Mémoire présenté le : 08 / 06 / 2021

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA
et l'admission à l'Institut des Actuaires

Par : Yohan GUEGADEN

Titre : Mesure du risque de résiliation dans un portefeuille de contrats multirisques habitation

Confidentialité : ☒ NON ☐ (Durée : ☐ 1 an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires *Signature*

. . . . . .

. . . . . .

. . . . . .

Membres présents du jury de l'ISFA

. . . . . .

. . . . . .

. . . . . .

Entreprise : Le Finistère Assurance
Nom : Yann DAGORN

Signature :

Directeur de mémoire en entreprise :
Nom :

Signature :

Invité :
Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

Mots clefs

Assurance, Multirisques habitation, Résiliation, Segmentation, *Score*, Modèles logistiques, Elasticité, *Bagging*, *Boosting*, Modèles de durée, Estimateur de Kaplan-Meier, Modèles de Cox, Modèles AFT, Modèle polytomique, Solvabilité 2.

Résumé

La résiliation des contrats d'assurance est un sujet transverse d'intérêt commercial, juridique, réglementaire et actuariel.
Commercial, à au moins deux niveaux : la tarification d'un produit et le commissionnement d'un apporteur d'affaires peuvent en tenir compte.
De plus, l'évolution récente de la réglementation juridique en la matière (lois Chatel et Hamon) impacte le comportement des assurés en leur permettant de faire jouer la concurrence en résiliant à tout moment leur contrat. L'étude de leur profil commercial permettrait de comprendre quelles sont les variables qui influent sur le risque.
En outre, la réglementation prudentielle (Solvabilité 2) impose aux compagnies d'évaluer le comportement futur de leurs assurés, et notamment d'estimer plus finement le risque de résiliation associé aux primes incluses dans la frontière des contrats. Les résiliations doivent, en effet, être déduites de la Meilleure Estimation des engagements liés à ces primes.
Enfin, ces thématiques peuvent être appréhendées par des méthodes statistiques, ou actuarielles, permettant de comprendre pour mieux expliquer ou prévoir les évolutions à venir.

Keywords

Insurance, Home Multirisk, Termination, Segmentation, Score, Logistic Models, Elasticity, Bagging, Boosting, Duration Models, Kaplan-Meier Estimator, Cox Models, AFT Models, Polytomous Model, Solvency2.

Abstract

The termination of insurance contracts is a transversal subject of commercial, legal, regulatory and actuarial interest.
Commercial, on at least two levels : the pricing of a product and the commissioning of a business provider can take this into account.
Moreover, recent developments in legal regulations in this area (Chatel and Hamon laws) have an impact on the behavior of policyholders by allowing them to compete in terminating their contract at any time. The study of their commercial profile would help to understand the variables that influence risk.
In addition, the prudential regulation (Solvency 2) requires companies to assess the future behavior of their policyholders, and in particular to estimate more precisely the risk of termination associated with the premiums included in the contract boundary. The terminations must, in fact, be deducted from the Best Estimate of the commitments related to these premiums.
Finally, these themes can be apprehended by statistical or actuarial methods, making it possible to understand in order to better explain or predict future developments.

Remerciements

Je voudrais commencer par remercier l'équipe pédagogique de l'ISFA pour la qualité et la variété de ses cours, ainsi que pour l'organisation de la Formation Continue d'Actuaire.

Je tenais ensuite à remercier particulièrement Christian Y. ROBERT et Xavier MILHAUD pour leurs conseils et l'intérêt porté à mon travail.

Un grand merci également à mes relecteurs, sans oublier mes collègues du Finistère Assurance pour leur professionnalisme et leur bonne humeur.

Note de synthèse

Contexte

Nous assistons depuis quelques années à une évolution de la réglementation, qui permet au consommateur de faire jouer la concurrence en facilitant la résiliation de son contrat d'assurance (loi Chatel, loi Hamon).
Il devient ainsi fondamental, pour la compagnie, de pouvoir cerner les profils de sociétaire les plus susceptibles de résilier.
La résiliation constitue ainsi le risque que nous souhaitons mesurer. Ce travail cherchera donc à décrire, expliquer et, si possible, prévoir le comportement des détenteurs de contrats d'assurance multirisques habitation.
Nous tacherons, en chemin, de nous interroger sur les indicateurs de mesure de la performance de certains types de modèle ainsi que sur l'appréciation de la stabilité temporelle supposée des variables exogènes.

Mesure, échantillonnage et horizon

L'échantillonnage de nos données va dépendre de ce que nous souhaitons mesurer. Nous venons de voir que le risque de résiliation est l'objet principal de notre étude. Après avoir défini ce qu'est une résiliation, nous avons cherché à en mesurer le risque selon plusieurs axes :

1. L'ensemble des polices (zone grise dans la Figure 1) enregistrées dans l'historique de la mutuelle (qu'elle soit en cours à la date d'étude, ou pas).
2. Un ensemble de polices (zone bleue), depuis leur date de souscription (entre le 01/01/2010 et le 31/12/2014) jusqu'à un horizon temporel fixé (12 ou 36 mois).
3. Un ensemble de polices (zone rouge), en cours à une date fixée (31/12/2017) jusqu'à un horizon temporel fixé à 12 mois.

Ces trois axes d'analyse ne poursuivent pas le même objectif.

Le **premier axe** va nous amener à nous faire une **idée générale** du risque mesuré avec un bonne robustesse, mais sans forcément tenir compte de l'évolution de la structure du portefeuille.

Le **second axe** cible plus précisément le risque de résiliation dès la **souscription** d'un contrat. Nous voyons a minima deux finalités pratiques à cette mesure : la tarification et le commissionnement des apporteurs d'affaires.

Enfin, le **troisième axe** va chercher à mesurer le risque sur un stock de contrats en cours à une date donnée. L'idée est d'estimer le risque à n'importe quel moment de la vie d'un contrat. Les finalités peuvent être diverses et variées : actions de fidélisation des sociétaires ou calcul réglementaire du taux de chute dans Solvabilité 2.

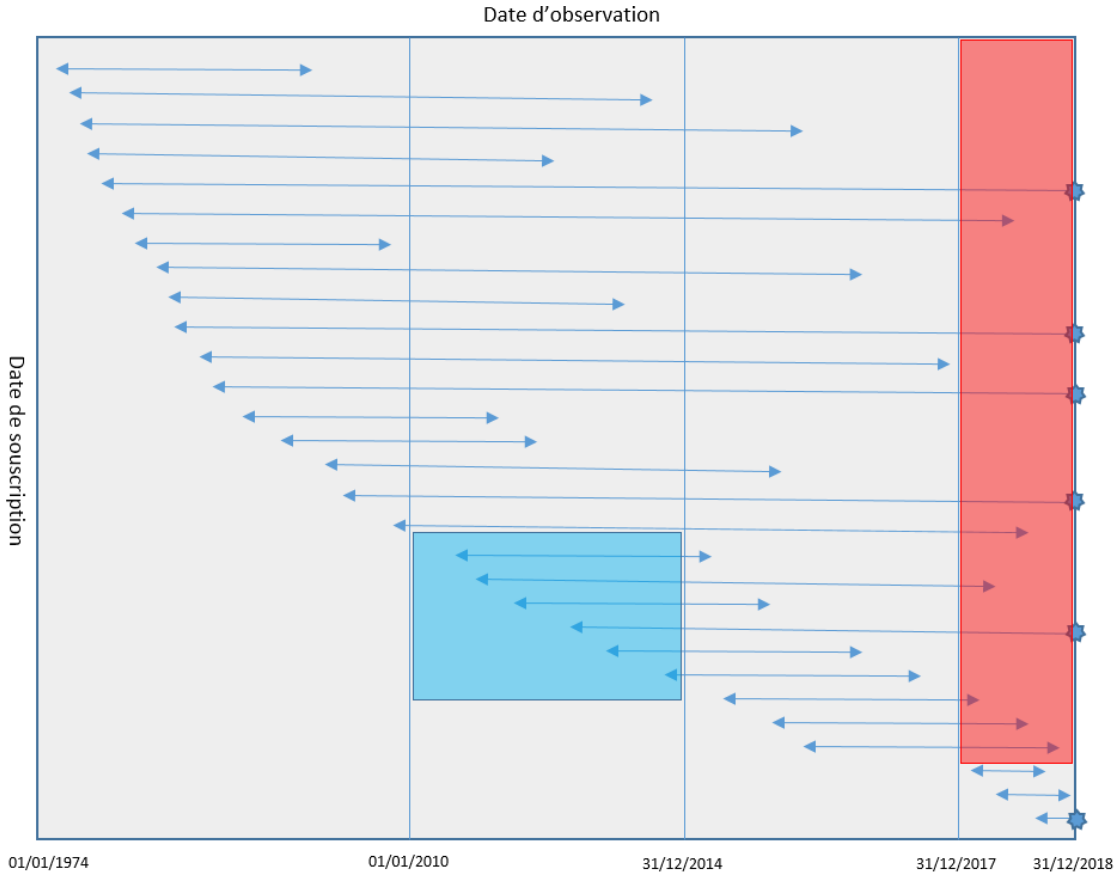


FIGURE 1 – Schéma des différents axes d'analyse du risque.

Une multitude de méthodes statistiques existe pour tenter de parvenir à nos fins. La FIGURE 2 ci-dessous schématise celles mises en œuvre dans ce travail :

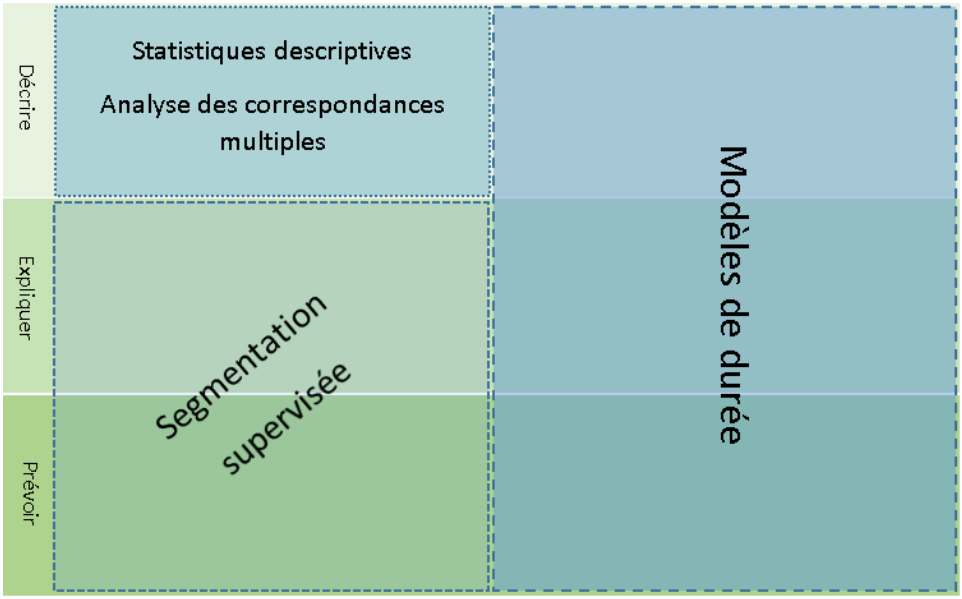


FIGURE 2 – Objectif et méthodes proposées.

Nous adopterons la démarche classique qui est d'observer (ici un ensemble de polices) afin de comprendre et d'interprêter le phénomène d'intérêt (ici la mesure d'un risque ciblé). Enfin, nous confronterons nos

hypothèses théoriques à la réalité afin d'en tirer d'éventuelles conclusions.

Décrire

Décrire le risque revient à compter son nombre d'occurences dans le temps, ou du moins sur une période donnée. Nous remarquons déjà deux choses à ce stade :

- Soit nous apprécions le risque avéré par rapport à une absence de risque : un indicateur de mesure adapté est le **taux de résiliation**, i.e. le ratio entre le nombre de polices résiliées et l'ensemble des polices du portefeuille.
- Soit nous apprécions le délai de survenance du risque : un indicateur de mesure adapté est la **durée de vie** moyenne des polices du portefeuille.

L'étude, en Figure 3, du taux de résiliation en fonction de la maturité des polices montre une forte intensité du risque lors de la première année qui suit la souscription :

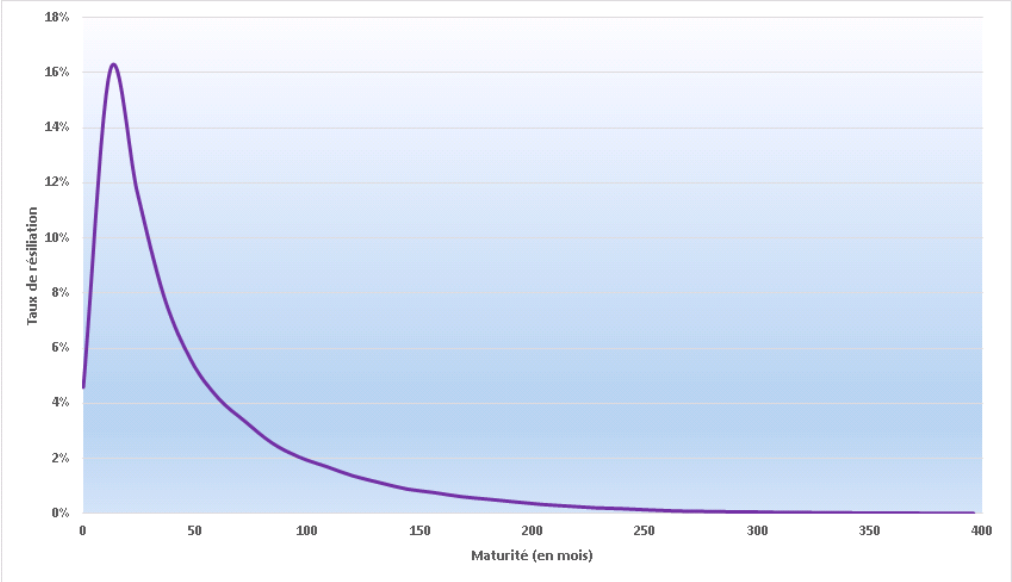


Figure 3 – Évolution mensuelle du taux de résiliation.

Bien que cet indicateur soit assez parlant et synthétique, nous avons choisi de le compléter par la durée de vie moyenne d'un contrat (qu'il soit résilié ou pas) :

Portefeuille	**Global**	**Résiliés**	**Censurés**
Agricole	6,74	6,46	7,28
Commerce	4,09	3,96	4,34
Caravane	4,03	3,94	4,27
Plaisance	3,61	3,45	3,88
Habitation	4,61	4,20	5,46
Global	**4,49**	**4,15**	**5,19**

Table 1 – Durée de vie moyenne (en années).

Nous mettons en évidence dans le tableau 1 que la durée de vie, et donc la résiliation, n'est pas homogène dans le portefeuille de la mutuelle. Ainsi, trouver des variables pertinentes permettrait de mieux comprendre et donc expliquer le risque étudié.

Nous nous intéresserons plus particulièrement aux contrats **multirisques habitation** car ces derniers sont très largement majoritaires dans le portefeuille de la mutuelle.

Expliquer

Les méthodes de **segmentation supervisée** telles que les **modèles de régression logistique** offrent des solutions relativement simples, robustes et répandues pour expliquer l'hétérogénéité des taux de résiliation présentée dans la FIGURE 4 :

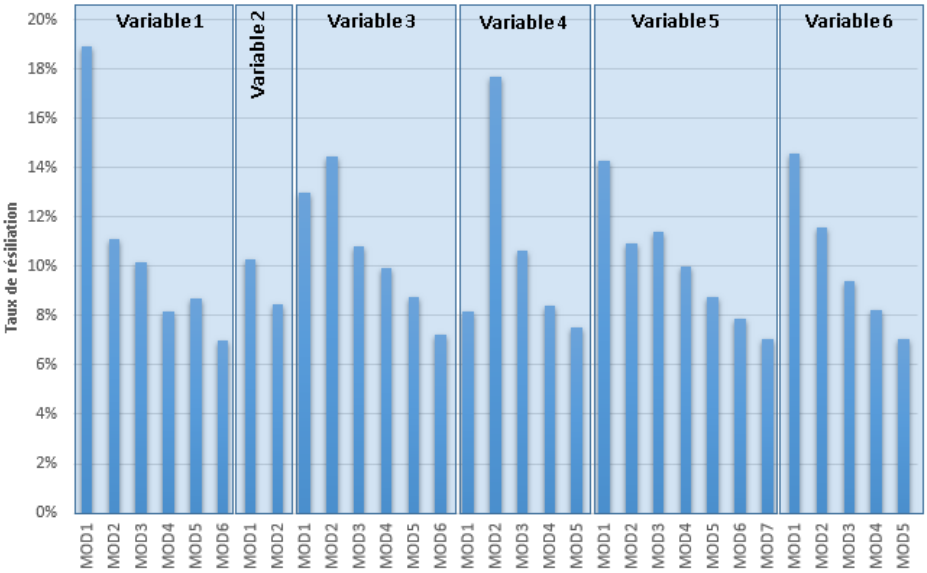


FIGURE 4 – Hétérogénéité des variables explicatives.

Nous verrons notamment comment optimiser le découpage des variables continues et le regroupement des modalités des variables qualitatives pour mettre en valeur leur pouvoir discriminant.

L'intégration de variables exogènes pertinentes (cf. tableau 2) permet également de mettre en évidence leur éventuel effet significatif sur la durée de vie des contrats :

Variable	**Modalité**	**Global**	**Résiliés**	**Censurés**
Nb pièces	2 pièces ou -	3,11	2,83	4,47
Nb pièces	3 ou 4 pièces	4,43	4,03	5,48
Nb pièces	5 ou 6 pièces	5,31	4,92	6,03
Nb pièces	7 pièces ou +	5,57	5,39	5,79
Qualité Juridique	Locataire	2,82	2,72	3,42
Qualité Juridique	PNO	4,78	4,33	5,21
Qualité Juridique	Propriétaire	6,14	6,06	6,26
Qualité Juridique	Inconnue	3,57	3,57	2,00

TABLE 2 – Durée de vie moyenne (en années) sur MRH.

Nous chercherons ainsi à mesurer l'effet de telles variables dans l'estimation de la fonction de survie via :

- Des modèles à **hasard proportionnel** (PH), en particulier le modèle semi-paramétrique de Cox. Nous vérifierons, en détail, l'hypothèse fondamentale de proportionnalité des risques à l'aide des courbes LLS et des résidus de Schœnfeld.
- Des modèles à **hasard accéléré** (AFT), dont l'hypothèse principale est que l'effet des variables explicatives est multiplicatif par rapport à la durée de survie. Nous sélectionnerons la fonction de survie la mieux adaptée à nos données grâce au test d'adéquation graphique des résidus de Cox-Snell.

Prévoir

Enfin, nous vérifierons le pouvoir prédictif (à horizon un an) des modèles testés, à partir du stock de polices MRH en cours à la fin de l'année 2017. Nous comparerons le taux prédit par nos modèles au taux réellement observé au 31/12/2018 dans le tableau 3 :

Type de Modèle	**Prise en compte de l'hétérogénéité**	**Variable(s) explicative(s)**	**Taux prédit à 1 an**	**Taux observé à 1 an**
Temps continu Kaplan-Meier	non		13,13%	9,52%
Temps discret Bernoulli	non		13,13%	9,52%
Hasard proportionnel : Cox	oui	Qualité juridique	11,31%	9,52%
Hasard accéléré (AFT)	oui	Qualité juridique	11,52%	9,52%
Hasard proportionnel : Cox	oui	Type de bien	12,43%	9,52%
Hasard accéléré (AFT)	oui	Type de bien	13,03%	9,52%
Hasard proportionnel : Cox	oui	Fréquence de paiement	12,19%	9,52%
Hasard accéléré (AFT)	oui	Fréquence de paiement	12,79%	9,52%
Logistique	oui	Sélection stepwise	10,73%	9,61%
Polytomique (Solvabilité 2)	oui	Sélection stepwise	10,32%	9,61%
Binaire (Solvabilité 2)	oui	Sélection stepwise	10,59%	9,61%

Table 3 – Synthèse des résultats obtenus sur le stock au 31/12/2017.

Un des principaux enseignements de cette étude est que la prédictivité des modèles à court terme peut, pour certains d'entre eux, être considérée comme correcte mais pas excellente. Cette conclusion était pressentie, en partie grâce aux indicateurs de mesure de la **performance** et de la **stabilité** des données.

Performance

La performance a été mesurée sur les modèles de régression logistique. Nous avons choisi de nous concentrer sur l'indicateur qui nous semblait le plus synthétique : l'AUC.
Nous retenons les éléments suivants :

- Le choix de l'échantillon de validation peut nuancer le jugement que l'on peut avoir des résultats (d'acceptable à excellent).
- Des méthodes plus complexes, telles que le *bagging* ou le *boosting* de **modèles logistiques**, ne semblent pas apporter d'amélioration significative de la performance.

Stabilité

Avant d'appliquer nos modèles, nous avons cherché à proposer des méthodes afin de mesurer (ou du moins apprécier) la stabilité temporelle :

- Des modèles de durée prenant en compte l'hétérogénéité du portefeuille : modèles PH et AFT.
- Des variables explicatives candidates à la construction du score de résiliation dès la souscription à horizon 3 ans.

Bien qu'assez contrastés, nous avons considéré que les résultats plaident globalement pour une relative stabilité des modèles de durée et des variables explicatives étudiées.

Summary

Context

For the past few years, we have been witnessing an evolution in regulation, which allows consumers to compete by facilitating the termination of their insurance contract (Chatel law, Hamon law). It thus becomes fundamental for the company to be able to identify the corporate profiles most likely to terminate. Termination is the risk we want to measure. This work will therefore seek to describe, explain and, if possible, predict the behaviour of property and casualty policyholders. Along the way, we will examine the performance measurement indicators of certain types of models as well as the assessment of the assumed temporal stability of exogenous variables.

Measurement, sampling and horizon

The sampling of our data will depend on what we want to measure. We have just seen that the risk of termination is the main focus of our study. After defining what a termination is, we sought to measure its risk along several lines :

1. The set of policies (grey area in the Figure 5) recorded in the history of the mutual (whether it is in progress on the study date, or not).
2. A set of policies (blue zone), from their subscription date (between 01/01/2010 and 12/31/2014) to a set time horizon (12 or 36 months).
3. A set of policies (red zone), in progress on a fixed date (12/31/2017) up to a time horizon of 12 months.

These three lines of analysis do not pursue the same objective.

The **first axis** will lead us to make us a **general idea** of the measured risk with a good robustness, but without necessarily taking into account the evolution of the portfolio structure.

The **second axis** more precisely targets the risk of termination from the **subscription** of a contract. We see at least two practical purposes for this measure : the pricing and commissioning of business providers.

Finally, the **third axis** will seek to measure the risk on a stock of contracts in progress on a given date. The idea is to estimate the risk at any point in the life of a contract. The aims can be varied and varied : loyalty actions of members or regulatory calculation of the fall rate in Solvency 2.

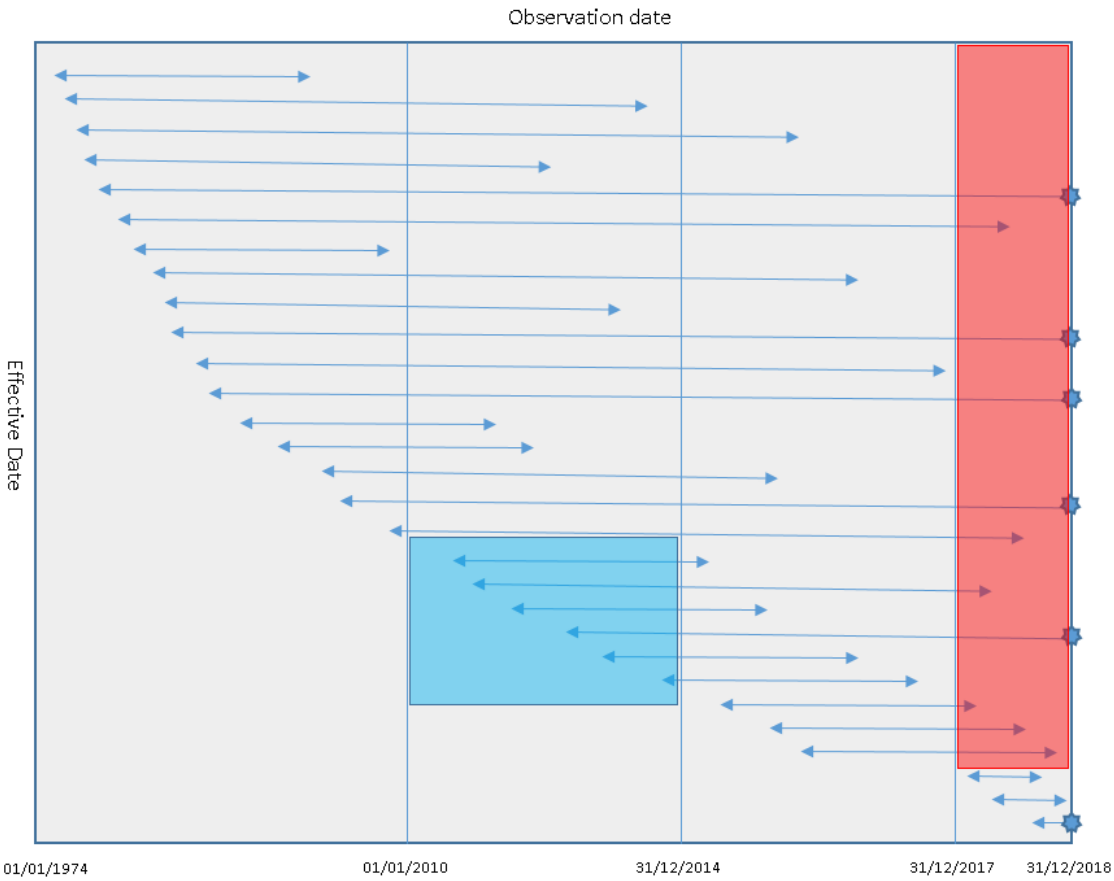


FIGURE 5 – Diagram of the different risk analysis axes.

A multitude of statistical methods exist to try to achieve our purposes. The FIGURE 2 below outlines those implemented in this work :

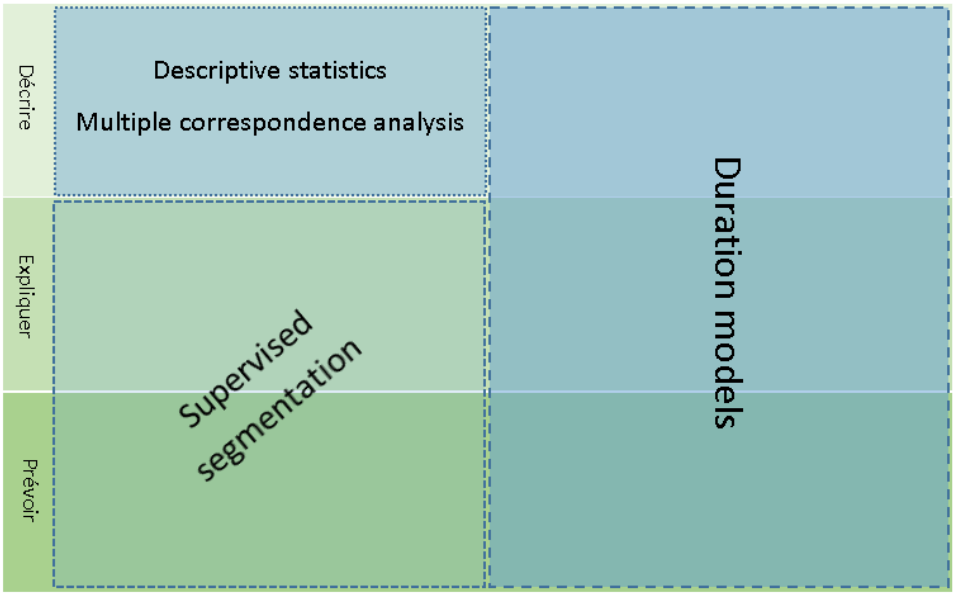


FIGURE 6 – Objective and proposed methods.

We will adopt the classical approach which is to observe (here a set of policies) in order to understand and interpret the phenomenon of interest (here the measure of a targeted risk). Finally, we will compare our

theoretical hypotheses with reality in order to draw possible conclusions.

Describe

To describe the risk is to count its number of occurrences over time, or at least over a given period of time. We are already noticing two things at this point :

- Either we assess the proven risk in relation to an absence of risk : a suitable measure indicator is the **rate of termination**, i.e. the ratio between the number of terminated policies and the total portfolio policies.
- Either we assess the risk occurrence time : a suitable measurement indicator is the average **lifetime** of the portfolio policies.

The study, in Figure 7, of the termination rate according to the maturity of the policies shows a high risk intensity in the first year following the subscription :

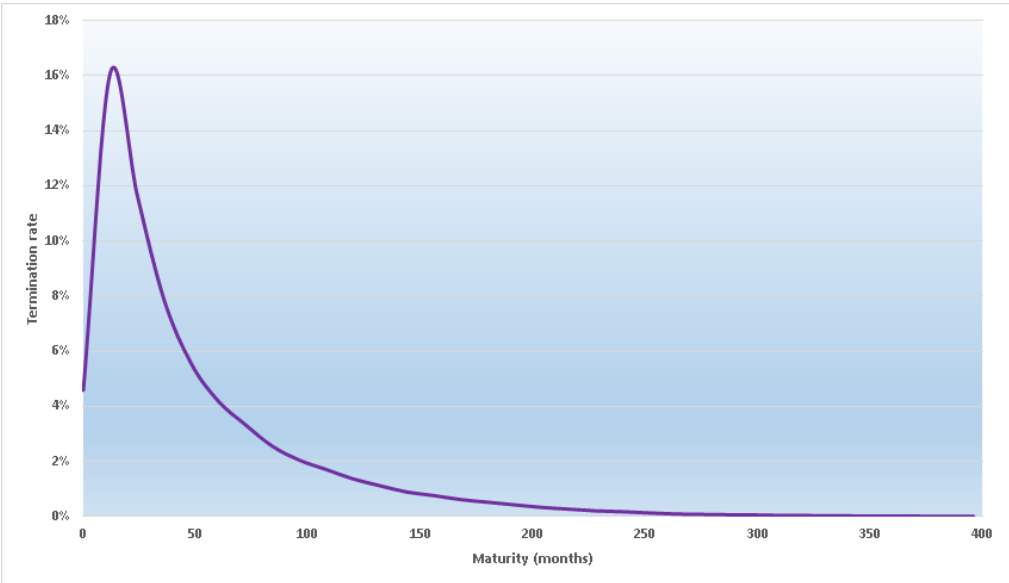


Figure 7 – Monthly evolution of the termination rate.

Although this indicator is quite telling and synthetic, we have chosen to supplement it with the average life of a contract (whether terminated or not) :

Portfolio	Global	Terminated	Censored
Agricultural	6,74	6,46	7,28
Commerce	4,09	3,96	4,34
Caravan	4,03	3,94	4,27
Boating	3,61	3,45	3,88
Inhabitation	4,61	4,20	5,46
Global	**4,49**	**4,15**	**5,19**

Table 4 – Average life (years).

We highlight in the table 4 that the lifetime, and therefore the termination, is not homogeneous in the mutual's portfolio. Thus, finding relevant variables would help to better understand and therefore explain the risk studied.

We will be particularly interested in **multi-risk housing** contracts because they are the vast majority in the mutual's portfolio.

