

INSEA

Projet de Fin d'Etudes

**Elaboration d'un Zonier automobile
Modélisation de risque et étude d'impact**

Préparés par : **BENABDERRAHMAN HOUDA
EL KHALIFA SALAH EDDINE**

Sous la direction de : **M.MARRI FOUAD (INSEA)**

M.ALOUAN IMAD EDDINE (AXA)

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

Diplôme d'Ingénieur d'Etat

Filière : ACTURIAT FINANCE

Devant le jury composé de :

- **M. EL HAJ TIRARI MOHAMMED (INSEA)**
- **M. MARRI FOUAD (INSEA)**
- **M.ALOUAN IMAD EDDINE (AXA)**

Résumé

Le marché de l'assurance automobile étant très concurrentiel, les compagnies d'assurance cherchent aujourd'hui à mettre en place une tarification basée sur une segmentation de plus en plus performante. La variable correspondant au critère spatial est l'une des variables tarifaires nécessitant une segmentation plus fine.

Nous disposons pour chaque assuré de son lieu de résidence mais aucune mesure ne nous permet de comparer ces lieux entre eux et de les classer par risque. Une approche à dire d'expert a été utilisée pour construire le zonier c'est-à-dire le regroupement des communes en risques homogènes actuel d'AXA. Cependant dans ce mémoire nous faisons la refonte de ce zonier par les méthodes de Machine Learning. En essayant de voir dans quelle mesure l'ajout des caractéristiques sociodémographiques des différentes communes, exogènes aux risques, pouvait-nous aider à affiner la segmentation de la zone géographique.

La première approche consiste à construire un arbre CART et la seconde approche utilise un autre algorithme de Machine Learning le Random Forest.

Nous avons créé par la suite des modèles de fréquence et coût moyen n'incluant aucun facteur susceptible de contenir de l'information géographique, et d'autres incluant les zoniers réalisés par les deux méthodes afin d'extraire la part du risque expliqué par les zoniers. Nous avons étudié aussi si les zoniers élaborés par les deux méthodes apportent bien un gain significatif au modèle de tarification.

MOTS-CLES : Zonier, Apprentissage supervisé, Tarification, Données externes, Classification, Modèle linéaire généralisé

ABSTRACT

As the auto insurance market is becoming more competitive, insurance companies are now seeking to adapt their pricing based on an efficient segmentation. The Territory is one of the rating variables requiring a finer segmentation.

For each insured person we have the geographical area where he lives however, there is no measure that allows us to compare these areas with each other and to classify them by risk. Axa insurance used risk underwriter's expertise to classify cities codes into four different groups. This was achieved by grouping cities codes in a subjective manner based on the insurance risk underwriter's belief about their claims experience .The aim of this study is to update the current spatial risk zoning not according to an expert opinion but according to machine learning, using HCP external data : Sociodemographic characteristics to ensure a maximal precision.

The first approach consists on building a CART decision tree, while the second approach used random forest which is an aggregation of decision trees.

Finally, we extracted the amount of risk explained by our zoning by creating models (frequency and severity) without including any factor that could contain geographical information, and others including the zoning carried out by the two methods. We studied also whether the zoning developed by the two methods bring a significant gain to the current pricing model

KEYWORDS: Zoning, Supervised learning, Pricing, External data, Classification, general linear models.

Dédicace

Je remercie Allah, le tout puissant, le miséricordieux, de m'avoir appris ce que j'ignorais, de m'avoir donné la santé et tout dont je necessitais pour l'accomplissement de ce projet de fin d'études.

Je dédie ce travail :

A mes chers parents,

Qui ont pris soins de me guider, me conseiller et qui m'ont toujours entouré d'affection et de soutien. Les résultats que j'obtiens aujourd'hui leur reviennent en grande partie. En témoignage de ma reconnaissance Je leur offre ce travail.

A ma sœur Farah et mon petit frère Mohammed

Je leur dédie ce travail en témoignage de l'attachement, de l'amour et de l'affection que je porte pour eux

A mes chers ami (e)s

Je leur dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite.

A tous ceux que j'aime et ceux qui m'aiment

Puisse Dieu, le tout puissant, vous préserver et vous accorder santé, longue vie et bonheur

Houda BENABDERRAHMAN

Dédicace

Dédicace,

A mon père qui m'aide à chaque fois que je tombe par terre,

A ma mère, mes frères, qu'ils sont pour moi les êtres les plus chers.

EL KHALIFA Salah Eddine

SEEK

Remerciements

Nous tenons à exprimer nos sincères remerciements à **M. Dbich Abderrahim** qui nous a accordé l'opportunité d'effectuer notre stage de fin d'études au sein de son équipe.

Nous exprimons aussi notre profonde reconnaissance à nos encadrants, **M.Alouan Imad Eddine**, **M. Bensouna Adil**, **M.Boukharsa Hicham**, qui par leurs Compétences techniques, leurs directives et leurs qualités humaines nous ont épaulés et orientés durant toute la période de stage pour l'accomplissement de ce modeste travail.

Qu'ils trouvent ici le témoignage de notre profonde gratitude.

Nous adressons nos vifs remerciements à notre professeur encadrant **Monsieur Marri Fouad**, qui par sa patience, ses conseils, et sa disponibilité notre travail a pu être mené au bon port.

Nos remerciements vont également :

- Aux membres du jury pour avoir accepté d'évaluer notre projet de fin d'études et pour toutes leurs remarques et leurs propositions enrichissantes.
- A tous les professeurs qui nous ont enseigné et qui par leurs compétences nous ont soutenu dans la poursuite de nos études.
- A tous le personnel d'AXA assurance qui nous a facilité l'insertion dans le milieu du travail.
- A tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce travail.

Merci à vous tous

Table des matières

ABSTRACT	3
<i>Devoirs</i>	4
<i>Devoirs</i>	5
Remerciements	6
Liste des Tableaux.....	12
Liste des Figures.....	13
Introduction	14
Chapitre 1 : AXA Assurance Maroc	16
I. Présentation de l’organisme d’accueil	17
I.1 Metier :	17
I.2 Vision :	18
I.2.1 Mission	18
I.2.2 Valeurs.....	18
I.2.3 Attitudes.....	18
Chapitre 2 : Marché de l’automobile au Maroc	19
I. Présentation du marché de l’automobile.....	20
I.1 Assurance Automobile	20
I.2 Vision globale du Marché	20
I.3 Axa assurance.....	22
Chapitre 3 : Présentation des données et statistiques descriptives	23
I. Description des données	24
I.1. Base de données interne	24
I.2. Base de données externe	24
I.3 Jointure des deux bases de données	26

II. Analyses descriptives	26
II.1. Base de données externe	26
II.1.1 Etude des corrélations entre les variables externes	26
II.2. Base de données interne.....	29
II.2.1 Epurement de la base de données.....	29
II.2.2 Etude des corrélations entre les variables tarifaires	30
II.2.3 Distribution de la fréquence matérielle en fonction des variables tarifaires	31
III. Etude préliminaire : analyse du zonier actuel	33
III.1. Présentation du zonier actuel.....	33
III.2. Etude de la fréquence et du coût moyen des sinistres et du ratio S / P.....	33
Chapitre 4 : Elaboration du zonier par les méthodes du Maching Learning.....	36
I. Méthodologie de Travail.....	37
I.1. Méthodes utilisées	37
I.2. Choix de l'indicateur de sinistralité	38
II. Partie théorique.....	38
II.1. 1 ^{ère} approche : Arbre de Régression (CART)	38
II.1.1. Algorithme	39
II.1.2. Nécessité de l'élagage de l'arbre.....	39
II.1.3. Validation croisée.....	40
II.2. 2 ^{ème} approche : Random Forest (Forêts aléatoires)	41
II.2.1. Présentation de la méthode.....	42
II.2.2. Avantages de la méthode.....	43
II.2.3. Algorithme	43
II.2.4. L'erreur OOB	43
II.2.5. L'importance des variables	44
II.2.6. Dépendances partielles	45
II.2.7. Sélection des variables	45

II. Elaboration du Zonier par les 2 approches	46
II.1 Zonier par la méthode des arbres de décisions (CART)	46
II.2. Zonier par la méthode des forêts aléatoires	46
II.2.1. Matrice de proximité	47
II.2.2. Le Multidimensional Scaling	47
II.2.3. Classification ascendante hiérarchique (CAH)	48
II.2.4. Choix du nombre optimum de classes.....	49
III. Application	50
III.1. Présentation de la base de travail.....	50
III.2. Traitement des variables utilisées.....	50
III.3. Zonier pour la fréquence des sinistres matériels par Arbre de Régression	52
III.3.1. Enjeu de la Méthode de CART	52
III.3.2. Arbre de régression pour la fréquence matérielle	53
III.3.3 Projection des classes sur la carte à risque.....	56
III.4. Zonier pour la fréquence des sinistres matériels par Random Forest.....	56
III.4.1. Enjeu de la Méthode du Random Forest.....	56
III.4.2. Importance des variables.....	57
III.4.3. Dépendances partielles.....	58
III.4.4. Sélection des variables	59
III.4.5. Optimisation des paramètres	60
III.4.6. Matrice de proximité.....	61
III.4.7. Classification ascendante hiérarchique	62
III.4.8. Variables importantes pour chaque cluster	64
III.4.9. Projection des Classes de risques sur la carte	65
III.5 Validation des zoniers	66
III.5.1 Comparaison entre les deux zoniers	66
Chapitre 5 : Modélisation du risque	69

I. Aspect théorique	70
I.1. Théorie des modèles linéaires généralisées.....	70
I.1.1. Définition des Modèles linéaires généralisés	70
I.1.2. Hypothèses du modèle.....	70
I.1.3. Détermination des coefficients d'un modèle	70
I.1.4. Significativité des variables	71
I.1.5. Sélection des variables explicatives	71
I.1.6. Qualité d'ajustement.....	72
I.2. Apport significatif d'une variable dans le modèle	73
I.2.1. Statistique du Chi ²	73
II.1. Modélisation de la fréquence des sinistres :	74
II.1.1. Loi de poisson	74
II.1.2. Loi binomiale négative.....	75
II.2. GLM sur la fréquence des sinistres.....	76
II.2.1 Modèle de Poisson	76
II.2.2 Modèle Binomiale négative	77
Chapitre 6 : Analyse de la performance du zonier	79
I.1. Evaluation de la performance du zonier.....	80
I.1.1. Zonier obtenu par CART	80
I.1.2. Zonier obtenu par Random Forest	81
Conclusion.....	83
Bibliographie et Webographie	84

Liste des abréviations

AIC : Akaike Information Criterion

AP : Année Police

BD : Base de Données

BIC : Bayesian Information Criterion

CA : Chiffre d'Affaires

CAH : Classification Ascendante Hiérarchique

CART : Classification and Regression Trees

CID : Convention d'indemnisation directe

CP : Paramètre de complexité

ddl : degrés de liberté

GLM : General Linear Model

HCP : Haut-Commissariat au plan

MSE : Mean Squared Error

Mtry : Nombre de variables explicatives choisi à chaque arbre

Nbsplit : Nombre de découpages

OOB : Out Of Bag

RC : Responsabilité Civile

RF : Random Forest

rel error ; Erreur relative

S / P : Loss Ratio

xerror : Erreur de validation croisée

xstd : Ecart type de l'estimation de l'erreur de validation croisée

Liste des Tableaux

Tableau 1: Structure du chiffre d'affaire du secteur d'assurance au Maroc	21
Tableau 2: Evolution du chiffre d'affaire des différentes branches du secteur d'assurance	21
Tableau 3: Position d'AXA assurance Maroc en terme du chiffre d'affaire Non vie.....	22
Tableau 4: Variables externes	25
Tableau 5: Statistiques sur les valeurs manquantes et aberrantes de la base des données	30
Tableau 6: Les variables tarifaires retenues	30
Tableau 7: Sortie SAS corrélation entre les variables tarifaires quantitatives	31
Tableau 8: Discrétisation des variables externes retenues	51
Tableau 9: Critères associés à chaque arbre créé	53
Tableau 10: Tranches de fréquence.....	57
Tableau 11: Classement des variables selon leurs valeurs d'importance dans chaque Cluster	64
Tableau 12: Statistique de Khi-deux pour la loi de Poisson et la loi Binomiale négative	75
Tableau 13: Estimation des paramètres pour la loi de Poisson	76
Tableau 14: Evaluation de la qualité d'ajustement du GLM pour la loi de poisson	77
Tableau 15: Estimation des paramètres pour la loi Binomiale Négative	77
Tableau 16: Evaluation de la qualité d'ajustement du GLM pour la loi Binomiale Négative .	78
Tableau 17: Estimation des paramètres pour la loi Binomiale Négative (CART).....	80
Tableau 18 : Evaluation de la qualité d'ajustement du GLM pour la Binomiale Négative (CART).....	80
Tableau 19: Estimation des paramètres pour la loi Binomiale Négative (RF).....	81
Tableau 20: Evaluation de la qualité d'ajustement du GLM pour la Binomiale Négative (RF)	82

Liste des Figures

Figure 1: AXA assurance dans le monde	17
Figure 2: Métiers	17
Figure 3 : Triangle inférieur de la matrice de corrélation	27
Figure 4: Graphique représentant la répartition de la fréquence par âge de l'assuré	31
Figure 5: Répartition de la fréquence en fonction du CRM	32
Figure 6 : Répartition de la fréquence en fonction de l'âge du véhicule	32
Figure 7: Carte du risque actuel : Répartition en 4 zones à risque	33
Figure 8: Evolution de la Fréquence des sinistres matériels sur la période 2010 à 2014	34
Figure 9: Evolution du coût moyen des sinistres matériels sur la période 2010 à 2014	34
Figure 10: Evolution du S/P des sinistres matériels sur la période 2010 à 2014	35
Figure 11: Séparation des données pour la méthode validation croisée ex du 5-fold	41
Figure 12: CART (sans élagage) appliquée aux données avec Fréquence matérielle comme Target	53
Figure 13: Graphique de l'erreur de la validation croisée en fonction du paramètre de complexité	54
Figure 14: Arbre élagué pour la Fréquence matérielle comme Target	55
Figure 15: Cartographie des zones de risque (CART)	56
Figure 16 : Importance des variables explicatives	57
Figure 17: Dépendances partielles entre les variables explicatives et la variable à expliquer	58
Figure 18: Sélection des variables pertinentes	59
Figure 19: L'erreur OOB en fonction du nombre d'arbres dans la forêt	60
Figure 20: Tracé de l'erreur OOB en fonction du Mtry	61
Figure 21: Graphique de proximité	62
Figure 23: Graphique des pertes d'inertie relatives	63
Figure 24 Carte des zones de risque (Random Forest)	65
Figure 27: Ajustement par loi de Poisson	74
Figure 28: Ajustement par Binomiale négative (sortie R)	75

Introduction

En tarification automobile, La zone géographique constitue une variable tarifaire importante qui segmente la population en groupe de risque homogène. Toutefois la technique de définition de chaque zone de risque reste propre à chaque assureur. Notre étude cherche à faire la refonte du zonier actuel. En effet, nous essayons de déterminer une analyse du risque et donc une segmentation plus fine des communes , en tenant compte de leurs données externes « sociodémographiques » jugées pertinentes pour déterminer le risque sous-jacent à chacune d'elles.

Nous disposons pour cela d'une base de données RC automobile détaillant le profil couvert ainsi que les sinistres de chaque assuré sur la période 2010 - 2014. Et nous essayons de mettre en évidence l'influence des caractéristiques sociodémographiques du territoire sur la sinistralité observée.

Les étapes menant à la construction du zonier ont été réparties en cinq chapitres :

Le 1^{er} chapitre présente l'organisme d'accueil Axa Assurance Maroc.

Le 2^{ème} chapitre traite le cadre assurantiel dans lequel le zonier s'inscrit. On y présente les aspects généraux de l'assurance automobile ainsi qu'une vision sur le marché de l'automobile. Le lecteur puisse ainsi acquérir une compréhension de l'environnement de travail.

Le 3^{ème} chapitre porte sur la présentation des données et les statistiques descriptives.

Nous y détaille la constitution de la base de données externe, et celle d'interne ainsi que les statistiques descriptives portant sur les deux bases. Une analyse du zonier actuel sera également mise en œuvre, pour examiner sa pertinence et sa performance pour faire face au risque actuel.

Dans le 4^{ème} chapitre nous introduisons les deux méthodes de Machine Learning retenues pour la construction du zonier, à savoir les arbres de décision CART et les forêts aléatoires. On y présente d'abord les théories propres à chaque méthode ainsi que leurs mises en pratique.

Le 5^{ème} chapitre nous permet de modéliser le risque. Nous modélisons à l'aide du GLM la fréquence des sinistres matériels.

Le 6^{ème} et dernier chapitre nous permet de juger la pertinence de nos zoniers réalisés par les deux méthodes. En étudiant si les deux zoniers apportent bien un gain significatif au modèle de tarification.

Afin de fluidifier la lecture du mémoire et d'éviter des analyses redondantes, ainsi que la répétition des calculs, les résultats de l'étude sont principalement présentés pour les sinistres matériels, ceux corporels s'obtenant de manière analogue.

NB :

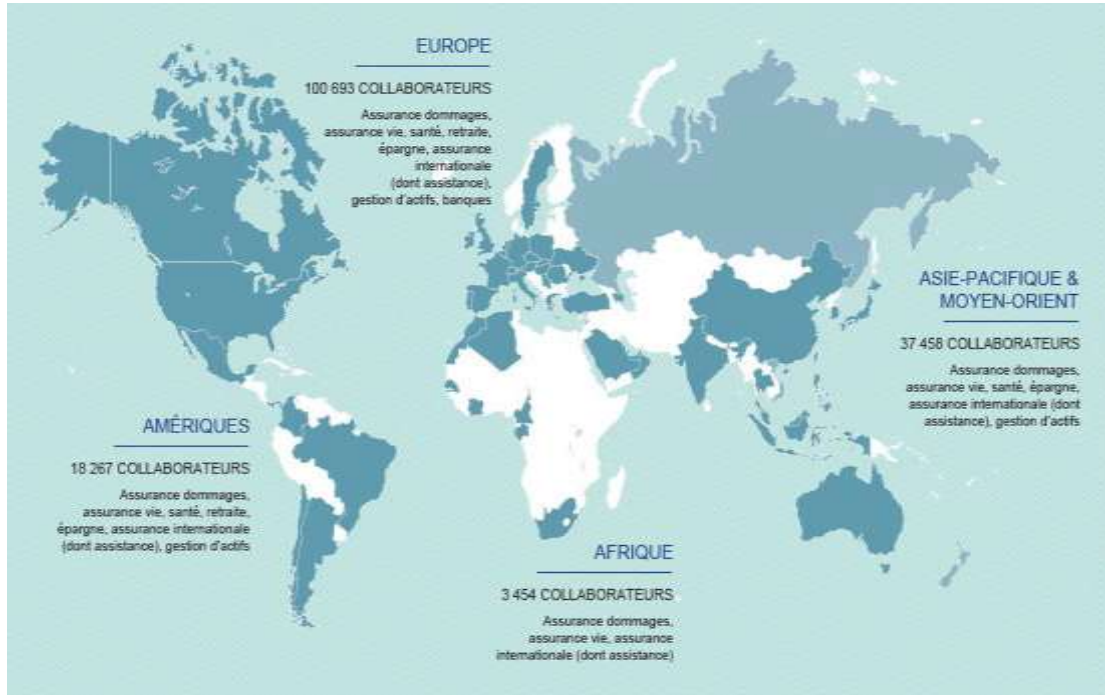
Pour des raisons de confidentialité, nous avons dépersonnalisé la base. Par exemple, nous avons masqué les données confidentielles, nous n'avons pas explicité les variables finales utilisées dans l'arbre de régression CART et les variables importantes dans chaque classe du zonier élaboré par le Random Forest, nous avons décoloré aussi les zones de risque projetées sur la carte et nous avons effectué des homothéties des variables quantitatives comme pour les sorties de SAS.

Chapitre 1 : AXA Assurance Maroc

I. Présentation de l'organisme d'accueil

AXA assurance est un groupe international français spécialisé dans l'assurance depuis sa création, et dans la gestion d'actifs depuis 1994. Elle est issue de la fusion de plusieurs sociétés d'assurance, dont la plus ancienne date de 1817.

Figure 1: AXA assurance dans le monde



I.1 Metier :

Figure 2: Métiers



I.2 Vision :

I.2.1 Mission

AXA Assurance Maroc aide ses clients à vivre confiants jour après jour, en les protégeant, en protégeant leurs familles et leurs biens contre les risques, et en gérant leur épargne.

I.2.2 Valeurs

Cinq valeurs fondent alors la culture du Groupe AXA et expriment la manière dont chacun agit au sein de l'entreprise : professionnalisme, respect de la parole donnée, innovation, esprit d'équipe et réalisme. Ces valeurs sont le fondement de l'ambition d'AXA Assurance Maroc. Elles servent de guide pour chaque collaborateur et inspirent ses actions et décisions.

I.2.3 Attitudes

Les collaborateurs d'AXA Assurance Maroc font tout ce qui est en leur pouvoir pour satisfaire leurs clients en étant :

– Disponible :

Nous sommes disponibles pour nos clients à tout moment et sommes réellement à leur écoute pour les accompagner dans leur quotidien.

– Attentionné :

Nous traitons nos clients avec tous les égards qui leur sont dus, avec compréhension et considération. Nous répondons à leurs besoins avec des services personnalisés, les conseillons tout au long de leur vie et récompensons leur fidélité.

– Fiable :

Nous sommes sincères et logiques dans notre démarche avec nos clients. Nous réalisons nos promesses et nous tenons nos clients continuellement informés, afin qu'ils puissent nous faire confiance. AXA Assurance Maroc aide ses clients à vivre confiants jour après jour, en les protégeant, en protégeant leurs familles et leurs biens contre les risques, et en gérant leur épargne.

*Chapitre 2 : Marché de
l'automobile au Maroc*

I. Présentation du marché de l'automobile

I.1 Assurance Automobile

L'assurance automobile comporte deux types de garanties, la garantie « responsabilité civile » et la garantie « garanties annexes ».

