**Assurances et gestion des risques**
**Insurance and Risk Management**

# Une approche locale de la gestion des sinistres graves en assurance automobile

Michel Grun-Rehomme, Noureddine Benlagha and Olaga A. Vasechko

Volume 75, Number 3, 2007

URI: https://id.erudit.org/iderudit/1092101ar
DOI: https://doi.org/10.7202/1092101ar

See table of contents

Publisher(s)

Faculté des sciences de l'administration, Université Laval

ISSN

1705-7299 (print)
2371-4913 (digital)

Explore this journal

Cite this document

Grun-Rehomme, M., Benlagha, N. & Vasechko, O. (2007). Une approche locale de la gestion des sinistres graves en assurance automobile. *Assurances et gestion des risques / Insurance and Risk Management*, *75*(3), 409–429. https://doi.org/10.7202/1092101ar

Article abstract

This article proposes to study the stability of homogeneous classes of risk of ensured in automobile insurance using an indicator of the atypical policy-holder detection. We distinguish two types of atypical points (or serious accidents). The first one corresponds to the extreme values of accident cost distribution in the whole portfolio (outliers), and the second one affects the stability of the premium in a class of risk and modify the class hierarchy built by the premium (inliers, local point of view). This stability is necessary to obtain a good adequacy between the accidents and the insurance pricing in the context of a strongly competing market. In each homogeneous class, the risk is measured in term of frequency and average cost, then the premium corresponds to the esperance of the losses is given like the product of these two indices. This indicator of premium, on the one hand, makes it possible to treat the classes on a hierarchical basis, and on the other hand, it is used as a base for the reference premium calculation. The final insurance pricing is necessarily determined by multiplying the amount of the actuarial premium by the reduction increase coefficient (the bonus-malus in France), where past accidents are used to fix the premium in the next period. The presence of serious accidents comes to disturb this hypothesis of collective risk differentiation from one class to another and the temporal stability of the premium indicator. These indicators are very sensitive to extreme values. The approach suggested for the inlier's detection, is based on the premium variance estimation to precise a fixed risk of error. A numerical application on real data of insurance is presented to put this approach into practice.

*Assurances et gestion des risques, vol. 75(3), octobre 2007, 409-430*
*Insurance and Risk Management, vol. 75(3), October 2007, 409-430*

# Une approche locale de la gestion des sinistres graves en assurance automobile

**par Michel Grun-Rehomme, Noureddine Benlagha et Olga A. Vasechko**

**RÉSUMÉ**

Cet article se propose d'étudier la stabilité de classes de risque homogènes d'assurés en assurance automobile à l'aide d'un indicateur de détection des assurés atypiques. On distingue deux types de points atypiques (ou de sinistres « graves ») : ceux qui correspondent aux valeurs extrêmes de la distribution du coût des sinistres dans le portefeuille (« outliers ») et ceux qui affectent la stabilité de la prime pure dans une classe de risque et qui engendrent une modification de la hiérarchie des classes établie avec la prime pure (« inliers », point de vue local). Cette stabilité est nécessaire pour obtenir une bonne adéquation entre la sinistralité et les cotisations des assurés dans ce contexte d'un marché fortement concurrentiel. Dans chaque classe homogène, le risque est mesuré en terme de fréquence et de coût moyen, puis la prime pure qui correspond à l'espérance des pertes est déterminée comme le produit de ces deux indices. Cet indicateur de prime pure permet d'une part de hiérarchiser les classes et d'autre part, il sert de base au calcul de la prime de référence. La prime payée par l'assuré est égale à la prime de référence multipliée par le coefficient réduction majoration (bonus-malus) de l'assuré. La présence de sinistres graves vient perturber cette hypothèse de différenciation du risque collectif d'une classe à l'autre et la stabilité temporelle de cet indicateur de prime pure. Ces indicateurs, calculés en quelque sorte par des moyennes, sont très sensibles aux valeurs extrêmes. La détection d'assurés atypiques permet de les isoler dans le calcul de la prime pure. La démarche proposée pour détecter les « inliers » est basée sur une estimation de la variance de l'indicateur de prime pure, pour une précision souhaitée (calculée sur la différence entre classes successives) et un risque d'erreur fixé. Une application numérique sur des données réelles d'assurance est présentée pour mettre en pratique cette démarche.

***Mots-clés*** : Assurance automobile, sinistres graves, classes de risque, prime pure, variance

**Les auteurs :**

Michel Grun-Rehomme, Université Paris 2, ERMES-UMR7017-CNRS, 92 rue d'Assas, 75006 Paris; Noureddine Benlagha, Université Paris 2, ERMES-UMR7181-CNRS, 92 rue d'Assas, 75006 Paris, France; Olga A. Vasechko, Research Institute of Statistics, 3 Shota Rustaveli str., 01023 Kyiv, Ukraine.

**ABSTRACT**

This article proposes to study the stability of homogeneous classes of risk of ensured in automobile insurance using an indicator of the atypical policy-holder detection. We distinguish two types of atypical points (or serious accidents). The first one corresponds to the extreme values of accident cost distribution in the whole portfolio (outliers), and the second one affects the stability of the premium in a class of risk and modify the class hierarchy built by the premium (inliers, local point of view). This stability is necessary to obtain a good adequacy between the accidents and the insurance pricing in the context of a strongly competing market. In each homogeneous class, the risk is measured in term of frequency and average cost, then the premium corresponds to the esperance of the losses is given like the product of these two indices. This indicator of premium, on the one hand, makes it possible to treat the classes on a hierarchical basis, and on the other hand, it is used as a base for the reference premium calculation. The final insurance pricing is necessarily determined by multiplying the amount of the actuarial premium by the reduction-increase coefficient (the bonus-malus in France), where past accidents are used to fix the premium in the next period. The presence of serious accidents comes to disturb this hypothesis of collective risk differentiation from one class to another and the temporal stability of the premium indicator. These indicators are very sensitive to extreme values. The approach suggested for the inlier's detection, is based on the premium variance estimation to precise a fixed risk of error. A numerical application on real data of insurance is presented to put this approach into practice.

