EXPLOITATION DU SONDAGE AUTOMOBILE 1971 EN FRANCE PAR UNE MÉTHODE D'ANALYSE MULTIDIMENSIONNELLE

L'Association Générale des Sociétés d'Assurances contre les Accidents

Paris

Sous la responsabilité de l'Association Générale des Sociétés d'Assurances contre les Accidents, un sondage au 1/50e a été effectué en 1971 dans les portefeuilles d'un grand nombre de Compagnies ou Mutuelles pratiquant en France l'assurance de responsabilité civile automobile.

Trois objectifs étaient visés:

Etudier l'influence de critères de tarification tels que l'âge du véhicule, l'âge du souscripteur ou l'ancienneté du permis, zone, groupe et usage.

Déterminer, pour chaque classe du tarif, les primes pures.

Tirer les enseignements des résultats de ces travaux, en particulier, étudier les possibilités d'amélioration du tarif à la fois sur le plan technique (choix des critères et méthodes de calcul des primes) et sur le plan politique (réalisation effective du tarif).

Les informations apportées par le sondage se présentaient comme suit:

l'unité statistique était constituée par l'ensemble souscripteur — véhicule — durée d'observation. Au cours de cette durée, les caractéristiques du souscripteur, du véhicule et de l'environnement, étaient inchangées. A chaque unité statistique étaient rattachés les renseignements concernant la zone, le groupe de tarification du véhicule, les clauses d'usage et de garantie souscrites, l'état matrimonial, le sexe et l'âge du souscripteur, l'année de première mise en circulation du véhicule, l'année d'obtention du permis de conduire. On disposait, d'autre part, du nombre de sinistres de chaque sorte (matériels ou mixtes) et le coût au titre de la responsabilité civile de ceux-là.

1. Definition des Differents Criteres

Zones

En France, les communes sont classées en 5 zones numérotées de 1 à 5. La zone, classification géographique, est fonction:

du lieu de garage habituel du véhicule;
de la résidence principale du souscripteur.

Dans certaines Sociétés, la zone peut être fonction du lieu de travail habituel pour les souscripteurs garantis en usage „Affaires-Commerce" ou en usage „Promenade et trajet".

Groupes

Il existe 16 groupes numérotés de 0 à 15. Plus ce groupe est élevé, plus le risque en R.C. est important. Le groupe est déterminé par les assureurs de la manière suivante:

Dès la sortie du véhicule, un groupe est calculé d'après une formule basée sur les expériences passées. Cette formule tient compte, entre autre, de la puissance réelle du véhicule, de sa vitesse de pointe, de sa conception mécanique (freins à disques ou non, freinage assisté ou non, emplacement du moteur, propulsion arrière ou avant, essieu rigide ou roues indépendantes ...). Ce groupe défini a priori peut être modifié si les résultats statistiques en font apparaitre la nécessité.

Usages

Six usages principaux sont employés:

Affaires-Commerce
Salariés
Fonctionnaires et assimilés
Artisans
Agriculteurs
Autres

Age du conducteur

Trois classes sont régulièrement employées:

Célibataires masculins âgés de moins de 25 ans ou autres souscripteurs âgés de moins de 25 ans, permis de conduire de moins de deux ans.

https://doi.org/10.1017/S051503610001134X Published online by Cambridge University Press

Tous autres souscripteurs âgés de moins de 25 ans, permis de conduire d'au moins deux ans ou autres souscripteurs âgés d'au moins 25 ans, permis de moins de deux ans.

Tous souscripteurs âgés d'au moins 25 ans, permis de conduire d'au moins deux ans.

Age du véhicule

Trois classes d'âge de véhicule ont été formées lors de l'analyse:

véhicules de 1967 à 1970
véhicules de 1963 à 1966
véhicules antérieurs à 1963

2. Methode Utilisee

La méthode d'hypothèse linéaire généralisée est à la base de la méthode d'analyse multidimensionnelle qui a été choisie. C'est celle qui permet notamment de traiter les informations d'un sondage représentatif caractérisé par l'inexistence de données dans certains classes de tarification et, plus généralement, par une distribution non uniforme des effectifs dans les cases. Cette méthode et les tests qui lui sont rattachés ne sont valables que si les données analysées sont de variance constante et de distribution aussi proche que possible d'une distribution normale. Nous avons, de ce fait, été conduits à effectuer des changements de variable. Cette opération a alors suscité un nouveau problème, celui de l'estimation non biaisée des moyennes des anciennes variables (fréquences annuelles et coûts des sinistres), en fonction d'éléments relatifs aux nouvelles.