La responsabilité Civile au Maroc

La responsabilité civile est obligatoire, cette assurance permet de couvrir la responsabilité civile du souscripteur du contrat, du propriétaire du véhicule et de toute personne ayant, la garde ou la conduite du véhicule. Ils sont couverts par cette garantie les dommages matériels et les dommages corporels.

Depuis 2006 les compagnies ont acquis le droit de calculer eux-mêmes la prime sans intervention réglementaire, malgré la libéralisation des prix, Le tarif RC est fixé à l'ancien niveau réglementaire.

En 2015, le CA d'affaire de la RC a représenté 86,1% du chiffre d'affaire globale pour l'usage automobile. Et la rentabilité du portefeuille a été portée essentiellement par la RC.

Les Garanties Annexes Facultatives

En complément de la RC obligatoire, les compagnies d'assurance proposent une panoplie de garanties permettant une meilleure protection (garantie incendie, vol, dommage tous accidents, dommage collision...), ces garanties annexes représentent 13,8% du CA automobile en 2015.

I.2 Vision globale du Marché

Le secteur d'assurance au Maroc est en plein essor, Il présente un potentiel de développement important en termes d'offres et de volume. La branche automobile est une composante importante de ce secteur, En effet, elle s'accapare de 28,4% de part de marché comme en témoigne le tableau ci-dessous tiré du rapport annuel de la Fédération marocaine des sociétés d'assurances et de réassurances (FMSAR) au titre de l'exercice 2016.

Tableau 1: Structure du chiffre d'affaire du secteur d'assurance au Maroc

Structure du Chiffre d'Affaires			
	Chiffre d'Affaires	Contribution	Evolution 2015/2016
Assurances Vie et Capitalisation	14 292,6	40,7%	35,4%
Automobile	9 953,8	28,4%	4,6%
Accidents Corporels	3 652,8	10,4%	8,7%
Accidents du Travail	2 174,1	6,2%	4,0%
Incendie	1 318,4	3,8%	0,5%
Assistance - Crédit - Caution	1 331,1	3,8%	12,5%
Transport	578,0	1,6%	4,7%
Autres Opérations Non Vie	734,5	2,1%	4,7%
Responsabilité Civile Générale	550,2	1,6%	1,1%
Risques Techniques	329,4	0,9%	-16,3%
Acceptations en réassurance	187,0	0,5%	-29,7%
Total	35 101,9	100%	15,4%

En millions de dirhams

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

Evolution du Chiffre d'Affaires					
	2014	2015	2016	Evolution 2015/2016	Evolution 2014/2015
Assurances Vie & Capitalisation	9 399,1	10 560,8	14 295,8	35,4%	12,4%
Assurances Individuelles	5 641,3	6 308,5	7 591,1	20,3%	11,8%
Assurances de Groupes	2 061,4	2 106,5	2 222,5	5,5%	2,2%
Assurances Populaires	-0,02	-	-0,15	NS	-100,0%
Capitalisation	1 368,2	1 684,3	3 985,6	136,6%	23,1%
Contrats à Capital Variable	326,7	460,1	493,6	7,3%	41%
Acceptations Vie	1,4	1,4	3,2	135,5%	-5,8%
Assurances Non Vie	19 022,5	19 862,9	20 806,1	4,7%	4,4%
Accidents Corporels	3 224,0	3 359,5	3 652,8	8,7%	4,2%
Accidents du Travail	2 213,5	2 090,9	2 174,1	4,0%	-5,5%
Automobile	9 033,7	9 514,2	9 953,8	4,6%	5,3%
Responsabilité Civile Générale	509,3	544,4	550,2	1,1%	6,9%
Incendie	1 159,3	1 312,1	1 318,4	0,5%	13,2%
Risques Techniques	416,0	393,7	329,4	-16,3%	-5,4%
Transport	568,5	552,3	578,0	4,7%	-2,9%
Autres Opérations Non Vie	606,3	701,2	734,5	4,7%	15,7%
Assistance - Crédit - Caution	1 091,1	1 183,2	1 331,1	12,5%	8,4%
Acceptations Non Vie	200,9	211,5	183,8	-13,1%	5,3%
Total	28 421,6	30 423,7	35 101,9	15,4%	7,0%

Fédération Marocaine des Sociétés d'Assurances et de Réassurance - Mars 2017

Source D.03

Comme mentionné précédemment la branche automobile contribue fortement au chiffre d'affaire global du marché, cela est dû majoritairement à l'obligation de la garantie « responsabilité civile ».

I.3 Axa Assurance

Tableau 3: Position d'AXA assurance Maroc en terme du chiffre d'affaire Non vie

Assurances Non Vie (y compris les acceptations en réassurance)					
	2014	2015	2016	Evolution 2014/2015	Part marché
Atlanta	1 410,8	1 550,7	1 699,1	9,6%	8,2%
Axa Assistance Maroc	125,2	152,4	106,8	-29,9%	0,5%
Axa Assurance Maroc	2 762,0	2 795,2	2 730,3	-2,3%	13,1%
CAT	636,1	631,2	634,0	0,5%	3,0%
Coface Maroc	-	17,9	45,6	NS	0,2%
Euler Hermes ACMAR	96,3	108,3	116,5	7,5%	0,6%
Saham Assistance	298,4	333,2	469,5	40,9%	2,3%
MAMDA	759,9	846,9	857,6	1,3%	4,1%
Maroc Assistance Internationale	436,5	432,6	435,3	0,6%	2,1%
Marocaine Vie	68,5	83,0	82,4	-0,7%	0,4%
MATU	247,0	265,1	293,0	10,5%	1,4%
MCMA	471,2	551,8	594,0	7,6%	2,9%
Mutuelle Taamine Chaabi	-	-	-	-	-
RMA	2 726,2	2 874,7	3 001,1	4,4%	14,4%
Saham Assurance	3 309,0	3 410,3	3 617,8	6,1%	17,4%
Sanad	1 337,8	1 420,0	1 452,3	2,3%	7,0%
Wafa Assurance	3 059,0	2 985,0	3 271,3	9,6%	15,7%
Wafa Ima Assistance	133,8	176,7	209,0	18,3%	1,0%
Zurich Assurance Maroc	1 144,7	1 228,0	1 190,3	-3,1%	5,7%
Total	18 135,0	19 862,9	20 806,1	4,7%	100,0%

Fédération Marocaine des Sociétés d'Assurances et de Réassurance - Mars 2017

En millions de dirhams

AXA Assurance Maroc occupe dans la globalité des activités non vie la 4^{ème} place par rapport aux autres acteurs de marché avec une part de marché d'ordre de 13,1%. En Automobile Axa Assurance se positionne 4^{ème} avec 12% de part de marché.

*Chapitre 3 : Présentation des
données et statistiques descriptives*

I. Description des données

I.1. Base de données interne

Notre analyse porte sur la garantie RC automobile matériel. La base de données fournie couvre les cinq exercices de 2010 à 2014.

La base de données interne est une jointure entre la base de données production (contenant les caractéristiques des assurés, les caractéristiques du contrat, les primes) et la base de données sinistres (contenant les charges, nombre de sinistres et les caractéristiques des sinistres).

Elle comporte le champ « code ville » qui contient les différentes communes présentes dans le portefeuille de la compagnie.

La base de données interne va servir d'une part à fournir l'information sur la sinistralité des différentes communes : nombres et charges des sinistres et d'autre part pour établir les modèles de tarification (fréquence coût moyen)

I.2. Base de données externe

La base de données externe contient les différentes communes d'AXA avec leurs code HCP et leurs caractéristiques sociodémographiques.

- **Sélection des variables externes explicatives**

Nous avons collecté diverses données sociodémographiques - jugées pertinentes - relatives à chaque commune, pour expliquer la sinistralité dans les différentes communes au Maroc.

Ces données sont les suivantes :

Présentation des données et statistiques descriptives

Tableau 4: Variables externes

	Variable	Signification	Calcul
	Densite	Densité de la population	Nombre d'habitants/surface
	Prop_tot15_59	Proportion de la population âgée entre 15 et 59 ans	Population âgée entre 15 et 59 ans / Population totale
	Tx_activite	Taux d'activité	population active (actifs occupés et chômeurs) âgée de 15 ans et plus / population totale du même âge
	Pourc_menagesPV	Pourcentage des ménages possédant une voiture	Nombre des ménages possédant une voiture / nombre des ménages
Structure selon le statut socio- professionnel	Pourc_salarieTot	Proportion des salariés	Nombre des salariés/population active occupée
	Pourc_indepTot	Proportion des indépendants	Nombre des indépendants/population active occupée
	Pourc_emplTot	Proportion des employeurs	Nombre des employeurs/population active occupée
Structure selon la branche d'activité	Agriculture_foret_peche	Pourcentage des actifs occupés dans le secteur de l'agriculture pêche et forêt	Nombre d'actifs travaillant dans le secteur de l'agriculture pêche et forêt/population active occupée
	Btp	Pourcentage des actifs occupés dans le secteur du bâtiment et travaux public	Nombre d'actifs travaillant dans le secteur du bâtiment travaux public/population active occupée
	Services	Pourcentage des actifs occupés dans le secteur des services	Nombre d'actifs travaillant dans le secteur des services/population active occupée

I.3 Jointure des deux bases de données

Afin de rendre le passage entre les 2 bases accessible, Nous avons utilisé une matrice de passage entre la base de données interne et celle externe.

Les différentes informations sur la sinistralité présente dans la BD interne (nombre de sinistres responsables, charge des sinistres, exposition, les primes) au niveau de chaque assuré ont été transposées au niveau communal. En effet les assurés ayant le même code ville sur la période analysée leurs différentes informations vont être agrégées et donc la base de données externe va être transformée en une base plus agrégée contenant pour chaque commune identifiée par son code ville ou son code HCP ses informations relatives à la sinistralité. Au final on est en possession d'une base de données contenant pour chaque commune (parmi les 229) ses caractéristiques sociodémographiques ainsi que ses informations relatives à la sinistralité.

La base de données ainsi construite va être utilisée pour réaliser la classification des communes en d'autres termes élaborer le zonier.

II. Analyses descriptives

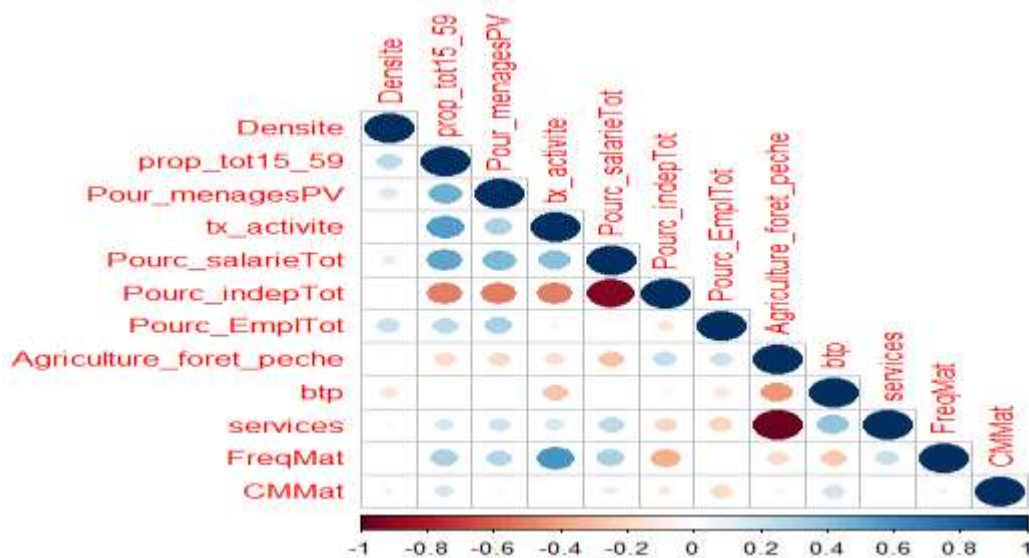
II.1. Base de données externe

II.1.1 Etude des corrélations entre les variables externes

A partir de la dernière base de données construite, on étudie les différentes corrélations entre les variables retenues à l'aide du coefficient de "Spearman "

Présentation des données et statistiques descriptives

Figure 3 : Triangle inférieur de la matrice de corrélation



Le graphique suivant met en relief les différentes corrélations entre les variables retenues et aussi avec notre indicateur de sinistralité : la fréquence des sinistres matériels. On tient à préciser que dans ce graphique les cercles en rouge réfèrent aux corrélations négatives, et celles en bleu aux corrélations positives, L'intensité de la couleur et la taille des cercles sont proportionnelles aux coefficients de corrélation.

1) Les corrélations les plus marquantes

- On constate que la proportion de la population âgée entre 15 et 59 est **très corrélée positivement** avec le pourcentage des ménages possédant une voiture, avec le taux d'activité et avec la proportion des salariés. En effet cette tranche d'âge représente la tranche la plus active dans la population est donc celles susceptibles d'avoir un emploi et par la suite posséder une voiture pour se déplacer au travail.
- La proportion des indépendants est **très corrélée négativement** à la proportion des salariés et au pourcentage des ménages possédant une voiture. Ce qui est logique, puisque c'est la proportion de la population des non salariés travaillant pour leurs propres comptes et constituée majoritairement des indépendants moyennement aisés.
- Le pourcentage des actifs occupés travaillant dans le secteur des services, et dans le secteur bâtiment et travaux publics sont **très corrélés négativement** au pourcentage des actifs occupés travaillant dans le secteur de l'agriculture Forêt et pêche. En effet plus

une commune est à dominance du secteur agricole moins d'activités du secteur primaire et secondaire y aura.

2) Corrélation des variables avec la fréquence des sinistres matériels

La fréquence des sinistres matériels est **corrélée positivement** avec la densité, la proportion de la population âgée entre 15 et 59 ans, le pourcentage des ménages possédant une voiture, taux d'activité et la proportion des salariés, le pourcentage d'actifs occupés travaillant dans le secteur des services.

- On pense que plus la densité de la population est grande dans une commune, plus il est probable d'avoir des sinistres et donc une fréquence de sinistres grande.
- Cette tranche d'âge 15 ans et 59 ans surtout de 18 ans à 59 ans est la tranche d'âge la plus présente dans le portefeuille d'assurance et donc la plus exposée au risque. Plus cette tranche est représentative dans une commune, plus il y aura de sinistres enregistrés.
- Plus le nombre de voitures est élevé dans une commune, plus le risque d'avoir des sinistres augmente.
- Le taux d'activité représente la part des actifs travaillant dans une commune, et donc c'est la part de la population qui est susceptible d'utiliser le plus souvent sa voiture pour se déplacer et donc plus d'exposition au risque d'avoir un sinistre. Plus ce taux est élevé plus la fréquence de sinistres est grande.
- Une proportion assez grande dans une commune (% d'actifs occupés travaillant dans le secteur des services) signifie que plus la population est tournée vers les activités de production des biens et des services, et donc plus exposée au risque d'avoir des sinistres lors de ces déplacements pour le travail.

Toutefois la fréquence matérielle est **corrélée négativement** à la proportion des indépendants, au pourcentage des actifs occupés travaillant dans le secteur de l'agriculture Forêt et pêche et dans le secteur bâtiment et travaux publics.

- Plus cette proportion (% d'actifs occupés travaillant dans le secteur de l'agriculture Forêt et pêche) est grande dans une commune, plus cette commune est tributaire du secteur agricole, et donc le nombre de voitures est faible ce qui diminue la probabilité d'avoir des sinistres.

Présentation des données et statistiques descriptives

- Cette proportion (% d'actifs occupés travaillant dans le secteur du bâtiment et travaux publics et la proportion des indépendants) est constituée majoritairement de la catégorie à plus au moins faible revenu et donc ne peuvent même pas se procurer une voiture et donc par la suite une fréquence de sinistres faible.

3) **Corrélation des variables avec le coût moyen des sinistres matériels**

On constate que le coût moyen en comparaison avec la fréquence est un peu moins corrélé aux variables externes.

Le coût moyen est **corrélé positivement** avec la proportion de la population âgée entre 15 et 59 ans, le pourcentage d'actifs occupés travaillant dans le secteur du bâtiment et travaux publics, la proportion des salariés et à la densité .

Il est **corrélé négativement** à la proportion des indépendants, la proportion des employeurs et au pourcentage des actifs occupés travaillant dans le secteur de l'agriculture Forêt et pêche.

L'explication des corrélations positives ou négatives qui vont dans le même sens avec la fréquence est la même avec le coût moyen. Cependant en ce qui concerne la corrélation positive avec le pourcentage d'actifs occupés travaillant dans le secteur du bâtiment et travaux publics, nous avons vu auparavant que cette variable est corrélée négativement avec la fréquence et donc plus ce pourcentage est élevé plus la fréquence des sinistres est faible toutefois il se peut que la charge engendrée par ces sinistres soit très élevée et qui peut expliquer une telle corrélation. Quant à la corrélation négative avec la proportion des employeurs pourrait éventuellement s'expliquer par le fait que cette catégorie représente la catégorie la plus cultivée et donc la plus averse au risque.

II.2. Base de données interne

II.2.1 Epurement de la base de données

Le processus d'épurement et de traitements des bases de données, objet de l'étude, s'avère nécessaire avant toute mise en œuvre des résultats. Nous présentons dans le tableau suivant l'ensemble des anomalies détectées et les solutions adoptées pour les résoudre.

Présentation des données et statistiques descriptives

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

	Age du Véhicule	Age assuré	Puissance	CRM	Sexe	Carburant
Nombre de valeurs aberrantes	14705	11142	17520	9372	0	0
Pourcentage de valeurs aberrantes	0,94%	0,91%	1,10%	0,59%	0	0
Nombres de valeurs manquantes	1020	0	0	0	9036	3059
Pourcentage de valeurs manquantes	0,06%	0,00%	0,00%	0,00%	0,51%	0,17%

Afin de rendre notre base de données exploitable, Nous avons envisagé la solution d'affecter les observations aberrantes et manquantes à la classe la plus risquée. Du fait que dans la majorité des cas les assurés ne déclarant pas une information sont eux qui font preuve d'une sinistralité si élevée.

II.2.2 Etude des corrélations entre les variables tarifaires

Les variables tarifaires retenues sont les suivantes :

Tableau 6: Les variables tarifaires retenues

Libellé	Variable
G	Le type de combustion
F	Le sexe du conducteur
ageass	L'âge du conducteur
VEPUI	La puissance fiscale
age_vehic	L'âge du véhicule
Zone	La zone géographique

Présentation des données et statistiques descriptives

Tableau 7: Sortie SAS corrélation entre les variables tarifaires quantitatives

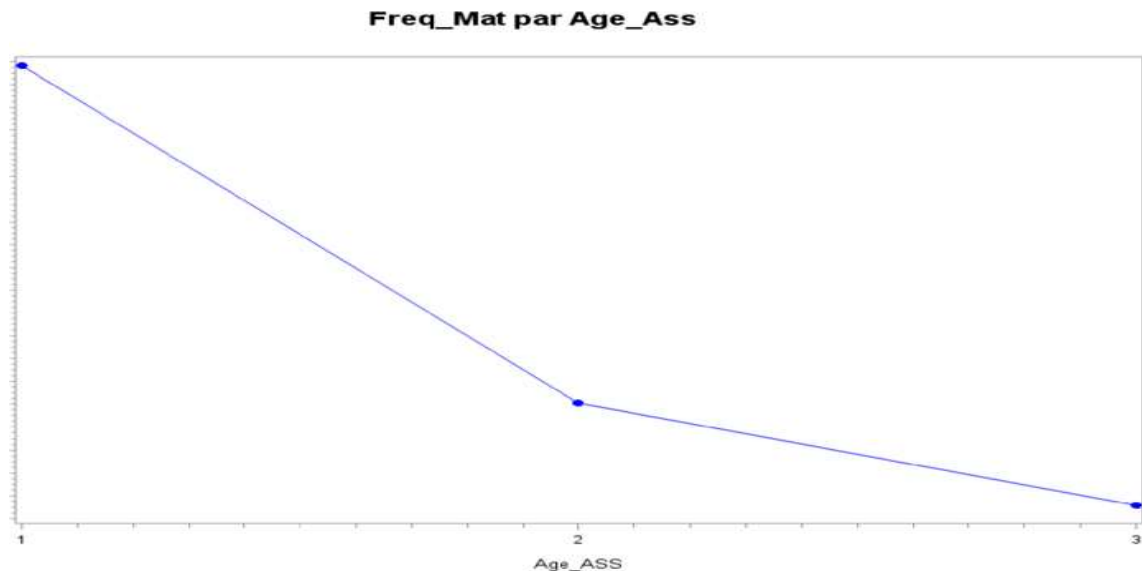
Coefficients de corrélation de Pearson, N = 1890904 Proba > r sous H0: Rho=0				
	VEPUI	ageass	age_vehic	VETXCR
VEPUI VEPUI	1.00000	0.09326	0.04755	-0.00999
ageass	0.09326	1.00000	-0.07284	-0.11833
age_vehic	0.04755	-0.07284	1.00000	0.03957
VETXCR VETXCR	-0.00999	-0.11833	0.03957	1.00000

En utilisant la procédure « PROC FREQ », qui nous permet d'effectuer la corrélation entre les variables quantitatives On constate que les corrélations entre nos variables sont faibles.

