

ROYAUME DU MAROC
..*.*
HAUT COMMISSARIAT AU PLAN
..*.*.*.*.*

INSTITUT NATIONAL
DE STATISTIQUE ET D'ECONOMIE APPLIQUEE

INSEA



Projet de Fin d'Etudes

*Analyse actuarielle du portefeuille automobile.
GLM et Crédibilité*

Préparé par : *Mme Itre Mtalai*

Mlle Kawtar Bouifadden

Sous la direction de : *Mr Kamal Benchekroun (INSEA)*
Mr Abderrahim Dbich (AXA Assurance Maroc)

Soutenu publiquement comme exigence partielle en vue de l'obtention du

Diplôme d'Ingénieur d'Etat

Option : Actuariat-Finance

Devant le jury composé de :

- *Mr Kamal Benchekroun (INSEA)*
- *Mr Said Ramadan Nsiri (INSEA)*
- *Mr Abderrahim Dbich (AXA Assurance Maroc)*

Juin 2012

Résumé

Le présent mémoire a pour objectif l'analyse statistique et actuarielle du portefeuille automobile d'AXA Assurance Maroc. Pour ce faire, nous présentons en premier temps le marché de l'automobile au Maroc et les différents types de contrats et de garanties.

Ensuite, nous procédons par une étude statistique uni-variée et multi-variée afin d'aboutir à une segmentation du portefeuille.

Nous enchainons par l'application du modèle linéaire Généralisé sur la fréquence et le coût moyen. Ces derniers ayant fait objet d'une modélisation à priori.

Finalement un traitement correctif est envisageable via la théorie de crédibilité qui fournira des conclusions vis-à-vis de la rentabilité des contrats et des décisions à envisager en termes de l'estimation de leur sinistralité.

Mots clés :

Segmentation, analyse de données, écrêtement, modèle linéaire généralisé, ratio S/P, crédibilité.

Dédicace

*A mon très cher oncle
Saber Ali*

*A ma précieuse petite
famille*

Fre

*A mes très chers parents
A mon unique sœur
A mes frères Jawad et
Othmane
A mon fiancé*

Kawtar

Remerciements

Au terme de ce travail, nous tenons à exprimer notre gratitude à tous ceux qui ont prêtés main forte de près ou de loin pour la réalisation de ce projet de fin d'études.

Nous adressons avec tout le respect et l'estime que cela se doit de requérir, nos remerciements à nos encadrants: Mr Kamal Benchekroun et Mr Abderrahim Dbich qui nous ont épaulés durant le stage et qui ont fait preuve de disponibilité, d'ouverture d'esprit et de professionnalisme.

Un grand merci est adressé à Mr Jamal Harmouch, Mr Fouad Mari et Mr Adil Bensouna pour leur collaboration inestimable.

Finalement, notre reconnaissance va à toute personne ayant contribué à achever à bon port ce modeste travail.

Liste des abréviations

- **AAM** : AXA Assurance Maroc
- **M3T5** : véhicule de poids inférieur à 3 .5T
- **P3T5** : véhicule de poids supérieur à 3 .5T
- **ACP** : analyse en composantes principales
- **CAH** : classification ascendante hiérarchique
- **AFD** : analyse factorielle discriminante
- **CRM** : coefficient de réduction majoration
- **DAPS** : Direction des Assurances et de la Prévoyance Sociale
- **FMSAR** : fédération Marocaine des Sociétés d'Assurances et de Réassurances
- **SP ou S/P** : ratio entre les coûts de sinistres d'un contrat et ses primes acquises
- **GLM** : modèle linéaire généralisé
- **RC** : la garantie responsabilité civile
- **QQ-plot** : quantile-quantile plot

Liste des tableaux

Tableau 1 : structure du chiffre d'affaire du secteur d'assurance au Maroc.....	13
Tableau 2 : évolution du chiffre d'affaire des différentes branches du secteur d'assurance....	13
Tableau 3 : position d'Axa assurance Maroc en termes du chiffre d'affaire non vie.....	15
Tableau 4 : types de garanties offertes par l'AAM.....	17
Tableau 5 : statistiques sur les valeurs manquantes et aberrantes de la base de données.....	24
Tableau 6 : codification des groupements de zones.....	27
Tableau 7 : codification des groupements de l'usage du véhicule.....	30
Tableau 8 : codification des variables âge du véhicule, carburant, âge du conducteur, sexe....	30
Tableau 9 : statistiques sur le portefeuille des contrats Mono.....	31
Tableau 10 : statistiques sur le portefeuille mono par usage de véhicule.....	33
Tableau 11 : libellé des variables de l'ACP.....	38
Tableau 12 : matrice des corrélations des variables de l'ACP.....	39
Tableau 13 : valeurs propres et inertie de l'ACP.....	40
Tableau 14 : coordonnées, contribution et qualité de représentation des unités d'observation de l'ACP.....	40
Tableau 15 : coordonnées, contribution et qualité de représentation des variables de l'ACP....	41
Tableau 16 : classification des variables par contribution et signe de coordonnées vis-à-vis du premier axe.....	42
Tableau 17 : classification des variables par contribution et signe des coordonnées.....	42
Tableau 18 : classification des unités d'observations par contribution et signe de coordonnées vis-à-vis du deuxième axe.....	43
Tableau 19 : classification des variables par contribution et signe des coordonnées vis-à-vis du deuxième axe.....	43
Tableau 20 : étapes de la CAH et mesures du R^2 de chaque étape.....	48
Tableau 21 : corrélation entre âge conducteur et âge permis.....	50
Tableau 22 : résultat du test khi2 d'indépendance entre carburant et usage.....	51
Tableau 23 : résultat du test khi2 d'indépendance entre sexe et zone.....	52
Tableau 24 : résultat du test khi2 d'indépendance entre zone et usage.....	52
Tableau 25 : test d'ajustement de la charge à la distribution Gamma.....	52
Tableau 26 : test d'ajustement de la charge à la distribution Gamma.....	57
Tableau 27 : fonctions de liens usuelles.....	59
Tableau 28 : comparaison des qualités d'ajustement du GLM sur la fréquence.....	62
Tableau 29 : comparaison des Déviations entre Poisson et Binomiale Négative.....	63
Tableau 30 : comparaison des qualités d'ajustement du GLM sur le coût moyen.....	66
Tableau 31 : comparaison des Déviations entre Gamma et log Normale.....	66
Tableau 32 : les facteurs de crédibilité : minimal, maximal et moyen / Bühlmann Straub	76
Tableau 33 : les classes de cotisation.....	82
Tableau 34 : les facteurs de crédibilité : minimal, maximal et moyen pour la classe au sein de la catégorie.....	85
Tableau 35 : exemple de comparaison entre S/P réel et S/P crédibilisé.....	86

Liste des figures

Figure 1 : Graphique représentant la répartition de la fréquence par région et par exercice.....	26
Figure 2 : Graphique représentant la répartition de la fréquence par région.....	26
Figure 3 : Graphique représentant la répartition de la fréquence par région et par exposition au risque.....	27
Figure 4 : cartographie du risque automobile au Maroc.....	28
Figure 5 : Graphique représentant la répartition de la fréquence par usage et par année.....	28
Figure 6 : Graphique représentant la répartition de la fréquence par région.....	29
Figure 7: Graphique représentant la répartition de la fréquence par région et par exposition au risque.....	29
Figure 8 : fréquence des sinistres pour chaque groupement de zone et par année.....	32
Figure 9 : Le graphique représentant la charge des sinistres pour chaque zone de circulation et pour chaque année	32
Figure 10 : graphique représentant la fréquence des sinistres pour chaque usage et pour chaque année	33
Figure 11 : Graphique suivant représente la charge des sinistres pour chaque usage et pour chaque année.....	34
Figure 12: représentation des unités d'observations sur le premier axe factoriel	44
Figure 13 : arbre de la CAH.....	49
Figure 14 : ajustement de la fréquence des sinistres à la loi binomiale négative.....	53
Figure 15 : ajustement de la fréquence des sinistres à la loi poisson.....	53
Figure 16 : ajustement de la charge des sinistres à la loi Gamma.....	56
Figure 17 : ajustement de la charge des sinistres à la loi log-normale.....	57
Figure 18 : GLM avec la loi binomiale négative.....	63
Figure 19 : test de significativité de Wald pour le GLM avec la loi Binomiale Négative.....	63
Figure 20 : Evaluation de la qualité d'ajustement du GLM pour la Binomiale Négative.....	64
Figure 21 : prob-plot des résidus pour la loi normale dans le cas d GLM sur la fréquence....	65
Figure 22 : GLM avec la Normale.....	66
Figure 23 : test de significativité de Wald pour le GLM avec la loi Normale.....	67
Figure 24 : test de significativité de Wald pour le GLM avec la loi Normale après regroupement.....	67
Figure 25 : prob-plot des résidus pour la loi normale dans le cas d GLM sur le coût moyen..	68
Figure 26 : S/P annuels crédibilisés avec le modèle de Bühlmann Straub.....	75
Figure 27 : aperçu des S/P contrats par contrats crédibilisés avec le modèle de Bühlmann Straub.....	75
Figure 28 : la variation de la crédibilité des flottes en fonction de leur prime.....	76
Figure 29 : la variation de l'erreur de classification en fonction du nombre de subdivision...	81
Figure 30 : la subdivision du portefeuille en classes de cotisations.....	82
Figure 31 : Les S/P par catégorie crédibilisés avec le modèle de Jewell à un niveau.....	83
Figure 32 : les S/P crédibilisés de la catégorie dans la classe de cotisation avec Jewell à 2 niveaux.....	83
Figure 33 : les S/P crédibilisés de la classe au sein du portefeuille par Jewell à 2 niveaux....	84
Figure 34 : aperçu des S/P crédibilisés contrat par contrat pour Jewell à 2 niveaux.....	85

Sommaire

Introduction	11
---------------------------	----

Partie Préliminaire : Le marché de l'assurance automobile12

I. Marché de l'assurance Automobile au Maroc.....	13
1. <i>Vision Globale du marché</i>	13
2. <i>AXA Assurance Maroc</i>	15
II. Types de contrats automobiles à AXA Assurance Maroc :.....	16
1. Contrats mono véhicules.....	16
2. Contrats flottes.....	16
III. Types de garanties automobiles présentés par AXA.....	17

Section I : Tarification de la garantie Responsabilité civile par le GLM.....19

Chapitre 1 : Préliminaire à la modélisation du risque automobile.....20

I. Principes de tarifications.....	20
1. Principe du Bonus-malus.....	20
2. Segmentation du portefeuille.....	21
II. Etude descriptive.....	22
1. Traitement et épurement des bases de données.....	22
2. Analyse descriptive des variables.....	25
III. Analyse de données.....	34
1. Analyse en composantes principale ACP.....	35
2. Les techniques de classification automatique.....	44

Chapitre 2: Le modèle linéaire généralisé.....50

I. Analyse des corrélations entre les variables explicatives.....	50
1. Les variables quantitatives.....	50
2. Les variables qualitatives.....	51
II. Etude des lois des variables dépendantes et tests d'adéquation.....	53
1. Les lois de la fréquence.....	53
2. Les lois du coût moyen.....	54

III.	Le modèle linéaire généralisé.....	58
A.	Aspect théorique.....	58
1.	Logique du modèle.....	58
2.	Estimation des paramètres.....	59
B.	Applications aux données.....	62
•	GLM sur la fréquence.....	62
1.	Qualité d'ajustement du modèle	
2.	Comparaison des déviations	
3.	Test de significativité de Wald	
4.	Qualité d'ajustement	
5.	Résidus	
•	GLM sur le coût moyen des sinistres.....	65
1.	Qualité d'ajustement du modèle	
2.	Comparaison des déviations	
3.	Test de significativité de Wald	
4.	Qualité d'ajustement	
5.	Résidus	
	Conclusion de la première section.....	69

Section II : La crédibilité sur les flottes automobiles d'AAM.....70

I.	Les modèles de Bühlmann Straub.....	72
1.	La description du modèle.....	72
2.	L'application aux données.....	74
II.	Le modèle hiérarchique.....	76
1.	La description du modèle.....	77
2.	L'application aux données.....	80
	Conclusion.....	87
	Bibliographie.....	89
	Annexes.....	90

Introduction

Le marché des assurances est en évolution perpétuelle. Il fait objet, désormais, d'une forte mondialisation qui engendre une libéralisation de ce secteur et oblige ainsi toutes les compagnies d'assurance à innover constamment dans le but de garantir la confiance d'une large clientèle et d'acquérir une forte position sur le marché dans lequel elles opèrent.

C'est dans cette perspective qu'AXA Assurance Maroc tient à assurer les meilleures indemnisations pour ses clients, et à évaluer son portefeuille et reconnaître le profil d'assurés qu'elle a intérêt à prendre en charge.

Afin de mener à bien sa mission, AAM doit accorder une importance particulière à la tarification de ses produits. Néanmoins, le calcul du tarif appelé prime nécessite la construction de classes homogènes. Pour ce, l'assureur doit tenir compte de tous les paramètres qui influencent le risque et mesurent son impact.

Ceci dit, en pratique, l'assureur ne parvient pas à construire des classes parfaitement homogènes. Ainsi des corrections à posteriori s'avèrent nécessaires, en tenant cette fois-ci compte de la sinistralité de l'assuré.

Cette correction est envisageable via deux approches : le mécanisme « Bonus Malus » dont la logique est de favoriser les conducteurs prudents et responsables. La deuxième approche est l'exploitation de la théorie de crédibilité et c'est la méthode suivie dans la deuxième partie de notre projet.

En effet, notre travail ventile deux grandes sections principales :

La première porte sur les contrats mono véhicules et ayant pour objectif majeur d'établir des tarifs adéquats pour ce type de contrats. Pour ce faire on débutera l'étude par une analyse descriptive dont l'objectif est d'éclaircir les spécificités du portefeuille objet de l'étude. Nous allons enchaîner par une analyse de données dans le but d'établir des conclusions sur les différentes liaisons entre les variables. Cette dernière sera, par la suite, établie via l'application du modèle linéaire généralisé et l'analyse de sa pertinence et sa robustesse.

La deuxième section portera sur la théorie de crédibilité appliquée au portefeuille des flottes automobiles. Nous débuterons par la présentation des différents modèles de la théorie de crédibilité puis nous appliquerons les différents modèles sur le portefeuille d'AAM et analyserons les résultats obtenus.

La réalisation du présent mémoire nécessite l'utilisation des outils informatiques suivants: les logiciels SAS et R et le tableur EXCEL.

*Partie
Préliminaire*

**Le marché de l'assurance
automobile**

- Présentation du marché de l'assurance automobile au Maroc
- Les différents contrats d'AXA Assurance Maroc
- Les types de Garanties AAM

I. Marché de l'assurance Automobile au Maroc

1. Vision Globale du marché :

Le secteur d'assurance au Maroc est en pleine Mutation, présentant un potentiel de développement important à la fois en termes de volumes qu'en termes d'offres. La branche automobile représente le pilier le plus important dudit secteur, comme en témoigne le rapport annuel de la Fédération marocaine des sociétés d'assurances et de réassurances (FMSAR) au titre de l'exercice 2011.

Tableau1 : structure du chiffre d'affaire du secteur d'assurance au Maroc

Structure du Chiffre d'Affaires

	Chiffre d'Affaires	Part Marché	Evolution 2010/2011
Assurances Vie et Capitalisation	7 650,6	32,0%	15,9%
Automobile	7 531,3	31,5%	6,4%
Accidents Corporels	2 799,6	11,7%	2,7%
Accidents du Travail	1 957,3	8,2%	3,3%
Incendie	1 062,7	4,4%	2,9%
Assistance - Crédit - Caution	763,7	3,2%	3,4%
Transport	730,0	3,1%	9,2%
Responsabilité Civile Générale	490,5	2,1%	7,0%
Autres Opérations Non Vie	405,6	1,7%	11,5%
Risques Techniques	339,5	1,4%	113,7%
Acceptations en réassurance	163,2	0,7%	-23,0%
Total	23 893,9	100%	9,2%

En millions de dirhams

Source : rapport annuel de l'exercice 2011 de la FMSAR.

Tableau 2 : évolution du chiffre d'affaire des différentes branches du secteur d'assurance

Evolution du Chiffre d'Affaires

	2009	2010	2011	Evolution 2010/2011	Evolution 2009/2010
Assurances Vie & Capitalisation	6 718,8	6 659,5	7 717,0	15,9%	-0,9%
Assurances Individuelles	4 496,5	4 303,2	4 626,1	7,5%	-4,3%
Assurances de Groupes	1 323,3	1 532,0	1 727,2	12,7%	15,8%
Assurances Populaires	-	-	0,2	N.S	N.S
Capitalisation	480,8	468,0	967,0	106,6%	-3%
Contrats à Capital Variable	342,8	287,7	330,1	14,7%	-16%
Acceptations Vie	75,5	68,6	66,4	-3,2%	-9,2%
Assurances Non Vie	14 220,8	15 213,3	16 176,9	6,3%	7,0%
Accidents Corporels	2 623,7	2 726,8	2 799,6	2,7%	3,9%
Accidents du Travail	1 831,0	1 804,3	1 957,3	3,3%	3,5%
Automobile	6 587,7	7 075,8	7 531,3	6,4%	7,4%
Responsabilité Civile Générale	424,3	458,2	490,5	7,0%	8,0%
Incendie	941,9	1 032,6	1 062,7	2,9%	9,6%
Risques Techniques	264,6	304,4	339,5	11,5%	15,1%
Transport	691,7	706,3	730,0	3,4%	2,1%
Autres Opérations Non Vie	164,2	189,8	405,6	113,7%	15,6%
Assistance - Crédit - Caution	635,9	699,3	763,7	9,2%	10,0%
Acceptations Non Vie	56,0	125,8	96,8	-23,0%	124,5%
Total	20 939,6	21 872,8	23 893,9	9,2%	4,5%

Fédération Marocaine des Sociétés d'Assurances et de Réassurance - Mars 2012

Source D.03

Commentaire :

On constate que la part de la branche automobile est de 31,5 % et continue à avoir une part importante dans le chiffre d'affaire global du marché. Cette part de la branche automobile était de 32,3 % en 2010 et 31,45 % en 2009.

Bien que le tarif RC ait été libéralisé depuis 2006, le niveau du tarif de la RC n'a pas subi une réelle modification et la concurrence reste limitée au niveau du tarif des garanties annexes. Il s'agit d'un marché en développement, dont le potentiel reste important au vu de la croissance continue du parc et le faible taux d'équipement des ménages en termes de véhicules.

L'importance de cette branche d'automobile s'explique en grande partie par l'obligation d'assurance responsabilité civile, mais également par la volonté des assurés prudents considérés comme de bons risques de se couvrir au mieux contre ce risque quotidien, tout en voyant leurs primes diminuer en récompense de leur bon comportement.

En termes de rentabilité, et depuis 15 ans, cette branche présente un niveau de rentabilité satisfaisant participant largement à l'équilibre technique global de l'entreprise.

Ce niveau de rentabilité en croissante amélioration, permet aux principaux acteurs du marché de mener une différenciation par les prix, à travers une politique tarifaire de plus en plus agressive.

Ainsi, la branche automobile constitue un axe de développement stratégique pour toutes les compagnies d'assurances et leur réseau d'agents généraux.

Le tableau suivant présente l'évolution du chiffre d'affaire du secteur d'assurance sur la période 2009-2011.

2. AXA ASSURANCE MAROC

Chiffre d'affaire Non Vie par entreprise d'assurance :

Tableau3 : position d'AXA assurance Maroc en termes du chiffre d'affaire Non vie

Assurances Non Vie (y compris les acceptations en réassurance)					
	2009	2010	2011	Evolution 2010/2011	Part marché
Atlanta	1 065,1	1 101,8	1 105,0	0,3%	6,8%
Axa Assistance Maroc	23,9	29,9	36,3	21,4%	0,2%
Axa Assurance Maroc	2 477,8	2 381,3	2 546,3	6,9%	15,7%
CAT	628,0	662,3	662,2	0,0%	4,1%
Cnia Saada Assurance	2 318,1	2 464,0	2 567,0	4,2%	15,9%
Euler Hermes ACMAR	48,5	55,4	82,3	48,5%	0,5%
ISAAF Assistance	283,5	297,5	308,7	3,8%	1,9%
MAMDA	329,0	405,8	554,9	36,8%	3,4%
Maroc Assistance Internationale	275,4	313,7	326,0	3,9%	2,0%
Marocaine Vie	44,7	47,7	52,2	9,4%	0,3%
MATU	203,8	215,8	215,2	-0,3%	1,3%
MCMA	354,3	365,0	394,2	8,0%	2,4%
RMA Watanya	2 514,7	2 599,3	2 637,7	1,5%	16,3%
Sanad	1 142,3	1 148,8	1 207,9	5,1%	7,5%
Wafa Assurance	1 716,8	2 237,5	2 451,6	9,6%	15,2%
Wafa Ima Assistance	-	-	6,6	N.S	N.S
Zurich	794,9	887,5	1 022,9	15,3%	6,3%
Total	14 220,8	15 213,3	16 176,9	6,3%	100,0%

En millions de dirhams

Fédération Marocaine des Sociétés d'Assurances et de Réassurance - Mars 2012

Commentaire :

AXA Assurance Maroc occupe la troisième place dans le secteur des assurances non vie avec une part de marché d'ordre de 15,7 % Soit un chiffre d'affaire de 2,54 milliards de dirhams avec une évolution par rapport à 2010.

II. Types de contrats automobiles à AXA Assurance Maroc :

En assurance automobile, on distingue deux types de contrats auto entreprise : les contrats mono-véhicules (ou mono) et les contrats flottes.

1. Contrats mono véhicules :

L'offre mono destinée aux particuliers représente environ 89 % du chiffre d'affaire Automobile. Comme leur nom l'indique, ces contrats couvrent un seul véhicule. On peut également faire appel à ce type de contrats dans le cas d'entreprises ayant peu de véhicules.

2. Contrats flottes :

L'offre flotte destinée aux entreprises représente environ 11 % du chiffre d'affaire Automobile avec plus de 1200 contrats.

Un contrat flottes permet de couvrir la totalité ou une partie du parc automobile d'une entreprise. Les contrats flottes peuvent être scindés en plusieurs familles selon divers critères (mode de gestion, poids, usage..), à savoir:

a. Subdivision par mode de gestion :

Selon le mode de gestion, les flottes peuvent être scindées en flottes fermées (ou à véhicules dénommés) et ouvertes.

→ Les flottes fermées: On connaît le nombre de véhicules de la flotte ainsi que leurs caractéristiques.

→ Les flottes ouvertes: L'assureur ne connaît de façon précise ni le volume ni la composition de la flotte.

b. Subdivision par poids :

Il est d'usage de distinguer les véhicules de plus de 3.5 tonnes (camions, bennes...) des autres. Une flotte sera dite plus de 3.5 tonnes (ou P3T5) si la majorité de ses cotisations vient des véhicules P3T5 et moins de 3.5 tonnes (M3T5) sinon. La distinction entre M3T5 et P3T5 vient d'une différence au niveau des coûts de sinistres : les P3T5 ont des sinistres généralement plus élevés.

c. Subdivision par usage :

Les flottes diffèrent par leur usage. La classification AXA comporte plusieurs dizaines de catégories. Les trois subdivisions essentielles sont l'usage tourisme, l'usage commercial inférieur à 3.5 T et l'usage commercial supérieur à 3.5T.

III. Types de garanties automobiles présentés par AXA :

Les garanties de la branche Automobile sont diverses. Leur multitude émane de la diversité des risques couverts, de la couverture maximale que chaque compagnie détermine pour chaque produit et également des franchises que les entreprises d'assurance adoptent dans leurs normes de tarification. Nous résumons ci-dessous les différentes garanties Auto assurées par AXA Assurance Maroc en termes de risque couvert, de couverture maximale et de franchises adoptées par la compagnie :

Tableau 4 : types de garanties offertes par AXA Assurance Maroc

Risque par sinistre	Couverture maximale	Franchises
Votre responsabilité et votre défense		
Responsabilité civile		
- Dommages corporels	50 000 000 Dh	
- Dommages matériels	50 000 000 Dh	
Défense et recours	Capital choisi figurant aux Conditions particulières	
Les dommages causés au véhicule		
Bris de Glaces (Article 6)	Valeur de remplacement ⁽¹⁾ dans la limite du capital assuré choisi	Taux figurant aux conditions particulières
⁽¹⁾ Y compris frais de dépose et pose		
Incendie (Article 7)		
- Véhicule et accessoires livrés par le constructeur	Valeur d'achat ou valeur vénale	
- Auto radio & remorque	Capital assuré déclaré	
- Aménagements professionnels	Capital assuré déclaré (valeur d'achat et frais de pose)	
Vol (Article 8)		
- Véhicule et accessoires livrés par le constructeur	Valeur d'achat ou valeur vénale	Taux figurant aux conditions particulières
- Auto radio & remorque	Capital assuré déclaré	
- Aménagements professionnels	Capital assuré déclaré (valeur d'achat et frais de pose)	
Evénements climatiques et naturels (Article 9)		
- Véhicule et accessoires livrés par le constructeur	Valeur d'achat ou valeur vénale	Taux figurant aux conditions particulières
- Auto radio & remorque	Capital assuré déclaré	
- Aménagements professionnels	Capital assuré déclaré (valeur d'achat et frais de pose)	
Dommage collision plafonnée (Article 10)		
- Véhicule et accessoires livrés par le constructeur	Capital choisi déclaré	Taux figurant aux conditions particulières
- Auto radio & remorque	Capital assuré déclaré	
- Aménagements professionnels	Capital assuré déclaré (valeur d'achat et frais de pose)	

Risque par sinistre	Couverture maximale	Franchises
Les dommages causés au véhicule		
Dommege collision <u>déplafonnée</u> (Article 10) - Véhicule et accessoires livrés par le constructeur Valeur d'achat ou valeur vénale - Auto radio & remorque Capital assuré déclaré - Aménagements professionnels Capital assuré déclaré (valeur d'achat et frais de pose)		Taux figurant aux conditions particulières
Dommege tous accidents (Article 11) - Véhicule et accessoires livrés par le constructeur Valeur d'achat ou valeur vénale - Auto radio & remorque Capital assuré déclaré - Aménagements professionnels Capital assuré déclaré (valeur d'achat et frais de pose)		Taux figurant aux conditions particulières
Dommege tous accidents éco plus (Article 12) - Véhicule et accessoires livrés par le constructeur Capital assuré déclaré - Auto radio & remorque Capital assuré déclaré - Aménagements professionnels Capital assuré déclaré (valeur d'achat et frais de pose)		Taux figurant aux conditions particulières
Dommege tous risques (Article 13) - Véhicule et accessoires livrés par le constructeur Valeur d'achat ou valeur vénale - Auto radio & remorque Capital assuré déclaré - Aménagements professionnels Capital assuré déclaré (valeur d'achat et frais de pose)		Taux figurant aux conditions particulières
Perte totale (Article 14) - Véhicule et accessoires livrés par le constructeur	Valeur d'achat ou valeur vénale calculée selon le barème figurant aux conditions particulières	
Rachat de vétusté (Article 15)		
Les dommages causés aux personnes		
Protection familiale, conducteur et passagers (Article 20) - Capital Décès } - Capital invalidité permanente } - Frais médicaux } Capitaux choisis figurant aux Conditions particulières		
Individuelle accidents conducteur habituel (Article 21) - Capital Décès } - Capital invalidité permanente } - Frais médicaux } - Indemnité journalière d'hospitalisation } Capitaux choisis figurant aux Conditions particulières		

Remarque :

Un même contrat peut faire l'objet de plusieurs garanties à la fois.

Tarification de la garantie Responsabilité civile par le modèle linéaire généralisé

Section I

- Analyse descriptive uni-variée et multi-variée
- Application GLM

Chapitre 1 : Préliminaire à la modélisation du risque automobile

Dans ce chapitre nous traitons en premier lieu les principes et la nécessité de tarification et le principe de bonus malus. Ensuite, nous procédons par une analyse descriptive uni-variée sur le portefeuille RC automobile après avoir effectué un traitement d'épure sur les tables de données. En dernier lieu, nous effectuons une analyse de données plus précisément l'ACP et la CAH.

I. Principes de tarifications :

1. Principe du Bonus-malus :

L'assurance automobile joue un rôle important en lien avec la sécurité routière et l'incitation à la prudence et la prévention. Dans ce sens le règlement des assurances au Maroc a introduit en 2005 la notion du coefficient de réduction majoration (CRM) qui a pour objectif de récompenser les conducteurs responsables et de pénaliser les mauvais conducteurs. Ce coefficient CRM est déterminé par la Direction des Assurances et de la Prévoyance Sociale (DAPS).

