Influence de la partition homme/femme et de l'expérience kilométrique dans l'assurance automobile

Alexandre Mornet, Patrick Leveillard, Stéphane Loisel

Influence de la partition homme/femme et de l'expérience kilométrique dans l'assurance automobile.

A.Mornet* P.Leveillard† S.Loisel‡

2014

Résumé

Dans le secteur de l'assurance automobile, la décision de la Cour de Justice de l'Union européenne selon laquelle il n'est plus possible de pratiquer des tarifs selon le sexe de l'assuré, ainsi que la diffusion des dispositifs d'assurance au kilomètre, entraînent nécessairement une évolution de la gestion du risque. Dans cet article, on se base sur le portefeuilles d'Allianz en France pour répondre à ces deux problématiques. On propose de caractériser la partition homme/femme en explorant différentes méthodes statistiques comme la procédure logistique, l'analyse des correspondances multiples (ACM) ou les arbres de classification (CART). On montre qu'il est possible de compenser l'absence de la variable sexe par d'autres informations spécifiques à l'assuré ou à son véhicule et en particulier l'utilisation de relevés kilométriques [11]. On revient ensuite à l'utilisation des modèles linéaires généralisés (GLM) pour valider ces résultats.
Dans une deuxième partie, on s'interesse à l'expérience acquise par les conducteurs novices. Cette catégorie d'assurés compte parmi les plus sensibles aux critères homme/femme dans sa tarification. Ici, les modèles additifs généralisés (GAM) permettent d'exploiter les variables numériques comme le kilométrage annuel parcouru. Nous proposons finalement d'étudier le comportement sur la route de l'assuré durant ces trois années de noviciat pour créer de nouvelles catégories de risques.

1 Introduction

La décision de la cour européenne de justice stipule que l'article 5(2) de la Directive 2004/113 n'est pas compatible avec l'article 6(2) du traité de l'Union Européenne. En clair, les assureurs ne sont plus autorisés depuis le 21 décembre 2012 à ajuster leurs tarifs en fonction du sexe. Une telle distinction n'étant pas compatible avec les principes d'égalité entre les hommes et les femmes. La variable sexe a pourtant été largement utilisée par les assureurs avant cette directive car elle leur permettait de déterminer facilement deux profils de risque différents. En assurance automobile par exemple, il s'avère que les femmes ont statistiquement moins d'accidents que les hommes. Dans ce contexte, nous avons voulu décrire ce qui caractérise la partition homme/femme à partir des autres variables explicatives que l'assureur est en droit d'utiliser. Nous avons aussi montré qu'un modèle de prédiction des sinistres peut fonctionner sans la variable sexe et fournir d'aussi bons résultats. Parmi les informations complémentaires à la disposition de l'assureur, la connaissance du kilométrage annuellement parcouru constitue une nouvelle variable très significative [11].

Ces dernières années, l'assurance au kilomètre ou *pay as you drive* ([1], [3]) s'est largement diffusée dans le système de l'assurance automobile. Elle ne constitue pas une prime supplémentaire mais plutôt un pas de plus vers une tarification plus proche du comportement de l'assuré. Grace à des boîtiers embarqués, on peut désormais mieux connaître l'utilisation que le conducteur fait de son véhicule. En plus du kilométrage, il est possible de savoir la période d'utilisation du véhicule (jour/nuit), le type de routes empruntées (autoroute/ville/campagne) ou même le type de conduite (vitesse/accélération). De façon pratique, il faut cependant noter que certaines limites doivent être observées dans le respect des libertés individuelles.

Bien que la relation entre la fréquence des sinistres et le kilométrage annuel puisse sembler évidente, peu d'études ont jusqu'à présent insisté sur l'importance de travailler avec des données fiables sur le

* Université de Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, Lyon F-69007, France and ALLIANZ, Coeur Défense, 82 Esplanade du Général de Gaulle, Courbevoie F-92400, France.

† ALLIANZ, Coeur Défense, 82 Esplanade du Général de Gaulle, Courbevoie F-92400, France.

‡ Université de Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, Lyon F-69007, France

kilométrage parcouru par les assurés ([2], [12]). Pourtant, la précision de cette information s'avère essentielle à la construction de catégories de risque lors de la tarification. Le facteur principal dans le choix d'une assurance automobile par les consommateurs demeure le prix. Par conséquent, une tarification plus flexible devrait constituer une offre plus intéressante et aurait en plus l'avantage écologique d'inciter à un usage moins systématique des véhicules individuels lorsque cela est possible [5]. Une des caractéristiques du portefeuilles Allianz en France réside dans son partenariat avec la société française des compteurs automobiles (SOFCA) qui nous a permis de travailler sur des relevés exacts de kilométrage sur trois années consécutives. Dans nos recherches, nous avons considéré la distribution des sinistres directement liés à la circulation routière comme les dommages matériels responsables, mais aussi ceux qui peuvent en sembler déconnectés comme le vol ou l'incendie. Historiquement, comme le montre Roger Roots dans son rapport sur les dangers de l'automobile [16], depuis les premiers moyens de transport à cheval, la tendance était à une diminution linéaire de la fréquence des sinistres pour chaque kilomètre parcouru. Dans une perspective plus actuelle, la sinistralité se reflète dans l'expérience acquise au volant et dans le nombre d'années depuis l'obtention du permis. Les habitudes de conduite que l'on retrouve dans le kilométrage annuel peuvent nous aider à mieux comprendre le risque d'accident.

Dans une première partie, on présente les données d'assurance à notre disposition. Nous observons la sinistralité des différentes catégories de risques selon la partition homme/femme. L'approche graphique permet de mettre en évidence des différences statistiques. Partant de ce constat, nous essayons dans une deuxième partie de caractériser le sexe du conducteur en fonction d'autres variables spécifiques à l'assuré, à son environnement ou à son véhicule. Nous utilisons les modèles linéaires généralisés (GLM) pour comparer l'efficacité des modèles avec et sans la variable sexe [13]. La troisième partie est consacrée aux novices et à l'acquisition d'expérience selon le kilométrage. Nous utilisons ici les modèles additifs généralisés (GAM) pour explorer à travers différentes tranches kilométriques le comportement et l'évolution des jeunes conducteurs durant leurs trois années de noviciat.

2 Description des données d'assurance

Pour cette étude, nous avons eu accès aux portefeuilles automobiles d'Allianz en France sur la période 2008-2010. Plusieurs catégories d'informations relatives à l'assuré sont disponibles. Elles relèvent de différents sources. La table **Sinistres** nous renseigne sur la nature de l'événement, sa date, son coût détaillé, elle contient autant de lignes par numéro de police que d'événements sur le contrat. La table **Polices x risques** contient les informations sur l'assuré et son contrat d'assurance auto. L'association **SRA** (Sécurité et Réparations Automobiles) fournit toutes les informations relatives aux véhicules en commercialisation. Le partenariat entre la compagnie d'assurances et l'entreprise **SOFCA** (Société Française des Compteurs Automobiles) nous fournit les relevés kilométriques de l'ensemble des conducteurs ayant équipé leurs véhicules du compteur additionnel. Nous utilisons aussi des données de l'**INSEE** (recensement démographique) pour connaître l'environnement démographique de l'assuré.

2.1 Répartition de la fréquence des sinistres selon le sexe

L'absence de la distinction homme / femme dans la tarification comme le stipule la directive européenne va surtout poser problème aux assureurs pour deux catégories de population, les conducteurs novices et les conducteurs seniors.

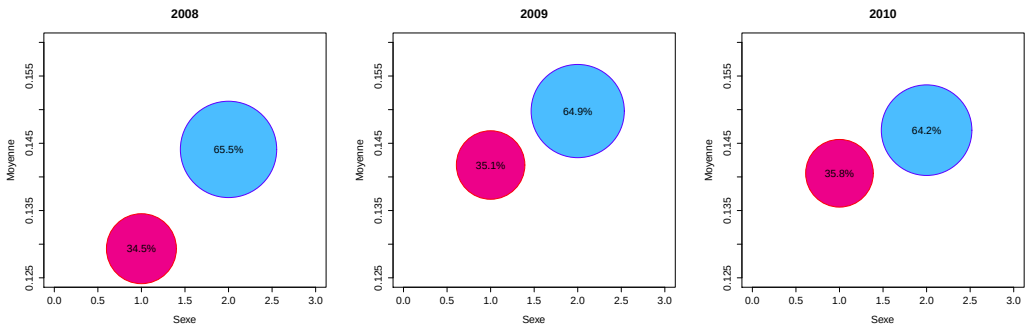

FIGURE 1 – Nombre moyen de sinistres pour les hommes et les femmes novices entre 2008 et 2010

En effet, on observe statistiquement une sinistralité plus forte chez les novices hommes et chez les

seniors femmes [17]. Dans cet article, on s'intéresse aux conducteurs **novices** (client avec moins de 3 ans d'assurance automobile). Chez les conducteurs novices on remarque une plus forte sinistralité chez les hommes toutes catégories de risques confondues que chez les femmes. Sur la Figure 1, on peut comparer l'évolution du nombre moyen de sinistres selon le sexe sur 3 années consécutives. La taille des cercles bleus et roses correspond à la proportion respective d'hommes et de femmes assurés chez Allianz. Le portefeuille a légèrement diminué en taille durant la période 2008-2010, en revanche les proportions sont relativement les mêmes avec pour les hommes une part comprise entre 65.5% et 64.2% et pour les femmes entre 34.5% et 35.8%. L'écart de sinistralité évolue aussi à la baisse avec un écart relatif qui passe de 10.2% à 4.4%. Pour plus de précisions on compare la moyenne de sinistres pour les différentes catégories de risques prises séparément.