II.2.3 Distribution de la fréquence matérielle en fonction des variables tarifaires

Nous avons discrétiser nos variables tarifaires afin de faciliter l'interprétation

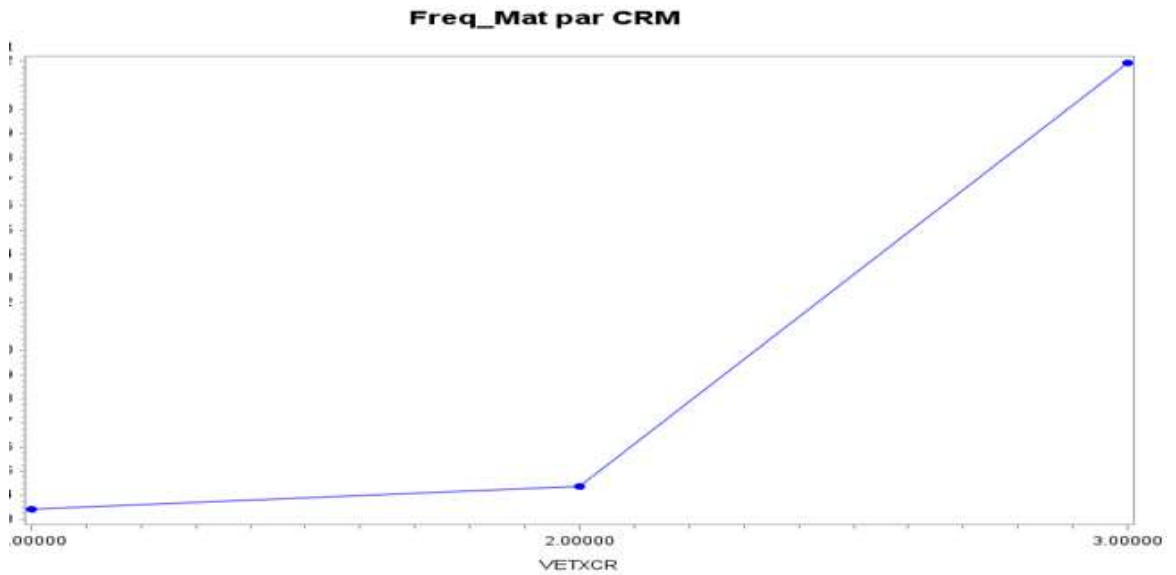
Figure 4.: Graphique représentant la répartition de la fréquence par âge de l'assuré



La première classe des jeunes conducteurs représente la classe ayant la plus grande fréquence matérielle. En effet les jeunes conducteurs sont encore novices et donc font plus de sinistres que les autres classes d'âges. Au fur et à mesure que l'âge augmente la fréquence des sinistres diminue.

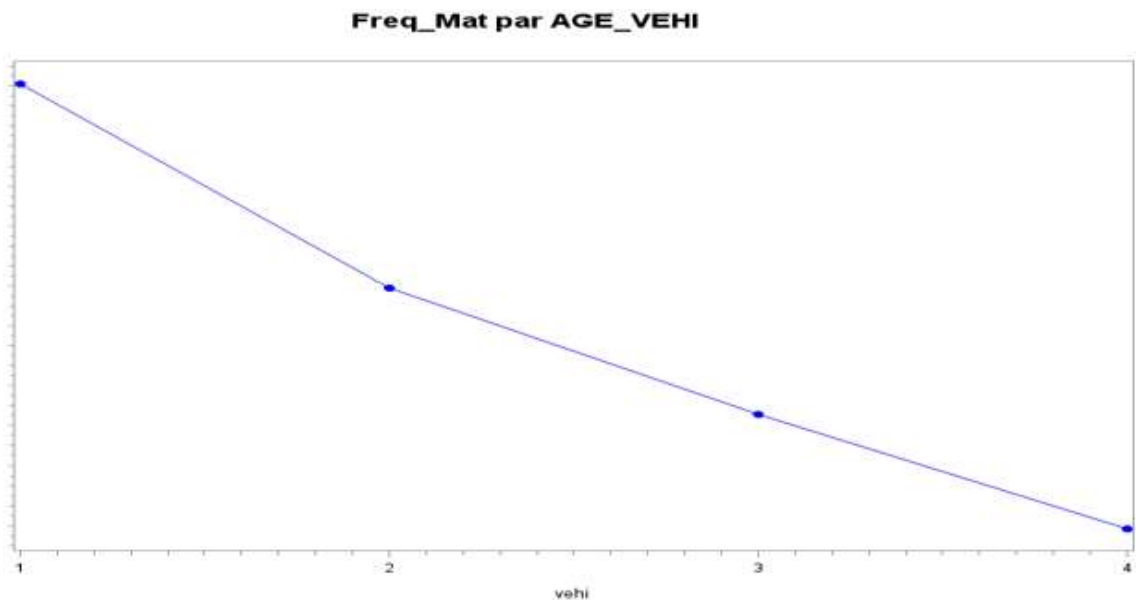
Présentation des données et statistiques descriptives

Figure 5: Répartition de la fréquence en fonction du CRM



Pour la classe où le CRM est majoré, la fréquence des sinistres est grande. En effet un CRM majoré signifie que l'assuré a eu des sinistres engageant ou susceptible d'engager totalement ou partiellement sa responsabilité durant la période d'assurance et donc c'est logique que la fréquence soit assez grande et vice versa.

Figure 6 : Répartition de la fréquence en fonction de l'âge du véhicule



La fréquence matérielle diminue avec l'âge du véhicule. Une explication possible de ce constat, Les véhicules neufs ont tendance à être conduits plus souvent et pour de longues

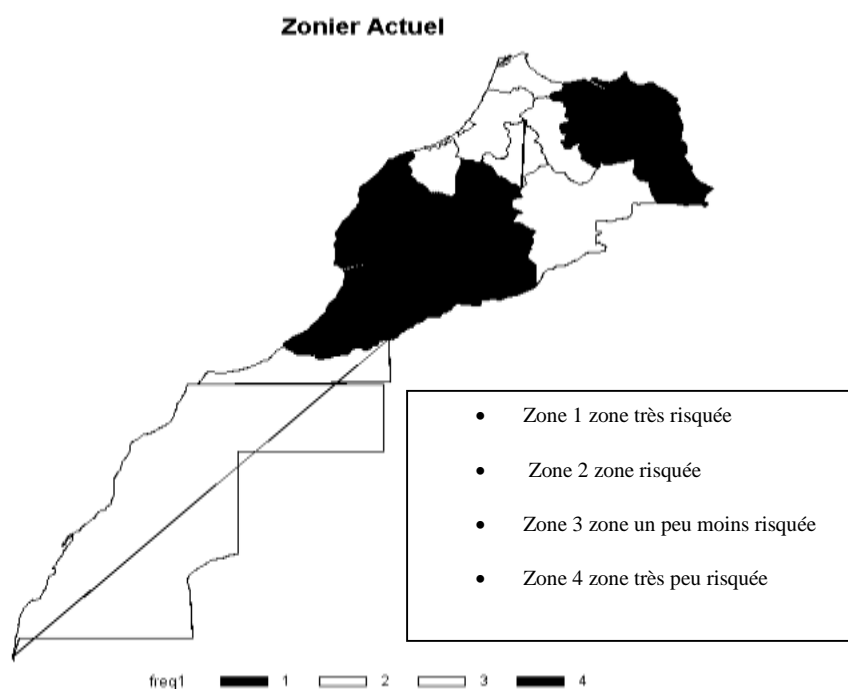
distances, Alors que les véhicules plus âgés ont tendance à être conduits moins souvent et à des distances plus courtes.

III. Etude préliminaire : analyse du zonier actuel

III.1. Présentation du zonier actuel

Le Zonier actuel est réalisé à dire d'expert c'est-à-dire que la compagnie a classé les communes en 4 zones de risques - de la 1^{ère} zone considérée comme la zone la moins risquée jusqu'à la 4^{ème} zone définie comme étant la zone la plus risquée - d'une manière subjective basée sur la croyance des experts sur le comportement de sinistralité des différentes communes. Toutefois Le zonier actuel nécessite d'être analysé afin d'en mesurer sa performance. On étudie donc l'évolution du zonier actuel sur la période analysée de 2010 à 2014 en prenant les deux indicateurs de sinistralité retenus dans ce mémoire à savoir la fréquence matérielle et le coût moyen matériel.

Figure 7: Carte du risque actuel : Répartition en 4 zones à risque

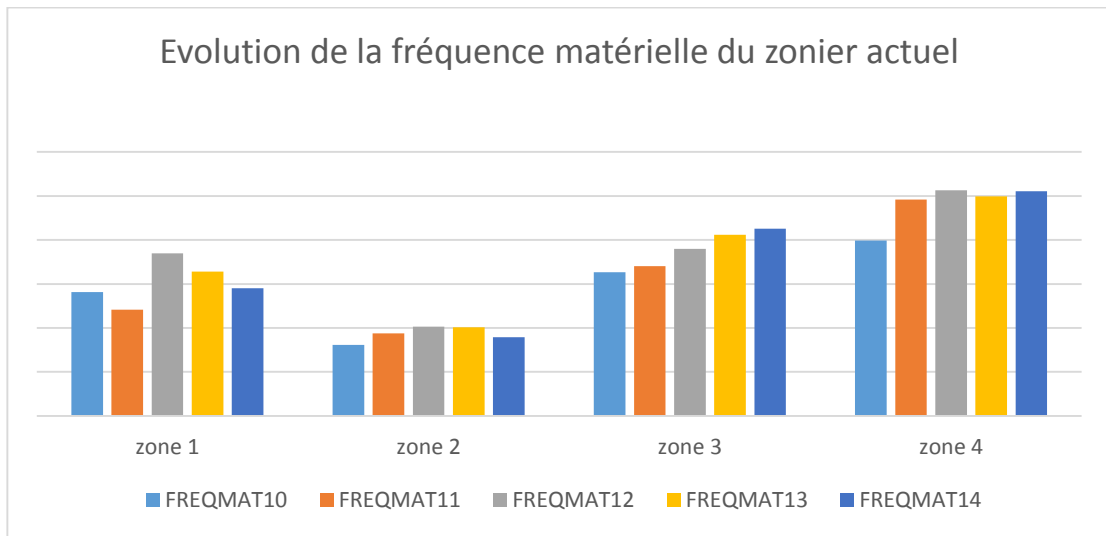


III.2. Etude de la fréquence et du coût moyen des sinistres et du ratio S / P

Indicateur de sinistralité : Fréquence matérielle

Présentation des données et statistiques descriptives

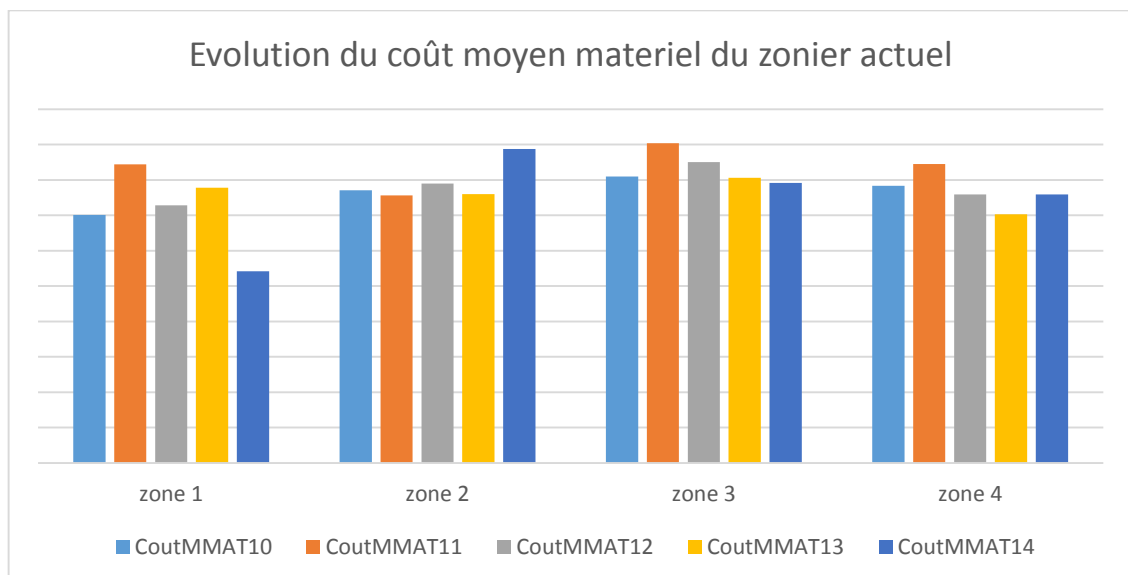
Figure 8: Evolution de la Fréquence des sinistres matériels sur la période 2010 à 2014



D'après l'analyse de ce graphique on conclut que les fréquences sont visiblement stables. Elles conservent le même ordre de grandeur d'une année à une autre pour une même zone. Cependant le zonier semble moins cohérent sur la période analysée. En effet, on constate une baisse de la fréquence des sinistres matériels dans la 2^{ème} zone qui se confirme pour chaque année de la période analysée

Indicateur de sinistralité le coût moyen matériel

Figure 9: Evolution du coût moyen des sinistres matériels sur la période 2010 à 2014

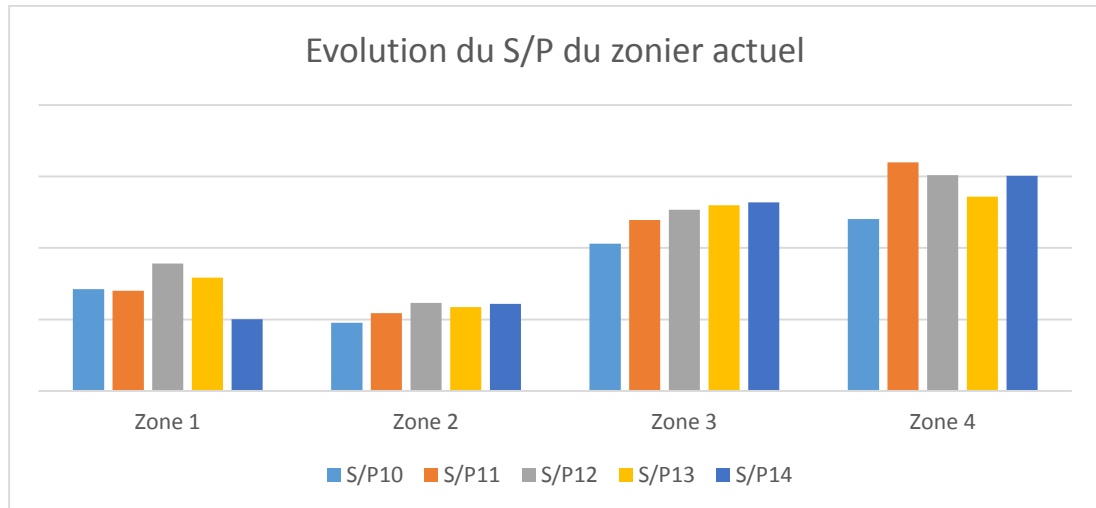


D'après le graphique ci-dessus on constate que tout comme la fréquence matérielle, le coût moyen matériel conserve le même ordre de grandeur d'une année à l'autre pour une même

zone. Toutefois nous remarquons que le coût moyen matériel en zone 4 est moins élevé qu'en zone 3. Cette tendance baissière se confirme pour chaque année tout au long de la période étudiée.

Indicateur de sinistralité le S/P

Figure 10: Evolution du S/P des sinistres matériels sur la période 2010 à 2014



Le ratio S/P est indicateur important permettant de mettre en évidence d'éventuelles sur-tarififications ou sous-tarififications selon les zones. En effet, On constate que la zone 2 un peu plus risquée que la 1^{ère} zone est sous tarifée par rapport à cette dernière.

Conclusion

Le zonier actuel évalué par les différents indicateurs de sinistralités révèle qu'il comporte des incohérences au regard de la sinistralité actuelle, ce qui confirme notre démarche à explorer de nouvelles méthodes afin d'élaborer un nouveau zonier en prenant compte de nouveaux critères : les caractéristiques sociodémographiques afin de faire une segmentation plus fine (par commune) et pour assurer un maximum de précision.

*Chapitre 4 : Elaboration du zonier
par les méthodes du Maching
Learning*

I. Méthodologie de Travail

I.1. Méthodes utilisées

Comme mentionné au début, le but de la classification des communes dans ce mémoire est à double finalité : une classification des communes pas seulement basée sur les similarités de leurs caractéristiques sociodémographiques mais aussi une classification basée sur leur comportement vis-à-vis de la sinistralité. Une 1^{ère} approche de type supervisé par arbres de décisions a été choisie dans ce cas pour 3 raisons :

Les arbres de décisions constituent une approche d'apprentissage supervisé. En effet, elle consiste à effectuer la modélisation en la présence de variable à expliquer contrairement aux méthodes de Clustering qui sont des approches de type non supervisé effectuant une classification des individus dans des classes homogènes purement basée sur leurs variables explicatives et en absence de variable à expliquer.

Les arbres de décisions permettent de traiter tout type de variables que ça soit quantitatif ou qualitatif Alors que d'autres techniques de classification (réseaux de neurones par exemple) nécessitent des données numériques.

Les arbres de décisions ont un grand avantage celui de la facilité d'interprétation des résultats. Ils classifient les observations sous la forme d'un arbre, dans lequel chaque nœud est soit un nœud final, soit un nœud interne (nœud de décision). Le nœud de décision correspond à un groupe d'observations auquel une autre division doit être effectuée. Si aucune division supplémentaire ne peut être faite, un nœud de décision devient un nœud final. Chaque nœud final peut être défini par un ensemble de règles selon lesquelles les divisions ont été réalisées. Par conséquent, le résultat de la technique DT est un certain nombre de segments (nœuds finaux) et leurs définitions.

Nous avons choisie dans ce mémoire l'algorithme CART pour faire notre Classification. CART divise les communes dans des groupes basées sur les similarités de leurs variables explicatives ainsi que de leur variable à expliquer. Cet algorithme est présenté en détail dans la section II.1.

La 2^{ème} approche proposée dans le cadre de ce mémoire est la méthode des forêts aléatoires une méthode d'ensemble qui agrège un grand nombre d'arbres de décisions (CART)

et qui vient combler les différents désavantages des arbres de décisions. Cette approche est présentée en détail dans la section II.2.

Nous adoptons pour notre objectif de classification une approche semi-supervisée qui détecte les Clusters qui apparaissent dans la classification de nos communes, en se basant sur une caractéristique importante des Random Forests à savoir « les valeurs de la matrice de proximité » (détaillée ci-dessous).

I.2. Choix de l'indicateur de sinistralité

Le choix de l'indicateur de sinistralité constitue une étape cruciale pour élaborer le Zonier, En effet il permettra de déterminer si le risque de faire un sinistre en RC est élevé ou non dans une zone donnée.

Le Zonier à mettre en place doit être stable dans le temps. En effet un zonier stable dans le temps serait fiable et pertinent dans plusieurs années et n'est pas nécessaire de le reproduire à chaque fois.

Dans la section traitant le zonier actuel nous avons met en évidence que les fréquences ainsi que les coûts moyens des sinistres matériels gardent le même ordre de grandeur et donc sont stables d'une année à une autre. Par conséquent on effectue un zonier par fréquence matérielle et un autre par coût moyen.

II. Partie théorique

II.1. 1^{ère} approche : Arbre de Régression (CART)

La méthode CART (Classification and Regression Trees) est une méthode non paramétrique d'apprentissage statistique, récursive qui se base uniquement sur les données pour réaliser des modèles de régression dans le cas où la variable réponse est quantitative ou bien des modèles de classification dans le cas où la variable à expliquer est qualitative.

La méthode CART découpe l'espace étudié en R régions. Au sein desquels la valeur de la variable modélisée est homogène.

En se basant sur un critère de division binaire récursif pour trouver la meilleure séparation des nœuds dans l'espace des variables descriptives indépendantes

Et donc au niveau de la racine de l'arbre ainsi qu'au niveau de chaque nœud intermédiaire de l'arbre partent deux branches distinctes. Les observations sont successivement

découpées en deux parties, créant à chaque nouvelle itération une partition plus précise de la population. Cette procédure de découpage est continuée jusqu'à ce qu'une règle d'arrêt s'applique. Dans chacune de ces régions une bonne valeur prédictive de Y est sa moyenne (calculée sur l'ensemble d'apprentissage).

II.1.1. Algorithme

L'algorithme est expliqué ci-dessous pour le cas où l'espace étudié est R^p :

- 1) Débuter avec $R=1$ une seule région. On a alors $P = \{R\} = R^p$
- 2) On redéfinit R en $R_{droite} R_{gauche}$ où

$$R_{gauche} = R * R * R * \dots (-00, d] * R * \dots R$$

$$R_{droite} = R * R * R * \dots (-00, d] * R * \dots R$$

seuil d, où d appartient à l'ensemble des valeurs prises par la variable à expliquer. Le choix de la variable ainsi que du seuil d sont déterminés de façon à effectuer la minimisation

$$\min_{\alpha_1} \sum_{x_l \in R_1(j,d)} (y_l - \alpha_1)^2 + \min_{\alpha_2} \sum_{x_l \in R_2(j,d)} (y_l - \alpha_2)^2$$

$R_1(j, d) = R_{gauche}$, $R_2(j, d) = R_{droite}$. Ces deux régions dépendent du choix de la variable et du point de séparation-seuil-d utilisés pour effectuer la séparation.

- 3) Affiner la partition comme en 2) via le découpage d'une des régions de la partition actuelle. Cela implique donc de rechercher la région à affiner, ainsi que la variable à expliquer, et le point de séparation. On met ensuite à jour la partition

$$\mathcal{P}_n = \mathcal{P}_{n-1} \setminus \text{cellule de partition affinée} \cup \{R_{droite}, R_{gauche}\}$$

- 4) On itère alors l'étape 3 pour un nombre important de cellules de partition.

A ce point, on aura construit un arbre maximal ce qui veut dire que la division des nœuds a été effectuée jusqu'aux dernières observations dans la base d'apprentissage.

II.1.2. Nécessité de l'élagage de l'arbre

L'arbre maximal peut être très grand où chaque variable à expliquer peut résulter dans un nœud séparé. Cependant le découpage doit s'arrêter et donc on doit élaguer l'arbre pour éviter le phénomène de sur-apprentissage.

L'élagage de l'arbre s'effectue selon un critère d'arrêt est donc une méthode pour trouver un équilibre entre un arbre trop complexe et sur apprend les données et un arbre trop simple et par conséquent supprime les détails importants. Le but est de trouver un compromis entre l'impureté (mesurée par l'erreur de généralisation) et la complexité de l'arbre.