***Keywords*:** Car insurance, outliers, inliers, risk management, variance

# 1. INTRODUCTION

## Le contexte français

Rappelons qu'en France, on distingue deux catégories d'assurance automobile, les sociétés d'assurances regroupées au sein de la FFSA (Fédération Française des Sociétés d'Assurances) et les mutuelles regroupées dans le GEMA (Groupement des Entreprises Mutuelles d'Assurance). Ces assureurs doivent maintenant faire face à un concurrent de taille : la bancassurance. Les banques qui traitent sans intermédiaire avec leurs clients à partir de leurs agences ou en utilisant des courtiers commencent à proposer des contrats d'assurance.

Les cotisations 2005 s'élevaient pour l'assurance automobile à 17,9 milliards d'euros, soit presque 43 % des assurances de dommages aux biens et de responsabilité (assurance non vie). La société Axa et la mutuelle Macif sont en tête du marché. Six assureurs dépassaient en 2005 le milliard d'euros de chiffre d'affaires en assurance automobile : Axa, Macif, Groupama-Gan, Maaf, Maif-Filiamaif, AGF. Des mouvements de rapprochement entre assureurs et banques sont annoncés.

Ces dernières années, le marché a subi un phénomène de déflation, attisé par le gouvernement, pour répondre à la baisse de la mortalité

sur les routes. Les tarifs de 2007 ne sont en baisse que de 2 % contre plus de 5 % les deux années précédentes. La baisse des prix a posé des problèmes de rentabilité. Une surenchère marketing s'installe pour préserver les marges. Les assureurs optimisent donc leurs dépenses (coûts de réparation) et se lancent dans des stratégies de différenciation de l'offre. La concurrence très forte sur le marché de l'assurance automobile conduit les acteurs à proposer des prix très serrés. Les alliances entre assureurs deviennent stratégiques dans ce processus d'industrialisation de la gestion des sinistres.

## Le problème de l'antisélection

Le caractère obligatoire de l'assurance, lié au risque de responsabilité civile de l'assuré, a pour avantage de limiter les effets de l'antisélection susceptibles de conduire à l'inassurabilité du risque, au cas où seuls les conducteurs certains d'être sinistrés s'assureraient. Mais des études empiriques récentes (Dionne, Doherty et Frombaron (2001), Dionne, Gouriéroux et Vanasse (2001)) sur le marché des assurances automobiles montrent l'existence de preuves en contradiction avec la théorie de l'antisélection. On s'était déjà interrogé (Grun-Réhomme et Joly, 2003) sur la question de savoir si le choix de la formule de garantie par l'assuré révèle une information sur les risques que ce dernier fait courir à l'assureur (et à lui-même) et sur son aversion au risque dans cette situation d'échange en information asymétrique (risque moral et sélection adverse).

À partir des données de l'ensemble du portefeuille d'une mutuelle française (3,3 millions d'observations), on avait montré que l'assuré, à travers le choix de la formule de garanties traduit principalement son aversion à la perte de l'investissement qu'il vient de réaliser. Il y est sensible en valeur absolue et non en terme de perte patrimoniale. L'aversion au risque de l'individu correspond non pas à l'espérance mathématique de la perte, mais au niveau de perte maximale. L'hypothèse faite souvent par la théorie économique, que le risque de perte du véhicule est relativisé par son poids dans le patrimoine global de l'assuré est mise en défaut.

Il faut toutefois noter que les assurés ont conscience de leur risque objectif puisque, à valeur de véhicule constante, ils cherchent à se couvrir plus fortement lorsque ce dernier augmente. Par contre, le coefficient réduction majoration (CRM), censé représenter l'expérience de conduite, joue de façon inverse : les assurés les moins risqués (ou les plus expérimentés), à valeur de véhicule constante, privilégient une forte couverture, à l'inverse des moins expérimentés qui souscrivent des formules moins protectrices. On retrouvait aussi les résultats de Dionne, Doherty et Frombaron (2001), à savoir que

les assurés à bas risques ont plus d'aversion pour le risque que les assurés à hauts risques, ainsi que le montrent les résultats de Dionne, Gouriéroux et Vanasse (2001) sur l'absence de sélection adverse résiduelle dans les classes de risque.

## Les classes de risque homogènes

Le coût du risque individuel de chaque assuré n'est pas prévisible et n'est connu qu'à 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 et la représentativité de leurs portefeuilles.

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. À 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.

Le taux de sinistre varie d'une classe à l'autre d'environ 5 à 8 % et les facteurs retenus pour constituer les classes de risque (le risque collectif) expliquent environ 75 % de la variance totale des coûts des sinistres, le quart restant étant attribué à la composante individuelle. Les sinistres corporels représentent environ 3 % du nombre total de sinistres indemnisés, mais plus du quart du coût total des sinistres.

Après une présentation de l'indicateur de prime pure qui sert de base pour le calcul de la prime payée par l'assuré (en section 2). Nous rappelons en section 3, rappelle la problématique des sinistres graves. Nous proposons, en section 4, une approche locale de détection des sinistres graves, en fonction des classes de risque. Cette approche est basée sur une estimation de la variance de l'indicateur

de prime pure. Nous appliquons cet indicateur sur des données réelles échantillonnées d'un portefeuille d'assurance. Une conclusion termine cette présentation.

## 2. LA PRIME PURE

Dans chaque classe, le risque est mesuré en terme de fréquence et de coût moyen, puis la prime pure est déterminée comme produit de ces deux indices. Plus précisément, notons $k$ une classe de risque ($k = 1, \ldots K$) et $n_k$ le nombre de véhicules années dans la classe $k$. La prime pure dans la classe $k$ est alors définie par :

$$P_k = \frac{\sum_{i=1}^{n_k} c_{k,i}}{\sum_{k=1}^{n_k} w_{k,i}}$$

où

$c_{k,i}$ correspond au coût du sinistre du véhicule assuré $i$ de la classe $k$.