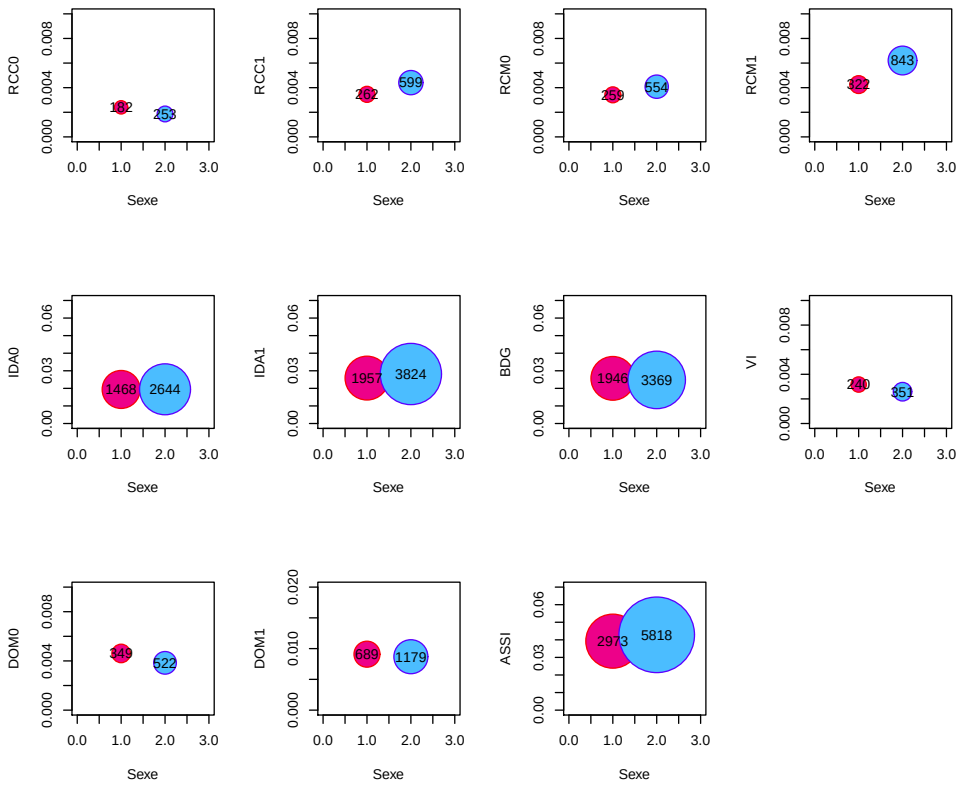

FIGURE 2 – Fréquence de sinistres hommes et femmes novices en 2010

Sur la Figure 2, on observe la moyenne des sinistres pour les hommes et les femmes selon la nature du sinistre (cf. Table 27 en annexe). Les catégories les plus représentées en terme de nombre de sinistres sont les indemnisations directes (IDA), les bris de glaces (BDG) et l'assistance (ASSI). Les hommes ont une sinistralité plus forte dans une majorité de catégories de risques, cependant ce n'est pas le cas pour les bris de glace, les vols et incendies auxquels on peut rajouter en 2010 les sinistres corporels non responsables et les dommages. On peut noter sur la période 2008-2010 que la sinistralité plus élevée chez les femmes (catégories en rouge sur Table 1) concerne majoritairement des sinistres que l'on peut considérer comme non responsables.

2010	RCC0	RCC1	RCM0	RCM1	IDA0	IDA1	BDG	VI	DOM0	DOM1	ASSI
F	0.24%	0.35%	0.34%	0.43%	1.94%	2.58%	2.57%	0.32%	0.46%	0.91%	3.92%
H	0.19%	0.44%	0.41%	0.62%	1.95%	2.82%	2.48%	0.26%	0.38%	0.87%	4.28%
2009	RCC0	RCC1	RCM0	RCM1	IDA0	IDA1	BDG	VI	DOM0	DOM1	ASSI
F	0.29%	0.36%	0.38%	0.37%	2.09%	2.55%	2.55%	0.33%	0.54%	0.82%	3.91%
H	0.17%	0.44%	0.42%	0.64%	1.93%	2.91%	2.52%	0.32%	0.47%	0.84%	4.31%
2008	RCC0	RCC1	RCM0	RCM1	IDA0	IDA1	BDG	VI	DOM0	DOM1	ASSI
F	0.25%	0.37%	0.36%	0.37%	1.95%	2.50%	2.23%	0.31%	0.35%	0.67%	3.58%
H	0.20%	0.48%	0.43%	0.63%	2.02%	2.90%	2.19%	0.31%	0.38%	0.76%	4.13%

TABLE 1 – Fréquence de sinistres hommes et femmes novices entre 2008 et 2010

2.2 Répartition du coût des sinistres selon le sexe

Pour la modélisation des coûts, on travaille généralement sur les sinistres fermés. A priori, les différences hommes/femmes apparaissent surtout au niveau des fréquences des sinistres. Sur la Figure 3 on a représenté la charge moyenne pour l'ensemble des garanties ainsi que la proportion en nombre de sinistres selon le sexe. Pour la charge moyenne, il y a bien une différence, cependant elle est relativement faible : 3% de plus pour les hommes. Pour le nombre de sinistres, on retrouve un ordre de grandeur correspondant aux constatations de la section précédente avec 68.3% des sinistres du côté des hommes et 31.7% du côté des femmes.

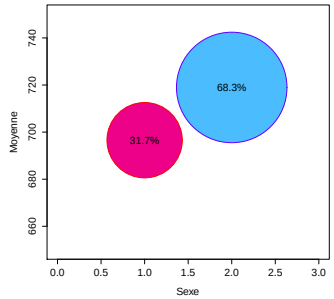

FIGURE 3 – Charge moyenne et proportion en nombre de sinistres chez les hommes et les femmes novices de 2008 à 2010

La Table 2 montre les coûts moyens des sinistres chez les hommes et les femmes novices selon les types de garanties. Lorsque l'on observe cette différence garantie par garantie, on voit qu'elle reste assez faiblement supérieure chez les hommes avec la plus forte différence observée pour les bris de glace avec un pic à 7%. On peut noter que les femmes ont un coût moyen supérieur dans le cas des sinistres matériels non responsables mais la différence reste marginale (1%).

	RCC0	RCC1	RCM0	RCM1	IDA0	IDA1	BDG	VI	DOM0	DOM1	ASSI
F	791.8	1089.2	693.3	976.7	697.8	1041.7	388.9	776.6	887.8	1043.5	438.6
H	808.7	1104.2	683.9	995.4	727.5	1051.1	418.1	809.6	907.9	1052.7	448.4
Diff	2%	1%	-1%	2%	4%	1%	7%	4%	2%	1%	2%

TABLE 2 – Fréquence de sinistres hommes et femmes novices entre 2008 et 2010

3 Caractérisation de la partition Homme/Femme

Le choix de tarifer par genre offre de nombreux avantages : simplicité de recueil de l'information, existence de corrélation entre risque et variable choisie. Mais il ne faut jamais oublier que ces critères ne sont que des indicateurs de l'exposition de l'assureur aux risques [4]. On se propose ici de caractériser la partition homme-femme à partir des variables catégorielles disponibles à la fois dans le portefeuilles Allianz, dans les tables SRA et les tables INSEE.

3.1 La procédure logistique

Plusieurs approches sont envisagées, la première est une modélisation de la variable sexe via une régression logistique. On utilise la procédure `logistic` du logiciel SAS et l'ensemble des variables sélectionnées ensuite pour les GLM. Sur un échantillon de 100 000 assurés, la méthode stepwise utilisée considère que 32 des 35 variables explicatives sont significatives. Les résultats de cette procédure selon les tests de Hosmer et Lemeshow sont dans la Table 3.

Lors de ces tests, les données sont rangées par ordre croissant des probabilités calculées à l'aide du modèle, puis partagées en 10 groupes. Le test est largement positif mais il est aussi malheureusement peu puissant : Le test du khi-deux est utilisé pour comparer les effectifs observés aux effectifs théoriques. Les prédictions globales sont assez précises. On observe dans la Table 4 les taux d'erreurs individuelles de la matrice de confusion. La procédure s'applique à 125 000 assurés, les résultats sont concluants pour 71 138 hommes et 15 485 femmes ce qui représente un taux d'erreur

		SEXE = H		SEXE = F	
Groupe	Total	Observé	Attendu	Observé	Attendu
1	5640	1547	1422.91	4093	4217.09
2	5641	2094	2147.43	3547	3493.57
3	5640	2617	2576.62	3023	3063.38
4	5641	2859	2930.25	2782	2710.75
5	5640	3154	3253.82	2486	2386.18
6	5640	3503	3572.71	2137	2067.29
7	5640	3890	3905.29	1750	1734.71
8	5640	4307	4246.56	1333	1393.44
9	5640	4675	4630.5	965	1009.5
10	5641	5164	5123.92	477	517.08

TABLE 3 – Partition pour les tests de Hosmer et de Lemeshow

inférieur à 31%. On comparera par la suite cette précision avec celle obtenue par les arbres de décision.

Error rate		0,307	
Confusion matrix			
	F	H	Total
F	15485	28423	43908
	12.39	22.74	35.13
H	9954	71138	81092
	7.96	56.91	64.87
Total	25439	99561	125000
	20.35	79.65	100

TABLE 4 – Proc FREQ

3.2 Exploration de données

3.2.1 ACM

Dans un deuxième temps on s'intéresse aux techniques de datamining. Une solution adaptée à la prise en charge de variables catégorielles consiste à faire une analyse des correspondances multiples (ACM). L'ACM permet techniquement de projeter et donc représenter un nuage de points initialement situé dans un espace de très grande dimension (le nombre de modalités moins le nombre de variables) dans l'espace de dimension plus réduite dans lequel la distance des points deux à deux est maximale, donc l'espace qui conserve le mieux la richesse de l'information de départ. On peut alors élaborer des variables quantitatives que sont les coordonnées des individus sur les principaux axes de l'analyse et préparer une classification des individus, à partir leurs coordonnées sur une partie des axes de l'ACM. [10].
Pour notre projet de partition des individus selon la distinction homme-femme, on va s'intéresser particulièrement au poids et à la contribution de la variable sexe dans le nuage de points et la constitution des axes. On pourra alors si par exemple la modalité "homme" a une contribution forte sur un des axes principaux associer à cette modalité celles des autres variables catégorielles qui influencent également l'axe auquel on s'intéresse. On utilise la procédure `CORRESP` du logiciel SAS à laquelle in combine la macro fournit par l'INSEE : `AIDEACM` qui facilite l'interprétation des résultats de l'analyse. Les résultats de l'ACM apparaissent dans la Table 5.