La réglementation concernant la détermination du CRM est donnée par l'Article 19 : Réduction ou majoration de la prime de l'Arrêté du ministre des finances et de la privatisation n° 1053-06_du_28 rabii II 1427 :

« ...Pour la détermination de la prime, l'assureur doit tenir compte des antécédents de sinistralité de l'assuré en multipliant la prime de base, calculée indépendamment de ces antécédents, par un coefficient de réduction – majoration fixé comme suit :

- 0,9, si l'assuré n'a causé aucun sinistre engageant ou susceptible d'engager totalement ou partiellement sa responsabilité durant une période d'assurance de vingt quatre (24) mois consécutifs précédant la souscription ou le renouvellement du contrat. Pour la détermination de la période d'assurance de vingt quatre (24) mois consécutifs susvisée, il est toléré une seule interruption d'assurance ne dépassant pas trente (30) jours.

- Si l'assuré a causé un ou plusieurs sinistres engageant ou susceptible d'engager totalement ou partiellement sa responsabilité durant la période d'assurance de douze (12) mois précédant la souscription ou le renouvellement du contrat, ce coefficient, qui ne peut excéder 2,5, s'obtient en ajoutant à un (1) pour chacun de ces sinistres :

. 0,15 pour l'usage transport public de voyageurs (TPV) ou 0,20 pour les autres usages si le sinistre est matériel ;

. 0,20 pour l'usage TPV ou 0,30 pour les autres usages si le sinistre est corporel, ou matériel et corporel à la fois.

- Dans les autres cas le coefficient de réduction – majoration est égal à un (1).

Lorsque l'assuré est garanti pour plusieurs véhicules, le coefficient de réduction – majoration est déterminé et appliqué séparément véhicule par véhicule.

Dans le cas où l'assuré apporte la preuve que sa responsabilité est entièrement et définitivement dégagee, l'assureur est tenu de restituer la portion de prime correspondant à la différence entre la prime perçue et celle qu'aurait payé l'assuré en étant non responsable du sinistre considéré. »

2. Segmentation du portefeuille :

a. Nécessité de segmentation

La segmentation désigne le fait de découper un portefeuille en plusieurs sous-ensembles homogènes et distincts composés d'individus ayant des comportements communs.

Au-delà du mécanisme de bonus malus, chaque assureur possède une tarification et une segmentation qui lui sont propres, en adéquation avec le profil de son portefeuille d'assurés. Dans un contexte de marché très concurrentiel, la segmentation des risques est une nécessité. Celle-ci consiste à différencier les assurés et le risque qu'ils portent. On obtient ainsi différentes catégories de risques en fonction des caractéristiques de l'assuré et des garanties consenties par l'assureur. Chaque catégorie se verra ainsi attribuer un tarif qui lui sera propre, en adéquation avec le risque associé.

En effet, la segmentation consiste à analyser et contrôler l'adaptation des primes aux sinistres suivant des classes de risque homogène, de façon à en tirer des conséquences du point de vue technique. La segmentation permettra de prendre des mesures techniques à chacun des niveaux de segmentation, tant en tarification, qu'en souscription.

b. Choix des critères de segmentation

On cherche à définir des classes de risque homogènes c.à.d. ayant le même coût du risque (prime pure). Il existe en effet deux grandes classes de variables :

Les variables exogènes : les informations relatives au risque à l'exclusion de toute donnée relative aux réalisations du risque par exemple des critères liés au véhicule, au conducteur...

Les variables endogènes : les informations relatives aux réalisations du risque (nombre de sinistre par garantie, le coût des sinistres...).

Le choix des variables de segmentation et de leurs modalités repose à la fois sur des objectifs commerciaux et sur des méthodes statistiques :

- Analyse descriptive : calcul des premiers moments (moyenne et variance), discrétisation des variables, modélisation de la distribution sur la base d'un histogramme...
- Analyse des données : analyse des composantes principales, classification hiérarchique ascendantes, examen des corrélations entre les variables...
- Le modèle linéaire généralisé (exemple de méthode de segmentation) : on part d'un modèle avec le maximum de variables et pas à pas on élimine celles qui sont les moins significatives, pour obtenir un modèle satisfaisant en ce qui concerne le nombre de variables, leur complémentarité et leur pouvoir explicatif.

Conclusion :

Afin d'aboutir à une segmentation, le passage par trois grandes étapes se voit primordial. La première sera une analyse descriptive, suivie d'une analyse des données et enfin l'application du modèle linéaire généralisé. Mais, avant toute analyse, il faut s'assurer tout d'abord de la qualité de la base de données et ensuite procéder à un processus d'actualisation et d'homogénéisation des données et ce qui fait l'objet de la partie suivante.

II. Etude descriptive :

1. Traitement et épurement des bases de données :

Avant la mise en vigueur des résultats des statistiques descriptives, un processus d'épurement et de traitements des bases de données, objet de l'étude, s'avère nécessaire. Pour ce faire, nous présentons en premier lieu les bases de données brutes, ensuite nous décrivons les opérations de jointures entre les différentes tables, enfin nous précisons l'ensemble des failles détectées et les solutions adoptées pour les résoudre.

a. Présentation des données brutes :

Cette partie du projet porte sur les contrats MONO de l'automobile, la base de données fournie couvre les trois exercices 2009, 2010 et 2011.

Le nombre d'observations sur ces trois exercices s'élève à un million d'enregistrements et les informations sont fournies sous forme de deux fichiers :

- Un fichier production
- Un fichier sinistre

→Fichier Production :

Ce fichier contient les informations relatives aux contrats des véhicules mono assurés par AXA assurance Maroc. Ces informations sont diverses et peuvent être ventilées ainsi :

Les identificateurs : numéro de la police, matricule du véhicule ...

Les caractéristiques du véhicule : usage, carburant, zone de circulation....

Types de garanties : Responsabilité civile, incendie, bris de glace....

Les primes acquises : donne l'information sur le montant des primes collectées auprès des assurés.

→Fichier sinistre :

Le fichier sinistre contient les informations liées à la sinistralité des contrats. On distingue :

Les identificateurs : immatriculation.

Les informations liées aux sinistres : date de survenance, nature du sinistre (matériel ou corporel), montant du sinistre, responsabilité de l'assuré (responsable, non responsable, partagée).

b. Jointure des deux fichiers :

La base de données finale qui fera l'objet de notre étude doit combiner les deux fichiers production et sinistre. Ainsi chaque contrat sera identifié avec la totalité des variables qui lui sont relatives.

La première étape consiste à trier le fichier sinistre par exercice et ce à travers la variable date de survenance.

Remarque : chez AXA l'usage est de travailler avec des années glissantes, ainsi l'exercice 2009 par exemple correspond à la période qui s'étale entre le 01/03/2009 et le 28/02/2010.

Une fois le fichier trié par exercice, on réalise la jointure de chaque fichier sinistre avec le fichier production qui lui correspond. La jointure se fait par une clé commune entre les deux fichiers. Dans notre cas la clé est le numéro de police. Il est à noter qu'avant de faire la jointure il faut trier les deux tables à combiner par la clé choisie. La jointure est réalisée via le code sas suivant :

```
Data table ;  
Merge table (in=a) table2 (in=b);  
By clé de jointure ;  
If a ;  
Run ;
```

1. Problèmes et failles

Deux types de problèmes sont mis en évidence : les valeurs manquantes et les valeurs aberrantes. Les valeurs manquantes sont des informations non déclarées par l'assuré ou non saisies par les gestionnaires de sinistres. Quant aux valeurs aberrantes, elles sont le résultat d'une erreur de saisie (exemple : âge du permis 278 ans). Le tableau ci-dessous résume le nombre de valeurs manquantes sur chaque variable et le pourcentage de ces valeurs par rapport à la totalité de la base de données.

Tableau 5 : statistiques sur les valeurs manquantes et aberrantes de la base des données.

Variable	nbr_valeurs manquantes	pourcentage	Valeurs aberrantes	pourcentage
garantie	0	0%	0	0
Prime acquise	0	0%	0	0
usage	2021	0%	0	0
carburant	1191	0%	0	0
sexe	13163	1%	9	0%
région	0	1%	0	0%
Age du véhicule	17552	0%	11398	1%
Age du conducteur	12802	1%	16510	1%
Age du permis	10361	1%	10234	1%

Deux solutions sont envisageables :

- Remplacer les valeurs manquantes et aberrantes par la moyenne s'il s'agit d'une variable quantitative.
- Mettre les observations en questions sur le compte de la classe la plus risquée. Ce choix relève du constat que généralement les assurés qui ne déclarent pas une information font preuve d'une sinistralité importante.

A ce stade, la base de données est exploitable, on peut donc entamer l'étude.

2. Analyse descriptive :

a. Les variables tarifaires :

En assurance automobile, le tarif peut dépendre de plusieurs paramètres tels les informations sur le conducteur (âge, sexe ou âge du permis) ou sur le véhicule (usage du véhicule, type de combustion ou zone de circulation) ...

Les variables tarifaires choisies pour le présent projet sont les suivantes :

- La région de circulation
- L'usage du véhicule
- Le type de combustion
- Le sexe du conducteur
- L'âge du conducteur
- L'âge du permis de conduire
- L'âge du véhicule

Avant de procéder à des statistiques descriptives sur l'ensemble de ces variables, il est convenable de regrouper les modalités de chacune sous forme de classes homogènes vis-à-vis du risque.

b. Regroupement des variables :

Par souci de lisibilité et tenant compte de la multitude des graphiques, nous avons choisi de présenter les résultats pour deux variables tarifaires (la zone de circulation et l'usage du véhicule). Toutefois, le lecteur trouvera en annexe la liste exhaustive des tableaux et des graphiques de toutes les autres variables.

i. Zone de circulation :

Les trois graphiques suivants mettent en évidence les regroupements des modalités de la variable zone qui peuvent avoir lieu et ce en visualisant ces regroupements en termes de fréquence, d'exposition et de progression du risque à travers le temps. Il est à noter que la variable exposition au risque est mesurée par le rayon des cercles sur le troisième graphique. Ainsi plus le rayon est grand plus l'exposition est importante.

Figure1 : Graphique représentant la répartition de la fréquence par région et par exercice

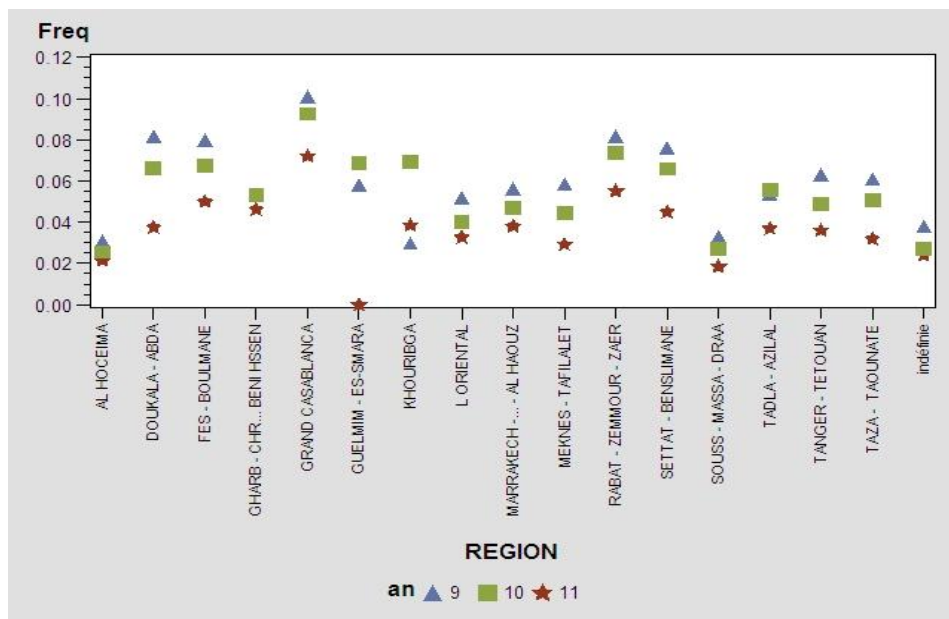


Figure 2 : Graphique représentant la répartition de la fréquence par région

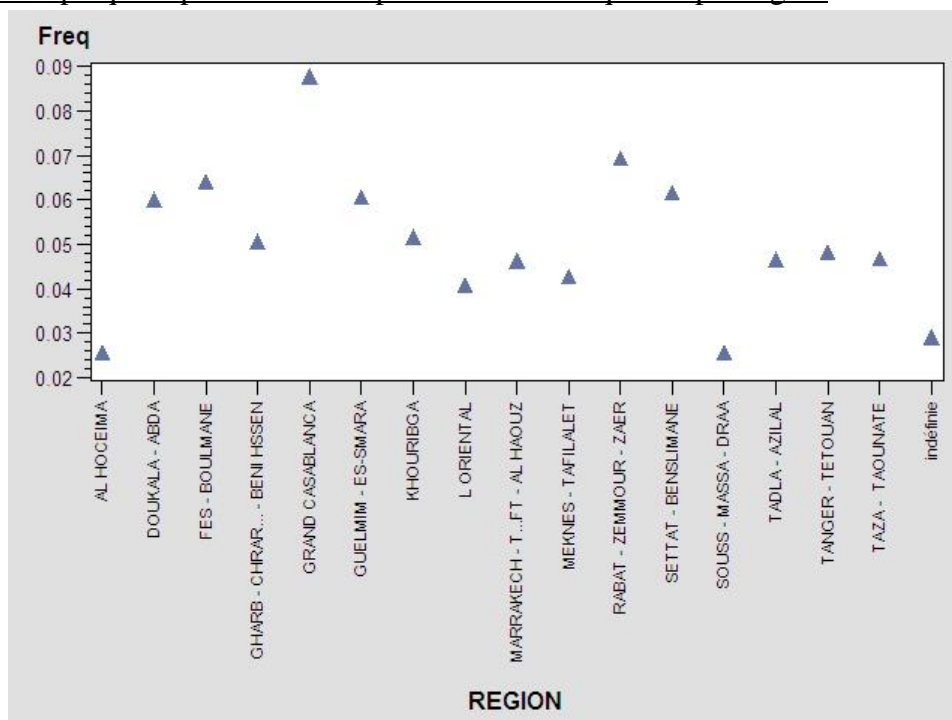
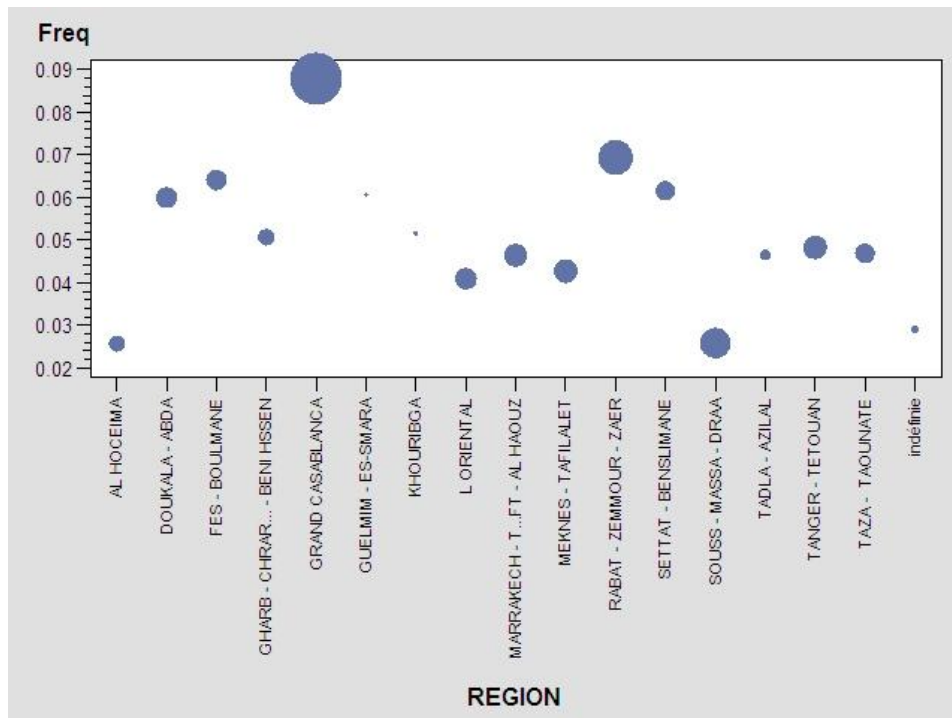


Figure 3 : Graphique représentant la répartition de la fréquence par région et par exposition au risque



Source : SAS

Commentaire et conclusion:

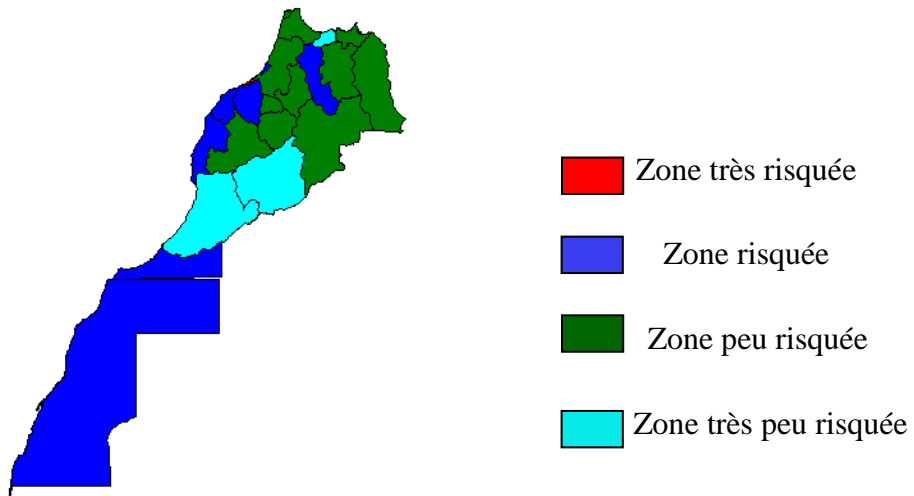
On remarque clairement que la région du Grand Casablanca a une très grande fréquence de sinistres (graphique 1) pour les trois années 2009,2010 et 2011 ainsi qu'une très grande exposition au risque (graphique 3). En deuxième position on retrouve les régions Rabat, Fès, Doukala, Gelmim et Settat. En 3ème position on trouve Marrakech, Tanger, Meknes, Taza, Gharb, Khouribga, Tadla et l'oriental représentant des zones à risque faible. Ainsi, nous pouvons donner le regroupement de la variable région :

Tableau 6 : codification des groupements de zones.

Zones	Code
Casablanca	1
Rabat + Fès + Doukala + Gelmim + Settat	2
Marrakech+ Meknes+Tanger+Taza+ gharb+ khouribga+ Tadla +l'oriental	3
El hoceima + Sous massa daraa	4

Ce regroupement dévoilé par le graphique peut être visualisé par une cartographie du risque évalué en termes de fréquence. La cartographie est faite sous SAS et se présente comme suit :

Figure 4 : cartographie du risque automobile au Maroc



ii. l'Usage du véhicule :

Figure 5 : Graphique représentant la répartition de la fréquence par usage et par année

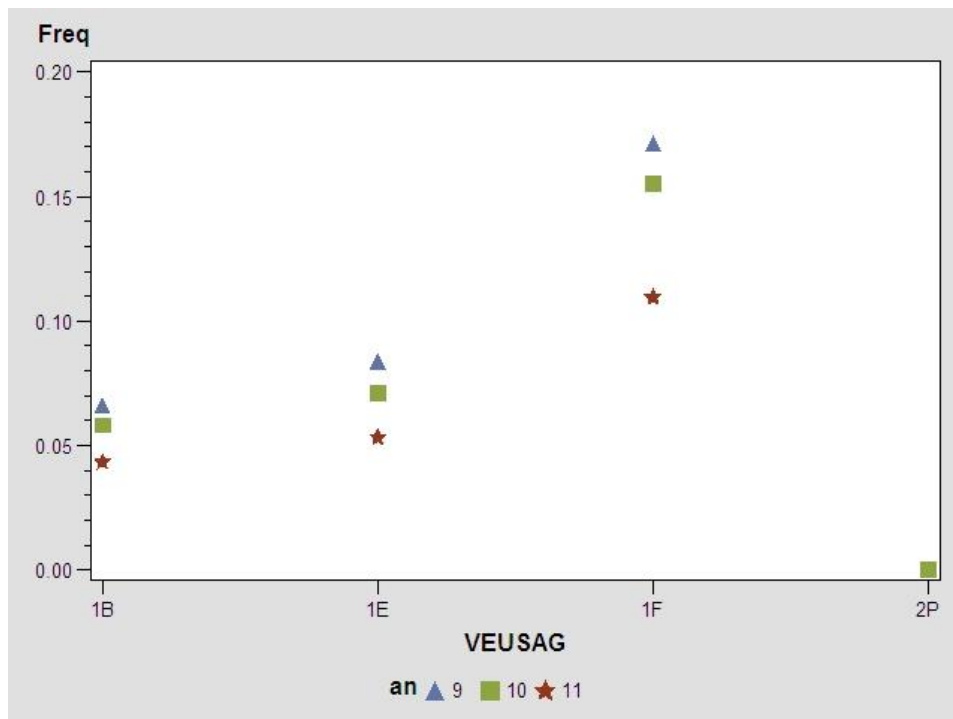
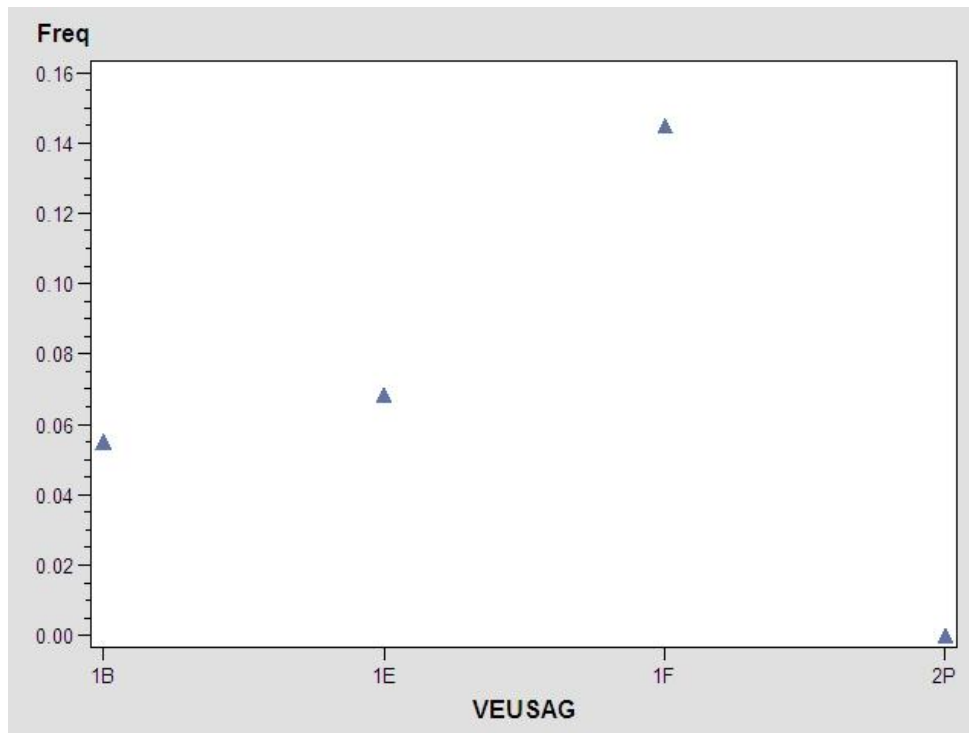
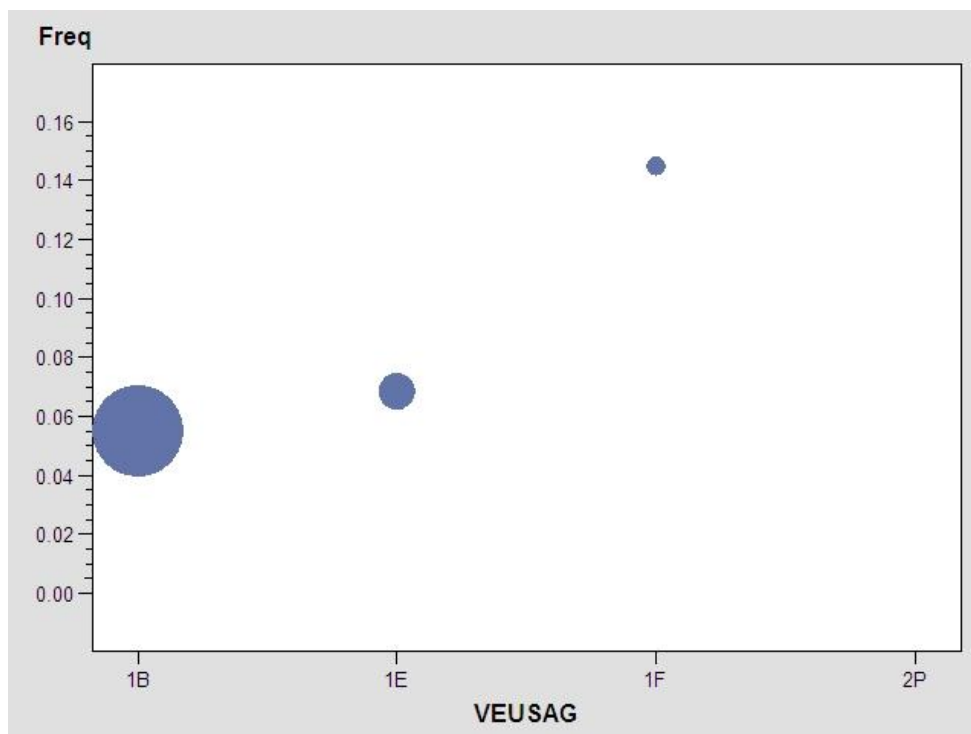


Figure 6 : Graphique représentant la répartition de la fréquence par région



Source : SAS

Figure 7: Graphique représentant la répartition de la fréquence par région et par exposition au risque



Source : SAS

Commentaire et conclusion :

D'après les graphiques, on peut détecter trois groupes essentiels (1B, 1 E et 1F) par rapport à l'exposition au risque et à la fréquence des sinistres. Ces trois usages ont connu une amélioration importante, en termes de fréquence des sinistres, au cours des trois années 2009, 2010 et 2011.

L'usage le plus risqué en terme de fréquence de sinistres est l'usage commercial supérieur à 3,5 T (1F) mais avec une exposition au risque très faible, suivi de l'usage commercial inférieur à 3,5 T (1 E) avec une exposition plus grande et une fréquence plus petite et enfin en troisième position l'usage tourisme (1B) avec une très grande exposition au risque et une fréquence plus ou moins faible.

Ainsi, nous pouvons donner le regroupement de la variable usage véhicule comme suit :

Tableau7 : codification des groupements de l'usage du véhicule

Groupe homogène	Code
Tourisme	1
Commercial <3.5T	2
Commercial >3.5T	3
Autres usages	4

iii. les autres variables

Comme nous l'avons cité avant, nous nous sommes contentées de détailler le processus pour deux variables seulement.