De nombreux coûts sont nuls.

$w_{k,i}$ correspond au poids du véhicule assuré $i$ de la classe $k$. En effet, au cours d'une année, le nombre d'assurés dans une classe varie, certains arrivent, d'autres résilient leur contrat ou changent de véhicule. Chaque observation $i$ est donc pondérée par $w_{k,i} \frac{1}{12} \times$ (nombre de mois où l'assuré $i$ est présent dans la classe $k$). Un changement de véhicule peut impliquer un changement de classe. Un assuré présent 3 mois dans l'année, aura donc un poids égal à 0,25 (sa cotisation ne correspond qu'à trois mois d'assurance). Par conséquent, $\sum_{i=1}^{n_k} w_{k,i} \leq n_k$ .

Ces indicateurs sont en général normés (en divisant chaque indicateur par la prime pure de l'ensemble du portefeuille) et multipliés par 100. Ainsi la prime pure du portefeuille est égale à 100 et les primes pures des classes sont facilement interprétables par rapport à la moyenne du portefeuille. 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.

Souvent on considère que le risque est mesurable dans la mesure où il est possible de calculer un risque moyen (coût moyen) qui caractérise la tendance de sinistralité du phénomène étudié. À travers ces indicateurs, on recherche une tendance qui dépend de la nature du produit assuré, de la régularité de l'environnement et du portefeuille, de la structure de l'entreprise, ainsi que de sa gestion et de sa stratégie commerciale.

Cet indicateur de prime pure permet d'une part de hiérarchiser les classes et d'autre part, il sert de base au calcul de la prime de référence. Cette dernière tient compte du taux de chargement de l'ordre de 33 % de la prime pure (frais de gestion, frais de production, d'encaissement des primes, …), des charges fiscales (33 % pour la garantie de responsabilité civile obligatoire et 18 % pour la garantie dommages facultative, soit en général une moyenne de 24 %) et de la stratégie de l'entreprise qui doit assurer la pérennité de l'entreprise dans un marché concurrentiel. Pour se protéger de sa méconnaissance a priori du montant total des sinistres et pour pouvoir résister à la volatilité des sinistres, il ajoute à la prime pure, un chargement de sécurité qui peut dépendre d'un certain quantile des pertes ou de l'écart-type des pertes ou être proportionnel à la prime pure. Les recettes des produits financiers peuvent diminuer le prix de l'assurance.

La prime payée par l'assuré est égale à la prime de référence multipliée par le coefficient réduction majoration (bonus-malus) de l'assuré (cf. Grun-Réhomme, 2000). Le système bonus-malus permet de garantir une solidarité minimale et suffisante entre assurés de classes de risque différentes.

Une remarque à propos des coûts des sinistres : tous les sinistres ne sont pas réglés en terme financier l'année où ils surviennent. On peut estimer qu'environ un tiers d'entre eux sont clos la même année, un petit tiers l'année suivante et de 10 à 15 % la troisième année. Les règlements sont en général plus longs pour les accidents corporels. Pour la gestion financière de la compagnie d'assurance, il est donc nécessaire d'effectuer des estimations des coûts des sinistres non réglés l'année en cours. L'assureur, en s'appuyant sur son expérience passée, établit des provisions pour frais au cas par cas. Dans le calcul de la prime pure de l'application numérique proposée par la suite, il s'agit donc des coûts réels ou de coûts estimés.

La fréquence des accidents n'est pas le seul facteur à prendre en compte dans la variation du prix de l'assurance, il faut tenir compte

aussi des coûts des sinistres (augmentation en moyenne de 5 % pour les sinistres matériels et de 6 % pour les sinistres corporels), de la capitalisation nécessaire et de la réassurance.

Pour les accidents corporels, l'assureur peut connaître le montant total de ses dépenses seulement après plusieurs mois, voire plusieurs années, d'où cette nécessité de faire une prévision de ses dépenses. Le coût des sinistres corporels comprend plusieurs composantes : indemnités aux personnes physiques (IPP), soins, tierce personne, préjudices personnels et économiques.

## 3. LES SINISTRES GRAVES

En pratique, on constate un décalage entre le moment où les indicateurs de risque sont calculés dans les classes et le moment où l'étude est prise en compte au niveau de la tarification. Au meilleur des cas, il faut compter un délai de deux ans.

D'une manière générale les risques des individus d'une classe homogène de risques dépendent de deux variables aléatoires indépendantes et équidistribuées : une variable structurelle qui caractérise l'hétérogénéité interindividuelle acceptée au sein de la classe et une variable endogène qui correspond au risque collectif de la classe. C'est cette dernière que l'assureur cherche à prévoir.

La présence de sinistres graves vient perturber cette hypothèse de différenciation du risque collectif d'une classe à l'autre et la stabilité temporelle des indicateurs. Ces indicateurs, calculés par des moyennes, sont très sensibles aux valeurs extrêmes. Il est nécessaire d'envisager des scénarios extrêmes qui tiennent compte des queues épaisses de la distribution du coût des sinistres.

La distribution du coût des sinistres est fortement asymétrique et la pertinence de la moyenne peut être mise en cause, non d'un point de vue comptable mais comme paramètre de localisation de la distribution. La moyenne est pertinente si la distribution est peu dispersée. La forme de la répartition de ces coûts est relativement semblable quel que soit l'effectif de la classe de risques. On a une invariance d'échelle. La fréquence de survenance d'un sinistre est aussi faible que l'amplitude est grande.

Un sinistre grave peut correspondre à un montant pour lequel un contrat de réassurance intervient ou qui est traité différemment par un autre service. La distribution des sinistres graves est modélisée

par une loi de Pareto généralisée. Le coût d'un sinistre grave est souvent moins bien modélisé que les sinistres usuels, tout simplement du fait d'un faible nombre de données disponibles.