L'exécution du programme nécessite la construction d'un tableau disjonctif complet. On choisit de lancer la procédure avec d'une part des variables actives issues des critères classiques, des critères spécifiques novices et des critères Client et d'autre part des variables supplémentaires issues des critères INSEE. Les variables actives serviront à la construction des axes de l'ACM et les variables supplémentaires seront seulement projetés sur ces axes pour d'éventuelles associations. Les résultats sur les 3 axes principaux ne sont malheureusement pas concluant pour la variable sexe qui, malgré son poids dans le nuage relativement important (10.2%), ne contribue que très faiblement à la constitution de chaque axe (jamais plus de 2.4%). De même, si on observe le rang des modalités homme-femme on peut voir que leurs contributions ne se placent pas avant la 20ième position sur 43. Il semble donc difficile d'utiliser ces résultats pour constituer des classes dont le sexe serait l'élément

```
N° Val.Pr. Diff. Pct Cum !
-------------------------------------------------------------------------------
1 0.3780 0.000 11.18 11.18 !*********************************************************
2 0.2394 0.1386 7.08 18.26 !*******************************
3 0.1843 0.0551 5.45 23.72 !******************
4 0.1616 0.0226 4.78 28.50 !*************
5 0.1437 0.0180 4.25 32.75 !**********
-------------------------------------------------------------------------------
```

TABLE 5 – ACM

discriminant. On peut néanmoins retenir de cette analyse que les variables "classe de prix", "groupe SRA" et "alimentation" sont celles qui contribuent le plus à la constitution des axes.

3.2.2 Tableaux de contingence

Une autre solution envisagée est l'utilisation de la macro `DESQUAL` de l'INSEE : La macro édite les tableaux de contingence croisant une variable de classe (ici le sexe de l'assuré) avec chacune des variables qualitatives du portefeuilles automobile. Elle effectue des tests statistiques permettant de caractériser les classes de la partition par les modalités des variables explicatives. On obtient davantage de résultats avec cette approche. Pour chacune des modalités, la macro retient dans un premier temps une vingtaine de variables explicatives qu'elle classe par niveau de significativité.

Modalité	Variable	Effectif	Fréquence	Fréquence	proba	Val.Test
TOPVIT2	TOPVIT	8018	94.2	87.9	0.0000	23.6971
CLASSVEHA	gpsra	3069	36	27.9	0.0000	19.9448
ClPrix1	clprix	2138	25.4	18.5	0.0000	19.1433
ETU	CSP	2162	26.1	19.1	0.0000	19.1354
TOPSPORT2	TOPSPORT	8028	94.3	90.1	0.0000	16.9284
INJINDIRECTE	ALIM	5372	63.8	58.4	0.0000	12.3909
ESSENCE	ENERGIE	4748	55.8	50.5	0.0000	11.8445
SPR	CSP	478	5.8	3.8	0.0000	11.2274
ClPrix2	clprix	3376	40	35.3	0.0000	11.1154
CLASSVEHC	gpsra	3478	40.9	37	0.0000	8.956
PORTES3	CARROS	1844	22.3	19.4	0.0000	8.0244
BERLINE5P	CARROS	1876	22.7	20.3	0.0000	6.6907
RENAULT	Marque	2570	33.2	30.4	0.0000	6.651
FIAT	Marque	513	6.6	5.5	0.0000	5.3752
FCT	CSP	305	3.7	3	0.0000	4.6434
PctChgtlogt2	PctChgtlogt	2287	27.5	25.8	0.0000	4.3727
PctChgtcom2	PctChgtcom	2287	27.5	25.8	0.0000	4.3727
TrAge5	TrAge	1185	14.3	13.2	0.0002	3.6012
cspmoins2	cspmoins	1919	23.1	21.9	0.0004	3.35
Pctsed3	Pctsed	2614	31.5	30.1	0.0006	3.242
MONOSPACE	CARROS	213	2.6	2.2	0.0024	2.8238
NAISSANCE1	NAISSANCE	461	5.6	5	0.0040	2.653

TABLE 6 – Modalités sur-représentées chez les femmes

Les Tables 6 et 7 contiennent les résultats de contingences d'abord pour les femmes puis pour les hommes. Dans la deuxième colonne on observe le classement des variables. La vitesse maximale du véhicule (`TOPVIT`), la catégorie socioprofessionnelle (`CSP`), le groupe SRA du véhicule (`gpsra`) et la classe de prix (`clprix`) sont les mieux placés, donc les plus discriminants à la fois pour les hommes et les femmes. L'ordre d'apparition est ensuite sensiblement le même pour les deux sexes. Pour voir apparaître les distinctions, il faut se référer à la première colonne de l'analyse qui caractérise à l'aide des modalités des variables les classes sur-représentées et sous-représentées chez les hommes et les femmes. On peut ainsi considérer par exemple, les critères `TOPVIT2`, la classe de véhicules A (`CLASSVEHA` : inférieur à 7608 euros), `ClPrix1`, et le statut étudiant (`ETU`) comme plutôt féminin sur notre panel. Le statut de salariés (`SAL`), les classes de véhicules D et E (`CLASSVEHD`, `CLASSVEHE` entre 10543 et 13042 euros) et `TOPVIT1` caractérisent quant à eux davantage les hommes assuré chez

Allianz.

Modalité	Variable	Effectif	Fréquence	Fréquence	proba	Val.Test
SAL	CSP	13245	78.9	74.2	0.0000	23.989
CLASSVEHD	gpsra	6224	35.6	31	0.0000	23.942
TOPVIT1	TOPVIT	2661	15.2	12.1	0.0000	23.6971
CLASSVEHE	gpsra	812	4.6	3.5	0.0000	16.9379
TOPSPORT1	TOPSPORT	2091	12	9.9	0.0000	16.9284
ClPrix4	clprix	2814	16.4	14	0.0000	16.6754
ClPrix3	clprix	5568	32.5	29.9	0.0000	13.2938
INJDIRECTESUR	ALIM	5350	31	28.5	0.0000	12.9835
GASOIL	ENERGIE	9081	52	49.5	0.0000	11.8445
ClPrix5	clprix	503	2.9	2.3	0.0000	11.154
FAMILIAL	CARROS	1082	6.5	5.6	0.0000	9.328
BMW	Marque	432	2.7	2.2	0.0000	9.2894
AUDI	Marque	339	2.1	1.7	0.0000	8.4062
COUPECABRIOLET	CARROS	749	4.5	3.9	0.0000	7.7786
PORTES4	CARROS	494	3	2.5	0.0000	7.5886
VOLKSWAGEN	Marque	1898	12	11	0.0000	7.2973
PctChgtlogt4	PctChgtlogt	5764	33.7	32.5	0.0000	5.8914
PctChgtcom4	PctChgtcom	5764	33.7	32.5	0.0000	5.8914
evol1	evol	2909	17	16.3	0.0000	4.2351
MERCEDES	Marque	201	1.3	1.1	0.0001	3.8412
BERLINE	CARROS	7648	46	45.2	0.0001	3.6491
Pctsed2	Pctsed	6893	40.3	39.6	0.0005	3.2861
Trpop4	Trpop	2911	17	16.5	0.0009	3.1057
cspmoins3	cspmoins	7871	46	45.4	0.0019	2.8989
PEUGEOT	Marque	4078	25.8	25.3	0.002	2.8741
NAISSANCE5	NAISSANCE	5773	33.8	33.2	0.0032	2.726

TABLE 7 – Modalités sur-représentées chez les hommes

3.2.3 Arbres de classification

L'idée inhérente à l'utilisation d'arbres est d'expliquer les modalités homme ou femme à partir de combinaisons de variables et non plus seulement par l'agencement de variables séparément significatives. Nous proposons donc ici de construire des arbres de décision selon la méthode `CART` proposée par Breiman et al. [6]. Ce type d'approche est aussi utilisé en assurance-vie pour expliquer le processus de décision d'un assuré voulant racheter son contrat [14]. Pour tester notre modèle sur le portefeuille Allianz, le logiciel gratuit `Tanagra` est utilisé. Plusieurs essais ont été réalisés sur différents types de variables et différentes tailles d'échantillons et même sur la totalité de la base de donnée puisque l'algorithme de tanagra supporte une grande quantité de données. La procédure commence par créer l'arbre maximum en regroupant les modalités lorsque nécessaire, puis vient la phase d'élagage pour obtenir l'arbre le plus performant de la plus petite taille possible. Pour ce faire, on minimise le taux d'erreur de l'arbre dans sa phase de construction puis on se fixe un intervalle de confiance pour produire un arbre plus simple tout en conservant un bon niveau de performances. Les arbres qui suivent on été obtenus en sélectionnant les variables les plus significatives issues de la régression logistique et des macros de l'INSEE.

La Table 8 présente les étapes successives de la construction de l'arbre. La performance maximum obtenue est d'environ 70 % de réussite (et ce à la fois sur la base qui sert de modèle et sur l'échantillon indépendant) mais il s'agit là d'une erreur individu par individu et non d'une comparaison des erreurs globales comme c'est le cas pour la régression logistique. Ces résultats sont donc intéressants. L'arbre maximal obtenu compte 602 feuilles, la performance est alors de 27% sur l'ensemble de construction et de 33% sur l'ensemble d'élagage. L'arbre le plus performant selon les deux critères précédents compte lui 22 feuilles, mais on arrive à des performances équivalentes en élaguant jusqu'à ne conserver que 16 feuilles, le taux d'erreur est alors de 31.6% sur l'ensemble de construction et de 32 % sur l'ensemble d'élagage.

La matrice de confusion apparaît sur la Table 9. Elle confronte les vraies valeurs et les valeurs prédites du sexe du conducteur sur les 50000 observations ayant participé à l'apprentissage (growing + pruning). Elle est accompagnée du taux d'erreur qui est de 0.3179 dans notre exemple.