Explain

Supervised segmentation methods such as **logistic regression models** offer relatively simple, robust and widespread solutions to explain the heterogeneity of termination rates presented in the FIGURE 8 :

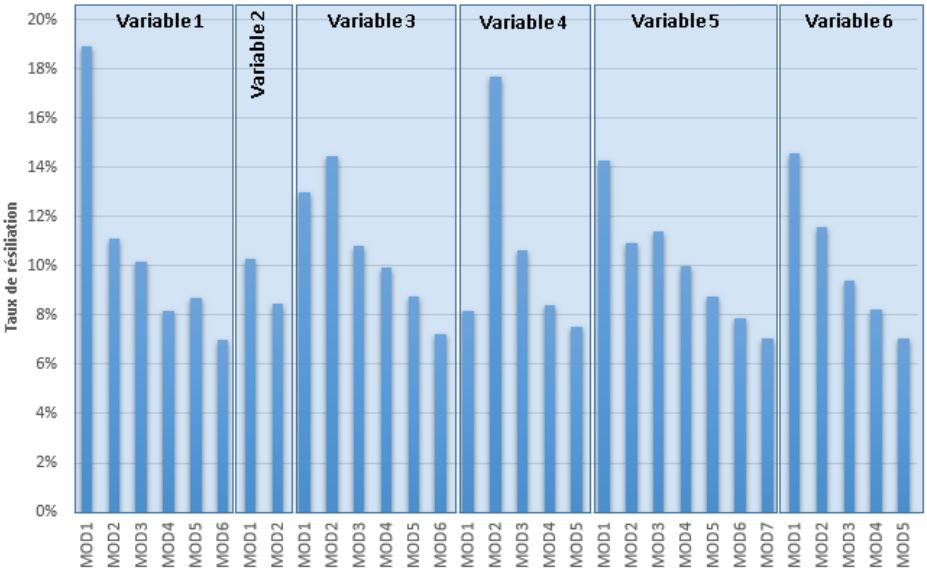


FIGURE 8 – Heterogeneity of explanatory variables.

In particular, we will see how to optimize the division of continuous variables and the grouping of qualitative variable modalities in order to highlight their discriminating power.

The integration of relevant exogenous variables (cf. table 5) also makes it possible to highlight their possible significant effect on the life of contracts :

Variable	Modality	Global	Terminated	Censored
Nb pieces	2 pieces or -	3,11	2,83	4,47
Nb pieces	3 or 4 pieces	4,43	4,03	5,48
Nb pieces	5 or 6 pieces	5,31	4,92	6,03
Nb pieces	7 pieces or +	5,57	5,39	5,79
Legal Quality	Tenant	2,82	2,72	3,42
Legal Quality	NOO	4,78	4,33	5,21
Legal Quality	Owner	6,14	6,06	6,26
Legal Quality	Unknown	3,57	3,57	2,00

TABLE 5 – Average life (years) on MRH.

We will thus seek to measure the effect of such variables in the estimation of the survival function via :

- **Proportional chance** (PH) models, in particular the semi-parametric Cox model. The basic hypothesis of proportionality of risk will be checked in detail using the LLS curves and Schœnfeld residues.
- Models with accelerated **chance** (AFT), whose main hypothesis is that the effect of explanatory variables is multiplicative in relation to survival time. We will select the survival function best suited to our data using the Cox-Snell Residue Graphical Suitability Test.

Predict

Finally, we will check the predictive power (within a year) of the models tested, from the stock of MRH policies in progress at the end of 2017. We compare the rate predicted by our models to the rate actually

observed at 12/31/2018 in the table 6 :

Model Type	**Taking into account heterogeneity**	**explanatory variable(s)**	**Predicted rate**	**Observed rate**
Continuous time Kaplan-Meier	no		13,13%	9,52%
Discreet time Bernoulli	no		13,13%	9,52%
Proportional Hazards : Cox	yes	Legal Quality	11,31%	9,52%
Accelerated failure time (AFT)	yes	Legal Qualitye	11,52%	9,52%
Proportional Hazards : Cox	yes	Type of property	12,43%	9,52%
Accelerated failure time (AFT)	yes	Type of property	13,03%	9,52%
Proportional Hazards : Cox	yes	Payment frequency	12,19%	9,52%
Accelerated failure time (AFT)	yes	Payment frequency	12,79%	9,52%
Logistic	yes	Stepwise selection	10,73%	9,61%
Polytomous (Solvency 2)	yes	Stepwise selection	10,32%	9,61%
Binaire (Solvency 2)	yes	Stepwise selection	10,59%	9,61%

Table 6 – Summary of results obtained on the stock at 12/31/2017.

One of the main lessons of this study is that the prediction of short-term models may, for some of them, be considered correct but not excellent. This conclusion was anticipated, in part due to the measurement indicators of **performance** and **stability** of the data.

Performance

Performance was measured on logistic regression models. We chose to focus on the indicator that seemed the most synthetic to us : the AUC. We retain the following elements :

- The selection of the validation sample may qualify the judgment that results can be obtained (from acceptable to excellent).
- More complex methods, such as bagging or boosting from **logistic models**, do not appear to provide significant performance improvement.

Stability

Before applying our models, we sought to propose methods to measure (or at least assess) temporal stability :

- Duration models taking into account the heterogeneity of the portfolio : PH and AFT models.
- Explanatory variables candidates for the construction of the termination score within 3 years.

Although quite contrasted, we considered that the results generally plead for a relative stability of the duration models and the explanatory variables studied.

Table des matières

Note de synthèse **8**

Summary note **13**

Introduction **18**

1 Généralités **19**
1.1 Présentation de la mutuelle . . . 19
1.1.1 Le Finistère Assurance . . . 19
1.1.2 La SGAM Covéa . . . 20
1.2 Aspects juridiques et opérationnels . . . 20

2 Présentation des données **24**
2.1 Volumétrie de départ . . . 24
2.2 Risques du portefeuille . . . 24
2.2.1 Risque de résiliation . . . 25
2.2.2 Durée de vie des contrats . . . 25
2.3 Zoom sur le risque Habitation . . . 26
2.3.1 Motifs de résiliation . . . 26
2.3.2 Évolution des résiliations . . . 26
2.3.3 Résiliation en fonction de la maturité . . . 28
2.3.4 Durée de vie des contrats . . . 28
2.4 Statistiques univariées . . . 29
2.4.1 Traitements des données manquantes et aberrantes . . . 29
2.4.2 Variables quantitatives . . . 30
2.4.3 Variables qualitatives . . . 31
2.4.4 Synthèse des variables . . . 32
2.4.5 Perspectives . . . 32

3 Étude de la durée de vie des contrats MRH **33**
3.1 Modèle en temps continu ou en temps discret ? . . . 34
3.1.1 Modèle en temps continu . . . 34
3.1.2 Modèle en temps discret . . . 35
3.1.3 Comparaison des approches . . . 36
3.2 Intégration de variables exogènes . . . 38
3.2.1 Justification de l'hétérogénéité . . . 38
3.2.2 Modèles à hasard proportionnel . . . 39
3.2.3 Modèles paramétriques AFT . . . 43
3.3 Stabilité temporelle des modèles . . . 44
3.3.1 Démarche proposée . . . 45
3.3.2 Modèle PH . . . 45
3.3.3 Modèle AFT . . . 47
3.4 Synthèse des résultats . . . 47

4 Segmentation supervisée à la souscription **51**
4.1 Définition d'une variable cible . . . 51
4.2 Statistiques bivariées des variables explicatives . . . 52
4.2.1 Découpage des variables quantitatives . . . 52
4.2.2 Regroupement des variables qualitatives . . . 54
4.2.3 Croisements de variables . . . 56
4.3 Mesure de la liaison et de la stabilité des variables qualitatives . . . 56
4.3.1 Corrélation des variables explicatives avec la variable cible . . . 57
4.3.2 Corrélation des variables explicatives entre elles . . . 58
4.3.3 Stabilité temporelle des variables explicatives . . . 59
4.4 Analyse des Correspondances Multiples . . . 61
4.5 Échantillonnage . . . 62
4.5.1 Nombres pseudo-aléatoires . . . 62
4.5.2 Méthode d'échantillonnage appliquée . . . 63
4.5.3 Compromis Biais-Variance . . . 63
4.6 Modèle considéré . . . 64
4.6.1 Sélection des variables . . . 64
4.6.2 Qualité de la discrimination . . . 65
4.6.3 Résultats de la discrimination . . . 66
4.6.4 Diagnostic de la modélisation . . . 67
4.7 Discussion sur l'horizon du risque mesuré . . . 69
4.7.1 Valeurs moyennes des coefficients à 12 et 36 mois . . . 69
4.7.2 Bootstrap des AUC à 12 et 36 mois . . . 69

5 Segmentation supervisée sur stock **72**
5.1 Périmètre de l'étude . . . 72
5.1.1 Définition de la variable cible . . . 72
5.1.2 Base de construction . . . 72
5.2 Résultats de la régression . . . 73
5.2.1 Application du modèle au stock du 31/12/2017 . . . 73
5.2.2 Comparaison avec les résultats obtenus par les modèles de durée . . . 74
5.3 Autres méthodes de segmentation . . . 75
5.3.1 Bagging . . . 75
5.3.2 Boosting . . . 76
5.3.3 Application à nos données . . . 77
5.4 Élasticité du taux de résiliation à la prime . . . 81
5.4.1 Généralités . . . 81
5.4.2 Application aux contrats d'assurance . . . 82

6 La résiliation dans Solvabilité 2 **85**
6.1 Rappels sur Solvabilité 2 . . . 85
6.1.1 Notion de frontière des contrats . . . 85
6.1.2 Provisions techniques . . . 86
6.2 Exemple illustratif . . . 86
6.3 Méthode actuelle . . . 87
6.4 Evolution possible de la méthode actuelle . . . 88

Conclusion **91**

A Annexe - Généralités **94**

B Annexe - Présentation des données **95**

C Annexe - Etude de la durée de vie des contrats MRH **105**

D Annexe - Segmentation supervisée à la souscription **122**

E Annexe - Segmentation supervisée sur stock **131**

F Annexe - La résiliation dans Solvabilité 2 **141**

Liste des figures **144**

Introduction

Pour des raisons stratégiques et réglementaires, une attention particulière doit être apportée au phénomène de résiliation des polices d'assurance.
En effet, nous assistons depuis quelques années à une évolution de la réglementation juridique à ce sujet, afin notamment de permettre au consommateur de faire jouer la concurrence en résiliant, de plus en plus à sa guise, son contrat (loi Chatel, loi Hamon...).
Pour optimiser la tarification voire le commissionnement, il est fondamental de bien cerner les profils de clients les plus appétants à résilier afin d'ajuster les chargements selon l'intérêt de la mutuelle.
Ce travail cherchera donc à décrire, expliquer et, si possible, prévoir [1] le comportement des détenteurs de contrats d'assurance.

Le comportement humain est difficile, pour ne pas dire impossible, à prévoir parfaitement. Cela peut sembler rassurant de penser qu'une part de libre arbitre subsistera toujours malgré la volonté affichée par divers acteurs (états, entreprises, etc.) de mesurer en permanence la qualité et l'utilisation de leurs services. Ce contrôle s'améliore régulièrement grâce à l'apport conjugué de la technologie et du consentement des usagers [2].
Ainsi, les énormes masses de données collectées ces dernières années, conséquences du développement de l'économie numérique, permettent de mieux comprendre les consommateurs. Cependant, d'une part ces volumes engendrent beaucoup de bruit inutile dans les données ; d'autre part, leur obsolescence est en général très rapide et peut amener les analystes à de mauvaises prédictions. C'est pour cette dernière raison que notre étude mettra souvent l'accent sur la notion de stabilité temporelle des données [3].

Après une brève présentation du contexte juridique dans le chapitre 1 et du portefeuille de la mutuelle dans le chapitre 2, nous nous intéresserons plus précisément aux contrats multirisques habitation car ces derniers y sont très largement majoritaires.

La résiliation constitue ainsi le risque que nous souhaitons mesurer. Comme pour tout risque étudié, nous devons définir une durée ou horizon d'observation, c'est-à-dire une date de départ et une date de fin.
Nous verrons deux façons d'appréhender la problématique du point de départ de l'observation : soit dès la souscription d'un contrat dans le chapitre 4, soit à une date fixée [4] pour l'ensemble d'un stock de polices dans les chapitres 3, 5 et 6.
Concernant la durée d'observation, l'objectif est de récolter suffisamment de résiliations pour pouvoir en mesurer le risque. Nous comparerons donc, dans le chapitre 4 les résultats obtenus à partir d'une durée à moyen terme (3 ans) avec ceux obtenus avec une durée à court terme (1 an).
Pour les autres chapitres (3, 5 et 6), la durée d'observation sera annuelle.

Pour comprendre quels sont les assurés les plus susceptibles d'arrêter leur contrat, plusieurs options statistiques complémentaires seront proposées :

- Des méthodes faisant intervenir des modèles de durée, dans le chapitre 3 : nous discuterons, notamment du type de modèle à utiliser (en temps discret ou continu), de l'apport de la prise en compte de l'hétérogénéité dans les modèles et enfin de leur prédictivité.

1. En référence au livre de M. Tenenhaus [Ten07].
2. L'aspect juridique de la collecte et du traitement des données personnelles est traité au niveau européen par le règlement général sur la protection des données (ou RGPD) en vigueur depuis le 25 mai 2018, en complément de la loi française Informatique et Liberté de 1978.
3. Nous le verrons lors de la présentation des données, nous ne sommes pas vraiment concernés par la problématique de « malédiction de la dimension » car assez peu de données clients sont récoltées.
4. Au 31/12 par exemple.

- Comme il s'agit de classer et d'affecter les individus au bon groupe homogène, nous envisagerons quelques méthodes de segmentation :
  - Des méthodes classiques faisant intervenir les modèles linéaires généralisés comme la régression logistique dans les chapitres 4 et 5 : largement implémentée dans les logiciels d'analyse, cette dernière constitue une référence car elle allie interprétabilité des résultats et robustesse.
  - Enfin, nous nous intéresserons à des méthodes de plus en plus utilisées dans le cadre du « *machine learning* » : le *bagging* et le *boosting*. Ces deux dernières méthodes compléteront celles vues auparavant (chapitre 5), dans le but de tenter de remédier à leurs insuffisances.

Afin de lier la tarification des produits et le phénomène de résiliation, nous essaierons d'étudier l'impact d'une augmentation de la cotisation sur l'attrition des assurés à la fin du chapitre 5. Cet impact est connu sous le nom d'**élasticité** du taux de résiliation à la prime.

Pour finaliser ce travail (chapitre 6), nous nous placerons dans le cadre de Solvabilité 2. Au niveau du premier pilier [5], il est demandé, lors du calcul de la « meilleure estimation des provisions pour primes » de tenir compte du comportement des assurés. Un taux de résiliation [6] est ainsi calculé pour les contrats du portefeuille inclus dans la « frontière des contrats ».
Actuellement, deux types de taux sont estimés dans une logique de calcul a posteriori. Nous allons donc, d'une part calculer ces taux a priori à l'aide d'un modèle de régression logistique multinomial. D'autre part, nous vérifierons que les taux obtenus sont cohérents entre les visions *ex post* et *ex ante*. Si tel est le cas, alors l'estimation des taux de chute pourrait être envisagée plusieurs mois avant l'inventaire annuel Solvabilité 2.

5. Exigences quantitatives.
6. Aussi appelé : taux de chute, ou taux de rachat.

Chapitre 1

Généralités

Après avoir très brièvement présenté la mutuelle et son écosystème, nous reviendrons sur le contexte juridique de la clôture d'un contrat d'assurance ainsi que sur quelques impacts opérationnels.

1.1 Présentation de la mutuelle

1.1.1 Le Finistère Assurance

Le Finistère Assurance [1] est une société d'assurance mutuelle à cotisations fixes, exerçant en France depuis 1874 une activité d'assurance non-vie. Les risques des particuliers et des professionnels constituent ses principaux marchés, comme nous le montre le diagramme de la Figure 1.1.

- Sur le marché des **risques des particuliers**, la société commercialise des contrats d'assurance Multirisques Habitation, Caravanes et Plaisance.
- Sur le marché des **risques des professionnels**, elle gère un portefeuille de contrats Multirisques Professionnels et Multirisques Agricoles.

Le Finistère Assurance s'appuie sur un réseau de distribution d'environ 1 400 intermédiaires indépendants (courtiers, intermédiaires avec ou sans mandat) répartis sur 16 départements du Grand Ouest. Au siège quimpérois, une équipe de 60 personnes travaille notamment au service de ces apporteurs d'affaires.

Notons que, dès 1903, Le Finistère s'est rapproché de la Mutuelle Générale Française Accidents (à la base de MMA). Au fil des années ces liens se sont renforcés et ont amené le Finistère Assurance à adhérer, en 2005, à la SGAM Covéa.

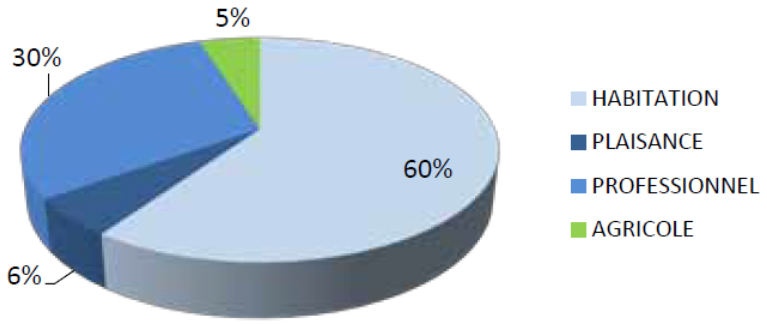


Figure 1.1 – Répartition de l'encaissement par produit pour 2019

1. Le rapport annuel 2019 est disponible à l'adresse suivante https://www.finistere-assurance.com/web/upload/directupload/1591773408-rapport-annuel-pour-ag-2020.pdf. Le SFCR est consultable ici https://www.finistere-assurance.com/web/upload/directupload/1586504964-sfcr-le-finistere-2019.pdf

1.1.2 La SGAM Covéa

Qu'est-ce-qu'une SGAM [2] ?

Une Société de Groupe d'Assurance Mutuelle est une société créée dans le but de permettre à ses membres d'établir des liens de solidarité financière importants et durables entre eux, tout en conservant leur indépendance, leur marque et leur réseau commercial [3].
La directive Solvabilité 2 a eu pour effet de renforcer les pouvoirs de ce type de structure sur les organismes affiliés.

Covéa, première SGAM créée en France

Covéa est la première SGAM créée en France en 2004 pour permettre aux mutuelles de s'unir afin de rivaliser avec les sociétés anonymes, incontournables sur le marché assurantiel, dont les structures juridiques permettaient ces regroupements depuis longtemps.
Covéa a pour mission de veiller à la pérennité et au développement de la solidarité financière entre ses adhérents. Dix entités appartenant au monde mutualiste, réparties en quatre familles, sont affiliées à Covéa :

- **Famille MMA** : MMA IARD Assurances Mutuelles, MMA Vie Assurances Mutuelles, DAS Assurances Mutuelles, Le Finistère Assurance, SMI.
- **Famille MAAF** : MAAF Assurances, MAAF Santé, APGIS.
- **Famille AM** : Assurances Mutuelles de France.
- **Famille GMF** : La Garantie Mutuelle des Fonctionnaires.

Parmi ces entreprises, certaines sont actionnaires de Covéa Coopérations, société de réassurance de droit français qui détient, à son tour, directement ou indirectement, tout ou partie du capital des sociétés opérationnelles du Groupe. Un organigramme simplifié de la structure est visible en FIGURE A.1, page 94.
Covéa Coopérations constitue ainsi la structure pivot entre les entreprises affiliées à la SGAM et les sociétés opérationnelles du groupe.
Le Finistère Assurance occupe dans l'organisation du groupe Covéa, la position de **mutuelle affiliée**.

1.2 Aspects juridiques et opérationnels

Cette partie est inspirée de l'ouvrage [AJB13] pour la partie juridique et de [TT17] pour la partie gestion opérationnelle.

Nous présenterons chronologiquement l'évolution des lois concernant la possibilité offerte au souscripteur de dénoncer son contrat d'assurance. Les trois dernières lois seront plus particulièrement détaillées car elles nous semblent les plus à même d'influencer le comportement de résiliation des consommateurs. Enfin, l'aspect opérationnel de la résiliation d'un contrat sera abordé succinctement.
Pour simplifier le propos, nous nous référerons, si possible, aux noms des ministres ou secrétaires d'état sous l'égide desquels les lois ont été promulguées.

Rappels sur le droit du contrat d'assurance

Pour plus de clarté, nous nous plaçons dans le cadre d'un **contrat souscrit par un particulier** d'une durée de validité d'un an avec tacite reconduction.

Rappelons qu'un contrat d'assurance est une convention synallagmatique [4] impliquant généralement deux parties : un assureur qui s'engage à garantir au niveau d'indemnisation le plus juste, en cas de sinistre, et un souscripteur en échange du paiement d'une cotisation [5].

2. On parle de SGAM dans le cas de mutuelles de type code des assurance, de SGAPS pour les IP, d'UMG pour les mutuelles 45.
3. Article R322-160 du code des assurances.
4. Convention par laquelle les parties s'obligent réciproquement l'une envers l'autre.
5. Cotisation, pour les sociétés d'assurance de type mutuelle ; Prime, pour les sociétés commerciales d'assurance.

Le contrat est donc généralement bilatéral, ce qui implique qu'il peut prendre fin à l'initiative d'une des deux parties.
Mais cela n'a pas toujours été aussi simple. Faisons un retour quelques dizaines d'années en arrière.

La loi Godart (1930) La loi du 13 juillet 1930, concernant les contrats d'assurance terrestre, a permis à chacune des deux parties de résilier un contrat tous les dix ans, sous réserve de prévenir l'autre partie au moins six mois à l'avance.
Avant l'application de cette loi, la durée imposée aux contrats d'assurance était très longue et obligeait bien souvent les assurés à payer une indemnité de résiliation très importante.

La loi du 11 **juillet** 1972 Les modalités d'application de cette loi sont assez alambiquées.
En effet, dans un premier temps, l'assuré ne pouvait résilier son contrat qu'au terme de deux périodes de trois ans succédant la souscription.
Dans un second temps, les deux parties étaient en mesure de résilier à l'issue de chaque année.

La loi Evin (1989**)** La loi du 31 décembre 1989 est venue modifier l'article L.113-12 du Code des assurances et est à la base de la philosophie actuelle concernant la possibilité de résilier : en fonction du risque garanti, les deux parties ont désormais le droit de résilier un contrat annuellement dès la souscription.

Les trois dernières lois qui vont être présentées ci-après sont plus emblématiques de la volonté du législateur de mettre en compétition les différents acteurs proposant des produits d'assurance, dans une logique de protection du consommateur.

La loi Chatel (2005) La loi du 28 janvier 2005 [6] a donné lieu à l'article L.113-15-1 du Code des assurances.
La plupart des contrats d'assurance pour les particuliers sont à reconduction annuelle automatique, et ce, afin que le consommateur ne se retrouve pas sans protection à la date d'échéance (et également afin que l'assureur ne perde pas la gestion d'un contrat).
Cette loi a imposé à l'assureur d'informer le souscripteur avant le début de préavis de résiliation de son contrat car en général ce dernier ne pense éventuellement à résilier que lorsqu'il reçoit son avis d'échéance de prime.
Le délai est habituellement d'un ou deux mois [7] ; l'assureur est donc tenu d'informer l'assuré quinze jours avant le début du délai de préavis.

Précisons que cette loi ne concerne pas les contrats non tacitement reconductibles tels que les assurances-vie, les assurances collectives, les assurances de risques professionnels, les contrats groupes ou encore les contrats associés à des flottes de véhicules.

La loi Lagarde (2010) La loi du 1er juillet 2010 [8] a surtout impacté le domaine du crédit à la consommation [9].
Un de ses effets a été la mise en place d'une libre concurrence pour le choix par l'emprunteur de son assurance de prêt (sous réserve de garanties équivalentes) lors de son octroi.
Il est estimé, trois à quatre ans après cette mesure, que cette possibilité a été largement sous-exploitée par les consommateurs. En effet, selon l'Inspection Générale des Finances [10], les banques possèdent environ 80 % des contrats de groupe.
Le rapport de l'IGF pointe également le risque de démutualisation croissant lié à une segmentation des clients de plus en plus fine mettant de coté, via des tarifs plus élevés, les risques les plus forts et donc bien souvent les populations les plus fragiles (d'un point de vue de la santé ou des revenus par exemple).
Les résultats mitigés de cette loi ont donc poussé les pouvoirs publics à la compléter en 2014 par la loi dite « Hamon ».

6. Loi n° 2005-67.
7. Si le délai n'est pas précisé contractuellement, il est légalement de deux mois.
8. Loi n° 2010-737.
9. Transposition en droit français de la directive européenne sur les crédits à la consommation.
10. Cf. la synthèse du rapport « Assurance-emprunteur » de novembre 2013 à télécharger à l'adresse http://www.igf.finances.gouv.fr/files/live/sites/igf/files/contributed/IGF%20internet/2.RapportsPublics/2013/2013-M-086.pdf.

La loi Hamon (2014) La **loi consommation** [11], promulguée le 17 mars 2014, a pour but de **redonner du pouvoir au consommateur** [12] en créant de nouveaux droits pour les assurés et en renforçant les obligations des assureurs [13].
Nous dénombrons en effet plus d'une vingtaine de mesures susceptibles d'informer ou de protéger le consommateur de pratiques commerciales pouvant parfois être considérées comme abusives ou, du moins, non justifiées.
Parmi ces mesures, quelques-unes concernent le secteur de l'assurance :

- **Résiliation de l'assurance emprunteur** : valable seulement pendant toute la première année, sous réserve que les garanties du nouveau contrat soient au moins équivalentes à celles proposées initialement par la banque octroyant le crédit. La loi a été complétée en 2017 par l'amendement Bourquin, qui permet aux assurés de résilier à chaque date anniversaire du contrat.
- **Résiliation des assurances tacitement reconductibles** : valable après un an de contrat pour tout particulier ayant souscrit une assurance automobile ou multirisque habitation.
- **Renonciation assurances affinitaires** : valable pour toute assurance associée à un bien ou service (voyage, téléphonie, ...) pour laquelle l'assuré est déjà couvert.
- **Action de groupe** : cette possibilité d'adhérer à une action de groupe pour obtenir la réparation des dommages matériels subis lors d'un acte de consommation ne cible pas spécifiquement les assureurs. Elle induit cependant un **risque d'image** ou de réputation (difficilement mesurable) pour les compagnies en cas de litiges récurrents.

Limites des lois censées protéger les consommateurs

Paradoxalement, bien que les évolutions législatives contribuent à stimuler la concurrence entre les assureurs, il n'est pas certain, à ce jour, que l'intérêt à long terme des consommateurs soit préservé.
Nous avons déjà abordé plus haut le risque de démutualisation des risques, mais notons également que faciliter les résiliations augmente considérablement la volatilité des portefeuilles ce qui risque d'induire *in fine* une augmentation notable des coûts de fidélisation des clients ainsi qu'une hausse des coûts d'acquisition de nouveaux contrats.
Il semblerait que la loi Hamon ait profité plus aux bancassureurs qu'aux assureurs classiques. Les bancassureurs ont en effet davantage de contacts avec leurs clients que les assureurs ainsi que bien souvent un système d'information plus enrichi, permettant de les segmenter plus finement.

Concrètement : impacts opérationnels

Assez logiquement, la résiliation se fait soit à l'initiative de l'assuré soit à l'initiative de l'assureur.

Résiliation par l'assuré S'il s'agit d'un contrat pour un professionnel, la résiliation ne se fait habituellement qu'à l'échéance anniversaire du contrat, après avoir respecté le délai de préavis contractuel (souvent un mois).
S'il s'agit d'un contrat pour un particulier ayant la qualité de consommateur, alors les lois Chatel et Hamon s'appliquent et nous distinguons deux cas de figure :

- **Résiliation à l'échéance** : en vertu de la loi Chatel, l'assureur est tenu d'adresser à l'assuré l'avis d'échéance au moins quinze jours avant le délai de préavis contractuel ou légal.
  Si cet avis parvient tardivement, c'est-à-dire moins de quinze jours avant la date limite, alors l'assuré aura vingt jours [14] pour dénoncer la reconduction.
- **Résiliation hors échéance** : en vertu de la loi Hamon, l'assuré peut résilier à tout moment son contrat après un an. La résiliation prend effet un mois et un jour [15] après la date notifiée sur le cachet de la poste de la lettre recommandée. Le principe de divisibilité de la prime s'applique [16].

11. Loi n° 2014-344.
12. cf. https://www.economie.gouv.fr/loi-consommation
13. Elle rajoute l'article L.113-15-2 au Code des assurances.
14. A partir de la date d'envoi de l'avis d'échéance tardif par l'assureur.
15. Le lendemain à minuit.
16. La portion de prime à rembourser par l'assureur se fait au prorata temporis de l'échéance.

Résiliation par l'assureur L'article L.113-9 précise qu'en cas de **fausse déclaration non intentionnelle** découverte avant sinistre, l'assureur peut demander la résiliation du contrat. Cette dernière prendra effet dix jours après notification adressée à l'assuré par lettre recommandée et le principe de divisibilité de la prime s'applique.
Notons qu'en cas de découverte de la fausse déclaration non intentionnelle après un sinistre, l'assureur règle une indemnité réduite en application de la règle proportionnelle de primes. Il conserve par la suite une faculté conventionnelle de résiliation après sinistre.
L'article L.113-3 précise qu'en cas de **non-paiement de la prime**, et suite à un process de recouvrement favorable à l'assuré, l'assureur peut résilier dix jours après sa suspension un contrat d'assurance de dommages.

Résiliation par l'assuré ou l'assureur La résiliation peut se faire par l'une des deux parties en cas de :

- **Modification du risque couvert** : aggravation ou diminution du risque, aliénation [17] de la chose assurée.
- **Modification de la situation personnelle de l'assuré** : changement de domicile, de situation ou de régime matrimonial, de profession, retraite ou cessation définitive d'activité professionnelle.

Ces précisions montrent que la dynamique de protection du consommateur vient quelque peu chambouler les habitudes commerciales prises par les compagnies d'assurance depuis de nombreuses décennies.
L'arrivée d'autres acteurs (bancassureur, fintech, gafam, etc.) vient également concurrencer les assureurs historiques dans le cœur même de leur métier, ce qui les incite à s'adapter à ces nouvelles contraintes.
Intéressons-nous désormais à ce qui se passe dans notre compagnie. Le chapitre suivant va présenter les données mises à notre disposition ainsi que quelques indicateurs stratégiques de suivi des résiliations.

17. Transmission qu'une personne fait d'une propriété ou d'un droit.

Chapitre 2

Présentation des données

La qualité des résultats d'une modélisation dépend, plus que des techniques statistiques utilisées (voir [Tuf05]), en grande partie du soin apporté à toutes les étapes de **nettoyage** et d'analyse exploratoire des données réalisées en amont. Ces étapes, non exhaustives dans ce mémoire, seront décrites au fur et à mesure.

2.1 Volumétrie de départ

Le **système de gestion** des polices est composé d'un **extranet**, permettant aux courtiers de saisir des projets, et d'un **intranet** permettant aux équipes de l'entreprise de traiter les projets transmis (affaires nouvelles, refontes, avenants, résiliations, etc.).
Nous avons ainsi extrait 733 676 observations (polices) récoltées à partir de la base de données issue de l'intranet au 31/12/2017. Ces données ont subi un nettoyage préliminaire.

Exclusions Ce premier nettoyage nous a permis de nous familiariser avec la base de données et d'effectuer un élagage assez important des polices jugées non pertinentes. Nous avons ainsi supprimé 222 762 observations, dont :

- Les contrats **sans effet** : 195 536 observations. Il s'agit des projets transmis n'ayant pas abouti à la création de polices.
- Les contrats **hors périmètre** : 29 961 observations (contrats de type **polices collectives**, **protection juridique**, etc.).
- Les produits avec **peu de volume** : 1 235 observations. Ces produits ne sont plus vendus ou trop spécifiques (comme par exemple, les assurances pour instruments de musique).
- Diverses **anomalies** : 189 observations. Sans rentrer dans un inventaire à la Prévert, des contrats peuvent être segmentés en professionnels alors que ce sont des contrats pour les particuliers et inversement. D'anciens contrats sont concernés.

Précisons que :

⇨ Une police peut se trouver dans plusieurs cas (sans effet et hors périmètre par exemple).

⇨ Au final, il reste 510 914 polices.

⇨ Nous n'avons pas effectué de retraitement, voire d'exclusion, des données manquantes ou aberrantes **à ce stade**.

2.2 Risques du portefeuille

Le portefeuille est segmenté en cinq segments ou types de risques : le risque habitation est, de loin, le plus représenté (cf. tableau 2.1).
Le contrat le plus ancien en portefeuille date de 1974 et environ 50% des polices ont une date d'effet antérieure à 2006.