Parmi les méthodes utilisées afin d'élaguer l'arbre on trouve la méthode « élagage coût-complexité » (cost-complexity pruning), qui consiste à créer un arbre de taille élevée dans un premier temps, puis à l'élaguer selon un critère de complexité $C_\alpha(T)$ où T correspond à l'arbre étudié, avec
$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m * Q_m(T) + \alpha * |T|$$

Dans cette formule, N_m correspond au nombre d'observations appartenant au nœud m, $|T|$ correspond au nombre de nœuds terminaux de l'arbre, et Q_m est l'estimation de la MSE au sein du nœud m :
$$Q_m(T) = \frac{1}{N_m} * \sum_{x_i \in R_m} (y_i - c_m)^2$$

α un paramètre à optimiser réalisant un arbitrage entre la taille de l'arbre et la qualité de l'ajustement. Pour α grand, l'arbre conservé est de faible taille, et vice versa.

La procédure consiste à réaliser des sous arbres T de l'arbre maximal, en enlevant à chaque fois un certain nombre de nœuds internes à l'arbre qui produisent la plus petite augmentation du terme $\sum_{m=1}^{|T|} N_m * Q_m(T)$, avec pour objectif de trouver le sous arbre optimal $T_\alpha(T)$ qui minimise $C_\alpha(T)$ pour chaque valeur de α , c'est-à-dire $T(\alpha) = \operatorname{argmin}_\alpha (C_\alpha(T))$ et on continue l'élagage jusqu'à ce qu'il ne reste que la racine de l'arbre. Cette procédure produit ainsi une série de sous arbres. Le sous arbre recherché est contenu dans cette série d'arbres. Le choix sur la valeur de α à utiliser se fait en mettant en œuvre une procédure de validation à l'issue de laquelle nous choisissons la valeur de α qui minimise l'erreur de validation.

II.1.3. Validation croisée

Comme a été vu dans la partie précédente, nous utilisons la validation croisée dans le cadre de l'élaboration de modèles CART afin d'optimiser le paramètre α

Les données sont découpées de façon aléatoire en K parties de tailles aussi égales que possible.

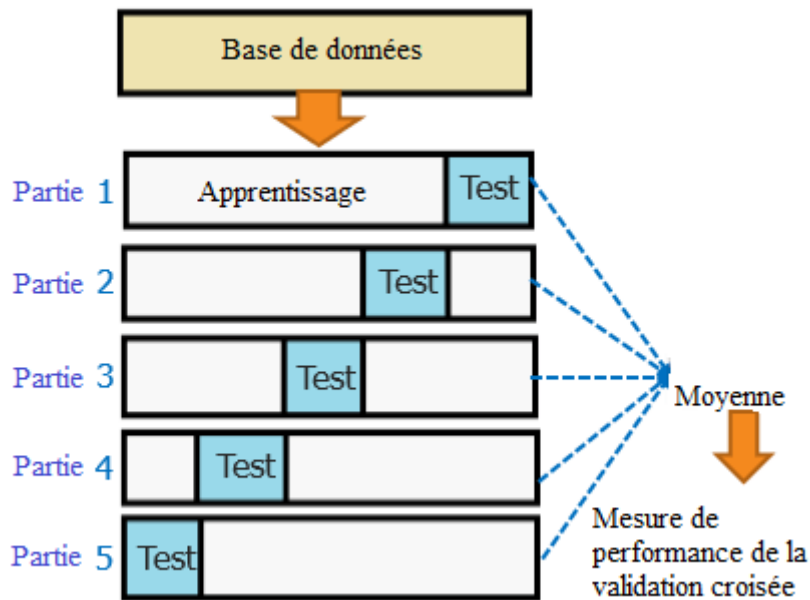
La méthode utilise alors les données dans une des partie B_k comme données test, et les autres comme données d'apprentissage.

Le modèle de régression est alors ajusté aux données K fois, et pour chacune d'entre eux, on calcule le terme $\frac{1}{|B_k|} * \sum_{x_i \in B_k} (Y_i - \hat{m}^{(-B_k)}(X_i))^2$

\hat{m}^{-B_k} correspond à la prédiction de la variable à expliquer sur les données d'apprentissages en enlevant la partie B_k . La performance de validation croisée se mesure par

$$\frac{1}{K} * \sum_{k=1}^K \frac{1}{|B_k|} * \sum_{i \in B_k} (Y_i - \hat{m}^{(-B_k)}(X_i))^2$$

Figure 11: Séparation des données pour la méthode validation croisée ex du 5-fold



II.2. 2^{ème} approche : Random Forest (Forêts aléatoires)

Bien que la montée du «Big data» ait rendu les algorithmes d'apprentissage en machine plus visibles et pertinents, ils sont encore largement considérés comme des «boîtes noires» qui ne conviennent seulement que pour faire la prédiction. Toutefois dans ce mémoire nous soutenons que ladite méthode ne peut être utilisée que pour cette finalité, en mettant l'accent sur son application pratique pour réaliser notre Zonier.

II.2.1. Présentation de la méthode

Forêt aléatoire ou bien Random Forest est une méthode statistique d'ensemble non paramétrique qui consiste à agréger un ensemble d'arbres aléatoires indépendants pour faire soit de la prédiction lorsque la variable à expliquer est quantitative ou bien de la Classification lorsque la variable à expliquer est qualitative.

Cette méthode fut introduite par Breiman pour combler les désavantages des arbres de décisions et de régression (CART) à savoir les valeurs prédites ont une grande variance, c'est-à-dire qu'il existe un risque de surapprentissage. Les valeurs prédites peuvent être instables, produisant donc des structures d'arbres différents lorsque des modifications sont apportées aux données utilisées.

L'idée centrale du Random Forest et les méthodes d'ensemble en général (Bagging, Boosting) est de faire diminuer la variance des prédictions, en générant plusieurs modèles et donc on explore grandement l'espace des solutions et ensuite en les combinant on récupère un prédicteur qui rend compte de cette exploration. Afin d'obtenir une variété de modèles qui ne surapprennent pas la base de données, chaque modèle est ajusté sur la base d'un échantillon Bootstrap. Un échantillon Bootstrap est un échantillon de même taille que la base de données d'origine mais conçu avec remplacement. Par conséquent, chacun de ces échantillons excluent une partie des données, qui est appelée données «out-of-bag» (OOB). Le Random Forest a une autre particularité celle du tirage aléatoire parmi les variables explicatives. Cela signifie que, au lieu de choisir la division à partir de toutes les variables, seulement un sous-ensemble aléatoire des variables explicatives est sélectionné. Cela peut sembler contre-intuitif au début, mais il a pour effet de diversifier les découpages à travers les arbres. S'il existe des variables très importantes, ils pourraient dissimuler l'effet de prédicteurs plus faibles car l'algorithme recherche la division qui entraîne la plus grande réduction de la fonction de perte. Si à chaque division seulement un sous-ensemble de prédicteurs est disponible pour être choisi, les prédicteurs plus faibles ont la possibilité d'être sélectionnés plus souvent. Ce qui permet à un très grand ensemble de variables explicatives d'être analysé. Et par conséquent les arbres obtenues seront différents les uns des autres ce qui évite le problème de corrélation des arbres et donc d'avoir une variance des prédictions plus faible.

II.2.2. Avantages de la méthode

Les forêts aléatoires détectent l'interaction et la non-linéarité sans pré-spécification, elles ont une faible erreur de généralisation dans les simulations et dans de nombreux problèmes du monde réel, et peuvent être utilisées avec beaucoup de corrélation entre les prédicteurs, même s'il existe plus de prédicteurs que d'observations.

II.2.3. Algorithme

Afin de construire une forêt aléatoire, un CART est appliqué à chacun de ces échantillons Bootstrap. Ensuite les prédictions par chaque arbre sont effectuées en utilisant les données OOB. Ainsi chaque observation aura une prédiction faite par chaque arbre sans cette dernière en d'autre terme où cette observation fût OOB "en dehors du bootstrap". Les valeurs prédites pour chaque observation sont combinées pour produire une estimation agrégée qui a une variance plus petite que celle prédite par un seul CART. Le prédicteur d'ensemble dans le cas où la variable réponse est de type quantitative est la moyenne des prédictions, sinon dans le cas qualitatif c'est le vote majoritaire.

II.2.4. L'erreur OOB

L'algorithme du Random Forest calcule en plus du prédicteur une estimation de l'erreur de généralisation : l'erreur Out-Of-Bag (OOB).

La méthode de calcul de cette erreur est la suivante :

Soit une observation (x_i, y_i) de l'échantillon d'apprentissage

- On considère l'ensemble des arbres construits sur les échantillons Bootstrap ne contenant pas cette observation.
- On agrège les prédicteurs de ces arbres (ne contenant pas cette observation) pour construire \hat{y}_i prédicteur de y_i .
- On calcule l'erreur quadratique moyenne

Dans le cas de la régression (MSE) : $\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$

Dans le cas de classification la proportion d'observations mal classées (*Mean decrease in accuracy*) : $\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\hat{y}_i \neq y_i}$

L'avantage de l'erreur OOB par rapport aux autres estimateurs classiques (échantillon test, validation croisée) c'est qu'elle ne nécessite pas de découper l'échantillon d'apprentissage. Ce découpage est en quelque sorte inclus dans la génération des différents échantillons bootstrap. Car par construction les données prédites sont des données qui n'ont pas été rencontrées au préalable par le prédicteur utilisé. Cependant, pour chaque observation ce n'est pas le même ensemble d'arbres qui est agrégé. Et donc cette erreur estime l'erreur de généralisation d'une forêt, mais elle n'utilise jamais les prédictions de la forêt elle-même, mais plutôt celles de prédicteurs qui sont des agrégations d'arbres de cette forêt.

II.2.5. L'importance des variables

1) Importance agrégée

Pour un objectif d'interprétation des données ainsi que pour la sélection des variables effectives qui expliquent le lien entrée-sortie. Un indice d'importance spécifique aux forêts est calculé.

Pour chaque échantillon Bootstrap l parmi les q échantillons

- On détermine son échantillon OOB_l associé (ensemble des observations n'appartenant pas à l'échantillon Bootstrap)
- Calcul de l'erreur sur l' OOB_l par l'arbre construit sur l'échantillon Bootstrap.
- On permute aléatoirement les valeurs de la variable dans l'échantillon OOB_l . On obtient un nouvel échantillon
- On calcule encore une fois l'erreur sur l'échantillon perturbé \widetilde{OOB}_l^j

L'importance de la variable X^j est obtenue comme suit

$$VI(X^j) = \frac{1}{q} \sum_{l=1}^q \left(\widetilde{errOOB}_l^j - errOOB_l \right)$$

Plus les permutations aléatoires engendrent une forte augmentation de l'erreur plus la variable est importante.

2) Importance locale

Par Forêt aléatoire on peut aussi mesurer l'importance de chaque variable dans chaque observation. Cette mesure s'appelle : l'Importance locale.

Le calcul de l'importance locale pour chaque variable dans chaque observation s'effectue de la même manière par l'algorithme précédent à une différence près au lieu de prendre de façon agrégée l'Out Of Bag de l'échantillon on prend en particulier l'Out Of Bag de cette observation.

II.2.6. Dépendances partielles

Les Dépendances partielles permettent de visualiser la relation entre la variable à expliquer et une ou bien plusieurs variables explicatives comme détectées par le Random Forest.

L'algorithme des dépendances partielles est le suivant :

Pour chaque valeur x^j de la variable à expliquer X^j $\{V = \{x^j\}_{i \in \{1, \dots, n\}}, |V| = K\}$, une nouvelle base de données k est créée où toutes les observations se voient assigner la même valeur x^j pour cette variable d'intérêt alors que les autres variables restent inchangées. Ensuite cette base de données est modélisée par une forêt aléatoire par la suite on obtient une prédiction pour chaque observation \widehat{y}_i^k . Et à la fin on fait la moyenne sur ces valeurs prédites pour chaque nouvelle base de données $\widehat{Y}^k = \sum_{i=1}^n \widehat{y}_i^k$

On visualise la relation entre y et la variable explicative en traçant \widehat{Y} en fonction des valeurs de la variable d'intérêt, ainsi on peut visualiser les non-linéarités potentielles entre la variable à expliquer et la variable explicative.

II.2.7. Sélection des variables

Pour un but d'interprétation comme dans notre cas on cherche à déterminer les variables importantes fortement reliées à la variable à expliquer

La Procédure de sélection des variables importantes s'effectue comme suit

1. Etape : Classement par VI index

Trier les variables selon leurs moyennes d'indice d'importance dans un ordre décroissant.

2. Etape : Elimination préliminaire

On trace la courbe de l'écart type de l'indice d'importance des différentes variables explicatives.

On définit le Seuil comme étant la plus petite valeur prédite par l'algorithme CART modélisant cette courbe.

On ne retient que les variables ayant une moyenne d'importance supérieure au seuil.

3. Etape : Sélection des variables pour le but d'interprétation

On calcule l'erreur OOB (ainsi que l'écart type de l'erreur OOB) des modèles emboîtés en commençant par celui contenant que la variable la plus importante et en terminant par celui contenant toutes les variables déterminées dans l'étape précédente.

Le modèle retenu est celui ayant la moyenne erreur OOB inférieur à la plus petite moyenne OOB augmentée de l'écart type de l'erreur OOB du modèle atteignant la plus petite erreur OOB)

II. Elaboration du Zonier par les 2 approches

II.1 Zonier par la méthode des arbres de décisions (CART)

En raison de la visualisation directe et la facilité d'interprétation de l'arbre de décision (CART) Le zonier est obtenu à partir des segments -nœuds finaux- générés par l'arbre, où chaque segment regroupe un nombre de communes qui ont des caractéristiques sociodémographiques et un comportement de sinistralité similaire.

II.2. Zonier par la méthode des forêts aléatoires

Comme on peut le voir dans la section précédente, le zonier par un seul arbre est relativement facile à déduire. Cependant comment est-il possible d'interpréter et donc d'élaborer le zonier à partir d'un millier d'arbres (constituant par excellence la spécificité des Random Forests), chacun ne correspondant qu'à un seul échantillon de données et à l'utilisation d'un échantillon aléatoire de variables explicatives à chaque division ?

Pour répondre à cette question plusieurs méthodes ont été développées pour extraire plus d'informations que de simples prédictions et rendre donc les forêts interprétables d'une manière substantiellement pertinente avec des mesures d'importance variable, dépendance partielle, matrices de proximité.

Comme mentionné précédemment, On a adopté une approche semi supervisée se basant sur une fonctionnalité intéressante générée par les forêts aléatoires : la matrice de proximité afin de l'interpréter nous avons construit une matrice distance par la suite on a effectué un positionnement multidimensionnel sur cette matrice pour obtenir ses principales coordonnées

Dans le but de regrouper les communes en exploitant ces principales coordonnées, nous avons utilisé la classification ascendante hiérarchique en prenant la distance euclidienne entre ces principales coordonnées.

II.2.1. Matrice de proximité

Nous utilisons Les forêts aléatoires pour comprendre les similarités entre les communes dans l'espace des variables expliquées et en fonction de la variable à expliquer. Étant donné que les observations ayant des valeurs x similaires voyagent de la même façon, les découpages obtenus avec les arbres qu'avec des observations ayant des valeurs dissemblables, la cooccurrence dans les nœuds terminaux est une mesure appropriée de la similarité. Dans cette logique une **matrice de proximité** est calculée. C'est une matrice $(n \times n)$ où chaque entrée donne le nombre de fois où l'observation i tombe dans le même nœud terminal que l'observation j normalisée par le nombre d'arbres dans la forêt.

A partir de la matrice de proximité nous calculons une matrice distance comme étant

$$\text{Matrice distance} = 1 - \text{Matrice de proximité } (P_{ij})$$

Par la suite on effectue un positionnement multidimensionnel (Multidimensional Scaling) sur cette matrice distance.

II.2.2. Le Multidimensional Scaling

Le multidimensional scaling tout comme l'analyse factorielle envoie des observations d'un espace R^p dans un espace de dimension inférieur, avec toutefois quelques différences.

L'analyse factorielle cherche des relations entre les variables alors que le positionnement multidimensionnel cherche des similarités entre les observations. Les axes de l'analyse factorielle peuvent être interprétés à l'aide des variables alors que le multidimensional scaling ne le permet pas (tout simplement parce qu'il n'y a plus de variables).

Avantage des méthodes de positionnement multidimensionnel :

- Elles s'appliquent à des problèmes plus généraux, dans lesquels on ne connaît pas nécessairement les coordonnées des observations $x_i \in \mathbb{R}^p$, mais seulement leurs distances ou similarités.
- Elles préservent mieux les distances en comparant avec les autres méthodes.

L'affichage du graphique de proximité permet de repérer d'éventuelles observations atypiques ou regroupés en classes. Chaque observation est positionnée dans un point x_i du graphique de telle façon que la distance entre deux individus sur le graphique soit la plus proche possible de la matrice de dissimilarité (1 – Matrice de proximité).

Selon la méthode des moindres carrés (méthode de Kruskal Shepard) on cherche les $\{x_i\}$ qui minimisent la fonction $\sum_{i \neq j} (1 - P_{ij} - \|x_i - x_j\|)^2$. La minimisation de cette fonction est obtenue par un algorithme d'optimisation du type descente du gradient. L'optimisation peut parfois être un peu lente.

Et par la suite pour obtenir un Clustering en groupes homogènes de nos communes. On effectue une classification ascendante hiérarchique en utilisant la distance euclidienne entre les axes du positionnement multidimensionnel

II.2.3. Classification ascendante hiérarchique (CAH)

La classification ascendante hiérarchique est une méthode de partitionnement itérative qui produit des suites de partitions emboîtées d'hétérogénéités croissantes, entre la partition en n classes où chaque objet est isolé, et la partition en 1 classe qui regroupe tous les objets. On utilise la CAH dès que l'on dispose d'une notion de distance que ce soit dans l'espace des individus ou dans l'espace des variables

L'algorithme de la CAH se présente comme suit

1. Les classes initiales sont des observations ;
2. On calcule les distances entre classes ;
3. On regroupe les deux classes les plus proches pour les remplacer par une seule classe ;
4. On reprend en 2 jusqu'à n'avoir plus qu'une seule classe, qui contient toutes les observations.

Comme déjà mentionné l'algorithme se base sur la notion de distance pour chercher à chaque étape les classes les plus proches pour les fusionner. La définition de distance retenue dans notre cas est celle de la méthode de Ward .

Distance de Ward de deux classes A et B, de barycentres a et b et d'effectifs N_A et N_B

$$d(A, B) = \frac{d(a,b)^2}{\frac{1}{N_A} + \frac{1}{N_B}} \quad \text{Avec } d(a, b) \text{ la distance euclidienne}$$

Pour faire une bonne classification il faut que l'inertie interclasse soit élevée, Toutefois le passage d'une classification en K+1 classes à une classification à K classes ne peut que faire baisser l'inertie interclasse. On cherche donc à fusionner les deux classes qui font baisser le moins l'inertie interclasse et la méthode de Ward correspond à cette objectif.

II.2.4. Choix du nombre optimum de classes

La CAH permet de choisir facilement un nombre optimum de classes. L'arbre (le dendrogramme) est coupé à une hauteur bien définie afin d'optimiser des critères de qualité statistique. Le principal est la perte d'inertie interclasse qui est représentée par la hauteur des deux branches jointes : comme cette perte doit être la plus faible possible, on coupe le dendrogramme à un niveau où la hauteur des branches est élevée.

III. Application

On retient dans un 1^{er} temps **la fréquence des sinistres matériels** comme indicateur de sinistralité.

III.1. Présentation de la base de travail

Notre base de données contient 229 communes. Chaque commune a sa fréquence de sinistres (fréquence moyenne sur 5ans) et ses caractéristiques sociodémographiques qui lui correspondent.

On ne retient que les sinistres responsables ainsi que les communes pour lesquelles on a une exposition ce qui réduit le nombre de communes à 143.

On calcule les différents indicateurs de sinistralité et cela en faisant une moyenne sur les 5 années

III.2. Traitement des variables utilisées

On a décidé de discrétiser nos variables quantitatives en différentes classes en vue de prendre en compte la répartition par commune et par année police¹ et éviter par la suite des découpages effectués par les arbres ne tenant pas compte de ces 2 critères.

¹ L'année police représente la durée pendant laquelle court la police d'assurance

Tableau 8: Discrétisation des variables externes retenues

Variables	Signification	Modalités	Description des modalités	%AP	%communes
DENS	Densité	A	■	■	■
		B	■■■■■	■	■
		C	■	■	■
PROP1559	Proportion de la population âgée entre 15 et 59 ans	A	■	■	■
		B	■	■	■
Tauxactiv	Taux d'activité	A	■	■	■
		B	■■■■■	■	■
		C	■	■	■
PV	Pourcentage des ménages possédant une voiture	A	■	■	■
		B	■■■■■	■	■
		C	■	■	■
PS	Proportion des salariés	A	■	■	■
		B	■■■■■	■	■
		C	■	■	■
PI	Proportion des indépendants	A	■	■	■
		B	■■■■■	■	■
		C	■	■	■
PE	Proportion des employeurs	A	■	■	■
		B	■	■	■
AGRI	Pourcentage des Travailleurs dans le secteur de l'agriculture pêche et forêt	A	■	■	■
		B	■■■■■	■	■
		C	■	■	■
BTP	Pourcentage des Travailleurs dans le secteur du bâtiment et travaux public	A	■	■	■
		B	■	■	■
SERVICES	Pourcentage des Travailleurs dans le secteur des services	A	■	■	■
		B	■	■	■

III.3. Zonier pour la fréquence des sinistres matériels par Arbre de Régression

III.3.1. Enjeu de la Méthode de CART

1) 1^{er} Enjeu

On utilise l'algorithme de CART afin de partitionner en fonction de l'espace des variables explicatives les communes en différentes classes homogènes en termes de risque. Et assigner par la suite à chaque classe la fréquence moyenne. Toutefois par construction l'algorithme de CART prend la fréquence moyenne de chaque classe comme étant la valeur moyenne des fréquences ce qui est erroné et rend l'approche inappropriée pour notre problème !