N	Leaves	Err (growing set)	Err (pruning set)	SE (pruning set)	x
67	1	0,3524	0,3559	0,0037	10,258368
62	16	0,3164	0,3208	0,0036	0,584760
59	22	0,3121	0,3187	0,0036	0,000000
1	602	0,2685	0,3347	0,0037	

TABLE 8 – Trees sequence (67)

Error rate			0.3179			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		H	F	Sum
H	0,8674	0,2928	H	28034	4287	32321
F	0,3435	0,4138	F	11606	6073	17679
			Sum	39640	10360	50000

TABLE 9 – Confusion matrix

TANAGRA nous indique (Table 10) que parmi les 50000 observations dédiées à l'apprentissage, il a réservé 33500 observations pour l'expansion de l'arbre (growing set), 16500 pour le post élagage (pruning set). La partition a été effectuée de manière aléatoire. L'arbre associé à ce modèle est représenté dans la Table 11.

Growing set	33500
Pruning set	16500

TABLE 10 – Data partition

Sur les 21 variables descriptives retenues pour caractériser le sexe, le programme n'en retient que 11. Le premier noeud correspond à la variable SEGMENT qui serait donc la plus discriminante sur notre panel. La décision est prise directement si on n'appartient pas au SEGMENT B. Sinon, on regarde la CSP qui divise le panel en deux groupes de taille équivalentes, c'est ensuite la combinaison de choix parmi les classes de kilométrages, la marque, l'usage, la classe SRA, la vitesse, l'âge d'obtention du permis, la parenté avec un assuré, la puissance et le prix du véhicule qui répartissent les genres avec plus de complexité que l'approche des variables prises individuellement. Par exemple, une FIAT de classe de kilométrage A ou F, dont la vitesse n'est pas élevée et qui appartient à un salarié enfant d'assuré sera plutôt conduite par un homme, alors que les FIAT étaient plutôt féminines prises séparément. Les hommes qui sont plus nombreux dans le portefeuilles Allianz comptent 9 feuilles pour seulement 7 pour les femmes.

On teste la solidité du du modèle de prédiction avec un ensemble test de 250 000 polices indépendantes des ensembles growing et pruning qui participent, chacun à leur manière, à l'élaboration de l'arbre et donnaient donc une estimation optimiste. La Table 12 donne les résultats du test. Nous obtenons un taux d'erreur de 0.3211 calculé sur les individus que nous avons mis de côté initialement.

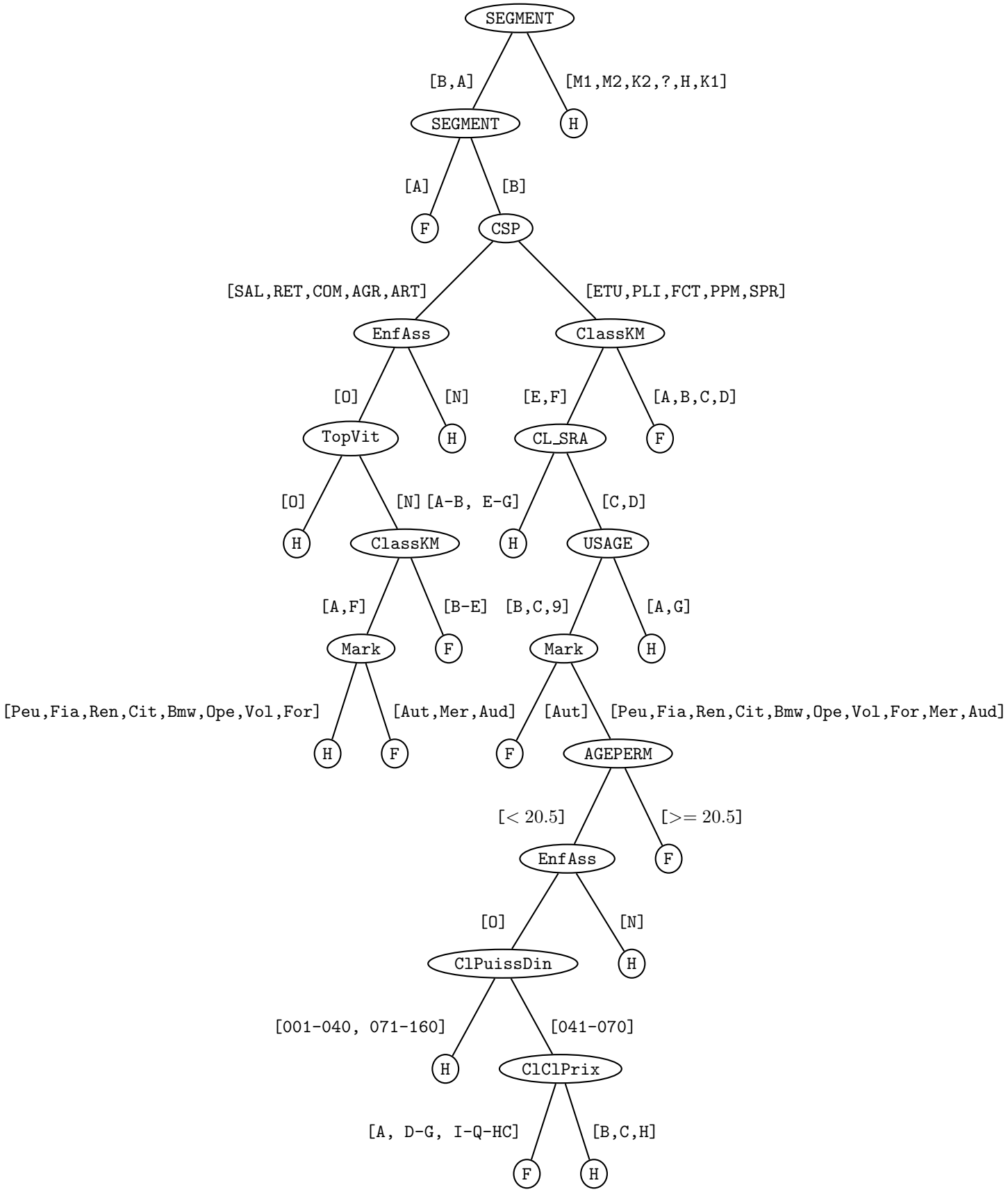

TABLE 11 – Classification tree

Error rate			0,3211			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		H	F	Sum
H	0,8640	0,2929	H	140449	22108	162557
F	0,3347	0,4303	F	58176	29267	87443
			Sum	198625	51375	250000

TABLE 12 – Pred Spv Instance

3.3 Utilisation des GLMs

Différentes catégories de critères entrent en jeu lors de la création d'un modèle linéaire généralisé (GLM). Une bonne gestion de ces catégories doit permettre à l'assureur d'obtenir une modélisation à la fois flexible, précise et adaptée à sa tarification. On distingue d'abord les critères classiques comme la formule de garanties x l'âge du véhicule, la CSP x Usage, l'âge du conducteur x l'âge à l'obtention du permis, les 3 zones VI-DOM et BDG, la présence d'un garage, les niveaux de franchises VI-DOM et BDG, les antécédents de sinistralité en nombre de sinistres et les critères SRA : groupe, classe de prix, énergie, carrosserie, segment, note de sécurité. On a ensuite des critères spécifiques aux novices : Sexe selon le tarif, Ancienneté de noviciat, Conduite accompagnée, Enfant d'assuré, Variables relatives aux parents si enfant d'assurés. On utilise aussi des information issues de l'INSEE : Taille de la population, Age de la population, Pourcentage de (artisans + commerçants + chefs entreprise + cadres + prof intellect sup), Pourcentage de (employés + ouvriers), Evolution de la population, Taux de naissance, Densité, Pourcentage de femmes entre 15 et 29 ans, Nombre de personnes de 5 ans ou plus habitant 5 ans auparavant le même logement, Nombre de personnes de 5 ans ou plus habitant 5 ans auparavant un autre logement de la même commune, Pop 5 ans ou plus habitant 5 ans avant autre commune même département, Nombre de personnes de 5 ans ou plus habitant 5 ans auparavant une région (une autre région) de France métropolitaine, Nombre de personnes de 5 ans ou plus habitant 5 ans auparavant hors de France métropolitaine ou d'un Département d'outre-mer. Et enfin les critères client comme le nombre de contrats auto, l'existence de contrats multirisques habitation (MRH) et autres.

Parmi ces variables, certaines sont numériques comme l'âge du conducteur ou le kilométrage d'autres, la plupart, sont catégorielles comme le sexe ou la CSP. Dans notre modèle, les variables numériques sont transformées en variables catégorielles dont les ajustements en différentes classes se font sur la base des variations de la composante non linéaire des GAM précédemment réalisés. Avant toute chose, la multicolinéarité entre les différentes variables est testée par le facteur d'inflation de la variance (VIF) qui se base sur la méthodes des moindres carrés ordinaires. La Table 13 présente la tolérance et la VIF pour les variables de base du modèle.

Variable	Tolérance	VIF
AGECOND	0.23759	4.20889
AGEVEH	0.87713	1.14008
ANCPERM	0.23957	4.17418
GPSRA	0.68133	1.46771
KMPar	0.97307	1.02768

TABLE 13 – Facteur d'inflation de la variance : variables de base

Une tolérance inférieure à 0.2 correspondant à une VIF supérieure à 5 peut indiquer des problèmes de multicolinéarité. Dans la Table 13, on peut voir que lorsqu'on se limite aux variables de base il n'y a pas de multicolinéarité. En revanche lorsque l'on rajoute des variables spécifiques aux caractéristiques du véhicule (Table 14), la VIF dépasse le seuil de 5 et même 10, il faudra donc se limiter dans le choix des variables à utiliser dans nos modèles.