Type de risque	**Fréquence**	**%**
Habitation	383 344	75,03
Commerce	67 321	13,18
Plaisance	33 326	6,52
Caravane	17 602	3,45
Agricole	9 321	1,82
Total	**510 914**	**100**

TABLE 2.1 – Volumétrie du portefeuille.

2.2.1 Risque de résiliation

Le tableau 2.2 nous montre les taux de résiliation, observés à partir de la souscription, selon différents horizons, par risque.

Type de risque		**Résiliations**			**Total**
		à 12 mois	**à 36 mois**	**à fin 2017**	
Agricole	*Nb*	682	2 133	6 114	
	%	*7,32*	*22,88*	*65,59*	**9 321**
Commerce	*Nb*	9 556	24 574	43 803	
	%	*14,19*	*36,50*	*65,07*	**67 321**
Caravane	*Nb*	2 593	6 909	12 565	
	%	*14,73*	*39,25*	*71,38*	**17 602**
Plaisance	*Nb*	4 411	11 848	20 680	
	%	*13,24*	*35,55*	*62,05*	**33 326**
Habitation	*Nb*	59 909	146 593	259 114	
	%	*15,63*	*38,24*	*67,59*	**383 344**
Total	***Nb***	**77 151**	**192 057**	**342 276**	
	%	***15,10***	***37,59***	***66,99***	**510 914**

TABLE 2.2 – Taux de résiliation par risque.

À l'exception du risque agricole (dont la volumétrie est relativement faible), les taux constatés sur les autres risques ne semblent ni trop homogènes ni trop hétérogènes. **Nous allons cependant nous intéresser au rique habitation, car il constitue la part la plus importante du portefeuille de l'entreprise** [1].

2.2.2 Durée de vie des contrats

Quelques statistiques descriptives Afin de nous faire une première idée de la distribution de la durée de vie des contrats du portefeuille, commentons quelques statistiques élémentaires, sur l'ensemble du portefeuille. Trois types de durée seront distingués :

- La durée de vie sachant que le contrat est résilié.
- La durée de vie sachant que le contrat est toujours en cours à la fin de l'année (cas de **censure à droite**, cf. annexe C, page 105, pour une définition précise).
- La durée de vie globale (sans distinguer la résiliation de la censure).

Ce premier tableau 2.3 nous présente les volumétries en jeu dans notre portefeuille de départ. Rappelons qu'il s'agit des polices observées entre le 08/07/1974 et le 31/12/2017. La part des polices résiliées est ainsi très importante dans l'historique du portefeuille (environ les deux tiers).

Le tableau 2.4 présente la durée de vie moyenne, en années, des polices avant résiliation sur le même périmètre que le tableau précédent. Deux phénomènes sont observés :

1. Et afin de ne pas trop nous disperser dans l'étude des variables disponibles.

Echantillon	Variable	Modalité	Global	Résiliés	Censurés
PTF	**GLOBAL**		**510 914**	**342 276**	**168 638**
PTF	RISQUE	Agricole	9 321	6 114	3 207
PTF	RISQUE	Commerce	67 321	43 803	23 518
PTF	RISQUE	Caravane	17 602	12 565	5 037
PTF	RISQUE	Plaisance	33 326	20 680	12 646
PTF	**RISQUE**	**Habitation**	**383 344**	**259 114**	**124 230**

TABLE 2.3 – Comptages sur l'ensemble du portefeuille.

- La durée de vie moyenne des polices censurées est supérieure à celles des résiliées.
- La durée de vie moyenne semble significativement différente pour chacun des risques du portefeuille (sauf pour les risques Commerce et Caravane).

Echantillon	Variable	Modalité	Global	Résiliés	Censurés
PTF	**GLOBAL**		**4,49**	**4,15**	**5,19**
PTF	RISQUE	Agricole	6,74	6,46	7,28
PTF	RISQUE	Commerce	4,09	3,96	4,34
PTF	RISQUE	Caravane	4,03	3,94	4,27
PTF	RISQUE	Plaisance	3,61	3,45	3,88
PTF	**RISQUE**	**Habitation**	**4,61**	**4,20**	**5,46**

TABLE 2.4 – Durée de vie moyenne (en années) avant résiliation.

2.3 Zoom sur le risque Habitation

2.3.1 Motifs de résiliation

Le tableau 2.5 synthétise les principales raisons pour lesquelles un sociétaire décide de résilier son contrat. Pour des raisons de lisibilité, nous avons choisi d'agréger dans la modalité « Autres causes » les motifs de faibles volumétries ou non explicites.

Motifs	Fréquence	%
Décès de l'assuré	3 096	1,19
Non paiement de la prime	38 306	14,78
Vente du bien assuré	42 658	16,46
Déménagement	51 334	19,81
Autres causes	123 720	47,75
Total	**259 114**	**100**

TABLE 2.5 – Motifs de résiliation pour les contrats MRH.

Sans grande surprise, il est observé que les principales causes de résiliation identifiées sont le déménagement et la vente du bien (selon que le sociétaire soit locataire ou propriétaire).
Les départs pour la concurrence sont noyés dans les « autres causes ».

2.3.2 Évolution des résiliations

La FIGURE 2.1 montre l'évolution annuelle, en nombre, des affaires nouvelles et des résiliations ainsi que la proportion $\frac{\text{Nb Resiliations}}{\text{Nb Affaires Nouvelles}}$ correspondante. Idéalement, une compagnie cherche à ce que ce taux diminue dans le temps [2].

2. La stratégie peut être différente au niveau produit, notamment en cas de mauvaise mutualisation des risques.

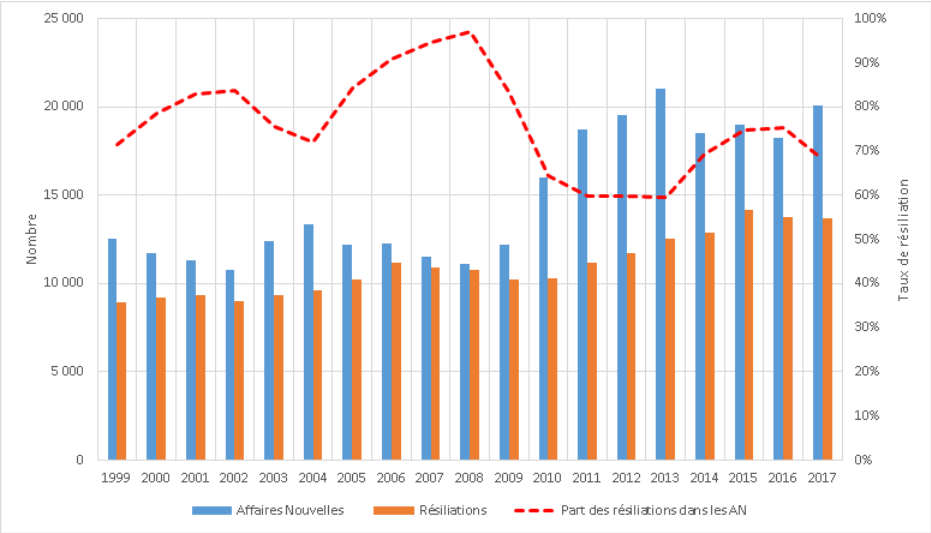


FIGURE 2.1 – Évolution annuelle des affaires nouvelles et résiliations.

Ce graphique met en évidence la dynamique de la production de contrats à partir de 2010. En effet, c'est à partir de cette époque que le développement commercial d'autres régions du grand ouest s'est intensifié. Pour la période [2011 – 2016], nous notons une constance de la part des résiliations jusqu'en 2013, puis une nette augmentation jusqu'en 2015/2016 (période post loi Hamon) suivie d'une baisse en 2017.

La FIGURE 2.2 montre l'évolution annuelle, en nombre, du stock de polices MRH et des résiliations ainsi que la proportion $\frac{\text{Nb Resiliations}}{\text{Stock Polices}}$ correspondante.
Nous avons également rajouté, pour information, la part des affaires nouvelles résiliées parmi les résiliations de l'année.

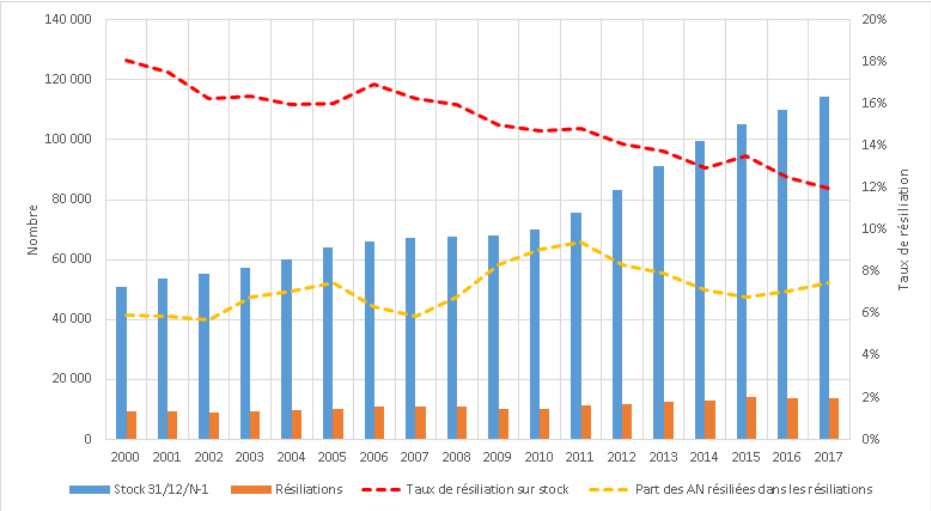


FIGURE 2.2 – Évolution annuelle du stock et des résiliations.

Une baisse tendancielle du taux de résiliation entre 2000 et 2017 est très clairement observée.
L'année 2015 (période post loi Hamon) est intéressante : nous constatons une augmentation relativement atypique du taux. Cette tendance ne se confirme ni en 2016 ni en 2017, au contraire le taux de résiliation chute assez significativement.

2.3.3 Résiliation en fonction de la maturité

La FIGURE 2.3 représente la distribution des polices résiliées (en nombre) en fonction de leur maturité (en mois ou en années, depuis la souscription).

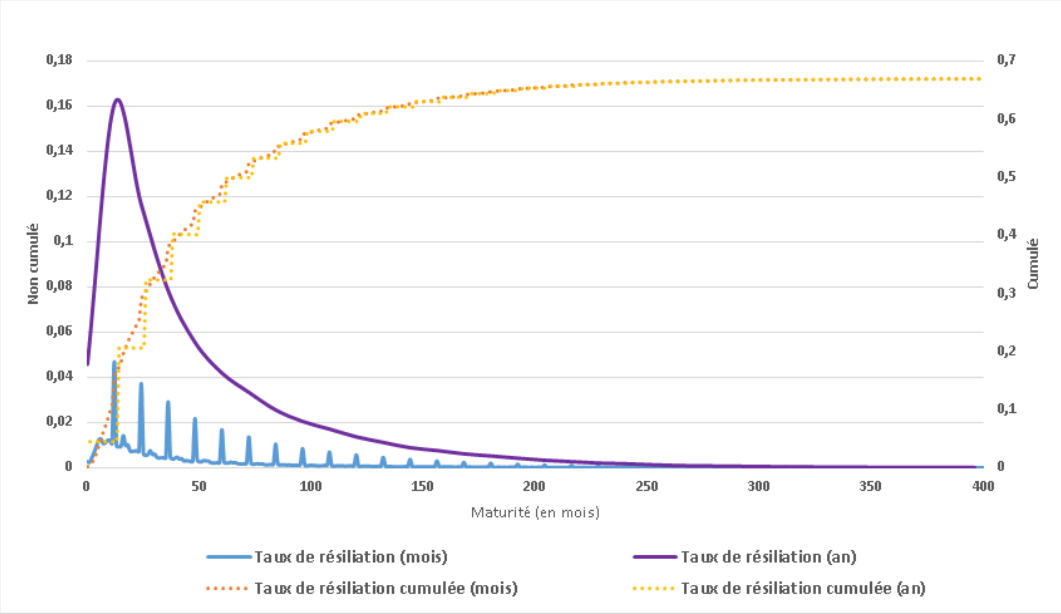


FIGURE 2.3 – Évolution mensuelle et annuelle du taux de résiliation.

Le graphique des résiliations mensuelles (en bleu) met en valeur le poids des résiliations à l'échéance anniversaire du contrat, notamment lors de la première année de souscription. Pour une visualisation plus synthétique, cette même distribution est présentée avec un pas annuel pour l'abscisse (en violet).

Un « pic » de résiliations est observé dès le premier anniversaire du contrat. La distribution décroît ensuite progressivement.
En effet, 20% des contrats sont déjà résiliés à l'horizon annuel, 40% à horizon trois ans. Il faut ensuite attendre 5 ans pour avoir 50% des contrats résiliés, puis 10 ans pour en avoir 60%.

2.3.4 Durée de vie des contrats

Segmentons un peu plus finement le portefeuille afin de présenter les résultats pour cinq variables exogènes :

- La **fréquence de paiement** des primes, qui est déclinée en trois modalités : annuelle, mensuelle ou autre (trimestrielle ou semestrielle).
- Le **nombre de pièces** du logement : quatre modalités (de deux pièces ou moins à sept pièces et plus)
- La **qualité juridique** du client : propriétaire, locataire ou PNO [3].
- Le **type de bien** assuré : maison, appartement ou autre.
- La **zone géographique** du risque assuré : trois zones sont proposées.

Le tableau 2.6 présente les durées de vie moyennes sur les différentes modalités des variables [4]. Nous notons ainsi une assez grande hétérogénéité des résultats, laissant supposer qu'un modèle de durée multivarié serait intéressant à tester.

Les tendances suivantes sont mises en évidence :

3. Propriétaire Non Occupant.
4. Les volumétries associées à ces contrats sont disponibles en annexe page 95 dans le tableau B.1.

- Une durée de vie plus élevée pour les propriétaires, maisons, grands logements, paiements annuels, zone géographique 3.
- Une durée de vie plus faible pour les locataires, appartements, petits logements, paiements mensuels, zone géographique 1.

Ce bref aperçu du portefeuille, et en particulier des polices MRH, va désormais être complété, dans la section suivante, par une analyse détaillée des principales variables disponibles dans le système d'information de la mutuelle.

Echantillon	Variable	Modalité	Global	Résiliés	Censurés
PTF	**RISQUE**	**Habitation**	**4,61**	**4,20**	**5,46**
MRH	Fréquence	Annuel	4,92	4,40	6,29
MRH	Fréquence	Mensuel	3,28	2,92	3,61
MRH	Fréquence	Autre	4,53	3,99	5,32
MRH	Nb pièces	2 pièces ou -	3,11	2,83	4,47
MRH	Nb pièces	3 ou 4 pièces	4,43	4,03	5,48
MRH	Nb pièces	5 ou 6 pièces	5,31	4,92	6,03
MRH	Nb pièces	7 pièces ou +	5,57	5,39	5,79
MRH	Qualité Juridique	Locataire	2,82	2,72	3,42
MRH	Qualité Juridique	PNO	4,78	4,33	5,21
MRH	Qualité Juridique	Propriétaire	6,14	6,06	6,26
MRH	Qualité Juridique	Inconnue	3,57	3,57	2,00
MRH	Type Bien	Appartement	3,17	2,82	4,31
MRH	Type Bien	Maison	5,05	4,42	5,96
MRH	Type Bien	Autre	3,76	3,00	4,20
MRH	Zone Géo	Zone 1	3,67	3,27	4,23
MRH	Zone Géo	Zone 2	4,21	3,81	5,01
MRH	Zone Géo	Zone 3	5,20	4,72	6,33

TABLE 2.6 – Durée de vie moyenne (en années) avant résiliation sur MRH.

2.4 Statistiques univariées des variables explicatives MRH

Nous allons présenter la **démarche d'analyse** des statistiques descriptives des variables enregistrées et disponibles dans le système d'information suite à la souscription d'un contrat [5] multirisques habitation (ou MRH). Précisons qu'elles sont retraitées des données absentes et aberrantes [6]. **Les données sont arrêtées au** 31 **décembre** 2017 et contiennent des contrats dont la date d'effet est comprise entre le 1er janvier 1974 et le 31 décembre 2017.
Rappelons que les contrats sont apportés au Finistère Assurance par des courtiers ou des agents. Un nombre minimal d'informations leur est demandé et ne concerne pour ainsi dire que les risques à assurer. Nous n'avons donc pas beaucoup d'informations concernant les clients, si ce n'est leur nom et prénom (l'âge, la CSP ne sont par exemple pas connus).
Cependant, ces quelques données de souscription sont fiables et, pour la plupart d'entre elles, bien renseignées dans les bases.

2.4.1 Traitements des données manquantes et aberrantes

Le SID [7], à partir duquel sont réalisées les études statistiques, accueille des données issues du SIO [8]. Le passage des données du SIO au SID nécessite généralement de nombreux traitements informatiques pouvant

5. D'autres variables (de comportement) pourront être intégrées par la suite, lorsque nous nous intéresserons à la probabilité de résilier sur un stock de contrats à une date donnée.
6. Les valeurs aberrantes (*outliers*, en anglais) ont été exclues de l'étude car peu nombreuses au regard de la volumétrie en jeu.
7. Système d'Information Décisionnel : bases de données structurées par thème (ou métier).
8. Système d'Information Opérationnel : bases de données brutes issues des progiciels de gestion.

conduire à des erreurs dont l'origine est potentiellement diverse :

- Évolution du SI opérationnel (exemple : nouvelles données à renseigner pour un produit).
- Migration d'un SI (et non reprise de la totalité de l'historique des données).
- Erreurs diverses : format, saisie opérationnelle (exemple : taper des chiffres en plus dans un âge ou un montant, etc.).

Les erreurs opérationnelles peuvent être évitées a priori en limitant les saisies au maximum et en contraignant les gestionnaires à valider les données suspectes (via des fenêtres de type pop-up par exemple), mais ce n'est pas le cas de tous les progiciels de gestion.

Il faut aussi être vigilant à ne pas considérer comme aberrantes toutes les données atypiques car il se peut très bien que la donnée soit correcte (par exemple, un client avec un très haut revenu).

De nombreuses méthodes d'estimation des seuils de détection des *outliers*[9] sont recensées dans la littérature. Par exemple :

- **Méthode à dire d'expert** : le seuil est choisi arbitrairement par les experts métier à partir de leur expérience du portefeuille et de l'analyse des tableaux de bord de pilotage stratégique.
- **Méthode des pourcentages** : sélection d'un seuil tel qu'il représente un certain pourcentage de la somme totale de la variable d'intérêt.
- **Méthode des quantiles** : le seuil est identifié à un certain quantile de la distribution de la variable considérée (typiquement 0,5 %, 1 % ou 2 %).
- **Méthodes TVE** : à partir de la théorie des valeurs extrêmes (fonction d'excès moyen, estimateur de Hill, etc.).

L'objectif de ce travail n'est pas de détailler l'ensemble de ces méthodes. Précisons simplement que le choix de nos seuils est un mélange de la méthode à dire d'expert et de celle des quantiles.

Dans le cas des données réellement manquantes ou aberrantes, plusieurs options peuvent être mises en œuvre :

- **Exclusion des individus** : si la volumétrie est faible.
- **Application d'une méthode d'imputation** (simple, multiple) : si la volumétrie est conséquente.
- **Pas de traitement** : si l'on souhaite mettre en évidence une classe de données manquantes, ou lorsque que les *outliers* sont justifiés.

Notons que la robustesse des modèles peut être mise à l'épreuve en faisant varier la part de valeurs manquantes ou aberrantes sur certaines variables discriminantes.

Ainsi, après traitement global de ces données problématiques[10], la base de données se réduit et passe de 368 216 individus à 366 198. La perte de données (-0.55%) reste donc très mesurée au regard de la volumétrie de départ.

2.4.2 Variables quantitatives

Ce type de variable, aussi appelée **numérique**, est représentée par un nombre (décimal si la variable est continue, ou entier si la variable est discrète) censé décrire une quantité comme par exemple : le montant d'une prime d'assurance, un nombre de pièces ou une surface. Les éléments que l'on souhaite généralement étudier sont les suivants :

9. Ces méthodes sont surtout utiles pour déterminer des seuils d'écrêtement en tarification, ou des seuils de sinistres graves en provisionnement des sinistres tardifs (IBNyR).

10. Nous n'avons effectué des exclusions que sur les *outliers*, les données manquantes ont été mises dans une classe à part.

- La mesure de la tendance centrale via la moyenne et la médiane.
- La mesure de la dispersion via les éléments classiques que sont la variance et l'écart-type [11].
- L'adéquation de la distribution des données à une loi de probabilité connue.

La représentation graphique des données est très importante afin de synthétiser l'ensemble de ces caractéristiques. Les deux principaux outils de visualisation de la dispersion sont l'**histogramme** et la **boîte à moustaches**. Pour des raisons de lisibilité, nous privilégierons la visualisation par histogramme.

Pour ne pas alourdir la lecture de ce mémoire, le détail des variables étudiées est présenté en annexe B, à partir de la page 95.

2.4.3 Variables qualitatives

Ce type de variable, aussi appelée **catégorielle**, est décrit par des **modalités** alphanumériques qui peuvent être soit **nominales** [12] soit **ordinales** [13]. Les variables qualitatives présentées sont toutes nominales.
Le tri à plat d'une variable qualitative est, dans un premier temps, représenté dans un tableau synthétisant, pour chaque modalité, son effectif et sa fréquence. Pour une meilleure visualisation des résultats, les données pourront être présentées sous forme de **diagramme à barres** (verticales ou horizontales) ou de **diagramme circulaire à secteurs**. La visualisation en diagrammes à barres verticales sera privilégiée.

Comme pour les variables quantitatives, le détail des variables qualitatives étudiées est présenté en annexe B, à partir de la page 99.

11. On peut aussi rajouter : l'étendue, les quartiles, la distance interquartiles, etc.
12. Si les modalités ne possèdent pas de structure particulière.
13. Si les modalités sont ordonnées (par exemple : Très content, Neutre, Pas content).

2.4.4 Synthèse des variables

Type	Nom	Synthèse succincte
Quantitatives	Nombre de pièces	La majorité des contrats souscrits concerne des logements avec moins de 8 pièces principales (4 pièces en moyenne).
	Surface du bien	La tarification se fait à la superficie lorsque le client déclare plus de 15 pièces principales (faible proportion du portefeuille).
	Valeur du mobilier garanti	Le montant médian observé est relativement faible (environ 8 000 €) pour une moyenne de l'ordre de 20 000 €.
	Valeur du SMP	Le Sinistre Maximum Possible est estimé en moyenne à environ 163 000 €.
	Montant de la cotisation nette	Le montant observé de cotisation nette annuelle est en moyenne de 206 €(la médiane est à 161 €).
Qualitatives	Département du risque assuré	Les quatre départements bretons (Finistère, Morbihan, Île-et-Vilaine et Côte d'Armor) ainsi que la Loire-Atlantique sont les plus représentés.
	Qualité juridique	Nous observons une majorité de propriétaires (occupant ou pas le logement).
	Type de bien assuré	Plus de la moitié des polices concerne des maisons.
	Type de produit	Le produit phare de la mutuelle est le produit Habitation Globale (>70% des polices du portefeuille).
	Type de formule	Nous observons un nombre très important de formules. Cependant, trois formules concernent plus de la moitié du portefeuille.
	Nature des garanties	La grande majorité des contrats comporte des garanties pour le bâtiment et le mobilier (+70 %).
	Fréquence de paiement	La fréquence de paiement est majoritairement annuelle (environ 80%).
	Réseau	Nous observons un nombre important de valeurs manquantes (données historisées dans le SI à partir de la fin des années 2000). La majorité des contrats provient des courtiers purs.
	Type d'agent	Cette variable peut être vue comme un regroupement métier de la variable Réseau.

Table 2.7 – Synthèse des variables étudiées lors de la souscription.

2.4.5 Perspectives

L'ensemble de ces variables va désormais enrichir notre étude afin de tester la pertinence de modèles explicatifs ou prédictifs du phénomène de résiliation.
Avant d'être intégrées dans un éventuel modèle, les variables **continues** peuvent être découpées (de manière statistique ou pas). Nous verrons dans un prochain chapitre, consacré à la segmentation supervisée, qu'un des avantages de cette discrétisation est d'atténuer l'effet des valeurs les plus extrêmes en les diluant dans une classe plus « moyenne ».
De même, les modalités de chaque variable **qualitative** pourront être, si besoin, regroupées dans une logique d'homogénéisation afin d'en réduire leur nombre.

Dans le chapitre qui va suivre, les quelques variables explicatives sélectionnées seront présentées avec un découpage ou un regroupement effectué manuellement, sur la base de considérations métier.

Chapitre 3

Étude de la durée de vie des contrats MRH

Notre étude peut être ramenée à l'étude du délai [1], avant la survenance de l'événement de résiliation. Pour cela, nous allons construire une table de survie (annuelle) à partir de laquelle il sera possible d'estimer la probabilité qu'une police reste dans le portefeuille durant un an et donc, par complémentarité, estimer le nombre de résiliations sur cet horizon.
Afin d'y parvenir, une partie des outils théoriques développés dans le cadre des modèles de durée sera utilisée. Par abus de langage, nous parlerons de modèles de durée de vie. La durée de vie peut être humaine (médecine, assurance vie) ou pas (fiabilité de systèmes, crédits). Dans notre cas, nous allons nous intéresser à la durée de vie des contrats d'assurance (non-vie) en portefeuille.

Techniquement, trois informations principales sont nécessaires pour observer et tenter de résoudre notre problème de départ :

- Une **date de début d'observation** des polices, qui sera la date d'effet de chacune des polices.
- Une **date de fin d'observation** des polices au 31/12/2017.
- Une **définition de l'événement de résiliation** : voir la fin de la sous-section 1.2, page 22.

Les rappels théoriques, nécessaires à la formalisation du problème, sont introduits en annexe C, à partir de la page 105.

Nous allons tout d'abord chercher à trancher entre l'utilisation d'un modèle à valeurs **continues** ou **discrètes**. En effet, nous avons plutôt dans l'idée d'observer la durée de détention d'un contrat de façon annuelle, et donc à valeurs temporelles discrètes. Nous devons donc vérifier si la théorie « continue » s'applique correctement à nos données.

En parallèle, la distribution de la durée de vie des contrats sera étudiée plus en détail. La notion de durée implique que la loi est supposée prendre des valeurs positives ou nulles. Cette loi peut être caractérisée à l'aide des mesures équivalentes suivantes : fonction et densité de survie, fonction de hasard (cumulée ou pas), espérance de vie résiduelle.

Ensuite, nous étudierons l'opportunité d'ajouter des covariables dans le modèle afin de tenir compte de l'hétérogénéité du portefeuille pour affiner les premiers résultats obtenus.

Enfin, il conviendra de vérifier si ces premiers modèles construits sont stables dans le temps et ont un bon pouvoir prédictif à horizon annuel.

1. Qui peut être considéré comme un temps d'arrêt.

3.1 Modèle en temps continu ou en temps discret ?

Précisons que l'approche « continue » est pressentie afin de traiter notre problème. Comme le pas de temps choisi est annuel, nous pouvons nous demander si une approche par des modèles en temps discret ne serait pas meilleure ou, du moins, mieux adaptée à notre étude. Ces deux approches vont ainsi être comparées afin d'estimer la fonction de survie :

- Un modèle en **temps continu** dans lequel la fonction de survie est estimée à l'aide de l'estimateur de Kaplan-Meier.
- Un modèle logistique dit **Bernoulli** [2] en **temps discret** dans lequel la fonction de survie est estimée indirectement à l'aide de la fonction de hasard.

Dans l'approche en temps continu, nous pouvons parler de **taux** de résiliation instantanée (à l'instant t), tandis que dans celle en temps discret nous parlerons plutôt de **probabilité** de résiliation brute (à la maturité k).

3.1.1 Modèle en temps continu

La FIGURE 3.1 montre l'allure de la fonction de survie, estimée à l'aide de l'estimateur de Kaplan-Meier, sur le portefeuille MRH [3].

Nous supposons que $t_0 = 0 < t_1 < t_2 < ... < t_n$ représentent les différents temps d'événement (ici la maturité annuelle). L'estimateur de la durée moyenne de survie est le suivant [4] :

$$\hat{\mu} = \sum_{i=1}^{n-1} \hat{S}(t_{i-1})(t_i - t_{i-1}) = \sum_{i=1}^{n-1} \hat{S}(t_{i-1}) \approx 7,6 \text{ ans}$$

Où $\hat{S} : [0, +\infty[\to [0, 1]$ est un estimateur de la fonction de survie S.
Notons que la durée de vie médiane est estimée à 4 ans, ce qui est plus cohérent avec les résultats du tableau 2.4, page 26 (4, 61 ans) [5].

La version discrétisée de l'**espérance de vie résiduelle** (voir équation (C.2), page 105) est formalisée ainsi :

$$r_X(t_k) = \frac{1}{S(t_k)} \sum_{i \geq 0} S(t_k + i)$$

En reprenant des notations plus spécifiques à l'assurance-vie, nous pouvons écrire :

$$r_X(k) = \sum_{i \geq 1} {}_ip_k \text{ avec } {}_ip_k = \mathbb{P}[X > i + k | X > k]$$

Il est souvent intéressant d'observer graphiquement cet indicateur par maturité afin de mieux cerner la distribution des résiliations. Ainsi, la FIGURE 3.2 propose une courbe dont l'allure est assez remarquable :

- Elle commence par croître assez rapidement pendant environ cinq ans : cela laisse penser que la compagnie fidélise correctement ses clients les premières années.
- Une phase de stagnation, à un niveau de durée de vie résiduelle de l'ordre de huit à dix ans, intervient ensuite (pendant une quinzaine d'années).
- Enfin, nous observons une décroissance progressive de la durée de vie moyenne restante, s'accentuant avec l'ancienneté. Il faut noter que, passé la maturité de vingt ans, le nombre de polices observées diminue fortement (cf. le tableau C.2, page 111 en annexe).

2. Appliqué individuellement pour chaque contrat.
3. Issu de la procédure LIFETEST de SAS, avec les intervalles de confiance à 95% de Hall-Wellner.
4. Ce résultat provient de la version discrétisée de la formule de l'espérance d'une durée de vie T : $\mathbb{E}[T] = \int_0^{+\infty} td(1 - S(t)) = \int_0^{+\infty} S(t)dt$
5. Ne perdons pas non plus de vue que cet estimateur est biaisé positivement.