Le regroupement de la variable âge du véhicule :

Tableau 8 : codification des variables âge du véhicule, carburant, âge du conducteur, sexe

Groupe homogène	Code
Véhicule neuf (âge<2)	1
Véhicule normal (2<âge<9)	2
Véhicule ancien (âge>9)	3

La codification de la variable combustion :

Groupe	Code
Gasoil	1
Essence	2

Le regroupement de la variable région :

Classes d'âges	Groupe homogène	Code
18-34	Age à grand risque	1
>34	Age à risque faible	2

Le regroupement de la variable âge :

Classes d'âges	Groupe homogène	code
0-2	Conducteur débutants	1
>2	Conducteur expérimenté	2

La codification de la variable sexe :

Groupe	Code
Femme	1
Homme	2

a. statistiques générales sur le portefeuille :

Le tableau ci-dessous définit les quantités statistiques calculées sur l'ensemble des variables :

<i>quantité</i>	<i>Définition</i>
<i>Nombre de sinistre</i>	<i>Donnée</i>
<i>Charge totale</i>	<i>Donnée</i>
<i>Coût moyen fréquence</i>	<i>Charge totale/nombre de sinistres</i>
<i>Exposition au risque</i>	<i>Nombre de sinistres /l'exposition au risque</i>
<i>S/P</i>	<i>Nombre de jour objet d'une garantie/360</i>
	<i>Charge totale/prime</i>

Pour les mêmes raisons citées en dessus, nous nous contentons de donner les résultats pour deux variables seulement.

ii. [La variable zone](#)

Tableau 9 : statistiques sur le portefeuille des contrats Mono

La table Excel suivante résume les caractéristiques des zones de circulation par année :

date	code	exposition au risque	prime	nombre sinistres	charge	freq	cm	sp
2009	1	61331,811	1,7E+08	6202	1,12E+08	0,101122	18051,67	0,659297
2009	2	55724,5507	1,62E+08	4481	1,32E+08	0,080413	29451,73	0,814796
2009	3	28359,9945	84428245	929	30604485	0,032757	32943,47	0,362491
2009	4	64271,5836	2E+08	3699	1,25E+08	0,057553	33780,73	0,625576
2010	1	66629,8603	1,84E+08	6169	1,04E+08	0,092586	16895,98	0,566609
2010	2	61596,989	1,78E+08	4309	1,06E+08	0,069955	24708,97	0,599676
2010	3	31107,1945	92347417	836	21611310	0,026875	25850,85	0,234022

2010	4	73752,6767	2,28E+08	3493	1,08E+08	0,047361	30933,81	0,474332
2011	1	70595,7726	1,95E+08	5079	71435791	0,071945	14064,93	0,367044
2011	2	67862,1096	1,94E+08	3348	63192876	0,049335	18874,81	0,326166
2011	3	33028,9589	96495988	638	15409438	0,019316	24152,72	0,15969
2011	4	83041,663	2,51E+08	2911	71265847	0,035055	24481,57	0,28374

Le graphique suivant représente la fréquence des sinistres pour chaque zone de circulation et pour chaque année :

Figure 8 : fréquence des sinistres pour chaque groupement de zone et par année

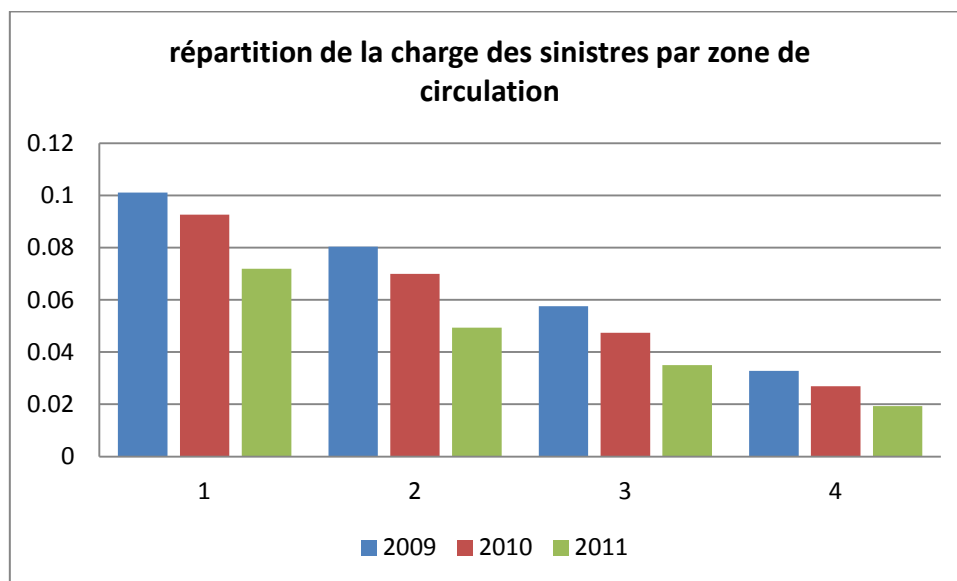
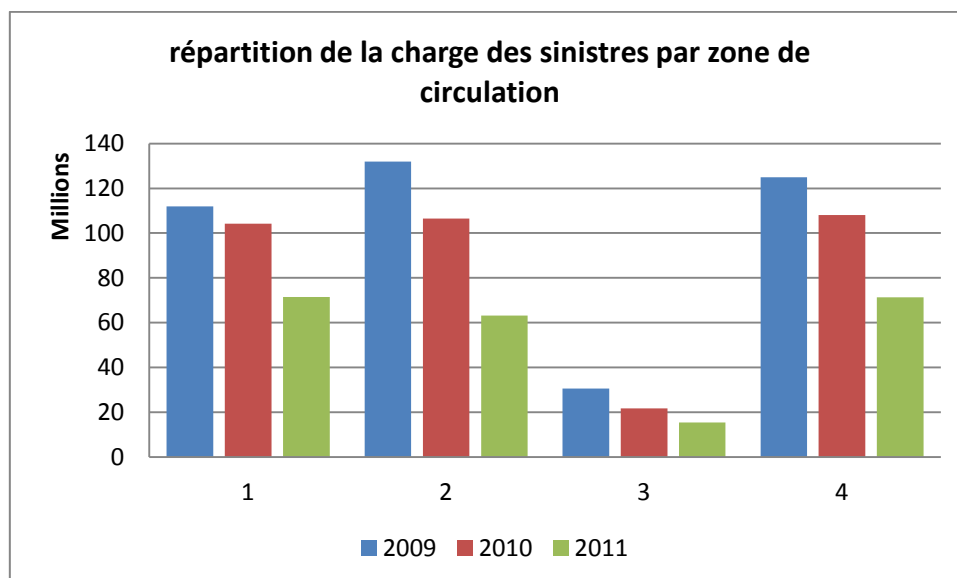


Figure 9 : Le graphique représentant la charge des sinistres pour chaque zone de circulation et pour chaque année :



Commentaire et conclusion :

Les répartitions finales de la fréquence et de la charge confirment le choix des groupements. Cependant, le groupe 1 qui garde toujours une grande fréquence par rapport aux autres groupes ne se comporte pas de la même manière vis-à-vis de la charge laquelle est importante pour les groupes 2 et 4.

iii. Usage du véhicule :

La table Excel suivante résume les caractéristiques des différents usages de véhicule par année :

Tableau 10 : statistiques sur le portefeuille mono par usage de véhicule

date	code	type usage	exposition au risque	prime	charge	freq	cm	sp
2009	1	tourisme	171888,386	4,49E+08	2,71E+08	0,066235	23764,15	0,602798
2009	2	com<3.5T	29185,0877	1,14E+08	78531697	0,083913	32066,84	0,686582
2009	3	com>3.5T	8613,46575	52738924	50402554	0,171476	34124,95	0,955699
2010	1	tourisme	192352,816	5,04E+08	2,45E+08	0,058096	21882,95	0,48559
2010	2	com<3.5T	31980,5644	1,25E+08	61625284	0,071043	27123,8	0,494696
2010	3	com>3.5T	8752,33973	53475905	34198160	0,155387	25145,71	0,639506
2011	1	tourisme	211081,074	5,48E+08	1,58E+08	0,043372	17246,46	0,288003
2011	2	com<3.5T	34373,0055	1,33E+08	42784392	0,053152	23417,84	0,322043
2011	3	com>3.5T	9074,42466	54949156	20628232	0,109539	20752,75	0,375406

Figure 10 : graphique représentant la fréquence des sinistres pour chaque usage et pour chaque année :

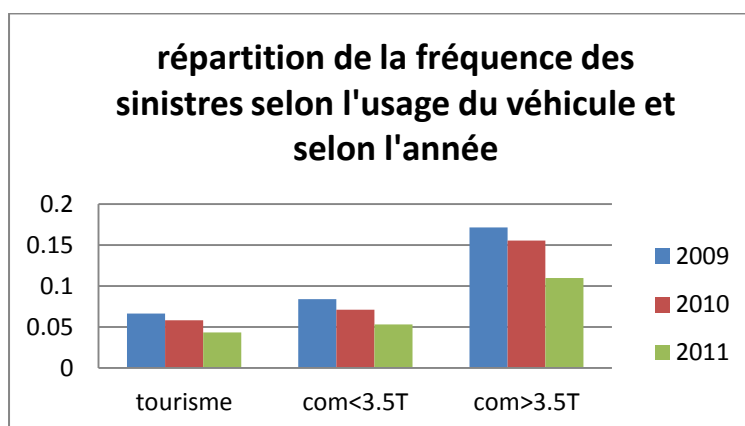
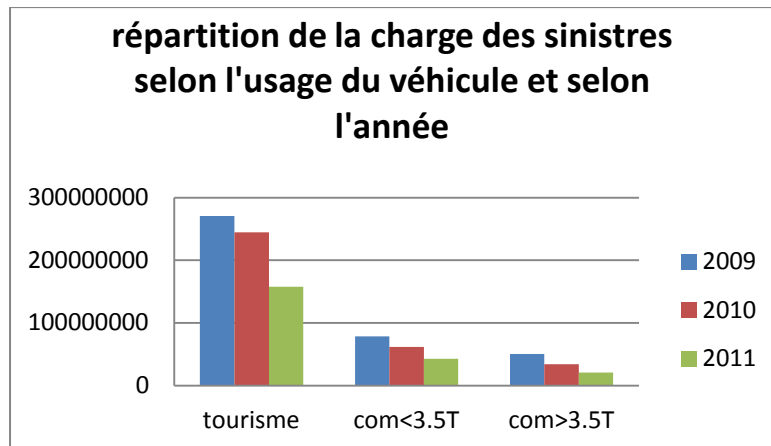


Figure 11 : Graphique suivant représente la charge des sinistres pour chaque usage et pour chaque année :



Commentaire et conclusion :

En ce qui concerne le premier graphique, les résultats sont les mêmes qu'avant : une amélioration au cours du temps en terme de fréquence pour les trois usages, et l'usage le plus risqué est le commercial supérieur à 3,5 T. Par contre en termes de charge, on a certes une amélioration au cours du temps mais l'usage le plus risqué en termes du montant total des sinistres est le tourisme qui peut être justifié par le nombre important de contrats inscrits en tourisme par rapport aux autres usages.

III. Analyse de données :

Introduction :

Tenant compte de la nature des données disponibles et les objectifs attendus du présent travail, nous allons utiliser les techniques de l'analyse des données, en particulier l'analyse en composantes principales (ACP) et la classification ascendante hiérarchique (CAH).

Ainsi, nous présenterons en premier lieu l'aspect théorique de chacune des méthodes suivi de l'application au portefeuille.

Il est à noter que la totalité des démonstrations des résultats et des propriétés énoncés dans partie aspect théorique est disponible sur le livre :

Saporta G. 2006, « probabilités, analyse des données et statistique », ED technip.

Consultable partiellement en recherche sur <http://books.google.fr>.

1. Analyse en composantes principales ACP :

A. Aspect théorique

L'objet de l'analyse en composantes principales est d'élaborer et de figurer géométriquement sur un plan les informations les plus diverses contenues dans un tableau $n \times p$ dont les lignes représentent n individus et les colonnes correspondent à p variables mesurées sur ces individus. Ce tableau peut être de très grande taille et les données étudiées peuvent être très hétérogènes.

Dans la littérature, on trouve deux approches différentes de l'ACP :

- Une approche anglo-saxonne qui présente l'ACP comme la recherche d'un ensemble réduit de variables non corrélées, combinaisons linéaires des variables initiales, résumant avec précision les données. cela revient donc à répondre aux questions suivantes :
 - Existe-t-il des groupes de variables fortement corrélées ?
 - Peut-on faire une typologie de variables ?
 - Peut-on résumer l'ensemble des variables par un nombre de variables synthétiques appelées composantes principales ?
- La deuxième approche repose sur la représentation des données initiales à l'aide de nuage de points dans un espace géométrique. L'objectif étant de trouver des sous espaces (droite, plan,...) qui représentent au mieux le nuage initial.

Avant de commencer l'étude, il convient de présenter un ensemble de notions et de définitions clés nécessaire pour la compréhension et l'interprétation des résultats de l'ACP. C'est l'objet de la partie suivante :

i. Préliminaires : notions et définition clés :

a. Données de L'ACP

L'analyse en composantes principales s'applique à des tableaux à deux dimensions, celui-ci croise des individus et des variables.

Les données mises en jeu en analyse en composantes principales sont relatives à des variables quantitatives, continues, homogènes ou non, à priori corrélées entre elles deux à deux.

On note $X = [X_{ij}] = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$ le tableau individus-variables.

Les données de l'ACP peuvent faire objet de quelques transformations. en effet, le meilleur axe de projection doit passer par le centre de gravité du nuage. Ceci revient à considérer le tableau X centré.

b. Recherche des axes factoriels :

Considérons un nuage pondéré $N = \{(X_1, m_1), \dots, (X_n, m_n)\}$ avec :

- X_i : La $i^{\text{ème}}$ observation.
- m_i : La masse de l'observation X_i .

La détermination d'un axe factoriel consiste à chercher l'axe $F_1 = \langle u_1 \rangle$ qui s'ajuste le mieux au nuage de point en question. On peut montrer que cela revient à chercher l'axe qui maximise la quantité $\sum_{i=1}^n d_i^2$.

d_i^2 étant la distance qui sépare la $i^{\text{ème}}$ observation de sa projection orthogonale sur l'axe F_1 .

Résultat :

Le premier axe factoriel correspond à l'axe engendré par le vecteur propre qui correspond à la plus grande valeur propre de la matrice $X'X$.

Le résultat est généralisé. Ainsi, le $k^{\text{ème}}$ axe factoriel n'est autre que l'axe engendré par la $k^{\text{ème}}$ plus grande valeur propre de la matrice $X'X$.

c. notion d'inertie :

Considérons un nuage pondéré $N = \{(X_1, m_1), \dots, (X_n, m_n)\}$ avec :

On définit l'inertie totale comme suit : $I(N) = \sum_{i=1}^n m_i * d_i^2$

d_i^2 étant la distance séparant l'observation X_i du centre de gravité du nuage de point.

Remarque : en ACP on prend $m_i = 1$

L'inertie reflète l'information disponible, on cherche donc à la maximiser.

L'inertie expliquée par un axe

Soit $F = \langle u \rangle$ un axe de vecteur directeur unitaire \vec{u} , l'inertie expliquée par l'axe F est donnée par :

$$I(N/F) = \alpha$$

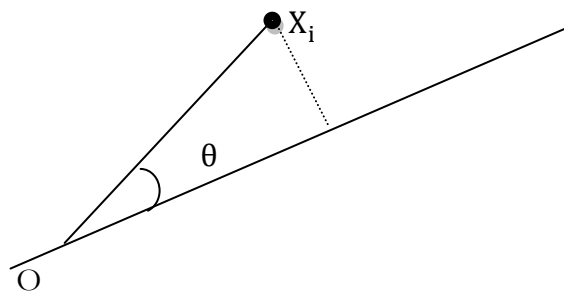
Avec : α la valeur propre correspondant au vecteur u .

Ainsi l'inertie totale est considérée comme la somme des inerties expliquées par tous les axes.

Sur un sous espace de dimension p l'inertie totale s'écrit alors : $I(N) = \sum_{i=1}^p \alpha_i$

d. Qualité de représentation :

Lors de l'analyse des données, il est indispensable de s'assurer que la représentation des points sur le plan factoriel est de bonne qualité. Cette dernière est mesurée par le *cosinus carré* de l'angle formé par le plan factoriel et le vecteur défini par le point.



On écrit alors : $QLT_{\alpha(i)} = \cos^2 \theta$

Remarque

- Graphiquement, une variable est jugée mal représentée si sa représentation s'avère proche du centre du repère.
- La qualité de représentation de tout le nuage des points est donnée par la somme des qualités de représentation sur l'ensemble des axes.

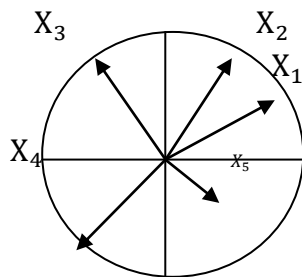
e. Corrélation entre les variables

Soit deux variables X_j, X_k , la traduction mathématique de la corrélation entre les deux variables est donnée par :

$$\rho_{jk} = \cos (X_j, X_k)$$

Ainsi la corrélation est traduite en termes de l'angle que font les axes portant les variables.

Exemple :



La conclusion quant à la relation entre les 5 variables se résume ainsi :

- X_1 et X_2 sont fortement corrélées positivement.
- X_3 et X_1 sont peu corrélées.
- X_4 et X_1 sont fortement corrélées négativement
- X_5 est mal représenté.

ii. Interprétations des axes graphiques en ACP

a. Graphique des variables

On a vu que la liaison entre deux variables s'interprète en terme de l'angle que font les axes portant ces deux variables. Ainsi, en ACP, il faut considérer les variables en tant qu'axes et non en tant que point (voir l'exemple ci-dessus).

b. Graphique des individus

S'agissant des individus, l'origine des axes constitue leur centre de gravité. De ce fait, le nuage des individus est bien réparti autour de ce centre. Ainsi :

- La proximité entre deux points individus (bien représentés) signifie un comportement similaire vis-à-vis l'ensemble des variables.

- Les individus qui sont au centre du graphique et qui sont bien représentés vont avoir un comportement qui ressemble à la moyenne.
- Les éléments se trouvant à l'extrémité du nuage vont avoir un comportement particulier et vont jouer un rôle important dans l'analyse.

c. Représentation simultanée individus-variables

On peut considérer la représentation simultanée des individus et variables sur le même graphique. C'est une représentation artificielle étant donné que les deux nuages de points appartiennent à deux espaces différents. De ce fait, La proximité entre points variables et points individus n'a pas de signification.

Facteur de taille :

Un axe est qualifié de taille si toutes les variables sont corrélées positivement et forment un faisceau autour de cet axe. Ainsi les individus qui se trouvent du côté droit de l'axe en question auront de grandes valeurs pour l'ensemble des variables, ceux du côté gauche par contre vont avoir des valeurs petites pour l'ensemble des variables.

B. Application aux données :

i. Présentation des données :

Nous disposons pour notre étude de contrats d'assurance automobile groupés selon plusieurs critères notamment la zone de circulation du véhicule, le carburant, l'usage et finalement le sexe du conducteur. Compte tenu des modalités de chaque critère la totalité des observations est d'ordre de 17 unités d'observations.

L'étude porte sur six variables quantitatives mesurées sur les 17 observations à savoir :

Tableau 11 : libellé des variables de l'ACP

Variable	libellé
Exposition au risque	Exposition
Nombre de sinistres	Nbr_sin
Fréquence	Freq
Coût moyen	CM
Ratio de sinistralité	SP
Charge	Charge

ii. [Matrice des corrélations entre variables :](#)

Le tableau suivant présente la matrice de corrélation entre les 6 variables étudiés :

Le Système SAS

La procédure CORR

6 Variables : exposition nbr_sin Freq CM SP charge

Tableau 12 : matrice des corrélations des variables de l'ACP

Coefficients de corrélation de Pearson						
Nombre d'observations						
	exposition	nbr_sin	Freq	CM	SP	charge
exposition	1.00000	0.96004	-0.23232	-0.18565	-0.09598	0.94283
	17	17	17	16	17	17
nbr_sin	0.96004	1.00000	-0.11657	-0.26332	-0.00609	0.98398
	17	17	17	16	17	17
Freq	-0.23232	-0.11657	1.00000	-0.04659	0.82373	-0.13340
	17	17	17	16	17	17
CM	-0.18565	-0.26332	-0.04659	1.00000	0.08355	-0.11131
	16	16	16	16	16	16
SP	-0.09598	-0.00609	0.82373	0.08355	1.00000	0.00212
	17	17	17	16	17	17
charge	0.94283	0.98398	-0.13340	-0.11131	0.00212	1.00000
	17	17	17	16	17	17

Les valeurs en gras sur le tableau représentent les corrélations fortes entre les six variables

[Commentaire et explication :](#)

On constate que la variable exposition au risque est positivement corrélée avec la variable charge. En effet, un assuré d'une exposition au risque importante bénéficie d'une longue durée de couverture et donc risque de coûter à l'assureur plus cher qu'un assuré dont la durée de couverture est petite et donc le sinistre a moins de chance pour se produire.

On constate également que le ratio de sinistralité est parfaitement corrélé avec la fréquence du sinistre. En effet plus un contrat fait preuve d'une sinistralité fréquente, plus sa rentabilité pour l'assureur est faible et son ratio de sinistralité est alors important.

iii. [Valeurs propres et taux d'inertie :](#)


Les out put suivants sont fournis par le logiciel SAS.ils résument quelques statistiques simples sur les six variables, puis présentent les huit premiers axes factoriels, les valeurs propres qui leur sont associées et l'inertie expliquée par chaque axe.

The PRINCOMP Procedure

Observations	17
Variables	6

Tableau 13 : valeurs propres et inertie de l'ACP

Eigenvalues of the Correlation Matrix				
	Valeur propre	Différence	Proportion	Cumulée
1	3.45316005	2.13995432	0.5755	0.5755
2	1.31320573	0.37541610	0.2189	0.7944
3	0.93778962	0.70129786	0.1563	0.9507
4	0.23649176	0.18036179	0.0394	0.9901
5	0.05612997	0.05290711	0.0094	0.9995
6	0.00322286		0.0005	1.0000

Le taux d'inertie 

Commentaires et conclusions

D'après le tableau ci-dessus, les trois premiers axes factoriels expliquent 95,07% de l'inertie total. La part du premier axe factoriel (F1) est 57,55% de l'inertie total. Le deuxième axe (F2) contribue avec 21,89% de l'inertie total, finalement le troisième axe contribue à 15,63% de l'inertie totale. Ainsi l'inertie expliquée par le premier plan factoriel est donc 79,44% de l'inertie total. On peut donc en conclure que la valeur de l'inertie expliquée par ces axes nous permet de mener l'analyse sur le premier plan factoriel ainsi que le plan (F1, F3). Nous effectuerons notre analyse sur le premier plan factoriel vu que son inertie est plus importante que celle du plan (F1, F3).

iv. Coordonnées, contributions et qualité de représentation

a. Pour les individus

Nous présenterons ici les coordonnées des 16 individus sur les deux premiers axes factoriels, leurs contributions ainsi que leur qualité de représentation pour s'assurer de la validité des conclusions que nous allons en tirer. Les résultats fournis par SAS et EXCEL sont comme suit :

Tableau 14 : coordonnées, contribution et qualité de représentation des unités d'observation de l'ACP

	Coordonnées des colonnes		Contribution à l'inertie		Qualité de représentation
	Prin1	Prin2	Prin1	Prin2	
ag_risq	-0.1223	1.1629	0.0024	0.2406	0.4176
ag_secu	0.0374	-0.3552	0.0007	0.0735	0.4176

essence	0.3051	0.2425	0.0477	0.0330	0.4281
gasoil	-0.8599	-0.6835	0.1343	0.0930	0.4281
debutant	-0.3057	2.5173	0.0044	0.3283	0.4702
experime	0.0224	-0.1841	0.0003	0.0240	0.4702
zone_risq	0.5888	-0.1125	0.0402	0.0016	0.0721
zone_risq_f	-0.2068	0.1489	0.0066	0.0038	0.0186
zone_risq_m	-0.5783	0.1727	0.0887	0.0087	0.2256
zone_t_risq	0.7405	-0.3527	0.0867	0.0216	0.1986
hom	2.2439	0.2486	0.2894	0.0039	0.4605
fem	-0.2027	-0.0225	0.0261	0.0004	0.4605
anci	-0.2995	0.2954	0.0390	0.0416	0.2966
neuf	0.3619	-0.4532	0.0062	0.0106	0.0246
norm	0.5331	-0.5043	0.0602	0.0591	0.2370
M3T5	-0.9897	-0.1913	0.0936	0.0038	0.1623
P3T5	-1.3430	-1.3167	0.0447	0.0471	0.1311
autre	0.7454	-0.6481	0.0005	0.0004	0.0012
tourisme	0.2222	0.0900	0.0024	0.0051	0.2714

Les valeurs encadrées correspondent aux observations qui contribuent significativement aux inerties expliquée par le premier ou le deuxième axe factoriel et qui sont très bien représentées.

b. Pour les variables :

Tableau 15 : coordonnées, contribution et qualité de représentation des variables de l'ACP

	coordonnés		contribution		qualite de representation
	Prin1	Prin2	Prin1	Prin2	
exposition_risque	-0,8766173	-0,263566	0,16545296	0,43079611	0,8507096821
chg_rc	0,41625076	0,783906	0,416911576	0,18874206	0,7500000462
Freq	-0,6854642	-0,26158	0,16720363	0,3258202	0,50063976
Cm	0,5837674	0,185068	0,16686715	0,0263095417	0,490981213
Sp	-0,5854640	0,26193	0,1672036	0,03267039	0,50063847

Commentaire et interprétations :

D'après le tableau ci-dessus on constate que la variable charge contribue fortement à l'inertie expliquée par le premier axe factoriel (41,629%). En plus elle jouit d'une excellente qualité de représentation (75%).le reste des variables contribuent par une quantité similaire à l'inertie expliquée par le premier axe. Pour le deuxième axe factoriel, c'est la variable exposition au risque qui emporte la part majoritaire de la contribution à l'inertie expliquée par l'axe en question et ce, avec une part de 43% et une qualité de représentation de 85%.

- v. Analyse des résultats dans le premier plan factoriel :
 - a. Analyse du premier axe factoriel
 - En termes de variables :

Ici on cherche les variables ayant une bonne qualité de représentation et dont les contributions sont significatives. On les classera en fonction de leur signe des coordonnées sur les deux axes factoriels. Deux ensembles sont alors mis en évidence. On les notera E^+ et E^- . les résultats sont résumés dans le tableau suivant :

Tableau 16 : classification des variables par contribution et signe de coordonnées vis-à-vis du premier axe

Axes	E^+	E^-
	charge (41,69%)	Exposition (16,54%)
Premier axe factoriel	Cm(16,68%)	Freq (16,72%)
		S/P (16,7%)

Les chiffres entre parenthèses indiquent la valeur de la contribution des variables. Les variables formant les éléments de ces deux ensembles contribuent pour plus de 90% à l'inertie expliquée par cet axe.

