de la sinistralité future.

Assurances et gestion des risques, vol. 78(1-2), avril-juillet 2010, 71-91
Insurance and Risk Management, vol. 78(1-2), April-July 2010, 71-91

L'impact de la sinistralité passée sur la sinistralité future : Approche empirique en assurance automobile

par Olga A. Vasechko et Michel Grun-Réhomme

RÉSUMÉ

Le risque individuel de chaque assuré automobile n'est pas prévisible et n'est connu qu'a posteriori, à l'inverse du risque collectif qui est prévisible dans la mesure où l'on dispose de l'expérience du passé le plus récent observé sur une population assez grande comparable à celle du portefeuille actuel. Dans cet article, on souhaite examiner de façon empirique, si la sinistralité passée (avant l'année de référence) et la sinistralité actuelle constituent un bon indicateur prévisionnel de la sinistralité future, conditionnellement aux caractéristiques de la classe de risque (ou case tarifaire) de l'assuré. On suppose, en fonction de la sinistralité passée, que chaque classe de risques est constituée de deux catégories de conducteurs : les assurés à bas risques et ceux à hauts risques. A l'aide d'une loi binomiale négative et d'une approche bayésienne, on montre que la probabilité d'être un conducteur à bas risques est plus importante en l'absence de sinistres (ou avec un seul sinistre) et qu'à l'inverse la probabilité d'être un assuré à hauts risques augmente fortement dès que l'assuré a 2 ou 3 sinistres au cours de l'année de référence. Bien sûr le niveau de probabilité varie selon les classes de risque. Dans presque tous les cas, la sinistralité passée est un bon indicateur de la sinistralité future.

Mots clés : Assurance automobile, sinistralité, loi binomiale négative, formule de Bayes.

Les auteurs :

Olga A. Vasechko, Research Institute of Statistics, 4-6 Esplanadna str., 01023 Kyiv, Ukraine, O.Vasechko@ukrstat.gov.ua.

Michel Grun-Réhomme, ERMES (EAC7181 CNRS)

UNIVERSITE PARIS II PANTHEON-ASSAS, 12 place du Panthéon, 75005 Paris, France grun@u-paris2.fr.

ABSTRACT

The risk of each insured motor vehicle is not predictable and is known only retrospectively. Unlike the collective risk is predictable when we have the experience of the past later observed over a relatively large population comparable to the existing portfolio. In this article, we would like to examine empirically on data from a French insurer, if the accident record (before the reference year) and the current claims are a good predictor of future claims, provided the requirements of the class of risk (or case pricing) of the insured. It is assumed, depending on the accident record, that each class of risk has two categories of drivers insured low-risk and high risk. Using a negative binomial distribution and a Bayesian approach, we show that the probability of being a low risk driver is more important in the absence of losses (or with a single claim) and that conversely the probability of an insured high risk greatly increases when the insured loss to 2 or 3 during the reference year. Of course the level of probability varies according to risk classes. In all cases, the previous accidents are a good indicator of future claims.

Keywords: Solvency 2, ORSA, ERM, *Risk Appetite.*

1. INTRODUCTION

Dans l'échange entre assureur et assuré, un contrat de garanties contre une rémunération (prime ou cotisation), la compagnie d'assurance fait face à un risque qui est directement lié à l'asymétrie d'information qui existe entre elle et l'assuré. En effet, l'information sur les risques n'est pas partagée et compromet l'optimalité de l'échange.

La segmentation sur les marchés d'assurance devient alors un moyen d'adapter la tarification aux caractéristiques de risque de chaque individu. Quand il y a antisélection, une tarification uniforme désavantage les consommateurs de moindre risque, qui paient plus cher que ne l'exigent les impératifs liés au coût de leur risque. Les assurés plus exposés aux sinistres, quant à eux, se trouvent avantagés, puisque la tarification est fondée sur le coût moyen des sinistres de l'ensemble de la population. Une classification des risques efficace permet alors de rapprocher chaque type d'individu de sa réalité actuarielle, instaurant plus d'équité dans la tarification de l'assurance.

L'assureur doit répartir la charge de sinistralité de façon équitable entre tous les assurés, en même temps qu'il mutualise les risques entre les assurés qui présentent des caractéristiques semblables personnelles et de véhicule. L'assureur procède donc à une recherche minutieuse de tous les facteurs disponibles et susceptibles d'expliquer le risque. Des classes de risque sont constituées à partir de ces

facteurs comme l'ancienneté de permis, l'usage du véhicule, la zone d'habitation, la puissance du véhicule, la gamme du véhicule,...

Dans la constitution des classes de risque, où l'information doit être disponible et fiable, un équilibre doit être trouvé entre la granularité et la robustesse. Si la granularité (ou la segmentation) est trop grossière, certes la robustesse temporelle des indicateurs de sinistralité est assurée, mais la mutualisation est trop large et un concurrent peut très bien attirer les bons risques de cette classe en proposant une cotisation plus faible grâce à une segmentation plus fine. A l'inverse une granularité trop fine ne permet pas d'avoir cette robustesse. Au sein d'une mutualisation des risques, il existe une volatilité résiduelle. Du fait de cette asymétrie d'information, les assureurs ne savent pas précisément dire s'ils ont affaire à un client (ou un sociétaire) à hauts ou à bas risques. La discrimination sur les marchés d'assurance devient alors un moyen d'adapter la tarification aux caractéristiques de risque de chaque individu.

Une autre sorte de discrimination consiste à créer une institution d'assurance spécifique à un groupe particulier de la population, désireuse de s'isoler du reste de la population, afin de mettre en place son propre système d'assurance. Les mutuelles d'assurance illustrent ce procédé dans la mesure où elles sont fréquemment rattachées à un groupe socioprofessionnel déterminé. Ainsi, si le degré de risque se trouve corrélé au fait d'appartenir à une catégorie socioprofessionnelle donnée, des mutuelles d'assurance organisées par profession effectuent de fait une classification (imparfaite) des risques. Créer une telle institution permet au groupe en question d'établir un sous marché d'assurance sur lequel l'antisélection s'atténue du fait de l'homogénéité de la clientèle concernée face au risque. Une frange de population donnée aura d'autant plus intérêt à segmenter le marché par le biais d'une mutuelle que son risque moyen est faible et/ou son homogénéité forte.