L'avantage d'une méthode de modélisation avancée sur une méthode plus simple réside souvent dans une meilleure modélisation des petites particularités de l'échantillon, mais cet avantage peut être illusoire du fait de plusieurs éléments (évolution de la population, choix de l'échantillon, …).

La pratique actuelle des assureurs consiste souvent à écrêter la distribution des coûts des sinistres uniformément selon les classes de risque (ou avec quelques aménagements) et à répartir cette charge supplémentaire sur l'ensemble du portefeuille. Cette démarche a l'avantage d'être rapide, simple et de garder un équilibre entre la sinistralité et les cotisations (ou primes payées) dans le ratio : sinistralité / cotisations. Mais elle présente l'inconvénient d'une part de ne pas prendre en compte la particularité des classes (plus ou moins risquées) et d'autre part cette troncature uniforme trop grossière peut venir perturber la hiérarchie des classes de risque et donc l'adéquation entre la prime de référence et la sinistralité.

## 4. APPROCHE LOCALE DE LA DÉTECTION DES SINISTRES GRAVES

Les évènements rares sont des évènements chers pour l'assureur.

On peut distinguer deux types de points atypiques (ou de sinistres graves). D'un point de vue général pour une distribution donnée $X$ (ici la variable du coût des sinistres).

Un point est dit **atypique global** pour $X$ si sa valeur est extrême pour cette distribution, c'est à dire si elle se situe en fin de distribution au-delà d'un certain seuil.

Un point est dit **atypique local** pour $X$ si sa valeur en $X$ est extrême à l'intérieur de la distribution, dans la sous distribution de ses points voisins. Un point atypique local affecte un sous-ensemble de données mais pas forcément l'ensemble des données.

Bien sûr tout point atypique global est un point atypique local. Une valeur est un point atypique local non parce qu'elle appartient à la catégorie des grands nombres, mais parce qu'elle l'est relativement aux valeurs prises par des observations de la même catégorie.

Dans la terminologie anglo-saxonne, un point atypique global se nomme « outlier » et un point atypique local, un « inlier » (Winkler, 1977).

Notre démarche est semblable, mais la notion de voisinage d'un point est remplacée par la notion de classe de risque. Un sinistre peut être considéré comme grave pour une classe et non pour l'ensemble du portefeuille. Une distinction entre les sinistres atypiques au niveau global (pour le portefeuille) et au niveau local (pour une classe de risque) est donc à faire.

Un sinistre sera considéré comme grave (ou atypique) pour une classe de risque donnée si son montant engendre une modification de la hiérarchie des classes établie avec la prime pure. Il faut neutraliser les valeurs extrêmes.

Le seuil à partir duquel un sinistre sera considéré comme grave localement dépend :

- de la taille $n_k$ de la classe $k$,
- de la variance de l'indicateur,
- de la précision souhaitée,
- du risque d'erreur fixé.

En considérant que l'ensemble des assurés d'une classe de risque constitue un échantillon aléatoire de la population des assurables ayant les mêmes caractéristiques, on peut par linéarisation du ratio (Cochran, 1977, Grun-Rehomme, 1998), montrer que la variance de la prime pure $P_k$ dans la classe $k$, est estimée par :

$$Var(P_k) \cong \frac{n_k}{\left(\sum_{i=1}^{n_k} w_{k,i}\right)^2} Var(z_k) \qquad (1)$$

où pour toute observation $i$, $z_{k,i} = c_{k,i} - P_k\, w_{k,i}$ (cf. la démonstration en annexe 1).

La variance de $z_k$ peut être calculée dans chaque classe $k$. Pour une précision $\varepsilon_k$ et un risque d'erreur $\alpha_k$ donnés, on montre d'après l'inégalité de Biénaymé-Tchebychev que la variance de l'indicateur doit vérifier :

$$Var(P_k) \leq \alpha_k \times \varepsilon_k^2 \qquad (2)$$

ou encore,

$$Var(z_k) \leq \frac{\left(\sum_i w_{k,i}\right)^2}{n_k} \times \alpha_k \times \varepsilon_k^2 . \qquad (3)$$

On suppose que les classes sont ordonnées selon la valeur de la prime pure, $P_k < P_{k+1}$ .

On fixe, par exemple, un risque d'erreur $\alpha_k = \alpha = 0{,}05$ ou $0{,}1$ pour toutes les classes et on peut également fixer $\varepsilon_k$ pour la classe $k$ par :

$$\varepsilon_k \leq P_{k+2} - P_k \qquad (4)$$

ou

$$\varepsilon_k \leq P_{k+1} - P_k \qquad (4b)$$

Cette différence sur les indicateurs est calculée dans le passé récent. On peut également se contenter d'une précision moins grande, en choisissant

$$\varepsilon_k \leq (P_{k+2} - P_k) + z_{\alpha/2} \frac{\sqrt{Var(z_{k+2})}}{n_{k+2}} \qquad (5)$$

ou

$$\varepsilon_k \leq (P_{k+1} - P_k) + z_{\alpha/2} \frac{\sqrt{Var(z_{k+1})}}{n_{k+1}} \qquad (5b)$$

où $z_{\alpha/2}$ correspond à la valeur de la loi normale centrée réduite pour $\alpha$ (ou $\alpha_k$). Dans cette démarche, on contrôle le dépassement de la prime pure de la classe $k + 1$ ou $k + 2$ par celle de la classe $k$. Il nous semble, à l'examen des données, que le respect strict de la hiérarchie des classes est trop exigeant si l'effectif de la classe de risque est faible et qu'il est alors préférable d'un point de vue pratique de respecter la hiérarchie à une unité près. On pourrait également contrôler les comparaisons dans l'autre sens.

On ordonne également dans chaque classe $k$, la variable $z_k$ ($z_{k,i} \leq z_{k,i+1}$), ce qui correspond presque toujours au même ordonnancement sur les coûts.

Posons $V(i) = Var\{z_{k,j} / j \leq i\}$, la variance des $i$ premières observations de la classe $k$.