3. Analyse des Frequences de Sinistres

Soit K le nombre de sinistres d'un certain type associé à une unité statistique (véhicule assuré — période de T — unité de temps) prise au hasard dans une classe de tarif.

Loi de probabilité de K: soit m la fréquence annuelle moyenne de l'unité statistique échantillonnée; supposons que la loi de k liée par t et m soit une loi de POISSON:

$$\text{Prob}\ (K = k;\ T = t \text{ et } M = m) = \frac{e^{-mt} \cdot (mt)^k}{k!}$$

https://doi.org/10.1017/S051503610001134X Published online by Cambridge University Press

Considérant m comme une variable aléatoire (notée M) liée au tirage de l'unité statistique, il est logique d'introduire la loi a priori de M.

Sous l'hypothèse que celle-ci est une loi de PEARSON du type III, nous avons:

$$\text{Prob}\,(m < M < m + dm) = \frac{1}{\Gamma(b)}\, e^{-m/a} \left(\frac{m}{a}\right)^{b-1} d\left(\frac{m}{a}\right) \quad m > 0$$

Par application du théorème des probabilités composées et après plusieurs changements de variables, nous obtenons la relation:

$$\text{Prob}\,(K = k;\, T = t) = \frac{\Gamma(b+k)}{\Gamma(b)} \cdot \frac{1}{k!} \left(\frac{1}{1+at}\right)^{b} \cdot \left(\frac{at}{1+at}\right)^{k}$$

ou mieux encore:

$$\boxed{\text{Prob}\,(K = k;\, T = t) = C^{k}_{k+b-1} \cdot \left(\frac{1}{1+at}\right)^{b} \cdot \left(\frac{at}{1+at}\right)^{k}}$$

Ceci n'est autre que l'expression de la loi binominale négative $(b + k)$ souvent utilisée dans l'étude des dénombrements. D'une façon générale, l'espérance mathématique et la variance d'une telle loi valent respectivement:

$$E(X) = \frac{c}{p} \text{ et } V(X) = \frac{qc}{p^2}$$

Dans notre cas, en posant $p = \dfrac{1}{1+at}$; $q = \dfrac{at}{1+at}$; $b = c$ et $k = n - b$ nous obtenons:

d'une part: $E(b + k) = b + E(k) = b(1 + at)$

par suite $\boxed{E(k) = abt}$

d'autre part: $V(b + k) = \boxed{V(k) = at(1 + at)b}$

Ceci montre que:

la fréquence annuelle expérimentale (K/t) est une bonne estimation de la fréquence annuelle moyenne théorique ab. la variance de (K/t) dépend de la moyenne ab.

https://doi.org/10.1017/S051503610001134X Published online by Cambridge University Press

De nombreux chercheurs se sont préoccupés de trouver une fonction $U(K/t)$, telle que sa variance $V(U)$ soit sensiblement constante.

La transformation $U = \log\left(\frac{K}{t} + \frac{b}{2}\right)$ convient quand la quantité b est constante dans le domaine exploré, sans que d'ailleurs ceci soit une hypothèse très forte. Les travaux de M. P. DELAPORTE ont montré que pour les risques étudiés, nous avons $1 < b < 2$. Il a donc été retenu le changement de variable:

$$U = \log\left(\frac{K}{t} + 1\right)$$

4. Modele d'Hypothese Lineaire Generalisee

Convenons de désigner par réponse, soit l'une quelconque des composantes du risque, soit une fonction de ces composantes. Les critères sont, dans le cas le plus général, qualitatifs ou quantitatifs. Le modèle repose sur le corps d'hypothèses suivant:

1) La réponse est une variable aléatoire qui, tous critères fixés, c'est-à-dire pour une combinaison des modalités des critères qualitatifs et pour des valeurs données des critères quantitatifs, suit une loi de GAUSS de variance constante égale à σ^2.