- En termes d'observations :

De même, on classe les observations qui contribuent significativement à l'inertie expliquée par le premier axe factoriel en deux pôles différents en termes de signe des coordonnées sur cet axes. Les résultats sont comme suit :

Tableau 17 : classification des variables par contribution et signe de coordonnées :

Axes	E^+	E^-
	homme (28,94%)	M3T5 (9,36%)
Premier axe factoriel	essence(4,77%)	Zone à risque moyen(8,87%)
	Zone très risquée(8,67%)	P3T5(4,47%)
	Véhicule normal(6,02%)	Ancien(3,9%)
	Zone risquée(4%)	Gasoil(13,43%)

Commentaires et interprétations :

D'après les deux tableaux précédents, on conclut que le premier axe factoriel oppose deux pôles d'individus : les véhicules anciens à usage commercial, à combustion gasoil circulant dans les zones à risque moyen qui ont des fortes valeurs pour la variable fréquence, exposition au risque et S/P cela est du en grande partie au fait que l'usage commercial se caractérise par une fréquence de sinistre importante ainsi que l'emploi du gasoil comme carburant. Et le pôle des conducteurs homme qui circulent dans des zones risquées et très risquées qui ont des valeurs fortes pour les variables charge et cout moyen. Ce résultat est prévu vu que les zones risquées coutent cher à l'assureur en termes de charge.

b. Analyse du deuxième axe factoriel

- [En termes d'observations :](#)

Tableau 18 : classification des unités d'observations par contribution et signe de coordonnées vis-à-vis du deuxième axe

Axes	E^+	E^-
	essence(3,3%)	Age- sécurisé(7,35%)
Deuxième axe factoriel	débutant (32 ,83%)	gasoil(9,30%)
	Ancien(4,16%)	Expérimenté(2,4%)
	Age risq(24,06%)	Zone très risquée(2,16%)
		Normal(5,91%)
		P3T5(4,71%)

- [En termes de variables :](#)

Tableau 19 : classification des variables par contribution et signe de coordonnées vis-à-vis du deuxième axe

Axes	E^+	E^-
	Charge(18,87%)	Exposition au risque (43,07%)
Deuxième axe factoriel	Cm(2,6%)	Fréquence (32,85%)
	S/P(3,26%)	

[Commentaires et interprétations :](#)

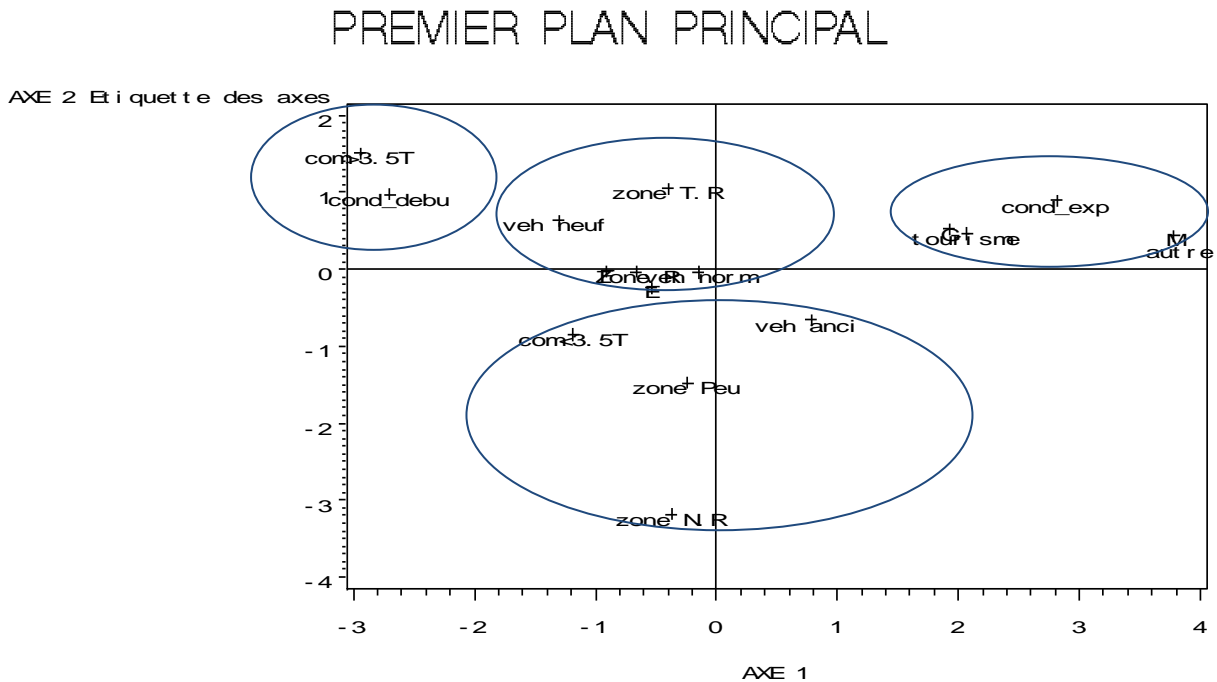
On conclut donc à partir des deux tableaux précédents que le deuxième axe factoriel oppose le pôle des conducteurs débutant d'âge risqué possédant un véhicule à combustion essence qui ont une grande valeur pour la variable charge, cout moyen et S/P avec le pôle des conducteurs expérimentés à âge sécurisé dont la fréquence des sinistres est importante.

c. Analyse dans le premiers plan factoriel (F_1 , F_2)

- [Représentation des individus :](#)

Le graphe suivant (figure), représente les 17 unités d'observations de l'étude sur le premier plan factoriel.la représentation est faite sous le logiciel SAS.

Figure 12: représentation des unités d'observations sur le premier axe factoriel :



La proximité de deux observations dans un plan factoriel signifie que ces deux observations ont un comportement similaire vis-à-vis de l'ensemble des variables. En effet, dans le premier plan factoriel (F1, F2), on peut construire quatre regroupements homogènes en terme de comportement vis-à-vis la sinistralité (en termes de fréquence, coût, charge et expositions au risque).

Ceci dit le risque de manque d'information ou de l'information imparfaite est toujours présent à cause du pourcentage parfois insuffisant d'inertie expliquée par le premier axe factoriel. En effet, la proximité entre deux points, sur la carte factorielle, ne reflète pas nécessairement qu'ils ont un comportement similaire vis-à-vis de l'ensemble des variables. Pour pallier à ce problème et pour fiabiliser les résultats de l'ACP on a procédé à une classification par la méthode de classification ascendante hiérarchique (CAH).

2. Les techniques de classification automatique :

A. Aspect théorique

L'objectif des techniques de classification est de produire d'une manière automatique des groupements des classes d'individus décrit par un certain nombre de variables quantitatives. Il s'agit donc de chercher une partition de la population des individus en k classes $P = \bigcup_{i=1}^k P_i$. Contrairement à l'AFD où les variables permettaient de décrire la partition connue a priori, Ici on va essayer d'utiliser les variables pour obtenir cette partition.

Les techniques de classification sont basées sur une démarche algorithmique et non sur le calcul algébrique valable pour les méthodes factorielles.

Le nombre de classes dépend des objectifs. Il y a plusieurs types de techniques de classification.

- Classification ascendante hiérarchique CAH :

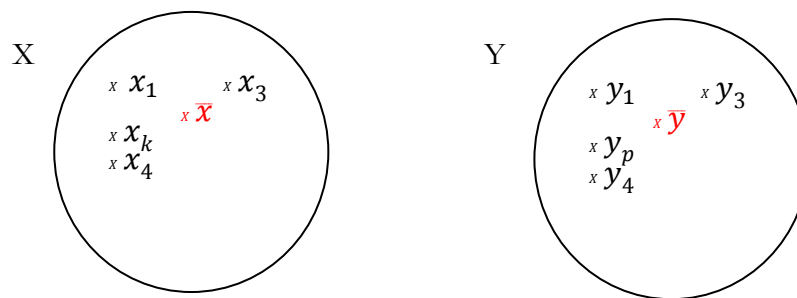
Elle produit une suite de partitions les unes plus fines que les autres (l'une est obtenue en agrégeant deux classes de l'autre). On part initialement d'une partition triviale à n classes : $\{x_1\}, \{x_2\}, \dots, \{x_n\}$ pour arriver à la fin à la partition évidente en une classe $\{x_1, x_2, \dots, x_n\}$. Il existe également des classifications descendantes qui font l'inverse.

- CAH utilisant une pseudo-distance :

Distance entre classes :

On parlera d'une pseudo-distance lorsque l'inégalité triangulaire n'est pas nécessairement exigée.

Problème : on dispose de la distance entre deux éléments, comment définir la distance entre deux ensembles ?



On peut définir $d(X, Y) = \min_{x \in X, y \in Y} d(x, y)$ distance du saut minimal

$D(X, Y) = d(\bar{x}, \bar{y})$ distance entre les centres

$D(X, Y) = \text{moyenne}_{x \in X, y \in Y} d(x, y)$ distance moyenne

- a) Algorithme de la CAH :

Etape 0 : on a la partition $\{\{x_1\}, \{x_2\}, \dots, \{x_N\}\}$, $k=N$

Etape 1 : on calcule les distances entre les k classes (il y a C_k^2 distances à calculer).

Etape 2 : on agrège les deux classes les plus proches.

Etape 3 : on pose $k \leftarrow k - 1$ on va à étape 1 l'algorithme s'arrête lorsque $k=1$

Remarque : l'algorithme produit une hiérarchie de N partitions (classifications). La première à N classes $P_1 = \{\{x_1\}, \{x_2\}, \dots, \{x_N\}\}$ la dernière à une classe $P_N = \{x_1, \dots, x_N\}$.

Exemple :

Afin de visualiser la procédure, Nous illustrons ces étapes en prenant comme objets à classer cinq points du plan, et comme distance entre ces objets le carré de leur distance usuelle en centimètres (un carreau sur la figure 1 est censé représenter 1cm). La matrice des distances ainsi définies est donnée par le tableau 1A. La règle de calcul pour cette règle sera, pour cet exemple, le saut minimal.

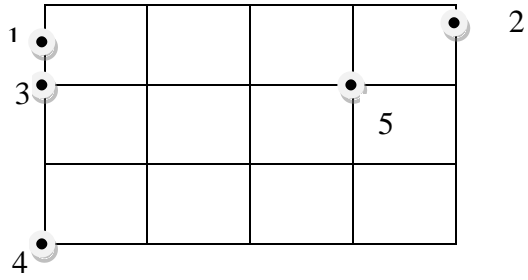


Fig.1.

Tableau 1.

	(1)	(2)	(3)	(4)	(5)
(1)	0	16	1	9	10
(2)	16	0	17	25	2
(3)	1	17	0	4	9
(4)	9	25	4	0	13
(5)	10	2	9	13	0

Tableau 1A

	(1)	(4)	(5)
(1)	0	9	4
(4)	9	0	13
(5)	4	13	0

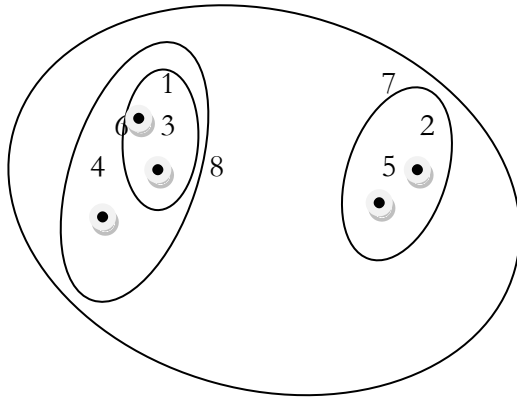
Tableau 1C

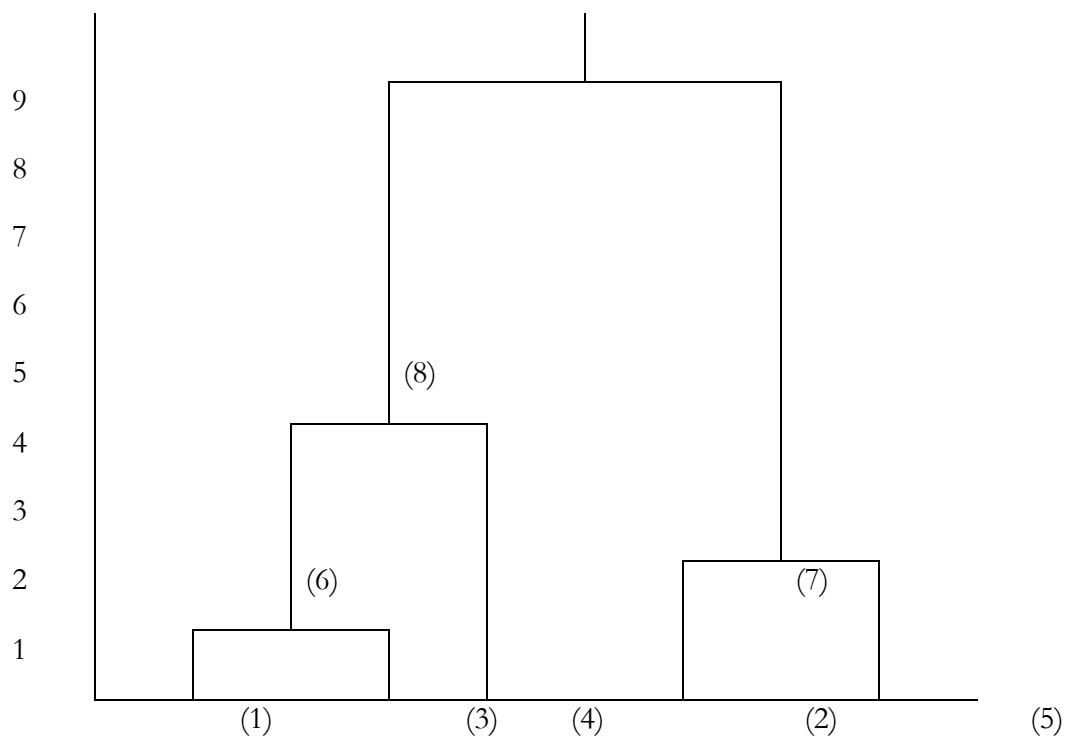
	(6)	(2)	(5)	(4)
(6)	0	16	9	4
(2)	16	0	25	25
(5)	9	25	0	13
(4)	4	25	13	0

tableau 1B

	(8)	(7)
(8)	0	9
(7)	9	0

tableau 1D





B. Application aux données :

Nous visons fiabiliser et forger les résultats fournis par l'ACP. On applique alors la classification hiérarchique ascendante aux données et ce, via deux étapes principales : La première a pour objectif de classifier les individus et aboutir à des classes à variance intra-classes minimales et variances inter-classes maximales. Pour ce faire, on fait appel à la procédure cluster du logiciel SAS dont la syntaxe est la suivante :

```
PROC CLUSTER METHOD=nom de méthode options ;
VAR variables ;
RUN ;
```

La deuxième établit une arborisation de cette classification permettant ainsi de visualiser les étapes de la classification des individus plus clairement. Cette arborisation est faite grâce à la procédure proc tree de SAS on donne sa syntaxe :

```
PROC TREE DATA=sortie de cluster OUT= fichier NCLUSTERS=nombre
de classes;
COPY variables;
RUN;
```

Les résultats des deux procédures se présentent comme suit :

Le Système SAS

The CLUSTER Procédure

Tableau 20 : étapes de la CAH et mesures du R^2 de chaque étape

Historique des classifications						
NCL	Classifications jointes		FREQ	SPRSQ	RSQ	T i e
15	OB6	OB8	2	0.0028	.997	
14	OB7	OB15	2	0.0028	.994	
13	CL15	OB17	3	0.0068	.988	
12	OB4	OB9	2	0.0073	.980	
11	OB12	OB14	2	0.0080	.972	
10	OB11	OB13	2	0.0103	.962	
9	CL14	CL11	4	0.0149	.947	
8	CL12	CL10	4	0.0219	.925	
7	OB1	CL9	5	0.0220	.903	
6	OB2	CL13	4	0.0288	.874	
5	OB5	OB16	2	0.0335	.841	
4	CL8	OB10	5	0.0618	.779	
3	CL7	CL4	10	0.1469	.632	
2	CL3	CL5	12	0.1916	.441	
1	CL2	CL6	16	0.4406	.000	

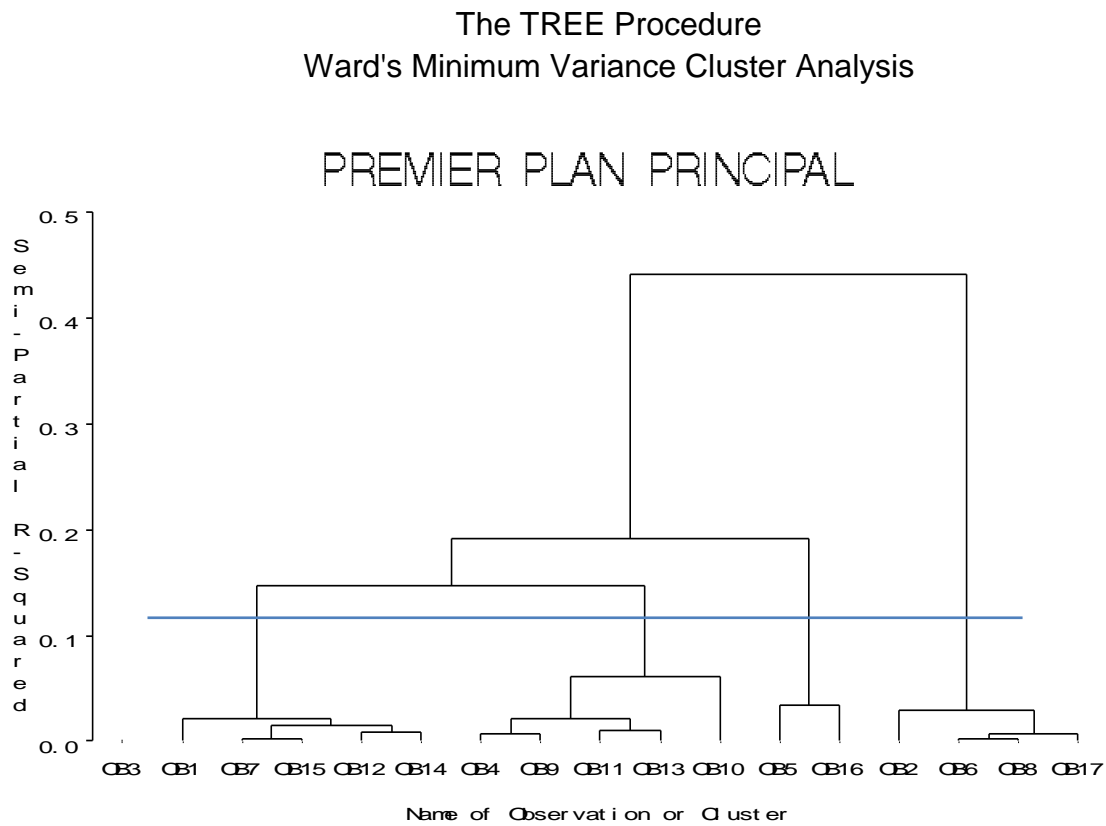
Saut du R^2

Interprétations et commentaires

La variable NCL représente le nombre de classes dans chaque étape de classification, la répartition devrait s'arrêter lorsque le R^2 fait un saut de valeur remarquable. Dans notre cas le R^2 passe d'une valeur de 0,0618 à une valeur de 0,146 au niveau de la quatrième étape de la classification ce qui correspond à un nombre de classes de quatre catégories.

On visualiser ces quatre classes via une arborisation des classes. Le résultat se présente ainsi :

Figure 13 : arbre de la CAH



Conclusion :

Le recours à la classification hiérarchique avait pour objectif la fiabilisation des résultats obtenus par l'analyse en composantes principales. Le nombre de classes obtenu est le même que celui visualisé sur le premier plan factoriel et l'identification des codes sur l'arborisation coïncide effectivement avec les unités d'observations regroupées.

Chapitre 2: Le modèle linéaire généralisé :

Dans ce chapitre et après avoir effectué les différentes analyses uni-variées et multi-variées sur les variables de l'étude, nous passons maintenant à l'application du modèle linéaire généralisé.

En premier lieu, nous s'assurons des corrélations entre les variable et ce en appliquons la matrice des corrélations pour les variable quantitative et le test de chi deux pour les qualitatives.

Dans la seconde partie, nous analysons l'ajustement de la fréquence et le coût moyen aux différentes lois candidates.

Dans la dernière partie, nous appliquons le GLM sur la fréquence et le coût moyen pour enfin déterminer les différentes primes de la garantie Responsabilité Civile Automobile.

- I. Analyse des corrélations entre les variables explicatives :
 1. Les variables quantitatives

Les variables quantitatives mise en jeu dans notre étude sont : l'âge de l'assuré, l'âge du véhicule et l'âge du permis. Les corrélations sont mesurées par le coefficient de Pearson résumé sur la matrice de corrélation ci-dessous :

Tableau 21 : corrélation entre âge conducteur et âge permis

Coefficients de corrélation de Pearson, N = 1125376			
Prob > r under H0 : Rho=0			
	Age conducteur	permis	véhicule
Age conducteur	1.00000	0.70752 <.0001	-0.05113 <.0001
permis	0.70752 <.0001	1.00000	-0.12568 <.0001
véhicule	-0.05113 <.0001	-0.12568 <.0001	1.00000

Commentaire et interprétation :

On constate une corrélation importante entre l'âge du conducteur et celui du permis ce qui est normal puisque l'ancienneté du permis suppose l'avancement de son titulaire dans l'âge. Par conséquent la variable ancienneté du permis sera exclue du modèle.

2. Les variables qualitatives :

Nous disposons de quatre variables qualitatives objet de l'étude à savoir : La zone de circulation, l'usage, le sexe du conducteur et le carburant du véhicule.

Afin de mettre en évidence les corrélations existantes entre ces variables nous ferons appel au test de khi2 pour l'indépendance. Sous SAS ce test est donnée par la procédure « proc freq ». Les résultats sont comme suit :

a) Khi2 pour l'indépendance entre le carburant et l'usage :

Tableau 22 : résultat du test khi2 d'indépendance entre carburant et usage

Statistique	DF	Valeur	Proba.
Khi-2	2	39886.8904	<.0001
Test du rapport de vraisemblance	2	53303.8599	<.0001
Khi-2 de Mantel-Haenszel	1	39008.1841	<.0001
Coefficient Phi		0.1883	
Coefficient de contingence		0.1850	
V de Cramer		0.1883	

Commentaire et interprétation :

Le test met en évidence une corrélation entre le carburant du véhicule et son usage, cette corrélation s'explique par l'usage de gasoil par exemple comme carburant dans tout usage relatif à l'activité commercial.

b) Khi2 pour l'indépendance entre le sexe et la zone :

Tableau 23 : résultat du test khi2 d'indépendance entre sexe et zone

Statistique	DF	Valeur	Proba.
Khi-2	16	36613.5077	<.0001
Test du rapport de vraisemblance	16	40335.2122	<.0001
Khi-2 de Mantel-Haenszel	1	5001.8191	<.0001
Coefficient Phi		0.1804	
Coefficient de contingence		0.1775	
V de Cramer		0.1804	

Commentaire et interprétation :

Le test met en évidence une corrélation entre la variable sexe et la zone de circulation. Cette corrélation peut être expliquée par le constat que les femmes qui conduisent au Maroc sont généralement celle circulant des les grandes villes du pays : Casablanca, Marrakech, rabat....

c) Khi2 pour l'indépendance entre l'usage et la zone :

Tableau 24 : résultat du test khi2 d'indépendance entre zone et usage

Statistique	DF	Valeur	Proba.
Khi-2	32	24061.0412	<.0001
Test du rapport de vraisemblance	32	23811.9066	<.0001
Khi-2 de Mantel-Haenszel	1	291.6813	<.0001
Coefficient Phi		0.1462	
Coefficient de contingence		0.1447	
V de Cramer		0.1034	

Commentaire et conclusion :

Le test de l'indépendance est rejeté. Ainsi, les deux variables zone et usage ne sont pas indépendantes. Cependant, vu que ces variables sont importantes, nous allons appliquer le glm sur chaque usage pour garder les deux variables.

II. Etude des lois des variables dépendantes et tests d'adéquation :

1) Lois de la fréquence :

A présent, on cherche à faire des intuitions quant à la distribution de la fréquence des sinistres avant de lui appliquer le modèle linéaire généralisé. Le but étant de tester un nombre minimum de modèle en se limitant à ceux les plus susceptibles d'être performants et valides.

En assurance automobile, les lois candidates pour modéliser la fréquence sont la loi poisson et la loi binomiale négative. On donne au dessous les box plots relatifs à la loi binomiale négative et la Loi poisson successivement. Les box plot sont déterminés sous le logiciel R.

Figure 14 : ajustement de la fréquence des sinistres à la loi binomiale négative

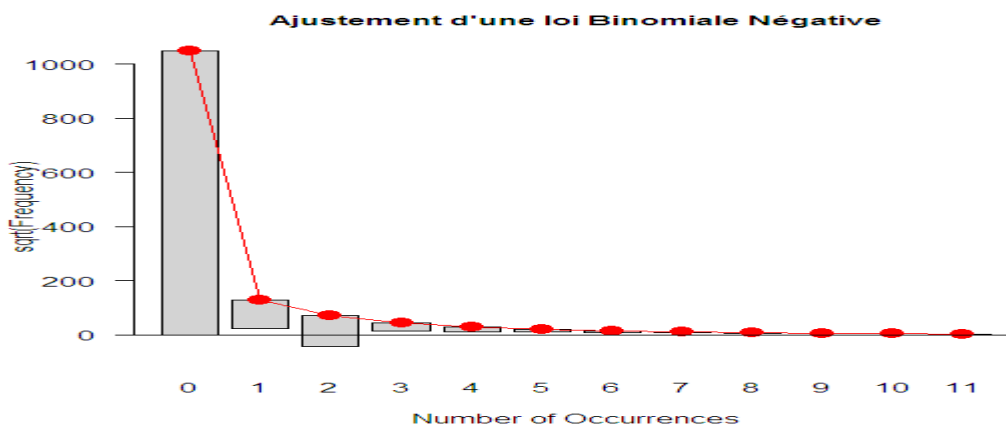
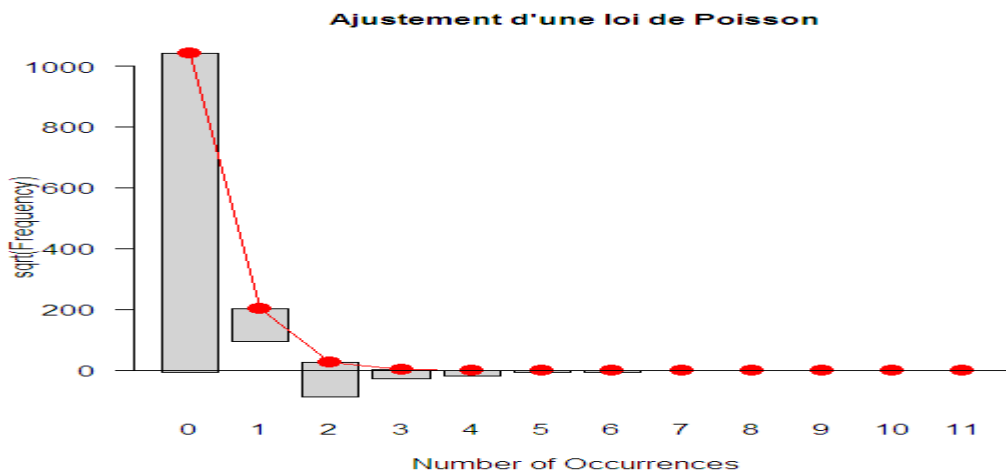


Figure 15 : ajustement de la fréquence des sinistres à la loi poisson



Commentaire et conclusion :

On constate que la fréquence des sinistres s'ajuste mieux à la loi binomiale négative qu'à la loi de poisson. En effet, les statistiques sur la fréquence des sinistres montrent que la variance de la fréquence est supérieure à sa moyenne. De ce fait, la loi de poisson s'avère incompatible d'où le recours à une loi à queue épaisse notamment la loi binomiale négative.

Le résultat fourni par le box plot doit être renforcé et validé par un test d'ajustement. Ici on a recouru à un test χ^2 d'adéquation le résultat est le suivant (output du logiciel R):

Pour la loi poisson :

```
Chi-squared test for given probabilities
data: freq.empirique
X-squared = 554882486, df = 12, p-value < 2.2e-16
```

Pour la loi Binomiale Négative:

```
Chi-squared test for given probabilities
data: freq.empirique
X-squared = 0.0138, df = 30, p-value = 0,974
```

A partir des résultats ci-dessus, le test ne peut être rejeté et la fréquence des sinistres semble effectivement suivre une loi binomiale négative.

2) Les lois du coût moyen :

Pour déterminer la loi de la charge, on fait appel à de nombreuses approches et outils, notamment le QQ-plot, les box-plot, les proba-plots et bien évidemment les tests d'adéquations pour confirmer la validité d'une loi.

En règle générale, les montants de sinistres sont modélisés à partir d'une loi Gamma ou d'une loi log normale. En effet, les ces derniers correspondent bien à une distribution continue, définie sur les réels positifs, et ayant une variabilité qui augmente avec la moyenne.

Néanmoins, Un problème classique en assurance non vie est le poids très important des sinistres extrêmes. Le ratio S/P est très sensible à la présence de gros sinistres. Ainsi, il ne peut y avoir de pénalisation si par exemple un très gros sinistre atteignant 500 ou 1000 fois la prime annuelle frappe un des clients car il s'agit d'un phénomène aléatoire. Telle est l'origine de la politique de l'écrêtement.

Afin d'écrêter les sinistres nous devons les plafonner à un niveau M. La charge excédentaire (ou sur crête) S-M est répartie entre tous les assurés.

Deux questions se posent : comment déterminer le seuil M et comment répartir l'excédent (sur crête) entre les assurés ?

Il y a plusieurs méthodes pour déterminer le seuil d'écrêtement, par exemple choisir un quantile du montant des sinistres ou déterminer un seuil tel que la charge écrêtée représente un pourcentage précis de la charge totale...

Reste à savoir comment répartir la charge excédentaire sur l'ensemble du portefeuille. Plusieurs approches sont envisageables :

- ✓ Une première méthode consiste à répartir la sur crête entre tous les assurés. L'idée sous jacente est que ces gros sinistres sont dus à la malchance et qu'ils pouvaient toucher n'importe quel assuré. Il est donc évident de faire supporter à tous les assurés une partie de la charge.
- ✓ Une deuxième approche (l'approche retenue) consiste à partager la sur crête des sinistres sur tous les assurés mais pondérés par des poids.

On est donc amené à choisir des poids ω_i pour pondérer la répartition de la charge excédentaire. Pour ce faire on a recours à deux méthodes classiques. Une première consiste à choisir des poids proportionnels soit à la charge que doit payer l'assuré après l'écrêtement (sous crête) soit au nombre de sinistres. Cette méthode présente deux défauts. Elle attribue tout d'abord plus de charge aux contrats déjà touchés ce qui va engendrer une double pénalisation. Le deuxième défaut réside dans le fait qu'un contrat qui n'a aucun sinistre va se trouver avec une charge nulle ce qui contredit la logique de l'approche qui suppose qu'un sinistre à une probabilité non nulle de toucher n'importe quel assuré. Nous avons choisit finalement de pondérer proportionnellement à la part de la prime du contrat de la totalité de la prime.