Le coût du risque individuel de chaque assuré n'est pas prévisible et n'est connu qu'a posteriori, à l'inverse du risque collectif qui lui est prévisible dans la mesure où l'on dispose de l'expérience du passé le plus récent observé sur une population assez grande comparable à celle du portefeuille actuel. Les grandes sociétés d'assurance sont donc avantagées dans la prévision du risque collectif de par la représentativité de leurs portefeuilles.

Dans cet article, on souhaite examiner de façon empirique, si la sinistralité passée est un bon indicateur prévisionnel de la sinistralité future et ceci conditionnellement aux caractéristiques de la classe de

risque (ou case tarifaire) de l'assuré. Les données utilisées sont celles d'un assureur français.

La section 2 développe les aspects formels de cette problématique, la section 3 présente les données d'assurance utilisées et la section 4 donne les résultats de cette étude empirique. Enfin une conclusion et une bibliographie terminent cet article.

2. ASPECTS MÉTHODOLOGIQUES

Certes les classes de risque sont supposées homogènes en terme de sinistralité (mesurée en fréquence et en coût), mais dans la réalité d'un portefeuille, on constate une hétérogénéité individuelle qui provient de facteurs inobservés ou inobservables, comme le nombre de kilomètres parcourus, l'état du véhicule, le comportement du conducteur au volant (respect du code de la route, rapidité des réflexes, agressivité au volant..), les conditions climatiques, l'état du réseau routier...

Partons de l'hypothèse réaliste, qu'une classe de risque donnée c, est composée de deux sous-populations :

– Les bons conducteurs, ceux qui ont fait des efforts de conduite, qui ont respecté les termes du contrat, les assurés à bas risques.

– Les mauvais conducteurs qui ne respectent pas les termes du contrat, qui provoquent des accidents et qui correspondent à des assurés à hauts risques.

Cette différentiation entre ces deux catégories peut être effectuée à partir de la sinistralité antérieure, mesurée par le Coefficient Réduction Majoration (CRM) ou Bonus-Malus, qui traduit le passé de sinistralité du conducteur. Rappelons que la première année d'assurance, le Coefficient Réduction Majoration est égal à 1.00, et qu'il est réévalué chaque année. Si l'assuré n'a eu aucun sinistre responsable, le coefficient est multiplié par 0.95 avec une règle pour les arrondis en utilisant la partie entière. Il est multiplié par 1.25 pour chaque sinistre 100 % responsable et par 1.125 pour les sinistres à torts partagés. En aucun cas, le coefficient ne pourra dépasser 3.50 ou passer en dessous de 0.50. Pour plus de détails, cf. Grun-Réhomme, 2000, Denuit, Charpentier, 2005, Denuit et *al.*, pp. 326-329, 2007.

Ce CRM est alors corrigé par l'ancienneté de permis du conducteur; en effet un jeune de 18 ans ne peut pas avoir un CRM égal à 50. Un ajustement peut éventuellement être mis en œuvre en

considérant l'âge du conducteur. Ceci peut être considéré comme une bonne estimation de la sinistralité passée en l'absence de données longitudinales.

On suppose que ces deux catégories de conducteurs, « toutes choses égales par ailleurs » (à caractéristiques identiques de construction des classes de risque) sont homogènes au niveau de la sinistralité. On fixe une clase de risque quelconque c et pour simplifier les notations, on n'indique pas cet indice. Fixer une classe de risque, revient à fixer les caractéristiques tarifaires de cette classe, à savoir la prime pure. La prime pure correspond au coût du sinistre moyen auquel devra faire face l'assureur. Elle est égale à l'espérance des pertes. Le calcul de la prime pure a pour objectif d'évaluer pour chaque assuré (selon ses caractéristiques) le montant attendu des sinistres pour la période d'assurance concernée, en général une année.

La variable CRM n'est pas utilisée comme facteur a priori pour segmenter le portefeuille.

2.1 Les modèles de comptage

Dans cette présentation, la sinistralité actuelle (pour l'année de référence considérée) est mesurée par le nombre de sinistres (et non les coûts).

Dans la modélisation de ces processus de comptage, ici de la fréquence des sinistres, deux sortes de modèle sont couramment mis en œuvre; le modèle de poisson et le modèle binomial négatif. On trouve une littérature abondante sur l'utilisation de ces modèles : Greene (1994), Wooldridge (1997), Cameron et Trivedi (1998), Winkilmann (2000), Yau et *al*. (2003), Yang et *al.* (2007).

Notons B (resp. H), la catégorie des bons conducteurs, des bas risques (resp. des hauts risques, des mauvais conducteurs). La probabilité qu'un bon conducteur (resp. mauvais conducteur) ait k sinistres durant l'année donnée d'observation de la sinistralité, peut suivre une loi de Poisson de paramètre λ_B (resp. λ_H).

Le modèle de Poisson fait partie des modèles linéaires généralisés. Dans un modèle de Poisson, la probabilité pour que la variable aléatoire N, nombre de sinistres responsables (totalement ou à torts partagés) déclarés durant l'année de référence (resp. N_B pour les bons conducteurs et N_H pour les hauts risques) prenne la valeur y_i (y_i=0,1,2...) pour un assuré i est donnée par :

$$P(N = y_i / X_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \qquad (1)$$

où le paramètre λ_i (λ_B ou λ_{H_i} selon les cas) dépend du vecteur X_i des caractéristiques (régresseurs) de l'assuré *i* par une équation log linéaire, à savoir : In $\lambda_i = X_i'\beta$, où β est le vecteur des coefficients de régression à estimer.