2) Les réponses attachées aux unités statistiques de l'échantillon sont indépendantes en probabilité.

3) L'influence des critères s'exprime sous une forme linéaire. Par exemple, dans le cas de deux critères qualitatifs A et B, prenant respectivement les modalités A_i et B_j, on écrit, pour une réponse appartenant à la case i, j:

$$y_u = \mu + \alpha_i^A + \beta_j^B + \varepsilon_u$$

avec comme contrainte:

$$\sum_i \alpha_i^A = \sum_j \beta_j^B = 0 \qquad \text{où:}$$

y_u est la valeur de la réponse associée à la $u^{\text{ième}}$ observation ($u = 1$ à n).

ε_u est la réalisation d'une variable aléatoire suivant une loi de GAUSS, centrée et de variance σ^2.

μ est l'ordonnée à l'origine.

α_i^A et β_j^B sont les coefficients différentiels liés respectivement aux modalités A_i et B_j.

5. ESTIMATION DES COEFFICIENTS DU MODELE

D'une facon générale, le modèle mathématique que l'on pose est le suivant:

$$y_{ijkl\ldots} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l \ldots\ldots + \varepsilon_{ijkl\ldots}$$

Il est commode alors d'introduire la notation matricielle; l'expression précédente se trouve alors être un élément de la forme plus générale suivante:

$$\vec{y} = X \cdot \vec{\beta} + \vec{\varepsilon}$$

avec:

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} \quad \vec{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_i \\ \vdots \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \delta_p \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1^1 & x_1^2 & x_1^3 & \ldots & x_1^p \\ 1 & \ldots & & & & \\ 1 & \ldots & & & & \\ 1 & \ldots & & & & \\ 1 & \ldots & & & & \\ 1 & \ldots & & & & \\ 1 & x_n^1 & x_n^2 & x_n^3 & \ldots & x_n^p \end{bmatrix} \quad \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\vec{y}$ est le vecteur colonne dont les composantes sont les réponses observées.

$\vec{\beta}$ est le vecteur colonne dont les composantes sont les estimations des coefficients.

X est le tableau des conditions d'observations à $(p + 1)$ colonnes et n lignes.

$\vec{\varepsilon}$ est le vecteur colonne résiduel.

Le schéma des coefficients théoriques est donné par la formule:

$$\vec{y} = X \cdot \vec{\beta} + \vec{\varepsilon}$$

L'estimation des coefficients de $\vec{\beta}$ conduit au schéma:

$$\vec{y'} = X \cdot \vec{\beta'}$$

Le vecteur $\vec{\beta}'$ est déterminé par la méthode du maximum de vraisemblance, c'est-à-dire qu'il nous faut minimiser la norme euclidienne de $\overrightarrow{y' - y}$ soit:

$$\Omega^2 = || y' - y ||^2 = \sum_{u=1}^{n} (y'_u - y_u)^2$$

$$\frac{\partial \Omega^2}{\partial \vec{\beta}'} = 2 \sum_{u=1}^{n} (y'_u - y_u) \cdot \frac{dy'_u}{d\vec{\beta}'} = 0$$

$$= \sum_{u=1}^{n} (y'_u - y_u) \cdot x_u^\alpha = 0$$

ou encore:

$$X^t \cdot \vec{y}' = x^t \cdot \vec{y}$$

Comme $\vec{y}' = X \cdot \vec{\beta}'$ il vient:

$$\vec{\beta}' = (X^t \cdot X)^{-1} \cdot X^t \cdot \vec{y}$$

expression dans laquelle X^t est la transposée de X.