Soit $q_k$ le plus petit entier tel que la variance de la distribution cumulée $z_k$ jusqu'à cet ordre dépasse la quantité $\frac{\left(\sum_i w_{k,i}\right)^2}{n_k} \times \alpha_k \times \varepsilon_k^2$. Cette valeur constitue un seuil au dessus duquel un sinistre peut être considéré comme grave pour la classe de risque correspondante.

Les sinistres graves de la classe *k* correspondent à l'observation de rang $q_k$ et aux observations de rang supérieur. Il est possible de ne pas avoir de sinistres graves dans une classe si la variance totale ne dépasse cette quantité.

En pratique, $V(i)$ est une fonction croissante de $i$ (cf. annexe 2).

## 5. APPLICATION NUMÉRIQUE

Avant de présenter une application numérique, rappelons la démarche pratique habituelle des assureurs et comment notre approche locale des sinistres graves s'inscrit dans cette démarche.

La modélisation de la sinistralité se fait, en général, à partir de deux échantillons indépendants et souvent sur deux exercices consécutifs, afin de limiter les biais d'échantillonnage et d'obtenir une certaine robustesse des indicateurs. Les variables retenues, explicatives de la sinistralité, (et conformes aux stratégies de l'entreprise) servent à construire les classes de risque pour le portefeuille.

Pour mettre en oeuvre cette approche de détection locale des sinistres graves, nous disposons d'un fichier de 60 000 observations d'une compagnie française d'assurance. Ces données concernent des véhicules 4 roues de tourisme assurés durant l'année 2004.

Les classes de risque sont construites à partir de caractéristiques du conducteur (ancienneté de permis, le type de conducteur), de caractéristiques du véhicule (ancienneté et puissance) 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 (mais cette dimension est déjà prise en compte par l'ancienneté de permis et le bonus-malus) et dans une moindre mesure pour les conjoints. Nous disposons aussi de la période de couverture : période, exprimée en mois, au cours de laquelle l'assuré est couvert par la police qu'il a souscrit, le plus souvent cette période est d'une année.

Pour des raisons de confidentialité, toutes les variables de construction des classes ne sont pas utilisées et la description précise des classes n'est pas donnée. La connaissance des classes et des primes pures associées permettrait à un concurrent de connaître les ratios sinistres / cotisations de cet assureur, puisqu'il est toujours possible de se renseigner sur le montant de la cotisation.

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 et informatique reste la même. Il est bien sûr impossible de « sortir » les données individuelles, qui sont nécessaires pour cette étude, de l'ensemble du portefeuille.

## Analyse exploratoire des données

Le fichier contient environ autant d'hommes que de femmes. Dans 30 % des cas, l'assuré n'est pas le conducteur principal.

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

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

2- Dommages au véhicule (DV1) : RC + Garantie Dommage au véhicule toutes causes avec une franchise importante (par exemple, 600 Euros pour une Renault Clio et 830 pour une Renault Laguna).

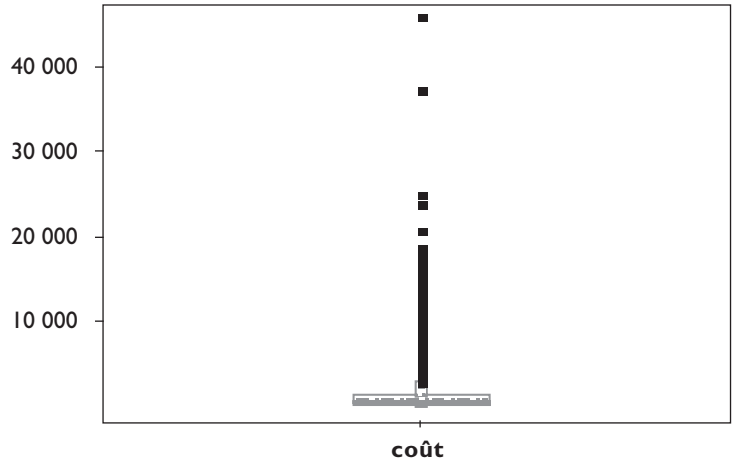
3- DV2 : RC + Garantie Dommage au véhicule toutes causes avec une franchise moyenne (200 Euros pour une Renault Clio et 250 pour la Laguna).

4- DV3 : RC + Garantie Dommage au véhicule toutes causes avec une franchise faible (70 Euros pour une Renault Clio et 83 pour une Laguna).

La répartition des assurés selon ces garanties est la suivante : RC (45 %), DV1 (16 %), DV2 (8,5 %) et DV3 (30,5 %).

Le graphique boxplot montre qu'avec un écrêtage à 20 000, l'échantillon comporte 5 sinistres graves (« outliers »).

**GRAPHIQUE 1**
**BOXPLOT DE LA DISTRIBUTION DU COÛT DES SINISTRES NON NULS**



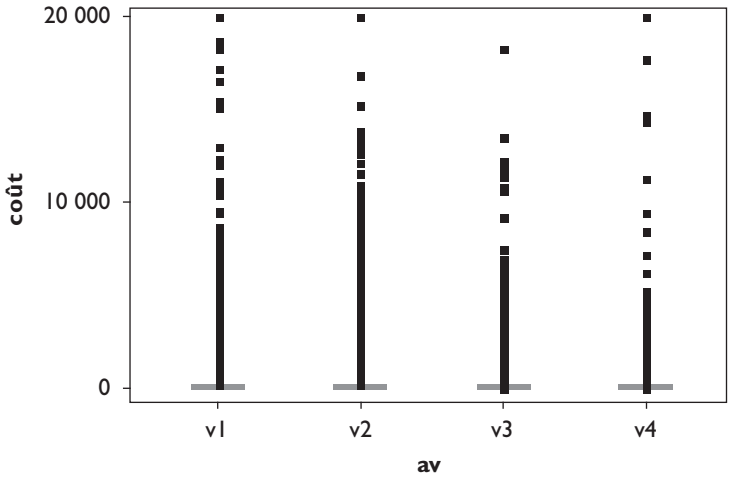

Quelques statistiques sur les variables numériques.