Afin de rendre l'approche adaptée à notre problème on modifie le code de l'algorithme de CART en réajustant la valeur moyenne au sein de chaque groupe i comme étant

$$\bar{y}_i = \frac{\sum_{R_j} w_j * y_j}{\sum w_j}$$

Avec w_j : le poids assigné à chaque observation i dans le groupe j .

y_j : La variable à expliquer (la fréquence Matérielle) dans notre cas.

On prend w_j égale à l'exposition au risque de chaque commune et donc l'algorithme calculera à chaque fois la fréquence moyenne de chaque classe comme étant la somme des sinistres matériels sur la somme des expositions.

NB: On a utilisé le package « rpart » de Therneau et al.(2009) car il nous permet en langage R de recoder les modifications à apporter sur l'algorithme pour intégrer le poids.

2) 2^{ème} Enjeu

Notre base de donnée doit être subdivisée en 2 parties : une partie " apprentissage " servant à construire le modèle et la 2^{ème} partie "test" pour évaluer sa performance. Toutefois comme notre base de données est de petite taille (143 observations) plus nous réservons des données pour la construction du modèle plus l'estimation de l'erreur test sera moins précise. et plus nous favorisons la partie test plus nous pouvons retenir de l'information importante pour la construction d'un modèle efficace. Dans ce contexte où la base est de petite taille plusieurs méthodes ont été développées pour répondre à cet enjeu entre elles La méthode de validation croisée (déjà expliquée dans la section 1.3)

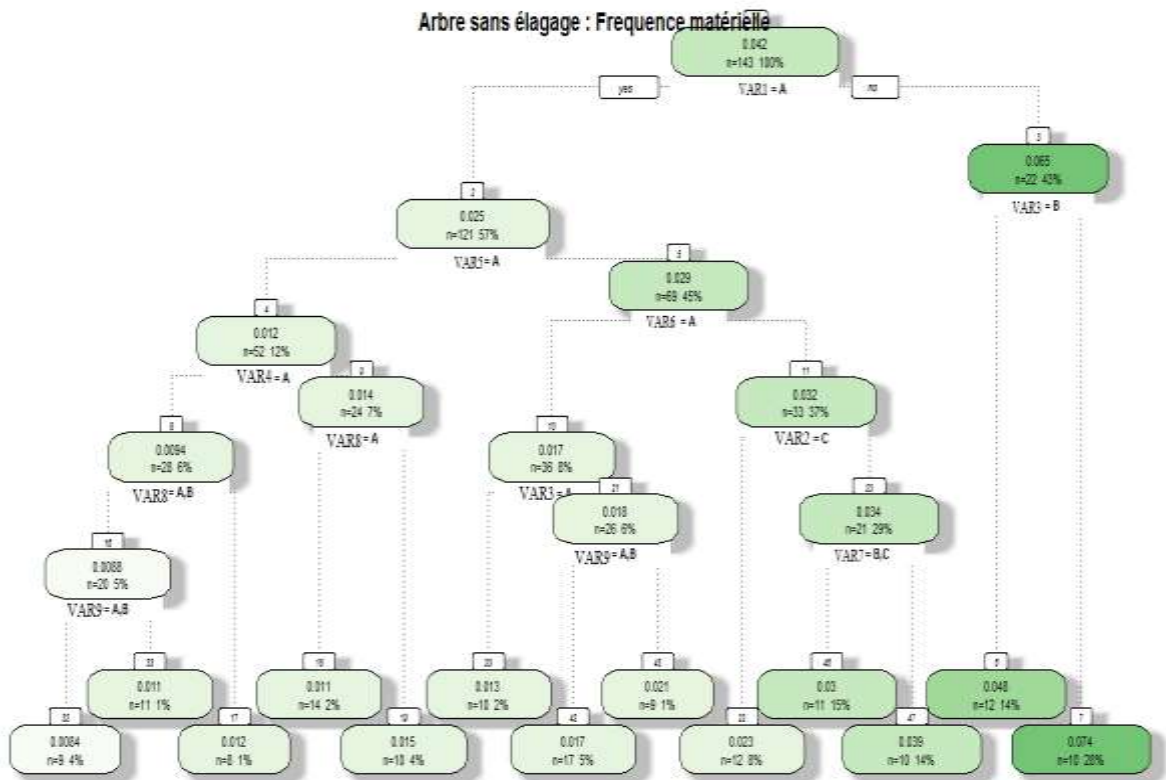
Nous consacrons donc la totalité de notre base de données pour la construction du modèle et on recourt à ladite méthode pour évaluer sa performance.

III.3.2. Arbre de régression pour la fréquence matérielle

1) Arbre Maximal

L'arbre réalisé est obtenu sans élagage. Il est développé à son niveau maximal.

Figure 12: CART (sans élagage) appliquée aux données avec Fréquence matérielle comme Target



2) Elagage de l'arbre

Tableau 9: Critères associés à chaque arbre créé

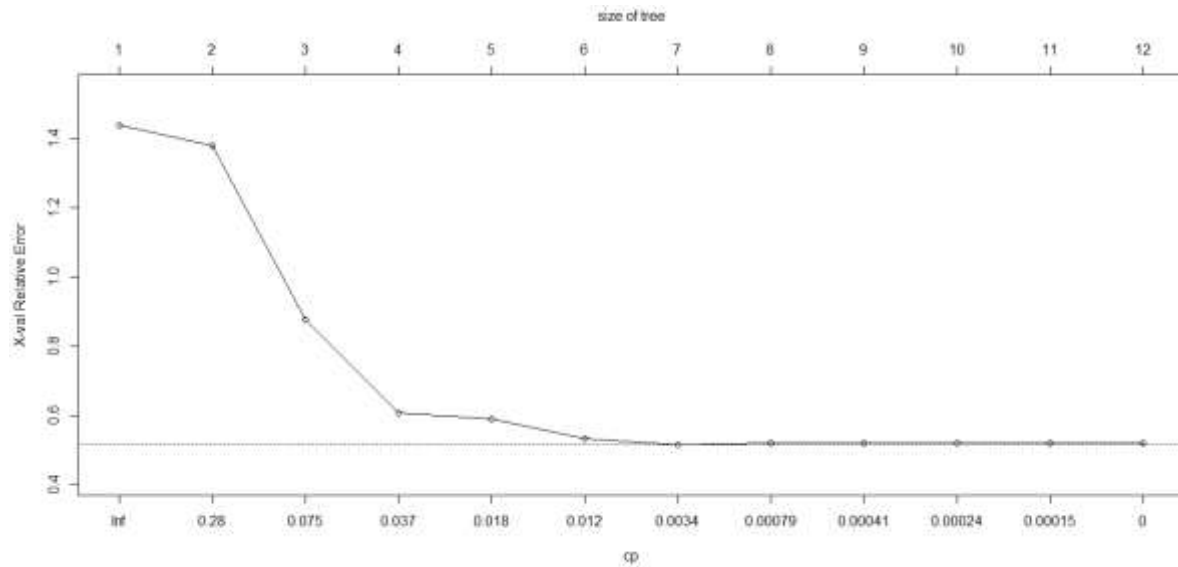
CP	nsplit	rel error	xerror	xstd
0.70392880	0	1.000000	1.39531	0.00129599
0.10864693	1	0.296071	1.39102	0.00132036
0.05181640	2	0.187424	0.89796	0.00082053
0.02678528	3	0.135608	0.60874	0.00064100
0.01269471	4	0.108823	0.52962	0.00066316
0.01141201	5	0.096128	0.52683	0.00066432
0.00098564	6	0.084716	0.51948	0.00066633
0.00063202	7	0.083730	0.52314	0.00066478
0.00026389	8	0.083098	0.52314	0.00066494
0.00021577	9	0.082834	0.52282	0.00066502
0.00010682	10	0.082619	0.52384	0.00066443
0.00000000	11	0.082512	0.52406	0.00066437

Comme mentionné dans la partie ci-dessus, le choix du critère d'arrêt a été fixé en choisissant la valeur du coefficient de complexité et donc le nombre de nœuds qui minimise l'erreur sur la base de validation (l'X.erreur). La valeur min de l'X erreur est obtenu pour la valeur de CP=0,00098564 .En l'élaguant avec ce critère. On garde donc l'arbre ayant 7 nœuds finaux.

Chaque ligne de ce tableau est associée à une valeur différente pour le paramètre α . Le tableau présente les valeurs du critère de complexité $C_\alpha(T)$ des différents arbres associés à cette valeur de α (CP), le nombre de découpages effectués (nsplit), ainsi que trois champs d'informations relatives à la validation croisée : le champ rel.error correspond au rapport de deux estimations de MSE: Au numérateur se trouve l'estimateur de la MSE que l'on obtient si l'on n'effectue aucune séparation dans l'arbre, et au dénominateur l'estimateur de la MSE obtenu en utilisant l'arbre correspondant à la ligne du tableau.

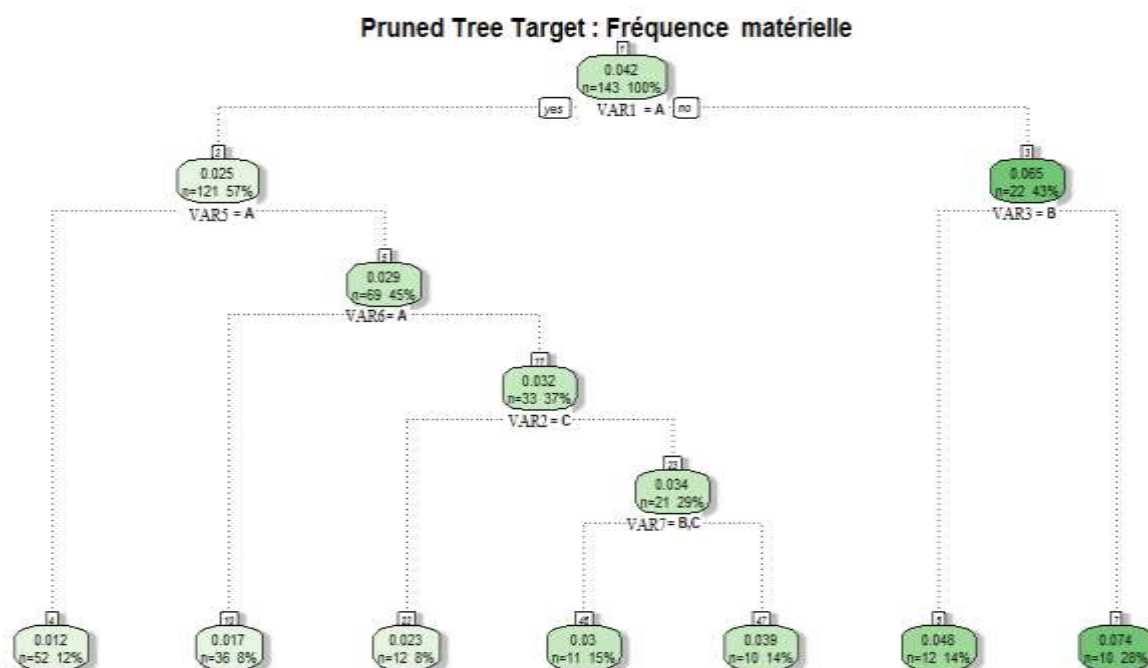
$$\text{Avec } \widehat{MSE} = \frac{1}{l} * \sum_{i=1}^l (Y_i^* - \widehat{m}(X_i^*))^2$$

Figure 13:: Graphique de l'erreur de la validation croisée en fonction du paramètre de complexité



3) Arbre élagué

Figure 14: Arbre élagué pour la Fréquence matérielle comme Target

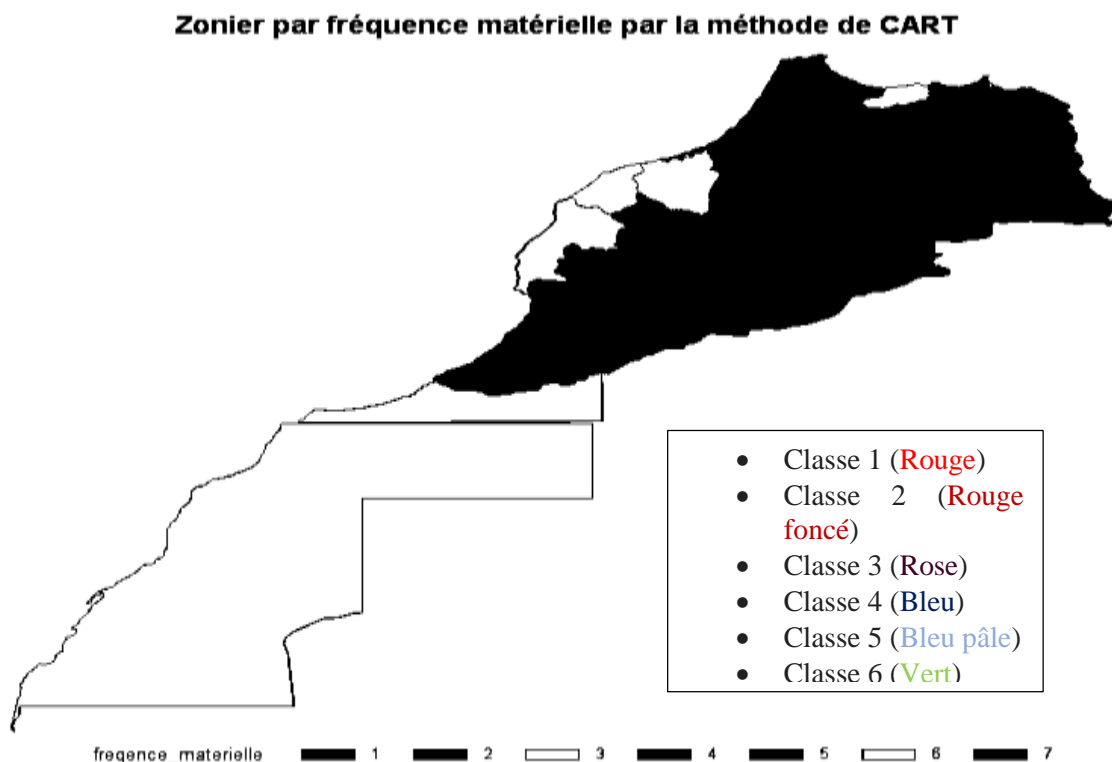


Le résultat graphique produit par l'algorithme nous indique que la fréquence moyenne du portefeuille est de 4,2% et concerne 143 communes (valeurs au sommet de l'arbre). Parmi l'ensemble des variables explicatives sociodémographiques retenues la variable qui effectue la meilleure séparation binaire de la base de données est la variable 1. La meilleure séparation à lieu pour la variable1 égale à la classe A : classe la plus faible. L'algorithme crée alors deux nœuds et cherche parmi ces deux nœuds celui qui nécessite le plus d'être séparé en deux et ainsi de suite.

On se retrouve au final avec 7 classes de risques présentés dans le tableau dans l'annexe

III.3.3 Projection des classes sur la carte à risque

Figure 15: Cartographie des zones de risque (CART)



En utilisant SAS, On projette sur la carte du Maroc les différentes zones de risques obtenues par CART pour nos 143 communes. On effectue par la suite un lissage pour les communes pour lesquelles on n'a pas l'information sur la sinistralité en prenant comme hypothèse que les communes voisines -les plus proches géographiquement- ont plus de chance d'avoir un risque spatial similaire et donc le même comportement vis-à-vis de la sinistralité.

III.4. Zonier pour la fréquence des sinistres matériels par Random Forest

III.4.1. Enjeu de la Méthode du Random Forest

1) 1^{er} Enjeu

Nous effectuons notre Random Forest en prenant la variable réponse : la fréquence des sinistres matériels. Toutefois dans cette approche il s'est avéré inexistant à notre connaissance un package du Random Forest sur R permettant de recoder les modifications à apporter sur l'algorithme pour intégrer le temps d'exposition ie (pondérer les observations) comme dans le cas de CART .Pour y remédier et éviter que notre algorithme effectue la moyenne des

fréquences pour estimer la fréquence prédite. Nous avons donc segmenté notre variable explicative en 3 tranches en prenant en considération le poids par commune et par année police.

Tableau 10: Tranches de fréquence

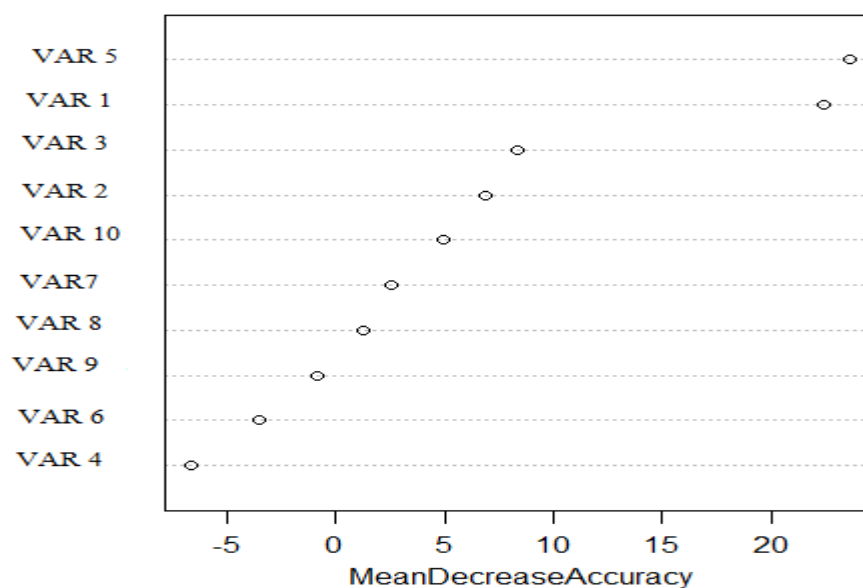
Tranches de fréquence des sinistres matériels	Intervalle
Tranche A	<2%
Tranche B	Entre 2% et 4%
Tranche C	> 4%

2) 2^{ème} Enjeu

Le 2^{ème} enjeu est celui relatif à la petite taille de notre base de données. Tout comme dans CART on prend la totalité de la base comme base d'apprentissage, et on utilise ici l'erreur OOB une estimation de l'erreur de généralisation sans avoir recours à un échantillon test supplémentaire pour évaluer la performance du modèle. (L'erreur OOB est déjà présentée dans la section II.2.4)

III.4.2. Importance des variables

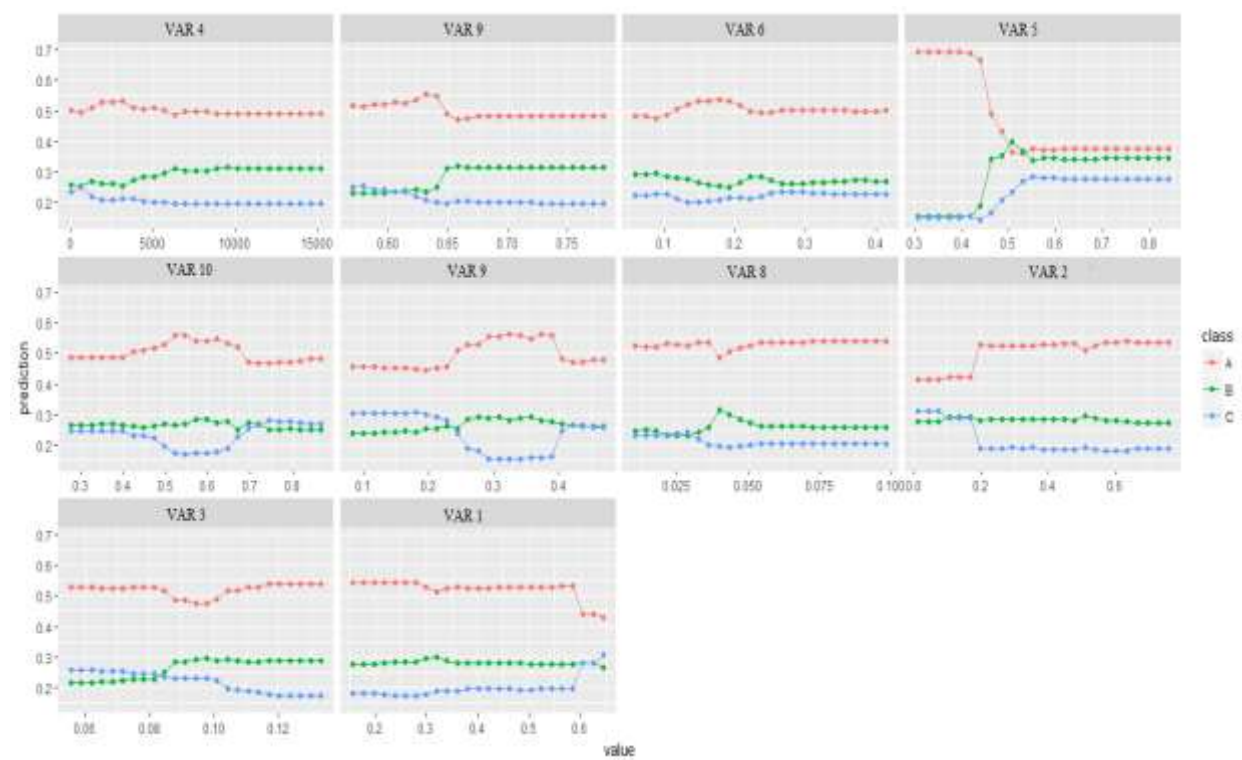
Figure 16 : Importance des variables explicatives



Les variables sont classées selon leurs ordres d'importance dans le modèle. On constate que la variable taux d'activité est la variable la plus importante dans le modèle (celle qui engendre le plus grand taux d'erreur lorsque ces valeurs sont permutées), suivie par la proportion de la population dans le secteur des services et puis les autres variables

III4.3. Dépendances partielles

Figure 17: Dépendances partielles entre les variables explicatives et la variable à expliquer



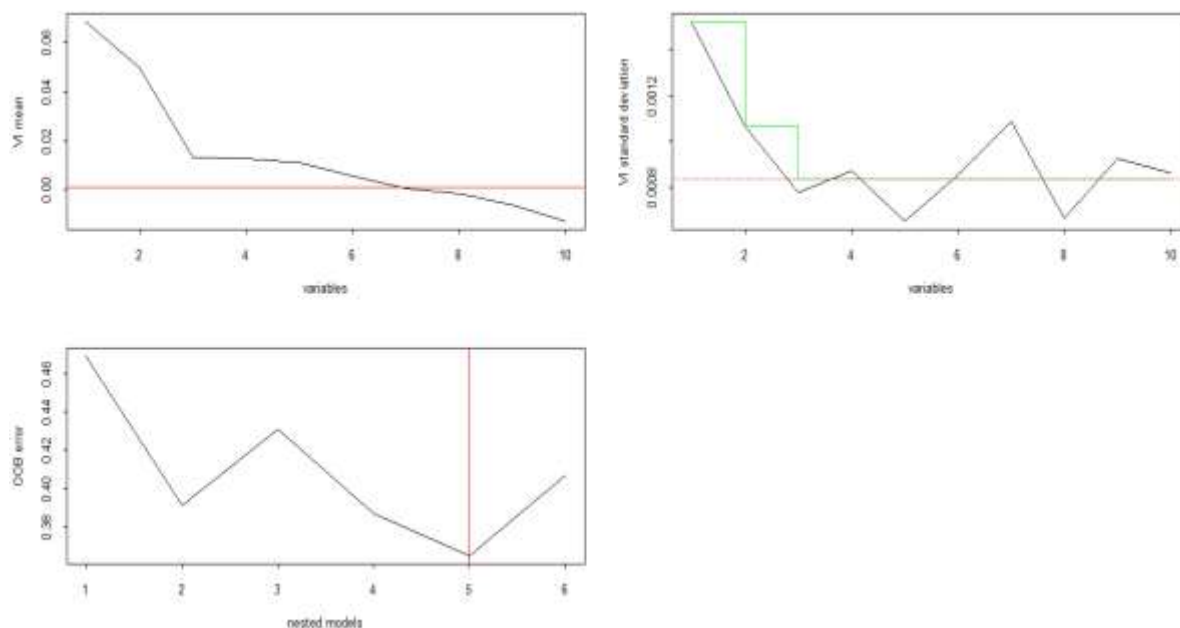
La figure ci-dessous met en relief les dépendances partielles entre les 10 variables explicatives et la variable à expliquer : la fréquence des sinistres matériels. Où chaque valeur de la variable explicative lui correspond la proportion de vote prédite pour chaque tranche de fréquence.