On vérifie aisément que dans la loi de Poisson (1), l'espérance est égale à la variance.

$$E(N_i / X_i) = Var(N_i / X_i) = \lambda_i = e^{X_i'\beta} \quad (2)$$

Plus précisément, $\lambda_i = \exp(\beta_0)\prod_{j=1}^{m}\exp(\beta_j x_{ij})$, où *m* correspond au nombre de variables dans le modèle de régression et $\exp(\beta_0)$ correspond à la fréquence annuelle de l'individu de référence. Lorsque toutes les variables sont qualitatives, les composantes du vecteur X_i valent 0 ou 1 (Denuit, Charpentier, tome II, 2005, pp. 130-145). Les caractéristiques X_i sont maintenant fixées.

On a, pour les moments d'ordre 1 et 2 pour chaque catégorie d'assurés :

$$E(N_B) = Var(N_B) = \lambda_B \quad (3)$$

et $$E(N_H) = Var(N_H) = \lambda_H \quad (4)$$

Cette hypothèse d'équidispersion (homogénéité du portefeuille par rapport au risque) est très restrictive.

Dans la pratique, du fait d'une abondance de valeurs nulles et de la présence de quelques valeurs extrêmes, la variance est souvent significativement supérieure à la moyenne. Dans ce cas, on parle d'une surdispersion de la variable *N*. Cette situation implique une sous estimation des écarts-types et on rejette trop souvent l'hypothèse nulle de non significativité des coefficients β du modèle.

D'où l'idée d'utiliser un modèle de comptage alternatif, basé sur la loi binomiale négative, qui prend en compte cette surdispersion par l'introduction d'un paramètre supplémentaire (α) qui permet, en outre, de capter l'hétérogénéité inobservée de la variable endogène (qui peut impliquer la surdispersion observée).

Notons w_B (resp. w_H) le pourcentage d'assurés à bas risques (resp. à hauts risques) dans la classe considérée. Bien sûr, ces pourcentages peuvent varier d'une classe à l'autre.

Dans une classe de risque :

$$E(N) = w_B E(N_B) + w_H E(N_H) = m \quad (5)$$

La variance de *N* est égale à (Denuit et *al.*, 2007, pp. 79-83) :

$$Var(N) = w_B E(N_B) + w_H E(N_H) + w_B (E(N_B) - m)^2 + w_H (E(N_H) - m)^2 > m \quad (6)$$

Ce qui implique une surdispersion dans la classe de risque et N suit alors une loi binomiale négative.

Dans un modèle binomial négatif, on définit la probabilité pour que N prenne la valeur y_i par :

$$P(N = y_i / X_i) = \frac{\Gamma(y_i + \nu)}{\Gamma(y_i + 1) \cdot \Gamma(\nu)} \cdot \left(\frac{\nu}{\nu + \lambda_i}\right)^{\nu} \left(\frac{\lambda_i}{\nu + \lambda_i}\right)^{y_i} \quad (7)$$

En posant, $\nu = 1/\alpha$, l'espérance et la variance s'expriment ainsi :

$$E(y_i / X_i) = \lambda_i = e^{X_i\beta}, \quad Var(y_i / X_i) = \lambda_i (1 + \alpha\lambda_i) \quad (8)$$

La variance est donc différente de l'espérance et le paramètre α traduit une surdispersion (ou une sousdispersion) des données. Si $\alpha = 0$, le modèle binomial se réduit au modèle de Poisson. Si $\alpha > 0$, le modèle de Poisson est rejeté au profit du modèle binomial négatif. La surdispersion peut être testée soit par le ratio $D/(n-p)$, où D désigne la déviance, n le nombre d'observations et p le nombre de paramètres dans le modèle, soit par le ratio $X^2/(n-p)$, où X^2 correspond à la statistique du chi-deux de Pearson. La déviance est définie comme 2 fois la différence entre le maximum possible de la log vraisemblance et le maximum atteint sur le modèle estimé (Mc Cullagh, Nelder, 1989). Le X^2 de Pearson correspond à la somme des carrés des écarts à la moyenne.

Si ces ratios sont supérieurs à 1, les données présentent une surdispersion (et une sous dispersion si ces rations sont inférieurs à 1).

De nombreux auteurs ont comparé différents modèles qui peuvent rendre compte d'un processus de comptage. Citons simplement, deux articles récents : Klugman et *al.* (2004), Karlis (2005).

2.2 Sinistralité passée et sinistralité actuelle

A partir de la sinistralité passée (deux catégories de conducteurs), comment estimer la sinistralité future ?

A la fin de l'année en cours, on dispose d'une information sur la sinistralité actuelle. Donc nous disposons d'une sinistralité passée (avant l'année de référence) et d'une sinistralité actuelle (mesurée par N), conditionnellement aux caractéristiques de l'assuré.

Considérons un assuré qui a eu k sinistres au cours de l'année actuelle de référence, il est alors possible de calculer sa probabilité d'être un bon conducteur en utilisant la formule de Bayes :

$$\Pr(B/N=k)=\frac{\Pr(N=k/B)\times\Pr(B)}{\Pr(N=k/B)\times\Pr(B)+\Pr(N=k/H)\times\Pr(H)} \quad (9)$$

ou encore

Pour une loi de Poisson :

$$\Pr(B/N=k)=\frac{\exp(-\lambda_B)\lambda_B^k\Pr(B)}{\exp(-\lambda_B)\lambda_B^k\Pr(B)+\exp(-\lambda_H)\lambda_H^k\Pr(H)} \quad (10)$$

Et pour une loi binomiale négative, $\Pr(B/N=k)$ est égale à :

$$\frac{\frac{\Gamma(k+\nu_B)}{\Gamma(k+1)\Gamma(\nu_B)}\times\left[\frac{\nu_B}{\nu_B+k}\right]^{\nu_B}\times\left[\frac{\lambda_B}{\lambda_B+\nu_B}\right]^{k}\Pr(B)}{\frac{\Gamma(k+\nu_B)}{\Gamma(k+1)\Gamma(\nu_B)}\times\left[\frac{\nu_B}{\nu_B+k}\right]^{\nu_B}\times\left[\frac{\lambda_B}{\lambda_B+\nu_B}\right]^{k}\Pr(B)+\frac{\Gamma(k+\nu_H)}{\Gamma(k+1)\Gamma(\nu_H)}\times\left[\frac{\nu_H}{\nu_H+k}\right]^{\nu_H}\times\left[\frac{\lambda_H}{\lambda_H+\nu_H}\right]^{k}\Pr(H)} \quad (11)$$