D'autre part, on démontre que $\vec{\beta}'$ est une estimation non biaisée de $\vec{\beta}$, obéissant à une gaussienne à k dimensions ayant comme matrice des variances — covariances l'expression $(X^t X)^{-1} \sigma^2$, σ^2 étant la variance de la population estimée par la relation suivante:

$$\sigma^2 = \frac{1}{n-k} \sum_u (y'_u - y_u)^2$$

6. Test des Influences des Coefficients

Le principe du test de l'influence d'un critère est le suivant:

dans un premier stade, on effectue un ajustement du modèle linéaire avec tous les coefficients, ce qui fournit un vecteur $\vec{\beta}'_{(0)}$ et une variance résiduelle $VR_{(0)}$ à d_0 degrés de liberté ($d_0 = n$ — rang de la matrice $X^t X$).

ensuite, on annule a priori les coefficients correspondants au critère testé et on effectue un ajustement analogue. Nous obtenons alors une nouvelle variabilité résiduelle notée $VR_{(h)}$, l'indice h servant à repérer les tests successifs. Nous avons:

$$\vec{\beta}'_{(h)} = (X^t_{(h)} \cdot X_{(h)})^{-1} \cdot X^t_{(h)} \cdot \vec{y}$$

$$VR_{(h)} = \vec{y}^t \cdot \vec{y} - \vec{y}^t \cdot X_{(h)} \cdot \overrightarrow{\beta'_{(h)}}$$

expression dans laguelle $X_{(h)}$ se déduit de X par suppression des colonnes correspondant aux coefficients annulés. Le nombre de d.d.l. de $VR_{(h)}$ est égal $_p$:

$$d_{(h)} = n - \text{rang de la matrice } X^t_{(h)} \cdot X'_{(h)}$$

On montre que si le critère testé n'a pas d'influence réelle, la quantité:

$$F = \frac{\dfrac{VR_{(h)} - VR_{(0)}}{d_{(h)} - d_{(0)}}}{\dfrac{VR_{(0)}}{d_{(0)}}}$$

est une variable de FISCHER-SNEDECOR à:

$$\left.\begin{cases} n_1 = d_{(h)} - d_{(0)} \\ n_2 = d_{(0)} \end{cases}\right\} \text{degrés de liberté.}$$

Le test statistique en résulte; si F est supérieur au seuil de signification, on est en droit de conclure d l'influence significative du critère testé en notant bien que, d'un point de vue théorique, les influences testées sont des influences conditionnelles: on isole l'influence d'un critère quand celui-ci est introduit dans le modèle après que tous les autres l'aient été.

7. Estimation de la Frequence Annuelle Moyenne: f

L'ajustement des coefficients du modèle:

$$Y_u = \mu + \alpha_i + \ldots\ldots\ldots + r_u$$

conduit aux estimations des valeurs moyennes de Y dans chaque classe de tarif et de la variance résiduelle de Y.

Il ne serait pas satisfaisant de prendre, pour estimation de ab (voir paragraphe "Analyse des fréquences des sinistres"), la valeur de f telle que log $(f + 1)$ soit égale à l'estimation de Y. En effet, nous pouvons écrire:

$$Y = \log_{10} \frac{K}{t} + 1$$

Posant $\frac{K}{t} = x$ et $E(x) = f$, il vient:

$$Y = M \operatorname{Log}(x + 1) = M \operatorname{Log}(x - f + f + 1)$$
$$= M \left[\operatorname{Log}(f + 1) + \frac{x - f}{f + 1} - \frac{1}{2} \cdot \frac{(x - f)^2}{(f + 1)^2} + \dots\dots \right]$$

d'où:

$$E(Y) \# M \left[\operatorname{Log}(f + 1) - \frac{1}{2} E \left(\frac{x - f}{f + 1}\right)^2\right]$$

pratiquement

$$\boxed{E(Y) \# M \operatorname{Log}(f + 1)}$$

Quant à la variance de Y, elle se met sous la forme:

$$V(Y) = E(Y - E(Y))^2$$

soit après calcul:

$$V(Y) = \frac{M^2}{(f + 1)^2} \cdot E(x - f)^2$$

d'où l'on tire:

$$E(x - f)^2 = \frac{E(Y - E(Y))^2 \cdot (f + 1)^2}{M^2}$$

En reportant cette quantité dans l'expression complète de $E(Y)$, nous obtenons:

$$E(Y) = M \left(\operatorname{Log}(f + 1) - \frac{E(Y - E(Y)^2}{2M^2}\right)$$

Soit:

$$M \operatorname{Log}(f + 1) = E(Y) + \frac{E(Y - E(Y))^2}{2M}$$

Finalement, nous obtenons l'expression simple suivante:

$$\log_{10}(f + 1) = E(Y) + 1{,}151\ V(Y)$$

Cette formule permet de calculer f à partir de $E(Y)$ moyenne de Y et $V(Y)$ variance de Y.