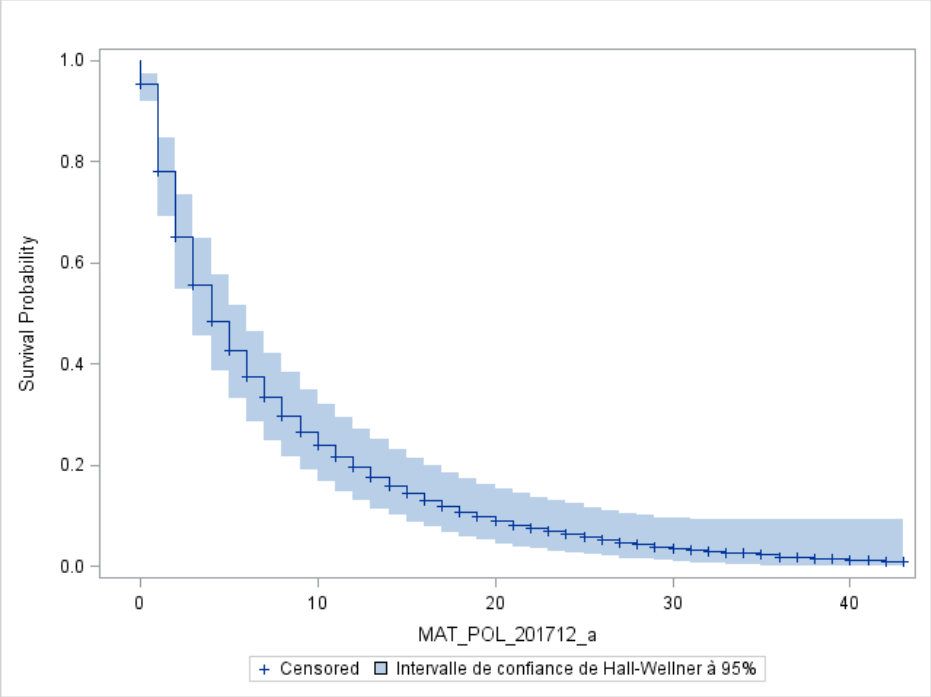


FIGURE 3.1 – Estimateur KM de la survie sur le portefeuille MRH

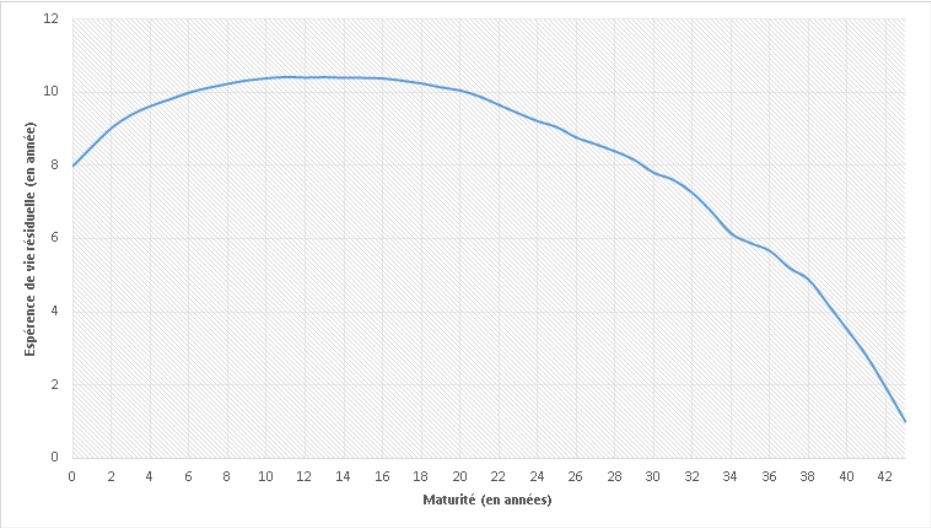


FIGURE 3.2 – Durée de vie résiduelle moyenne sur le portefeuille MRH

3.1.2 Modèle en temps discret

Le lecteur pourra lire avec intérêt les articles de [SW93] ou de [LG03]. Nous nous inspirerons des notations de [PT00].

Considérons que la variable aléatoire T, décrivant la durée de vie des polices, prend des valeurs discrètes et que sa distribution est entièrement décrite par les $p_k = \mathbb{P}[T = k]$, pour $k \geq 0$. Sa fonction de survie s'écrit :

$$S_T(k) = \sum_{m \geq k+1} p_m$$

Dans le cas discret, la fonction de hasard est définie comme telle :

$$h_T(k) = \mathbb{P}[T = k | T > k - 1] = \frac{p_k}{S_T(k-1)}$$

La fonction de hasard s'interprète ainsi comme le taux de résiliation à la maturité k. Comme $1 - h_T(k) = \frac{S_T(k)}{S_T(k-1)}$ nous avons, par récurrence :

$$S_T(k) = \prod_{m=1}^{k} [1 - h_T(m)] \tag{3.1}$$

La fonction de hasard peut donc s'identifier aux q_k, c'est-à-dire aux taux de résiliation à la maturité k estimés par $\hat{q_k} = \frac{d_k}{r_k}$ dans l'approche continue (cf. équation (C.3), page 106).

Il est donc ainsi possible de supposer que le nombre de résiliations d_k à la maturité k suit une loi binomiale $\mathcal{B}(r_k, q_k)$, où r_k est le nombre de polices à risque à la maturité k (r_k tient donc bien compte des données censurées).

Comme la loi binomiale fait partie de la famille exponentielle [6], nous nous plaçons dans le cadre d'un modèle linéaire généralisé. L'objectif est d'estimer q_k pour $k \in \{0, 1, 2, \ldots, k_{max}\}$. Dans cette configuration, le logit du taux de résiliation est supposé être une combinaison linéaire de la maturité (temps discrétisé) :

$$\text{logit}(q_k) = \eta_k = \sum_{m=0}^{k} \alpha_m \mathbb{1}_{[T=m]} = \alpha_k \mathbb{1}_{[T=k]} \tag{3.2}$$

Les q_k sont ainsi mis en évidence :

$$q_k = \frac{1}{1 + \exp(-\eta_k)}$$

La vraisemblance de ce modèle s'écrit :

$$\mathcal{L} = \prod_{k=0}^{k_{max}} C_{r_k}^{d_k} {q_k}^{d_k} (1 - q_k)^{r_k - d_k}$$

On dérive la vraisemblance (ou, par commodité de calcul, la log-vraisemblance) par rapport aux paramètres $\mathrm{q} = (\mathrm{q}_0, \ldots, \mathrm{q}_{k_{max}})$ afin d'obtenir un système d'équations qui se résout numériquement en appliquant un algorithme itératif de descente du gradient de type Newton-Raphson par exemple.

Application d'un modèle en temps discret Nous appliquons le modèle sur le portefeuille MRH. Techniquement, les données doivent être mises sous forme longitudinale afin de pouvoir calculer des probabilités de résiliation année par année.
Les résultats de la modélisation des taux de résiliation par régression logistique sont présentés en annexe dans le tableau C.1, page 110. Le tableau C.2, page 111, montre les volumétries en jeu.

Le passage des coefficients de la régression à la fonction de survie s'effectue en deux temps [7] :

- Nous déduisons les $\hat{q_k}$ des coefficients de la régression car par l'équation (3.2) $\hat{q_k} = \frac{1}{1+\exp(-\hat{\alpha_k})}$.
- Nous approximons la fonction de hasard par les $\hat{q_k}$ et par l'équation (C.1) nous obtenons la fonction de survie.

3.1.3 Comparaison des approches

Nous avons donc obtenu deux estimations de la fonction de survie : l'une par une approche en temps continu et l'autre par une approche en temps discret. La FIGURE 3.3 nous montre la **convergence des deux méthodes**.

Le comparatif des fonctions de survie entre le modèle de type Bernoulli et celui en temps continu, dans la FIGURE 3.3, est interpellant car les fonctions de survie semblent confondues, même en présence d'une censure loin d'être négligeable.

6. Cf. section C, en annexe page 107.
7. Ces deux étapes sont mises en évidence dans le tableau C.1, page 110.

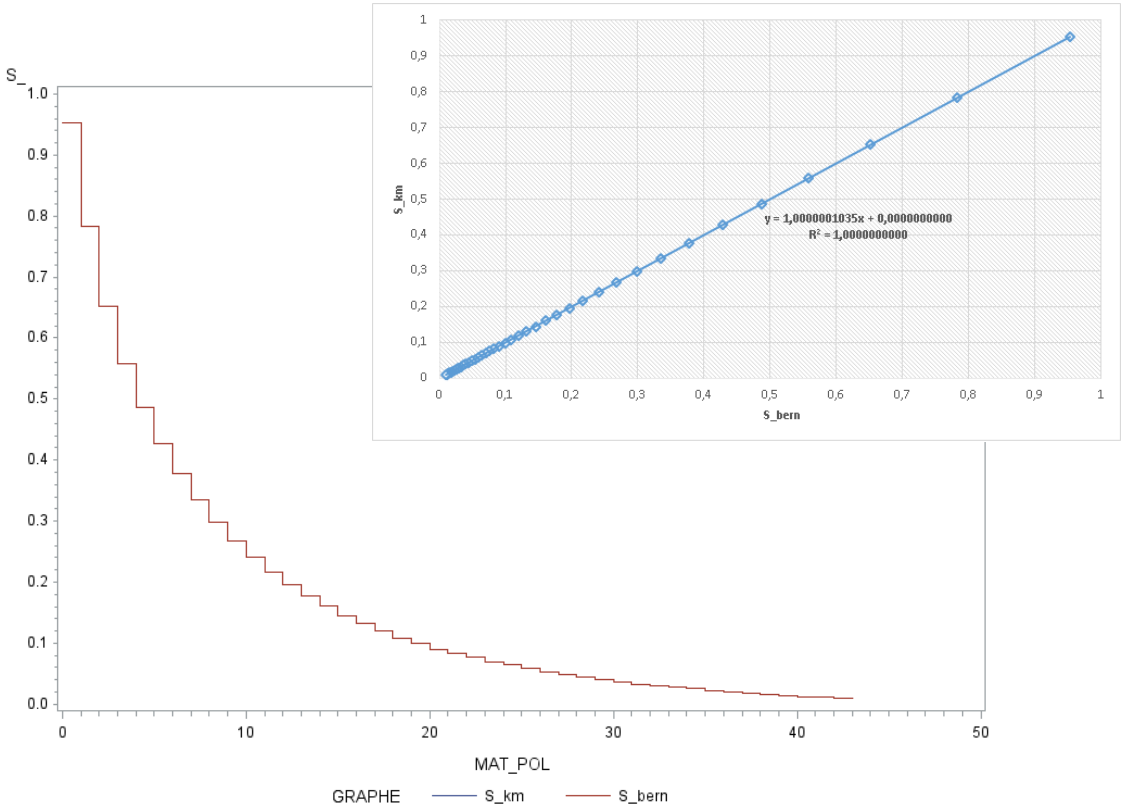


FIGURE 3.3 – Comparaison des fonctions de survie KM et Bernoulli.

En réalité, prouver la convergence des deux méthodes revient à prouver (d'après (C.3) en page 106 et (3.1)) que pour chaque maturité k on a :

$$\frac{1}{1+\exp(-\alpha_k)} \text{ est équivalent à } \frac{d_k}{r_k}$$

Avec α_k les coefficients du modèle Bernoulli, d_k les résiliés et r_k les polices à risque à la maturité k.

Nous avions déjà identifié cela lors de la définition de notre modèle discret. Nous remarquons, de plus, que le rapport $d_k/(r_k - d_k)$ est l'odd-ratio pour le risque de résiliation à la maturité k. Or, on sait que pour une regression binomiale avec lien logit, l'odd-ratio est égal à l'exponentielle de la combinaison linéaire des coefficients par la matrice des explicatives [8].

Remarques

- Des intervalles de confiance à un niveau $1-\alpha = 95\%$ ont été rajoutés afin d'avoir une idée de la précision des estimateurs obtenus par les deux méthodes. Les représentations graphiques et commentaires associés sont présentées en annexe à partir de la page 112.
- Des pas de temps **mensuels** puis **quotidiens** [9] ont aussi été testés. Les résultats ne nous ont pas semblé fondamentalement différents pour remettre en cause le pas **annuel** choisi initialement.

Nous concluons cette partie en considérant que l'application de méthodes destinées initialement à des durées continues est acceptable pour des durées discrétisées en années.

8. Nous verrons plus loin que le modèle logistique s'écrit : $\text{logit}(p) = X\beta \Leftrightarrow p/(1-p) = \exp(X\beta) \Leftrightarrow OR = \exp(X\beta)$ avec $OR = p/(1-p)$.
9. Dans l'approche continue (KM) uniquement.

3.2 Intégration de variables exogènes

Jusqu'à présent, la fonction de survie du portefeuille MRH a été étudiée sur un mode **descriptif**, c'est-à-dire en ne tenant compte que de l'effet du temps (via la maturité) sur elle.
Nous allons désormais étudier ce portefeuille sur un angle plus **explicatif**, c'est-à-dire, en intégrant, dans l'estimation de la fonction de survie, l'effet de variables explicatives afin de mieux décrire l'hétérogénéité observée dans le portefeuille (cf. section 2.3.4, page 28.).
Nous commencerons par présenter quelques pistes afin de justifier la prise en compte de variables exogènes. Dans un second temps, le modèle de Cox (dit à **hasard proportionnel**) sera étudié. Enfin, si les hypothèses d'application du modèle de Cox ne sont pas vérifiées, nous chercherons à décrire la survie par des modèles à **hasard accéléré**.
En conclusion, nous verrons comment essayer de mesurer l'hétérogénéité liée au temps sur les modèles afin de savoir si ces derniers sont suffisamment stables pour être utilisés avec optimisme.

3.2.1 Justification de l'hétérogénéité

Elle peut se faire **graphiquement**, en comparant les fonctions caractérisant la survie pour diverses modalités de variables. C'est ce que proposent les graphiques de la FIGURE 3.4, page 41, pour chacune des modalités des cinq variables, **étudiées séparément** : plus les courbes sont distantes l'une de l'autre, plus la prise en compte de la variable exogène dans un modèle de survie est justifiée. C'est particulièrement évident pour les trois premiers graphiques, un peu moins pour les deux derniers.

Cette incertitude de décision nous incite à objectiver notre intuition à l'aide de **tests statistiques**. Les deux principaux tests proposés par SAS sont celui du **log-rank** (basé sur la statistique de Mantel-Haenzel) et celui de **Gehan** (qui généralise le test de Wilcoxon aux données censurées). Pour plus d'informations, se référer à [PT00].
Ces deux tests, non-paramétriques et adaptés aux données avec censures, sont basés sur l'étude du rang des observations dans les différents échantillons considérés.

La philosophie de ce type de test est souvent la même : sous l'hypothèse [H_0 **: les distributions (de la survie) sont confondues**], la statistique d'intérêt est censée suivre une loi de type χ^2 (avec un niveau de degrés de liberté égal au nombre de paramètres - 1). Si la distance entre la statistique et le seuil de la loi (au niveau $1-\alpha$) est trop grande, alors on considère l'hypothèse H_0 improbable et on la rejette.

Les résultats appliqués à nos données sont synthétisés dans le tableau 3.1. L'hypothèse d'égalité des distributions de survie est presque toujours rejetée. Seul le test de Gehan-Wilcoxon pour la variable « Zone géographique [10] » n'est pas rejeté (le test du log-rank n'est pas loin, non plus, d'accepter H_0).

Nom de la variable	**Test d'égalité sur les niveaux de discrétisation**			
	Test	**khi-2**	**DDL**	**Pr > khi-2**
Qualité juridique	Log-rang	327.8460	2	<.0001
	Wilcoxon	205.3563	2	<.0001
Type de bien	Log-rang	109.8862	2	<.0001
	Wilcoxon	60.8602	2	<.0001
Nombre de pièces	Log-rang	148.1712	3	<.0001
	Wilcoxon	111.0237	3	<.0001
Fréquence de paiement	Log-rang	13.5505	2	0.0011
	Wilcoxon	8.7460	2	0.0126
Zone géographique	Log-rang	6.8879	2	0.0319
	Wilcoxon	4.1801	2	0.1237

TABLE 3.1 – Résultats des tests non paramétriques

10. Zones parmi les départements du Grand Ouest français.

Ainsi nos suppositions initiales, basées sur les observations graphiques, sont quasiment toutes confirmées par les tests. On peut considérer que la variable « Zone géographique » n'apportera pas grand chose à la modélisation de la survie, mais que la prise en compte des quatre autres variables permettra d'affiner l'estimation des taux de résiliation.

3.2.2 Modèles à hasard proportionnel

Cette classe de modèles (PH pour *Proportional Hazard*) se propose de décrire la fonction de hasard h conditionnellement à des variables explicatives X (matrice avec $p+1$ colonnes, indicatrice incluse). Formellement, on pose (pour $t > 0$) :

$$h(t|X;\theta) = h_0(t)\exp({}^tX\theta) \tag{3.3}$$

Où h_0 représente la fonction de hasard de base, ${}^tX$ est la transposée de X
L'idée est donc de partir d'un sous-groupe de référence décrit par h_0 pour déduire la survie des autres sous-groupes de la population.
Nous remarquons que le quotient des fonctions de hasard ne dépend pas du temps t mais uniquement de la différence intrinsèque entre individus. En effet, d'après l'équation (3.3) :

$$\frac{h(t|X_i;\theta)}{h(t|X_j;\theta)} = \frac{h_0(t)\exp({}^tX_i\theta)}{h_0(t)\exp({}^tX_j\theta)} = \exp\left({}^t(X_i - X_j)\theta\right)$$

Supposons que l'individu i diffère de l'individu j pour l'unique covariable binaire $X^{(k)}$ (valant 1 pour i et 0 pour j) :

$$\frac{h(t|X_i;\theta)}{h(t|X_j;\theta)} = \exp(\theta_k)$$

Concrètement, cela signifie que l'individu i a un risque de résilier son contrat $\exp(\theta_k)$ fois celui de l'individu j.

Notons que l'hypothèse de proportionnalité est forte et devra être préalablement vérifiée avant d'appliquer ce type de modèle.
L'hypothèse PH signifie en pratique que l'effet des variables explicatives est multiplicatif par rapport au hasard [11].

La fonction de hasard h_0 peut être connue (par exemple, suivre une loi de Weibull) ou pas.
Dans ce dernier cas, si h_0 n'est pas spécifiée (cas non-paramétrique), nous parlerons de **modèle de Cox**. Ce modèle est dit **semi-paramétrique** car la partie de la formule associée aux variables explicatives est bien paramétrique.
Dans ce cas-là, les paramètres du modèle sont ajustés à l'aide d'une **vraisemblance partielle**, dans laquelle la fonction h_0 est considérée comme un paramètre de nuisance et n'intervient pas directement dans les calculs.
La **vraisemblance de Cox** est ainsi donnée par la formule suivante (voir [PT00] pour plus de détails) :

$$L(\theta) = \prod_{i=1}^{n}\left[\frac{\exp({}^t\theta X_i)}{\sum_{j=1}^{n}\exp({}^t\theta X_j)\mathbb{1}_{T_i \le T_j}}\right]^{d_i} \tag{3.4}$$

Comme pour le modèle logistique décrit précédemment, les coefficients $\hat{\theta}$ qui maximisent la vraisemblance sont calculés numériquement via un algorithme itératif. Les formules présentées dans l'annexe C (page 105) permettent ainsi, à parti de $\hat{h}$, de calculer $\hat{S}$ et $\hat{q}$:

$$\hat{S}(t) = S_0(t)^{\exp({}^tX\hat{\theta})} \text{ et } \hat{q}(t) = 1 - (1 - q_0(t))^{\exp({}^tX\hat{\theta})}$$

La significativité des coefficients $\theta = (\theta_k)_{1\le k\le p}$ peut être testée à l'aide du test de Wald par exemple. Pour que ce test soit valide, il faut que l'**hypothèse de proportionnalité** soit respectée. Cette hypothèse peut notamment être vérifiée :

- **Graphiquement** via les représentations suivantes :

11. I.e. à la fonction de hasard.

  - Courbes LLS (logarithme de l'antilogarithme de la fonction de survie) pour chaque modalité de variable.
  - Résidus de Schœnfeld.

- **Analytiquement** : par un test d'hypothèse sur la tendance temporelle.

Courbes LLS

L'idée est de présenter $\log(-\log(S_t))$ en fonction de la maturité ou de son logarithme (si l'on travaille sur des durées très longues) pour chaque modalité de la variable : **si l'hypothèse de proportionnalité est vérifiée, alors les courbes associées à chacune des covariables doivent être parallèles**. En effet :

$$S(t) \overset{\text{Cox}}{=} S_0(t)^{\exp({}^tX\theta)} \Leftrightarrow \log(-\log(S(t))) = {}^tX\theta + \log(-\log(S_0(t)))$$

Cette méthode est intéressante lorsque nous travaillons sur les variables explicatives qualitatives avec peu de modalités [12]. Les courbes [13] sont présentées dans les graphiques de la FIGURE 3.4 pour les variables du tableau précédent 3.1, page 38. Pour plus de clarté, nous ne croiserons pas les variables entre elles et n'en considérerons qu'une seule à la fois.

Les courbes ne semblent pas vraiment très parallèles entre elles, sauf peut-être pour les variables « Zone géographique », voire « Nombre de pièces ». L'hypothèse PH semble donc en premier lieu compromise.

Résidus de Schœnfeld

Cette seconde représentation graphique se base sur l'observation de ces résidus : **aucune tendance temporelle ne doit être visible**.
Formellement, notons $R_{ik}^{\text{Sch}}(t) = X_{ik} - \bar{X}_k(t)$ le résidu, avec $\bar{X}_k(t) = \frac{\sum_{i=1}^n Y_i(t) X_{ik} \exp({}^tX_i\theta)}{\sum_{i=1}^n Y_i(t) \exp({}^tX_i\theta)}$, où $Y_i(t)$ est un booléen valant 1 si l'individu i est présent à la date t et X_{ik} est sa k-ième covariable.
Les résidus sont présentés, pour chaque variable [14], dans les graphiques de la FIGURE 3.5, page 42.

Nous observons en plus des résidus, sur chaque graphe, une courbe de lissage de la tendance et ses intervalles de confiance associés (en pointillés dans nos exemples, peu visibles car très fins). Idéalement, comme les résidus sont centrés, la courbe lissée doit être proche de la droite $y = 0$ (autrement dit, de l'axe des abscisses).

Cela semble être plus ou moins le cas sur nos graphiques. Cependant, la plupart des résidus ne sont pas vraiment constants dans le temps à l'exception de la variable « Qualité juridique ». Pour les variables « Zone géographique » et « Nombre de pièces », précédemment potentiellement pressenties comme respectant l'hypothèse PH, l'hypothèse de constance n'est plus vérifiée pour les maturités élevées.

Test sur la tendance temporelle

Les deux méthodes graphiques présentées peuvent nous laisser assez hésitant pour valider l'hypothèse PH, ce qui nous amène à présenter succinctement une méthode complémentaire basée sur le test statistique dit *de Grambsch et Therneau*. Sans rentrer dans les détails de ce test, basé sur une statistique supposée suivre une loi du χ^2, la fonction *cox.zph* de R en fournit les diagnostics. Par exemple, pour la variable « Qualité juridique » :

```
> cox.zph(modcox1)
                          rho   chisq        p
QUA_JUR_2LOCATAIRE    -0.0659  1125.0  0.00e+00
QUA_JUR_2PNO          -0.0535   746.6  0.00e+00
QUA_JUR_2PROPRIETAIRE -0.0110    31.5  2.04e-08
GLOBAL                    NA  5087.6  0.00e+00
```

12. Les variables quantitatives éventuelles doivent être préalablement discrétisées.
13. La procédure LIFETEST du logiciel SAS permet, grâce à l'option *LLS*, d'obtenir de telles courbes.
14. La librairie survival du logiciel R fournit la fonction cox.zph qui permet de visualiser ce type de résidus et surtout leur allure dans le temps.

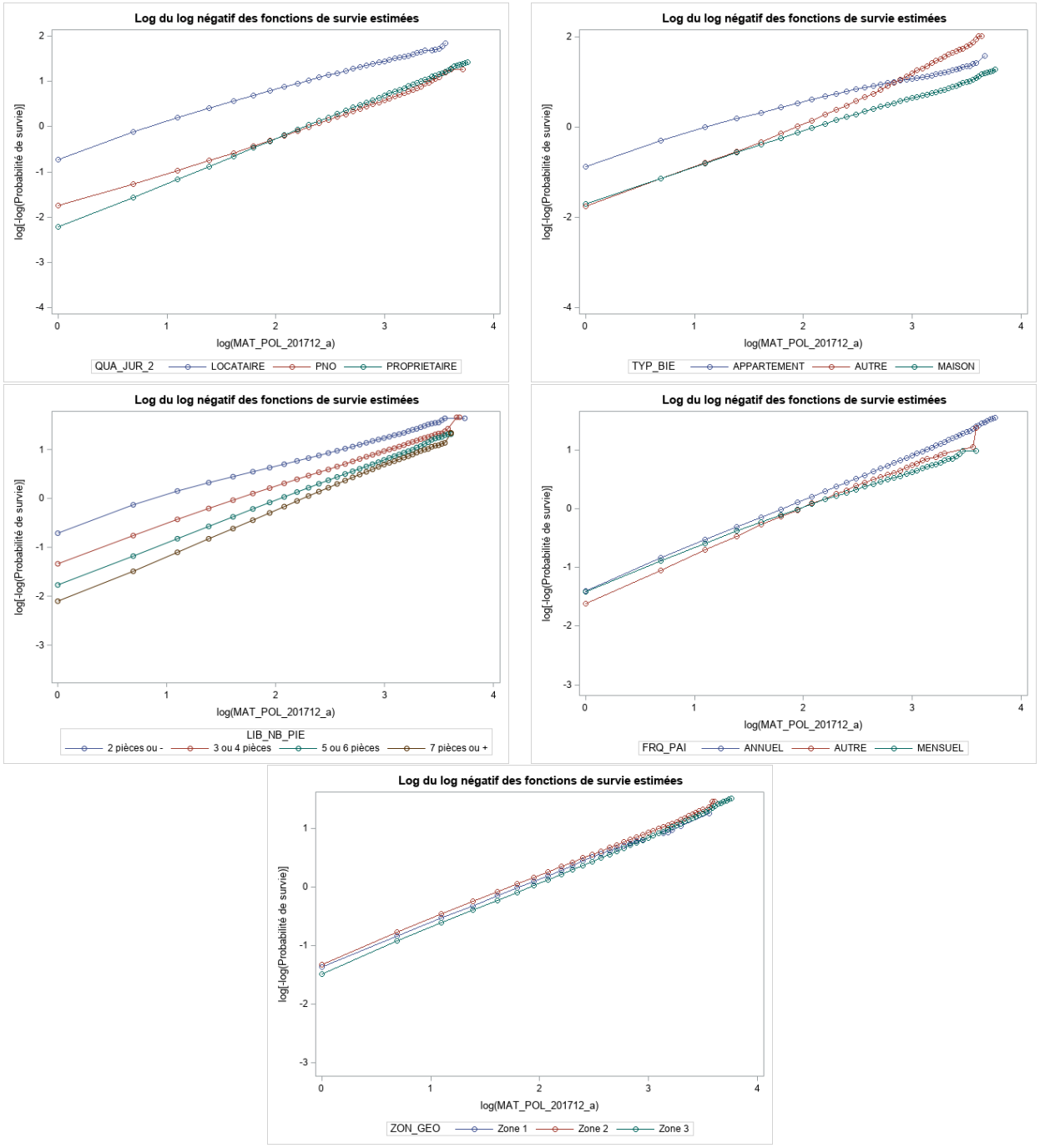


FIGURE 3.4 – Courbes LLS

Bien que les corrélations temporelles (mesurées pour chacune des modalités par le *rho*) ne soient pas élevées et assez proches de zéro en valeurs absolues, le test rejette l'hypothèse H_0 d'absence de tendance.
Notons qu'un test global sur la variable est aussi effectué (lui aussi est logiquement rejeté dans notre exemple).

Les résultats ne sont pas présentés pour toutes les variables. Nous observons que l'hypothèse PH est toujours rejetée au global. Cela nous amènera, à la sous-section suivante, à tester une approche alternative (par des modèles AFT) pour tenter de mieux appréhender l'hétérogénéité des données.

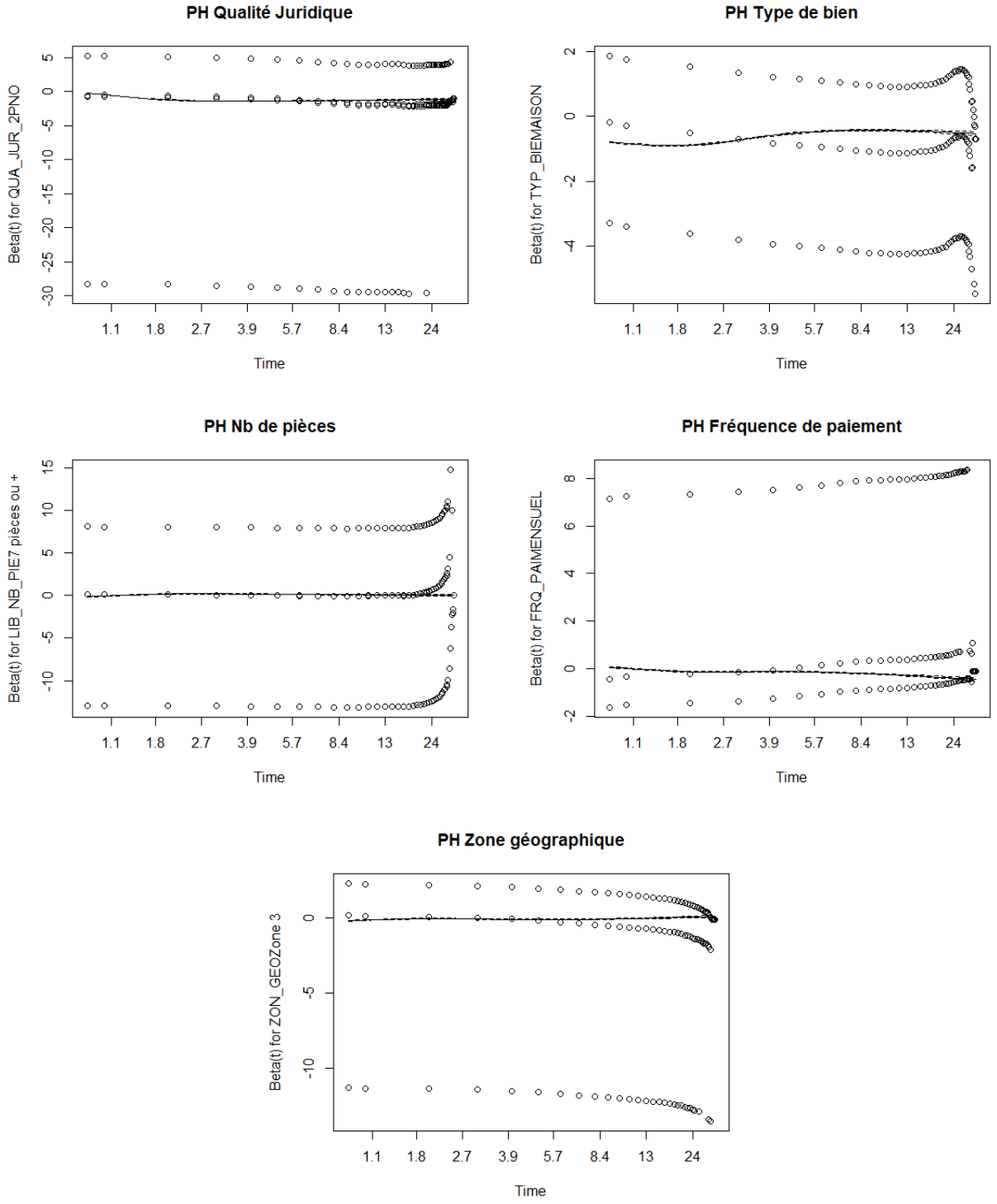


FIGURE 3.5 – Résidus de Schœnfeld

Application du modèle de Cox à nos données

Bien que l'hypothèse PH ne soit pas validée statistiquement, nous avons quand même calculé les probabilités de survie à horizon annuel, à titre indicatif, afin de comparer les résiliations prédites sur le portefeuille MRH à celles prédites précédement par l'approche continue.

Les taux de résiliation prévus [15] (à horizon un an) sont présentés sous la forme d'un tableau synthétique en 3.4 page 49 : une assez bonne disparité des taux prédits pour les trois premières variables (QJ, TB et NP) est observée. La discrimination semble moins bonne pour les deux dernières variables (FP et ZG).

3.2.3 Modèles paramétriques AFT

Nous venons de voir que l'hypothèse de proportionnalité (PH) n'est globalement pas respectée. Nous proposons donc de poursuivre notre étude à l'aide des modèles dits « à hasard accéléré » ou AFT pour *Accelerated Failure Time* [16].
L'hypothèse AFT signifie que l'effet des variables explicatives est multiplicatif par rapport à la durée de survie. Formellement, la fonction de survie d'un tel modèle s'écrit :

$$S(t|X;\theta) = S_0(\phi_i \cdot t) \tag{3.5}$$

où S_0 est la fonction de survie (à spécifier) d'un individu de référence (i.e. avec des valeurs nulles pour toutes les indicatrices associées aux variables explicatives) et $\phi_i = \mathrm{e}^{-(\theta_1 X_1 + \cdots + \theta_p X_p)}$.
Une démonstration de cette égalité est proposée en annexe C, page 112.

Si $\phi_i < 1$ nous interprétons cela comme un **rajeunissement** par rapport à l'individu de référence, dans le cas inverse ($\phi_i > 1$) on parle de **vieillissement**.

Choix d'une distribution

Un tel choix, après estimation, n'est pas si simple. Il faut, d'une part, s'assurer que les modèles testés s'ajustent bien aux données observées, d'autre part comparer les modèles testés entre eux. Ce choix peut se faire à l'aide :

- De tests de spécification, comme par exemple, le test du rapport de vraisemblance.
- En analysant graphiquement les résidus de Cox-Snell.