Pour $k=0$, $\Pr(B/N=0)=\Pr(B)$

et pour $k=1$, $\Pr(B/N=1)$ est égale à :

$$\frac{\nu_B\times\left[\frac{\nu_B}{\nu_B+1}\right]^{\nu_B}\times\left[\frac{\lambda_B}{\lambda_B+\nu_B}\right]\Pr(B)}{\nu_B\times\left[\frac{\nu_B}{\nu_B+1}\right]^{\nu_B}\times\left[\frac{\lambda_B}{\lambda_B+\nu_B}\right]\Pr(B)+\nu_H\times\left[\frac{\nu_H}{\nu_H+1}\right]^{\nu_H}\times\left[\frac{\lambda_H}{\lambda_H+\nu_H}\right]\Pr(H)} \quad (12)$$

Les distributions des deux catégories d'assurés (bas risques et hauts risques) diffèrent par les valeurs des deux paramètres λ et v. Les relations entre les fonctions de répartition de deux distributions binomiales (ici, Bas risques et Hauts risques) ont été étudiées par Tobin (1965) et Rothschild, Stiglitz (1970).

3. APPROCHE EMPIRIQUE

Nous disposons d'un échantillon aléatoire de 44 038 observations (après apurement du fichier initial) du portefeuille d'une mutuelle française d'assurance. Ces données concernent des véhicules 4 roues de tourisme assurés durant l'année 2005 (année de référence).

3.1 Les données

Pour chaque assuré de notre échantillon, nous disposons de 4 groupes de variables : les caractéristiques du conducteur, les caractéristiques du véhicule, le type de contrat et la sinistralité de l'année de référence. Plus précisément,

Caractéristiques du conducteur

Sexe : Il s'agit du sexe du conducteur principal déclaré.

Type du conducteur, il exprime la qualification du conducteur principal déclaré au regard du véhicule (le conducteur principal déclaré est ou non l'assuré).

Age du conducteur, exprimé en années.

Numéro de département : numéro de département de l'habitat principal du conducteur.

Ancienneté de permis : exprimée en années.

Coefficient Réduction Majoration (CRM) ou Bonus Malus : il est compris entre 0.50 et 3.50 inclus (exprimé en %), conformément à la législation française en vigueur.

Caractéristiques du véhicule

Ancienneté de véhicule : elle exprime le millésime de l'année du modèle du véhicule. *Puissance réelle du véhicule* : elle exprime la puissance du moteur en chevaux Din (Deutsch Industrie Normen). Cette mesure donne une vision plus réaliste de la puissance effective au niveau des roues (1 ch. Din = 0,735 Watt).

La variable sur l'usage du véhicule n'a pas été retenue puisque la quasi-totalité des sociétaires en avait un usage promenade-trajet et non un usage professionnel.

Marque : indique le constructeur automobile.

Les Contrats

Cette assureur propose quatre type de garanties pour l'assurance d'un véhicule 4 roues de tourisme :

– Responsabilité Civile (RC, assurance minimale obligatoire); sont inclus dans cette formule des garanties Défense-recours, Attentats, Catastrophes naturelles, Dommages corporels du conducteur et Assistance.

– Dommages au Véhicule (DV1) : RC + Garantie Dommage au véhicule toutes causes avec une franchise importante.

– DV2 : RC + Garantie Dommage au véhicule toutes causes avec une franchise moyenne.

– DV3 : RC + Garantie Dommage au véhicule toutes causes avec une franchise faible.

Cette variable n'est pas utilisée dans cette problématique, mais elle pourrait intervenir dans une étude plus fine qui prendrait en compte la question de la sélection adverse.

Caractéristiques des sinistres

Nombre de sinistres déclarés : pour l'année de référence (2005).

Responsabilité du conducteur : variable binaire qui indique si la responsabilité du conducteur est engagée, ou non, en cas de sinistre.

3.2 Les classes de risque

Les classes de risque sont construites à partir de caractéristiques du conducteur (type de conducteur), de caractéristiques du véhicule (ancienneté, puissance, marque) et du lieu d'habitation. Le type de conducteur correspond au fait que l'assuré est, ou n'est pas, le conducteur principal du véhicule assuré; cette distinction est pertinente pour les jeunes conducteurs et dans une moindre mesure pour les conjoints.

Cette application numérique n'a valeur que d'exemple puisque l'on travaille sur un échantillon et non sur l'ensemble du portefeuille, mais la démarche méthodologique reste la même. Par ailleurs, nous n'avons pas utilisé toutes les variables du fichier pour construire les classes, car d'une part certaines n'étaient pas significatives (comme le sexe), et d'autre part, la retenue des autres (moins significatives) aurait produit un trop nombre de classes avec des effectifs trop petits, perdant ainsi toute robustesse de la méthode. Il est bien sûr impossible de « sortir » les données individuelles, qui seraient nécessaires pour cette étude, sur l'ensemble du portefeuille.

Plus précisément, les classes de risques sont construites à partir du croisement de trois variables :

– La zone géographique de garage habituelle du véhicule (3 modalités; milieu rural, villes moyennes et grandes villes),

– La gamme du véhicule (4 modalités du bas de gamme au haut de gamme qui dépendent de la puissance du véhicule et du prix du véhicule. Cette variable est donnée par les constructeurs à l'assureur et elle est ici modulée par l'ancienneté du véhicule, et

– Le type de conducteur selon que l'assuré est, ou non, le conducteur principal déclaré.

Dans le fichier, 65,5 % des assurés se déclarent comme conducteur principal (type=t2).