Ainsi, chaque assuré se voit attribuer une charge sur crête*primes de l'assuré/Prime totale de la classe.

Application :

Le calcul sur Excel donne montre que le quantile 95% de la loi de la charge correspond à un pourcentage de 35% de la charge. Le résultat étant jugé satisfaisant, on plafonne la charge de tout sinistre à un maximum de 80 000DH. Ainsi, tout montant de sinistre supérieur à 80 000 DH fera l'objet d'un écrêtement et la charge excédentaire et sera réparti sur le reste des contrats en fonction de leur attribution à la prime totale.

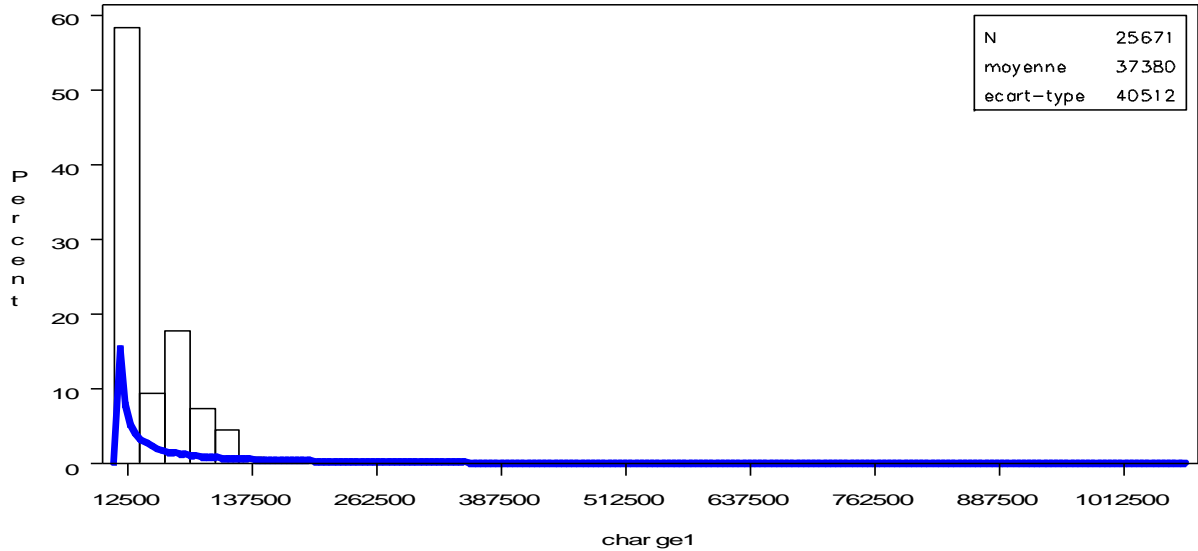
Une fois l'écrêtement est effectué, on passe au choix de la loi qui s'ajuste le mieux aux données.

Afin de déterminer laquelle des deux lois (gamma ou log-normal) s'ajuste mieux à la distribution de la charge on exploite l'histogramme des deux distributions pour trancher avant de valider les résultats par un test d'ajustement adéquat.

Les histogrammes des deux lois se présentent comme suit :

➤ La loi Gamma

Figure 16 : ajustement de la charge des sinistres à la loi Gamma



Commentaire :

L’histogramme dévoile un décalage considérable entre la distribution de la charge et celle de la loi gamma. Pour confirmer ce constat on fait appel aux tests d’ajustement. Le résultat est comme suit :

Tableau 25 : test d’ajustement de la charge à la distribution Gamma

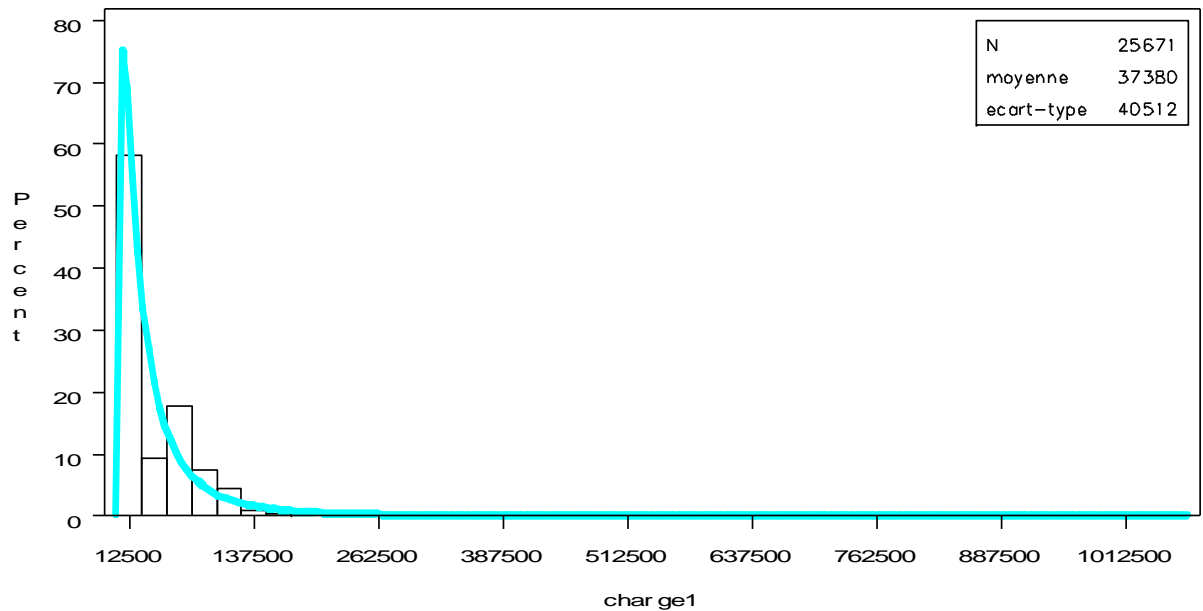
Goodness-of-Fit Tests for Gamma Distribution			
Test	Statistique		p Value
Kolmogorov-Smirnov	D	0.24013	Pr > D <0.001
Cramer-von Mises	W-Sq	235.78988	Pr > W-Sq <0.001
Anderson-Darling	A-Sq	1170.82901	Pr > A-Sq <0.001

Conclusion :

Le test d’ajustement kolmogorov-smirnov est rejeté et confirme ainsi le constat de l’histogramme. Par conséquent, la loi gamma est éliminée de l’étude. Reste à tester la loi log normale.

➤ La loi log normale :

Figure 17 : ajustement de la charge des sinistres à la loi log-normale



Commentaire :

L'histogramme montre un ajustement satisfaisant entre la distribution théorique et celle de la charge augmentant ainsi les chances de la loi log normale d'être la loi adéquate pour la distribution de la charge.

Test d'ajustement :

Tableau 26 : test d'ajustement de la charge à la distribution Gamma

Goodness-of-Fit Tests for Lognormal Distribution				
Test	Statistique		p Value	
Kolmogorov-Smirnov	D	0.162780	Pr > D	0.046

Le test semble rejeter l'hypothèse d'ajustement de la distribution de la charge à une loi log normale pour un seuil supérieur à 4,6% mais on l'accepte pour tout seuil inférieur à ce dernier.

Conclusion :

Nous pouvons ainsi conclure que la loi log-normale est la loi qui s'ajuste le mieux aux données des montants des sinistres.

III. Le modèle linéaire généralisé :

A présent, nous disposons d'une vision préliminaire sur les corrélations entre les variables explicatives et les intuitions quant à la distribution de la fréquence et du coût. On applique alors le modèle linéaire généralisé en premier temps à la fréquence, en suite au coût des sinistres. Une appréciation de la validité et la robustesse des modèles fera objet d'une étude postérieure.

A. Aspect théorique :

1) Logique du modèle

Un modèle linéaire généralisé a pour but de relier des variables explicatives $X = (X_1, X_2, \dots, X_p)$ à une variable à expliquer Y . La logique sous-jacente à un tel modèle peut alors être résumée à travers le schéma suivant :

à expliquer Composante aléatoire	← Lien →	← Explicatif →
<p>Y suit une loi de la famille exponentielle et sa densité est de la forme :</p> $f(x/\theta, \phi) = \exp \left\{ \frac{x\theta - b(\theta)}{a(\phi)} + c(x, \phi) \right\}$ <p>Nous savons alors que :</p> $E(Y) = \mu = b'(\theta)$ $\text{Var}(Y) = b''(\theta) \cdot a(\phi) = V(\mu) \cdot a(\phi)$	<p>L'espérance de Y noté μ dépend de $\eta(X)$ à travers une fonction de lien noté $g()$, monotone et dérivable, donc inversible.</p> $g(\mu) = \eta(X)$ <p>La fonction de lien canonique est une fonction de lien particulière qui vérifie la relation :</p> $\mu = g^{-1}(\theta) \Leftrightarrow \theta = \eta(X)$	<p>Soit $x = (x_1, \dots, x_p)$ une observation des variables explicatives. On définit le prédicateur linéaire associé à cette observation par :</p> $\eta(X) = \sum_{i=1}^p x_i \beta_i$ <p>Les coefficients $(\beta_1, \dots, \beta_p)$ doivent être estimés.</p>

On conclut qu'il est nécessaire d'effectuer deux choix pour construire un modèle linéaire généralisé. Le premier concerne la loi de la variable à expliquer, ce choix peut être orienté par le type de la variable et des connaissances préalables. Le deuxième choix porte sur la fonction lien. Nous reprenons ci-dessous à travers un tableau les fonctions de liens classiquement utilisées. Le choix de la densité peut alors dépendre de la loi.

- Si Y est binaire, on préférera utiliser les liens logit, probit ou cloglog, si Y est un comptage, on utilisera classiquement le lien log, et enfin
- si Y est continue, on pourra utiliser les liens canoniques de la loi normale et de la loi gamma.

Le tableau ci-dessous résume les fonctions de lien classiques dans les modèles linéaires généralisés :

Tableau 27 : fonctions de liens usuelles

Nom du lien	Fonction du lien
Lien identité	$g(\mu) = \mu$
Lien log	$g(\mu) = \ln(\mu)$
Lien cloglog	$g(\mu) = \ln(-\ln(1 - \mu))$
Lien logit	$g(\mu) = \ln\left(\frac{\mu}{1 - \mu}\right)$
Lien probit	$g(\mu) = \Phi(\mu)$ Φ fonction inverse de la fonction de répartition d'une loi normale
Lien réciproque	$g(\mu) = -1/\mu$
Lien puissance	$g(\mu) = \mu^\gamma$ avec $\gamma \neq 0$ $g(\mu) = \ln(\mu)$ avec $\gamma = 0$
Aranda Ordaz (asymétrique)	$g(\mu) = \ln\left(\frac{(1 - \mu)^{-\lambda} - 1}{\lambda}\right)$

2) Estimation des paramètres :

Nous considérons une variable à expliquer Y , pour laquelle nous possédons des observations pour n individus notées (Y_1, \dots, Y_n) . Nous cherchons donc à expliquer cette variable à partir de p variables explicatives notées (X_1, \dots, X_p) .

Pour estimer les paramètres β , nous procédons par la méthode du maximum de vraisemblance dont la fonction s'écrit :

$$L(\theta/Y) = \prod_{i=1}^n f_{\theta}(Y_i)$$

Où f_{θ} est la fonction de densité

La maximisation de cette vraisemblance se fait par le logiciel SAS, celui-ci à recours à des méthodes itératives telles que celle de Newton Raphson.

Le logarithme de la vraisemblance s'écrit de la manière suivante :

$$L(\theta/Y, \phi) = \sum_{i=1}^n \ln \left(f_{\theta} \left(Y_i/\theta_i, \phi \right) \right) = \frac{\sum_{i=1}^n Y_i \theta_i - \sum_{i=1}^n b(\theta)}{a(\phi)} + \sum_{i=1}^n c(Y_i, \phi)$$

$$\text{où } E(Y_i) = \frac{\partial b(\theta)}{\partial \theta} \text{ et } V(Y_i) = a(\phi) \frac{\partial^2 b(\theta)}{\partial \theta^2}$$

Le but étant d'obtenir les estimateurs du maximum de vraisemblance des paramètres de régression β , en maximisant $L(\theta/Y, \phi)$.

L'estimateur du maximum de vraisemblance est solution du système suivant :

$$\begin{cases} U_j = \frac{\partial L(\theta/Y, \phi)}{\partial \beta_j} = 0 & \text{pour } j = 1 \dots p \\ U_j \text{ le score correspondant à la variable explicative } X_j \end{cases}$$

On peut encore écrire sous forme matricielle:

$$U_j = X^t W^{-1}(\hat{\beta}) (Y - \mu(\hat{\beta})) = 0$$

Où $W^{-1}(\hat{\beta})$ est une matrice donnée par : $V(Y)g'(\mu)$

La difficulté de ce problème de maximisation est l'inexistence d'une solution explicite, ce qui n'est pas très possible dans le cas où les équations de vraisemblances ne sont pas linéaires. Une façon de maximiser la vraisemblance est d'appliquer par exemple la méthode de Newton Raphson, dont l'algorithme est donné par :

$$\hat{\beta}_{k+1} = \hat{\beta}_k - H^{-1}(\hat{\beta}_k) U(\hat{\beta}_k)$$

Avec $H(\beta)$ la matrice Hessienne de $L(\beta/Y)$ i.e $H(\beta) = \frac{\partial^2 L(\beta/Y)}{\partial \beta_i \partial \beta_j}$ Et $U(\beta) = \frac{\partial L(\beta/Y)}{\partial \beta}$

La matrice Hessienne est aussi définie par la matrice variance covariance de la manière suivante : $H(\beta) = -\Gamma(\beta)$

Après plusieurs calculs on aboutit à la formule suivante :

$$\hat{\beta}_{k+1} = (X^t W^k X)^{-1} X^t W^k \left[X \beta^k + \text{diag} \left(\frac{\partial \eta_i}{\partial \mu_i} \right) (Y - \mu) \right]$$

Ainsi obtient-on les estimateurs du maximum de vraisemblance par la même logique du logiciel SAS utilisé dans le cadre de notre étude.

3) Adéquation du modèle et test de significativité :

Afin de juger la validité et la robustesse du modèle linéaire généralisé on fait appel à des critères et des analyses en mesures de confirmer la pertinence du modèle. On présente ci-dessous ces outils et leurs apports au jugement des modèles.

a) Test de Wald

Le but de ce test est de juger la significativité des variables explicatives et décider donc les quelles doivent être éliminées et les quelles expliquent significativement le modèle.

Analytiquement On cherche à tester : $H_0: A\beta = 0$ contre $H_1: A\beta \neq 0$

Où A est une matrice (q, p) donnée de plein rang avec $q < p$.

L'hypothèse H_0 implique l'existence de q liaisons linéaires entre les paramètres de régression β_j .

Pour tester H_0 contre H_1 on utilise la statistique de Wald donnée par :

$$W = \hat{\beta}^t (\hat{V}(\hat{\beta}))^{-1} \xrightarrow{\text{sous } H_0} \chi^2_{(q)}$$

Pour le test individuel sur les paramètres, on rejette H_0 au seuil α lorsque $W_{\text{obs}} > \chi^2_{(\alpha, q)}$.

b) Résidus

L'analyse des résidus est indispensable pour vérifier si les données ne contredisent pas les hypothèses du modèle. Cependant, dans le cadre du modèle linéaire généralisé les résidus bruts ne sont pas pertinents. On utilise essentiellement ceux de Pearson et de la déviance.

Pour $i=1, \dots, n$ on a :

- Résidu brut : $r_i = Y_i - \hat{\mu}_i$
- Résidu (non standardisé) de Pearson : $r_i^{(p)} = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$
- Résidu (non standardisé) de la déviance : $r_i^{(D)} = \text{sgn}(Y_i - \hat{\mu}_i) \sqrt{d_i}$ où

où $d_i = 2\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]\}$ Est le ième terme de la déviance, en posant :
 $\tilde{\theta}_i = b'^{-1}(Y_i)$ et $\hat{\theta}_i = b'^{-1}(\hat{\mu}_i)$

L'analyse de ces résidus est un indicateur de la validité du modèle. Il s'agit de vérifier la normalité de ces derniers pour confirmer que le modèle ne contredit pas les hypothèses de l'étude.

c) La déviance :

La déviance est une statistique qui reflète l'écart entre la vraisemblance de la distribution théorique et celle de la variable dépendante. Ainsi moins cette statistique est grande plus le modèle est pertinent.

B. Application aux données :

- GLM sur la fréquence :

On a vu dans la partie préparatoire au modèle linéaire généralisé que la fréquence des sinistres s'ajuste mieux à une loi binomiale négative. Néanmoins le GLM sera appliqué avec les deux distributions poisson et binomiale négative en vue de crédibiliser notre choix de distribution d'une part et de s'assurer mieux que le GLM retenu est le modèle le plus pertinent pour l'explication de la fréquence des sinistres. Ainsi nous allons analyser les résultats obtenus et mettre les GLM à l'épreuve via les outils de validation des modèles afin de s'assurer que le modèle et la distribution choisis sont pertinents et adéquats.

Remarque : nous avons choisi d'appliquer le modèle linéaire généralisé aux deux variables coût et fréquence de la manière suivante : fixer à chaque fois une modalité de la variable Usage et tester la significativité des autres variables explicatives. Pour la présentation des outputs nous nous contenterons d'étaler les résultats obtenus pour l'usage tourisme en raison de la multitude des outputs et leur aspect répétitif. Néanmoins le lecteur intéressé par les détails des modèles pourra se référer à la partie Annexe pour la liste exhaustive des résultats pour usage.

L'application est réalisée sous le logiciel SAS via la procédure **PROC GENMOD**.

Le tableau suivant établit une comparaison entre les résultats obtenus par le modèle avec la loi poisson et celui avec la loi binomiale négative. Nous présenterons d'abord les résultats permettant d'apprécier le modèle à retenir entre les deux, par la suite nous présenterons le détail du modèle retenu en termes de significativité des variables.

1) Qualité d'ajustement du modèle :

Il s'agit de comparer le rapport déviance/ degré de liberté à la valeur 1

Tableau 28 : comparaison des qualités d'ajustement du GLM sur la fréquence

Déviance/degré de liberté	
Modèle à la loi binomiale négative	Modèle à la loi poisson
0,2499	1,0281

Commentaire :

Le test sur la qualité d'ajustement confirme que le modèle à distribution poisson n'est pas le bon. En effet le rapport en question dépasse 1. nous en concluons que le modèle à distribution binomiale négative est le plus susceptible d'être le modèle approprié.

2) Comparaison de la déviance :

On rappelle que la déviance mesure l'écart entre la vraisemblance théorique et celle de la distribution. Ainsi lors d'une comparaison, le modèle qui aura une déviance minimale sera jugé meilleur. Le résultat de notre comparaison est le suivant :

Tableau 29 : comparaison des Déviiances entre Poisson et Binomiale Négative

Déviance	
Modèle à la loi binomiale négative	Modèle à la loi poisson
235535 ,3890	741212,44

Commentaire :

De même le test sur les deux déviiances qualifie le modèle à distribution binomiale. Par conséquent nous continuerons l'étude uniquement sur le modèle à distribution binomiale négative pour vérifier la significativité des variables explicatives et tester la normalité des résidus.

Figure 18 : GLM avec la loi binomiale négative

The GENMOD Procedure

Informations sur le modèle

Data Set	STAT.CHARGE1
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	nb_rc
Offset Variable	ldur
Number of Observations Read	942460
Number of Observations Used	942460

3) Test de significativité de wald

Figure 19 : test de significativité de Wald pour le GLM avec la loi Binomiale Négative

Analyse des résultats estimés de paramètres

Paramètre	DF	Estimation	Erreur standard	Wald 95Limites de confiance %		Chi 2	Pr > Chi 2
Intercept	1	-3.4106	0.0134	-3.4369	-3.3844	64773.4	<.0001
NÂge	1	0.4229	0.0130	0.3974	0.4483	1057.88	<.0001
NÂge	2	0.0000	0.0000	0.0000	0.0000	.	.
NVEH	1	0.4194	0.0151	0.3898	0.4490	770.90	<.0001
NVEH	2	0.1965	0.0129	0.1712	0.2217	232.41	<.0001
NVEH	3	0.0000	0.0000	0.0000	0.0000	.	.
NZone	1	0.6076	0.0149	0.5784	0.6367	1668.73	<.0001
NZone	2	0.3687	0.0159	0.3377	0.3998	541.18	<.0001
NZone	3	-0.5171	0.0267	-0.5694	-0.4648	375.84	<.0001
NZone	4	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion	0	0.0000	0.0000	0.0000	0.0000	.	.

NOTE: The negative binomial dispersion parameter was held fixed.

Commentaire :

On constate que le test ne rejette pas l'hypothèse de significativité des variables. On en conclut que les variables zone, usage, âge de l'assuré et âge du véhicule contribuent significativement à l'explication de la variable fréquence des sinistres.

4) Qualité de l'ajustement

Il s'agit d'évaluer le rapport de la déviance standardisé sur les degrés de liberté. Le modèle est jugé bon si ce rapport est inférieur à 1. les résultats sous SAS se présentent ainsi :

Figure 20 : Evaluation de la qualité d'ajustement du GLM pour la Binomiale Négative

Critère pour évaluer la qualité de l'ajustement			
Critère	DF	Valeur	Valeur/DF
Deviance	94E4	235535.3890	0.2499
Scaled Deviance	94E4	235535.3890	0.2499

Commentaire :

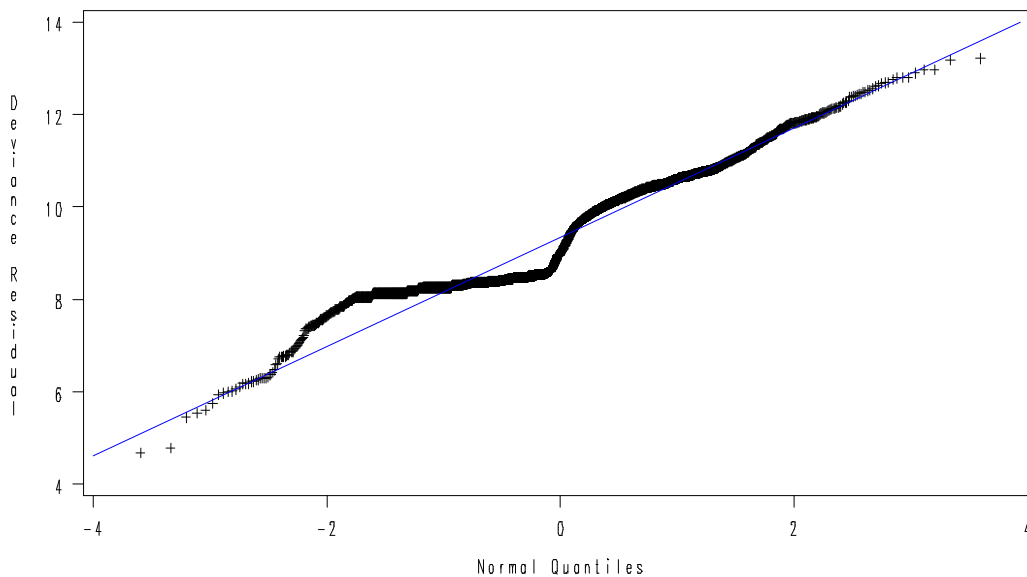
Les résultats sont satisfaisants. En effet, on constate que toutes les valeurs du rapport déviance/ddl sont suffisamment inférieures à 1. On peut donc en conclure que l'ajustement de notre modèle est bon.

5) Résidus

Pour lisser notre jugement sur la qualité et la validité de notre modèle, le point sera mis cette fois sur les résidus. On rappelle que l'objectif est de vérifier la normalité des résidus de la déviance .sinon la conformité de notre modèle aux hypothèses des modèles linéaires généralisés sera mise en cause.

Pour ce faire nous avons tracé proba-plot des résidus sur la distribution normale. Le proba-plot se présente ainsi :

Figure 21 : prob-plot des résidus pour la loi normale dans le cas d GLM sur la fréquence



Commentaire :

La distribution des résidus de la déviance s'ajuste de façon acceptable avec la loi normale. Ce qui nous rassure au sujet de la validité de notre modèle.

Conclusion :

Le modèle a prouvé la significativité des quatre variables explicatives : usage, zone, âge du véhicule et âge du conducteur. Sa validité a été mise à l'épreuve ; la qualité de son ajustement est qualifiée bonne vu que le rapport déviance/ddl est inférieur à un ce qui prouve que le choix de la loi binomiale négative était sage. De plus la normalité des résidus de la déviance a été vérifiée être validé via le proba_plot au dessus. Ainsi on peut confirmer la robustesse et la pertinence de notre modèle.

- GLM sur le coût moyen des sinistres:

Lors de la modélisation de la charge, deux lois candidates ont fait objet de l'étude Gamma et la log normale. L'ajustement par l'histogramme et le test de Kolmogorov Smirnov ont qualifié la loi log-normal au détriment de Gamma. A présent, nous allons appliquer le modèle linéaire généralisé sur les deux distributions et vérifier le quel des deux modèles expliquera mieux le cout des sinistres. Ce serait également une occasion pour fiabiliser ou remettre en cause notre choix de distribution.

Comme dans le GLM sur la fréquence, nous présenterons uniquement les résultats des modèles sur l'usage tourisme. Les sorties en totalité seront détaillé dans la partie Annexe. Le tableau suivant établit une comparaison entre les résultats obtenu par le modèle avec la loi gamma et celui avec la loi log normale. Nous présenterons d'abord les résultats permettant d'apprécier le modèle à retenir entre les deux, par la suite nous présenterons le détail du modèle retenu en termes de significativité des variables.

- 1) Qualité d'ajustement du modèle :

Il s'agit de comparer le rapport déviance/ degré de liberté à la valeur 1

Tableau 30 : comparaison des qualités d'ajustement du GLM sur le coût moyen

Déviance/degré de liberté	
Modèle à la loi log normale	Modèle à la loi Gamma
0,8023	2,0239

Commentaire :

Le test sur la qualité d'ajustement confirme que le modèle à distribution Gamma n'est pas le bon. en effet le rapport en question dépasse amplement 1. nous en concluons que le modèle à distribution log normale est le plus susceptible d'être le modèle approprié.

2) Comparaison de la déviance :

On rappelle que la déviance mesure l'écart entre la vraisemblance théorique et celle de la distribution. Ainsi lors d'une comparaison, le modèle qui aura une déviance minimale sera jugé meilleur. Le résultat de notre comparaison est le suivant :

Tableau 31 : comparaison des Déviations entre Gamma et log Normale.

Déviance	
Modèle à la loi log normale	Modèle à la loi Gamma
5714	78122,44

Commentaire :

De même le test sur les deux déviations qualifie le modèle à distribution log normale au détriment de celui à distribution Gamma. Par conséquent nous continuerons l'étude uniquement sur le modèle à distribution log – normal pour vérifier la significativité des variables explicatives et tester la normalité des résidus.

Figure 22 : GLM avec la Normale :

```

The GENMOD Procedure
  Informations sur le modèle
  Data Set          STAT.CHARGE1
  Distribution       Normal
  Link Function     Identity
  Dependent Variable      cout

  Number of Observations Read      942460
  Number of Observations Used      942460
  
```

3) Test de significativité de WALD

Figure 23 : test de significativité de Wald pour le GLM avec la loi Normale

Analyse des résultats estimés de paramètres

Paramètre	DF	Estimation	Erreur standard	Wald 95Limites de confiance %		Khi 2	Pr > Khi 2
Intercept	1	10.2861	0.0135	10.2598	10.3125	583720	<.0001
NÂge	1	0.0706	0.0130	0.0451	0.0961	29.45	<.0001
NÂge	2	0.0000	0.0000	0.0000	0.0000	.	.
NVEH	1	-0.1879	0.0151	-0.2175	-0.1582	153.96	<.0001
NVEH	2	-0.1657	0.0129	-0.1910	-0.1403	164.21	<.0001
NVEH	3	0.0000	0.0000	0.0000	0.0000	.	.
NZone	1	-0.5481	0.0149	-0.5773	-0.5188	1349.55	<.0001
NZone	2	-0.1485	0.0158	-0.1795	-0.1175	87.91	<.0001
NZone	3	-0.0162	0.0267	-0.0685	0.0361	0.37	0.5433
NZone	4	0.0000	0.0000	0.0000	0.0000	.	.
Scale	0	1.0000	0.0000	1.0000	1.0000	.	1.0000

NOTE: The scale parameter was held fixed.