Seule la seconde option sera présentée, pour ne pas alourdir inutilement l'exposé.

Résidus de Cox-Snell

Nous avons vu précédemment, avec les **résidus de Schœnfeld**, que la notion de **résidu** peut être différente entre un modèle de durée et un modèle multilinéaire classique. Elle diffère dans le sens où il ne s'agit plus de mesurer directement l'écart entre une valeur théorique (observée) et une valeur prédite (par le modèle) mais plutôt de vérifier certaines propriétés propres à cet objet.
Nous parlons alors de **résidu généralisé**, dont l'un des plus connus est le **résidu de Cox-Snell**. Il est défini pour l'individu i par :

$$\begin{aligned} r_i^{CS} &= H[T_i|X_i;\theta] \\ &= -\log[S(T_i|X_i;\theta)] \end{aligned}$$

Si la distribution sélectionnée pour décrire les données est correcte, alors le résidu suit une loi exponentielle de paramètre unitaire $\mathcal{E}(1)$ [ce résultat découle du théorème d'inversion de la fonction de répartition pour les variables continues : si X est de loi F_X alors $F_X(X)$ est de loi $\mathcal{U}[0,1]$ car $\mathbb{P}[F_X(X) \leq t] = \mathbb{P}[X \leq F_X^{-1}(t)] = F_X[F_X^{-1}(t)] = t$].
Ainsi, lors du passage au logarithme du résidu représenté en fonction du temps t, nous sommes censés observer une droite très proche de la première bissectrice.

15. Techniquement, le calcul des paramètres θ décrits dans l'équation (3.4) peut s'effectuer à l'aide de SAS et de sa procédure PHREG. Des tests (Rapport de vraisemblance, Wald et Score) basés sur les propriétés asymptotiques de θ sont proposés par défaut en sortie, afin de s'assurer de la significativité du modèle (sous l'hypothèse $H_0 : \theta = 0$).

16. La procédure LIFEREG de SAS permet d'ajuster de tels modèles. Les distributions suivantes sont prises en charge : Exponentielle, Gamma généralisée, Log-logistique, Log-normale et Weibull. Des modèles AFT peuvent être PH (par exemple, les modèles Exponentiel ou Weibull).

Test graphique d'adéquation des résidus de Cox-Snell

Les graphiques de la FIGURE C.3, en annexe page 114, présentent les résultats [17] obtenus sur la variable « Qualité juridique ». L'idée est d'apprécier la distance entre la droite d'équation $Y = t$ et les résidus représentés par les croix bleues.
Les deux premiers graphiques du haut représentent (respectivement) un ajustement à une loi de Weibull et à une loi Exponentielle. Les ajustements ne sont pas bons pour les maturités élevées, nous ne retiendrons donc pas ces deux lois.
Le dernier graphique représente un ajustement à une loi Gamma. Bien que les points soient assez proches de la première bissectrice, nous ne retiendrons pas non plus cette distribution car les deux autres graphes représentés en troisième et quatrième position semblent bien mieux décrire la propriété attendue.
En effet, les ajustements log-normal et log-logistique (respectivement) suivent assez bien la droite d'équation $Y = t$.

Nous avons ensuite effectué une régression linéaire simple entre les résidus et une droite d'équation $Y = at$. L'idée est de privilégier les ajustements où le paramètre $\hat{a}$, la pente de la droite, est proche de 1. Nous avons également calculé le coefficient de détermination R^2 correspondant. Les résultats sont décrits pour l'ensemble des cinq variables testées dans le tableau 3.2.
Comme nous le pressentions avec la variable « Qualité juridique », ce sont bien les lois log-normale et log-logistique qui s'adaptent le mieux à nos données : les pentes et le R^2 sont généralement très proches de 1, surtout pour les modèles log-logistiques [18].
Attention cependant, car bien que ces indicateurs soient meilleurs avec le modèle log-logistique pour les variables « Fréquence de paiement » et « Zone géographique », nous avons choisi au final le modèle log-normal (cf les graphes FIGURE C.4, en annexe page 115).

Modèle	**Qualité juridique**		**Type de bien**		**Nombre de pièces**		**Fréquence de paiement**		**Zone géographique**	
	$\hat{a}$	R^2	$\hat{a}$	R^2	$\hat{a}$	R^2	$\hat{a}$	R^2	$\hat{a}$	R^2
Weibull	0,67	0,92	0,64	0,93	0,71	0,95	0,80	0,99	0,79	0,99
Exponentiel	0,84	0,96	0,73	0,96	0,80	0,97	0,85	0,99	0,85	0,99
Log-normal	0,96	1,00	0,88	1,00	0,93	1,00	0,95	1,00	0,95	1,00
Log-logistique	1,01	1,00	0,92	1,00	0,97	1,00	0,98	0,99	0,98	0,99
Gamma	1,06	0,99	1,07	0,99	1,12	0,99	1,10	0,94	1,18	0,95

TABLE 3.2 – Résultats des ajustements AFT

Application des modèles AFT à nos données

Comme pour les modèles de Cox, nos résultats sont présentés sous la forme d'un tableau synthétique C.3, en annexe page 117.
Les résultats sont comparables à ceux obtenus par le modèle de Cox : la discrimination est meilleure pour les trois premières variables (QJ, TB et NP) que pour les suivantes (FP et ZG). L'estimation des taux de résiliation à un an est également comparable. Toutefois, au niveau global du portefeuille MRH, l'approche AFT propose des taux de résiliation un peu plus élevés que l'approche PH.

3.3 Stabilité temporelle des modèles

Nous proposons de vérifier que les modèles sont indépendants du temps. La démarche proposée est d'analyser l'échantillon selon quatre dates de vue : à fin 2005, 2009, 2013 et 2017.
Dans chaque cas, nous nous placerons au dernier jour de l'année et reconstituerons ainsi les contrats résiliés et censurés.

17. La procédure LIFEREG du logiciel SAS permet, grâce à l'option CRES, d'obtenir les résidus de Cox-Snell pour la distribution spécifiée. Une seconde étape est ensuite nécessaire afin de linéariser la distribution des résidus (procédure LIFETEST et option graphique LS).
18. Quelques précisions sur les deux distributions choisies sont proposées en annexe C, page 115.

L'idée est de comparer les paramètres obtenus dans les différents modèles afin de s'assurer qu'ils ne varient pas trop dans le temps. Si tel n'est pas le cas, alors une approche prospective pourrait être envisagée.

Un rappel du principe des **tests d'hypothèses** est proposé en annexe C, page 117.

3.3.1 Démarche proposée

Chaque paramètre des modèles testés va être calibré aux quatre dates différentes définies ci-dessus. Chacune de ces estimations va donner lieu à un calcul de dispersion et donc, avec des hypothèses de convergence asymptotique, à des estimations d'intervalles de confiance. Par simplicité, nous choisissons de fixer le seuil de confiance à un niveau $1-\alpha$ avec $\alpha = 5\%$.
Afin de ne pas être trop restrictif, nous allons comparer la valeur des paramètres entre deux dates consécutives [19].

Soit θ_A le paramètre d'intérêt pour l'année A. Les hypothèses à vérifier seront donc $[H_0^{'} : \theta_{2005} = \theta_{2009}]$, $[H_0^{''} : \theta_{2009} = \theta_{2013}]$ et $[H_0^{'''} : \theta_{2013} = \theta_{2017}]$ [20].

Soit $\theta_A^{\inf}$ (resp. $\theta_A^{\sup}$) la borne inférieure (resp. supérieure) de l'intervalle de confiance du paramètre θ_A.

Nous allons considérer que l'hypothèse H_0 n'est pas rejetée si :

$$\left[\theta_A^{\inf}, \theta_A^{\sup}\right] \bigcap \left[\theta_{A-1}^{\inf}, \theta_{A-1}^{\sup}\right] \neq \emptyset \text{ pour } A \in \{2009, 2013, 2017\} \text{ et } A-1 \in \{2005, 2009, 2013\}$$

En d'autres termes, si l'intersection des intervalles de confiance du paramètre θ entre deux périodes consécutives est vide, alors nous rejetons l'hypothèse H_0 d'indépendance temporelle des paramètres.

Remarque Les variances ont tendance à diminuer dans le temps (car la taille de l'échantillon augmente et les données sont assez homogènes). Il n'est donc plus forcément possible de supposer que les écarts-types sont connus et constants.

3.3.2 Modèle PH

La comparaison des paramètres du modèle (cf. section 3.2.2) va se faire en deux temps, pour chaque variable :

- Comparaison des paramètres pour les **fonctions de survie de référence**.
- Comparaison des paramètres associés aux **autres modalités** de la variable.

Pour la **fonction de survie de référence**, l'intervalle de confiance proposé, à un niveau $1-\alpha$, est le suivant [21] :

$$IC_{1-\alpha}[S(t)] = \exp\left[\log(\hat{S}(t) \pm z_{\frac{\alpha}{2}}\hat{\sigma_0}(t)\right] \tag{3.6}$$

où $\hat{\sigma_0}$ est l'écart-type proposé par Kalbfleish-Prentice-Tsiasis (1980) et $z_{\frac{\alpha}{2}}$ le percentile de niveau $1-\frac{\alpha}{2}$ d'une gaussienne standard.

Le FIGURE 3.6 synthétise les résultats obtenus variable par variable et en fonction de la maturité annuelle. Un code couleur permet de voir plus facilement le nombre de tests, parmi les trois effectués [22], où l'hypothèse H_0 n'est pas rejetée : vert si plus de 50% des tests sont bons, orange si 50 % et rouge si moins de 50%. Dans ce dernier cas, nous pourrons considérer que l'hypothèse de stabilité temporelle des paramètres n'est pas vérifiée.

Nous observons que l'hypothèse de stabilité temporelle n'est jamais vérifiée pour la maturité 0. Ce résultat confirme une certaine instabilité de la clientèle à l'issue de la première année où nous notons un pic dans le risque de résiliation (cf. FIGURE 2.3, page 28).

19. Nous aurions, en effet, pu considérer l'ensemble des quatre dates.
20. Contre $[H_1^{'} : \theta_{2005} \neq \theta_{2009}]$, $[H_1^{''} : \theta_{2009} \neq \theta_{2013}]$ et $[H_1^{'''} : \theta_{2013} \neq \theta_{2017}]$ resp.
21. L'option BASELINE de la procédure PHREG de SAS offre en sortie ses bornes inférieures et supérieures grâce aux sous-options LOWER et UPPER.
22. Voire un ou deux pour les maturités élevées.

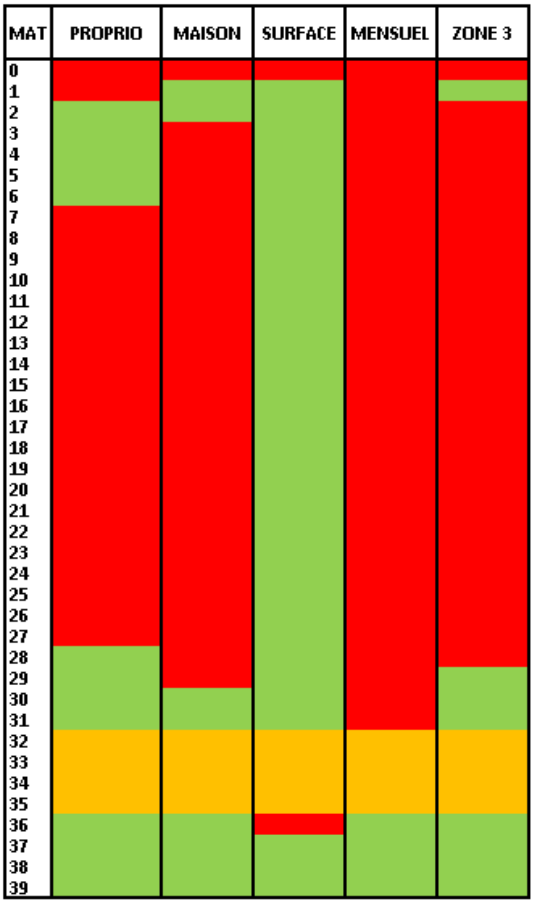


FIGURE 3.6 – Résultats des tests pour les survies de référence

À l'inverse, une relative stabilité temporelle est observée sur les maturités plus élevées (à partir de 32 ans). Mais cette stabilité n'est guère étonnante, car ces classes ont un faible effectif et donc potentiellement une forte variance, ce qui incite le modèle semi-paramétrique à plus de prudence et donc à élargir ses intervalles de confiance.
Au niveau des variables, nous observons que seule la modalité SURFACE [23], relative au nombre de pièces du bien assuré, peut être considérée comme stable. La même raison que précédemment peut être évoquée : le modèle est prudent, car peu de biens sont enregistrés opérationnellement avec seulement la surface comme information [24].
Ainsi, globalement les paramètres relatifs à la population de référence ne semblent pas vraiment stables dans le temps. Nous observons en effet plus de rouge que de vert dans la FIGURE 3.6 (traduisant le fait que plus de tests sont rejetés qu'acceptés).

Pour les **autres paramètres** de la variable, des intervalles de confiance du ratio de risque e^{β_i}, relatif à la modalité i de la variable, peuvent être obtenus [25]. En effet, si $\hat{\beta}$ est le vecteur associé aux paramètres et $\hat{V}(\hat{\beta})$ sa matrice de variance-covariance [26], alors nous avons :

$$IC_{1-\alpha}[e^{\beta_i}] = \exp\left[\hat{\beta}_i \pm z_{\frac{\alpha}{2}}\sqrt{\hat{V}_{ii}(\hat{\beta})}\right] \tag{3.7}$$

Les résultats obtenus sont détaillés dans le tableau C.4, en annexe page 119. La dernière colonne, nommée

23. Cette variable est un croisement des variables **nb de pièces** et **surface** : vaut SURFACE si la surface est renseignée ; nombre de pièces sinon.
24. Nous pourrions sans doute modifier les résultats en forçant les modèles PH à prendre comme population de référence celle avec la plus faible volumétrie.
25. Via l'option MODEL de la procédure PHREG et sa sous-option RISKLIMITS.
26. Estimée par la méthode de Newton-Raphson.

« Test », permet de visualiser le résultat du test d'hypothèse entre l'année en cours et l'année précédente [27]. Nous observons ainsi une bonne stabilité temporelle :

- **Fréquence de paiement** : seul un test sur les six ne rejette pas H_0.
- **Nombre de pièces** : aucun des douze tests ne rejette H_0.
- **Qualité juridique** : deux tests sur six sont rejetés.
- **Type de bien** : deux tests sur six sont rejetés.
- **Zone géographique** : deux tests sur six sont rejetés.

A l'exception de la variable **Fréquence de paiement**, **les résultats sont plutôt favorables à l'hypothèse de stabilité temporelle**.
L'instabilité sur cette variable montre une dynamique à l'œuvre depuis quelques années sur le portefeuille : les clients ont de plus en plus tendance à mensualiser le paiement de leur cotisation d'assurance (au détriment de l'annualisation).
Nous notons également que le test sur la variable **Zone géographique** est rejeté pour la comparaison des années 2013/2017, ce qui s'explique par le fort développement commercial du portefeuille sur de nouveaux départements à partir de 2012.

3.3.3 Modèle AFT

Le tableau C.5, en annexe page 120, présente les résultats obtenus pour les paramètres propres aux cinq variables explicatives :

- **Fréquence de paiement** : deux tests sur les six ne rejettent pas H_0.
- **Nombre de pièces** : aucun des douze tests ne rejette H_0.
- **Qualité juridique** : trois tests sur six sont rejetés.
- **Type de bien** : trois tests sur six sont rejetés.
- **Zone géographique** : deux tests sur six sont rejetés.

Les résultats sont assez contrastés mais plaident, globalement, pour une relative stabilité temporelle des paramètres. Nous constatons qu'ils sont aussi assez proches de ceux obtenus pour le modèle PH.

3.4 Synthèse des résultats

Estimation de la survie moyenne à un an Pour une meilleure lisibilité, reprenons quelques notations actuarielles propres aux tables de mortalité.

Soit $p_k = {}_1p_k = \mathbb{P}[X > k+1 | X > k]$ la probabilité d'être en vie à la maturité $k+1$ sachant qu'on est en vie à k. Les p_k sont déduits directement de la fonction de survie estimée par Kaplan-Meier :

$$p_k = \frac{S(k+1)}{S(k)}$$

Nous en déduisons $q_k = 1 - p_k$ qui est la probabilité de résilier dans l'année sachant que le contrat n'était pas résilié à la maturité k.

27. Vaut 1 si l'hypothèse H_0 est acceptée, 0 sinon.

L'idée principale est d'appliquer les taux de résiliation annuels q_k à l'ensemble des polices en stock, c'est à dire les polices censurées à la date d'observation. Appliquons donc ces taux par maturité aux nombres de polices censurées afin d'obtenir une estimation du nombre de polices résiliées dans l'année :

$$\text{Nb de polices résiliées dans l'année} = \sum_{i \geq 0} \text{Nb Censurées}_i \times q_i$$

Nous pouvons donc, en pondérant par le nombre de polices en stock au 31/12, en déduire un taux de résiliation (en nb). Pour l'approche continue par Kaplan-Meier, nous avons :

$$\text{Taux de résiliation annuel} = \frac{\sum_{i \geq 0} \text{Nb Censurées}_i \times q_i}{\sum_{i \geq 0} \text{Nb Censurées}_i} \approx \frac{16\,311}{124\,230} \approx 13,13\%$$

Ce premier taux semble cohérent avec ceux observés sur la FIGURE 2.2, page 27. Cependant, nous verrons, à la fin de ce chapitre, le faible pouvoir prédictif de cette méthode d'estimation (avec nos données, du moins).

Comparaison des taux à un an Nous allons enfin effectuer une première synthèse des résultats obtenus par les diverses approches présentées en amont (KM [28], DT [29], PH et AFT) en les comparant avec les taux de résiliation observés à horizon annuel.

Le tableau 3.3 présente les résultats obtenus **avant** la prise en compte de variables explicatives dans la durée de survie.
Nous observons ainsi une surestimation des taux réels observés sur l'année 2018 : la surestimation est forte pour les deux approches testées (KM et Bernoulli).
Notons que le taux observé n'est pas non plus inclus dans des intervalles de confiance à 95% car ces derniers sont trop fins.

Type de modèle	**Nb de polices MRH au 31/12/2017**	**Nb résiliations observées en 2018**	**Taux de résiliation observé**	**Taux de résiliation prédit**
Modèle continu KM	124 230	11 825	9,52%	13,13%
Modèle discret BERN	124 230	11 825	9,52%	13,13%

TABLE 3.3 – Synthèse des résultats avant prise en compte de l'hétérogénéité

Le tableau 3.4 page 49 présente les résultats obtenus **après** la prise en compte de variables explicatives dans la durée de survie.
La variable **Nombre de pièces** semble être la plus prédictive pour l'estimation du risque de résiliation à un an : le taux prédit au global est de 11,04% pour le modèle de Cox et de 10,15% pour le modèle AFT log-logistique. Ces taux sont assez proches de la valeur réellement observée a posteriori (9,52%). La prédictivité au niveau des modalités de la variable est également plutôt correcte.
Les résultats obtenus à l'aide des autres variables ne sont pas beaucoup meilleurs que ceux obtenus en amont (sans prise en compte de l'hétérogénéité).

L'étude de nos résultats montre qu'il est assez délicat de démontrer statistiquement la stabilité temporelle des paramètres des différents modèles testés. Le point crucial à retenir est de ne pas faire preuve de naïveté excessive en supposant, à tort, que la stabilité temporelle est vérifiée (cf. FIGURE 3.7).
Cette instabilité impacte négativement la propriété de prédictivité des modèles testés. En effet, la structure du portefeuille évolue dans le temps : la contribution des anciens contrats au modèle semble donc trop importante.
Il aurait été intéressant de *backtester* les taux obtenus en 2018 sur une dizaine d'exercices antérieurs. Cela aurait permis de suivre, par modèle, l'écart entre le taux prédit et le taux réellement observé.

28. Continu : Kaplan Meier.
29. Discret Time : Bernoulli.

Variable	Modalité	Nb de polices MRH au 31/12/2017	Taux de résiliation observé	Taux de résiliation prédit COX	Taux de résiliation prédit AFT
Qualité jur.	**Locataire**	18 060	22%	24%	22%
	PNO	45 461	8%	10%	10%
	Propriétaire	60 709	7%	9%	9%
	TOTAL	124 230	9,52%	11,31%	11,52%
Type de bien	**Appartement**	27 034	13%	20%	21%
	Maison	87 621	8%	10%	11%
	Autre	9 575	9%	13%	9%
	TOTAL	124 230	9,52%	12,43%	13,03%
Nb de pièces	**2 pièces ou -**	13 607	18%	21%	21%
	3 ou 4 pièces	29 272	10%	13%	13%
	5 ou 6 pièces	34 300	8%	10%	7%
	7 pièces ou +	25 709	7%	9%	7%
	Surface	21 342	7%	7%	9%
	TOTAL	124 230	9,52%	11,04%	10,15%
Frq paiement	**Annuel**	84 467	9%	12%	13%
	Mensuel	37 369	11%	12%	13%
	Autre	2 394	9%	10%	10%
	TOTAL	124 230	9,52%	12,19%	12,79%
Zone géo.	**Zone 1**	16 589	10%	13%	14%
	Zone 2	55 113	10%	13%	14%
	Zone 3	52 528	9%	12%	12%
	TOTAL	124 230	9,52%	12,60%	13,06%

TABLE 3.4 – Synthèse des résultats après prise en compte de l'hétérogénéité

Cependant, sur une durée prospective d'un an et en supposant une certaine stabilité du portefeuille, avec un modèle ajusté récemment [30], nous considérons que nous pouvons raisonnablement faire confiance aux résultats attendus avec la variable « nombre de pièces ».

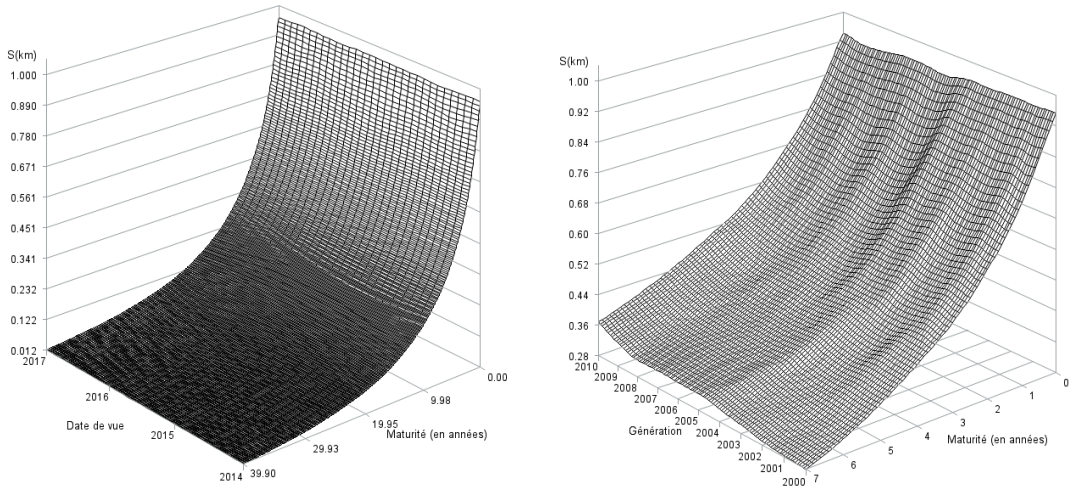


FIGURE 3.7 – Evolution de la survie en fonction de la maturité.

Rappelons que seule une variable a été introduite dans les modèles présentés afin de respecter le **principe de parcimonie**. En complément de l'étude des résiliations par des modèles de durée, il semble ainsi néces-

30. I.e. dans les 12 mois qui précèdent.

saire de comparer la pertinence des résultats issus de méthodes de segmentation supervisées. Ces méthodes permettent d'interpréter des modèles multivariés plus facilement, en tenant compte des corrélations entre les différentes variables [31]. C'est précisément l'objet des trois derniers chapitres.

31. Par exemple, le nombre de pièces et la qualité juridique du client.

Chapitre 4

Segmentation supervisée des profils de résiliation MRH à la souscription

Le lecteur intéressé par les méthodes de segmentation sur variables qualitatives pourra se référer avec intérêt à l'ouvrage [Tuf17] pour un exposé plus complet. L'auteur y présente de manière très pédagogique un catalogue de méthodes statistiques diverses et variées en termes d'applications techniques sans rentrer dans des détails trop théoriques.

Nous avons vu, en introduction, qu'une des motivations du mémoire est la nécessité pour la mutuelle de faire évoluer sa politique tarifaire (soit au travers d'une meilleure connaissance des risques assurés, soit au travers d'une optimisation du commissionnement des apporteurs d'affaires). La construction d'un outil estimant la probabilité a priori d'une résiliation pourrait donc s'avérer intéressante.

Dans un premier temps, l'idée serait de partir d'une estimation de la probabilité de résilier dès la souscription d'un contrat. Une réflexion s'impose sur l'horizon d'observation du risque de résiliation, sachant que :

- Trop court, nous risquons de ne pas avoir assez d'observations du risque (de résiliation).
- Trop long, les données d'entrées (souscription) peuvent ne plus être à jour ou avoir trop évolué (déformation du portefeuille).

Nous proposons donc d'engager la discussion sur les deux horizons suivants :

- **Horizon 36 mois** : car à partir de ce seuil nous constatons une progression relative des résiliations bien moindre (voir la FIGURE 2.3, page 28).
- **Horizon 12 mois** : tout simplement car c'est la fréquence de la supervision Solvabilité 2.

Dans un second temps (cf. chapitre 5), l'idée serait non plus d'estimer la probabilité de résilier à la souscription, mais la probabilité de résilier (fixée à horizon 1 an) sur un stock de contrats en cours à la date d'inventaire.

Nous parlerons de discrimination ou de segmentation supervisée si la variable $Y \in \mathcal{D}$, que l'on souhaite expliquer à l'aide de p variables $X = (X_1, \ldots, X_p)$, est **qualitative**.
Un n-échantillon de même loi que (X, Y) sera donc considéré, à partir duquel on va chercher à prédire au mieux la sortie y à partir des observations x en entrée. L'objectif est donc de trouver une fonction mesurable $f : \mathbb{R}^p \mapsto \mathcal{D} = \{0, 1\}$ qui lie fonctionnellement la sortie $y = f(x)$ à l'entrée x.

4.1 Définition d'une variable cible

Nous cherchons à mesurer le **risque de résiliation**, que ce soit lors de la souscription d'un contrat d'assurance ou à n'importe quel moment de la vie du contrat. **Dans ce chapitre, nous nous placerons dans le cadre d'un score de résiliation à la souscription**.

Il faut préalablement définir une durée maximale sur laquelle le risque sera observé. Soit M cet horizon.

La valeur optimale de l'horizon d'observation du risque étudié n'est pas si simple : un horizon trop faible ne permet pas de récolter suffisamment de données positives pour l'observation du risque, tandis qu'un horizon trop élevé induit un risque d'instabilité temporelle de la structure du portefeuille et donc un potentiel problème de robustesse du modèle proposé.

Soit un ensemble de contrats (affaires nouvelles) sélectionnés. La détermination de l'horizon du risque nous amène à distinguer deux cas :

1. **Risque à horizon fixe** : les contrats sont observés pendant une durée fixe. À l'issue de cette durée, nous regardons si le contrat est résilié ou non.
2. **Risque à horizon variable** : les contrats sont observés jusqu'à une date T fixée. À T, nous regardons si le contrat est résilié ou non.

Nous nous plaçons dans le premier cas. Formellement, la définition de Y^M, la variable cible à M mois suivant la souscription, est la suivante :

$$Y^M = \mathbb{1}_{\text{Date Résiliation} \leq \text{Date Effet+M mois}}$$

Nous supposerons par la suite que $M \in \{12, 36\}$ et que Date Résiliation $\geq$ Date Effet.

Le problème se résume ainsi à estimer l'espérance $\mathbb{E}[Y^M|X]$ où X est un ensemble de covariables censées avoir été judicieusement choisies. Cette quantité est ainsi la probabilité de résiliation à la souscription sur un horizon de M mois.

4.2 Statistiques bivariées des variables explicatives

La mesure de risque à partir de laquelle la discrétisation des variables va être effectuée sera le taux de résiliation à horizon 36 mois.

Nous avons choisi, afin d'avoir des résultats plus pertinents et représentatifs de nos données actuelles, d'effectuer nos découpages et regroupements sur des polices dont la date d'effet est comprise entre le 1er **janvier** 2010 **et le** 31 **décembre** 2014**.**

4.2.1 Découpage des variables quantitatives

Comme les variables explicatives continues ne sont pas forcément linéaires par rapport à la variable à expliquer, elles ont été discrétisées. Cette étape consiste à découper la variable en plusieurs tranches (par exemple en déciles [1]) puis à les regrouper selon celles ayant la même mesure de risque [2] c'est-à-dire en fonction des tranches dont l'agrégation ne change pas statistiquement la mesure. Attention, comme il s'agit de variables continues, les regroupements se feront de manière **ordonnée** (de proche en proche donc).
Deux philosophies de discrétisation peuvent être envisagées : **manuelle** ou **automatisée**. Attardons-nous sur la seconde option.

Plusieurs méthodes automatiques de regroupement des classes de variables ordinales peuvent être proposées. Comme il est inutile de les décrire toutes [3], seule la méthode, décrite dans [NC03], dite des **indicatrices emboîtées** sera présentée et appliquée.

Son principe est le suivant : une variable ordinale X, avec K modalités, est transformée en $K-1$ indicatrices non disjointes telles que, $\forall k \in \{1, \ldots, K-1\}$:

$$\begin{cases} D_k = 0 \text{ si } X \in]1, k] \\ D_k = 1 \text{ si } X \in]k, K] \end{cases}$$

1. À ajuster en fonction des effectifs disponibles.
2. Dans notre étude, la mesure de risque est le taux de résiliation à 36 mois.
3. Nous pouvons raisonnablement supposer que les résultats seraient proches.

Ce codage respecte ainsi la structure d'ordre de la variable X et permet de proposer des regroupements de modalités en interprétant simplement les tests de Wald issus de la régression logistique associée :

$$\text{logit}\left(\mathbb{P}\left[Y^{36} = 1 | D_1, \ldots, D_{K-1}\right]\right) = \sum_{k=0}^{K-1} \beta_k D_k$$

Avec $D_0 = 1$ par convention.

Si le coefficient β_k est retenu au niveau de confiance $1-\alpha$, alors nous considérons que la plage sous-jacente est pertinente pour le découpage. Comme les coefficients correspondent à des plages incluses les unes dans les autres, alors le découpage de la variable correspondra aux bornes des plages sélectionnées. Nous allons illustrer cette technique sur la variable « **Nombre de pièces** ».

Concrètement, les seuils de découpage obtenus pour cette variable (et les autres) sont synthétisés dans le tableau D.1, en annexe page 123.

Le graphique de gauche en FIGURE 4.1 nous présente le taux de résiliation à 36 mois en fonction de centiles de la variable (un premier découpage en huit classes de même effectif est ici proposé). La constante en rouge pointillé représente la valeur moyenne du taux de résiliation observé sur l'échantillon [4].