TABLEAU 1
RÉPARTITION DE L'ÉCHANTILLON SELON LA GAMME DU VÉHICULE ET LA ZONE D'HABITATION

Gamme	Effectifs	%	Zone	Effectifs	%
1 (Bas)	13 303	30,21	1 (Rurale)	8 814	20,01
2 (Economique)	14 295	32,46	2 (Agglomérations moyennes)	18 126	41,16
3 (Moyen)	13 975	31,73	3 (Grandes agglomérations)	17 098	38,83
4 (Haut)	2 465	5,60			

Cette répartition de l'échantillon selon ces deux variables correspond « grosso modo » à la répartition de l'ensemble du portefeuille. Pour cet échantillon, l'espérance de la variable de comptage N du nombre de sinistres est égale à 0,74 et sa variance à 1,23.

La distribution du nombre de sinistres présente donc une surdispersion dans l'ensemble de l'échantillon du portefeuille. Le tableau 2 présente la distribution N.

La sinistralité passée, calculée à partir du CRM et modulée par l'ancienneté de permis et l'âge du conducteur, sert uniquement à distinguer, dans chaque classe de risque, les bons conducteurs des conducteurs à hauts risques.

Conformément à notre approche, il s'agit maintenant de s'interroger sur la pertinence de la sinistralité passée (et actuelle) comme un bon indicateur prévisionnel de la sinistralité future. Il est donc nécessaire de calculer (ou d'estimer) les différentes probabilités d'occurrence de sinistralité selon les classes de risque et la catégorie de conducteur. Le tableau 3 présente la probabilité (estimée à partir des fréquences) d'être un bon conducteur (B) pour chaque classe de risque de l'année de référence. Naturellement, pour les conducteurs à hauts risques : $\Pr(H) = 1 - \Pr(B)$.

TABLEAU 2
RÉPARTITION DE L'ÉCHANTILLON SELON LE NOMBRE DE SINISTRES

Nombre de sinistres *(N)*	Effectifs	Pourcentages
0	28 389	64,46
1	5 966	13,55
2	5 497	12,48
3	2 328	5,29
4	1 097	2,49
5	462	1,05
6	192	0,44
7	68	0,15
8	31	0,07
9	8	0,02
10	1	0,00

Par exemple, la classe « z2_g3_t2 » correspond à des assurés qui habitent dans des agglomérations moyennes, possèdent un véhicule de gamme moyenne et se déclarent comme conducteur principal.

Les résultats présentés dans ce tableau mettent en évidence les éléments suivants :

– La sinistralité actuelle est plus importante, en moyenne, pour les assurés à hauts risques, « toutes caractéristiques égales par ailleurs ».

– Le pourcentage de bons conducteurs est plus important quand l'assuré n'est pas le conducteur principal.

– La sinistralité est, en moyenne, plus importante en zone urbaine qu'en zone rurale. Ce résultat est conforme à la pratique des assureurs qui imposent une prime plus importante aux assurées des zones rurales, à caractéristiques de tarification identiques. On peut toutefois souligner que les accidents en zone rurale sont, en moyenne, plus graves et plus coûteux que les accidents en ville qui correspondent plus souvent à des « accrochages », où seuls des dégâts matériels sont en cause.

TABLEAU 3
RÉPARTITION DE L'ÉCHANTILLON SELON LES CLASSES DE RISQUE ('Z' INDIQUE LA ZONE, 'G' LA GAMME ET 'T' LE TYPE DE CONDUCTEUR) ET CALCUL DE PR(*B*) SELON CES CLASSES

Classe	Effectifs	Pourcentages	Effectifs cumulés	Pr (B)
z1_g1_t1	199	0,46	199	0,76
z1_g1_t2	297	0,67	496	0,70
z1_g2_t1	987	2,24	1483	0,69
z1_g2_t2	1873	4,25	3356	0,63
z1_g3_t1	924	2,10	4280	0,63
z1_g3_t2	1787	4,06	6067	0,58
z1_g4_t1	1068	2,43	7135	0,61
z1_g4_t2	1679	3,81	8814	0,58
z2_g1_t1	447	1,02	9261	0,74
z2_g1_t2	563	1,28	9824	0,69
z2_g2_t1	1878	4,26	11702	0,68
z2_g2_t2	3751	8,52	15453	0,61
z2_g3_t1	1952	4,43	17405	0,61
z2_g3_t2	4036	9,16	21441	0,55
z2_g4_t1	2222	5,05	23663	0,60
z2_g4_t2	3277	7,44	26940	0,54
z3_g1_t1	394	0,89	27334	0,71
z3_g1_t2	565	1,28	27889	0,68
z3_g2_t1	1873	4,25	29772	0,66
z3_g2_t2	3613	8,20	33385	0,58
z3_g3_t1	1804	4,10	35189	0,60
z3_g3_t2	3792	8,61	38981	0,53
z3_g4_t1	1888	4,29	40869	0,59
z3_g4_t2	3169	7,20	44038	0,50

3.3 Approche prévisionnelle de la sinistralité

La démarche est identique dans chaque classe de risque. A partir de la sinistralité passée (antérieure à l'année de référence) et de la sinistralité actuelle, il s'agit d'effectuer des prévisions de sinistralité selon que l'assuré appartient dans le passé à la catégorie des conducteurs à bas risques ou à hauts risques, Le tableau 4 présente les

valeurs des paramètres nécessaires (espérance empirique et variance) à cette étude.