| | Min | $Q_1$ (25 %) | Médiane | Moyenne | $Q_3$ (75 %) | Max |
|---|---|---|---|---|---|---|
| **Coût des sinistres non nuls** | 37 | 128 | 468 | 967 | 1 168 | 45 972 |
| **Ancienneté du véhicule** | 0 | 3 | 7 | 7.5 | 11 | 83 |
| **Âge du conducteur** | 18 | 35 | 48 | 47 | 57 | 98 |
| **Puissance du véhicule*** | 24 | 60 | 76 | 83 | 100 | 485 |

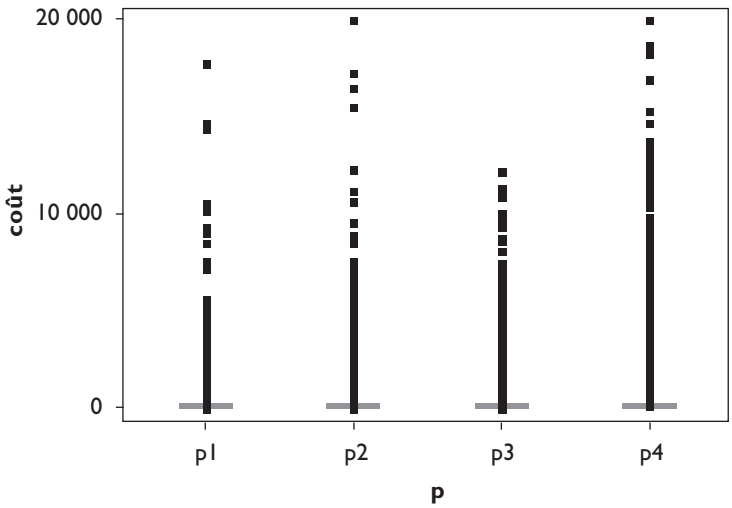
* La puissance (réelle) du véhicule exprime la puissance du moteur en chevaux Din (Deutsch Industrie Normen). Cette mesure donne une vision plus réaliste de la puissance effective au niveau des roues (1 ch. Din = 0,735 Watt). Par exemple, la Fiat Uno, la Renault Clio et la Peugeot 205 ont 60 ch. Din, et les Peugeot 206, Renault Laguna et Citroen Picasso, 110 ch. Din.

Les coûts augmentent en moyenne de façon logarithmique avec la puissance du véhicule et ils diminuent de façon parabolique (concave) avec l'ancienneté du véhicule. La dispersion de ces coûts suit les mêmes monotonies. Les véhicules puissants et les véhicules de marque étrangère sont plus fréquents chez les hommes que chez les femmes.

**GRAPHIQUE 2**
**BOXPLOT DES COÛTS DES SINISTRES SELON L'ANCIENNETÉ DU VÉHICULE (4 MODALITÉS)**




**GRAPHIQUE 3**
**BOXPLOT DES COÛTS DES SINISTRES SELON LA PUISSANCE DU VÉHICULE (4 MODALITÉS)**



L'étude de la sinistralité en fonction des caractéristiques du couple (conducteur, véhicule) a fait l'objet de nombreux travaux (cf. J.L. Martin, Y. Derrien, B. Laumon, 2003).

On aurait pu retenir l'âge du conducteur, mais cette variable est corrélée avec l'ancienneté de permis et n'est pas commercialement très appropriée. La prise en compte de l'âge fait apparaître une sur-sinistralité chez les moins de 25 ans.

Nous présentons ici une trentaine de classes de risque obtenues en regroupant plusieurs strates (croisements de modalités de variables); celles ci sont classées dans l'ordre croissant de la prime pure de la classe. Pour détecter les « inliers », nous avons retenu comme paramètres : $\alpha_k = \alpha = 0{,}1$ et $\varepsilon_k \leq P_{k+2} - P_k$ . Pour l'avant dernière classe, on a choisit $\varepsilon_k \leq P_{k+1} - P_k$ et pour la dernière classe, la détection, uniquement d'outliers, a été effectuée à l'aide du boxplot. Rappelons que, dans notre exemple, les effectifs dans les classes de risque sont plus faibles que dans celles d'un portefeuille et le nombre d'« intliers » détecté dans une classe dépend de cet effectif. Plus les écarts sont faibles entre les primes de classes consécutives, plus il est nécessaire d'avoir un effectif suffisant dans ces classes pour obtenir une stabilité de ces primes pures (cf. Grun-Rehomme, 1998).

| Classes | Prime pure | $\frac{\left(\sum w_i\right)^2}{n}$ | $\frac{\left(\sum_i w_{k,i}\right)^2}{n_k} \times \alpha_k \times \varepsilon_k^2$ | Nombre d'« inliers » |
|---|---|---|---|---|
| **1** | 45,63 | 1 015 | 23 853 | 1 |
| **2** | 58,02 | 1 701 | 74 301 | 1 |
| **3** | 60,96 | 755 | 77 990 | 0 |
| **4** | 78,92 | 2 377 | 210 238 | 0 |
| **5** | 93,10 | 2 539 | 211 473 | 2 |
| **6** | 108,66 | 1 109 | 53 043 | 1 |
| **7** | 121,96 | 2 311 | 50 894 | 3 (*) |
| **8** | 130,53 | 1 377 | 51 398 | 5 |
| **9** | 136,80 | 1 061 | 77 691 | 1 |
| **10** | 149,85 | 1 437 | 48 545 | 0 |
| **11** | 163,86 | 2 364 | 130 108 | 2 |
| **12** | 168,23 | 923 | 66 393 | 1 |
| **13** | 187,32 | 2 282 | 96 182 | 0 |
| **14** | 195,05 | 1 465 | 149 641 | 2 |
| **15** | 207,85 | 1 603 | 101 071 | 1 |
| **16** | 227,01 | 1 812 | 49 992 | 1 |
| **17** | 232,96 | 1 897 | 181 479 | 2 |
| **18** | 243,62 | 1 239 | 123 643 | 3 |
| **19** | 263,89 | 1 262 | 35 576 | 18 (**) |
| **20** | 275,21 | 3 492 | 93 121 | 0 |
| **21** | 280,68 | 1 437 | 91 618 | 2 |
| **22** | 291,54 | 2 937 | 256 807 | 1 (*) |
| **23** | 305,93 | 3 132 | 204 618 | 1 |
| **24** | 321,11 | 870 | 187 065 | 1 |
| **25** | 331,49 | 3 422 | 5 281 204 | 0 (*) |
| **26** | 367,48 | 1 148 | 2 803 811 | 0 (*) |
| **27** | 455,72 | 1 168 | . | 2 (*) |
| **Total** | 211,20 | 48 135 | | |