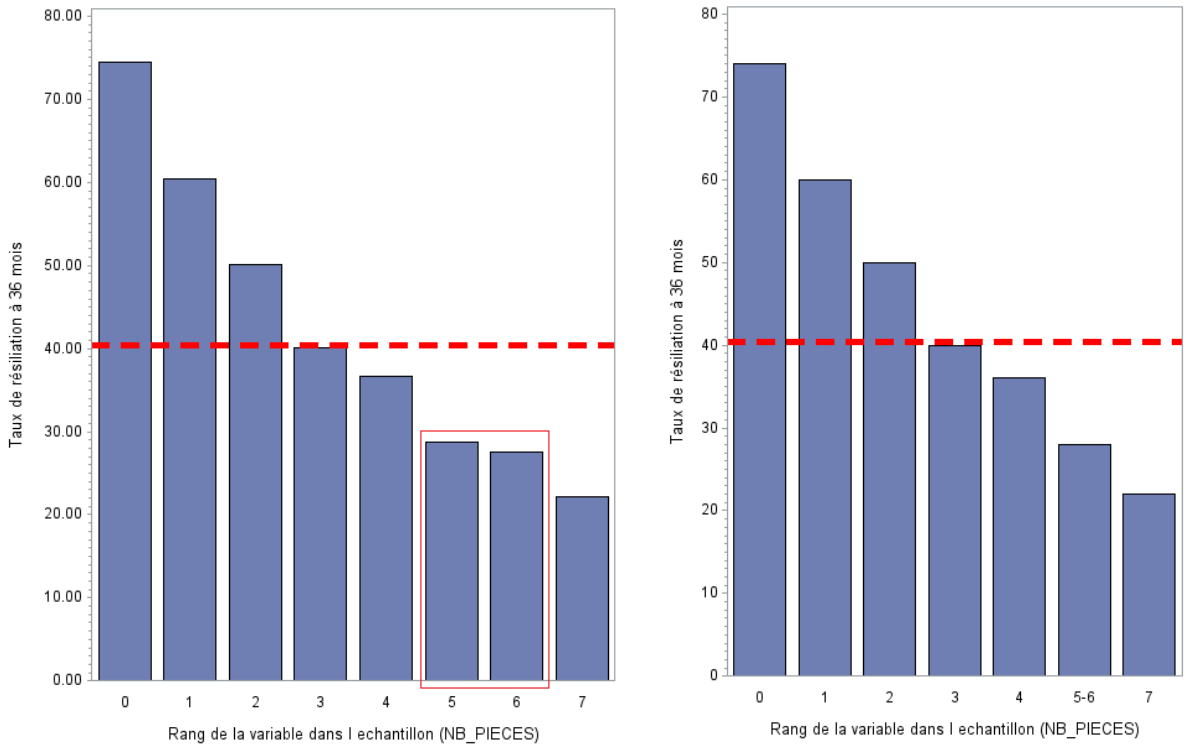


FIGURE 4.1 – Découpage de la variable Nombre de pièces

Nous observons une fréquence de résiliation plus élevée que la moyenne pour les logements dont le nombre de pièces est compris dans les trois premières modalités (moins de trois pièces).
La question à se poser désormais est la suivante : quel est le découpage le plus pertinent à effectuer pour mettre en évidence les plages de valeurs avec des taux de résiliation significativement différents ? Notons que, visuellement, les deux modalités 5 et 6 du diagramme semblent avoir une taille comparable.
La méthode présentée en début de paragraphe et appliquée à nos données nous donne le graphique de droite. Nous observons ainsi que les modalités 5 et 6 ne sont pas statistiquement différentes et sont donc ainsi

4. Nous avons exclu les données avec un nombre de pièces nul. La donnée est manquante et dans ce cas-là, la surface est renseignée.

regroupées (6 ou 7 pièces). Le tableau 4.1 nous donne les résultats de la régression logistique effectuée.

Analyse des valeurs estimées du maximum de vraisemblance					
Paramètre	**DDL**	**Estimation**	**Erreur type**	χ^2 **de Wald**	**Pr** $> \chi^2$
Intercept	1	-1,3275	0,0236	3166,22	<0,0001
D1	1	-0,1389	0,0282	24,34	<0,0001
D2	1	0,4032	0,0287	197,03	<0,0001
D3	1	0,4046	0,0262	238,11	<0,0001
D4	1	0,1489	0,0263	31,96	<0,0001
D5	1	0,3501	0,0283	153,39	<0,0001
D6	1	0,0817	0,0366	4,98	0,0256
D7	1	0,2751	0,0389	50,00	<0,0001

TABLE 4.1 – Résultat de la régression pour le découpage de la variable Nombre de pièces

Pour la classe 6, la valeur du χ^2 de Wald est en effet inférieure au seuil de significativité à 99% pour un χ^2 à un seul degré de liberté [5].
Au final, les résultats observés sur cette variable sont cohérents avec l'expérience que nous avons du portefeuille : le taux de résiliation diminue lorsque le nombre de pièces augmente. Nous pouvons, de plus, supposer que cette variable est corrélée au type de logement.

4.2.2 Regroupement des variables qualitatives

Comme pour les variables quantitatives, nous pouvons envisager deux approches : manuelle ou automatique. Notons que les modalités rares, pour lesquelles les effectifs sont faibles, doivent être regroupées afin d'éviter de leur donner un poids trop important dans la modélisation.

Comme l'objectif est de regrouper les classes dont le taux de résiliation est homogène, nous pouvons par exemple effectuer une classification, non supervisée, de type hiérarchique ascendante (ou CAH) sur chacune des classes de la variable. L'idée est la suivante : nous supposons l'existence d'une mesure de dissimilarité [6] entre les individus permettant, partant de la classification la plus fine, de regrouper entre eux les individus considérés comme proches [7].

La méthode de Ward permet, à l'aide d'une distance préalablement définie qui maximise l'inertie interclasse, de regrouper de manière optimisée (dans le sens de l'homogéneité) les classes (voir [NC05] ou [HLP09] pour plus de détails sur le sujet).
Mais nous devons préalablement définir le nombre de classes souhaité. Pour cela, différents indicateurs, permettant de mesurer le gain d'inertie en fonction du nombre de classes, existent. Nous présenterons le critère du R^2 dans notre exemple. L'objectif est de trouver le meilleur compromis entre une inertie forte et un nombre de classes limité.

Formellement, nous cherchons à optimiser la partition décrite selon le théorème de Huygens par :

$$\text{Inertie totale} = \text{Inertie inter-classse} + \text{Inertie intra-classe}$$

Ainsi le critère d'arrêt est appliqué lorsque l'ajout d'une nouvelle classe n'apporte pas un gain significatif à l'explication de l'inertie.

Nous allons illustrer la méthode pour la variable **Département du risque**. Nous observons, dans le graphique de gauche en FIGURE 4.2, pour chaque modalité de cette variable, le taux de résiliation à horizon 36 mois. La ligne pointillée rouge représente le taux de résiliation moyen sur l'ensemble de l'échantillon.
Un arbre issu d'une CAH est ensuite proposé sur le graphique de droite [8] et montre un partitionnement des

5. Ce seuil est égal à $6,63$.
6. Comme par exemple la distance euclidienne.
7. D'où le qualificatif de méthode « ascendante » : on part de la partition la plus fine et on agrège.
8. Avec les procédures **cluster** et **tree** de SAS.

modalités en fonction du critère du R^2. Plus cette distance est grande et plus le regroupement est justifié.

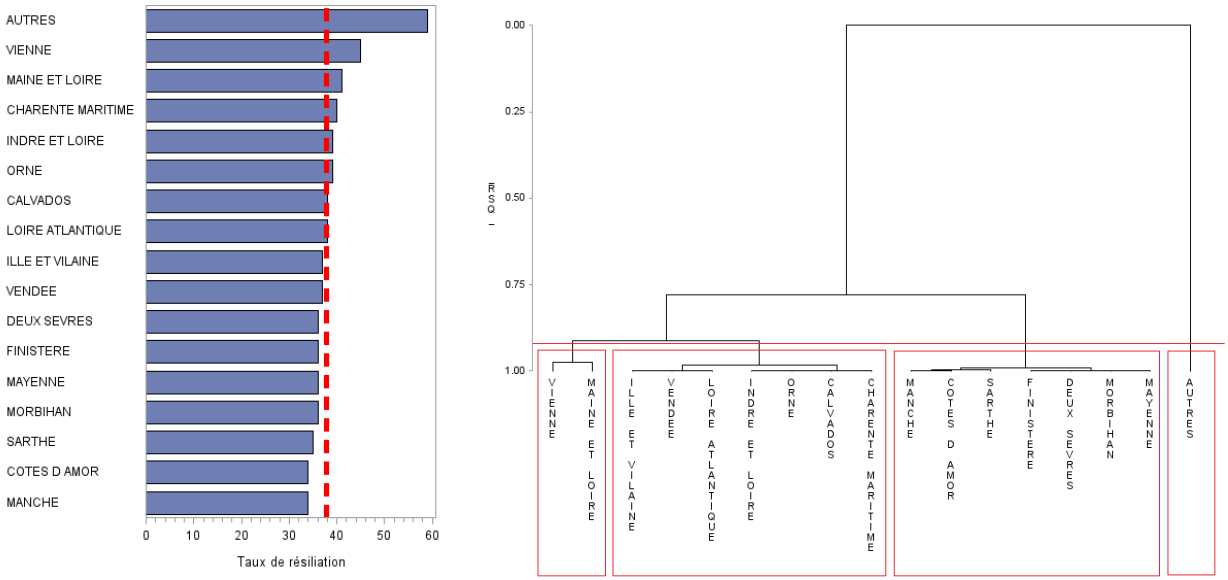


FIGURE 4.2 – Variable Département du risque - Arbre

Le dernier graphique en FIGURE 4.3 représente l'évolution du critère du R^2 en fonction du nombre de classes.

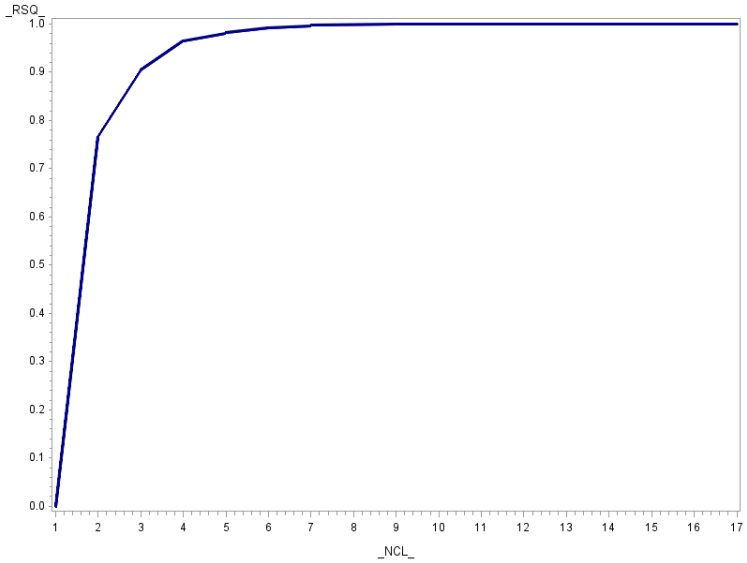


FIGURE 4.3 – Variable Département du risque - R^2

À partir de ces résultats, nous pouvons considérer que le nombre de classes optimales est de quatre (car l'ajout de classes supplémentaires n'apporte pas grand chose à la distinction du taux de résiliation entre les classes). Ceci est graphiquement mis en évidence par la ligne rouge du graphique en haut à droite qui coupe l'arbre en quatre branches. Enfin, la FIGURE 4.4, synthétise l'information finale de la variable avec le regroupement de ses modalités.

Comme pour les variables quantitatives, **le détail des regroupements est visible en annexe, page 124**. Nous présentons également en annexe, page 125, les variables ne nécessitant pas d'être regroupées statistiquement.

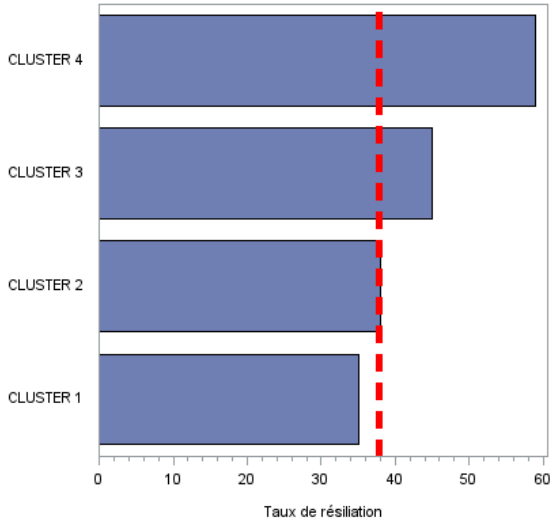


FIGURE 4.4 – Regroupements de la variable Département du risque

4.2.3 Croisements de variables

Croiser des variables entre elles peut sensiblement améliorer la prédictivité des modèles. Comme pour le découpage des variables, les approches manuelle ou automatique peuvent être mises en œuvre.
Nous proposons dans notre étude d'effectuer ces croisements manuellement, de telle sorte qu'ils aient un sens métier. Sept croisements de variables deux à deux ont ainsi été effectués :

- Qualité juridique X Type de bien
- Qualité juridique X Valeur de la garantie
- Qualité juridique X Nombre de pièces
- Surface X Nombre de pièces
- Qualité juridique X Montant de la cotisation
- Montant de la cotisation X Type de bien
- Qualité juridique X Montant du SMP [9]

4.3 Mesure de la liaison et de la stabilité des variables qualitatives

Le **V de Cramer** est une des mesures d'association les plus populaires, dans ce cas où les variables sont toutes qualitatives. C'est une normalisation de la statistique du χ^2 permettant [10] d'avoir une valeur (absolue) comprise entre 0 (indépendance) et 1 (liaison fonctionnelle).

La statistique du χ^2 est formellement décrite par l'expression suivante :

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

Où O est l'effectif observé et T l'effectif théorique sous l'hypothèse d'indépendance des deux variables [11].

9. SMP : Sinistre Maximum Possible.
10. D'autres mesures, comme le coefficient de contingence de Pearson (ϕ) ou de Tschuprow, sont également des fonctions du χ^2.
11. O_{ij} est l'effectif observé pour la modalité i de la variable explicative et la modalité j de la variable cible (même raisonnement pour T_{ij}).

Le V de Cramer est défini comme suit :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{N \times [\min(l,c) - 1]}} = \sqrt{\frac{\chi^2}{N}}$$

Où l (resp. c) est le nombre de lignes (resp. colonnes) du tableau de contingence associé, N est l'effectif total. Comme la variable explicative est binaire, nous avons forcément : $\min(l,c) = 2$.

Nous pouvons interpréter la statistique du χ^2 comme une distance entre deux variables : plus la distance est grande et plus la corrélation est forte, et à l'inverse plus la distance est faible et plus l'indépendance est forte.
Le V de Cramer est une mesure que nous pouvons comparer au coefficient de corrélation linéaire des variables continues. Notons toutefois que $V \in [0,1]$ en général, sauf dans le cas de deux variables avec deux modalités (2×2) où les valeurs peuvent être comprises dans l'intervalle $[-1,1]$ en fonction de l'agencement des modalités.

Pour résumer succinctement, l'objectif poursuivi désormais est triple :

- Avoir des variables explicatives les plus corrélées possible avec la variable cible.
- Avoir des variables explicatives les moins corrélées possible entre elles.
- Avoir des variables explicatives les moins corrélées possible entre elles-mêmes dans le temps.

4.3.1 Corrélation des variables explicatives avec la variable cible

Dans le cas où nous avons accès à un grand nombre de variables, il est très intéressant de quantifier la liaison entre les prédicteurs et la variable cible. Cela permet d'effectuer un premier tri et d'exclure les variables les moins pertinentes a priori.

Selon [Tuf17], il est possible d'apprécier qualitativement le pouvoir discriminant d'une variable à partir de la valeur du V de Cramer :

- Le pouvoir est **fort** si le V de Cramer est > 0.2.
- Le pouvoir est **bon** si le V de Cramer est compris entre 0.1 et 0.2.
- Le pouvoir est **moyen** si le V de Cramer est compris entre 0.05 et 0.1.
- Le pouvoir est **faible** si le V de Cramer est < 0.05.

Les résultats obtenus à partir des variables explicatives candidates pour la modélisation sont les suivants :

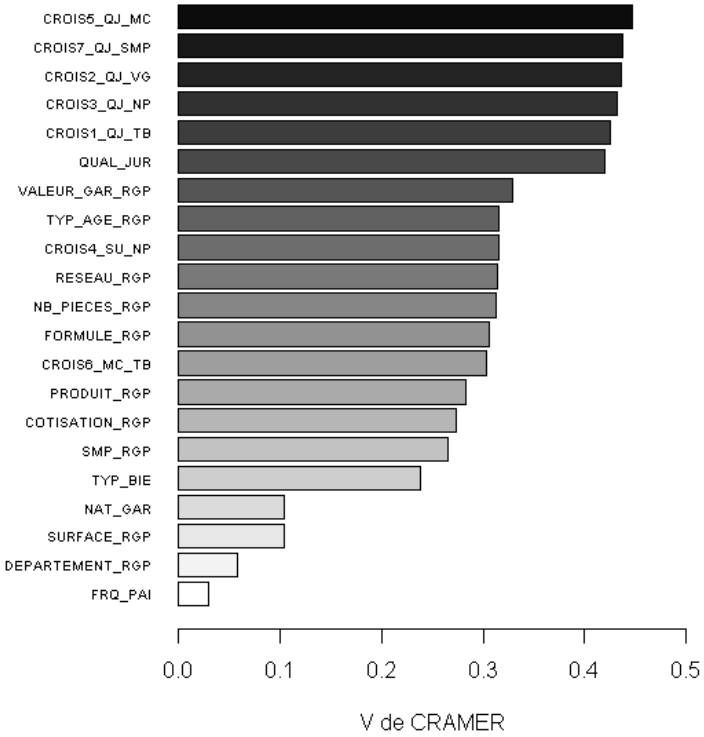


FIGURE 4.5 – Liaison des variables explicatives avec la variable cible

Nous observons tout d'abord que les cinq variables les plus discriminantes sont les croisements de « Qualité juridique » avec (dans l'ordre) : le montant de cotisation, la valeur du SMP, de la garantie, le nombre de pièces et le type de bien. Cependant, la différence de gain apportée par le croisement le plus discriminant et la variable seule (Qualité juridique) nous semble trop faible ($+ 6,48\%$) pour l'isoler dans un modèle. Pour des raisons d'interprétation, nous n'intégrerons aucun de ces cinq croisements dans les modèles.
En revanche, les croisements « Montant de cotisation » et « Type de bien » semblent apporter un gain assez important aux deux variables. Ce croisement sera donc conservé.
Nous avons également choisi de conserver le croisement des variables « Nombre de pièces » et « Surface » car lorsque l'une n'est pas renseignée, l'autre l'est.

Corrélation et causalité Il est très important de distinguer ces deux notions car elles peuvent être source de confusion ou pire, d'erreur d'analyse. Une causalité est une corrélation, mais la réciproque n'est pas vraie. Il est ainsi nécessaire de bien distinguer la cause de l'effet.
Nous pourrions être tentés de considérer qu'une causalité est une corrélation qui va se vérifier sur une durée très longue. Cependant, il faut prendre garde à d'éventuelles variables **confondantes** qui peuvent influencer chacune des deux variables dont le lien est mesuré [12].
Dans notre étude, nous supposons implicitement que les corrélations entre les variables explicatives et la variable cible sont des causalités. Comme le phénomène de résiliation est enregistré **après** l'enregistrement des données de souscription (cause), il peut être considéré comme l'effet.

4.3.2 Corrélation des variables explicatives entre elles

Il convient généralement, dans de nombreux types de modèle, de vérifier que les variables candidates soient indépendantes les unes des autres, ou du moins que les corrélations entre elles soient faibles. D'après [Tuf17], nous pouvons considérer que les corrélations mesurées avec un V de Cramer > 0.4 en valeur absolue sont trop fortes et doivent amener le statisticien à s'interroger sur la redondance des variables.

12. Par exemple, ce n'est pas parce que l'on a plus de (mal)chance de mourir à l'hôpital que chez soi, que l'hôpital est un lieu dangereux. On s'y rend car on est malade.

Un **corrélogramme** est une représentation graphique permettant de mettre en évidence d'éventuelles corrélations entre variables. La FIGURE 4.6, issue du logiciel R [13], synthétise ainsi l'ensemble des corrélations entre les variables.

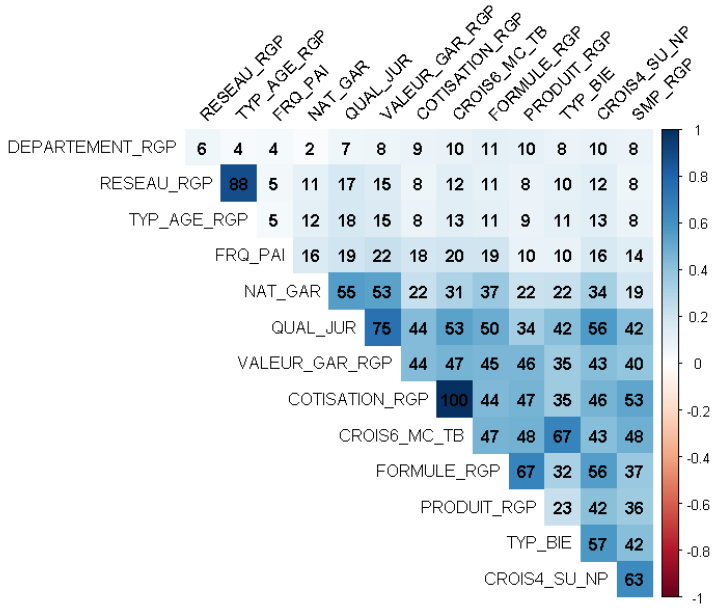


FIGURE 4.6 – Liaison des variables explicatives entre elles

Nous observons ainsi d'importantes corrélations pour les variables suivantes :

- Réseau et type d'agent
- Qualité juridique et valeur du mobilier garanti
- Formule et produit
- « Nombre de pièces / surface » et valeur du SMP

Ceci nous amène donc à être prudents lors de la modélisation, en particulier si ces couples de variables sont conservés dans un modèle. Il serait sans doute plus prudent d'en enlever une des deux (celle participant le moins à la variance du modèle par exemple).

4.3.3 Stabilité temporelle des variables explicatives

Nous avions suggéré l'intérêt de vérifier la dynamique temporelle des modalités des différentes variables. Cela est important dans le cadre de modèles censés prédire une mesure de risque **toute chose égale par ailleurs**. Ce leitmotiv est souvent évoqué, que ce soit dans la littérature scientifique ou dans les discussions professionnelles. La principale motivation est de travailler sur des données stables dans le temps qui vont pouvoir nous servir de **patron**. La métaphore photographique est également pertinente : nous pouvons difficilement reproduire un paysage à partir d'une photographie trop floue.
Il est bien sûr absurde de penser, d'une part que cette stabilité est parfaite sur notre échantillon de données, d'autre part qu'elle le restera dans le futur [14].

Nous allons ainsi proposer une méthode permettant de discuter la stabilité des variables. L'idée est de s'assurer que la distribution des variables, discrétisées précédemment, ne se déforme pas trop sur la profondeur temporelle de notre échantillon de construction (ici de début 2010 à fin 2014.). Nous pouvons voir cela de deux manières :

13. Package corrplot.
14. N'oublions pas que l'intérêt d'un modèle peut aussi être prédictif et non pas seulement explicatif.

- Qualitativement : en vérifiant que la part de chacune des modalités de la variable est stable dans le temps.
- Quantitativement : en effectuant des tests statistiques d'indépendance (ou d'adéquation) de type χ^2.

Illustrons cela à travers l'étude de la stabilité de la variable relative à la **qualité juridique** de l'assuré. La FIGURE 4.7 montre l'évolution, de 2010 à 2014, de la distribution de la variable.

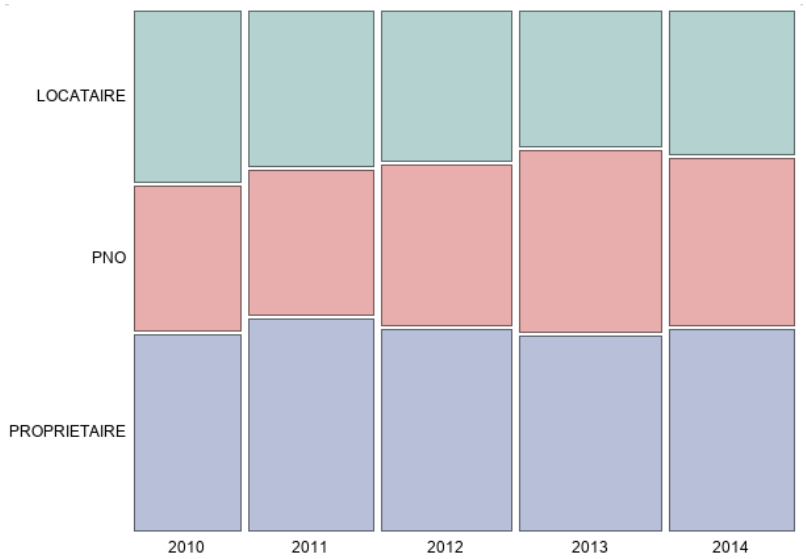


FIGURE 4.7 – Évolution temporelle de la variable Qualité juridique

Les modalités semblent assez stables dans le temps ; et en effet, si nous effectuons une régression linéaire simple pour chacune des trois modalités : le coefficient directeur (associé à la droite affine) contient à chaque fois zéro dans son intervalle de confiance à 95% (cf. FIGURE 4.8).

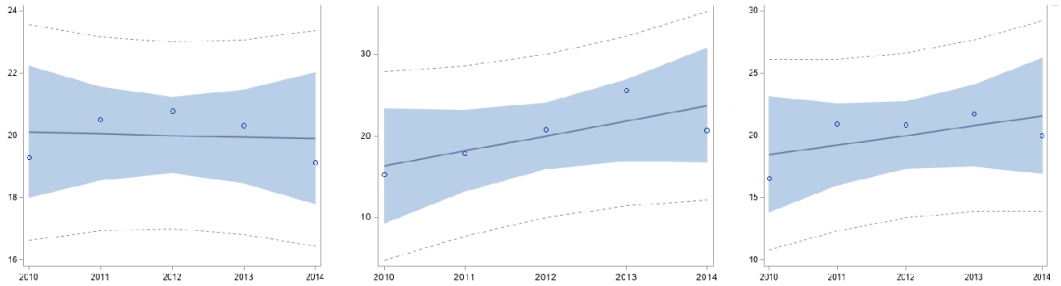


FIGURE 4.8 – Régression linéaire sur chacune des trois modalités

Nous complétons la discussion par la mesure du V de Cramer et deux tests du χ^2 : le test classique d'indépendance et le test de Cochran-Mantel-Haenszel [15]. Cinq périodes d'étude sont considérées : dans un premier temps sur deux années consécutives (2010 – 2011, 2011 – 2012, 2012 – 2013, 2013 – 2014) et enfin sur l'ensemble des périodes (2010 – 2014).
Le tableau 4.2 présente les résultats obtenus.

Observons tout d'abord que les différentes valeurs prises par le V de Cramer sont proches de zéro, ce qui suggère une relative stabilité au cours du temps.
Enfin, les deux tests ne donnent pas les mêmes conclusions :

- Test du χ^2 d'indépendance : l'hypothèse d'indépendance est toujours rejetée.

15. Notons qu'il n'existe pas, à notre connaissance, de test basé sur le V de Cramer car nous ne connaissons pas sa loi sous l'hypothèse d'indépendance.

Qualité juridique	V de Cramer	Khi^2		Khi^2 MH	
		Valeur	*Proba*	Valeur	*Proba*
2010 - 2011	**0,04**	46,35	*0,00*	46,18	*0,00*
2011 - 2012	**0,03**	43,08	*0,00*	1,36	*0,24*
2012 - 2013	**0,05**	93,61	*0,00*	2,95	*0,09*
2013 - 2014	**0,03**	44,11	*0,00*	0,12	*0,73*
2010 - 2014	**0,05**	443,05	*0,00*	39,65	*0,00*

TABLE 4.2 – Tests statistiques de la stabilité de la variable Qualité juridique

- Test du χ^2 de Mantel-Haenszel : l'hypothèse d'indépendance est vérifiée pour les périodes $2011-2012$, $2012-2013$ et $2013-2014$.

Les résultats pour les autres variables sont disponibles en annexe (cf. FIGURE D.2, page 126 et suivante). **Globalement, et même si les tests statistiques ne valident pas toujours la stabilité temporelle, nous allons considérer qu'elle est vérifiée.**

4.4 Analyse des Correspondances Multiples

En amont d'une modélisation sur variables qualitatives, il est intéressant de visualiser sur un plan les liens qui existent entre les deux modalités de la variable binaire à expliquer et les modalités des variables qualitatives candidates à la discrimination. Pour cela, l'analyse des correspondances multiples ou ACM est la méthode adéquate. Le lecteur pourra se référer à [LMP00] ou [HLP09] pour plus d'informations sur les méthodes dites d'**analyse des données**.

L'ACM est une généralisation de l'**analyse factorielle des correspondances**, qui s'applique à plus de deux variables qualitatives [16]. L'idée est de projeter orthogonalement un nuage de points sur des axes, dits **factoriels**, tout en maximisant leur inertie (ou **variance expliquée**). Ce nuage est représenté par les caractéristiques des individus dans le tableau disjonctif complet [17].
Dans ce nuage de points, deux individus sont proches, au sens de la distance du χ^2, s'ils possèdent des modalités identiques. Il est possible de superposer modalités et individus sur un même plan car le point représentant une modalité est le barycentre des individus possédant cette modalité (à un coefficient près) [18].
L'inertie totale du nuage de points se décompose par la somme des **valeurs propres** associées à la diagonalisation d'une matrice issue du tableau disjonctif complet. Ces valeurs propres caractérisent les axes factoriels et représentent, chacune, un pourcentage de variance expliquée.
En général, et c'est le cas dans la FIGURE 4.9, les données (individus ou modalités) sont représentées sur les deux premiers axes appelé **plan factoriel principal**.
Notons que le centre du plan factoriel correspond à l'individu moyen (i.e. au profil moyen des individus).
Chaque axe peut être interprété séparément, afin de mettre en évidence les oppositions entre individus.

Dans l'ACM effectuée avec nos données, la variable cible [19] n'a pas contribué à la construction des axes factoriels, elle a simplement été positionnée en tant que variable **illustrative**.
Nous observons ainsi que la modalité *RESIL* de la variable cible est plus proche des modalités relatives aux locataires, vivant dans un petit logement et ayant donc un montant de cotisation plus faible et un SMP plus réduit.
À l'inverse, la modalité *NON RESIL* est liée aux propriétaires, occupant un logement de plus grande taille avec un SMP plus important et devant donc s'acquitter d'une cotisation plus élevée.

16. À l'aide du tableau disjonctif complet et du tableau de Burt qui peut être vu comme un tableau de contingence pour l'ensemble des modalités.
17. Avec en lignes les individus et en colonnes leurs caractéristiques décrites par les variables qualitatives.
18. L'inverse est également vrai, d'après les relations de transition individus/modalités établies pour l'ACM.
19. Composée des modalités *NON RESIL* et *RESIL*, en vert dans le graphique.

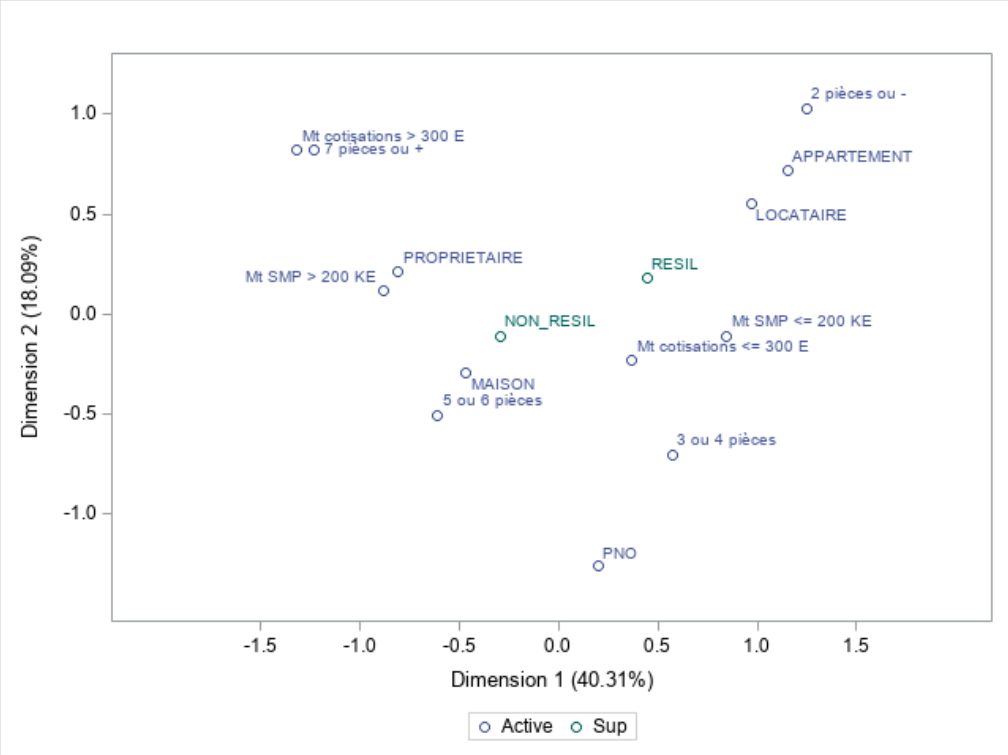


FIGURE 4.9 – Analyse des Correspondances Multiples

4.5 Échantillonnage

4.5.1 Nombres pseudo-aléatoires

Pour un exposé plus détaillé sur le sujet, le lecteur pourra se référer avec intérêt à [PTJ11].