Dans les différents tableaux, les valeurs sont arrondies pour une meilleure lecture, mais elles ne sont pas arrondies pour les calculs

TABLEAU 4
VALEUR DE L'ESPÉRANCE ET DE LA VARIANCE DE LA VARIABLE DE COMPTAGE DES OCCURRENCES DES ACCIDENTS SELON LES CLASSES DE RISQUE ET LA SINISTRALITÉ PASSÉE

Classes	Bas risques Moyenne	Bas risques Variance	Hauts risques Moyenne	Hauts risques Variance
z1_g1_t1	0,81	1,51	1,10	2,56
z1_g1_t2	0,88	1,85	1,05	2,25
z1_g2_t1	0,65	1,23	0,88	2,07
z1_g2_t2	0,67	1,39	0,99	2,13
z1_g3_t1	0,62	1,25	0,89	1,72
z1_g3_t2	0,57	1,28	0,70	1,44
z1_g4_t1	0,55	0,96	0,78	1,49
z1_g4_t2	0,62	1,69	0,70	1,56
z2_g1_t1	0,72	1,30	1,17	2,31
z2_g1_t2	0,78	1,37	1,09	1,99
z2_g2_t1	0,75	1,39	1,05	2,10
z2_g2_t2	0,72	1,42	0,98	2,07
z2_g3_t1	0,73	1,46	1,03	2,10
z2_g3_t2	0,65	1,28	0,78	1,64
z2_g4_t1	0,66	1,30	0,87	1,74
z2_g4_t2	0,50	0,98	0,54	1,21
z3_g1_t1	0,71	1,25	1,19	2,40
z3_g1_t2	0,75	1,64	1,05	2,28
z3_g2_t1	0,77	1,64	1,19	2,37
z3_g2_t2	0,73	1,51	1,02	2,04
z3_g3_t1	0,78	1,56	1,17	2,50
z3_g3_t2	0,67	1,39	0,78	1,64
z3_g4_t1	0,68	1,14	0,87	1,69
z3_g4_t2	0,52	1,00	0,60	1,32

intermédiaires. Les logiciels SAS (construction des classes) et Excel ont été utilisés.

Pour chaque classe de risque, la sinistralité (mesurée par le nombre de sinistres) actuelle est plus importante pour les assurés à bas risques (sinistralité passée) que les assurés à hauts risques. On vérifie également qu'il en est de même pour chaque variable de cette segmentation (cf. annexe 1). La sinistralité diminue avec la gamme du véhicule et le fait que l'assuré est le conducteur principal. Elle augmente avec la densité de la zone géographique.

On constate que la variance empirique excède significativement la moyenne empirique dans chaque classe. Dans ce cas, les observations sont surdispersées par rapport au modèle Poisson. On va donc utiliser la loi binomiale négative puisqu'elle possède un paramètre supplémentaire, qui peut être utilisé pour ajuster la variance indépendamment de la moyenne.

A partir des relations (8), les paramètres $\lambda = (E(N))$et ν de la loi binomiale négative peuvent aussi s'exprimer ainsi :

$$\nu = \frac{\lambda p}{1-p}, où\ p = \frac{\lambda}{Var(N)} \qquad (13)$$

Les valeurs de ν et p peuvent alors être obtenues à partir du tableau 4. Le tableau 5 suivant en présente les résultats,

Ainsi dans chaque classe, connaissant la catégorie du conducteur dans le passé (sinistralité passée) et sa sinistralité actuelle (nombre de sinistres), on peut calculer la probabilité d'être un conducteur à bas risques ou à hauts risques selon le nombre k de sinistres au cours de l'année de référence, en utilisant la formule de Bayes (12) et les propriétés classiques de récurrence de la fonction Gamma : $\Gamma(x+1) = x\,\Gamma(x)$ (pour $x > 0$) et $\Gamma(k+1) = k!$ (pour k entier naturel).

Le tableau 6 présente les résultats, pour k = 0, 1, 2, 3, pour une loi binomiale négative. Au-delà de 3 sinistres, les effectifs étant assez faibles (dans chaque clase de risque), la robustesse temporelle des résultats n'est plus assurée.

Les résultats de ce tableau 6 mettent en évidence les points suivants :

– La probabilité d'être un bon conducteur diminue avec le nombre de sinistres dans l'année de référence et à l'inverse la probabilité d'être un conducteur à hauts risques augmente avec le nombre d'accidents.

TABLEAU 5
ESTIMATION DES PARAMÈTRES λ ET ν DE LA LOI BINOMIALE NÉGATIVE SELON LES CLASSES DE RISQUE ET LA CATÉGORIE DU CONDUCTEUR

Classes	λ_B	ν_B	λ_H	ν_H
z1_g1_t1	0,81	1,56	1,10	2,42
z1_g1_t2	0,88	1,61	1,05	2,45
z1_g2_t1	0,65	0,92	0,88	1,38
z1_g2_t2	0,67	0,88	0,99	2,09
z1_g3_t1	0,62	0,77	0,89	1,89
z1_g3_t2	0,57	0,58	0,70	0,98
z1_g4_t1	0,55	0,70	0,78	1,38
z1_g4_t2	0,62	0,36	0,70	0,57
z2_g1_t1	0,72	1,23	1,17	3,91
z2_g1_t2	0,78	1,56	1,09	3,71
z2_g2_t1	0,75	1,31	1,05	2,76
z2_g2_t2	0,72	1,10	0,98	2,09
z2_g3_t1	0,73	1,11	1,03	2,53
z2_g3_t2	0,65	0,88	0,78	1,22
z2_g4_t1	0,66	0,91	0,87	1,68
z2_g4_t2	0,5	0,51	0,54	0,52
z3_g1_t1	0,71	1,23	1,19	3,93
z3_g1_t2	0,75	1,06	1,05	2,40
z3_g2_t1	0,77	1,16	1,19	4,05
z3_g2_t2	0,73	1,07	1,02	2,54
z3_g3_t1	0,78	1,29	1,17	3,34
z3_g3_t2	0,67	0,88	0,78	1,22
z3_g4_t1	0,68	1,19	0,87	1,76
z3_g4_t2	0,81	1,56	1,10	2,42

– Pour un assuré qui n'a pas eu de sinistre au cours de l'année de référence ou un seul sinistre, la probabilité d'être un bon conducteur est supérieure à la probabilité d'être un conducteur à hauts risques.

– Pour les assurés ayant eu deux sinistres, la probabilité d'être un conducteur à bas risques est supérieure (ou parfois inférieure, si Pr(B) est proche de 0,50) à celle d'être un conducteur à hauts risques selon les classes.