On constate que les dépendances partielles ont révélés des relations non linéaires entre la variable à expliquer et chaque variable explicative. Par exemple pour la variable la plus importante : Variable 5 Pour des valeurs faibles pour cette variable la proportion des votes pour la tranche C « tranche ayant la fréquence la plus élevée » est faible et pour des valeurs élevées pour la variable 5 la proportion des votes pour la tranche C est élevée.

Les dépendances partielles ont été calculées et visualisées à l'aide du package « edarf » permettant d'explorer les données à l'aide du Random Forest.

III.4.4. Sélection des variables

Figure 18: Sélection des variables pertinentes



Tout d'abord on est en disposition des 10 variables déjà présentées lors de la modélisation par l'arbre CART. On sélectionne (par la procédure déjà présentée dans la section X) que les variables pertinentes afin d'interpréter notre base de données. On utilise pour cet effet le package « VSURF ». Au final on se retrouve avec 5 variables jugées par le Random Forest les plus importantes pour le but d'interprétation de la base de données.

Les 5 variables retenues sont : la variable 1, la variable 2, la variable 3, la variable 5, la variable 10.

On a utilisé le « RandomForest » sur nos 5 variables retenues.

```
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 3

OOB estimate of error rate: 42.66%
Confusion matrix:
  A  B  C class.error
A 57 15  3  0.2400000
B 23 11  4  0.7105263
C  8  8 14  0.5333333
```

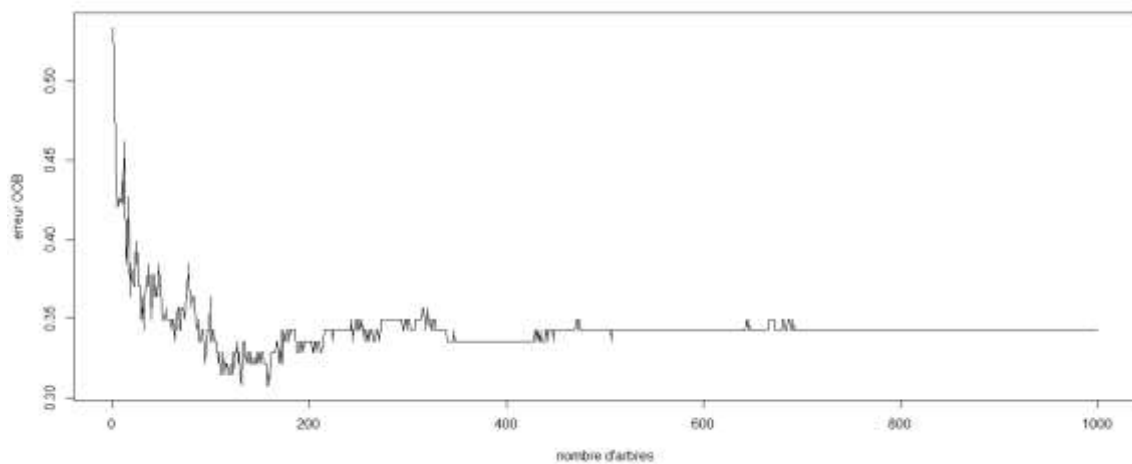
On obtient un taux d'erreur OOB de 42,66%, on cherche donc à optimiser les paramètres du Random Forest afin de la faire diminuer.

III.4.5. Optimisation des paramètres

1) Nombre d'arbres dans la forêt

On a choisi le nombre d'arbres minimisant l'erreur Out Of Bag (OOB)

Figure 19: L'erreur OOB en fonction du nombre d'arbres dans la forêt

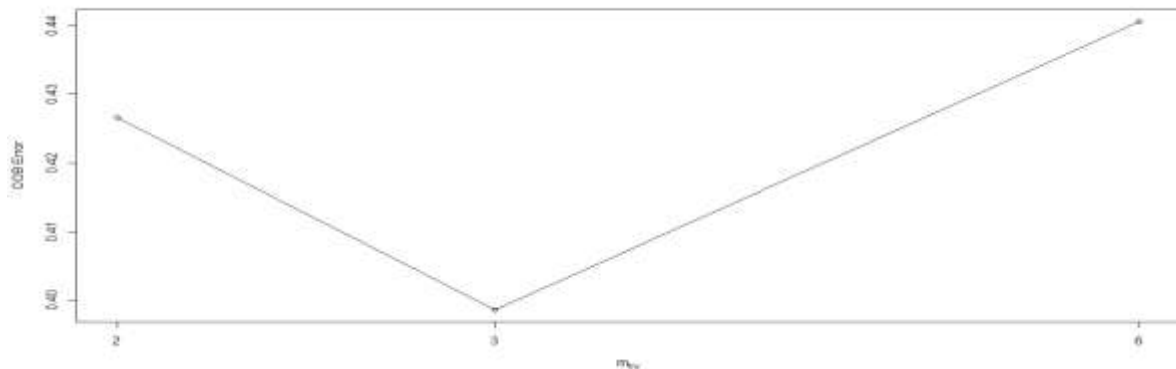


On constate qu'à partir de 800 arbres l'erreur OOB se stabilise. On a retenu un nombre d'arbres égal à 1000.

2) Nombre de variables explicatives choisi à chaque arbre

On cherche le nombre Mtry de variables optimal du point de vue du taux d'erreur OOB.

Figure 20: Tracé de l'erreur OOB en fonction du Mtry



Et donc le modèle Random Forest final optimisé est celui ayant 11 variables explicatives, un nombre d'arbres égal 1000 et un mtry égal à 3.

```

Type of random forest: classification
Number of trees: 1000
No. of variables tried at each split: 3

OOB estimate of error rate: 34.97%
Confusion matrix:
  A  B  C class.error
A 60 13  2  0.2000000
B 19 16  3  0.5789474
C  6  7 17  0.4333333
    
```

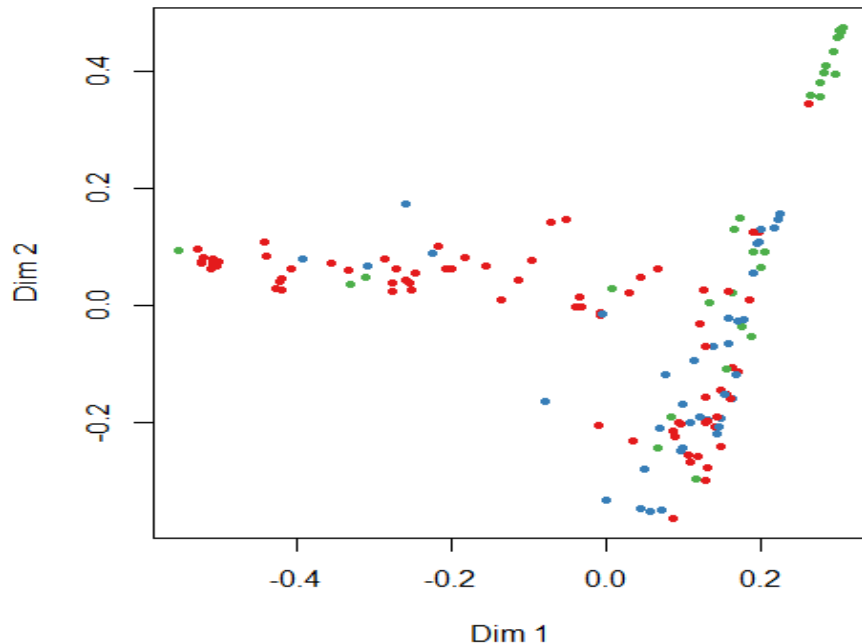
L'erreur OOB est donc diminuée après ce paramétrage.

III.4.6. Matrice de proximité

Le package « Random Forest » contient la fonctionnalité qui nous permet de calculer à partir de notre modèle Random Forest retenu la matrice de proximité et l'affichage du graphique de proximité qui en découle. Toutefois due à la nature stochastique des forêts aléatoires il est probable qu'il y a une certaine variation dans cette matrice. Par conséquent on calcule une

matrice de proximité agrégée sur 10 forêts aléatoires et on prend la médiane de cette matrice comme étant notre matrice de proximité.

Figure 21: Graphique de proximité



● Tranche A ● Tranche B ● Tranche C

Chaque point représente une commune, coloriée selon sa fréquence matérielle observée.

Plus les valeurs sont élevées dans la matrice de proximité, plus les communes sont proches dans ce graphique.

On n'observe pas des formes très nettes, si ce n'est peut-être la forme d'un V renversé dont la partie gauche contient presque toutes les communes ayant une fréquence de sinistres faible.

En haut à droite, on constate que la forêt aléatoire a bien identifié les communes risquées.

III.4.7. Classification ascendante hiérarchique

Comme expliqué dans la partie théorique nous allons effectuer un positionnement multidimensionnel sur notre matrice distance. On utilise pour ce fait le package « cmdscale » pour obtenir les principales coordonnées. Afin de regrouper les communes en exploitant ces

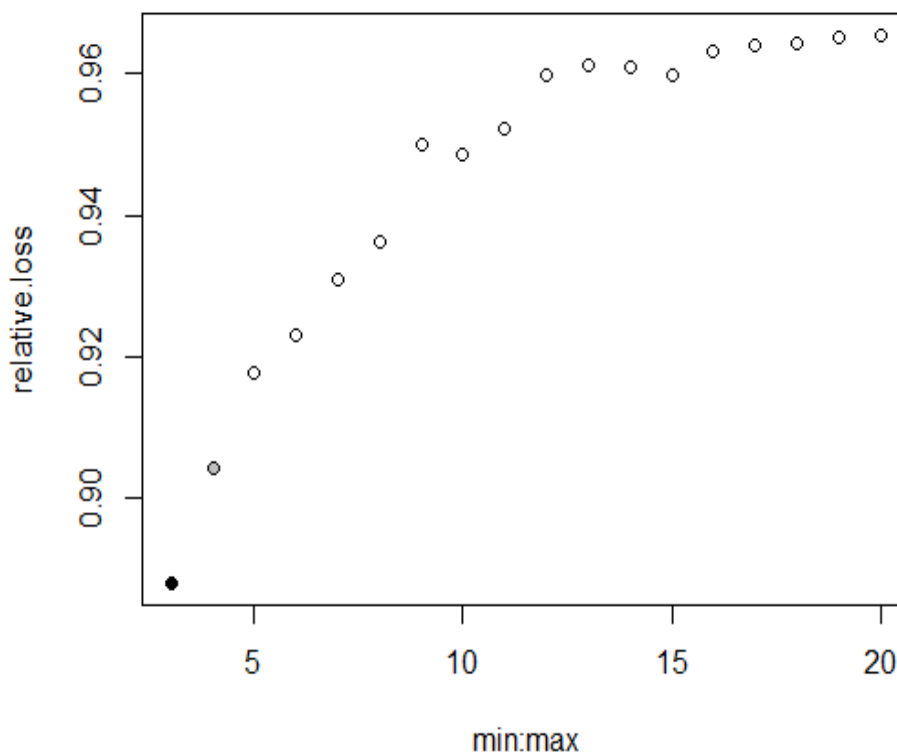
principales coordonnées, nous utilisons la classification ascendante hiérarchique (hclust du package “stats”) en prenant la distance euclidienne entre ces principales coordonnées.

En utilisant la fonction « best.cutree » du package « JLUtills » qui se base sur le critère de la perte d’inertie interclasse. On détermine le nombre de classes optimal.

On constate que la plus petite perte d’inertie relative est obtenue pour un nombre de classes égal à 3.

le tableau présentant les différentes classes avec leur perte d’inertie relatives se trouve dans l’annexe

Figure 22: Graphique des pertes d’inertie relatives



La meilleure partition étant représentée par un point noir obtenue pour un nombre de classe égal à 3 et la seconde par un point gris est obtenue pour un nombre de classe égal à 4.

On coupe donc le dendrogramme pour obtenir 3 classes.

Les Classes obtenues par Random Forest ainsi que leurs répartition par année police et par communes sont présentes dans l'annexe.

III.4.8. Variables importantes pour chaque cluster

En utilisant la matrice d'importance locale calculée elle aussi de façon agrégée sur 10 forêts aléatoires et en prenant la médiane. On obtient donc l'importance de chacune des 5 variables pour chaque observation (commune). On détermine les variables plus importantes pour chaque cluster comme étant la moyenne de la matrice d'importance locale pour chaque variable dans les observations appartenant au cluster.

Tableau 11: Classement des variables selon leurs valeurs d'importance dans chaque Cluster

	Cluster 1		Cluster 2		Cluster 3
VAR 5	0,17284561	VAR 5	0,10876486	VAR 5	0,30427603
VAR 1	0,14268356	VAR 10	0,08810456	VAR 1	0,08535316
VAR 10	0,0802492	VAR 3	0,06001578	VAR 2	0,05259316
VAR 3	0,07476955	VAR 1	0,05071058	VAR 3	0,01631211
VAR 2	0,01816798	VAR 2	0,04199237	VAR 10	0,00800224

Les variables les plus importantes dans le 1^{er} cluster sont la variable 5 la variable 1 et la variable 10.

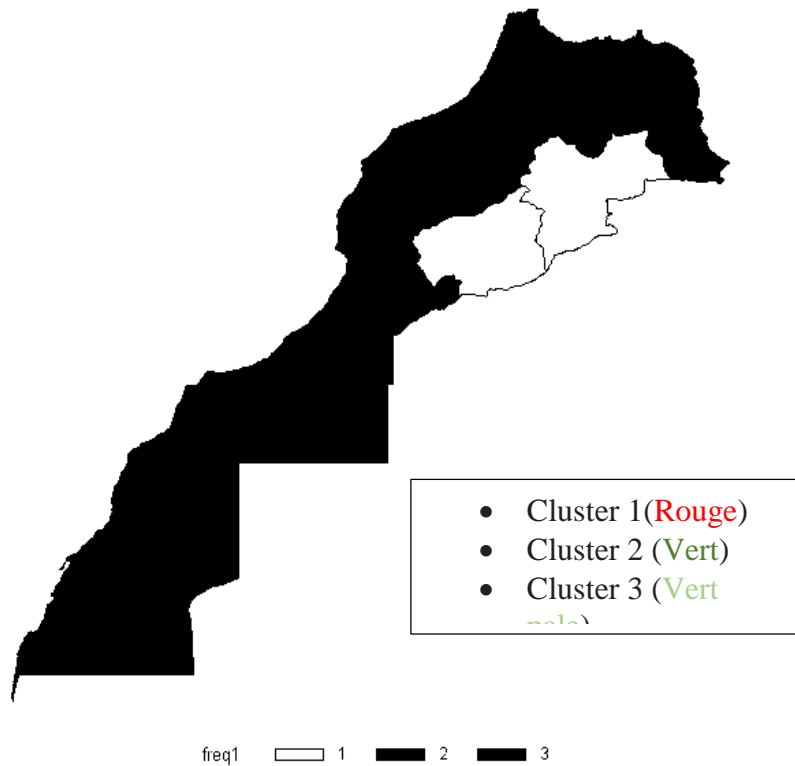
Les variables les plus importantes dans le 2^{eme} cluster sont les variable 5,10,3

Les variables les plus importantes dans le 3^{eme} cluster sont les variable 5 ,1,2

III.4.9. Projection des Classes de risques sur la carte

Figure 23 Carte des zones de risque (Random Forest)

Zonier par fréquence matérielle par la méthode Random Forest



On projette sur la carte du Maroc les différentes zones de risques obtenues par le Random Forest pour les 143 communes. On effectue par la suite un lissage pour les communes pour lesquelles on n'a pas l'information sur la sinistralité en prenant comme hypothèse que les communes voisines -les plus proches géographiquement- ont plus de chance d'avoir un risque spatial similaire et donc le même comportement vis-à-vis de la sinistralité.

III.5 Validation des zoniers

III.5.1 Comparaison entre les deux zoniers

Les deux méthodes permettent de mieux appréhender la variable 'zone' dans la sinistralité globale d'un automobiliste. Cependant nous avons voulu comparer les deux modèles entre eux et vérifier aussi à tel point ils sont cohérent avec la réalité.

Pour cela, nous avons utilisés des données sur la sinistralité des contrats automobiles d'AXA pour l'exercice 2015. Et on a comparé la sinistralité prédite par les deux méthodes (fréquence estimée) et la sinistralité observée (fréquence réelle).

Nous avons ainsi cherché à regarder le ratio fréquence/ fréquence estimée pour toutes nos communes

Ensuite, nous les avons regroupées selon leur pourcentage de précision :

0.7 : les contrats qui obtiennent un ratio fréquence observée/fréquence estimée inférieur à 0.8

0.8 : les contrats qui obtiennent un ratio fréquence observée/fréquence estimée entre 0.8 et 0.9

0.9 : les contrats qui obtiennent un ratio fréquence observée/fréquence estimée entre 0.9 et 1

1 : les contrats qui obtiennent un ratio fréquence observée/fréquence estimée entre 1 et 1.1

1.1 : les contrats qui obtiennent un ratio fréquence observée fréquence estimée entre 1.1 et 1.2

1.2 : les contrats qui obtiennent un ratio fréquence observée/fréquence estimée supérieur à 1.2

On a considéré que par exemple, notre meilleur modèle sera celui où le pourcentage est le plus grand dans la tranche A

Avec :

- **Tranche A** = tranche 0.9 + tranche 1

- **Tranche B** = tranche 0.8 + tranche 1.1

- **Tranche C** = tranche 0.7 + tranche 1

1) La garantie RC matérielle testée avec notre 1ere méthode (CART)

rapport	%AP	Nombre de communes
0,7	14,32%	32
0,8	15,64%	17
0,9	14,72%	30
1	28,57%	28
1,1	17,07%	11
1,2	9,68%	25
Total général	100%	143

Tranche	%AP	Répartition par communes
Tranche A	43,29%	40,55%
Tranche B	32,71%	19,58%
Tranche C	24,00%	39,86%

Ainsi, le zonier élaboré par CART permet de dire qu'on allait prévoir correctement (à une marge de 10%) 43,29% de nos années polices la fréquence de la sinistralité de nos assurés pour la garantie RC matérielle.

2) La garantie RC matérielle testée avec notre 2^{eme} méthode (RF)

rapport	%AP	Nombre de communes
0,7	10,41	21
0,8	20,60	14
0,9	16,26	38
1	29,08	23
1,1	14,11	32
1,2	09,54	15
Total général	0,9984	143

Tranche	%AP	Répartition par communes
Tranche A	45,34%	42,65%
Tranche B	34,71%	32,09%
Tranche C	19,95%	25,26%

Le zonier élaboré par Random Forest permet de dire qu'on allait prévoir correctement (à une marge de 10%) 45,34% de nos années polices la fréquence de la sinistralité de nos assurés pour la garantie RC matérielle. Par ailleurs on constate qu'avec le Random Forest dans la tranche C (des mal prédits) le pourcentage années police est plus petit que celui trouvé avec CART.

Toutefois on tient à préciser que les communes se trouvant dans la tranche C par les deux méthodes sont majoritairement des petites communes où on manque d'années police.

On peut conclure donc que les 2 zoniers élaborés par les deux méthodes sont relativement bons et sont cohérents avec la réalité.

Remarque

On prenant **le coût moyen des sinistres matériels** comme indicateur de sinistralité pour élaborer le zonier on constate qu'en le modélisant avec nos variables sociodémographiques il est non expliqué par lesdites variables. Une explication logique de cela c'est que le coût moyen est un indicateur biaisé en raison de la convention qui lie les assureurs entre eux la CID. En effet, lors d'un sinistre matériel responsable, quelque soit le montant réel du sinistre, un forfait sera payé. Le coût moyen sera donc toujours le même.