Commentaire :

Le test de Wald montre que toutes les variables explicatives contribuent significativement à l'explication des couts de sinistres. Néanmoins il met en évidence la nécessité de regrouper les deux zones : peur risquée et très peu risqué dans une seule modalité. Ceci fait, le résultat du test devient comme suit :

Figure 24 : test de significativité de Wald pour le GLM avec la loi Normale après regroupement

Analyse des résultats estimés de paramètres

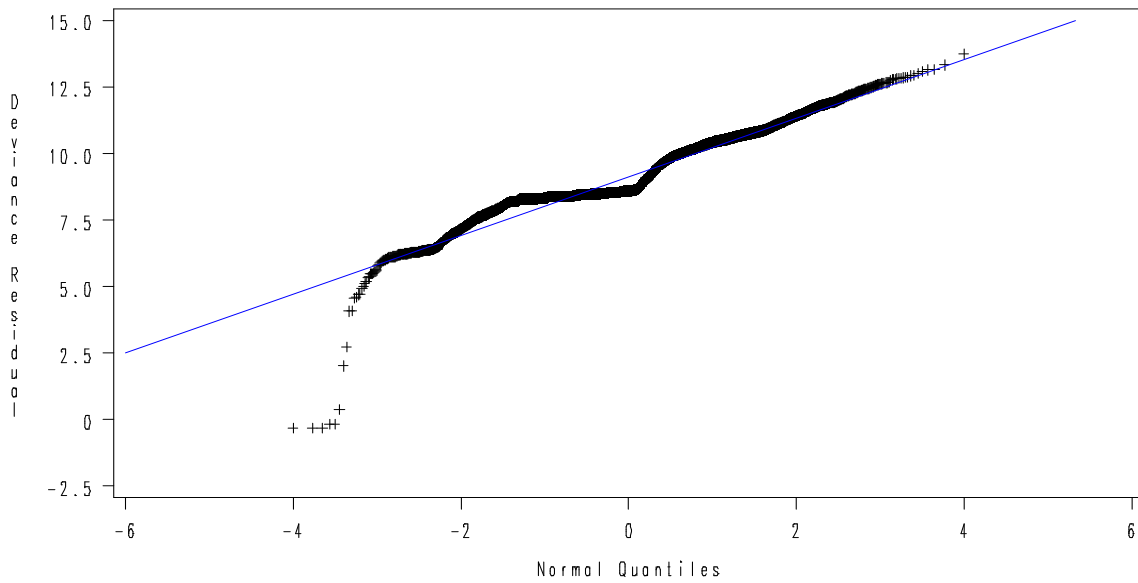
Paramètre	DF	Estimation	Erreur standard	Wald 95Limites de confiance %		Khi 2	Pr > Khi 2
Intercept	1	10.2828	0.0123	10.2587	10.3069	699543	<.0001
NÂge	1	0.0708	0.0130	0.0453	0.0963	29.59	<.0001
NÂge	2	0.0000	0.0000	0.0000	0.0000	.	.
NVEH	1	-0.1876	0.0151	-0.2173	-0.1579	153.66	<.0001
NVEH	2	-0.1655	0.0129	-0.1909	-0.1402	163.95	<.0001
NVEH	3	0.0000	0.0000	0.0000	0.0000	.	.
NZone	1	-0.5449	0.0140	-0.5724	-0.5175	1518.57	<.0001
NZone	2	-0.1453	0.0149	-0.1746	-0.1160	94.53	<.0001
NZone	3	0.0000	0.0000	0.0000	0.0000	.	.
Scale	0	1.0000	0.0000	1.0000	1.0000	.	1.0000

NOTE: The scale parameter was held fixed.

4) Analyse des résidus :

Pour tester la normalité des résidus de déviance et par suite s'assurer de la conformité du modèle aux hypothèses nous traçons le proba-plot de ce dernier sous SAS. La sortie est la suivante.

Figure 25 : proba-plot des résidus pour la loi normale dans le cas d GLM sur le coût moyen



Commentaire :

Les résidus se montrent parfaitement ajustés à une loi normale ce qui nous rassure que le modèle ne contredit pas les hypothèses du GLM.

Conclusion de la première section :

En guise de conclusion, les lois utilisées pour la modélisation de la fréquence et du coût moyen par le modèle linéaire généralisé sont respectivement la loi Binomiale Négative et la loi Log-Normale.

La prime de chaque segment sera donc calculée comme le produit du coût moyen et de la fréquence. Ainsi, les modèles finaux s'écrivent de la façon suivante :

- Pour le coût moyen :

$$\text{Coût moyen}_{Uijk} = \exp(\text{intercept} + \beta_{\text{âge } i} + \beta_{\text{véhicule } j} + \beta_{\text{zone } k})$$

Où

- U prend les valeurs (1, 2, 3) pour les usages.
 - i prend les valeurs (1, 2) pour les classes des âges.
 - j prend les valeurs (1, 2, 3) pour les classes des véhicules.
 - k prend les valeurs (1, 2, 3) pour les classes des véhicules.
- Pour la fréquence :

$$\text{fréquence}_{Uijk} = \exp(\text{intercept} + \beta_{\text{âge } i} + \beta_{\text{véhicule } j} + \beta_{\text{zone } k})$$

Où

- U prend les valeurs (1, 2, 3) pour les usages.
- i prend les valeurs (1, 2) pour les classes des âges.
- j prend les valeurs (1, 2, 3) pour les classes des véhicules.
- k prend les valeurs (1, 2, 3, 4) pour les classes des zones.

- Le calcul de la prime s'écrit donc:

$$\text{prime}_{Uijk} = \text{fréquence}_{Uijk} * \text{coût moyen}_{Uijk}$$

Section II

La crédibilité sur les flottes automobiles d'AAM

- Les différents modèles de la théorie de crédibilité
- L'application de la crédibilité au portefeuille Flottes Automobile AAM
- Les corrections à postériori

Introduction :

Dans un portefeuille d'assurance, certains présentent un profil plus dangereux que d'autres. Réclamer la même prime pour tout le monde pourrait donc paraître inéquitable car cela induirait nécessairement en la sur-tarification de certains assurés. Une solution pour atténuer l'hétérogénéité du portefeuille est de le partitionner en classes de risques homogènes. A cet effet, il subsistera le plus souvent une certaine hétérogénéité au sein de chaque classe. Il est donc naturel d'utiliser la sinistralité relative à un individu et donc à son historique pour réévaluer le montant de sa cotisation.

Dans cette section, nous cherchons à évaluer la rentabilité des contrats du portefeuille des flottes automobile matérialisée par le S/P (sinistre/prime). Nous aboutissons à deux cas de figure:

- Juger la rentabilité d'un contrat à partir de celle du groupe auquel il appartient. Cette méthode présente des inconvénients vu l'hétérogénéité présente au sein des classes. Certaines polices peuvent donc s'écarter de façon nette du comportement moyen de la classe. Leur donner un S/P collectif n'est pas adéquat.
- La deuxième approche consiste à juger le contrat par son historique propre, cependant, le problème de la volatilité du S/P se présente dans ce cas.

La logique derrière la théorie de la crédibilité est de chercher une solution intermédiaire entre les deux approches. A chaque contrat on associe un facteur Z appelé crédibilité qui mesure la fiabilité de son expérience propre ($0 \leq Z \leq 1$). La rentabilité d'un contrat donné s'écrit alors comme suit :

$$S/P_{retenu} = Z * S/P_{retenu} + (1 - Z) * S/P_{groupe}$$

Ainsi, si $Z = 0$, l'information du contrat est jugée comme absolument non fiable, par conséquent il faut tenir compte uniquement des résultats du groupe.

Si $Z = 1$, l'expérience du groupe est jugée parfaitement non fiable alors on se base uniquement sur l'expérience propre.

Il reste à savoir donc comment déterminer le facteur Z . Plusieurs critères paraissent naturels :

- Plus l'expérience d'un contrat est grande et plus il paraît crédible.
- certains contrats ont des résultats plus stables que d'autres et doivent donc être d'avantage crédibilisés.

Dans le cas des flottes d'entreprises, la taille de la flotte reste le facteur majeur influant la stabilité du ratio S/P. En effet, le nombre de sinistres varie peu dans le cas des grosses flottes et le coût moyen est plus stable vu le grand nombre de sinistres.

Notre étude donc devrait répondre à ces questions. Ainsi, nous allons en premier lieu présenter les différents modèles de crédibilité, ensuite le point sera mis sur la méthodologie du calcul des facteurs de crédibilité pour chaque modèle. Finalement nous allons analyser les résultats fournis par les différents modèles pour évaluer la rentabilité des contrats de notre portefeuille.

Les modèles de crédibilité :

Le premier modèle américain de crédibilité ‘ méthode des fluctuations limitées’ avait pour objet de décider si un contrat devait ou non être jugé à partir de son propre historique (i.e. $Z=1$) comme elle permet de gérer le cas de la crédibilité partielle (i.e. $0 < Z < 1$). Néanmoins cette méthode manque de fondements théoriques robustes et elle est par suite la moins utilisée de nos jours. L’approche actuelle repose sur le modèle établi par BUHLMAN en 1967, le modèle européen reste toutefois frustré vu qu’il suppose tous les contrats sont identiques à priori, il a été donc amélioré de manière à prendre en considération la différence de poids entre les contrats : c’est le modèle de BUHLMAN-STRAUB. Les contrats restent identiques à un facteur de taille près. En fin les modèles hiérarchiques introduits par JEWELL visent la prise en considération de la subdivision du portefeuille en catégories et sous catégories et résout ainsi le problème des contrats identiques.

Dans cette partie, nous présentons en premier lieu les différents modèles de crédibilité, nous les appliquons au portefeuille des flottes automobile en vue de comparer les résultats et opter finalement pour le modèle qui semble le plus approprié.

I. Modèle de BUHLMAN-STRAUB :

1. La description du modèle

En 1967, un premier modèle semi paramétrique a été proposé par BUHLMAN. Le modèle est fondé sur des bases mathématiques robustes et rigoureuses, néanmoins il présentait un défaut : il ignorait la différence entre les contrats en termes de taille. en 1970, Buhlman et STRAUB ont travaillé pour promouvoir le modèle et inclure l’information liée à la taille des contrats. Le modèle obtenu est décrit comme suit :

Soit un portefeuille de N contrats observés sur une période de T années. On note X_{jt} la grandeur (objet de l’étude) observé au niveau du $j^{ième}$ contrat à l’année t. cette grandeur peut éventuellement être le nombre de sinistres, le coût des sinistres, le taux de destruction...

Ici on s’intéresse au ratio de sinistralité S/P. on note alors ω_{jt} le poids associé à l’observation X_{jt} . Le poids ω_{jt} peut être déterminé selon plusieurs critères lié à la taille du contrat, à la sinistralité ou à l’exposition. De plus on associe à chaque contrat un paramètre de risque θ , la loi des X_{jt} est alors déterminée par θ et ω . néanmoins le paramètre θ est évidemment inobservable d’où la nécessité de supposer que la loi des X_{jt} ne varie pas dans le temps.

Hypothèses du modèle :

Le modèle de BUHLMAN-STRAUB repose sur un ensemble d’hypothèses, on les résume comme suit :

Hyp1 : Les variables X_{jt} sont de carrés intégrables (i.e. A variances finies).

Hyp2 : Les contrats du portefeuille sont deux à deux indépendants, mathématiquement cela est traduit par :

Les vecteurs $(\theta_j, X_{jt} \geq 0)$ sont indépendants.

Hyp3 : à priori, tous Les contrats ont le même risque, mathématiquement :

Les vecteurs $(\theta_j, X_{jt}, t \geq 0)$ ont la même loi de distribution.

Hyp4 La valeur moyenne ne dépend que du paramètre de risque :

$E(X_{jt} | \theta_j = \theta)$ est indépendante de j et de t .

Hyp5 : La variance du risque à θ fixé est inversement proportionnelle au poids :

$V(X_{jt} | \theta_j = \theta) = \frac{\sigma(\theta)^2}{\omega_{jt}}$; avec $\sigma(\theta)^2$ est indépendante de j et de t .

Hyp6 : conditionnellement au paramètre de risque, les différentes observation d'un même contrat sont indépendantes, autrement:

Conditionnellement à θ_j , les variables $X_{jt}, t \geq 0$ sont non corrélées.

Avant d'énoncer le théorème de Buhlman-Straub, il convient d'introduire un ensemble de notations qui nous seront utiles lors de l'étude qui suit.

- $\mu_0 = E(X_{jt})$. C'est le S/P à priori, autrement dit c'est le S/P que l'on doit attendre d'un contrat dont on ne dispose pas de données.
- $\mu(\theta) = E(X_{jt} | \theta_j = \theta)$. C'est le S/P probable d'un contrat de paramètre θ . Néanmoins on ne connaît pas en pratique le paramètre, par conséquent, on ne connaît pas non plus $\mu(\theta)$. On cherche donc à l'estimer.
- $a = V(E(X_{jt} | \theta))$. C'est la variance inter contrats. elle mesure l'hétérogénéité du portefeuille. en matière de crédibilité, plus cette variance est importante plus les contrats risquent de s'éloigner nettement de la moyenne du portefeuille, plus il faudrait donc se fier aux statistiques propres du contrat.
- $S^2 = E(V(X_{jt} | \theta))$. C'est la variance intra contrat Elle mesure la fluctuation du ratio S/P dans le temps. Ainsi, Plus S^2 est important, plus les S/P individuels sont volatils. l'expérience individuelle paraît donc peu fiable et on doit se fier plus à l'information collective.

On cherche à travers la méthode de crédibilité à déterminer le meilleur estimateur

\hat{X}_j possible du ratio de sinistralité S/P de chaque contrat j donc le meilleur estimateur de $\mu(\theta_j)$.

Le meilleur estimateur au sens quadratique est le X_j mesurable par rapport aux variables $X_{jt}, t < T$ qui minimise la quantité : $E((X_{jT+1} - X_j)^2 | X_{j1} \dots X_{jT})$.

La solution de ce problème est l'estimateur de Bayes $E(X_{jT+1} | X_{j1} \dots X_{jT})$. Néanmoins cet estimateur ne peut pas être calculé en pratique. Par conséquent il faut restreindre davantage la forme de l'estimateur. Deux approches sont possibles :

- La crédibilité homogène : consiste à chercher l'estimateur sous forme d'une fonction linéaire des observations c'est-à-dire du type : $\sum \alpha_{it} X_{it}$
- La crédibilité non homogène : vise à déterminer l'estimateur comme fonction affine de la forme :

$$\alpha_0 + \sum \alpha_{it} X_{it}$$

Sous les hypothèses du modèle de Buhlman-Straub, ces estimateurs prennent une forme plus simple, c'est l'objet du théorème qui suit :

Théorème : estimateurs de Buhlman-Straub

Sous les hypothèses (HYP1) à (HYP6), les estimateurs de Buhlman-Straub sont donnés \bar{X} par :

Cas homogène : $\hat{X}_j = Z_j \bar{X}_j + (1 - Z_j) \mu_0$.

Cas non homogène : $\hat{X}_j = Z_j \bar{X}_j + (1 - Z_j) \bar{X}$.

Un problème se pose : les estimateurs dépendent des paramètres structuraux a et s^2 , qui sont inconnus. Par conséquent, Il faut les estimer tout d'abord. Plusieurs estimations sont possibles. Ici on choisit l'estimateur classique de VYLLDER donné comme suit :

$$\hat{s}^2 = \frac{1}{N(T-1)} \sum_{it} (X_{it} - \bar{X}_i)^2$$

$$\hat{a} = \frac{1}{N(T-1)} \sum_i Z_j (\bar{X}_i - X)^2$$

L'estimation par \hat{s}^2 paraît simple, par contre celle des paramètres a et S^2 n'est pas évidente vu qu'elle fait intervenir les coefficients de crédibilité Z_j qui eux même dépendent des paramètres a et s^2 . Ainsi une procédure itérative est envisageable pour résoudre le problème.

Les étapes de la procédure se résument comme suit :

- 1) déterminer \hat{s}^2 .
- 2) choisir arbitrairement des valeurs μ^0 et a^0 pour a et μ_0 .
- 3) en tirer les coefficients de crédibilités Z_i .
- 4) en déduire les valeurs de $\hat{\mu}_0$ et \hat{a} .

On fait finalement des réitérations à partir de l'étape 3.

2. Application aux données :

L'application du modèle a été réalisée sous le logiciel R. en premier lieu on calcule les crédibilités des contrats dans le portefeuille par année, le code du programme est le suivant :

```
> fitt <- cm(~annee, data=x, ratios=ratio.1:ratio.543, weights=weight.1:weight.543,
method="iterative")
> summary(fitt)
Call:
cm(formula = ~annee, data = x, ratios = ratio.1:ratio.543, weights = weight.1:weight.543,
  method = "iterative")
```

Les résultats obtenus se présentent ainsi :

Figure 26 : S/P annuels crédibilisés avec le modèle de Bühlmann Straub

Detailed premiums				
Level: annee				
annee	Indiv. mean	Weight	Cred. Factor	Cred. premium
2009	0.5074080	0.3966889	0.5177991	0.4757190
2010	0.4422026	0.4328125	0.5395121	0.4419668
2011	0.3824484	0.5081772	0.5790572	0.4073860

↓ ↓
 Facteur de S/P crédibilisé
 Crédibilité

Commentaire :

On constate que les facteurs de crédibilités croient d'une année à l'autre. Cela s'explique par l'effet de l'ancienneté des contrats. En effet plus un contrat est ancien plus son expérience est importante et plus donc il est crédible. D'un autre coté on constat que la rentabilité augmentent d'une année à l'autre. Cela est traduit par la diminution du ratio de sinistralité d'une année à l'autre.

Ici on applique le modèle de Buhlman- Straub pour obtenir les coefficients de crédibilité de chaque contrat du portefeuille. le code R utilisé est le suivant :

```
fit1 <- cm(~police, data=x, ratios=ratio.1:ratio.3, weights=weight.1:weight.3)
> summary(fit1)
Call:
cm(formula = ~police, data = x, ratios = ratio.1:ratio.3, weights = weight.1:weight.3)
```

Nous rappelons que le nombre des contrats flotte de notre portefeuille s'élève à 1500 contrats. De ce fait nous ne présenterons pas la totalité des facteurs de crédibilité calculés par le modèle.

Les sorties du programme R se présentent comme suit :

Figure 27 : aperçu des S/P contrats par contrats crédibilisés avec le modèle de Bühlmann Straub

Detailed premiums				
Level: police				
police	Indiv. mean	Weight	Cred. factor	Cred. premium
*****	0.334307	0.08482288	0.833763527	0.345997
*****	0.382025	0.06826132	0.8014396404	0.386514
*****	0.615382	0.06338326	0.7893768008	0.570993
*****	0.203823	0.03603024	0.6805567836	0.26797
*****	0.347917	0.003641559	0.1771736519	0.394582
*****	0.945034	0.003637217	0.1769997751	0.500282

Afin de pouvoir commenter les résultats on présente ci-dessous quelques statistiques sur les crédibilités des contrats

Tableau 32 : les facteurs de crédibilité : minimal, maximal et moyen / Bühlmann Straub

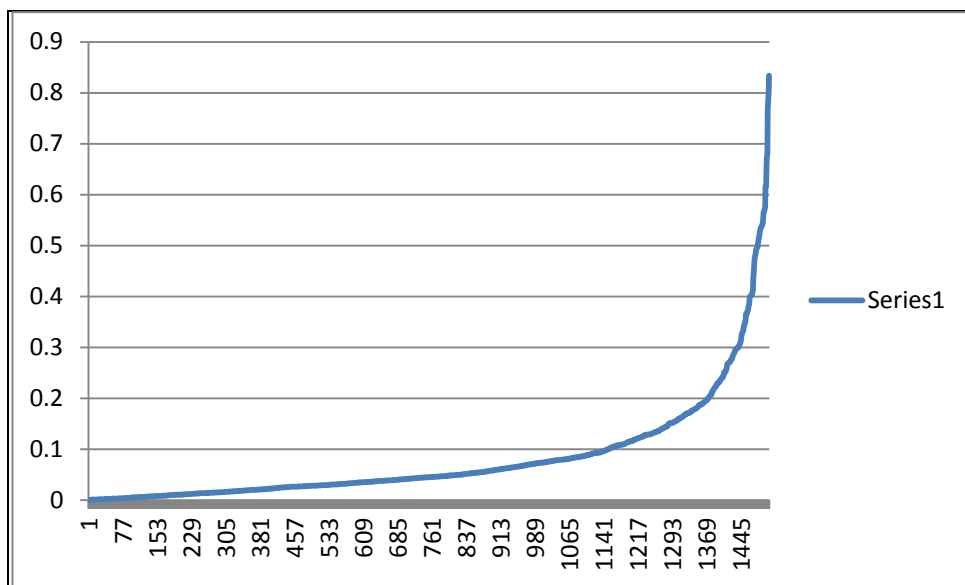
Crédibilité minimale (Z_{min})	0,00016
Crédibilité maximale (Z_{max})	0,833
Crédibilité moyenne (Z_{moyen})	0,16

Commentaires et interprétations :

Les contrats paraissent moyennement crédibles et l'étendue de la variable Z est très grande. Cela met en évidence l'existence d'un paramètre qui différencie les contrats et agit sur leur crédibilité. Lors du choix des poids, la taille et l'ancienneté des flottes sont jugés comme étant les facteurs majeurs qui différencient plus entre les flottes. De ce fait on propose de visualiser la variation des facteurs de crédibilité en fonction de la taille des flottes (matérialisée par leur prime) afin de trouver une explication à la grande étendue de la variable Z.

Le graphique suivant représente la variation de la crédibilité des flottes en fonction de leur prime :

Figure 28 : la variation de la crédibilité des flottes en fonction de leur prime



On constate que la crédibilité des flottes croit avec la prime, elle tend vers 0 pour les petites flottes et dépasse 80% pour les flottes de grande taille. Ce qui explique l'étendue importante de la variable Z. les résultats obtenus par le modèle de Buhlmann-Straub mettent en évidence la nécessité de subdiviser le portefeuille des flottes.

II. Le modèle hiérarchique :

Deux motivations majeures ont donné naissance aux modèles hiérarchisés de crédibilité, la première est de remédier à l'inconvénient des contrats identiques que le modèle Buhlman-Straub n'a pas réussi à résoudre amplement. La deuxième étant l'établissement d'un modèle spécifique au portefeuille des flottes automobile qui s'adapte et convient aux spécificités de ce type de portefeuille. En effet, les polices couvrant ce type de flotte comportent plusieurs niveaux : les flottes et les véhicules au sein de ces flottes. Ainsi, les

modèles hiérarchisées sont développés afin de séparer les effets aléatoires liés à la flotte et ceux propres à chaque véhicule au sein de celle-ci.

1. Description du modèle

Sur une période de T années, On observe un portefeuille de N contrats. Le portefeuille est divisé en M classes C_i . Celles-ci sont divisées en M_i catégories notées C_{ij} . On s'intéresse au ratio de sinistralité S/P de chaque contrat du portefeuille en question. Ainsi, on note X_{ijkt} le S/P du contrat C_{ijk} de la catégorie C_{ij} observé pendant l'année t. A chaque observation X_{ijkt} , on associe un poids W_{ijkt} . On suppose de plus que chaque niveau de subdivision du portefeuille possède un paramètre de risque. Celui de la classe C_i est noté θ_i ; celui de la catégorie C_{ij} est noté θ_{ij} et finalement celui du contrat C_{ijk} est noté θ_{ijk} . Les paramètres θ et W déterminent la loi des observations X_{ijkt} . on suppose à nouveau que la loi du ratio S/P ne varie pas au cours du temps.

Les hypothèses du modèle hiérarchisé se résument comme suit :

- (H1) : Les variables X_{ijkt} sont de carrés intégrables.
- (H2) : Les classes du portefeuille sont deux à deux indépendantes.
- (H3) : Les catégories d'une même classe n'influent l'une sur l'autre que par le paramètre de risque de la classe. Mathématiquement, cela se traduit par :
 - Les vecteurs $(\theta_{ij}, \theta_{ijk} \text{ et } X_{ijkt} / t \geq 0)$ sont indépendants conditionnellement à θ_i .
- (H4) : Les différents contrats d'une même catégorie n'influent l'un sur l'autre également que par le paramètre de risque de la catégorie en question.

Mathématiquement :

 - Les vecteurs $(\theta_{ijk}, X_{ijkt} / t \geq 0)$ sont indépendants conditionnellement à θ_{ij} .
- (H5) : Les différentes observations d'un même contrat n'influent l'une sur l'autre que par le paramètre de risque du contrat ; cela se traduit par :
 - Les variables $(X_{ijkt}, t \geq 0)$ sont deux à deux indépendantes conditionnellement à θ_{ijk} .
- (H6) Les contrats ont a priori le même niveau de risque. Formellement :
 - Les vecteurs $(\theta_i, \theta_{ij}, \theta_{ijk}, X_{ijkt} / t \geq 0)$ suivent a priori la même loi.
- (H7) La valeur moyenne des observations ne dépend que du paramètre de risque, i.e. :
 - $E(X_{ijkt} / \theta_{ijk} = \theta) = \theta$ On le note $\mu_3(\theta)$.
- (H8) à θ fixé, La variance du risque est inversement proportionnelle au poids. i.e. :
 - $V(X_{ijkt} / \theta_{ijk} = \theta) = \frac{\sigma^2(\theta)}{W_{ijkt}}$.

On présentera à ce stade quelques notations qui nous seront utiles dans ce qui suit :

- $\mu_0 = E(X_{ijkt})$. C'est le S/P à priori, autrement dit c'est le S/P que l'on doit attendre d'un contrat dont on ne dispose pas de données.
- $\mu_1(\theta) = E(X_{ijkt} / \theta_i = \theta)$ C'est la valeur à attribuer à un contrat si l'unique information dont on dispose sur ce dernier est son appartenance à une classe de paramètre θ .
- $\mu_2(\theta) = E(X_{ijkt} / \theta_{ij} = \theta)$ C'est la valeur que l'on doit attribuer à un contrat si l'unique information disponible sur ce dernier est son appartenance à une catégorie de paramètre θ .
- $\mu_3(\theta) = E(X_{ijkt} / \theta_{ijk} = \theta)$ C'est la valeur que l'on doit attribuer à un contrat dont le paramètre de risque est θ . C'est la quantité que le modèle vise estimer car elle correspond au ratio S/P probable du contrat.

Avant d'énoncer le théorème du modèle de Jewell il serait convenable de définir en premier lieu les paramètres structuraux du modèle.

- On note $S^2 = E(\sigma^2(\theta_{ijk}))$ c'est la variance due aux fluctuations dans le temps.
 - On note $a = E(V(\mu(\theta_{ijk}/\theta_{ij})))$ c'est la variance intra-contrat, elle mesure donc l'hétérogénéité entre les contrats de la même catégorie.
 - On note $b = E(V(\mu(\theta_{ij}/\theta_i)))$. C'est la variance inter-catégorie. Elle mesure donc l'hétérogénéité entre les différentes catégories des classes.
 - On note $c = V(E(\mu(\theta_{ij})))$: la variance interclasse.
- On rappelle alors que : $V(X_{ijkt}) = a + b + c + S^2$

Attribution des poids aux différents éléments du portefeuille :

Chaque observation a déjà été munie d'un poids W_{ijkt} . Ainsi :

- Chaque contrat est muni d'un poids égal à la somme des poids de ses observations sur la période $[0, T]$: $W_{ijk} = \sum_t W_{ijkt}$.
- Chaque catégorie est munie d'un poids W_{ij} égal à la somme des crédibilités de ses poids.
- Chaque classe est munie d'un poids W_i égal à la somme des crédibilités de ses catégories.
- Finalement Le portefeuille est muni d'un poids w égale à la somme des crédibilités de ses classes.

Attribution des crédibilités aux différents éléments du portefeuille :

- ✓ La crédibilité d'un contrat C_{ijk} :

$$Z_{ijk} = \frac{W_{ijk}}{W_{ijk} + \frac{S^2}{a}}$$

Cette formule illustre le constat qu'un contrat est d'autant plus crédible (i.e. Z_{ijk} grand) qu'il est stable dans le temps (i.e. S^2 faible) et que les catégories sont hétérogènes (i.e. a grand).