On choisit d'utiliser la procédure `GENMOD` du logiciel SAS pour modéliser l'ensemble des sinistres en RC responsable : RC0, non responsable : RC1 et bris de glace : BDG. La méthode retenue consiste à diviser notre panel d'environ 600 000 polices en deux : une base de 400 000 pour créer le modèle et un échantillon avec le reste des observations pour le tester. On se propose alors de comparer la solidité des modèles avec et sans la variable sexe, d'abord à partir des critères classiques comme la vraisemblance, et les critères d'information d'Akaike et bayésien, puis en appliquant les prédicteurs du modèle sur l'échantillon. Le portefeuille est aussi divisé selon que la police concerne un enfant

Variable	Tolérance	VIF
AGEVEH	0.72372	1.38176
GPSRA	0.09675	10.33565
KMPar	0.97453	1.02614
COUPLEMOTMAXI	0.87065	1.14856
VITMAXI	0.12134	8.24098
PTAC	0.4448	2.2482

TABLE 14 – Facteur d'inflation de la variance : variables du véhicule

d'assuré ou non, car les variables explicatives du modèle ne seront pas les mêmes.

On compare les modèles obtenus avec et sans la variable sexe du conducteur. Les critères d'évaluation de l'adéquation (Table 15) indiquent d'abord une sous dispersion car la valeur/DDL de la déviance est faible (0.2 comparé à 1). On choisit ici PSCALE pour traiter les problèmes de dispersion. Une sous estimation des écarts types surestime les statistiques de test et augmente la significativité de nos variables explicatives. On introduire donc un terme de bruit qui correspond à la variance du nombre de sinistres non expliquée par les variables.
La qualité des deux modèles est proche avec un léger avantage pour le modèle sans sexe selon la vraisemblance (-24569.37 contre -24582.93), l'AIC (49408.75 contre 49437.87) et le BIC (50817.88 contre 50857.43).

		sans Sexe		avec Sexe	
Critere	DDL	Valeur	Valeur/DDL	Valeur	Valeur/DDL
Deviance	250000	54127.3863	0.2147	54078.5844	0.2145
Scaled Deviance	250000	37799.6975	0.15	37812.6824	0.15
Pearson Chi-Square	250000	360950.9351	1.432	360500.2271	1.4302
Scaled Pearson X2	250000	252069	1	252068	1
Log Likelihood		-24451.6207		-24465.0321	
Full Log Likelihood		-24569.3747		-24582.9329	
AIC (smaller is better)		49408.7494		49437.8658	
AICC (smaller is better)		49408.8951		49438.0136	
BIC (smaller is better)		50817.8786		50857.4329	

TABLE 15 – Critères d'évaluation de l'adéquation

La statistique LR pour Analyse de Type III (Table 16 et 17) donne les p-values de chaque variable indépendamment de leur ordre d'apparition. Il s'agit de voir d'une part si la variable SEXE est significative en terme de p-value (inf. à 5%), d'autre part si son ajout au modèle diminue la significativité des autres variables.

Source	DDL	Khi-2	Pr>Khi-2	Khi-2	Pr> $Khi-2$
Cl_FormVeh	13	24.61	0.0259	24.1	0.0302
CSP	9	62.49	< .0001	53.45	< .0001
USAGE	4	68.86	< .0001	70.65	< .0001
Cl_age_perm	35	281.48	< .0001	270.38	< .0001
Cl_Energie	2	33.99	<.0002	33.58	<.0002
SEGMENT	7	22.32	0.0022	17.73	0.0132
Cl_PuissDin	15	37.45	0.0011	37.51	0.0011
Cl_Zone_RCDM	23	50.73	0.0007	49.41	0.0011
classkm	5	131.28	< .0001	126.9	< .0001
SEXE	1			34.12	< .0001
AncNov	2	35.07	< .0001	35.67	< .0001
FormPar	9	18.88	0.0263	18.8	0.027
DENSITE	5	44.6	< .0001	43.68	< .0001
Cl_NbAuto	5	145.1	< .0001	144.08	< .0001

TABLE 16 – RC1 : Statistique LR pour Analyse de Type III

Pour les sinistres responsables la variable SEXE est fortement significative, mais elle ne diminue que faiblement la p-value des autres variables. Seule le SEGMENT perd 1% de significativité.

Source	DDL	Khi-2	Pr>Khi-2	Khi-2	Pr> $Khi-2$
Cl_FormVeh	13	76.75	$<.0001$	76.3	$<.0001$
CSP	9	19.77	0.0194	19.66	0.0202
USAGE	4	56.12	$<.0001$	56.1	$<.0001$
Cl_age_perm	35	148.09	$<.0001$	147.39	$<.0001$
Cl_ClPrix	16	37.16	0.002	36.42	0.0025
Cl_Zone_RCDM	23	39.27	0.0185	39.27	0.0185
CRapPP	9	30.16	0.0004	30.16	0.0004
classkm	5	41.75	$<.0001$	41.72	$<.0001$
SEXE	1			0	0.9681
FormPar	9	22.13	0.0085	22.12	0.0085
TRPOP	5	28.41	$<.0001$	28.41	$<.0001$
CSPPLUS	4	13.14	0.0106	13.14	0.0106

TABLE 17 – RC0 : Statistique LR pour Analyse de Type III

Pour les sinistres non responsables la variable SEXE n'est plus significative.

On utilise alors un échantillon test indépendant des données ayant servi à calculer les prédicteurs pour vérifier la solidité du modèle. Le GLM pour les enfants d'assurés est obtenu sur une base de 250 000 polices. Il est ensuite testé sur un échantillon de 125 000 polices. Les résultats apparaissent dans la Table 18. Les prédictions du modèle sans la variable SEXE sont très proches de celles obtenues avec cette variable et sont même meilleures pour les garanties responsable et BDG.

	SINRC0	SINRC1	SINBDG
Modèles novices sans Sexe			
obs	3148	4124	3583
Pred	3180.36	4124.67	3512.6
Diff. Relative	1.028%	0.016%	1.965%
Modèles novices avec Sexe			
obs	3148	4124	3583
Pred	3180.34	4126.66	3512.12
Diff. Relative	1.027%	0.065%	1.978%

TABLE 18 – Résultats sur échantillon test

4 Comment définir l'expérience des conducteurs novices

4.1 Utilisation des GAMs

Plusieurs paramètres comme l'âge, la durée d'obtention du permis de conduire, le kilométrage parcouru ont une influence avérée sur l'expérience de l'assuré au volant de son véhicule. Ils peuvent avoir des répercussions sur sa sinistralité. Ces trois informations présentent l'avantage d'être quantifiables numériquement, ce qui les rends plus flexibles dans les modèles linéaires que nous utilisons. Nous poursuivons donc l'étude des risques à travers les variables explicatives numériques. Les modèles additifs généralisés (GAM) permettent une approche plus fine de l'influence de ces variables sur la sinistralité et aussi des éventuelles corrélations [9]. On propose ici des modèles univariés, l'objectif étant de mesurer la valeur explicative de la composante non linéaire (spline) de la variable. Lorsque le critère de 5% de significativité est rempli, on modélise graphiquement la spline dont les variations nous permettront de délimiter des classes de risques plus homogènes que celles obtenues avec les modèles linéaires [15]. On commence par l'âge du conducteur modélisé selon les catégories de risques entre 2008 et 2010. Seules les catégories pour lesquelles la spline s'avère significative sont représentées.

Valeurs estimées des paramètres				
Paramètre	Valeur estimée des paramètres	Erreur type	Valeur du test t	$Pr > \|t\|$
Intercept	-1.08495	0.03504	-30.97	$< .0001$
Linear(AGECOND)	-0.03564	0.00157	-22.74	$< .0001$
Analyse du modèle de lissage				
Récapitulatif d'ajustement pour composantes du lissage				
Composante	Paramètre de lissage	DDL	GCV	Obs unique num
Spline(AGECOND)	0.323351	9.244633	0.001423	18
Smoothing Model Analysis				
Approximate Analysis of Deviance				
Source	DDL	Khi-2	$Pr >$ Khi-2	
Spline(AGECOND)	9.24463	197.4399	$< .0001$	

TABLE 19 – Analyse du modèle de régression

La Table 19 présente les résultats du GAM pour le nombre total de sinistres en 2010. On a choisi de se focaliser sur la tranche 18 - 35 ans qui représente la majorité des novices (86%), soit un panel d'environ 190 000 polices. Les tests permettent de conclure que pour cette garantie, l'âge du conducteur est à la fois significatif par sa composante linéaire et par sa composante spline qui est représentée ci-dessous dans la figure 3.

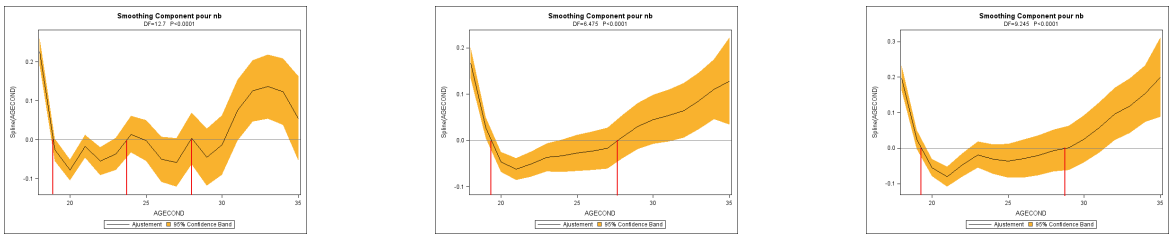

FIGURE 4 – Influence non linéaire de l'âge du conducteur sur la sinistralité totale entre 2008 et 2010

Sur la Figure 4 on peut voir entre 2008 et 2010 l'évolution de la relation entre l'âge du conducteur et la sinistralité toutes catégories confondues. Selon l'étude graphique des splines de l'âge du conducteur, les variations les plus fortes correspondent aux tranches 18-19 ans et 19-23 ans. Mis à part en 2008, on a une certaine stabilité de 23 à 28 ans, l'augmentation est ensuite régulière après 28 ans. Pour les conducteurs les plus âgés dont la proportion augmente chaque année, on pourra se référer à l'étude américaine de l'Oak Ridge Institute [8].

Pour l'ancienneté du permis (Figure 5), les plus fortes variations sont entre 0 et 5 ans avec un minimum vers 3 ans, on a ensuite en 2008 et 2010 deux zones stables comprises entre 6 et 11, puis entre 11 et 15 ans de permis.