La régression de $V(Y)$ en fonction de $E(Y)$ est sensiblement linéaire.

Conduite de l'Etude

Pour calculer la prime pure, nous avons appliqué l'analyse multidimensionnelle à chacune de ses quatre composantes.

$$PP = \overline{f_m} \cdot \overline{C_m} + \overline{f_c} \cdot \overline{C_c}$$

Aux fréquences des sinistres matériels et corporels, nous avons fait subir le changement de variable:

$$y = \log_{10} (f + 1)$$

Des dépouillements nous ont montré que les coûts des sinistres d'un certain type, relatifs p une classe de tarif, se distribuent suivant une loi logarithmo-normale de variance constante. Nous avons donc soumis à l'analyse les logarithmes des coûts S:

$$W = \log_{10} (S)$$

Après traitements pour repasser des logarithmes aux valeurs réelles, nous avons employé les formules suivantes:

pour l'étude des fréquences des sinistres matériels, la régression de $V(Y)$ en fonction de $E(Y)$ est:
$V(Y) = 0{,}319\, E(Y) - 0{,}00016$ avec un coefficient de corrélation de $v = 0{,}82$.
pour l'étude des fréquences des sinistres corporels, la régression de $V(Y)$ en fonction de $E(Y)$ est:
$V(Y) = 0{,}496\, E(Y) - 0{,}001$ avec un coefficient de corrélation de $v = 0{,}87$
Par suite, les formules à employer pour repasser en réel seront:
$\log_{10} (f_m + 1) = 1{,}367169\, E(Y)$ pour les matériels
$\log_{10} (f_c + 1) = 1{,}57089\, E(Y)$ pour les corporels

Quant aux coûts moyens matériels et corporels, les formules du type $\log_{10} (C) = E(W) + 1{,}151\, V(W)$ sont des formules exactes permettant d'estimer $C = E(S)$, c'est-à-dire les coûts moyens.

La variance calculée pour les coûts moyens matériels est:

$$V(W\ 1) = 0{,}2135;$$

https://doi.org/10.1017/S051503610001134X Published online by Cambridge University Press

celle des coûts moyens corporels vaut:

$$V(W\,2) = 0{,}7704.$$

Lors de cette étude, trois analyses ont été effectuées.
dans la première, nous avons exploité les critères suivants: groupes, zones, usages, âge des conducteurs;
dans la seconde, en plus des 4 critères de l'exploitation précédente, nous avons introduit l'âge du véhicule;
dans la troisième, en plus de 5 critères de l'exploitation précédente, nous avons introduit des groupes d'interactions qui nous semblaient être significatifs.

Finalement, les primes pures ont été calculées à l'aide des coefficients de l'exploitation à 5 facteurs, car dans la troisième analyse l'introduction des interactions n'a guère diminué la somme des carrés d'écart, et l'analyse à 4 facteurs était tout de même moins complète.

Dans l'exploitation retenue, nous avons extrait et exploité en sus 3 sous-populations formées pour les usagers suivants:

salariés
fonctionnaires
agriculteurs

Le modèle mathématique retenu était de la forme:

$$y_{ijklm} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \Psi_m + \varepsilon_{ijklm}$$

modèle dans lequel:

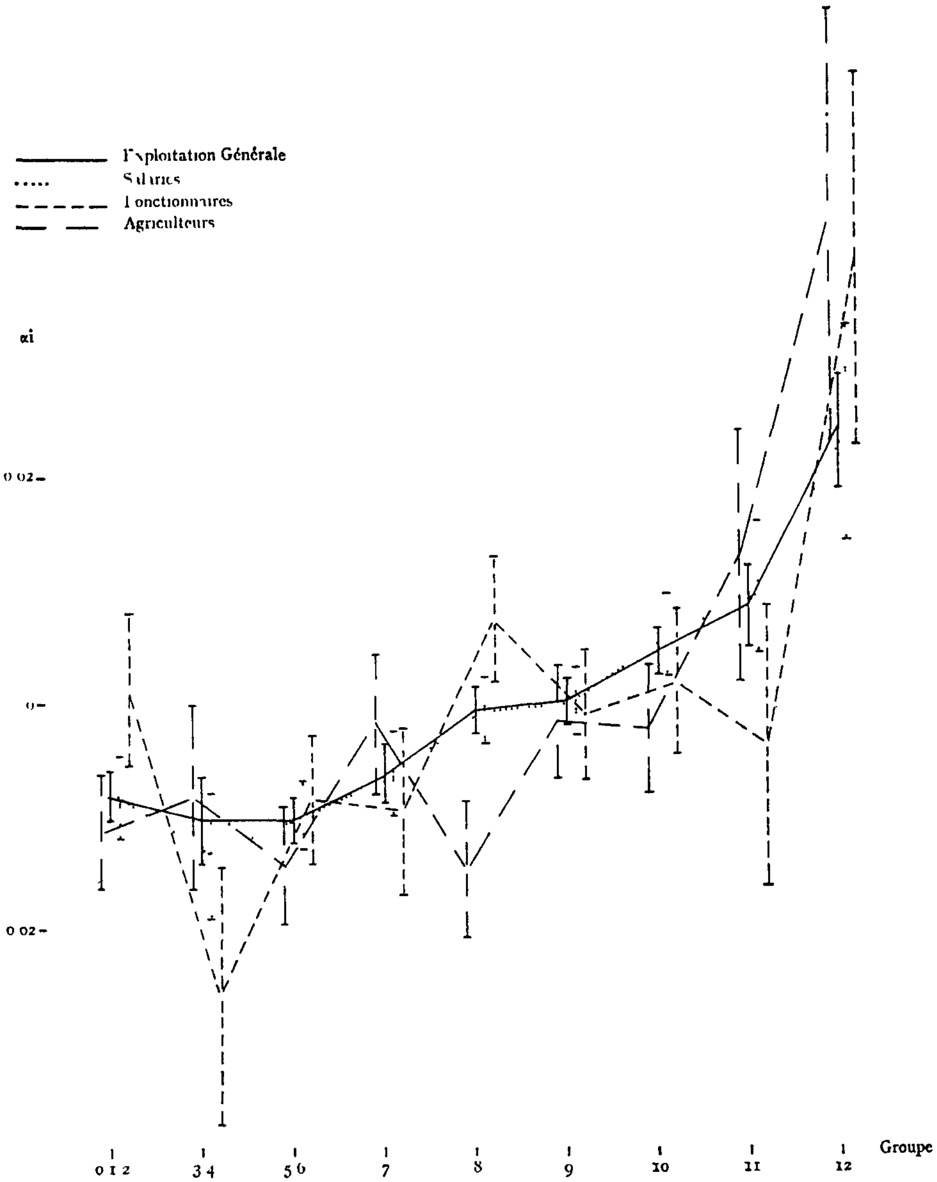
μ	: terme de centrage	
α_i	: coefficient relatif aux groupes	(9 niveaux)
β_j	: coefficient relatif aux zones	(4 niveaux)
ν_k	: coefficient relatif aux usages	(6 niveaux)
z_l	: coefficient relatif à l'âge des conducteurs	(3 niveaux)
Ψ_m	: coefficient relatif à l'âge des véhicules	(3 niveaux)
ε_{ijklm}	: résiduelle	

Les résultats pour les groupes (par exemple) étaient les suivants:

Facteurs		Effets des facteurs	Ecarts type
	μ	0,07079124	0,000847
		Exploitation Generale	
Groupes	i		
0 + 1 + 2	1	—0,00848394	0,001186
3 + 4	2	—0,01040909	0,001884
5 + 6	3	—0,01039765	0,001016
7	4	—0,00647201	0,001315
8	5	—0,00091265	0,001024
9	6	0,00013299	0,000997
10	7	0,00444214	0,001074
11	8	0,00827156	0,001837
12 et +	9	0,02382864	0,002515
		Sous-Exploitation „Salaries"	
	μ	0,06906843	0,001080
Groupes	i		
0 + 1 + 2	1	—0,00838255	0,001812
3 + 4	2	—0,01335252	0,002892
5 + 6	3	—0,00999472	0,001540
7	4	—0,00654063	0,001891
8	5	—0,00084569	0,001490
9	6	0,00012398	0,001531
10	7	0,00571994	0,001774
11	8	0,00969048	0,003088
12 et +	9	0,02358172	0,004871
		Sous-Exploitation „Fonctionnaires"	
	μ	0,05819486	0,002240
Groupes	i		
0 + 1 + 2	1	0,00133143	0,003381
3 + 4	2	—0,02579000	0,005675
5 + 6	3	—0,00857455	0,002801
7	4	—0,00966734	0,003710
8	5	0,00727953	0,002834
9	6	—0,00110694	0,002924
10	7	0,00149672	0,003234
11	8	—0,00398095	0,006221
12 et +	9	0,03901210	0,008244
		Sous-Exploitation „Agriculteurs"	
	μ	0,06039833	0,003562
Groupes	i		
0 + 1 + 2	1	—0,01135983	0,002490
3 + 4	2	—0,00837894	0,004073
5 + 6	3	—0,01431175	0,002573
7	4	—0,00195246	0,003101
8	5	—0,01489352	0,002915
9	6	—0,00184322	0,002537
10	7	—0,00229437	0,002760
11	8	0,01286054	0,005547
12 et +	9	0,04217354	0,009586

Graphique 1

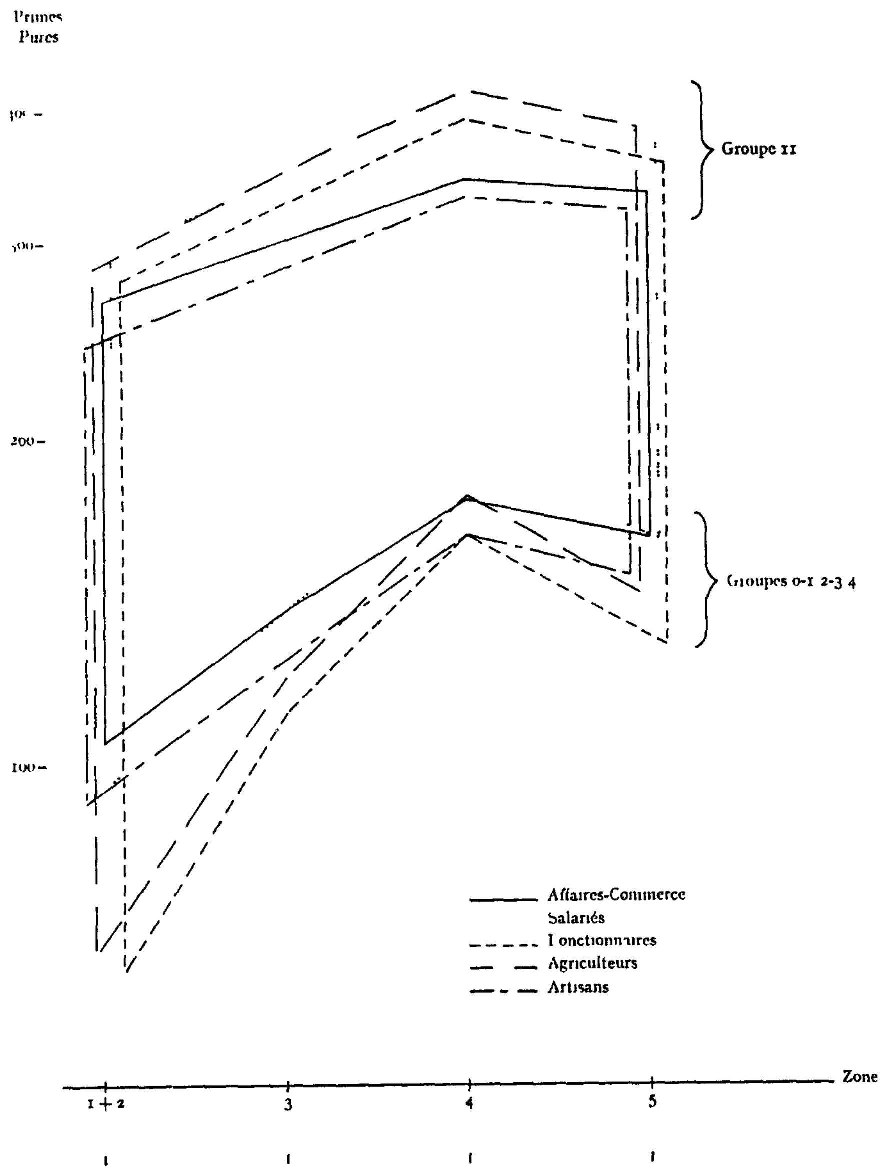
Influence du groupe sur la fréquence des sinistres matériels