(*) : Classe de risque où un sinistre grave (outliers) est présent pour l'ensemble des observations.

(**) : Dans cette classe, le nombre d' « inliers » est trop important, ce qui provient d'une part de l'effectif faible de cette classe et d'une relative petite valeur de $\varepsilon_k$ et d'autre part d'une petite sous population ayant une sinistralité plus importante.

Il faut encore rappeler que ces résultats concernent un échantillon du portefeuille, ce qui présente l'inconvénient d'avoir un effectif faible dans les classes de risque, et donc d'être moins robustes.

## 6. CONCLUSION

Il est nécessaire pour l'assureur d'avoir une bonne adéquation entre la hiérarchisation des primes pures (et par conséquent des cotisations payées par les sociétaires ou les primes payées par les assurés) et la sinistralité. Une sous-estimation des primes pures peut mettre en péril la pérennité de l'entreprise et une surestimation des primes peut entraîner un départ des bons risques vers un concurrent. Tous les assureurs calculent une prime pure, considérée comme relativement stable, dans chaque classe de risque du portefeuille, après un écrêtage des sinistres graves pour le portefeuille.

Notre démarche de détection des « inliers » permet ensuite de contrôler cette adéquation entre les primes et la sinistralité. Il existe de nombreuses méthodes générales, algébriques, graphiques ou probabilistes de détection d'unités atypiques ou de valeurs extrêmes (Reiss, Thomas, 2001, Nikulin, Zerbet, 2002, Sim, Gan, Chang, 2005, Benlagha, Grun-Rehomme, Vasechko, 2006). Mais la méthode proposée ici présente l'avantage d'être simple et surtout d'être appropriée à cette problématique d'assurance, basée sur le respect (strict ou à une unité près) de la statistique d'ordre des classes de risque.

Trois raisons principales peuvent expliquer une variance importante de $z_k$ dans une classe de risque:

- La présence d'un ou de quelques « inliers » dans cette classe,
- La présence d'une sous population plus risquée, d'une niche dans ce segment qu'il est donc nécessaire de suivre avec attention,
- Un manque d'homogénéité structurelle de la classe qui peut provenir de variables non retenues ou latentes.

Cette approche locale de détection des « inliers » propose un seuil, dont le niveau est fixé par l'écart de prime pure entre deux classes de risque voisines et un risque d'erreur de 5 % (classiquement). Ce seuil permet de contrôler la variance du surcoût de sinistralité occasionné par les valeurs extrêmes. Au dessus de ce seuil, un sinistre sera considéré comme grave pour cette classe de risque. Ainsi on

dispose d'une statistique pour détecter les « inliers » et donner une réponse à une situation d'hétérogénéité (variance importante de $z_k$) dans une classe de risque qui affecte la hiérarchie de ces classes.

Pour le traitement de ces « inliers », on peut envisager une troncature, puis une mutualisation de la partie excédentaire sur l'ensemble du portefeuille.

## ANNEXE 1
## ESTIMATION DE LA VARIANCE DE LA PRIME PURE

On considère, de façon générale, que $x = \sum_{i=1}^{n} x_i$ $\left( \text{resp. } w = \sum_{i=1}^{n} w_i \right)$ est une réalisation de la variable aléatoire $X = \sum_{i=1}^{n} X_i$ $\left( \text{resp. } W = \sum_{i=1}^{n} W_i \right)$ dans un sondage aléatoire simple. Aucune hypothèse n'est formulée sur les lois de *X* et *W*. La variable *X* correspond aux coûts des sinistres et *W* au poids. On utilise alors le procédé de linéarisation d'un ratio (valable pour $n > 30$).

On pose $x = X + \delta$ (resp. $w = W + \varepsilon$) où $\delta$ (resp. $\varepsilon$) correspond à l'erreur d'échantillonnage.

On obtient,

$$\frac{x}{w} = \frac{X+\delta}{W+\varepsilon} = \frac{X\left(1+\frac{\delta}{X}\right)}{W\left(1+\frac{\varepsilon}{W}\right)} = \frac{X}{W}\left(1+\frac{\delta}{X}-\frac{\varepsilon}{W}+\ldots\right).$$

Les termes du second ordre (et a fortiori d'ordre supérieur) étant négligeables par rapport aux termes du premier ordre, on obtient:

$$Var\left(\frac{x}{w}\right) \cong Var\left(\frac{X}{W}\left(\frac{\delta}{X}-\frac{\varepsilon}{W}\right)\right) = Var\left(\frac{1}{W}\left(\delta - P_k\varepsilon\right)\right)$$

*W* est estimé par $\frac{N}{n}\sum w_i$, d'où $Var(P_k) \cong \frac{n^2}{N^2\left(\sum w_i\right)^2} Var(\delta - P_k\varepsilon)$.