TABLEAU 6
PROBABILITÉ D'APPARTENIR À L'UNE DES DEUX CATÉGORIES DE CONDUCTEURS SELON LE NOMBRE DE SINISTRES DÉCLARÉS DURANT L'ANNÉE DE RÉFÉRENCE ET SELON LA CLASSE DE RISQUE

	Bas risques				Hauts risques (*)			
Classes	*k* = 0	*k* = 1	*k* = 2	*k* = 3	*k* = 0	*k* = 1	*k* = 2	*k* = 3
z1_g1_t1	0,76	0,74	0,72	0,39	0,24	0,26	0,28	0,61
z1_g1_t2	0,70	0,67	0,64	0,42	0,30	0,33	0,36	0,58
z1_g2_t1	0,69	0,66	0,64	0,40	0,31	0,34	0,36	0,60
z1_g2_t2	0,63	0,56	0,49	0,31	0,37	0,44	0,51	0,69
z1_g3_t1	0,63	0,55	0,48	0,28	0,37	0,45	0,52	0,72
z1_g3_t2	0,58	0,53	0,47	0,36	0,42	0,47	0,53	0,64
z1_g4_t1	0,61	0,56	0,50	0,29	0,39	0,44	0,50	0,71
z1_g4_t2	0,58	0,53	0,47	0,37	0,42	0,47	0,53	0,63
z2_g1_t1	0,74	0,67	0,60	0,23	0,26	0,33	0,40	0,77
z2_g1_t2	0,69	0,63	0,57	0,27	0,31	0,37	0,43	0,73
z2_g2_t1	0,68	0,63	0,58	0,31	0,32	0,37	0,42	0,69
z2_g2_t2	0,61	0,57	0,52	0,35	0,39	0,43	0,48	0,65
z2_g3_t1	0,61	0,55	0,48	0,31	0,39	0,45	0,52	0,69
z2_g3_t2	0,55	0,52	0,50	0,42	0,45	0,48	0,50	0,58
z2_g4_t1	0,60	0,55	0,50	0,34	0,40	0,45	0,50	0,66
z2_g4_t2	0,54	0,54	0,53	0,51	0,46	0,46	0,47	0,49
z3_g1_t1	0,71	0,64	0,56	0,22	0,29	0,36	0,44	0,78
z3_g1_t2	0,68	0,62	0,55	0,33	0,32	0,38	0,45	0,67
z3_g2_t1	0,66	0,57	0,47	0,26	0,34	0,43	0,53	0,74
z3_g2_t2	0,58	0,51	0,44	0,31	0,42	0,49	0,56	0,69
z3_g3_t1	0,60	0,53	0,46	0,30	0,40	0,47	0,54	0,70
z3_g3_t2	0,53	0,50	0,47	0,43	0,47	0,50	0,53	0,57
z3_g4_t1	0,59	0,57	0,54	0,32	0,41	0,43	0,46	0,68
z3_g4_t2	0,50	0,49	0,49	0,47	0,50	0,51	0,51	0,53

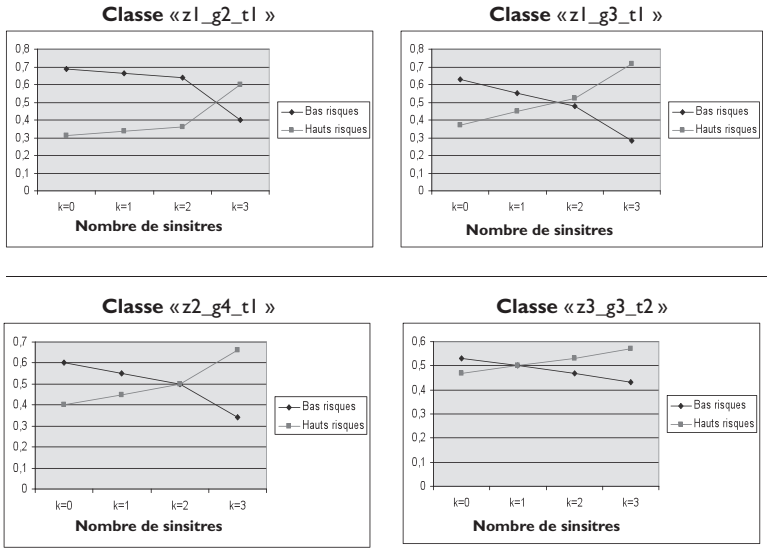
(*) Pour une valeur de *k* fixée, on a toujours : Pr (*H*/*k*) = 1 – Pr(*B*/*K*)

– Dès qu'un assuré a trois sinistres dans l'année de référence (et a fortiori plus de trois sinistres), la probabilité d'être un conducteur à hauts risques peut dépasser les 70 % dans certaines classes.

– La classe « z2_g4_t2 » est particulière dans la mesure où les paramètres de la distribution des sinistres pour les assurés à bas risques sont presque les mêmes que pour les assurés à hauts risques. Dans cette classe la sinistralité passée n'est pas un bon indicateur de la sinistralité future. D'autres hypothèses peuvent aussi être invoquées : des erreurs dans le fichier, une mauvaise estimation de la sinistralité passée,…La présence de 2 valeurs extrêmes (8 et 9 sinistres) parmi les assurés à hauts risques a peu d'influence sur les paramètres de la distribution du fait que cette classe a un effectif important.

La figure 1 compare les probabilités d'être un conducteur à bas risques ou hauts risques selon le nombre de sinistres pour deux classes représentatives de l'ensemble de ces classes.

FIGURE 1
PROBABILITÉ D'ÊTRE UN CONDUCTEUR À BAS RISQUES OU HAUTS RISQUES SELON LE NOMBRE DE SINISTRES (POUR QUATRE CLASSES)



Ces graphiques mettent en évidence une augmentation (en valeurs absolues) de la pente de la courbe des probabilités lorsque l'on passe de deux sinistres à trois sinistres (sauf pour la dernière classe).