https://doi.org/10.1017/S051503610001134X Published online by Cambridge University Press

Graphique 2

Primes pures par zone, groupe et usage
Exploitation à 5 facteurs (véhicles de moins de 3 ans)



Conclusion

Dans l'ensemble, les coefficients des sous-populations suivent relativement bien le sens des coefficients de l'exploitation générale, hormis ceux du groupe 8 où l'on note une divergence entre les coefficients relatifs aux fonctionnaires et aux agriculteurs. Cela est dû à l'hétérogénéité importante de ce groupe (proportion importante de véhicules âgés dans l'usage ,,Agriculteurs'').

Le graphique 2 représente l'étendue entre le maximum (groupe 11) et le minimum (groupes 0-1-2-3-4) des primes pures pour des véhicules de moins de trois ans, réparties suivant quatre zones. Il est à noter que les primes pures en zone 5 sont moins élevées que celles de la zone 4; cette anomalie s'explique par le fait que, si durant longtemps la zone 5 fut la plus dangereuse, elle est maintenant dépassée par la zone 4 qui continue à se développer, alors qu'en zone 5 la phénomène de saturation commence à se manifester. Ce point est très sensible en corporel: la zone 4 présente une fréquence très importante du fait du manque de transport en commun et des possibilités meilleures de circulation.

Les Assureurs envisagent, devant ces statistiques, une fusion des zones 4 et 5 (PARIS-LYON) et un relèvement à un niveau supérieur des villes qui s'avèreront les plus mauvaises.

Resume

Exploitation du sondage automobile 1971 en France par une methode d'analyse multidimentionnelle

Sous la responsabilité de l'Association Générale des Sociétés d'Assurances contre les Accidents, un sondage au 1/50ème a été effectué en 1971 dans les portefeuilles de 31 Sociétés d'assurances pratiquant en France l'assurance de responsabilité civile automobile.

Le modèle d'hypothèse linéaire généralisée est à la base de la méthode d'analyse multidimentionnelle qui a été choisie, afin de déterminer des primes pures pour chaque classe de tarif. C'est en effet celle qui permet notamment de traiter les informations d'un sondage représentatif caractérisé par l'inexistence de données dans certaines classes de tarification et plus généralement par une distribution non uniforme des effectifs dans les cases.

La nécessité de soumettre à l'analyse des quantités dont la variance est constante dans le domaine exploré nous a conduit à effectuer des changements de variables. Cette opération a alors suscité un nouveau problème, celui de l'estimation non biaisée des moyennes des anciennes variables (fréquences annuelles et coûts des sinistres) en fonction d'éléments relatifs aux nouvelles.

Summary

Use of a multidimensional analysis method to investigate the results of the French 1971 motor vehicle survey

Sponsored by the Association Générale des Sociétés d'Assurances contre les Accidents, an inquiry (approximation 1/50) was carried out in 1971 bearing on the portfolios of 31 Insurance Companies dealing with the Motor Vehicule Third Party Insurance in France.

The method of multidimensional analysis selected is based on the model of linear hypothesis taken as a rule, to try and estimate the pure premium in each class of rate. In fact, this method allows, in particular, to deal with the information gathered by means of representative sample to deal which features the non-availability of information within certain rating class and, more generally speaking, a highly diversified distribution of factors within each class.

Quantities entering into the surveys must have constant variance within the investigated field. So we applied a variable transformation. We were then confronted with another probleme i.e. a non-biased estimation of the previous variable averages (annual frequency rate of losses and costs of losses) in terms of data related with the new ones.