Chapitre 5 : Modélisation du risque

I. Aspect théorique

I.1. Théorie des modèles linéaires généralisés

I.1.1. Définition des Modèles linéaires généralisés

Les modèles linéaires généralisés sont une extension du modèle linéaire simple. En effet, au lieu de modéliser la variable à expliquer directement on estime plutôt l'espérance de la variable aléatoire réponse (variable à expliquer), en la représentant comme combinaison linéaire de ensemble de prédicteurs.

Ils sont utilisés pour modéliser la variable à expliquer lorsque la relation entre la variable à expliquer et les variables explicatives est non linéaire.

Dans ce mémoire la variable à expliquer Y sera, soit

- i. Discrète : estimation du nombre de sinistres.
- ii. Continue : modélisation du coût moyen d'un sinistre

I.1.2. Hypothèses du modèle

Notons $(Y_i)_{1 < i < n}$ l'ensemble des variables aléatoires à expliquer. Pour employer

Le modèle GLM, il nous faut poser les hypothèses suivantes :

1. (Y_1, \dots, Y_n) définit une famille de variables aléatoires indépendantes qui suivent une distribution appartenant à la famille exponentielle.
2. Les prédicteurs (X_1, \dots, X_p) correspondent aux composants déterministes du modèle sous la forme de combinaison linéaire.
3. Pour tout $i \in \{1, \dots, n\}$ la loi de Y_i est supposée appartenir à une famille de distributions dont les paramètres dépendent des variables explicatives à travers une fonction de lien, g , strictement monotone. $g(E[Y]) = \beta_0 + \sum_{i=1}^p \beta_i X_i$

I.1.3. Détermination des coefficients d'un modèle

Les paramètres $(\beta_0, \dots, \beta_p)$ d'un GLM sont estimés par la méthode du maximum de vraisemblance. En pratique, On recourt à des méthodes itératives pour maximiser la fonction de vraisemblance. Le log vraisemblance s'écrit :

$$L(\theta; y) = \sum_{i=1}^n \ln f(y_i, \theta_i, \varphi) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{\frac{\varphi}{\omega_i}} + \sum_{i=1}^n c(y_i, \varphi)$$

Nous recherchons le p-uplet $(\beta_0, \dots, \beta_p)$ qui maximise $L(\theta; y)$, i.e. les β_0, \dots, β_p solutions de $U_j = 0$ où

$$U_j = \frac{\partial L(\vartheta, y)}{\partial \beta_j} = \sum_{i=1}^n \omega_i (y_i - \mu_i) \frac{x_{ij}}{b''(\vartheta_i) g'(\mu_i)} \text{ pour } j=0, \dots, K$$

Avec

$$E(Y_i) = \mu_i = b'(\vartheta_i)$$

$$V(Y_i) = \mu_i = b''(\vartheta_i) \frac{\varphi}{\omega_i}$$

$$E(Y_i) = g(\mu_i) = x_i^t \beta$$

I.1.4. Significativité des variables

La significativité des coefficients associés aux variables explicatives peut être testée à l'aide du test de Wald. Soit le test suivant :

$$H_0: \beta_j = 0 \text{ contre } H_1: \beta_j \neq 0$$

La statistique de Wald s'écrit : $W = \frac{\widehat{\beta}_j}{\widehat{\sigma}(\widehat{\beta}_j)}$

Sous H_0 , la statistique du test suit approximativement une loi Normale $N(0,1)$

Le test de Wald peut aussi être défini ainsi : $(\frac{\widehat{\beta}_j}{\widehat{\sigma}(\widehat{\beta}_j)})^2$

Sous H_0 , cette statistique suit asymptotiquement une loi de khi-deux à un degré de liberté.

I.1.5. Sélection des variables explicatives

1) Sélection des variables explicatives tarifaires par les différentes méthodes

La variable à expliquer est estimée avec toutes les variables tarifaires qui sont significativement explicatives. Le modèle obtenu est dit modèle de référence. Toutefois il faut conserver uniquement les variables significativement explicatives de la variable réponse et ce grâce à une procédure de sélection pas à pas.

La sélection pas à pas consiste à effectuer des régressions successives afin de sélectionner les variables définitives du modèle par une des méthodes suivantes :

La méthode BACKWARD : On démarre avec le modèle complet (incluant toutes les variables ayant un effet significatif sur le risque à modéliser) puis de retirer la variable la moins significative, celle dont l'élimination entraîne la plus faible augmentation de la déviance.

La méthode FORWARD : On recherche la variable la plus significative au sens de la déviance. En partant de ce modèle à une variable, nous cherchons ensuite la variable qui, associée à la première, explique le mieux le risque et ainsi de suite.

La méthode STEPWISE : une combinaison et amélioration des méthodes BACKWARD et FORWARD. Puisque à chaque étape, nous réexaminons toutes les variables introduites précédemment dans le modèle. Il se peut qu'une variable considérée comme la plus significative à une étape de l'algorithme puisse à une étape ultérieure devenir non significative.

I.1.6. Qualité d'ajustement

Pour évaluer la qualité d'ajustement du modèle sur la base des différences entre observations et estimations plusieurs critères sont utilisés.

1) La déviance

On compare le modèle estimé au modèle saturé, c'est-à-dire le modèle possédant autant de paramètres que d'observations et estimant donc exactement les données.

La déviance est définie à partir de la log-vraisemblance de ces deux modèles :

$$D = -2(L - L_s)$$

La déviance D suit asymptotiquement une loi de khi 2 à n-p-1 degré de liberté.

Le test de rejet ou d'acceptation du modèle basé sur la déviance est le suivant

Si la déviance est supérieure au quantile d'une loi de khi 2 à n-p-1 degrés de liberté d'ordre $1-\alpha$, le modèle est jugé de mauvaise qualité.

D'autres critères peuvent être utilisés pour mesurer la qualité d'ajustement d'un modèle ainsi que de comparer deux modèles statistiques et de choisir par la suite le meilleur modèle, deux critères sont utilisés pour cette fin : AIC et BIC.

2) AIC

AIC est l'une des critères de mesure de la qualité d'un modèle statistique. L'AIC repose sur un compromis entre la vraisemblance du modèle et entre le nombre de paramètres. Plus l'AIC est petit plus l'ajustement est bon.

Il est défini par la formule mathématique :

$$AIC = 2k - 2\ln(L)$$

K : est le nombre de paramètres à estimer du modèle

L : est le maximum de la fonction de vraisemblance du modèle.

3) BIC

BIC tout comme l'AIC, permet aussi de juger et comparer les modèles statistiques. Il prend en considération en plus du nombre de paramètres, le nombre d'observations dans le modèle ce qui le rends plus robuste.

Sa formule mathématique s'écrit :

$$BIC = -2 \ln(L) + k * \ln(L)$$

I.2. Apport significatif d'une variable dans le modèle

On teste si l'apport d'une variable tarifaire est significatif dans le modèle en utilisant le test statistique suivant

I.2.1. Statistique du Chi²

Pour la variable X :

$$R = \frac{\text{Vraisemblance du modèle sans } X}{\text{Vraisemblance du modèle avec } X}$$

- $n = \dim(\text{modalités du modèle avec } X) - \dim(\text{modalités du modèle sans } X)$
= nombre de modalités de X
- $S_{n,95\%}$ = le seuil à 95% du Chi² à n ddl

Sous H_0 la variable X n'est pas influente dans le modèle, La statistique $S = -2\ln R$ suit asymptotiquement une loi de Chi² à n ddl

Si $P[S < s_{n,95\%}] > 5\%$ alors H_0 est vraie et les deux vraisemblances sont proches, donc X n'a pas d'apport significatif dans le modèle.

$$Deviance_{sans X} - Deviance_{avec X} = 2[\ln(V_{avec X} - V_{sans X})]$$

II. Modélisation du risque

On modélise dans un 1^{er} temps la fréquence des sinistres matériels incluant toutes les variables tarifaires sauf la zone géographique.

Tout d'abord on cherche la loi qui modélise la fréquence des sinistres par la suite on la modélise avec un GLM

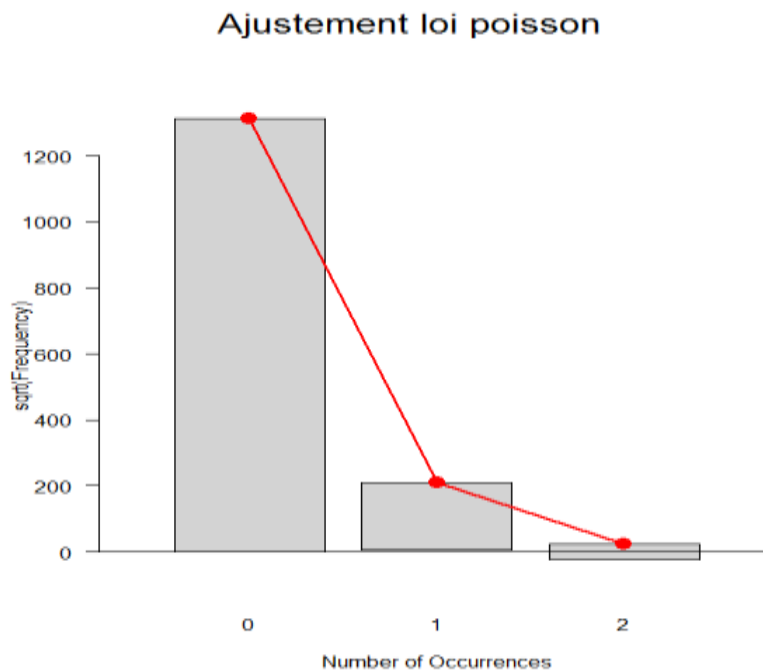
II.1. Modélisation de la fréquence des sinistres :

Avant de passer au GLM, le bon choix de la loi qui ajuste mieux la fréquence s'avère très important. Pour ce faire, on teste à l'aide des box plots déterminés sous le logiciel R les distributions possibles qui peuvent l'ajuster à savoir la loi de poisson et la loi binomiale

II.1.1. Loi de poisson

La loi de poisson se voit très classique dans ce type de modélisation. C'est pour cela qu'on va la présenter en premier.

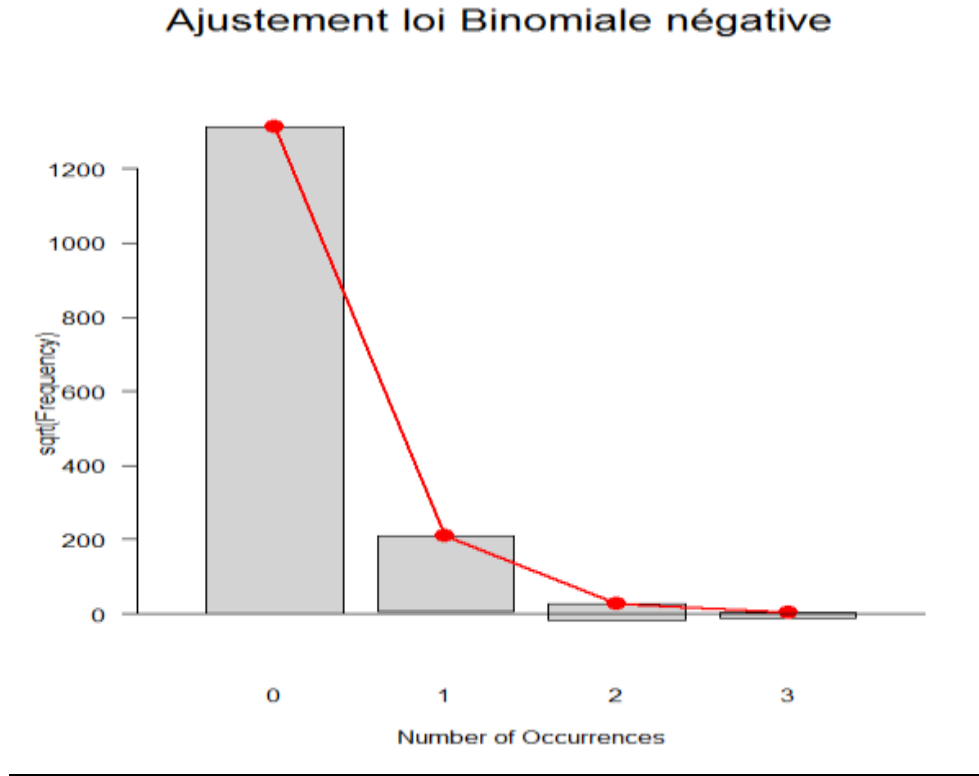
Figure 24: Ajustement par loi de Poisson



II.1.2. Loi binomiale négative

La loi binomiale négative se présente aussi comme une loi très classique dans ce type de modélisation surtout quand on se retrouve avec une variance supérieure à la moyenne comme il est le cas ici (légèrement supérieur).

Figure 25: Ajustement par Binomiale négative (sortie R)



On constate que la fréquence des sinistres s’ajuste mieux à la loi binomiale négative qu’à la loi de poisson. Toutefois pour confirmer le résultat fourni par le box plot, On recourt à un test khi2 d’adéquation le résultat est le suivant

Tableau 12: Statistique de Khi-deux pour la loi de Poisson et la loi Binomiale négative

Binomiale négative	Poisson
X-squared = 4333,5	X-squared= 4833,5

On observe que la statistique du test Khi 2 de la loi binomiale négative est plus petite que celles de la loi de poisson.

II.2. GLM sur la fréquence des sinistres

On modélise la fréquence des sinistres avec les variables tarifaires suivantes : CRM, puissance fiscale, âge véhicule, âge conducteur, sexe, type carburant, et en ne prenant pas la zone géographique comme variable tarifaire.

On applique le GLM avec les deux distributions poisson et binomiale négative pour s'assurer de ce qu'on a trouvé avant sous R et de s'assurer aussi que le modèle retenu est le modèle le plus pertinent pour expliquer la fréquence des sinistres

II.2.1 Modèle de Poisson

Tableau 13: Estimation des paramètres pour la loi de Poisson

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		DDL	Valeur estim	Erreur type	Wald 95% intervalle de confiance		Khi-2 de Wal	Pr > Khi-2
Intercept		1	-7,1733	0,0169	-8,1733	-6,1733	19499,1	<.0001
VEENR	E	1	0,5685	0,0131	-0,4315	1,5685	270,15	<.0001
VEENR	G	0	0	0	-1	1,	.	.
VEPUI		1	-1,2567	0,0177	-2,2567	-0,2567	472,23	<.0001
VEPUI		2	-1,0476	0,0161	-2,0476	-0,0476	391,22	<.0001
VEPUI		3	-0,7902	0,0084	-1,7902	0,2098	854,22	<.0001
VEPUI		4	0	0	-1	1,	.	.
DUSEX	F	1	0,6417	0,0099	-0,3583	1,6417	553,16	<.0001
DUSEX	M	0	0	0	-1	1,	.	.
Age_ASS		1	1,3188	0,0109	0,3188	2,3188	1798,95	<.0001
Age_ASS		2	0,3312	0,0088	-0,6688	1,3312	208,7	<.0001
Age_ASS		3	0	0	-1	1,	.	.
vehi		1	2,9097	0,0102	1,9097	3,9097	9480,64	<.0001
vehi		2	1,7799	0,0114	0,7799	2,7799	2932,98	<.0001
vehi		3	0,8859	0,0132	-0,1141	1,8859	590,15	<.0001
vehi		4	0	0	-1	1,	.	.
VETXCR		1	-4,677	0,0152	-5,677	-3,677	10074,6	<.0001
VETXCR		2	-4,1475	0,0146	-5,1475	-3,1475	8644,09	<.0001
VETXCR		3	0	0	-1	1,	.	.

D'après la figure ci-dessus, on repère que toutes nos variables tarifaires sont significatives (test de Wald).

Tableau 14: Evaluation de la qualité d'ajustement du GLM pour la loi de poisson

Criterion	DF	Value	ValueDF	pvalue
Deviance	1,80E+06	555920,851	0,3152	1
Scaled Deviance	1,80E+06	555920,851	0,3152	1
Log Likelihood		-286121,582		,
Full Log Likelihood		-327009,226		,
AIC (smaller is better)		654056,452		,
AICC (smaller is better)		654056,452		,
BIC (smaller is better)		654291,73		,

Avec une p-value > 0.05, le test de Khi-deux affirme que notre modèle est bon.

II.2.2 Modèle Binomiale négative

Tableau 15: Estimation des paramètres pour la loi Binomiale Négative

Paramètres estimés par l'analyse du maximum de vraisemblance									
Paramètre		DDL	Valeur estim	Erreur type	Wald 95% intervalle de confiance		Khi-2 de Wal	Pr > Khi-2	
Intercept		1	-7,1046	0,0392	-7,6046	-6,6046	3423,68	<,0001	
VEENR	E	1	0,5853	0,021	0,0853	1,0853	126,46	<,0001	
VEENR	G	0	0	0	-0,5	0,5	,	,	
VEPUI		1	-1,3836	0,0284	-1,8836	-0,8836	203,91	<,0001	
VEPUI		2	-1,2648	0,0256	-1,7648	-0,7648	210,37	<,0001	
VEPUI		3	-0,8643	0,0134	-1,3643	-0,3643	381,41	<,0001	
VEPUI		4	0	0	-0,5	0,5	,	,	
DUSEX	F	1	0,6705	0,0171	0,1705	1,1705	226,8	<,0001	
DUSEX	M	0	0	0	-0,5	0,5	,	,	
Age_ASS		1	1,38	0,0186	0,88	1,88	709,24	<,0001	
Age_ASS		2	0,3357	0,014	-0,1643	0,8357	99,66	<,0001	
Age_ASS		3	0	0	-0,5	0,5	,	,	
vehi		1	2,9652	0,0154	2,4652	3,4652	4401,42	<,0001	
vehi		2	1,7553	0,0168	1,2553	2,2553	1357,2	<,0001	
vehi		3	0,834	0,0188	0,334	1,334	279,36	<,0001	
vehi		4	0	0	-0,5	0,5	,	,	
VETXCR		1	-5,0406	0,0382	-5,5406	-4,5406	1768,66	<,0001	
VETXCR		2	-4,5399	0,0377	-5,0399	-4,0399	1460,51	<,0001	
VETXCR		3	0	0	-0,5	0,5	,	,	

D'après la figure ci-dessus, on repère que pour le modèle Binomiale négative, toutes nos variables tarifaires sont significatives (test de Wald).

Tableau 16: Evaluation de la qualité d'ajustement du GLM pour la loi Binomiale Négative

Criterion	DF	Value	ValueDF	pvalue
Deviance	1,80E+06	484 020,00	0,2689	1
Scaled Deviance	1,80E+06	484 020,00	0,2689	1
AIC (smaller is better)	_	310479,7376	_	,
BIC (smaller is better)	_	310479,13	_	,

Encore une autre fois, avec un p-value >0.05 , le test de Khi-deux affirme que notre modèle est bon.

Mais, la question qui se pose maintenant est la suivante : **quelle le modèle le plus pertinent et le mieux adéquat ?**

La réponse à cette question est donnée par le AIC ou le BIC. Rappelons que plus le AIC et BIC sont faibles, plus le modèle est bon et d'après les sorties données ci-dessus, le modèle le plus pertinent est celui de **Binomiale négative**. Cela vient confirmer ce qui a été trouvé auparavant sous R.

Remarque

Pour éviter la redondance des résultats, La même modélisation a été effectuée pour le coût moyen. On trouve que le coût moyen est modélisé par la loi Log normal. Les résultats de cette modélisation se trouvent en annexe.

*Chapitre 6 : Analyse de la
performance du zonier*

I.1. Evaluation de la performance du zonier

Afin d'évaluer la performance du zonier obtenu, on utilise dans un 1^{er} temps le modèle tarifaire n'incluant aucun facteur susceptible de contenir de l'information géographique, et un autre contenant à chaque fois le zonier élaboré par les deux méthodes. En utilisant le test de Chi² déjà présenté dans la partie théorique on étudie la significativité de l'apport du zonier.

I.1.1. Zonier obtenu par CART

On compare le modèle de fréquence sans variable zone et l'autre modèle de fréquence incluant les segments de zones obtenus par CART

Tableau 17: Estimation des paramètres pour la loi Binomiale Négative (CART)

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		DDL	Valeur estim	Erreur type	d 95% intervalle de confia		Khi-2 de Wal	Pr > Khi-2
Intercept		1	-5,484664	0,043	-7,484664	-3,484664	5952,49	<,0001
VEENR	E	1	0,163746	0,0209	-1,836254	2,163746	22,41	<,0001
VEENR	G	0	0	0	-2	2	,	,
VEPUI		1	-0,6283546	0,0281	-2,6283546	1,3716454	182,41	<,0001
VEPUI		2	-0,4882608	0,0254	-2,4882608	1,5117392	134,84	<,0001
VEPUI		3	-0,4202814	0,0133	-2,4202814	1,5797186	362,28	<,0001
VEPUI		4	0	0	-2	2	,	,
DUSEX	F	1	0,2591818	0,017	-1,7408182	2,2591818	85,16	<,0001
DUSEX	M	0	0	0	-2	2	,	,
segments		1	2,6698868	0,0243	0,6698868	4,6698868	4394,59	<,0001
segments		2	1,9915814	0,0266	-0,0084186	3,9915814	2052,13	<,0001
segments		3	1,6298516	0,0269	-0,3701484	3,6298516	1346,42	<,0001
segments		4	1,3286582	0,0265	-0,6713418	3,3286582	917,69	<,0001
segments		5	1,080889	0,0305	-0,919111	3,080889	459,8	<,0001
segments		6	0,505297	0,0322	-1,494703	2,505297	90,19	<,0001
segments		7	0	0	-2	2	,	,
Age_ASS		1	0,9004376	0,0185	-1,0995624	2,9004376	864,45	<,0001
Age_ASS		1	0,2492578	0,0139	-1,7507422	2,2492578	118,38	<,0001
Age_ASS		3	0	0	-2	2	,	,
vehi		1	1,2342148	0,0155	-0,7657852	3,2342148	2306,12	<,0001
vehi		2	0,6983188	0,0169	-1,3016812	2,6983188	625,95	<,0001
vehi		3	0,2899462	0,019	-1,7100538	2,2899462	85,55	<,0001
vehi		4	0	0	-2	2	,	,
VETXCR		1	-2,4894354	0,0367	-4,4894354	-0,4894354	1680,77	<,0001
VETXCR		2	-2,084867	0,0362	-4,084867	-0,084867	1209,94	<,0001
VETXCR		3	0	0	-2	2	,	,

Tableau 18 : Evaluation de la qualité d'ajustement du GLM pour la Binomiale Négative (CART)

Critère	DDL	Valeur	Valeur/DDL
Deviance		1,80E+06	422 460
Scaled Deviance		1,80E+06	422 460
AIC (smaller is better)		300217,1	
BIC (smaller is better)		300469,1	

La variable zone élaborée par CART est significative par le test de Wald

On constate que la déviance a diminué en ajoutant la variable zonier créée par CART, toutefois pour tester si l'apport de la variable zone élaborée par le Random Forest est significatif dans le modèle de fréquence on utilise la statistique du Chi² :

$Deviance_{sans\ zone} - Deviance_{avec\ zone\ CART} = 484020 - 422460 = 61560$ suit une loi de Khi-deux de 7 degrés de liberté qu'on teste pour trouver une P-value < 0.05. Par conséquent, l'ajout de cette variable améliore significativement notre modèle. De plus, on constate qu'en l'introduisant au modèle, AIC et le BIC diminuent.