Il est impossible pour un ordinateur, voire même un humain, de générer une (grande) suite de nombres qui soit purement aléatoire. C'est la raison pour laquelle, afin de s'en approcher, de nombreux algorithmes appelés **générateurs de nombres pseudo-aléatoires** [20] ont été proposés.
Leur objectif est de générer une suite pseudo-aléatoire, et considérée i.i.d., $u_{n+1} = f(u_n)$ à partir d'une graine u_0 (*seed* en anglais).
Ces générateurs présentent donc la capacité de reproduire le caractère aléatoire de ces nombres, ce qui est très intéressant pour au moins deux raisons :

- La plupart des lois continues générées pseudo-aléatoirement dans les logiciels sont construites par la méthode d'inversion de la fonction de répartition appliquée à une loi uniforme sur $[0, 1]$.
- Les exigences réglementaires font que l'actuaire doit intégrer dans ses processus une démarche d'auditabilité qui induit une parfaite reproductibilité des résultats obtenus.

Sans rentrer dans des détails techniques exhaustifs, notons que c'est le générateur de Mersenne-Twister qui est utilisé par défaut dans les logiciels R et SAS.

Nous admettrons par la suite, et pour être cohérent avec la théorie des méthodes appliquées (notamment celles utilisant la simulation), que ces suites de nombres ainsi générées sont parfaitement aléatoires et indépendantes (ce qu'il est, en réalité, abusif de penser).

20. Ou PRNG pour Pseudo Random Number Generator.

4.5.2 Méthode d'échantillonnage appliquée

Afin de ne pas être trop optimiste sur les résultats d'un modèle d'apprentissage supervisé, il est admis de scinder au moins en deux la **base de construction** afin d'obtenir les échantillons suivants :

- Un **échantillon d'apprentissage**, à partir duquel se fait la modélisation de la variable cible (qui consiste à optimiser l'étiquetage des données en deux groupes).
- Un **échantillon de validation** (indépendant du précédent), permettant de mesurer de manière objective la prédictivité du modèle obtenu à partir des observations de la base d'apprentissage.

Dans certains cas, un **échantillon hors du temps** [21] peut être proposé en complément. Ceci est intéressant dans le cadre de la création d'un score de risque amené à être utilisé sur une longue période.

L'intérêt de ce type d'échantillonnage est de s'assurer que le modèle proposé est robuste, c'est-à-dire que le phénomène de **surapprentissage** est maîtrisé. L'objectif est donc, idéalement, de maximiser les performances du pouvoir discriminant du modèle dans les différents types d'échantillon.
Nous ne pouvons pas faire grand chose, en réalité, dans l'échantillon de validation car le modèle est ajusté à partir des seules données présentes dans l'échantillon d'apprentissage.
Il peut même arriver que plus un modèle est performant en construction, moins il est prédictif (un cas pathologique simple de surapprentissage, pour une variable continue, est le modèle polynomial).
Ce genre de précaution méthodologique est fondamental, car bien souvent l'intérêt d'un tel modèle est d'être appliqué sur des données inconnues du SI (et donc n'ayant pas participé à sa conception).

Nous admettrons les points suivants :

- La base d'apprentissage comporte généralement deux fois plus d'observations que la base de validation (ce qui donne la réparation $2/3 - 1/3$) [22].
- La proportion préconisée pour l'échantillon d'apprentissage (2/3) correspond environ à la probabilité de tirer au moins une fois un individu donné dans un tirage aléatoire simple avec remise dans l'échantillon de construction (soit $0,632$) [23].
- 1 000 à 2 000 observations d'apprentissage suffisent pour construire un modèle correct (par la **loi forte des grands nombres** : plus les observations sont nombreuses, plus le modèle va converger vers la « bonne » proportion).
- Les groupes à modéliser peuvent être équi-répartis [24] ($50\% - 50\%$) dans chacun des échantillons (c'est même conseillé si l'événement d'intérêt est rare).

4.5.3 Compromis Biais-Variance

Ce dilemme a été initialement développé dans le cadre de modèles de régression, mais il s'applique aussi aux problèmes de classification.
L'**erreur de modélisation** peut se décomposer en la somme de l'**erreur de modèle** (appelée l'erreur de spécification) et de l'**erreur d'estimation** (due à une mauvaise estimation des paramètres ajustés sur le modèle choisi) [25]. L'erreur d'estimation se mesure généralement à l'aide de l'**erreur quadratique**.
L'erreur quadratique moyenne d'un estimateur $\hat{\theta}$ du paramètre θ se décompose ainsi :

$$\begin{aligned}\mathbb{E}[(\theta-\hat{\theta})^2] &= \Big(\mathbb{E}[\theta-\hat{\theta}]\Big)^2 + \mathbb{V}[\hat{\theta}] \\ &= \Big[\text{Biais}(\hat{\theta})\Big]^2 + \text{Variance}(\hat{\theta})\end{aligned} \tag{4.1}$$

21. Afin de s'assurer que le modèle, construit dans un cadre statique, est robuste dans un cadre plus dynamique.
22. Pour chaque observation, un nombre aléatoire est tiré suivant une loi uniforme sur $[0,1]$. Si la valeur de ce nombre est inférieur à 2/3, alors l'observation sera incluse dans la base d'apprentissage. Sinon elle sera incluse dans la base de validation.
23. $0,632 = \lim_{n\to\infty}[1-[1-1/n]^n] = 1-\mathrm{e}^{-1}$, suite à un développement limité au voisinage de zéro de la fonction logarithme.
24. Si le taux modélisé n'est pas représentatif de celui de la population de départ, nous pouvons procéder à un redressement formel de la constante pour un modèle de régression. Nous le verrons par la suite.
25. Nous supposerons par la suite, abusivement sans doute, que notre modèle est bien spécifié. L'erreur de modèle sera donc considérée nulle.

Pour une erreur quadratique constante, nous observons que chercher à diminuer le biais revient à augmenter la variance (et inversement). En régression, l'ajout de variables explicatives tend à diminuer le biais de la prédiction mais rajoute de la variance à cette prédiction. À l'inverse, un modèle plus parcimonieux aura tendance à mieux se généraliser au détriment de la précision (cf. FIGURE 4.10).
Au niveau de l'échantillonnage, augmenter la taille de l'échantillon d'apprentissage réduit la variance du modèle (mais le gain est faible au-delà d'une certaine taille).
Dans notre étude, nous cherchons à estimer la probabilité de résiliation d'une police d'assurance : le paramètre θ sera donc identifié par cette probabilité.
Nous verrons plus tard (cf. section 5.3) quelques méthodes permettant soit de diminuer le biais, soit de diminuer la variance des modèles.
L'objectif final est d'arriver à un bon compromis permettant de faire confiance à la généralisation du modèle sélectionné (schématisé par la ligne verticale rouge ci-dessous).

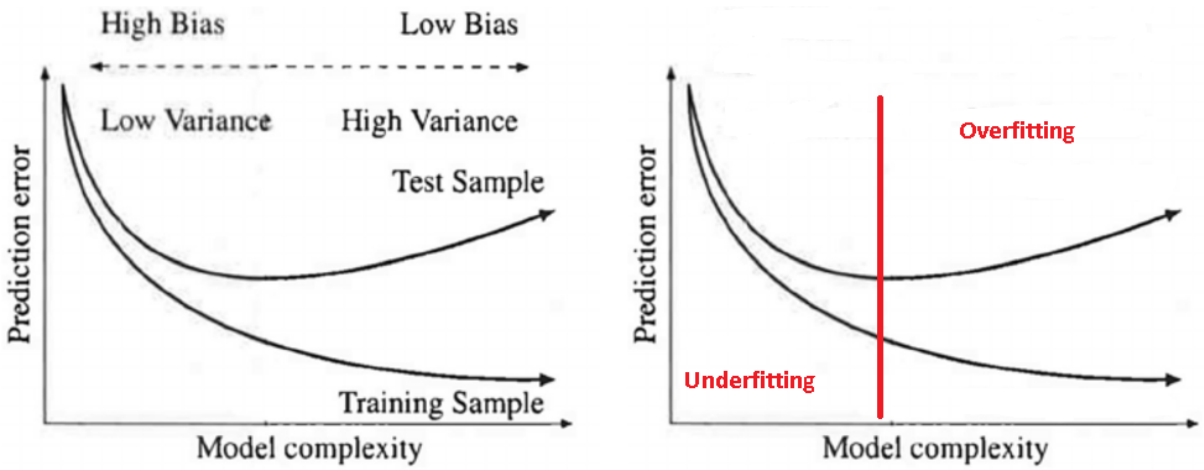


FIGURE 4.10 – Complexité et erreur (source :[BMTW12])

4.6 Modèle considéré

La taxonomie des modèles est très large, aussi nous nous intéresserons à l'un des plus populaires pour ce type de problème : le modèle de **régression logistique binaire** (cf. section C en annexe) avec une fonction de lien *logit*.
Cette fonction de lien permet d'interpréter les coefficients issus de l'ajustement en termes de rapports de côtes ou ***odd-ratio***, noté OR. Dans notre étude, l'*odd-ratio* d'une modalité est le rapport entre la côte de la probabilité de résilier et celle de ne pas résilier. Soit p la probabilité de résilier, $\text{OR} = \frac{p}{1-p}$ est donc, par définition, le rapport de côtes.
Nous rappelons que le modèle de régression s'écrit $\text{logit}(p) = X\beta$, ce qui revient à écrire : $\frac{p}{1-p} = \exp(X\beta) = \text{OR}$. Le rapport de côte peut donc simplement s'exprimer à l'aide de l'exponentielle des coefficients du modèle.
Notons que nous ne tiendrons pas compte directement des **interactions** entres les variables explicatives car leur interprétation est souvent délicate. Nous les avons introduites indirectement en construisant de nouvelles variables (cf. section 4.2.3).

4.6.1 Sélection des variables

Le nombre de modèles possibles, disposant de k variables explicatives, est de $\sum_{i=0}^{k} C_k^i = 2^k$ (d'après la formule du binôme de Newton). Les ordinateurs de base actuels ne permettent pas d'effectuer les calculs rapidement [26]. C'est pour cela que des méthodes alternatives de sélections automatiques des variables ont été développées comme la méthode de sélection pas à pas **descendante** (on part de toutes les variables puis on supprime celles qui sont les moins explicatives) et la méthode **ascendante** (on part d'une variable et on rajoute les variables les plus explicatives).

26. Pour 10 variables explicatives, 1 048 576 modèles sont à évaluer, pour 100 l'ordre de grandeur est de 10^{30}.

La méthode descendante présente l'inconvénient qu'il n'est plus possible de réintroduire une variable dès lors qu'elle a été supprimée. Inversement, la méthode ascendante, bien que la plus rapide, ne permet plus la suppression d'une variable dès lors qu'elle est conservée.
Nous avons choisi d'effectuer nos sélections de variables à l'aide de la méthode **stepwise** [27], qui peut être considérée comme un bon compromis entre les méthodes ascendante et descendante.
Cette méthode donne des résultats très corrects, mais il faut prendre garde à d'éventuelles corrélations entre les variables conservées dans le modèle final.
Précisons, enfin, que ces trois méthodes de sélection cherchent à optimiser un critère de qualité de l'ajustement (AIC, BIC ou Déviance par exemple.).

4.6.2 Qualité de la discrimination

Nous proposons au lecteur intéressé de consulter l'ouvrage [Ten07] pour plus de détails.

Beaucoup de mesures de la **qualité de l'ajustement** (AIC, BIC, déviance, R^2, etc.) ou de la **qualité de la discrimination** (taux de concordance et d'erreur, rappel, Gini, etc.) sont disponibles.
Des **indicateurs graphiques** existent également, tels que la courbe ROC ou la courbe de lift.
Nous allons présenter les deux indicateurs qui nous semblent les plus synthétiques : la courbe ROC et l'AUC.

Quelques notations Soit $\hat{Y} \in \{0, 1\}$ l'estimateur de l'acte de résiliation modélisé par une variable aléatoire Y. Considérons les deux événements suivants :

- $[\hat{Y} = 1|Y = 1]$: la résiliation est prédite à raison (vrai positif).
- $[\hat{Y} = 1|Y = 0]$: la résiliation est prédite à tort (faux positif).

La **sensibilité** α est définie comme étant la probabilité que la résiliation soit prédite à raison :

$$\alpha = \mathbb{P}[\hat{Y} = 1|Y = 1]$$

La **spécificité** β est définie comme étant la probabilité que la non résiliation soit prédite à raison :

$$\beta = \mathbb{P}[\hat{Y} = 0|Y = 0]$$

Dans la pratique, ces deux valeurs dépendent d'un **seuil** s, pouvant être interprété comme un score ou une probabilité, à partir duquel l'acte de résiliation est considéré comme réalisé.
Ce seuil permet de définir une **matrice de confusion** [28], qui mesure la qualité de la classification. Ce seuil peut varier, par exemple en fonction d'un ratio coût-bénéfice à définir [29].
Nous ne rentrerons pas dans plus de détails techniques, car les indicateurs présentés ci-dessous vont nous permettre de généraliser le concept de matrice de confusion grâce au couple $(\alpha(s), \beta(s))$.

Courbe ROC La courbe ROC, qui est l'acronyme de **Receiver Operating Characteristic** (cet indicateur provient de la théorie du signal), mesure la performance d'un classificateur binaire en comparant la probabilité a priori de l'évènement modélisé avec le résultat observé a posteriori. Elle représente, dans le plan euclidien, le taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (antispécificité) pour chaque seuil s, i.e. le couple $(1 - \beta(s), \alpha(s))$ défini ci-dessus.
La fonction ainsi définie est non décroissante de $[0, 1] \mapsto [0, 1]$.
Un bon modèle doit être sensible et spécifique : il est donc préférable que la courbe croisse très vite.
Cette courbe généralise ainsi la matrice de confusion, qui ne permet de visualiser la performance d'un modèle qu'à partir d'un seuil s unique : ici le seuil varie (presque [30]) continûment dans l'intervalle $[0, 1]$. Un des avantages de la courbe ROC est qu'elle permet de comparer assez facilement plusieurs modèles candidats à la résolution d'un problème.

27. Cette méthode, ainsi que les deux autres, est implémentée dans l'option SELECTION de la procédure LOGISTIC du logiciel SAS. Elle permet, notamment, de revenir sur des choix précédents d'ajout ou de suppression de variables.
28. Matrice 2×2 avec en ligne les résultats du test (binaire), en colonne les observations.
29. Par défaut, nous pouvons le fixer à 0,5 ou au même niveau que la proportion observée.
30. Les données, bien que nombreuses, sont discrètes.

AUC En complément, nous devons présenter l'AUC, qui quantifie le pouvoir discriminant d'un modèle de classification binaire.
L'AUC (pour *Area Under Curve*) est directement liée à la courbe précédente, car il s'agit de l'aire entre elle-même et l'axe des abscisses. Ainsi, plus l'AUC est grande (maximum 1), meilleur est le modèle.
Formellement, il s'agit d'une intégrale [31] dont la valeur est comprise entre 0 et 1 :

$$\text{AUC} = \mathbb{P}\left[[\hat{Y} = 1|Y = 1] > [\hat{Y} = 1|Y = 0]\right] = \int_{[0,1]} (1 - \beta(s))d\alpha(s)$$

Dans le cas idéal AUC = 1 : le modèle classe toujours un client qui va résilier devant un client qui ne va pas résilier.
En revanche, si AUC = 1/2, alors cela signifie que le modèle ne fait pas mieux que le hasard, i.e. qu'il considère que $\mathbb{P}[Y = 1|X] = \mathbb{P}[Y = 0|X]$ (cas représenté par la première bissectrice sur le graphe).

Les critères suivants sont souvent proposés dans la littérature afin de juger de la qualité d'un modèle :

- Si AUC $= 0,5$: pas de discrimination.
- Si $0,7 \leq$ AUC $< 0,8$: discrimination acceptable.
- Si $0,8 \leq$ AUC $< 0,9$: discrimination excellente.
- Si AUC $\geq 0,9$: discrimination exceptionnelle.

AUC et Gini Notons que l'AUC et l'indice de Gini sont liés par la relation suivante :

$$\text{AUC} = \frac{\text{Gini} + 1}{2}$$

L'indice de Gini est donc égal au double de l'aire entre la courbe ROC et la première bissectrice. En l'absence d'ex-aequo [32], l'indice de Gini est équivalent au **D de Somers**.

4.6.3 Résultats de la discrimination

Nous avons testé un premier modèle de régression logistique, pour le risque de souscription à horizon 36 mois, sur des polices MRH souscrites entre 2010 et 2014 (pour avoir une structure de portefeuille stable). La volumétrie de construction est détaillée dans le tableau 4.3.

	Résiliées	**Non résiliées**	**Total**	**% total**
Apprentissage	23 813	39 344	**63 157**	*66,61%*
Validation	11 989	19 667	**31 656**	*33,39%*
Total (Construction)	**35 802**	**59 011**	**94 813**	*100,00%*
% total	*37,76%*	*62,24%*	*100,00%*	

TABLE 4.3 – Volumétrie de l'échantillon de construction.

Les résultats de la régression sont détaillés dans la FIGURE D.5, en annexe page 129.
Leur interprétation est relativement aisée si l'on se base sur l'antépénultième colonne du tableau. En effet, cette dernière présente les *odd-ratio* (OR) des variables : si OR >1, cela signifie que le risque de résiliation est supérieur à la moyenne ; s'il est compris entre 0 et 1, le risque de résilier est moindre.
Ces résultats confirment et complètent ceux observés précédemment à l'aide de l'ACM.

Pour une modélisation binaire, nous représentons graphiquement la qualité de la discrimination via la courbe ROC. Grâce aux méthodes numériques implémentées dans le logiciel, il est possible d'estimer l'aire sous cette courbe (AUC). La FIGURE 4.11 ainsi que le tableau 4.4 nous fournissent ces indications.

31. Il existe de nombreuses méthodes numériques permettant d'estimer cette intégrale, notamment à l'aide des U-statistiques de Mann-Whitney.
32. Paire (d'individus) ni discordante ni concordante.

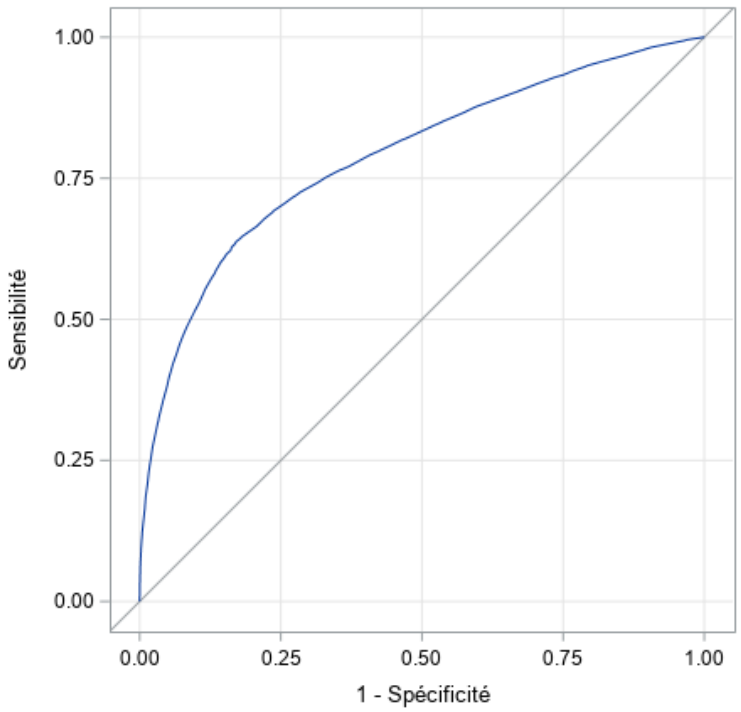


FIGURE 4.11 – Courbe ROC sur l'échantillon d'apprentissage.

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	78,3	**D de Somers**	0,576
Pourcentage discordant	20,7	**Gamma**	0,582
Pourcentage lié	0,9	**Tau-a**	0,271
Paires	936 898 672	**c**	0,788

TABLE 4.4 – Indicateurs de mesure de la qualité.

- ⇨ L'AUC est donnée par le c, qui prend ici comme valeur[33] $0,788$.
- ⇨ La discrimination à 36 mois peut ainsi être considérée comme **acceptable**, selon nos critères cités en amont.
- ⇨ Bien sûr, la robustesse a été testée sur un échantillon de validation : l'AUC est du même ordre de grandeur $(0,781)$, ce qui laisse penser que le modèle est robuste.

4.6.4 Diagnostic de la modélisation

En complément de ces deux indicateurs (courbe ROC et AUC), nous pouvons aussi visualiser la distribution du modèle sur les deux modalités de la variable à expliquer. Les graphiques de la FIGURE 4.12 nous proposent d'observer la densité de la distribution (à gauche) ainsi que sa fonction de répartition empirique (à droite).

À gauche, le graphique du bas représente la répartition de la probabilité de résilier pour les individus qui ont finalement résilié. Nous observons qu'il est plus étalé sur la droite que le graphique du haut, ce qui se mesure à l'aide du *skewness* (qui est donc négatif). Cela est logique car, idéalement, nous souhaitons avoir une probabilité de résiliation proche de 1 sur cette population.
À droite, la fonction de répartition de cette même population est toujours inférieure à celle de la population de ceux qui ne résilient pas : la part des individus qui ne résilient pas est plus importante sur les probabilités

33. Le Gini est donné par le *D de Somers*, aux ex-aequo près : il vaut $0,576$ $(= 2 \times AUC - 1)$.

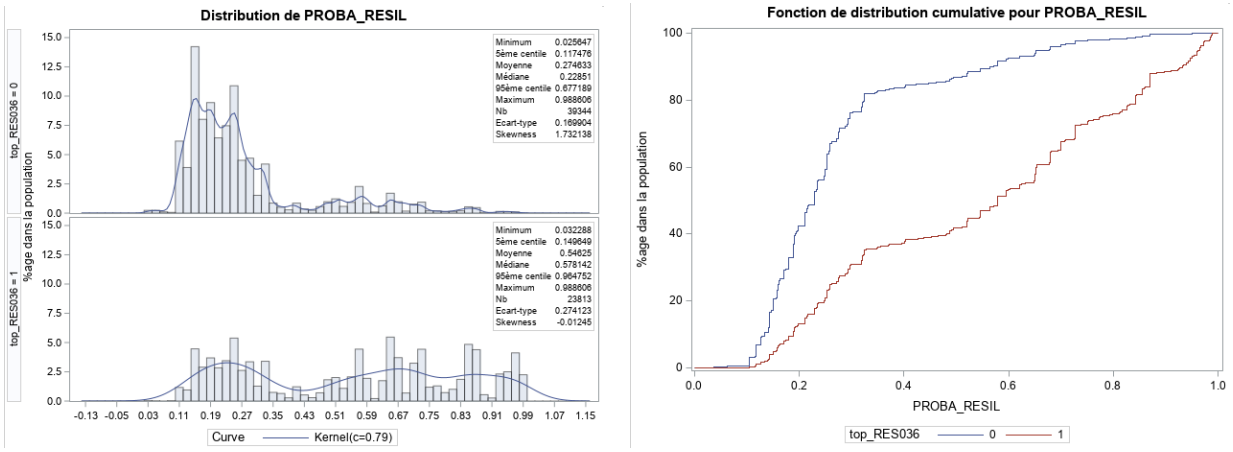


FIGURE 4.12 – Distribution des deux évènements modélisés.

faibles (à gauche de l'axe des abscisses) et diminue progressivement ensuite.

La FIGURE 4.13 nous montre, d'une part la répartition de l'ensemble de la population en fonction de sa classe de score [34], d'autre part le taux de résiliation observé sur ces mêmes classes de score.

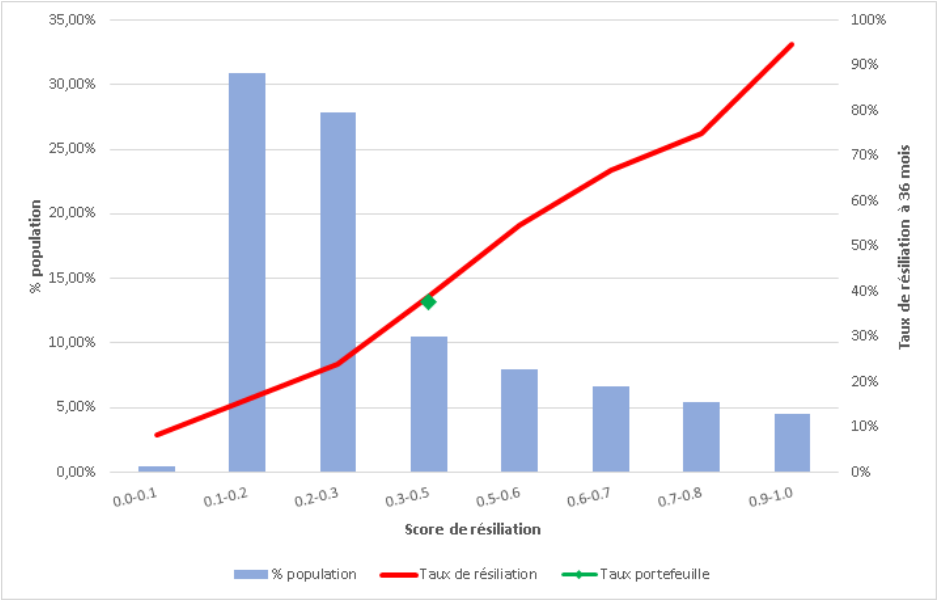


FIGURE 4.13 – Taux de risque en fonction de la classe de score.

Nous observons bien, et c'est logique, une croissance du taux de résiliation en fonction de la classe de score : la hausse du taux, modérée au départ, semble s'accélérer par la suite.
Nous notons, de plus, que la majorité de la population se situe dans les classes $[0.1-0.2]$ et $[0.2-0.3]$ (faible probabilité de résilier).

34. Ensemble ordonné, représenté ici par des tranches de **probabilité de résilier**.

4.7 Discussion sur l'horizon du risque mesuré

Nous allons, dans ce paragraphe, mener une réflexion concernant le choix de l'horizon d'observation de la résiliation. Nous souhaiterions savoir si les modèles obtenus par régression logistique avec une variable cible à 12 mois et une variable cible à 36 mois sont fondamentalement différents.
Pour cela, nous avons rééchantillonné B échantillons à partir de la base de construction décrite ci-dessous (horizon 36 mois) et B autres échantillons à partir d'une base de construction pour un modèle à horizon 12 mois ($B = 1000$, cette méthode est inspirée du bagging que nous allons décrire en section 5.3.1). Les variables cibles ont été stratifiées avec une équi-distribution des résiliés et des non résiliés (50%/50%).
Cela a permis d'obtenir deux fois 1000 modèles et donc d'obtenir un comptage du nombre de fois où une variable est sélectionnée dans un modèle ainsi que la valeur d'ajustement du coefficient. Ces coefficients sont ensuite « moyennisés ».

4.7.1 Valeurs moyennes des coefficients à 12 et 36 mois

Les résultats obtenus (sur 43 modalités de variables) sont représentés sur la FIGURE 4.14.

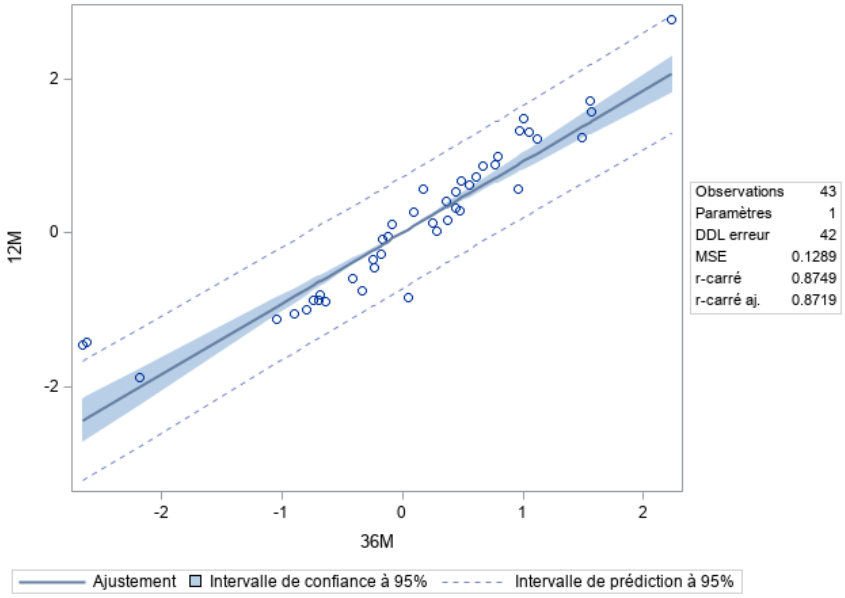


FIGURE 4.14 – Comparaison des coefficients 12M et 36M

La moyenne des coefficients à horizon 36 mois est représentée en abscisse ; en ordonnée celle des coefficients à horizon 12 mois : nous observons que les points ne sont pas situés trop loin de la première bissectrice (droite bleue).
Nous pouvons également calculer la corrélation (linéaire), entre 12 et 36 mois :

- De la valeur des coefficients moyennisé : $\rho_{\text{coef}} = 0,914$.
- Du nombre de fois où les mêmes variables apparaissent : $\rho_{\text{nb}} = 0,966$.

Cela laisse ainsi penser qu'un modèle à 36 mois explique assez bien le risque à 12 mois et inversement.

4.7.2 Bootstrap des AUC à 12 et 36 mois

Pour conclure, nous en avons profité pour obtenir des intervalles de confiance bootstrap sur les AUC des échantillons de construction. Le principe de base du **bootstrap des individus** est de tirer B séries de n-échantillons aléatoires et simples avec remise parmi le N-échantillon initial. Ces B échantillons bootstrappés vont ainsi permettre l'estimation du paramètre θ d'intérêt [35] et, si besoin, de proposer des intervalles de

35. Ici l'AUC.

confiance bootstrap (se référer à l'ouvrage de référence de [ET93] pour plus de détails).
La **méthode de l'erreur standard**, paramétrique, propose des limites de confiance données par le pourcentile $1-\alpha/2$ de la loi normale centrée réduite [36].
La **méthode des pourcentiles simples**, non paramétrique, propose des limites de confiance données par les pourcentiles $\alpha/2$ et $1-\alpha/2$ de la distribution d'échantillonnage empirique [37].

Le tableau 4.5 nous présente ainsi des indicateurs sur la dispersion des AUC des différents modèles obtenus. Les graphiques en FIGURE 4.15 nous montrent la distribution des AUC bootstappées sur les échantillons de validation.

STATISTIQUES	AUC APP 12	AUC VAL 12	AUC APP 36	AUC VAL 36
Nb Échantillons	*1000*	*1000*	*1000*	*1000*
Moyenne	0,7701	0,7527	0,7904	0,7733
Ecart-type σ	0,0111	0,0156	0,0115	0,0153
Moyenne - 1,96σ	0,7483	0,7221	0,7679	0,7433
Moyenne + 1,96σ	0,7919	0,7833	0,8129	0,8032
Coeff Variation	1,4424	2,0741	1,4508	1,9784
Ecart interquartile	0,0154	0,0214	0,0158	0,0204
Intervalle	0,0763	0,0915	0,0669	0,0976
Min	0,7302	0,7042	0,7572	0,7192
P1	0,7439	0,7129	0,7624	0,7373
P2.5	0,7481	0,7203	0,7658	0,7412
P5	0,7511	0,7264	0,7713	0,7475
P10	0,7554	0,7331	0,7758	0,7539
P25	0,7622	0,7424	0,7826	0,7635
P50	0,7710	0,7527	0,7906	0,7731
P75	0,7776	0,7638	0,7985	0,7839
P90	0,7837	0,7717	0,8055	0,7932
P95	0,7882	0,7775	0,8088	0,7981
P97.5	0,7910	0,7818	0,8125	0,8023
P99	0,7941	0,7869	0,8155	0,8069
Max	0,8065	0,7957	0,8241	0,8168

TABLE 4.5 – Indicateurs de distributions des AUC.