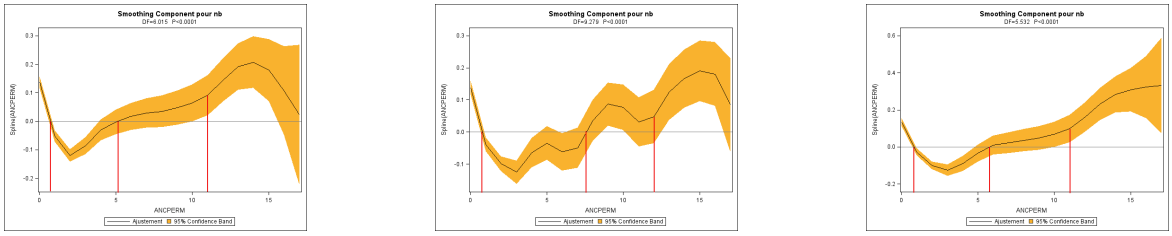

FIGURE 5 – Influence non linéaire de l'ancienneté du permis sur la sinistralité totale entre 2008 et 2010

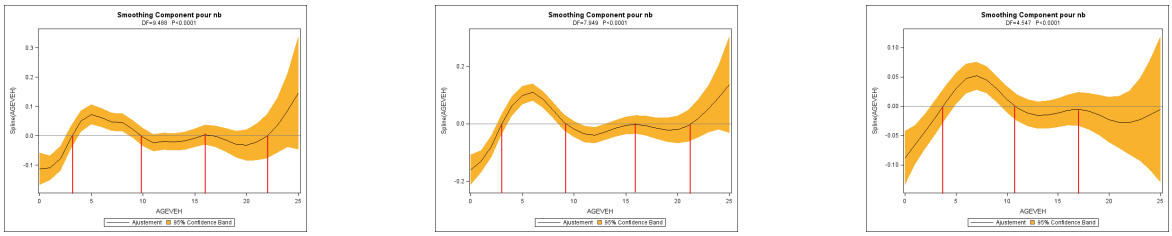

FIGURE 6 – Influence non linéaire de l'âge du véhicule sur la sinistralité totale entre 2008 et 2010

Pour l'âge du véhicule (Figure 6), on a choisi de se focaliser sur la tranche 0 - 25 ans qui représente la quasi totalité des novices (99%). Les plus fortes variations sont entre 0 et 10 ans avec un maximum vers 6 ans, on a ensuite une zone stable comprise entre 10 et 17, au delà incertitude due à la faible représentativité l'emporte.

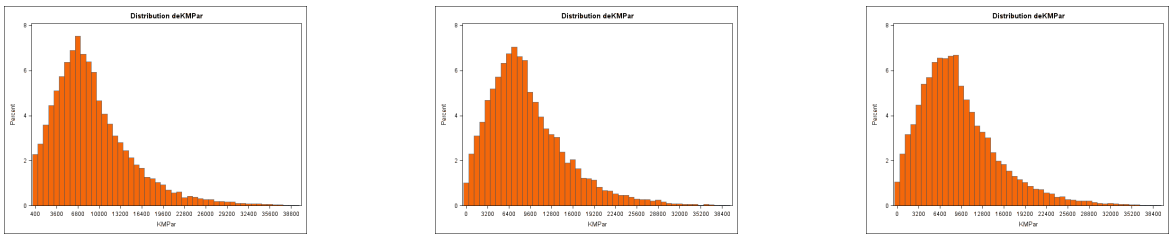

FIGURE 7 – Histogrammes du kilométrage annuel parcouru entre 2008 et 2010

En ce qui concerne le kilométrage (Figure 7 et Figure 8), la quantité de données disponible est limitée par la présence d'un relevé SOFCA pour le véhicule. On récupère ainsi une moyenne de 40 000 assurés par ans. Le kilométrage annuel moyen se situe autour de 9000 km et la grande majorité (99%) des assurés parcoure moins de 30 000 km chaque année.

Pour obtenir la convergence du GAM, nous avons du utiliser des variables centrées réduites. La composante spline est alors significative et présente des variations notables entre -1 et 0.5 puis une certaine stabilité jusqu'à 1.5 et enfin une tendance linéaire après 3, ce qui correspond aux paliers réels 3000, 12000, 18000 et 27000 km. Nous tenterons par la suite d'évaluer la part d'expérience acquise par les novices et donc la diminution de sinistralité induite selon leur tranche de kilométrage.

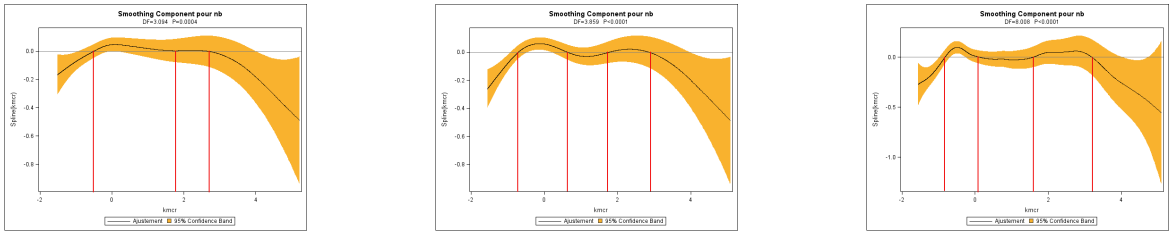

FIGURE 8 – Influence non linéaire du kilométrage annuel sur la sinistralité totale entre 2008 et 2010

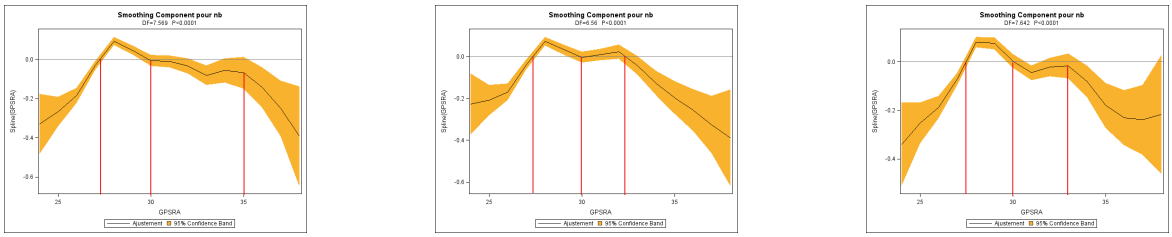

FIGURE 9 – Influence non linéaire du groupe SRA sur la sinistralité totale entre 2008 et 2010

Le groupe SRA (Figure 9) permet de classer les véhicules par type, on se focalise ici sur la tranche 25 - 35 qui représente la majorité des assurés. Sur les trois années, les courbes sont assez similaires avec une première tranche jusqu'à 28, puis à 30 et enfin après le groupe 33.

4.2 Acquisition de l'expérience au volant

On se pose ici la question de l'expérience acquise par les assurés au cours de leurs trois années de noviciat, et plus particulièrement l'évolution de leur sinistralité selon le nombre de kilomètre parcouru chaque année. D'autres rapports comme [3] ou [7] développent aussi la relation entre expérience et kilométrage dans le cadre du *Pay as You Drive*, mais nous nous focalisons ici sur les conducteurs novices. On peut en effet se poser la question de l'adaptation de cette période de 3 ans pour l'ensemble des assurés dont l'expérience au volant au cours de cette période est différente selon l'utilisation qu'ils font de leur véhicule [11]. D'après les résultats obtenus précédemment via les GAM, on choisit de créer différentes catégories de novices selon la durée mise par ces derniers pour passer le pallier des 12 000 km. Ce découpage particulier s'est avéré plus efficace qu'un découpage classique par tranche de kilomètres parcourus annuellement car il reflète davantage l'évolution de la conduite de l'assuré durant son noviciat. Les quatre catégories retenues sont les suivantes :

- A : Moins d'un an pour parcourir 12 000 km
- B : Moins de deux ans pour parcourir 12 000 km
- C : Moins de trois ans pour parcourir 12 000 km
- D : Plus de trois ans pour parcourir 12 000 km

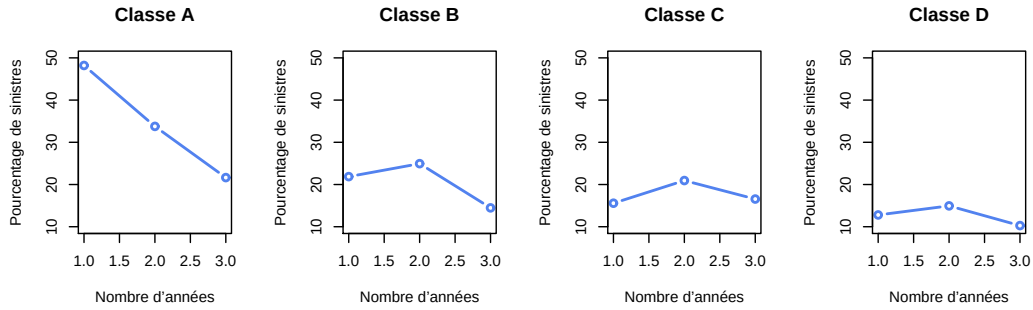

FIGURE 10 – Évolution de la sinistralité moyenne annuelle entre 2008 et 2010

Sur la Figure 10, on a représenté le nombre moyen de sinistres annuel. Un premier constat naturel est que les conducteurs parcourant un grand nombre de kilomètres (classe A) sont plus exposés à la sinistralité et comptent donc un pourcentage de sinistres bien supérieur aux autres classes. Pour autant, ce sont les assurés de la classe A qui progressent le plus au cours de leurs trois années de

noviciat avec une diminution de plus de 50%. Pour les autres classes l'amélioration est beaucoup plus mesurée et passe même par une régression en cours de deuxième année. On choisit donc d'observer à nouveau nos quatre classes mais cette fois-ci, en terme de nombre de sinistres par kilomètres parcourus pour faire apparaître l'exposition réelle du véhicule au risque sur une même distance.