- ✓ La crédibilité d'une catégorie C_{ij} :

$$Z_{ij} = \frac{W_{ij}}{W_{ij} + \frac{a}{b}}$$

Une catégorie est donc d'autant plus crédible qu'elle est homogène (a faible) et que les classes de cette catégorie sont hétérogènes (b grand).

- ✓ La crédibilité d'une classe C_i :

$$Z_i = \frac{W_i}{W_i + \frac{b}{c}}$$

Une classe est ainsi d'autant plus crédible qu'elle est homogène (b faible) et que le portefeuille est hétérogène (c grand).

Finalement on définit les moyennes pondérées à chaque

→ La moyenne du contrat C_{ijk} :

$$\bar{X}_{ijk} = \frac{1}{W_{ijk}} \sum_t W_{ijkt} X_{ijkt}$$

→ La moyenne de la catégorie C_{ij} :

$$\bar{X}_{ij} = \frac{1}{W_{ij}} \sum_k W_{ijk} \bar{X}_{ijk}$$

→ La moyenne de la classe C_i :

$$\bar{X}_i = \frac{1}{W_i} \sum_j W_{ij} \bar{X}_{ij}$$

→ La moyenne du portefeuille:

$$\bar{X} = \frac{1}{W} \sum_i W_i \bar{X}_i$$

A ce stade, nous disposons de tous les ingrédients pour déterminer les meilleures estimations des ratios de sinistralité à chaque niveau du portefeuille. Les estimations fournies par le modèle de Jewell se présentent comme suit :

Théorème : modèle de Jewell homogène:

L'estimateur de crédibilité homogène de la classe i est donné par :

$$\hat{X}_i = Z_i \bar{X}_i + (1 - Z_i) \mu_0$$

L'estimateur de crédibilité homogène de la catégorie j est donné par :

$$\hat{X}_{ij} = Z_{ij} \bar{X}_{ij} + (1 - Z_{ij}) \hat{X}_i$$

L'estimateur de crédibilité inhomogène du contrat k est donné par :

$$\hat{X}_{ijk} = Z_{ijk} \bar{X}_{ijk} + (1 - Z_{ijk}) \hat{X}_{ij}$$

Théorème : modèle de Jewell inhomogène:

L'estimateur de crédibilité inhomogène de la classe i est donné par :

$$\hat{X}_i = Z_i \bar{X}_i + (1 - Z_i) \bar{X}$$

L'estimateur de crédibilité inhomogène de la catégorie j est donné par :

$$\hat{X}_{ij} = Z_{ij} \bar{X}_{ij} + (1 - Z_{ij}) \hat{X}_i$$

L'estimateur de crédibilité inhomogène du contrat k est donné par :

$$\hat{X}_{ijk} = Z_{ijk} \bar{X}_{ijk} + (1 - Z_{ijk}) \hat{X}_{ij}$$

Comme dans le modèle de Buhlman-Straub, il faut estimer les paramètres structuraux a, b, c, μ_0 et s^2 . On résume les estimations des différents paramètres dans le tableau suivant :

Tableau 31 : Estimations des différents paramètres du modèle de Jewell

paramètre	estimateur	Nature de l'estimateur
a	$\frac{1}{\sum_{ij}(\text{card}(C_{ij}) - 1)} \sum_{ijk} (\bar{X}_{ijk} - \bar{X}_{ij})^2$	Non Biaisé consistant
b	$\frac{1}{\sum_i(M_i - 1)} \sum_{ij} (\bar{X}_{ij} - \bar{X}_i)^2$	Non Biaisé consistant
C	$\frac{1}{N(T - 1)} \sum_{ijk} (\bar{X}_{ijkt} - \bar{X})^2$	Non Biaisé consistant
μ_0	\bar{X}	Non biaisé consistant
s^2	$\frac{1}{M - 1} \sum_{ijk} (\bar{X}_i - \bar{X}_{ijk})^2$	Non biaisé consistant

Comme dans le modèle de Buhlman Straub, le recours à un algorithme itérative est inévitable vu que les coefficient de crédibilité sont déterminés à partir des poids qui sont eux même déterminés à partir des crédibilité. Le processus se présente comme suit :

1. fixer arbitrairement a, b et c
2. en déduire les crédibilités à chaque niveau
3. en déduire les poids à chaque niveau
4. en déduire les moyennes à chaque niveau
5. en déduire les estimateurs \hat{a} , \hat{b} , \hat{c}
6. réitérer à partir de la deuxième étape jusqu'à la convergence de l'algorithme.

2. Application aux données :

Comme son nom l'indique, le modèle hiérarchique suppose une subdivision du portefeuille en des niveaux (catégorie, classe,...).

Deux critères de subdivision sont jugés pertinents : l'usage des véhicules de la flotte et la taille de la flotte matérialisée par sa prime.

Subdivision du portefeuille par usage :

Le premier niveau de subdivision consiste à différencier les flottes par l'usage de leurs véhicules. Pour ce faire, nous avons calculé le nombre de véhicules de chaque flotte selon les usages : Tourisme, commercial inférieur à 3 tonnes et demi et commercial supérieur à trois tonnes et demi. Ainsi une flotte sera jugée à usage tourisme si la majorité de ses véhicules sont de types tourisme.

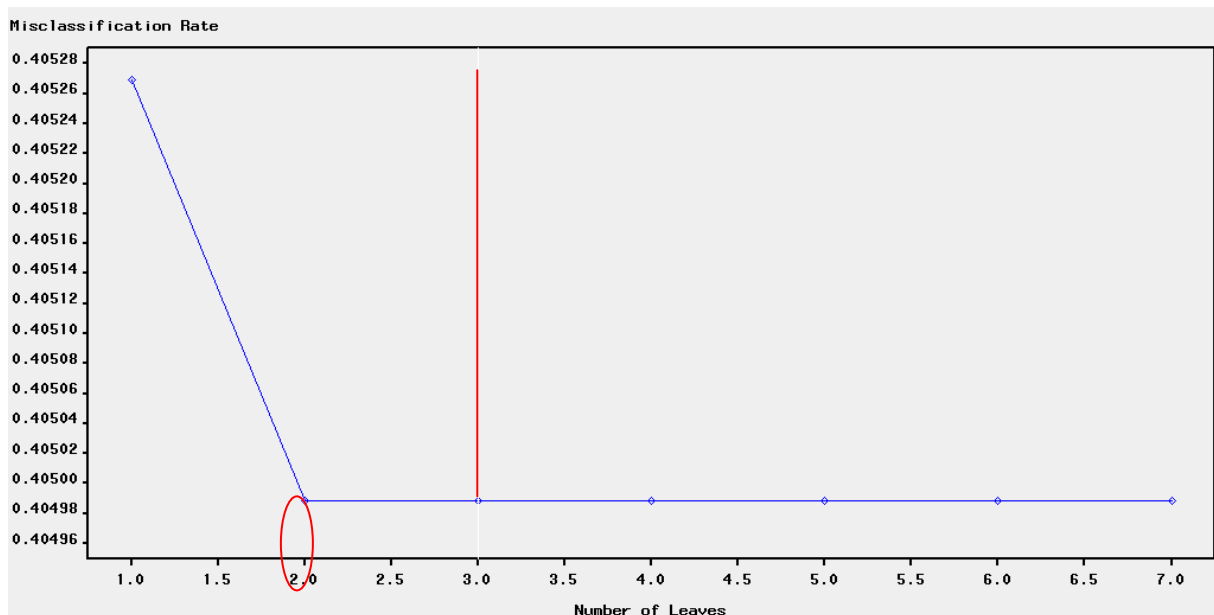
Subdivision du portefeuille par classe de cotisation :

La démarche de subdivision du portefeuille flotte est la même que celle adopté lors de la classification automatique détaillée dans la première section, l'objectif étant de construire des classes homogène en termes de cotisation. Techniquement parlant, on cherche des classes de cotisations à variance intra classe minimale et inter classes maximale.

La subdivision est réalisée sous le logiciel SAS Enterprise Miner Tools tree, le résultat est comme suit :

Le graphique suivant représente la variation de l'erreur de classification en fonction du nombre de subdivision.

Figure 29 : la variation de l'erreur de classification en fonction du nombre de subdivision

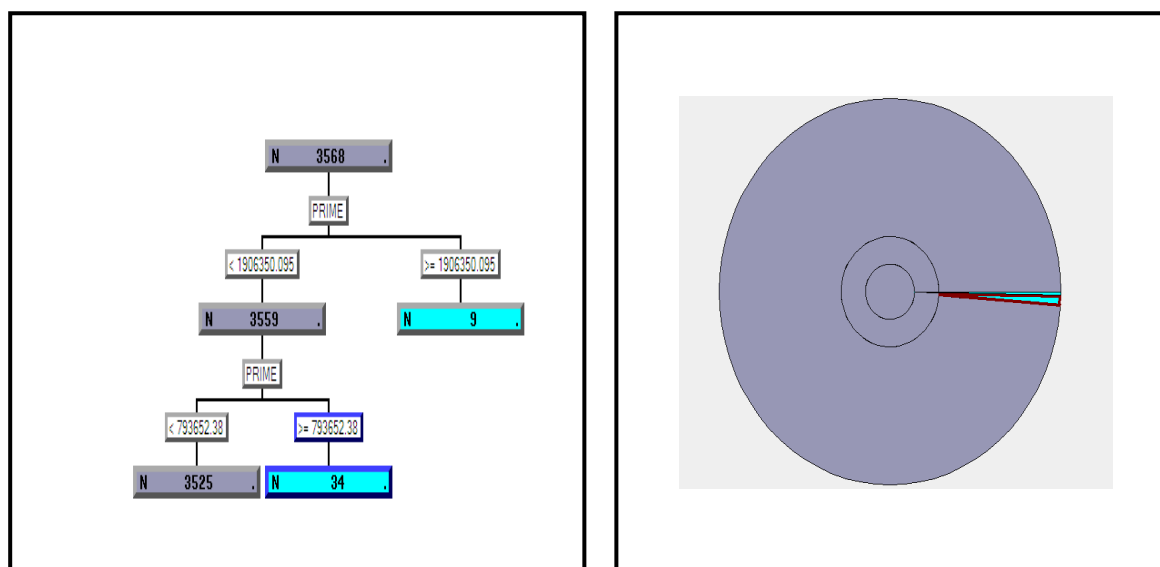


Commentaire :

Nous cherchons le niveau de subdivision qui minimise l'erreur de la classification. D'après le graphique nous constatons que ce taux d'erreur est minimisé au niveau de la deuxième subdivision. En effet, après la deuxième subdivision ce taux ne décroît plus et devient stable. Nous en concluons donc que notre portefeuille sera subdivisé en trois classes de cotisations.

C'est ce que nous allons visualiser via l'arbre suivant :

Figure 30 : la subdivision du portefeuille en classes de cotisation



L'arborisation a effectivement mis en évidence les trois classes de cotisations suivantes :

Tableau 33 : les classes de cotisation

Classe de cotisations en DH	Nombre de flottes
< 793652.38	3525
[793652.38 ; 1906350.095[34
>=1906350.095	9

A ce stade, le premier et le deuxième niveau sont obtenus.

Le modèle hiérarchique à un seul niveau :

Ici nous calculons les facteurs de crédibilité ainsi que les ratios de sinistralité estimés pour un seul niveau de subdivision. La subdivision est par catégorie d'usage.

On estime le S/P d'une catégorie par :

$$\widehat{S/P}_{catégorie} = Z_{cat}S/P_{catégorie} + (1 - Z_{cat})S/P_{portefeuille}$$

L'application du modèle de Jewell est réalisée sous R ; on présente le code du programme suivi des outputs obtenus :

```
> fit4 <- cm(~cat, data=z, ratios=ratio.1:ratio.3179, weights=weight.1:weight.3179,
method="iterative")
> summary(fit4)
Call:
cm(formula = ~cat, data = z, ratios = ratio.1:ratio.3179, weights = weight.1:weight.3179,
method = "iterative")
```

Les facteurs de crédibilité et les ratios estimés des catégories :

Figure 31 : Les S/P par catégorie crédibilisés avec le modèle de Jewell à un niveau

Detailed premiums					
Level: cat					
cat	Indiv. mean	Weight	Cred. factor	Cred. premium	
tourisme	0.389471	2.79501683	0.78967262	0.394215	
cominf	0.457725	0.05824855	0.07256628	0.415342	
comsup	0.500050	0.14675917	0.16467496	0.426521	

Commentaire :

Les résultats du modèle confirment qu'il existe bien un effet se rapportant à la catégorie. On remarque que l'usage Tourisme a un facteur de crédibilité important et par conséquent on se base en grande partie sur son expérience propre. En effet, les outputs montrent que les facteurs de crédibilité changent avec l'usage, ainsi l'usage tourisme possède la plus grande crédibilité soit de 0.78, suivi de l'usage commercial supérieur à 3.5 T avec une crédibilité 0.16 et enfin l'usage commercial inférieur à 3.5T.

Avant d'appliquer le modèle hiérarchique à deux niveaux, il serait judicieux de vérifier si l'ajout d'un deuxième niveau de subdivision à savoir les classes de cotisation influe sur les résultats de la crédibilité sinon l'application du modèle sera de nul intérêt. Pour ce faire, nous proposons de calculer les facteurs de crédibilité des différentes catégories relativement aux classes de cotisation. Ainsi nous pouvons vérifier si le coefficient d'une même catégorie diffère d'une classe de cotisation à l'autre ou pas et par conséquent nous pourrions conclure si la deuxième subdivision est significative ou sans intérêt.

Le résultat obtenu sous R est comme suit :

Figure 32 : les S/P crédibilisés de la catégorie dans la classe de cotisation avec Jewell à 2 niveaux

Level: cat					
classe	cat	Indiv. mean	Weight	Cred. factor	Cred. premium
1	tourisme	0.342694	0.27266126	0.21832833	0.363086
2	tourisme	0.323493	0.31581651	0.24443736	0.352367
3	tourisme	0.404694	2.20653900	0.69328358	0.405630
3	cominf	0.457725	0.05824855	0.05630893	0.410561
2	comsup	0.358258	0.04247552	0.04169691	0.361565
3	comsup	0.557803	0.10428364	0.09651592	0.422229

Commentaires :

Concernant les deux niveaux de subdivision, les résultats montrent que le facteur de crédibilité d'une même catégorie d'usage n'est pas le même dans toutes les classes de cotisation, à titre d'exemple le facteur de crédibilité de l'usage tourisme est d'ordre de 0,21 pour la première

classe de cotisation, dans la deuxième il est de 0,24 alors qu'il prend la valeur de 0,69 sur la troisième classe de cotisation.

On peut donc en conclure que la deuxième subdivision du portefeuille impact les résultats et par conséquent l'application du modèle de Jewell à deux niveaux ne serait pas sans intérêt.

Modèle de Jewell à deux niveaux :

On applique le modèle à deux niveaux au portefeuille afin de déterminer les facteurs de crédibilité relatifs aux contrats par rapport à la catégorie à laquelle ils appartiennent, puis ceux des catégories relativement à la classe de cotisations à laquelle elles appartiennent et finalement les facteurs de crédibilités des classes de cotisation relativement à la totalité du portefeuille.

On rappelle qu'ainsi les ratios de sinistralité de chaque étage s'écrivent ainsi :

$$\begin{aligned}\widehat{S/P}_{classe} &= Z_{classe}S/P_{classe} + (1 - Z_{classe})S/P_{portefeuille} \\ \widehat{S/P}_{catégorie} &= Z_{cat}S/P_{catégorie} + (1 - Z_{cat})\widehat{S/P}_{classe} \\ \widehat{S/P}_{contrat} &= Z_{contrat}S/P_{contrat} + (1 - Z_{cat})\widehat{S/P}_{catégorie}\end{aligned}$$

Le code R est le suivant :

```
> library(actuar)
> X<-cbind(classe=c(1,2,3,3,2,3),x)
> fit<-cm(~classe + classe:cat, data=X, ratios=ratio.1:ratio.3140, weights=weight.1:weight.3140,
method="iterative")
> summary(fit)
```

L'application sous R donne les résultats suivants :

Crédibilité de la classe dans le portefeuille :

Figure 33 : les S/P crédibilisés de la classe au sein du portefeuille par Jewell à 2 niveaux

Level: classe				
Classe	Indiv. Mean	Weight	Cred. factor	Cred. premium
1	0.342694	0.2183283	0.2895169	0.368782
2	0.328559	0.2861343	0.3481297	0.361709
3	0.425688	0.8461084	0.6122823	0.407746

Commentaire :

Les résultats du modèle par rapport au deuxième niveau à savoir les classes de cotisation confirme le constat que plus une flotte est de grande taille (sa cotisation est importante) plus elle est stable, et par conséquent on se base plus sur son expérience propre. En effet, les outputs montrent que les facteurs de crédibilité croient avec la classe de cotisation ; ainsi, la première classe possède la plus petite crédibilité évaluée par 0,28, suivie de la classe de cotisation moyenne avec un coefficient de crédibilité de 0,34 et finalement la plus grande classe possédant la plus grande crédibilité 0,61.

La crédibilité de la catégorie au sein de la classe a été donnée dans la partie au dessus pour mettre en évidence l'utilité d'un deuxième niveau de subdivision.

Crédibilité du contrat dans la catégorie :

Afin d'appliquer la crédibilité de Jewell à deux niveaux contrat par contrat on applique le code R suivant :

```
> fit10<-cm(~cat+ cat:police, data=l, ratios=ratio.1:ratio.3, weights=weight.1:weight.3)
> summary(fit10)
```

Remarque : **cat** est la catégorie dans la classe, par exemple tourisme1 c'est la catégorie tourisme dans la classe de cotisation 1.

Le résultat R est le suivant :

Figure 34 : aperçu des S/P crédibilisés contrat par contrat pour Jewell à 2 niveaux

```
Level: police
cat      police  Individ. mean Weight      Cred. factor  Cred. premium
tourisme1 ***** 0.334307 8.482288e-02 0.9549246368 0.337735
tourisme1 ***** 0.382025 6.826132e-02 0.9445943820 0.383596
tourisme1 ***** 0.615382 6.338326e-02 0.9405836479 0.603201
```

Vu la multitude des contrats on présentera un tableau qui récapitule les caractéristiques générales des facteurs obtenus :

Tableau 34 : les facteurs de crédibilité : minimal, maximal et moyen pour la classe au sein de la catégorie

		Zmin	Zmax	Zmoyen
tourisme	1	0,93348749	0,95492464	0,94339754
	2	0,65475114	0,89998787	0,77542134
	3	0,06590387	0,88603583	0,27051042
commercial> 3TT5	2	0,10635125	0,40997433	0,2249912
	3	0,00678154	0,80368933	0,24842891
commercial< 3T5	3	0,00578521	0,79597591	0,2461178

On constate que les contrats relatifs à l'usage tourisme sont ceux les plus crédibles spécialement ceux relatifs aux deux premières classes de cotisation.

Une manière d'analyser les S/P crédibilisés obtenus est de les comparer au S/P réels des contrats. Ainsi,

- Si $S/P_{créd} < S/P$: alors la prime crédibilisée est supérieure à la prime réelle. Cela veut dire que l'historique dudit contrat montre que ce dernier a fait preuve d'une sinistralité qui était sous estimée à priori. Par conséquent, la prise en considération de cet historique sera traduite par une majoration de la prime du contrat en question
- Si $S/P_{créd} > S/P$: alors la prime crédibilisée est inférieure à la prime réelle. Cela veut dire que l'historique dudit contrat montre que ce dernier a fait preuve d'une sinistralité surestimée à priori. Ainsi la prise en considération de cet historique sera traduite par une réduction de la prime du contrat en question.

Exemple :

Tableau 35 : exemple de comparaison entre S/P réel et S/P crédibilisé

police	S/P réel	S/P crédibilisé	constat
*****	0,291922	0,402581	Sinistralité surestimé à priori
*****	0,634308	0,490317	Sinistralité sous estimée

Conclusion :

Les résultats obtenus par le modèle de Bühlman-Straub ont dévoilé son inconvénient en matière de la supposition que les contrats sont identiques. En effet les crédibilités ne sont pas satisfaisantes et une nécessité de différencier entre les flottes était clairement prouvée. Le modèle hiérarchisé donne des résultats plus satisfaisants, les crédibilités sont plus grandes et le modèle prend en compte les spécificités des différentes flottes en termes de prime, usage... palliant ainsi à l'inconvénient du modèle Bühlmanien.

Conclusion

Nous avons entrevu à travers ce mémoire une méthodologie de l'analyse actuarielle du portefeuille automobile.

A travers la logique sous jacente à la mise en place d'une extraction de données, nous avons précisé l'importance de la détection des erreurs et de l'épurement des données à partir d'un périmètre d'étude clairement défini. Ainsi un traitement pour les valeurs manquantes ou aberrantes a été établi par le recours à la moyenne pour les variables quantitatives ou par la mise de ces observations en question sur le compte de la classe la plus risquée. La logique derrière étant qu'un assuré qui ne déclare par une information cherche à cacher un paramètre qui risque de coûter cher à l'assureur.

Nous avons après mis en avant des techniques de base concernant l'analyse descriptive uni-variée et multi-variée, qui reste un préliminaire à ne pas négliger de la modélisation du risque. En effet, les résultats obtenus de cette analyse nous ont fournis un ensemble d'intuitions et de conclusions sur les différentes liaisons qui unissent les variables d'études. En effet, nous avons pu mettre en évidence les groupements pertinents des différentes variables en prenant en considération leurs comportements vis-à-vis de la fréquence des sinistres. De plus, nous avons pu à travers l'ACP et la CAH construire une vision sur les profils des assurés qui ont un comportement similaire vis-à-vis des paramètres fréquence, charge, S/P ...

L'application du modèle linéaire généralisé a été précédée par une étude préparatoire dont le but est de forger et lisser les ingrédients du GLM. Ainsi, des tests sur l'indépendance entre les variables qui participeront au modèle ont détecté une corrélation entre l'usage du véhicule et le carburant utilisé, la zone de circulation et le sexe du conducteur et finalement entre l'âge de l'assuré et celui de son permis. Ainsi, la liste des variables explicatives a été réduite à l'usage, la zone de circulation et l'âge du conducteur. La deuxième phase préparatoire a porté sur la modélisation des deux variables dépendantes du GLM. Les histogrammes et les tests d'ajustement ont montré que la fréquence s'ajuste mieux à une loi binomiale négative et la charge à une distribution log-normale.

Ces résultats ont été crédibilisés par les apports de l'application du modèle linéaire généralisé. On effet, les sorties ont mis en évidence non seulement une qualité d'ajustement médiocre pour les lois poisson et Gamma mais aussi des déviations plus grandes que celles relatives à la loi binomiale négatives et la loi log-normale. La déviance étant une grandeur qui mesure l'écart entre la distribution réelle et théorique. Les résultats nous ont ainsi rassuré que nos choix de distributions étaient sages.

D'une autre part les tests de WALD fournis par le GLM ont prouvé la significativité de toutes les variables explicatives avec la précision de la nécessité de grouper deux modalités de la variable zone de circulation pour le GLM sur le coût moyen pour l'usage tourisme : la zone peu risquée et la zone très peu risquée. Les modèles retenus s'écrivent ainsi :

$$\text{Coût moyen}_{Uijk} = \exp(\text{intercept} + \beta_{\text{âge } i} + \beta_{\text{véhicule } j} + \beta_{\text{zone } k})$$

$$\text{fréquence}_{Uijk} = \exp(\text{intercept} + \beta_{\text{âge } i} + \beta_{\text{véhicule } j} + \beta_{\text{zone } k})$$

Finalement les modèles retenus ont été mis à l'épreuve dans une tentative de vérifier leur conformité aux hypothèses. Ainsi, une analyse des résidus a donné un ajustement satisfaisant à la loi normale. Ce qui nous rassure sur la robustesse des modèles retenus.

Enfin, nous avons fait appel à la théorie de crédibilité dans une tentative de conclure sur la rentabilité des contrats et fournir des apports correctifs au jugement préalable de leur sinistralité. Les résultats du modèle Buhlmanien ont fait preuve de l'inconvénient de ne pas différencier entre les contrats d'où le recours au modèle hiérarchisé. Ledit modèle appliqué avec un seul niveau de subdivision tenant compte de l'usage des véhicules a donné des résultats plus pertinents que ceux du modèle Buhlmanien. Toutefois, un recours à une deuxième subdivision tenant compte de la cotisation en termes de primes s'est avérée nécessaire. Ainsi le modèle de Jewell à deux niveaux de subdivision s'avère le modèle le plus pertinent. Ces résultats sont satisfaisants et son optique nous ont fourni un outil pour juger et corriger l'estimation du degré de risque d'un contrat flotte. La comparaison entre les S/P crédibilisés et ceux réels permet de déduire si la sinistralité dudit contrat a été sous ou surestimée préalablement. C'est ainsi que le présent mémoire s'est achevé à bon port.

Bibliographie

Livres:

- **Saporta G. (2006)**, « Probabilités, analyse des données et statistique », Ed. TECHNIP.
Consultable partiellement en recherche sur <http://books.google.fr/>
- **Denuit M. et Charpentier A. (2005)**, «mathématiques de l'assurance non-vie», tome II, Ed.Economica.
- **Partrat C. et Besson J. (2005)**, «assurance non vie: modélisation, simulation», Ed. Economica.

Mémoires :

- **Anne J. (1998)**, mémoire «méthodologie de l'analyse des données en assurance automobile : construction d'un tarif ».
- **Chamar B. (2008)**, mémoire «processus de majoration des contrats flottes d'entreprises d'AXA France».
Consultable sur www.ressources-actuarielles.net
- **Gonnet G. (2010)**, mémoire « étude de la tarification et la segmentation en assurance automobile ».
Consultable sur www.ressources-actuarielles.net

Rapports :

- Rapport annuel de la FMSAR (Fédération Marocaine des Sociétés d'Assurances et de réassurance), Mars 2012.
- l'Arrêté du ministre des finances et de la privatisation n° 1053-06_du_28 rabii II 1427.