D'autre part,

$$\delta - P_k\varepsilon = \sum_i \left((x_i - X_i) - P_k(w_i - W_i)\right) \cong \sum \left(\left(x_i - \frac{N}{n}x_i\right) - P_k\left(w_i - \frac{N}{n}w_i\right)\right)$$

$$= \frac{n-N}{n}\left(\sum (x_i - P_k w_i)\right).$$

Puis en posant $z_i = x_i - P_k w_i$, on obtient

$$Var(P_k) \cong \frac{n^2}{N^2\left(\sum w_i\right)^2} \times \left(\frac{n-N}{n}\right)^2 \times nV(z).$$

Finalement, l'estimation de la variance de l'indicateur est donnée par :

$$Var(P_k) \cong \frac{n}{\left(\sum w_i\right)^2} V(z).$$

## ANNEXE 2
## COMPARAISON DES MOYENNES ET DES VARIANCES

On considère trois distributions :

$X_n = (x_1, ..., x_n)$, $X_{n+1} = (x_1, ..., x_n, x)$ et $Y_{n+1} = (x_1, ..., x_n, y)$

où $0 \le x_i \le x_{i+1}$ pour tout $i = 1, ..., n-1$ et $x_n \le x \le y$.

On suppose que toutes les observations d'une même distribution ont un poids identique ($1/n$ ou $1/(n+1)$).

1. Pour les *moyennes* : $\overline{X}_n \le \overline{X}_{n+1} \le \overline{Y}_{n+1}$

Plus précisément : $\overline{Y}_{n+1} = \overline{X}_{n+1} + \dfrac{(y-x)}{n+1}$ et de façon générale si chaque observation est pondérée par $w_i$ strictement positif, avec $\sum_{i=1}^{n+1} w_i = 1$, on a : $\overline{X}_{n+1} = \overline{X}_n + w_{n+1}(x - \overline{X}_n)$ et l'expression $(x - \overline{X}_n)$ est positive.

2. Pour les *variances* :

$V(Y_{n+1}) \ge V(X_{n+1})$.

En effet,

$$V(Y_{n+1}) = \frac{1}{n+1}\left(\sum_{i=1}^{n} x_i^2 + y^2 - x^2 + x^2\right) - \left(\overline{X}_{n+1} + \frac{(y-x)}{n+1}\right)^2$$

$$= V(X_{n+1}) + \frac{y^2 - x^2}{n+1} - \frac{2(y-x)\overline{X}_{n+1}}{n+1} - \frac{(y-x)^2}{(n+1)^2}$$

et $V(Y_{n+1}) \ge V(X_{n+1})$ si et seulement si $n(y+x) \ge 2\left(\sum_{i=1}^{n} x_i\right)$, ce qui est vérifié puisque $x_n \le x \le y$.

D'autre part,

$$V(X_{n+1}) = E(X_{n+1}^2) - (E(X_{n+1}))^2 = \frac{n}{n+1} \times \frac{1}{n}\left(\sum_{1}^{n} x_i^2 + x^2\right) - \left(\frac{n}{n+1}\overline{X}_n + \frac{x}{n+1}\right)^2$$

et $$V(X_{n+1}) = \frac{n}{n+1}V(X_n) + \frac{n}{(n+1)^2}\overline{X}_n^2 + \frac{x^2}{n+1} - \frac{x^2}{(n+1)^2} - \frac{2n}{(n+1)^2}\overline{X}\,x$$

d'où $$V(X_{n+1}) = \frac{n}{n+1}V(X_n) + \frac{n}{(n+1)^2}(\overline{X}_n - x)^2$$

Pour $n$ grand et $x$ grand, strictement supérieur à $x_n$, ce qui est le cas en pratique pour les valeurs extrêmes de la distribution des coûts des sinistres dans les classes de risque, on a : $V(X_n) \le V(X_{n+1})$, la variance étant une fonction quadratique de $x$.

Mais cette relation n'est pas théoriquement vérifiée dans tous les cas. Par exemple, la distribution {1, 2, 2} a pour variance 32/144 et la variance de la distribution {1, 2, 2, 2} est 27/144.

## Références


Cochran W.G. (1977), *Sampling Techniques,* Ed. Wiley, New-York, 1977.

Dionne G., Doherty N., Fombaron N. (2001), *Adverse Selection in Insurance Markets,* Handbook of Insurance, Kluwer Academic Publishers, Boston, 185-243.

Dionne G., Gouriéroux C., Vanasse C. (2001), *Testing for Evidence of Adverse Selection in the Automobile Insurance Market,* Journal of Political Economy, vol. 109, no 2, 444-453.

Grun-Rehomme M. (1998), *Étude de la stabilité des indicateurs de risque en Assurance,* revue « Risques », Les Cahiers de l'Assurance, no 35, p. 111-119, 1998.

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

Grun-Rehomme M., Joly V. (2003), *Risque individuel et choix de contrat : Le cas de l'assurance automobile,* revue Assurances et gestion des risques, vol. 71-1, 145-162, Montréal, Canada.

Mandelbrot B. (1997), *Fractales, hasard et finance,* Ed. Flammarion, Paris.

Martin J.L., Derrien Y., Laumon B. (2003), *Estimating relative driver fatality and injury risk according characteristics of cars and drivers using matched-pair multivariate analysis,* ESV, Proceedings, Nagoya.

Nikulin M., Zerbet A. (2002), *Détection des observations aberrantes par des méthodes statistiques*, RSA, L (3), 25-51.

Reiss R., Thomas M. (2001), *Statistical Analysis of extreme values*, Birkhauser Verlag.

Sim C.H., Gan F.F., Chang T.C. (2005), *Outlier labelling with Boxplot Procedures*, JASA, vol. 100, no 470, 642-652.

Vaesechko O., Grun-Rehomme M., Benlagha N. (2006), *Panorama of methods for detecting outliers in structural business surveys: implementation on French and Ukrainian data,* European Conference on Quality in Survey Statistics (Q2006), 24-26 April, Cardiff, UK.

Winkler W.E., (1997), *Problems with inliers,* European Conference of Statisticians, October 14-17, 1997, Prague, Czech Republic.