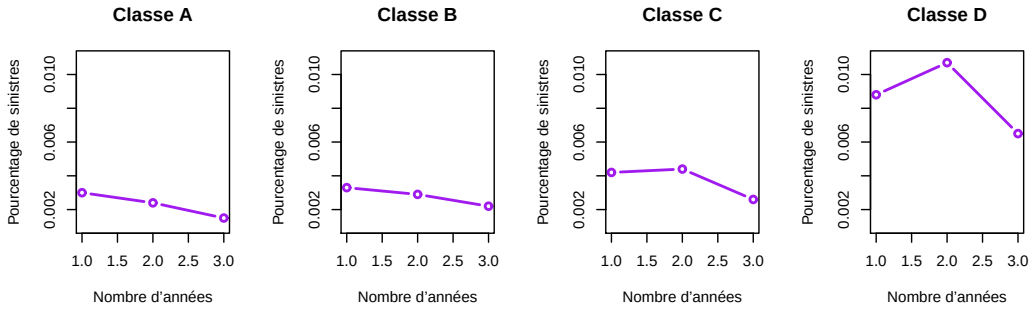

FIGURE 11 – Évolution de la sinistralité moyenne annuelle par kilomètre entre 2008 et 2010

Sur la Figure 11, la sinistralité moyenne est estimée en fonction du kilométrage. Une première constatation est l'inversion dans l'échelle de la sinistralité, la classe A devient la plus performante et la classe D présente un pourcentage de sinistres par kilomètre nettement supérieur aux trois autres classes. Concernant l'amélioration la classe A est aussi la meilleure avec une progression de 50% alors que les classes B, C et D progressent respectivement de 33, 38 et 26%. Les assurés qui roulent beaucoup deviennent donc a priori plus vite de bons conducteurs, ce qui pourrait justifier une sortie anticipée du noviciat, mais ils demeurent les individus les plus risqués pour l'assureur sur une période de contrat annuel. En terme de garanties, on peut séparer les sinistres pour déterminer si cette diminution du risque s'avère plus importante lorsque l'assuré est responsable que lorsqu'il est non responsable de son accident.

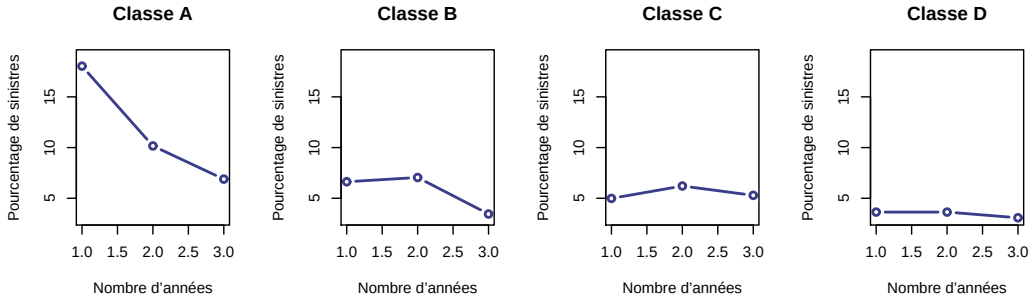

FIGURE 12 – Évolution du nombre moyen de sinistres responsables annuel entre 2008 et 2010

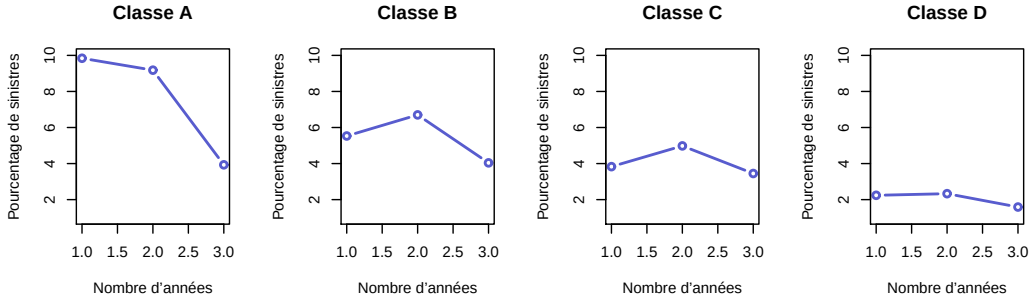

FIGURE 13 – Évolution du nombre moyen de sinistres non responsables annuel entre 2008 et 2010

Sur les graphiques 12 et 13, on observe une diminution plus rapide du nombre de sinistres responsables, la classe A offre toujours la meilleure progression et se positionne au même niveau de risque que les autres classes après deux ans de noviciat. Pour autant les sinistres non responsables finissent aussi par diminuer même si la deuxième année présente une légère régression pour les classes B et C.

On propose ensuite une nouvelle classification des conducteurs en lien direct avec leur kilométrage annuel. On commence par une approche statique, quelle que soit l'année de noviciat, le conducteur se voit attribuer une classe allant de A à D correspondant aux tranches 0-3000, 3000-6000, 6000-12000 et plus de 12000 kilomètres. Il n'y a donc pas de suivi des assurés qui peuvent migrer d'une classe à l'autre au cours de leurs trois années d'observation. On compare le nombre moyen de sinistres selon les classes, année par année.

	Fréquence annuelle					Nombre d'assurés			
année	A	B	C	D		A	B	C	D
1	6.63%	8.72%	11.39%	13.40%		2424	5092	11219	7360
2	6.81%	7.81%	8.83%	14.10%		2645	4898	10545	7140
3	4.16%	7.68%	8.45%	13.75%		2676	4230	7429	3696

TABLE 20 – Sinistres responsables

	Fréquence annuelle					Nombre d'assurés			
année	A	B	C	D		A	B	C	D
1	5.83%	5.65%	8.34%	9.68%		2424	5092	11219	7360
2	5.26%	5.89%	8.36%	12.79%		2645	4898	10545	7140
3	2.77%	4.37%	6.76%	9.56%		2676	4230	7429	3696

TABLE 21 – Sinistres non responsables

Dans les Tables 20 et 21 on peut comparer les fréquences de sinistres et le nombre d'assurés de chaque catégorie de kilométrage selon la nature du sinistre (responsable ou non responsable). Les résultats de cette approche statique ne font que confirmer la plus faible sinistralité des classes de kilométrage les moins élevées. La troisième année de noviciat reste la meilleure en terme de diminution du risque, mais la progression la plus forte est observée chez les conducteurs de la classe A, ce qui, contrairement au résultat précédent, ne va pas dans le sens d'une meilleure maîtrise du véhicule lorsque l'on roule beaucoup. On peut noter que la diminution du risque est plus forte en cas de sinistre non responsable dans la classe A (-3.06 contre -2.47%) et en cas de sinistre non responsable dans la classe C (-2.94 contre -1.58%). Cependant nous n'observons pas ici le kilométrage cumulé mais un état instantané qui ne reflète pas forcément l'expérience acquise par le conducteur au cours de ces trois ans. On opte alors pour une approche plus dynamique en considérant les tranches successives de kilométrage sur deux ans.

1ère année	2ème année	Classe	Moy1	Moy2	nb
$0-3000$	$0-3000$	A	5.82%	12.01%	1132
	$3000-6000$	B	4.52%	3.44%	315
	$6000-12000$	C	10.19%	7.33%	195
	>12000	D	15.81%	1.79%	56
$3000-6000$	$0-3000$	E	3.64%	2.29%	313
	$3000-6000$	F	8.86%	9.94%	2150
	$6000-12000$	G	6.85%	7.38%	973
	>12000	H	12.95%	9.28%	245
$6000-12000$	$0-3000$	I	9.72%	7.26%	137
	$3000-6000$	J	6.81%	6.75%	594
	$6000-12000$	K	11.91%	9.63%	5865
	>12000	L	7.76%	9.08%	736
>12000	$0-3000$	M	15.02%	0.00%	20
	$3000-6000$	N	28.34%	13.16%	19
	$6000-12000$	O	35.38%	36.64%	55
	>12000	P	14.40%	18.65%	4070

TABLE 22 – Fréquence annuelle de sinistres responsables en première et deuxième année

Les assurés sont ainsi classés en seize groupes de A à P, les tranches restant les mêmes, de 0 à 3000 km, de 3000 à 6000 km, de 6000 à 12000 km et plus de 12000 km, le kilométrage de la première année d'observation conditionnant l'appartenance au groupe de la deuxième année. Par exemple la

classe C correspond à une distance parcourue comprise entre 0 et 3000 km la première année et entre 6000 et 12000 la deuxième. Les résultats suivants sont présentés dans quatre tableaux successifs pour les observations de sinistres responsables (Tables 22 et 23) et non responsables (Tables 24 et 25) des novices entre leur première et leur deuxième année puis entre leur deuxième et leur troisième année. On s'intéresse particulièrement aux catégories pour lesquelles on observe une diminution de la sinistralité d'une année sur l'autre. Ces groupes apparaîtront en bleu.

2ème année	3ème année	Classe	Moy1	Moy2	nb
$0-3000$	$0-3000$	A	3.47%	5.10%	1176
	$3000-6000$	B	6.02%	1.56%	257
	$6000-12000$	C	3.22%	10.86%	141
	>12000	D	6.54%	11.32%	45
$3000-6000$	$0-3000$	E	0.77%	2.90%	261
	$3000-6000$	F	6.99%	9.97%	1891
	$6000-12000$	G	4.52%	6.59%	642
	>12000	H	8.52%	8.37%	123
$6000-12000$	$0-3000$	I	2.36%	0.00%	105
	$3000-6000$	J	3.34%	2.60%	444
	$6000-12000$	K	7.85%	7.99%	3932
	>12000	L	3.09%	8.69%	356
>12000	$0-3000$	M	0.00%	0.00%	8
	$3000-6000$	N	7.69%	8.37%	13
	$6000-12000$	O	4.17%	6.65%	32
	>12000	P	11.10%	16.33%	1811

TABLE 23 – Fréquence annuelle de sinistres responsables en deuxième et troisième année

Une première constatation est que les effectifs stables (A, F, K, P) sont les plus représentés : les assurés ont donc tendance à rester dans la même tranches de kilométrages d'une année sur l'autre. Les plus nombreux sur les deux premières années sont les groupes K et P qui correspondent aux tranches 6000-12000 km et plus de 12000 km avec combinés près de 10000 personnes. Entre la deuxième et la troisième année, les écarts sont moins marqués, le groupe K demeure le plus important. D'un autre côté les conducteurs ayant peu roulé une année et beaucoup la suivante (ou réciproquement) sont très peu nombreux. Les groupes D et M comptent seulement une cinquantaine et une vingtaine d'individus respectivement.