Plus l'écart à l'origine (k=0) entre les deux courbes est important, plus elles mettent de « temps » à se croiser. Autrement dit, plus le nombre de sinistres doit être important pour que la probabilité d'être un assuré à hauts risques dépasse celle d'un assuré à bas risques.

4. CONCLUSION

La probabilité d'être un bon conducteur dans le futur, pour ceux qui étaient des assurés à bas risques dans le passé, diminue avec le nombre de sinistres déclarés au cours de l'année de référence. Un assuré à bas risques qui n'a pas déclaré de sinistres l'année de référence augmente sa probabilité d'être un futur bon conducteur.

La situation est inverse pour les assurés à hauts risques. Ils diminuent leur probabilité d'être à hauts risques en l'absence de sinistre l'année de référence et l'augmentent dans le cas contraire.

Cette étude a ses limites pour deux raisons :

– La classification a priori des assurés en deux catégories (bas risques et hauts risques) n'est qu'une estimation obtenue à partir du CRM, de l'ancienneté de permis et de l'âge du conducteur, en l'absence de données longitudinales. Par exemple, il est possible qu'un assuré soit déclaré à hauts risques, en raison de son CRM élevé par rapport à son ancienneté de permis, non à cause d'une sinistralité antérieure, mais du fait de ne pas avoir contracté d'assurance durant plusieurs années. Une vérification longitudinale sur quelques données laisse à penser que ces erreurs sont minimes et n'influent pas sur les résultats, d'autant plus que l'échantillon est de grande taille et que la population cible de cette mutuelle est assez homogène.

– On sait qu'un certain nombre d'assurés ne déclarent pas tous les sinistres. La valeur zéro peut correspondre à deux sous-populations : les assurés qui n'ont eu aucun sinistre dans l'année (cas général) et ceux qui ont eu un accident responsable et qui ne l'ont pas déclaré à l'assureur. La non-déclaration peut correspondre à un accident mineur, en indemnisant directement la partie adverse pour éviter d'avoir un malus et une augmentation de sa prime. Elle peut aussi correspondre à un délit de fuite.

On pourrait envisager une étude similaire à celle de Boucher et Denuit (2008) avec des modèles à inflation de zéros.

Notre démarche est proche de celles utilisées en théorie de la crédibilité par les actuaires pour la tarification de contrats d'assurance, dans la mesure où l'on s'appuie sur l'historique des sinistres de l'assuré et sur la prime actuelle.

D'un point de vue pratique, l'assureur pourrait utiliser cette approche probabiliste pour se fixer un seuil (différent selon les deux catégories de conducteurs; par exemple $\Pr(H/k) > 70\,\%$ pour les conducteurs à hauts risques et $\Pr(B/k) < 45\,\%$ pour les conducteurs à bas risques, cf. tableau 6) au-delà duquel il pourrait envisager une action de sensibilisation à la sécurité routière destinée à ces conducteurs potentiellement à hauts risques.

Globalement, les assurés à bas risques demeureront de bons conducteurs et les assurés à hauts risques de mauvais conducteurs. « Le temps ne fait rien à l'affaire » comme disait G. Brassens.

Références

Boucher J.P., Denuit M. (2008) *Credibility premium for Zero-Inflated model and new hunger bonus interpretation*, Insurance: Mathematics and Economics, 42, 727-735.

Cameron A.C., Trivedi P.K. (1998) *Regression Analysis of Count Data*, Cambridge University Press.

Denuit M., Maréchal X., Pitrebois S., Wahlin J-F. (2007), *Actuarial Modelling of Claims Counts: Risk classification, Credibility and Bonus-malus Systems*, Ed Wiley.

Denuit M., Charpentier A. (2005), *Mathématiques de l'assurance non vie*, Tomes I et II, Economica, Paris.

Greene W.H. (1994) *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models* Working Paper EC-94-10, Department of Economics, Stern school of Business, New York University.

Grun-Réhomme M. (2000), Prévision du risque et tarification : le rôle du bonus-malus français, *Assurances et gestion des risques*, Montréal, Canada, 1, 21-30.

Karlis D. (2005) *Mixed Poisson Distribution*, International Statistical Review, 73(1).

Klugman S.A., Panjer H.H., Wilmot G.E. (2004) *Estimation, Evaluation and Selection in actuarial models* in Loss Models: from data to decisions, Ed. Wiley.

Mc Cullagh P., Nelder J.A. (1989) *Generalized Linea Models* Chapman and Hall.

Rothschild M., Stiglitz J. (1970) Increasing Risk, *Journal of Economic Theory*, 2, 225-243.

Tobin J. (1965) The theory of portfolio selection in Hahn and Brechling "The Theory of Interest Rates", MacMillan, London.

Winkelmann R. (2003) *Econometric Analysis of Count Data*, Springer-Verlag.

Wooldridge J.M., 1997 Multiplicative panel data models without the strict exogeneity assumption, *Econometric Theory* 13, 667–678.

Yang Z., Hardin J.W., Addy C.L., Vuong Q.H. (2007) Testing Approaches for Overdispersion in Poisson Regression versus the Generalized Poisson Model *Biometrical Journal*, 49, 565–584.

Yau K.K., Wang K., Lee A.H. (2003) Zero-Inflated Negative Binomial Mixed Regression Modelling of Over-Dispersed Count Data with Extra Zeros, *Biometrical Journal*, 45, 437-452.

ANNEXE 1
MOYENNE DU NOMBRE DE SINISTRES SELON LES VARIABLES DE LA SEGMENTATION ET POUR LES DEUX CATÉGORIES D'ASSURÉS

	Zone 1	Zone 2	Zone 3	Gamme 1	Gamme 2	Gamme 3	Gamme 4	Type 1	Type 2
Bas risques	0,61	0,67	0,69	0,77	0,72	0,67	0,56	0,70	0,64
Hauts risques	0,77	0,84	0,89	1,10	1,02	0,85	0,66	1,00	0,78