I.1.2. Zonier obtenu par Random Forest

On compare le modèle de fréquence sans variable zone et l'autre modèle de fréquence incluant les segments de zones obtenus par Random Forest

Tableau 19: Estimation des paramètres pour la loi Binomiale Négative (RF)

Paramètres estimés par l'analyse du maximum de vraisemblance									
Paramètre		DDL	Valeur estim	Erreur type	Wald 95% intervalle de	Khi-2 de Wal	Pr > Khi-2		
					confiance				
Intercept		1	-4,10364	0,0463	-5,10364	-3,10364	5163,81	<,0001	
VEENR	E	1	0,19944	0,021	-0,80056	1,19944	97,74	<,0001	
VEENR	G	0	0	0	-1	1	,	,	
VEPUI		1	-0,57624	0,0283	-1,57624	0,42376	225,18	<,0001	
VEPUI		2	-0,47904	0,0256	-1,47904	0,52096	186,54	<,0001	
VEPUI		3	-0,3594	0,0134	-1,3594	0,6406	415,81	<,0001	
VEPUI		4	0	0	-1	1	,	,	
DUSEX	F	1	0,2232	0,0171	-0,7768	1,2232	164,34	<,0001	
DUSEX	M	0	0	0	-1	1	,	,	
clusterCut		1	1,44156	0,0278	0,44156	2,44156	2036,21	<,0001	
clusterCut		2	0,91884	0,0288	-0,08116	1,91884	812,47	<,0001	
clusterCut		3	0	0	-1	1	,	,	
Age_ASS		1	0,55152	0,0186	-0,44848	1,55152	713,29	<,0001	
Age_ASS		2	0,12948	0,0139	-0,87052	1,12948	94,29	<,0001	
Age_ASS		3	0	0	-1	1	,	,	
vehi		1	1,05396	0,0154	0,05396	2,05396	3476,71	<,0001	
vehi		2	0,612	0,0168	-0,388	1,612	1042,32	<,0001	
vehi		3	0,26484	0,0189	-0,73516	1,26484	185,75	<,0001	
vehi		4	0	0	-1	1	,	,	
VETXCR		1	-1,95972	0,0375	-2,95972	-0,95972	1728,77	<,0001	
VETXCR		2	-1,72968	0,037	-2,72968	-0,72968	1367,61	<,0001	
VETXCR		3	0	0	-1	1	,	,	

Tableau 20: Evaluation de la qualité d'ajustement du GLM pour la Binomiale Négative (RF)

Criterion	DDL	Valeur	Valeur/DDL
Deviance	1,80E+06	385 200	0,214
Scaled Devia	1,80E+06	385 200	0,214
AIC (smaller)	_	294787,76	_
BIC (smaller)	_	294787,76	_

La variable zone élaborée par le Random Forest est significative par le test de Wald

On constate que la déviance a diminué en ajoutant la variable zonier crée par Random Forest, toutefois pour tester si cet apport est significatif dans le modèle de fréquence .On utilise la statistique du Chi² :

$Deviance_{sans\ zone} - Deviance_{avec\ Zone\ RF} = 484020 - 385200 = 98820$ suit une loi de Khi-deux de 3 degrés de liberté qu'on teste pour trouver une P-value<0.05. Par conséquent, l'ajout de cette variable améliore significativement notre modèle. De plus, on constate qu'en l'introduisant au modèle, AIC et le BIC diminuent.

Remarque :

Même si qu'on n'a pas pu obtenir un zonier par coût moyen (Pour la raison expliquée dans la remarque du chapitre 4), on a utilisé le même zonier élaboré par la fréquence et on a testé son amélioration par rapport au modèle tarifaire du coût moyen matériel.

L'étude de l'apport de la variable zone que ce soit celle élaborée par CART ou Random Forest révèle que la variable zone n'améliore pas le modèle tarifaire. En effet, son apport n'est pas significatif au modèle tarifaire par le coût moyen.

Les résultats explicites de cette étude sont donnés dans l'annexe

Conclusion

Nous sommes finalement arrivés à l'objectif initial qui était d'élaborer un zonier. Nous nous sommes tournés vers l'utilisation des méthodes de Machine Learning pour arriver à cette fin. En effet, nous avons élaboré le zonier par deux approches à savoir CART et Random Forest.

Nous avons souligné l'importance du rôle des variables sociodémographiques externes de l'HCP, permettant d'expliquer un effet que les variables tarifaires à elles seules n'auraient pas pu décrire et par conséquent d'affiner encore plus la segmentation.

Nous avons démontré aussi la pertinence de notre zonier élaboré par les deux méthodes et cela en voyant si le zonier apporte un gain significatif au modèle de tarification.

Nous espérons que ce travail constituera une base documentaire solide pour inciter à utiliser les données externes afin d'effectuer une segmentation au plus juste. En effet, on peut penser tout comme nous avons fait avec le zonier à utiliser encore une fois les données exogènes telles que la marque du véhicule, la capacité du moteur, type de carrosserie et le nombre de portes afin d'élaborer un véhiculier (groupement de véhicules en classes homogènes).

Bibliographie et Webographie

Bibliographie :

- Bristol Genetic Epidemiology. Laboratories. Using a Random Forest proximity measure for variable importance stratification in genotypic data, 2014.
- Elodie FERRAND et Benjamin TANGUY. Zonage Géographique d'un portefeuille d'assurés automobile. Mémoire ENSAE, 2007.
- Fédération marocaine des sociétés d'assurance et de réassurance - Situation liminaire 2011.
- Lucie MOISAN. Zonage d'une garantie Vol en Assurance Multirisques Habitation. Mémoire ISUP, 2012.
- Marri Fouad. Assurance non vie. INSEA. RABAT, 2016.
- Rasa Sipulskyte. Development of a Motor Vehicle Classification Scheme for a New Zealand Based Insurance Company. New Zealand Society of Actuaries Conference, 2012.
- Robin Genuer. Forêts aléatoires : aspects théoriques, sélection de variables et applications. UNIVERSITÉ PARIS-SUD XI, 2010.
- Stéphane Tufféry. Data mining et statistique décisionnelle. Editions TECHNIP 2005.
- Vehicle Postcode Zoning in Personal Lines Rating. General Insurance Convention, 1999.
- Zachary Jones and Fridolin Linder. Exploratory Data Analysis using Random Forests, 2015.

Webographie :

- Site officiel de R : <https://cran.r-project.org>
- Ressources Actuarielles : <https://Ressources-actuarielles.net>
- Site HCP : <https://rgphencartes.hcp.ma/>

ANNEXE

Définition des classes obtenues par CART

Classes	Définition
<p>Classe 1 (plus risquée)</p> <p>(Segment 7 obtenu par l'arbre CART)</p>	<ul style="list-style-type: none"> • Elle correspond à variable 1 appartenant à la catégorie B et par variable 3 appartenant à la catégorie A • la fréquence moyenne de cette classe est de 7,38% elle regroupe 7% du totale des communes qui représente 28,36% années polices du portefeuille de l'assurance.
<p>classe 2 (segment 6 obtenu par l'arbre CART)</p>	<ul style="list-style-type: none"> • Elle correspond à variable 1 appartenant à la catégorie B et par variable 2 appartenant à la catégories B • la fréquence moyenne de cette classe est de 4,84% elle regroupe 8,39% du totale des communes qui représente 14,16% années polices du portefeuille de l'assurance
<p>classe 3 (segment 47 obtenu par l'arbre CART)</p>	<ul style="list-style-type: none"> • définit par variable 1 appartenant à la catégorie. Par variable 5 appartenant aux catégories B et C. par variable 6 appartenant aux catégories B et C et par variable 2 appartenant aux catégories A,B et par variable 7 à la catégorie A .

	<ul style="list-style-type: none"> la fréquence moyenne de cette classe est de 3,90% elle regroupe 7% du totale des communes et représente 13,50% années polices du portefeuille de l'assurance.
classe 4 (segment 46 obtenu par l'arbre CART)	<ul style="list-style-type: none"> Définit par variable 1 appartenant à la catégories A ,par variable 5 appartenant aux catégories B et C et par variable 6 appartenant aux catégories B et C et par variable 2 appartenant aux catégories A,B et variable 7 appartient aux catégories B et C. La fréquence moyenne de cette classe est de 2,96% elle regroupe 7,69% du totale des communes et représente 15,46% années polices du portefeuille de l'assurance.
classe 5 (segment 22 obtenu par l'arbre CART)	<ul style="list-style-type: none"> définit par variable 1 appartenant à la catégorie A, par variable 5 appartenant aux catégories B et C et par variable 6 appartenant aux catégories B et C et par variable 3 appartenant à la catégorie C. la fréquence moyenne de cette classe est de 2,34% elle regroupe 8,39% du totale des communes et représente 8,10% années polices du portefeuille de l'assurance.
classe 6 (segment 10 obtenu par l'arbre CART)	<ul style="list-style-type: none"> définit par variable 1 appartenant à la catégories A, par variable 5 appartenant aux catégories B et C

	<p>et variable 6 appartenant à la catégorie A.</p> <ul style="list-style-type: none"> la fréquence moyenne de cette classe est de 1,66% elle regroupe 25,17% du totale des communes et représente 7,98% années polices du portefeuille de l'assurance.
Classe 7 (segment 4 obtenu par l'arbre CART)	<ul style="list-style-type: none"> défini par variable1 appartenant à la catégorie A et par variable 5 appartenant à la catégorie A la fréquence moyenne de cette classe est de 1,16% elle regroupe 36,36% du totale des communes et représente 12,28% années polices du portefeuille de l'assurance.

Les différentes classes obtenues par la CAH avec leur perte d'inertie relatives

Nombre de classes	Pertes d'inerties relatives
3	0.8879103
4	0.9042427
5	0.9176628
6	0.9229063
7	0.9309167
8	0.9362626
9	0.9499941
10	0.9486532
11	0.9523016
12	0.9598404
13	0.9612438
14	0.9609034
15	0.9598338
16	0.9631460
17	0.9639962
18	0.9643321
19	0.9650903
20	0.9654144

Classes obtenues par Random Forest

Clusters	%AP	Répartition par communes
Cluster 1	46,81%	54,55%
Cluster 2	11,72%	34,27%
Cluster 3	41,31%	11,19%

GLM coût moyen

1) Modèle de Gamma (Sans introduction de la Zone géographique dans le modèle) :

Tableau Modèle Gamma avant regroupement des variables

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		DDL	Valeur estim	Erreur type	Wald 95% intervalle de	Khi-2 de Wal	Pr > Khi-2	
					confiance			
Intercept		1	13,67205	0,0177	12,67205	14,67205	267075	<,0001
DUSEX	F	1	-0,1104	0,0108	-1,1104	0,8896	23,36	<,0001
DUSEX	M	0	0	0	-1	1		
VEENR	E	1	0,0237	0,0146	-0,9763	1,0237	9,22	0,0024
VEENR	G	0	0	0	-1	1		
VETXCR		1	-0,081	0,0151	-1,081	0,919	2,63	0,105
VETXCR		2	-0,02415	0,0149	-1,02415	0,97585	0,77	0,3787
VETXCR		3	0	0	-1	1		
Age_ASS		1	-0,0441	0,0138	-1,0441	0,9559	0,03	0,8662
Age_ASS		2	-0,0039	0,0102	-1,0039	0,9961	2,92	0,0875
Age_ASS		3	0	0	-1	1		
vehi		1	0,0531	0,0121	-0,9469	1,0531	23,83	<,0001
vehi		2	-0,06645	0,0135	-1,06645	0,93355	1,75	0,1865
vehi		3	-0,0306	0,0159	-1,0306	0,9694	0,45	0,5013
vehi		4	0	0	-1	1		
VEPUI		1	-0,4251	0,0205	-1,4251	0,5749	140,71	<,0001
VEPUI		2	-0,25425	0,0178	-1,25425	0,74575	56,99	<,0001
VEPUI		3	-0,2124	0,0096	-1,2124	0,7876	162,05	<,0001
VEPUI		4	0	0	-1	1		

Tableau Modèle Gamma après regroupement des variables

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		DDL	Valeur estim	Erreur type	Wald 95% intervalle de	Khi-2 de Wal	Pr > Khi-2	
					confiance			
Intercept		1	11,517912	0,0091	9,517912	13,517912	1022787	<,0001
DUSEX	F	1	-0,09513	0,0107	-2,09513	1,90487	26,18	<,0001
DUSEX	M	0	0	0	-2	2		
VEENR	E	1	0,021294	0,0146	-1,978706	2,021294	9,73	0,0018
VEENR	G	0	0	0	-2	2		
VETXCR		1	-0,067788	0,0086	-2,067788	1,932212	18,2	<,0001
VETXCR		2	0	0	-2	2		
vehi		1	0,059724	0,0087	-1,940276	2,059724	55,02	<,0001
vehi		2	0	0	-2	2		
VEPUI		1	-0,354186	0,0204	-2,354186	1,645814	139,42	<,0001
VEPUI		2	-0,212184	0,0178	-2,212184	1,787816	56,38	<,0001
VEPUI		3	-0,176778	0,0096	-2,176778	1,823222	159,27	<,0001
VEPUI		4	0	0	-2	2		
Scale		1	1,793988	0,0094	-0,206012	3,793988		

Tableau Déviance de Gamma après regroupement

Critères d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Deviance	3,90E+04	29853,5226	0,7705
Scaled Deviance	3,90E+04	43053,3797	1,1111
Pearson Chi-Square	3,90E+04	45714,3935	1,1798
Scaled Pearson X2	3,90E+04	65927,1996	1,7015
Log Likelihood		-389392,483	
Full Log Likelihood		-389392,483	
AIC (smaller is better)		778802,966	
AICC (smaller is better)		778802,97	
BIC (smaller is better)		778880,051	

2) Modèle de logNormal (Sans introduction de la Zone géographique dans le modèle) :

Log normal avant regroupement

Intercept		1	17,5072	0,0185	16,5072	18,5072	224702	<,0001
DUSEX	F	1	-0,132	0,0114	-1,132	0,868	14,74	0,0001
DUSEX	M	0	0	0	-1	1	,	,
VEENR	E	1	-0,0046	0,0155	-1,0046	0,9954	3,28	0,0703
VEENR	G	0	0	0	-1	1	,	,
VETXCR	1	1	-0,123	0,0158	-1,123	0,877	3,7	0,0544
VETXCR	2	1	-0,0742	0,0157	-1,0742	0,9258	0,17	0,6827
VETXCR	3	0	0	0	-1	1	,	,
Age_ASS	1	1	-0,0668	0,0146	-1,0668	0,9332	0,11	0,7396
Age_ASS	2	1	-0,0104	0,0107	-1,0104	0,9896	2,18	0,1396
Age_ASS	3	0	0	0	-1	1	,	,
vehi	1	1	-0,092	0,0127	-1,092	0,908	2,75	0,0973
vehi	2	1	-0,172	0,0142	-1,172	0,828	16,87	<,0001
vehi	3	1	-0,0896	0,0167	-1,0896	0,9104	0,53	0,4648
vehi	4	0	0	0	-1	1	,	,
VEPUI	1	1	-0,3646	0,0216	-1,3646	0,6354	42,01	<,0001
VEPUI	2	1	-0,236	0,0188	-1,236	0,764	18,55	<,0001
VEPUI	3	1	-0,1754	0,0101	-1,1754	0,8246	44,71	<,0001
VEPUI	4	0	0	0	-1	1	,	,

Tableau Lognormal après regroupement

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		DDL	Valeur estim	Erreur type	Wald 95% intervalle de confiance		Khi-2 de Wal	Pr > Khi-2
Intercept		1	11,3698	0,0087	9,3698	13,3698	1022376	<,0001
DUSEX	F	1	-0,09191	0,0111	-2,09191	1,90809	19,25	<,0001
DUSEX	M	0	0	0	-2	2	,	,
VEENR	E	1	-0,00299	0,0154	-2,00299	1,99701	3,28	0,0701
VEENR	G	0	0	0	-2	2	,	,
VETXCR		1	-0,05928	0,0091	-2,05928	1,94072	9,44	0,0021
VETXCR		2	0	0	-2	2	,	,
VEPUI		1	-0,2288	0,0215	-2,2288	1,7712	38,78	<,0001
VEPUI		2	-0,14911	0,0187	-2,14911	1,85089	17,3	<,0001
VEPUI		3	-0,10998	0,0101	-2,10998	1,89002	41,21	<,0001
VEPUI		4	0	0	-2	2	,	,
Scale		1	1,12918	0,0031	-0,87082	3,12918		

Déviance lognormal après regroupement

Critères d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Deviance	3,90E+04	29632,7609	0,7649
Scaled Deviance	3,90E+04	38755	1,0003
Pearson Chi-Square	3,90E+04	29632,7609	0,7649
Scaled Pearson X2	3,90E+04	38755	1,0003
Log Likelihood		-49790,4441	
Full Log Likelihood		-49790,4441	
AIC (smaller is better)		99608,8881	
AICC (smaller is better)		99608,8989	
BIC (smaller is better)		99728,7983	

3) Modèle de logNormal (avec Zonier élaboré par CART) :

Tableau Log normal avec zone CART

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		DDL	Valeur estim	Erreur type	Wald 95% intervalle de confiance		Khi-2 de Wal	Pr > Khi-2
Intercept		1	6,8612	0,0301	5,8612	7,8612	87582,4	<,0001
DUSEX	F	1	-0,0394	0,0111	-1,0394	0,9606	12,47	0,0004
DUSEX	M	0	0	0	-1	1	,	,
VEENR	E	1	0,0324	0,0154	-0,9676	1,0324	4,43	0,0352
VEENR	G	0	0	0	-1	1	,	,
VETXCR		1	-0,0265	0,0091	-1,0265	0,9735	8,58	0,0034
VETXCR		2	0	0	-1	1	,	,
VEPUI		1	-0,1333	0,0215	-1,1333	0,8667	38,59	<,0001
VEPUI		2	-0,0797	0,0187	-1,0797	0,9203	18,14	<,0001
VEPUI		3	-0,0618	0,0101	-1,0618	0,9382	37,61	<,0001
VEPUI		4	0	0	-1	1	,	,
segments		1	-0,245	0,0299	-1,245	0,755	40,85	<,0001
segments		2	-0,1909	0,0308	-1,1909	0,8091	8,1	<,0001
segments		3	-0,236	0	-1,236	0,764	,	0,0024
segments		4	0,0547	0,0031	-0,9453	1,0547		<,0001
segments		5	-0,1909	1,0031	-1,1909	0,8091		<,0001
segments		6	0,2647	2,0031	-0,7353	1,2647		0,0044
segments		7	0	3,0031	0	0		,

Tableau Déviance avec Zone CART

Critères d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Deviance	3,90E+04	29492,2435	0,7628
Scaled Deviance	3,90E+04	38670	1,0002

4) Modèle de logNormal (avec Zonier élaboré RandomForest)

Tableau Lognormal avec zone Random forest

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		DDL	Valeur estim	Erreur type	Wald 95% intervalle de confiance		Khi-2 de Wal	Pr > Khi-2
Intercept		1	6,8612	0,0301	5,8612	7,8612	87582,4	<,0001
DUSEX	F	1	-0,0394	0,0111	-1,0394	0,9606	12,47	0,0004
DUSEX	M	0	0	0	-1	1	,	,
VEENR	E	1	0,0324	0,0154	-0,9676	1,0324	4,43	0,0352
VEENR	G	0	0	0	-1	1	,	,
VETXCR		1	-0,0265	0,0091	-1,0265	0,9735	8,58	0,0034
VETXCR		2	0	0	-1	1	,	,
VEPUI		1	-0,1333	0,0215	-1,1333	0,8667	38,59	<,0001
VEPUI		2	-0,0797	0,0187	-1,0797	0,9203	18,14	<,0001
VEPUI		3	-0,0618	0,0101	-1,0618	0,9382	37,61	<,0001
VEPUI		4	0	0	-1	1	,	,
clusterCut		1	-0,1909	0,0299	-1,1909	0,8091	40,85	<,0001
clusterCut		2	-0,0876	0,0308	-1,0876	0,9124	8,1	0,0044
clusterCut		3	0	0	-1	1	,	,

Tableau Déviance en Random Forest

Critères d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Deviance	3,90E+04	29492,2435	0,7628
Scaled Deviance	3,90E+04	38670	1,0002
Pearson Chi-Square	3,90E+04	29492,2435	0,7628
Scaled Pearson X2	3,90E+04	38670	1,0002
Log Likelihood		-49631,7898	
Full Log Likelihood		-49631,7898	
AIC (smaller is better)		99283,5797	
AICC (smaller is better)		99283,5854	
BIC (smaller is better)		99369,2079	