En ce qui concerne les sinistres responsables, on remarque que les groupes progressent plus au cours de la première que de la deuxième année. Parmi les groupes stables, seul le K présente entre la première et la deuxième année une diminution significative de la fréquence annuelle des sinistres (environ 2%). Entre la deuxième et la troisième année de noviciat aucun grand groupe ne progressent et seuls les groupes B, H, I et J ont une diminution de la sinistralité. Contrairement à l'approche précédente, les plus gros rouleurs ne sont pas ceux qui progressent le plus, le groupe P notamment ne progresse ni en première ni en deuxième année.

Pour les sinistres non responsables aucun des grands groupes ne progresse entre la première et la deuxième année, seuls quelques petits groupes (B, C, E, G, I, L) équitablement répartis dans les trois premières tranches de kilométrage montrent une légère diminution de la sinistralité. Entre la deuxième et la troisième année, on note davantage de progression, en particulier dans les grands groupes comme A, F et P. Avec cette catégorisation des assurés, on voit ressortir deux sortes d'évolutions selon la nature (responsable ou non) du sinistre. La deuxième année, les conducteurs des principales catégories progressent en moyenne moins bien en ce qui concerne leurs nombres de sinistres responsables, mais ils ont moins de sinistres non responsables.

1ère année	2ème année	Classe	Moy1	Moy2	nb
$0-3000$	$0-3000$	A	4.00%	7.35%	1132
	$3000-6000$	B	6.14%	4.85%	315
	$6000-12000$	C	5.23%	2.34%	195
	>12000	D	4.59%	12.13%	56
$3000-6000$	$0-3000$	E	5.17%	3.39%	313
	$3000-6000$	F	6.71%	6.94%	2150
	$6000-12000$	G	5.20%	4.82%	973
	>12000	H	12.15%	12.98%	245
$6000-12000$	$0-3000$	I	7.03%	2.28%	137
	$3000-6000$	J	4.35%	4.84%	594
	$6000-12000$	K	8.05%	10.77%	5865
	>12000	L	8.97%	8.31%	736
>12000	$0-3000$	M	0.00%	0.00%	20
	$3000-6000$	N	7.27%	38.41%	19
	$6000-12000$	O	10.12%	42.18%	55
	>12000	P	11.00%	15.69%	4070

TABLE 24 – Fréquence annuelle de sinistres non responsables en première et deuxième année

2ème année	3ème année	Classe	Moy1	Moy2	nb
$0-3000$	$0-3000$	A	3.22%	2.12%	1176
	$3000-6000$	B	4.11%	4.08%	257
	$6000-12000$	C	1.51%	3.55%	141
	>12000	D	14.37%	4.45%	45
$3000-6000$	$0-3000$	E	1.67%	5.42%	261
	$3000-6000$	F	5.87%	4.47%	1891
	$6000-12000$	G	4.36%	5.29%	642
	>12000	H	5.36%	9.54%	123
$6000-12000$	$0-3000$	I	4.72%	4.76%	105
	$3000-6000$	J	3.73%	4.21%	444
	$6000-12000$	K	6.91%	7.28%	3932
	>12000	L	8.09%	6.61%	356
>12000	$0-3000$	M	0.00%	0.00%	8
	$3000-6000$	N	0.00%	7.69%	13
	$6000-12000$	O	3.13%	15.29%	32
	>12000	P	11.26%	10.66%	1811

TABLE 25 – Fréquence annuelle de sinistres non responsables en deuxième et troisième année

5 Conclusion

En conclusion de cette démarche, la variable sexe, si elle permet de distinguer avec facilité des fréquences et des coûts de sinistres différents, ne s'avère pas indispensable à la construction d'un modèle de prédiction en assurance automobile. D'autres variables sur le type de véhicule, le parcours de l'assuré et son comportement en terme de kilométrage apportent une information aussi complète.

Les résultats obtenus pour la partie "expérience" de conduite des novices confirment l'utilité d'une information précise sur le kilométrage annuel de chaque assuré. Nous avons ainsi pu analyser l'évolution de la sinistralité des novices durant trois années consécutives. Les conducteurs ayant parcouru le plus de kilomètres progressent plus vite que les autres. Il faut donc différencier la fréquence annuelle de sinistres qui est naturellement plus élevée lorsque l'on conduit souvent, de la fréquence de sinistres par kilomètre qui reflète davantage la maîtrise du véhicule. Ces constatations vont dans le sens des évolutions observées sur le marché autour de la conduite connectée et de la tarification au comportement, avec l'apparition de nouvelles options comme celle baptisée Allianz Conduite Connectée. Une tarification à la fois compétitive et en adéquation avec les habitudes routières des assurés devrait ainsi voir le jour si les compagnies font bon usage des nouvelles informations à leur disposition.

6 Lexique

Variables dans le portefeuilles	
AGECOND	age de la personne assurée
AGEVEH	ancienneté du véhicule
ANCPERM	nombre d'années depuis l'obtention du permis de conduire
ANCNOV	nombre d'années de noviciat
PUISSADM	puissance officielle du véhicule
NBPLACES	nombre de places
PUISS	puissance du véhicule
COUPLEMOTMAXI	taille du moteur du véhicule
TOPVIT	catégorie de vitesses
VITMAXI	vitesse maximale du véhicule
PTAC	poids total autorisé en charge
CDSEXCON	sexe du conducteur
CARROS	carrosserie
CSP	catégorie socioprofessionnelle
GPSRA	catégorie pour la sélection SRA
CDPRIME	catégorie de prime
CLPRIX	classe de prix du véhicule
CLASSKM	kilométrage annuel par catégorie
TXPRIME	taux de prime
ENERGIE	type d'énergie du véhicule
TRANSM	transmission
TYPE	type de véhicule
ALIM	alimentation
BOITEVIT	boîtier de vitesses
NBRAP	nombre de vitesses
SUSPENS	suspension
ASSISTFR	Assistance au freinage
ABS	anti-lock braking system
USAGE	type d'usage du véhicule
INSEE variables par communes	
TRPOP	Taille de la population
TRAGE	Age de la population
NAISSANCE	Taux de naissance
EVOL	Evolution de la population (2008/2010)
DENSITE	Nb habitants / superficie km^2
PCTCHGTLOGT	Pourcentage de changements de logement
PCTCHGTCOM	Pourcentage de changements de commune

TABLE 26 – Les variables explicatives utilisées

Sinistre RC corporel non responsable	RCC0
Sinistre RC corporel responsable	RCC1
Sinistre IDA non responsable	IDA0
Sinistre IDA responsable	IDA1
Sinistre RC matériel non IDA non responsable	RCM0
Sinistre RC matériel non IDA responsable	RCM1
Sinistre Bris de glace	BDG
Sinistre Vol Incendie	VI
Sinistre Dommage non responsable	DOM0
Sinistre Dommage responsable	DOM1
Sinistre Assistance	ASSI

TABLE 27 – Classification des sinistres

Références

[1] Pay as you drive (payd) insurance pilot program phase 2 mid-course project report. Technical report, 2007.

[2] Use less, pay less - a simple concept that reduces the cost of car insurance now available to michigan and oregon drivers. Technical report, 2007.

[3] Pay-as-you-drive vehicle insurance - converting vehicle insurance premiums into use-based charges. Technical report, 2014.

[4] Pierre Arnal and Romain Durand. Une vie sans sexes. comment le sexe devint genre, et comment le genre devint code : Le sort cruel d'une variable explicative. *Risques - Les cahiers de l'assurance*, 87 :1–7, 2011.

[5] Jason E. Bordoff and Pascal J. Noel. Pay-as-you-drive auto insurance : A simple way to reduce driving-related harms and increase equity. *The Brookings Intitution*, pages 1–58, 2008.

[6] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[7] Allen Greenberg. Designing pay-per-mile auto insurance regulatory incentives using the nhtsa light truck cafe rule as a model. Technical report, 2009.

[8] Patricia S. Hu, Donald W. Jones, Timothy Reuscher, Richard S. Schmoyer Jr., and Lorena F. Truett. Projecting fatalities in crashes involving older drivers, 2000-2025. Technical report, 2000.

[9] Nadja Klein, Michel Denuit, Stefan Lang, and Thomas Kneib. Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance Math. Econom.*, 55 :225–249, 2014.

[10] Romuald Le Lan. Analyse de donneées et classification sur données d'enquete - choix sur les variables, le nombre de classes et le nombre d'axes. Technical report, 2003.

[11] Jean Lemaire, Sojung Park, and Kili Wang. The use of annual mileage as a rating variable. *STAT Discussion Paper*, pages 1–26, 2014.

[12] Todd Litman. Pay-as-you-drive pricing in british columbia. *Victoria Transport Policy Institute*, pages 1–11, 2011.

[13] L.Dawn Massie, Paul E. Green, and Kenneth L. Campbell. Crash involvement rates by driver gender and the role of average annual mileage. *Accid. Anal. and Prev.*, 29 :675–685, 1997.

[14] Xavier Milhaud, Stéphane Loisel, and Véronique Maume-Deschamps. Facteurs explicatifs du rachat en Assurance-Vie : classification et prévisions du risque de rachat. In *42èmes Journées de Statistique*, Marseille, France, France, 2010.

[15] Sandra Pitrebois, Michel Denuit, and Jean-François Walhin. Personnalisation des primes-frequence en assurance automobile par regression poissonienne en presence de donnees longitudinales. *STAT Discussion Paper*, pages 1–29, 2001.

[16] Roger Roots. The dangers of automobile travel : A reconsideration. *American Journal of Economics and Sociology*, 66 :959–975, 2007.

[17] Andre Thepaut. Quelques reflexions sur la reforme du tarif français d'assurance automobile. *Astin*, 2 :109–124, 1961.