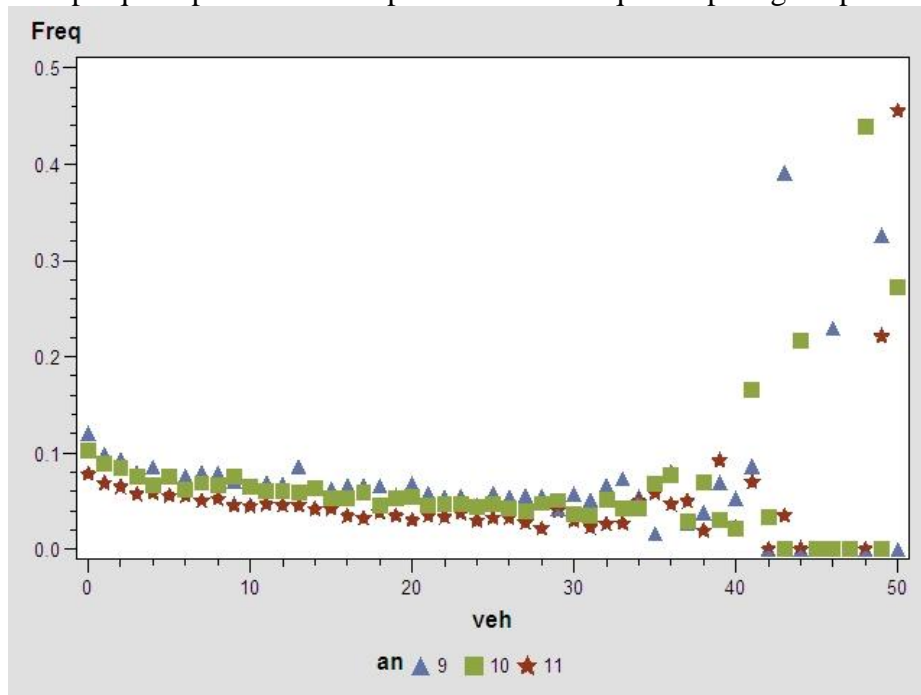
Documents :

- **Pred'homme E. (2006)**, cours de «SAS 9.1 for Windows»,
Téléchargeable sur www.stat.ucl.ac.be/cours/stat2020/
- **Burlot A. (1967)**, cours professé à l'institut de statistique de l'université de Paris «application de la statistique aux assurances accidents et dommages», Ed. BERGER-LEVRAULT .
- **Charpentier A. (2010)**, cours de l'assurance non vie « Statistique de l'assurance, STT 6705V: Statistique de l'assurance II ».
Téléchargeable sur le site :
<http://perso.univ-rennes1.fr/arthur.charpentier/stat-assurance-partie1-2010.pdf>
- Goulet V. (2010), cours d'actuariat université Laval « ACT 2008 Mathématiques actuarielles IARD II (Théorie de la crédibilité) ».
Téléchargeable sur le site : <http://libre.act.ulaval.ca>
- Tutoriel du package Actuar sous R « Credibility theory features of Actuar »
Téléchargeable sur cran.r-project.org/web/packages/actuar/actuar.pdf

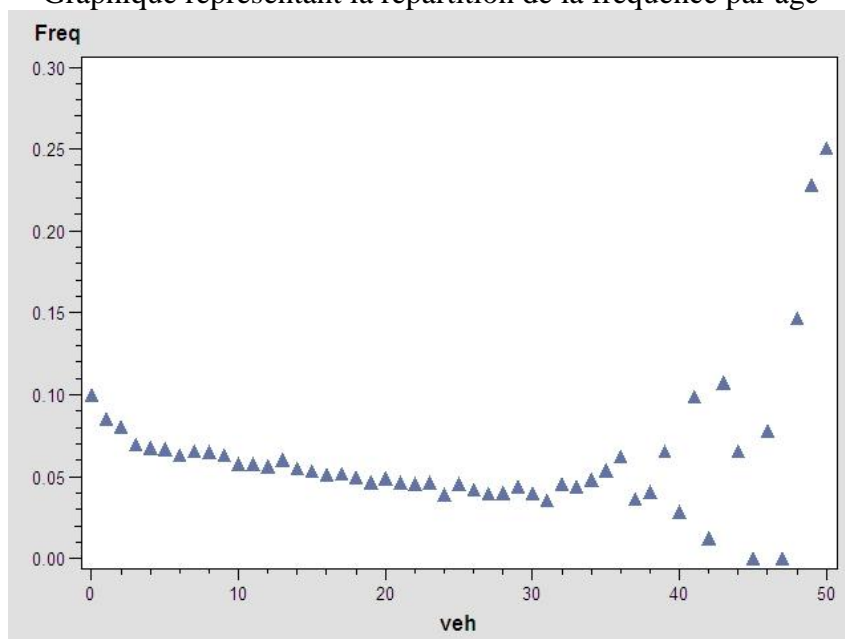
Annexes

- Regroupement :
I. L'âge du véhicule:

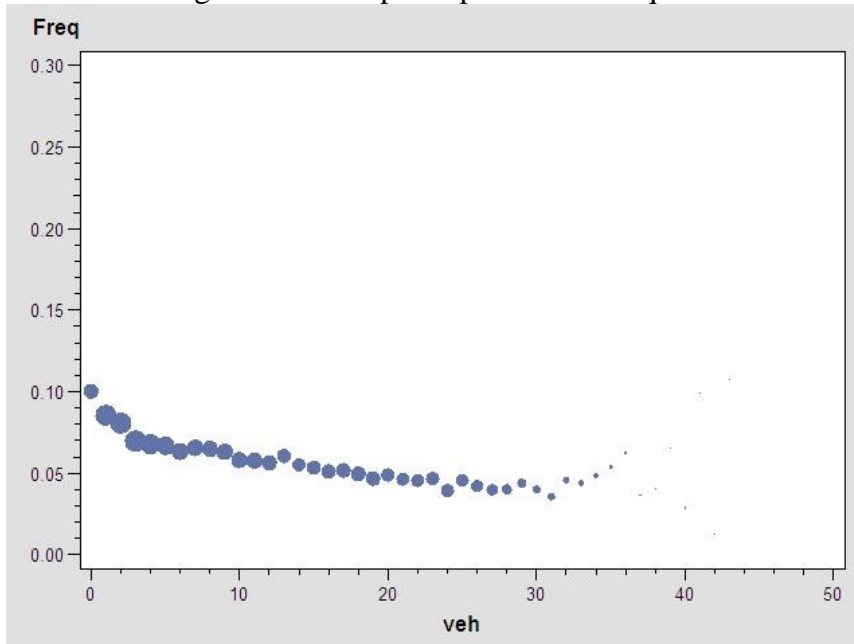
Graphique représentant la répartition de la fréquence par âge et par année



Graphique représentant la répartition de la fréquence par âge

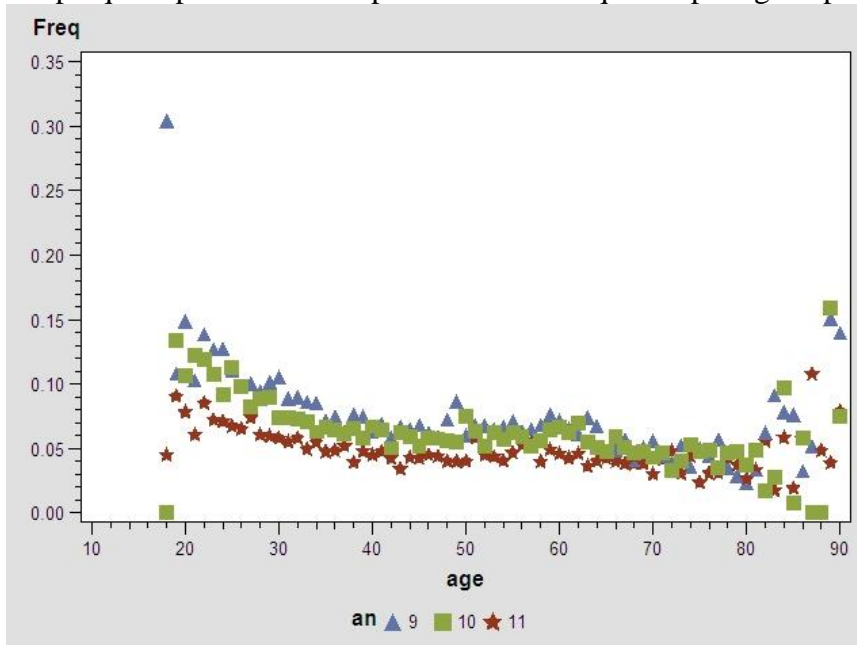


Graphique représentant la répartition de la fréquence par âge véhicule et par exposition au risque

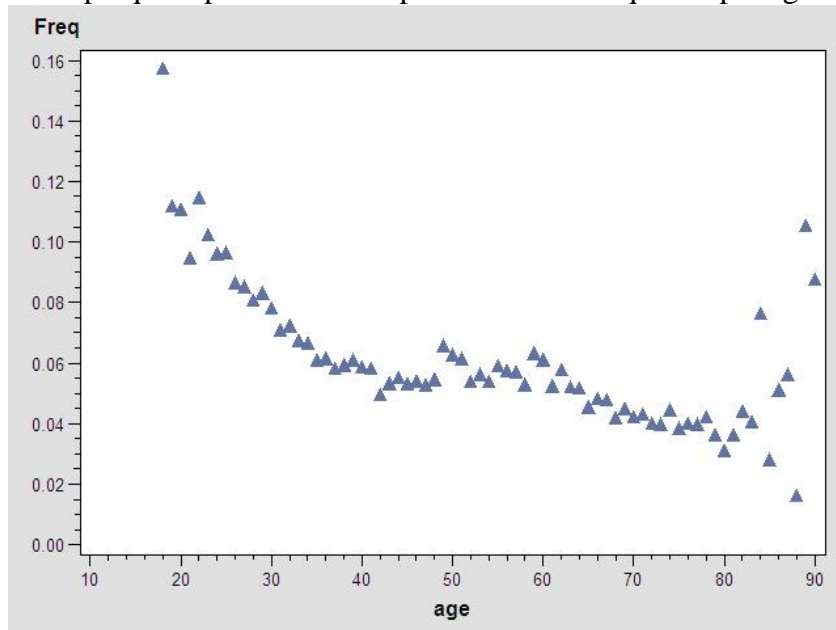


II. l'âge du conducteur :

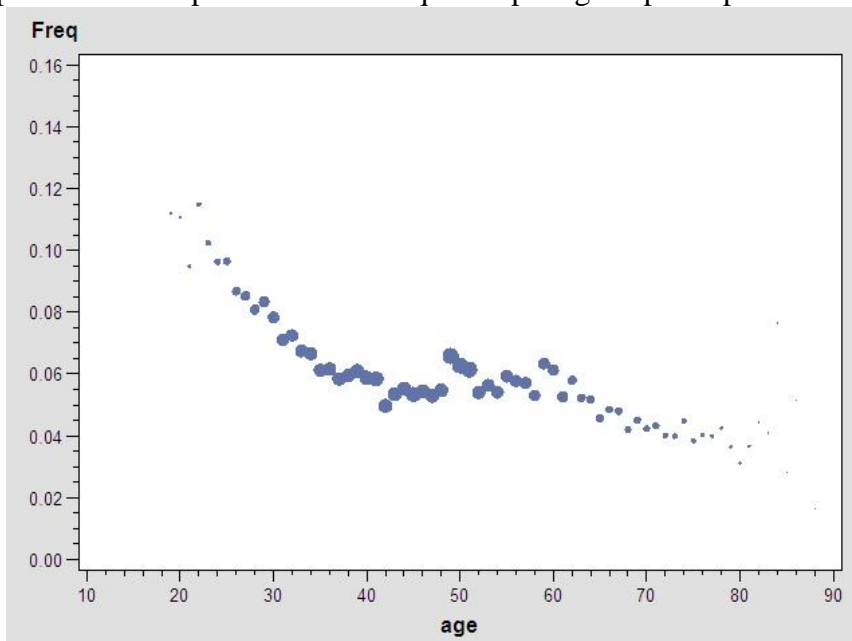
Graphique représentant la répartition de la fréquence par âge et par année



Graphique représentant la répartition de la fréquence par âge

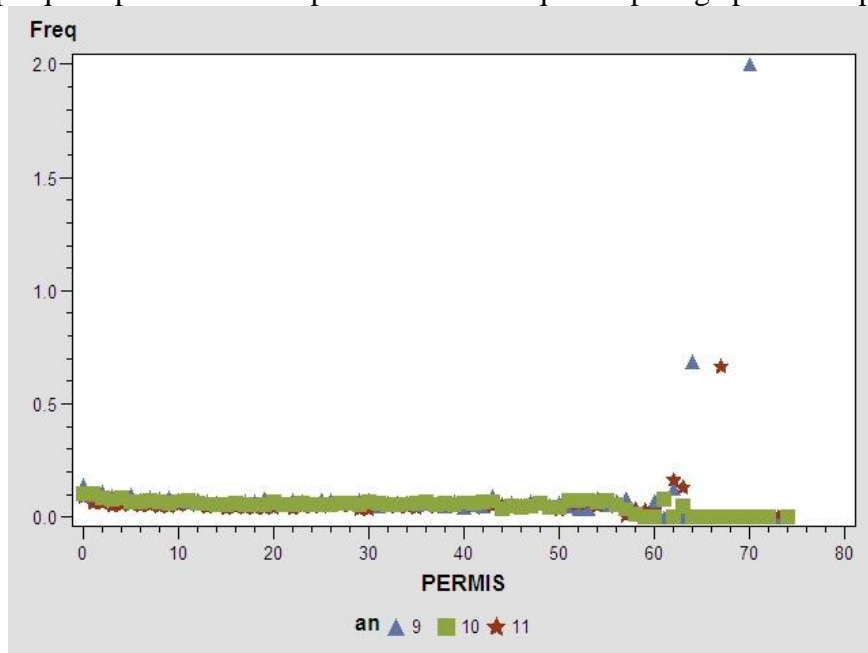


Graphique représentant la répartition de la fréquence par âge et par exposition au risque

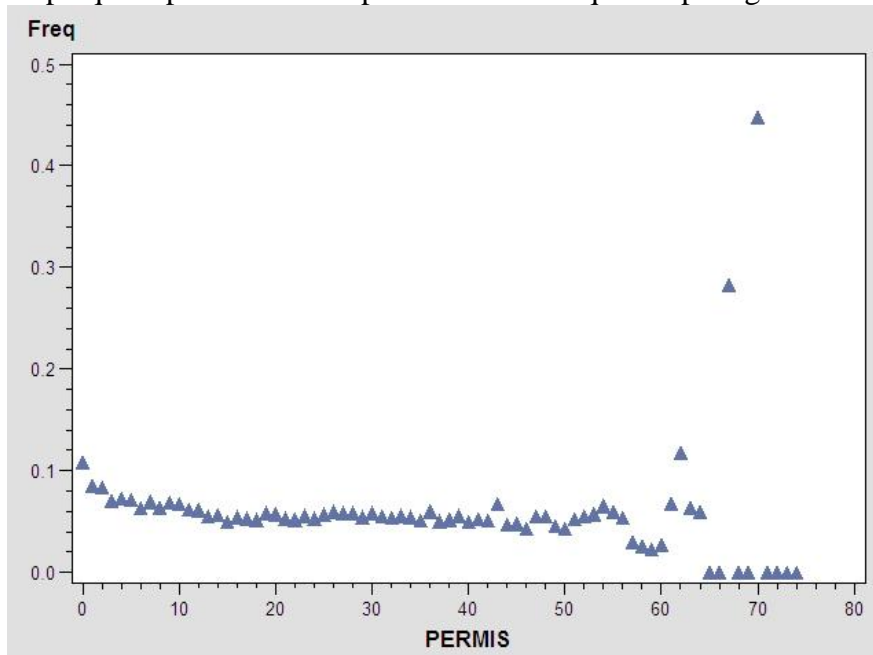


III. l'âge du permis:

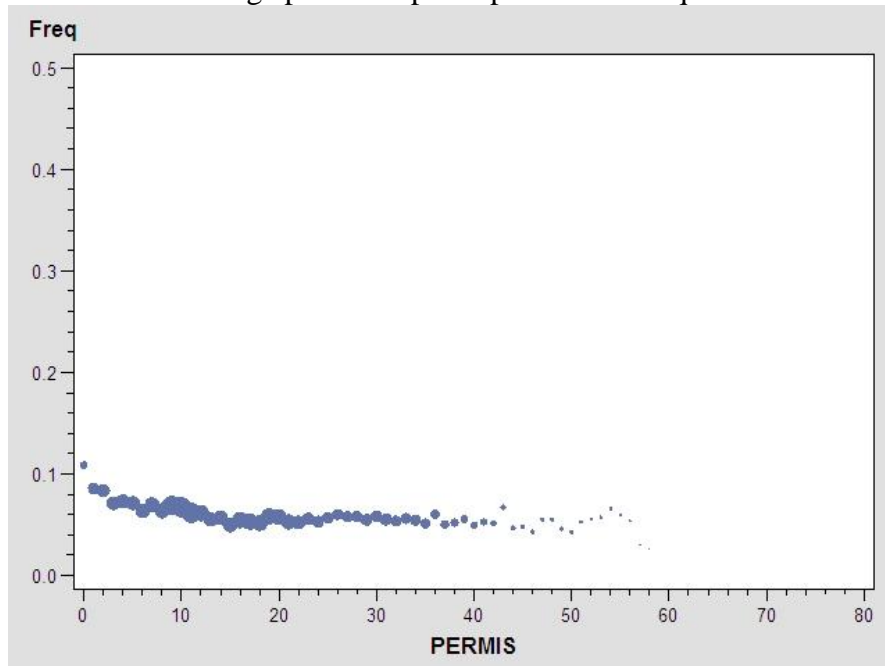
Graphique représentant la répartition de la fréquence par âge permis et par année



Graphique représentant la répartition de la fréquence par âge véhicule



Graphique représentant la répartition de la fréquence
Par âge permis et par exposition au risque



Source : SAS

- statistiques après regroupement :

IV. âge du conducteur :

date	Age	type	exposition au risque	prime	nombre sinistres	charge	freq	cm	sp
2009	1	âge risqué	40938,03288	123792526	4016	112864312	0,1	28103,7	0,912
2009	2	âge Nrisqu	168749,9068	492162128	11295	286624779	0,07	25376,3	0,582
2010	1	âge risqué	44746,87671	133819631	3648	93076370	0,08	25514,4	0,696
2010	2	âge Nrisqu	188339,8438	547829360	11159	247289008	0,06	22160,5	0,451
2011	1	âge risqué	46213,75616	135920777	2754	54834561	0,06	19910,9	0,403
2011	2	âge Nrisqu	208314,7479	600110087	9222	166469392	0,04	18051,3	0,277

V. combustion :

date	code	type combustion	exposition au risque	prime	nombre sinistre	charge	freq	cm	sp
2009	2	E	57523,493	127623006	3840	88819409	0,07	23130	0,7
2009	1	G	152164,45	488331648	11471	3,11E+08	0,08	27083	0,6
2010	2	E	60896,978	134577180	3592	73368347	0,06	20425	0,5
2010	1	G	172189,74	547071811	11215	2,67E+08	0,07	23807	0,5
2011	2	E	63788,178	139450159	2840	44523375	0,04	15677	0,3
2011	1	G	190740,33	596580705	9136	1,77E+08	0,05	19350	0,3

VI. âge permis :

date	code	type permis	exposition au risque	prime	nombre sinistres	charge	freq	cm	sp
2009	1	cond débutant	10163,81	3E+07	1113	34435369	0,11	30939,24	1,131
2009	2	cond expérimenté	199524,1	5,9E+08	14198	365053722	0,071	25711,63	0,623
2010	1	cond débutant	11005,9	3,2E+07	1027	27877875,2	0,093	27144,96	0,863
2010	2	cond expérimenté	222080,8	6,5E+08	13780	312487503	0,062	22676,89	0,481
2011	1	cond débutant	11437,29	3,3E+07	740	14529529,6	0,065	19634,5	0,442
2011	2	cond expérimenté	243091,2	7E+08	11236	206774423	0,046	18402,85	0,294

VII. sexe du conducteur :

date	code	type sexe	exposition	prime	nombre sinistres	charge	freq	cm	sp
2009	1	femme	23622,6877	56619630	1961	36406219	0,083	18565	0,643
2010	1	femme	26669,3068	64088152	1975	33293823	0,074	16858	0,52
2011	1	femme	29849,8712	71741841	1747	25570865	0,059	14637	0,356
2009	2	homme	186065,252	5,59E+08	13350	3,63E+08	0,072	27197	0,649
2010	2	homme	206417,414	6,18E+08	12832	3,07E+08	0,062	23930	0,497
2011	2	homme	224678,633	6,64E+08	10229	1,96E+08	0,046	19135	0,295

VIII. âge du véhicule :

date	code	type véhicule	exposition au risque	prime	nombre sinistres	charge	freq	cm	sp
2009	1	veh neuf	34370,263	104194255	3471	78123715	0,101	22508	0,7498
2010	1	veh neuf	35840,512	107008543	3218	64961988	0,09	20187	0,6071
2011	1	veh neuf	34753,386	105013038	2391	38256435	0,069	16000	0,3643
2009	2	veh norm	61445,737	173292764	4809	1,11E+08	0,078	23058	0,6399
2010	2	veh norm	71446,115	202246016	5008	99988440	0,07	19966	0,4944
2011	2	veh norm	84537,107	236955185	4593	74425072	0,054	16204	0,3141
2009	3	veh anci	113871,94	338467634	7031	2,1E+08	0,062	29936	0,6219
2010	3	veh anci	125800,09	372394433	6581	1,75E+08	0,052	26655	0,471
2011	3	veh anci	135238,01	394062641	4992	1,09E+08	0,037	21759	0,2756

IX. GLM sur le coût moyen des sinistres pour le commercial inférieur à 3.5T

The GENMOD Procedure

Informations sur le modèle

Data Set STAT.CHARGE2
 Distribution Normal
 Link Function Identity
 Dependent Variable cout

Number of Observations Read 145691
 Number of Observations Used 145691

Informations sur le niveau de classe

Classe	Niveaux	Valeurs
NÂge	2	1 2
NVEH	3	1 2 3
NZone	4	1 2 3 4

Critère pour évaluer la qualité de l'ajustement

Critère	DF	Valeur	Valeur/DF
Deviance	15E4	354751.9788	2.4351
Scaled Deviance	15E4	94699.0000	0.6500

Analyse des résultats estimés de paramètres

Paramètre	DF	Estimation	Erreur standard	Wald 95Limites de confiance %		Khi 2	Pr > Khi 2
Intercept	1	10.6388	0.0260	10.5879	10.6898	167602	<.0001
NÂge	1	0.1938	0.0584	0.0793	0.3082	11.00	0.0009
NÂge	2	0.0000	0.0000	0.0000	0.0000	.	.
NVEH	1	-0.2006	0.0316	-0.2626	-0.1386	40.21	<.0001
NVEH	2	-0.3184	0.0302	-0.3775	-0.2593	111.52	<.0001
NVEH	3	0.0000	0.0000	0.0000	0.0000	.	.
NZone	1	-0.6171	0.0322	-0.6803	-0.5539	366.52	<.0001
NZone	2	-0.2543	0.0328	-0.3185	-0.1901	60.27	<.0001
NZone	3	-0.2110	0.0553	-0.3194	-0.1025	14.54	0.0001
NZone	4	0.0000	0.0000	0.0000	0.0000	.	.
Scale	0	1.0000	0.0000	1.0000	1.0000	.	.

NOTE: The scale parameter was held fixed.

X. GLM sur le coût moyen des sinistres pour le commercial supérieur à 3.5T

The GENMOD Procedure

Informations sur le modèle

Data Set STAT.CHARGE3
 Distribution Normal
 Link Function Identity
 Dependent Variable cout

Number of Observations Read 37225
 Number of Observations Used 37225

Informations sur le
niveau de classe

Classe	Niveaux	Valeurs
NÂge	2	1 2
NVEH	3	1 2 3
NZone	3	1 3

Critère pour évaluer la qualité de l'ajustement

Critère	DF	Valeur	Valeur/DF
Deviance	37E3	170004.2118	4.5677
Scaled Deviance	37E3	28149.0000	0.7562

Analyse des résultats estimés de paramètres

Paramètre	DF	Estimation	Erreur standard	Wald 95Limites de confiance %		Khi 2	Pr > Khi 2
Intercept	1	10.3798	0.0319	10.3173	10.4423	106072	<.0001
NÂge	1	0.1701	0.0430	0.0858	0.2544	15.65	<.0001
NÂge	2	0.0000	0.0000	0.0000	0.0000	.	.
NVEH	1	-0.1706	0.0449	-0.2586	-0.0827	14.46	0.0001
NVEH	2	-0.1737	0.0384	-0.2490	-0.0985	20.49	<.0001
NVEH	3	0.0000	0.0000	0.0000	0.0000	.	.
NZone	1	-0.1903	0.0337	-0.2564	-0.1242	31.82	<.0001
NZone	3	0.0000	0.0000	0.0000	0.0000	.	.
Scale	0	1.0000	0.0000	1.0000	1.0000	.	.

NOTE: The scale parameter was held fixed.

XI. GLM sur la fréquence des sinistres pour le commercial inférieur à 3.5T

The GENMOD Procedure

Informations sur le modèle

Data Set	STAT.CHARGE2
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	nb_rc
Offset Variable	ldur

Number of Observations Read	145691
Number of Observations Used	145691

Informations sur le niveau de classe

Classe	Niveaux	Valeurs
NÂge	2	1 2
NVEH	3	1 2 3
NZone	4	1 2 3 4

Critère pour évaluer la qualité de l'ajustement

Critère	DF	Valeur	Valeur/DF
Deviance	15E4	13837.4945	0.0950
Scaled Deviance	15E4	13837.4945	0.0950

Analyse des résultats estimés de paramètres

Paramètre	DF	Estimation	Erreur standard	Wald 95Limites de confiance %	Khi 2	Pr > Khi 2
Intercept	1	-3.1859	0.0406	-3.2655 -3.1063	6157.25	<.0001
NÂge	1	0.2908	0.0451	0.2024 0.3792	41.56	<.0001
NÂge	2	0.0000	0.0000	0.0000 0.0000	.	.
NVEH	1	0.6632	0.0531	0.5590 0.7673	155.71	<.0001
NVEH	2	0.2903	0.0477	0.1969 0.3837	37.12	<.0001
NVEH	3	0.0000	0.0000	0.0000 0.0000	.	.
NZone	1	0.7465	0.0527	0.6431 0.8498	200.47	<.0001
NZone	2	0.3500	0.0503	0.2515 0.4486	48.43	<.0001
NZone	3	-0.5825	0.0730	-0.7256 -0.4395	63.67	<.0001
NZone	4	0.0000	0.0000	0.0000 0.0000	.	.
Dispersion	1	26.4232	0.7591	24.9353 27.9110	.	.

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

XII. GLM sur la fréquence des sinistres pour le commercial supérieur à 3.5T

The GENMOD Procedure

Informations sur le modèle

Data Set STAT.CHARGE3
 Distribution poisson
 Link Function Log
 Dependent Variable nb_rc
 Offset Variable ldur

Number of Observations Read 37225
 Number of Observations Used 37225

Informations sur le
niveau de classe

Classe	Niveaux	Valeurs
NAge	2	1 2
NVEH	3	1 2 3
NZone	3	1 2 3

Critère pour évaluer la qualité de l'ajustement

Critère	DF	Valeur	Valeur/DF
Deviance	37E3	6571.5812	0.1766
Scaled Deviance	37E3	6571.5812	0.1766

Analyse des résultats estimés de paramètres

Paramètre	DF	Estimation	Erreur standard	Wald 95Limites de confiance %	Chi 2	Pr > Chi 2
Intercept	1	-2.5660	0.0491	-2.6622 -2.4698	2733.05	<.0001
NAge	1	0.1638	0.0745	0.0178 0.3097	4.83	0.0279
NAge	2	0.0000	0.0000	0.0000 0.0000	.	.
NVEH	1	0.4831	0.0803	0.3258 0.6404	36.22	<.0001
NVEH	2	0.2357	0.0656	0.1071 0.3642	12.91	0.0003
NVEH	3	0.0000	0.0000	0.0000 0.0000	.	.
NZone	1	1.2609	0.0689	1.1259 1.3959	335.12	<.0001
NZone	2	0.5409	0.0679	0.4079 0.6739	63.55	<.0001
NZone	3	0.0000	0.0000	0.0000 0.0000	.	.
Dispersion	1	14.5803	0.5625	13.4778 15.6829	.	.

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

XIII. Code de la cartographie du risque automobile:

```
data sasuser.map1;
set maps.morocco;
if id=16 then id1=23;
if id=5 then id1=23;
if id=14 then id1=25;
if id=12 then id1=25;
if id=13 then id1=26;
if id=1 then id1=26;
if id=3 then id1=27;
if id=9 then id1=28;
if id=11 then id1=28;
if id=18 then id1=29;
if id=21 then id1=29;
if id=19 then id1=831;

if id not in ( 16 5 14 12 13 1 3 9 11 18 21 19 ) then id1=id; drop
id;rename id1=id;run;

data p1;set sasuser.map1;
x=-1*Long;
y=lat;
segment=segment;
cont=94;
country=id;
keep x y segment cont country;
run;
data p2;set maps.africa;if id=831;
x=-1*Long;
y=lat;
segment=segment;
cont=94;
country=id;
keep x y segment cont country;
run;
data mar;set p1 p2;run;
proc gproject data=mar out=mar dupok eastlong project=winkel2;id cont
country;
run;
data mydata;
input cont country region freq countryname $16.;
if freq ge 0 and freq le 0.03 then freq1=5;
if freq ge 0.03 and freq le 0.052 then freq1=2;
if freq ge 0.052 and freq le 0.07 then freq1=3;
if freq ge 0.07 and freq le 1 then freq1=1;
;
datalines;
94 26 1 0.038326817 Morocco
94 2 2 0.023335877 Morocco
94 27 27 0.027753799 Morocco
94 4 4 0.120835898 Morocco
94 23 5 0.05463516 Morocco
94 6 6 0.080651914 Morocco
94 7 7 0.043549748 Morocco
94 8 8 0.035332411 Morocco
94 28 9 0.038345632 Morocco
94 10 10 0.058804767 Morocco
94 28 11 0.038345632 Morocco
```

```
94 25 12 0.039853783 Morocco
94 26 1 0.038326817 Morocco
94 25 12 0.039853783 Morocco
94 15 15 0.090016088 Morocco
94 23 5 0.05463516 Morocco
94 17 17 0.063415326 Morocco
94 29 18 0.063976974 Morocco
94 831 19 0.034343747 Morocco
94 20 20 0.039772456 Morocco
94 29 18 0.063976974 Morocco
94 831 19 0.034343747 western sahara
```

```
;  
run;  
proc gmap data=mydata map=mar all;  
  id cont country;  
  choro freq1 / discrete  
  des="" name="name";  
run;  
quit;
```