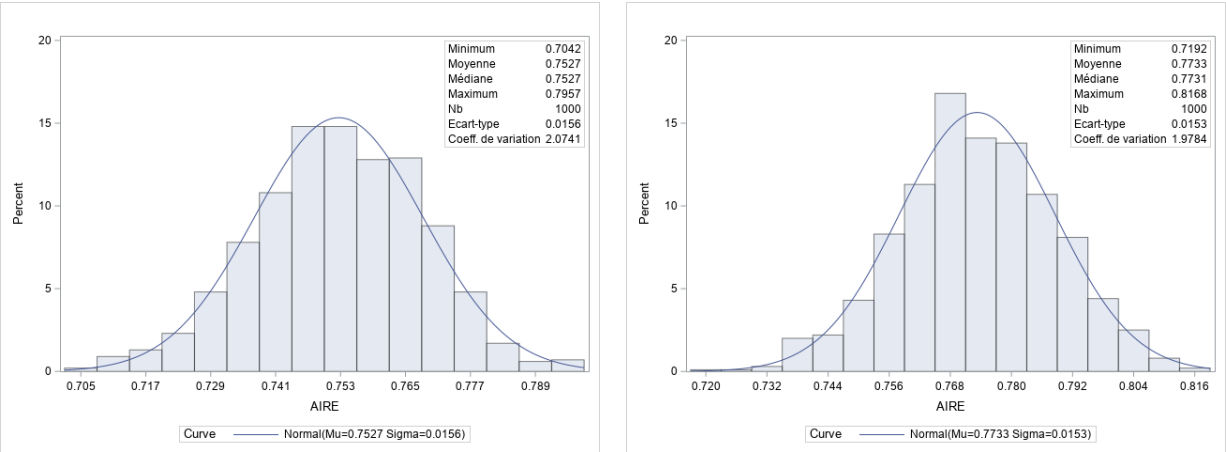


FIGURE 4.15 – Distribution bootstrap des AUC (validation) à 12 et 36 mois.

36. Formellement : $\theta \in \left[\hat{\theta} \pm u_{[1-\alpha/2]}\hat{\sigma}_{\hat{\theta}}\right]$
37. Formellement : $\theta \in \left[\hat{\theta}_{[\alpha/2]}\ ;\hat{\theta}_{[1-\alpha/2]}\right]$

Il est ainsi intéressant de noter que l'AUC du modèle à 36 mois est toujours supérieure à celle du modèle à 12 mois. Cela semble logique, si nous supposons que la structure du portefeuille ne change pas beaucoup (stabilité temporelle des variables explicatives) car le risque est observé sur une période trois fois plus longue.
Comme nous pouvions nous y attendre avec le bootstrap, la distribution des AUC est proche de celle d'une gaussienne (cf. FIGURE 4.15). Les tests de normalité [38] ne rejettent d'ailleurs quasiment jamais cette hypothèse.
Ces informations illustrent, de plus, assez bien la sensibilité que peut avoir l'indicateur de mesure de la performance (ici l'AUC) face au choix d'un échantillon (la qualité peut être jugée acceptable ou excellente) :

- La méthode de l'erreur standard nous donne les plages $[0,7221 - 0,7833]$ et [0,7433 - 0,8032] comme limites de confiance des AUC avec les modèles à 12 et 36 mois respectivement.
- La méthode des pourcentiles simples [39] nous donne des plages proches : $[0,7203 - 0,7818]$ et $[0,7412 - 0,8023]$ comme limites de confiance des AUC avec les modèles à 12 et 36 mois respectivement.
- Enfin, nous proposons en annexe page 130, avec la FIGURE D.6, une visualisation des courbes ROC associées aux 1 000 échantillons de validation (à 12 et 36 mois).

Nous venons ainsi d'étudier la résiliation des contrats **à partir de leur date de souscription**. Les principales méthodes de traitement des données, en amont de la modélisation par régression logistique binaire, ont été présentées de façon relativement détaillée afin de proposer des pistes pour élaborer un modèle robuste. L'étude de l'horizon d'observation a mis en évidence une convergence des modèles à 12 et 36 mois. Nous allons désormais, dès le chapitre suivant, étudier les résiliations sur **une durée de un an à partir d'un stock de polices** à une date donnée.

38. Proposés en sortie de SAS.
39. Au seuil $\alpha = 5\%$.

Chapitre 5

Segmentation supervisée des profils de résiliation MRH sur un stock

5.1 Périmètre de l'étude

Afin de nous rapprocher de notre problématique Solvabilité 2 [1], nous allons désormais tenter de modéliser la résiliation à horizon annuel à partir des stocks de contrats en cours vus au 31/12 des années 2014 à 2016. Nous ne reviendrons pas sur les techniques employées en amont de la modélisation car elles ont précédemment été décrites. Seuls les principaux résultats seront exposés.
Notons que les données comportementales suivantes sont désormais candidates au modèle : variables relatives à d'éventuels sinistres ou avenants au contrat, maturité de la police.

5.1.1 Définition de la variable cible

Il faut tout d'abord sélectionner des polices en cours [2] à la fin de l'année N :

$$\text{Date Effet} \leq 31/12/N < \text{Date Résiliation}$$

La valeur des variables explicatives est ainsi censée être la photo au $31/12/N$ de ce stock de polices.

La variable cible est, quant à elle, basée sur la photo au $31/12/N{+}1$ de l'ensemble des polices résiliées à cette date. Une simple jointure à gauche entre la base au $31/12/N$ et celle au $31/12/N{+}1$ va nous permettre de connaître l'état du stock de polices au $31/12/N$ vu un an plus tard. En conséquence, la définition de la variable cible Y est la suivante :

$$Y = \mathbb{1}_{\text{Date Résiliation} \in [01/01/N+1,\ 31/12/N+1]}$$

5.1.2 Base de construction

Les individus-contrats ont été mis sous forme longitudinale. Or, si un individu est présent plusieurs années en stock, il est observé sur plusieurs lignes. Ces observations ne sont donc plus indépendantes car il s'agit du même contrat.
Cette autocorrélation entre donc en contradiction avec l'hypothèse fondamentale que les observations doivent être i.i.d. dans un modèle de régression. Nous pouvons proposer plusieurs options face à ce problème :

- L'ignorer, si l'autocorrélation est jugée minime.
- Calculer la durée entre deux dates t_1 et t_2, et considérer qu'à partir d'un seuil (durée suffisamment longue) les observations sont indépendantes.
- Choisir aléatoirement une des observations apparaissant plusieurs fois dans l'échantillon.
- Si la structure de corrélation est considérée connue, l'introduire dans le modèle.

1. Pour rappel : estimation d'un taux de résiliation dans les 12 mois.
2. Nous supposerons que si la police n'est pas résiliée, alors sa date de résiliation est initialisée au 31/12/2099.

Comme le nombre de périodes observées est relativement faible (trois exercices), nous supposerons que les individus sont indépendants entre eux (ce qui est bien entendu faux). Nous verrons plus loin que ce choix n'impacte ni la cohérence ni les performances globales du modèle.

Les découpages des variables numériques et les regroupements des variables qualitatives ont été effectués à partir de ces données de stock. Les résultats sont présentés en annexe page 132.

Nous avons enfin scindé la base de construction en deux : apprentissage et validation, dans les mêmes proportions qu'au chapitre précédent. Un échantillon « hors-du-temps » est venu compléter celui de validation. Notons que les découpages et regroupements ont été effectués sur la base de construction entière (comme pour le score à la souscription). Nous pouvons supposer que les résultats n'auraient pas été fondamentalement différents si les traitements avaient été faits sur la base d'apprentissage uniquement.

5.2 Résultats de la régression

Les résultats obtenus sont loin d'être exceptionnels. Ils confirment la difficulté, observée lors de l'étude par des modèles de durée au chapitre 3, à obtenir un taux de résiliation très prédictif. Les trois courbes ROC sont présentées ci-dessous en 5.1.

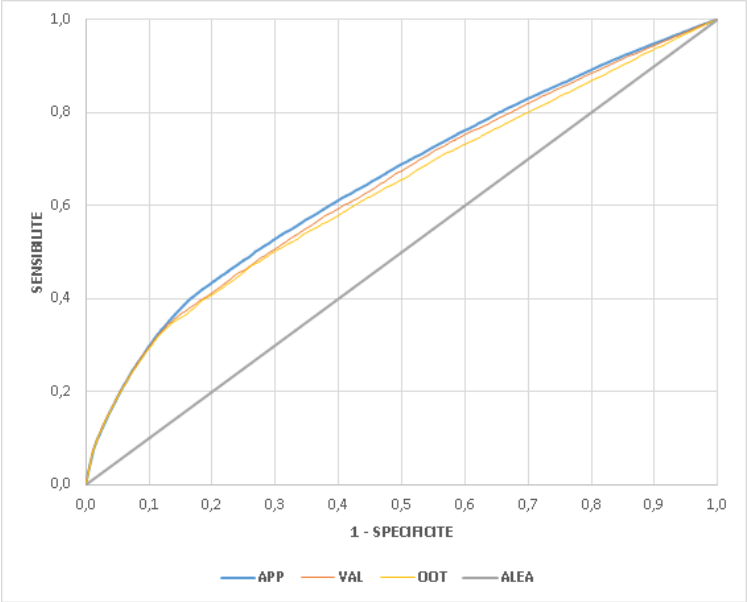


FIGURE 5.1 – Courbes ROC - Apprentissage, Validation et Out-Of-Time

L'AUC sur apprentissage est de $0,65$; sur validation de $0,64$. La discrimination est donc jugée moyenne (ni nulle, ni satisfaisante). Le modèle semble cependant assez robuste : l'écart de performance entre apprentissage et validation n'est pas très élevé ($-1,64\%$).

Échantillon « hors du temps » Le stock de polices non résiliées au 31/12/2017 peut être considéré comme un échantillon « *out-of-time* » (cf. section 4.5.2 du chapitre précédent).
Ce type d'échantillon va nous permettre de tester la robustesse du modèle dans un cadre plus dynamique. Nous observons ainsi une AUC de $0,63$: bien que témoignant d'une qualité de discrimination assez moyenne, elle est du même ordre de grandeur que les AUC sur apprentissage et validation. Cela confirme ainsi la relative robustesse du modèle.

5.2.1 Application du modèle au stock du 31/12/2017

Malgré ces résultats assez décevants, nous avons testé le modèle sur le stock de polices MRH au 31/12/2017 (via l'échantillon hors-du-temps). Nous avons comparé les probabilités moyennes et médianes obtenues par le modèle construit en amont avec les taux de résiliation observés au 31/12/2018. Le tableau 5.1 synthétise

les résultats obtenus pour quelques variables.

Variable	Modalité	Tx moyen estimé	Tx médian estimé	Tx observé
Nb de pièces	**mod1**	18,34%	15,34%	18,15%
	mod2	11,34%	9,24%	10,48%
	mod3	10,40%	9,60%	9,33%
	mod4	8,65%	8,25%	7,80%
	mod5	10,01%	9,75%	8,13%
	mod6	8,45%	8,32%	6,88%
Mnt cotisation nette	**mod1**	16,21%	11,70%	13,65%
	mod2	10,73%	7,96%	10,93%
	mod3	11,38%	9,31%	10,88%
	mod4	10,71%	9,37%	9,63%
	mod5	9,47%	8,67%	8,43%
	mod6	9,07%	8,80%	7,42%
	mod7	8,45%	8,36%	6,84%
Mnt garanti	**mod1**	8,92%	8,50%	8,08%
	mod2	17,50%	14,51%	16,93%
	mod3	11,17%	9,52%	9,55%
	mod4	9,06%	8,43%	8,06%
	mod5	8,85%	8,66%	6,94%
Maturité	**mod1**	15,26%	11,48%	14,36%
	mod2	12,83%	10,58%	10,83%
	mod3	10,14%	8,95%	9,11%
	mod4	8,84%	7,94%	7,69%
	mod5	7,36%	6,55%	7,04%
Nb avenants	**mod1**	11,63%	9,64%	10,20%
	mod2	9,75%	8,33%	8,83%
	mod3	9,06%	7,73%	8,59%
	mod4	8,77%	7,49%	9,22%
	mod5	8,77%	7,65%	8,23%
Portefeuille MRH		**10,73%**	**9,08%**	**9,61%**

Table 5.1 – Application du modèle au stock du 31/12/2017

Si nous choisissons le taux moyen comme indicateur de prédictivité, alors le modèle semble surestimer le phénomène de résiliation. En revanche, si nous choisissons le taux médian, la conclusion s'inverse.
Nous observons que le taux médian est plus proche de la réalité que le taux moyen. Cependant, comme nous ne souhaitons pas sous-estimer le risque de résiliation des clients, nous choisissons l'indicateur « taux moyen » pour la comparaison qui va suivre.

5.2.2 Comparaison avec les résultats obtenus par les modèles de durée

Le tableau 5.2 compare les différents modèles testés (régression et modèles de durée du chapitre 3) sur quelques variables complémentaires.

Le modèle logistique construit semble ainsi meilleur, globalement, que les modèles de durée (Cox et AFT).

Variable	Modalité	Modèle logistique	Modèle de Cox	Modèle AFT	Tx observés 31/12/2018
Frq de paiement	**Annuelle**	10%	12%	13%	9%
	Mensuelle	13%	12%	13%	11%
	Autres	11%	10%	10%	9%
Portefeuille MRH		**11%**	**12%**	**13%**	**10%**
Qualité jur.	**Locataire**	22%	24%	22%	22%
	PNO	9%	10%	10%	8%
	Propriétaire	9%	9%	9%	7%
Portefeuille MRH		**11%**	**11%**	**12%**	**10%**
Type de bien	**Appartement**	16%	20%	21%	13%
	Maison	9%	10%	11%	8%
	Autres	9%	13%	9%	9%
Portefeuille MRH		**11%**	**12%**	**13%**	**10%**

TABLE 5.2 – Comparaison de la prédictivité des différents modèles.

5.3 Autres méthodes de segmentation

Afin de tenter de remédier à la faible performance du score sur stock, nous pouvons tester d'autres méthodes de classification des individus.
Nous allons nous intéresser à deux méthodes complémentaires, toutes deux basées sur des techniques de rééchantillonnage permettant l'agrégation de modèles. Ces techniques, pas si récentes que cela, sont de plus en plus populaires en raison de l'engouement actuel pour le « *big data* » et le « *machine learning* ». La massification des données récoltées, couplées à l'explosion de l'économie numérique, a en effet permis d'augmenter les espaces de stockage des données ainsi que la vitesse des calculs.
La contrepartie inhérente à ces techniques est une perte de lisibilité quant au modèle final retenu. Certaines de ces techniques peuvent donner l'impression de fonctionner comme une boîte noire. Sans rentrer dans des détails théoriques trop complexes, nous allons expliquer les fondements du *bagging* et du *boosting*.

5.3.1 Bagging

Le *bagging*[3], présenté par Leo Breiman en 1996, consiste à moyenniser les résultats obtenus sur différents modèles bootstrappés.
L'objectif est surtout d'améliorer la robustesse d'un modèle, en réduisant sa variance dans l'échantillon de validation.
En effet, nous avons vu dans la section 4.5.3, page 63, que l'erreur quadratique moyenne pouvait se décomposer en termes de biais et de variance. Les échantillons bootstrappés permettent de préciser encore un peu plus la formule (4.1).
Nous considérons B échantillons bootstrappés, permettant chacun d'obtenir une estimation $\hat{\theta}_k$ de la quantité θ. Formellement, le *bagging* consiste à calculer un estimateur de $\hat{\theta}$, nommé $\hat{\theta}_B$, de la manière suivante :

$$\hat{\theta}_B = \frac{1}{B}\sum_{k=1}^{B}\hat{\theta}_k$$

Si nous considérons que les estimateurs $\hat{\theta}_k$ sont i.i.d. (de variance σ^2), alors nous pouvons écrire :

$$\mathbb{E}[\hat{\theta}_B] = \mathbb{E}[\hat{\theta}_1] \quad \text{et} \quad \mathbb{V}(\hat{\theta}_B) = \frac{1}{B}\mathbb{V}(\hat{\theta}_1) = \frac{\sigma^2}{B}$$

Le biais de l'estimateur ne bouge pas, tandis que sa variance diminue avec B. Mais malheureusement les $\hat{\theta}_k$ sont corrélés entre eux car ils proviennent du même échantillon : nous devons donc développer un peu plus la formule.
Nous considérons donc que le coefficient de corrélation (linéaire) entre les $\hat{\theta}_k$ est égal à $\rho > 0$.

3. Association des termes Bootstrap et AGGregatING.

Les $\hat{\theta}_k$ sont toujours supposés avoir la même loi : le biais de l'estimateur agrégé ne change donc pas. Par contre, la variance s'écrit désormais :

$$\mathbb{V}(\hat{\theta}_B) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Comme $\hat{\theta} = \lim_{B\to\infty} \hat{\theta}_B$, nous admettrons que $\mathbb{V}(\hat{\theta}) = \lim_{B\to\infty} \mathbb{V}(\hat{\theta}_B)$ et donc que $\mathbb{V}(\hat{\theta}) = \rho\sigma^2$.
Donc si $\rho < 1$, alors la variance de l'estimateur baggé est plus faible que celle des estimateurs agrégés [4].

Historiquement, cette méthode est surtout appliquée aux arbres de décisions (trop sensibles au surapprentissage), mais elle est aussi valable dans le cadre d'autres types de modèles [5]. Nous allons l'illustrer pour des modèles de régression logistique.
L'échantillon de départ est tout d'abord divisé en B échantillons par tirages aléatoires simples avec remise. Dans chacun des B échantillons bootstrappés, un modèle de régression logistique est ajusté. Toutes les variables candidates sont présentes au départ et seule leur sélection finale peut changer :

- **Bagging classique** : toutes les variables explicatives sont conservées.
- **Bagging stepwise** : sélection de variables [6] afin de ne conserver que les variables pertinentes (au niveau α).
- **Forêts aléatoires** : sélection aléatoire de variables explicatives. Cette méthode nécessite de fixer préalablement leur nombre. En général, pour p variables explicatives, il est admis d'en choisir $p/3$ en régression et $\sqrt{p}$ en classification.

Le modèle final est un vecteur de paramètres qui est la moyenne de la valeur des coefficients conservés dans chacun des B modèles, pour chacun des types de *bagging*.

5.3.2 Boosting

La dernière méthode d'agrégation de modèles que nous présenterons s'appelle le *boosting*. Sa théorisation a été popularisée par Robert Schapire et Yoav Freund à partir de 1996. Cette méthode propose, tout comme le *bagging*, de nombreuses variantes. Nous allons en expliquer le principe général.
Contrairement au *bagging*, les étapes du *boosting* s'effectuent séquentiellement. Il n'est donc malheureusment pas possible de paralléliser les calculs, ce qui implique un temps de traitement plus long.
Comme en régression classique, l'objectif est de minimiser une perte entre les données observées et les données modélisées. Mais contrairement à la régression, nous ne nous intéresserons pas forcément à la perte quadratique mais à d'autres fonctions de perte convexes, à paramétrer à l'initialisation du modèle.
À la première itération, un modèle est ajusté aux données. Les données mal classées vont ensuite être pondérées différemment des données bien classées.
À la seconde itération, un autre modèle est ajusté aux données en tenant compte des données mal classées précédement (grâce à leur pondération). Une nouvelle pondération des nouvelles données mal classées a alors lieu, et ainsi de suite.
Techniquement, à chaque itération, l'algorithme va chercher à minimiser la perte qui caractérise les mauvais classements de l'étape précédente. Le processus se réplique ainsi jusqu'au nombre d'itérations spécifiées lors de l'initialisation de l'algorithme. Sa stratégie est donc adaptative et non pas aléatoire comme le *bagging*.
Les schémas de la FIGURE 5.2 nous montrent de façon instinctive le principe de fonctionnement de la méthode. À chaque étape, les individus mal classés sont entourés. Au final, les modèles se complètent ou s'élaguent afin d'améliorer la prédiction.

Un des algorithmes de *boosting* les plus populaires s'appelle **Real AdaBoost** [7]. Son fonctionnement est présenté en annexe E, page 134.

La surpondération des individus mal classés ou mal ajustés se fait à l'aide d'un fonction de perte. Cette fonction, que l'on cherche à minimiser, possède généralement une forme convexe (afin d'éviter une multiplication des minima locaux lors de son optimisation).

4. Prendre B grand ne génère pas de surapprentissage car le biais reste le même.
5. Ou classifieurs.
6. Selon la méthode stepwise. Nous aurions pu tester la méthode backward ou forward également.
7. Dans le cadre d'une classification, l'algorithme **Discrete AdaBoost** sera privilégié à **Real AdaBoost** (qui s'applique généralement dans le cas d'une régression).

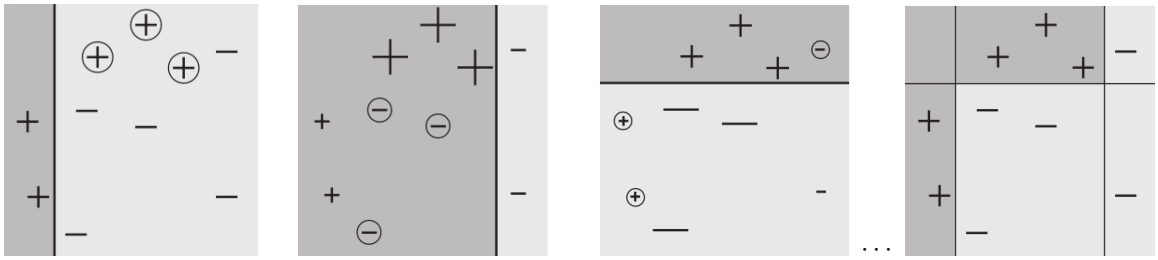


FIGURE 5.2 – Principe du *boosting* (source : [SF12]).

La plus connue est sans doute la fonction de perte **quadratique**, qui permet de solutionner le modèle de régression multilinéaire classique.
La FIGURE 5.3 nous en présente quelques autres, notamment celle choisie dans l'algorithme adaboost : la fonction **exponentielle**. Cette dernière pénalise plus fortement les individus mal modélisés (i.e. dont la valeur y.F est strictement négative dans le graphique) que les autres. En contrepartie, elle peut ne pas bien fonctionner en présence d'individus trop extrêmes.

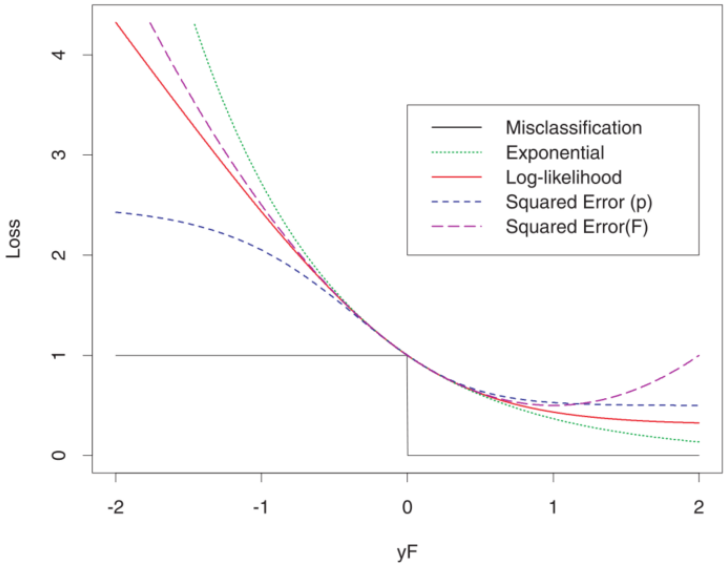


FIGURE 5.3 – Exemples de fonctions de pertes (source : [FHT00]).

L'ajout d'un dernier paramètre de pénalisation λ (ou de shrinkage) dans la fonction de perte, permet, lorqu'il est strictement inférieur à 1, d'atténuer la surpondération des individus mal classés (l'algorithme cherchera alors à optimiser λ.y.F à la place de y.F).

Ainsi le *bagging* est censé réduire la variance, le *boosting* quant à lui est censé réduire à la fois le biais et la variance des prévisions. Nous allons vérifier ces propriétés lors des applications qui vont suivre [8].

5.3.3 Application à nos données

Il convient de préciser que le biais est censé être représenté par l'AUC et la variance par l'écart-type des AUC obtenus.

Application du bagging Afin d'obtenir des résultats robustes, nous avons extrait par bootstrap 1 000 bases de constructions. Nous avons testé deux stratifications de la variable cible : soit la même répartition de

8. Pour des raisons de clarté, nous avons choisi de ne pas nous intéresser explicitement à l'importance des variables explicatives sélectionnées dans les différents modèles.

non résiliés-résiliés qu'observé dans l'échantillon réel (environ 90%-10%), soit une équidistribution (50%-50%) des deux risques. Chaque base est constituée de trois types d'échantillons [9].

- **Apprentissage** : 10 000 observations si répartition 90-10, 2 000 observations si répartition 50-50.
- **Validation** : 5 000 observations si 90-10, 1 000 observations si 50-50.
- **Hors-Du-Temps** : 10 000 observations si 90-10, 2 000 observations si 50-50.

Nous remarquons qu'il aurait pu être intéressant de faire varier **encore plus** la part des positifs/négatifs (voire leur volumétrie) dans les bases de construction afin d'observer l'évolution du pouvoir discriminant des modèles.

En plus du *bagging* classique et du *bagging stepwise*, nous avons testé deux forêts aléatoires, notées RF pour *Random Forest*, avec les paramètres suivants :

- RF14 : Nombre d'arbres = 1 000 et nombre de variables [10] = 14 ($\approx \sqrt{201}$).
- RF67 : Nombre d'arbres = 1 000 et nombre de variables = 67 ($\approx \frac{201}{3}$).

Le paramétrage du nombre de variables optimales peut se faire par simulation. En raison des importants temps de traitement, nous ne l'avons pas effectué dans ce mémoire. Nous nous sommes contentés des paramètres conseillés par défaut en classification et régression.

Le tableau 5.3 synthétise les résultats des performances obtenues sur les trois échantillons :

	Répartition 90-10			**Répartition 50-50**		
	App	**Val**	**Oot**	**App**	**Val**	**Oot**
BAGGING classique	**0,689**	**0,635**	0,628	**0,711**	0,623	0,618
BAGGING stepwise	0,647	0,625	0,621	0,649	0,618	0,614
BAGGING type RF (14)	0,606	0,596	0,594	0,611	0,595	0,592
BAGGING type RF (67)	0,668	0,633	**0,629**	0,683	**0,625**	**0,621**

TABLE 5.3 – Moyennes des AUC (bagging).

Sans surprise et quelle que soit la stratification retenue, en apprentissage l'AUC est maximisée sur le *bagging* classique. Par contre, la forêt aléatoire avec 67 variables est celle pour laquelle les performances sont les meilleures pour le test hors-du-temps. En validation, les meilleurs résultats, bien qu'assez proches, sont partagés entre le *bagging* classique et la forêt RF 67.
La forêt aléatoire avec 14 variables fait office de **classifieur faible** : les performances observées sont légèrement meilleures que le hasard [11].

Sur le tableau 5.4, nous constatons que le *bagging stepwise* semble minimiser les écarts-type sur les échantillons de validation et hors du temps. Les modèles ainsi créés sont sans doute plus homogènes en eux, ce qui laisse supposer que les variables pertinentes ont tendance à se répéter.

	Répartition 90-10			**Répartition 50-50**		
	App	**Val**	**Oot**	**App**	**Val**	**Oot**
BAGGING classique	**0,0097**	0,0157	0,0119	**0,0113**	0,0203	0,0143
BAGGING stepwise	0,0138	**0,0152**	**0,0114**	0,0171	**0,0190**	**0,0135**
BAGGING type RF (14)	0,0196	0,0231	0,0232	0,0207	0,0252	0,0256
BAGGING type RF (67)	0,0100	0,0160	0,0159	0,0123	0,0193	0,0199

TABLE 5.4 – Ecarts-type des AUC (bagging).

Remarque sur le redressement des échantillons non représentatifs Deux stratifications (90-10 et 50-50) ont été testées dans notre modèle logistique [12]. Les résultats obtenus semblent très proches. Voici la

9. Dont l'intérêt est décrit en amont, page 63.
10. Nous avons $p = 201$ variables indicatrices candidates aux modèles.
11. Cependant, les performances des autres types de *bagging* ne sont pas beaucoup meilleures.
12. Lorsque que le risque à modéliser est rare, une répartition 50-50 peut être intéressante.

distribution des AUC que l'on observe sur les échantillons hors-du-temps :

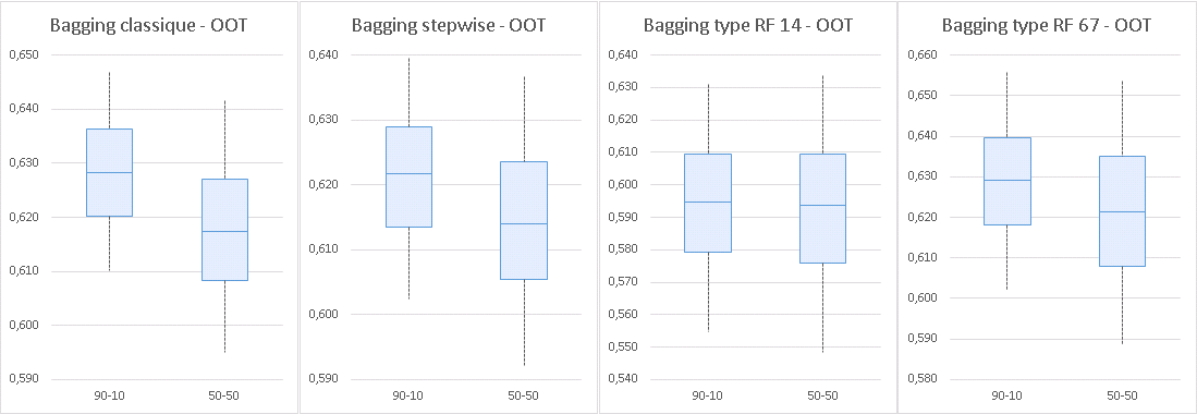


FIGURE 5.4 – Comparaison de la distribution des AUC pour 90-10 et 50-50.

Les *boxes plot* obtenues, dont les extrémités sont représentées par le 5ème et le 95ème centile, laissent penser que la stratification de la variable cible n'a pas d'impact significatif sur les performances des modèles.
Si on note C le logit obtenu avec des données de construction non représentatives, et C^* le logit théorique obtenu sur un échantillon représentatif, alors on montre que (voir [Ten07] pour les données cas-témoins) :

$$C^* = -\ln \frac{n_+}{n_-} + \ln \frac{p}{1-p} + C$$

Où n_+ (resp. n_-) est le nombre de polices résiliées (resp. non résiliées) dans la base de construction, et p le taux de résiliation dans la population globale.
Nous voyons donc que la correction apportée concerne uniquement la constante du modèle logistique et non les coefficients associés aux variables explicatives.
Le redressement de la constante est à privilégier lorsque l'on souhaite comparer plusieurs modèles ayant la même définition du risque à prévoir, car il uniformise ainsi la probabilité estimée.
Dans notre cas d'étude, nous nous intéressons au classement des données via la courbe ROC dont découle l'AUC. Or, nous voyons bien que le redressement de la constante n'agit pas sur le classement des données (en fonction de la probabilité d'occurrence du risque).
Les résultats de nos calculs tendent ainsi à confirmer qu'un échantillonnage non représentatif de la variable cible n'a pas vraiment d'impact sur les AUC baggués.

Ainsi, si l'apport du *bagging* est important sur un classifieur instable tel qu'un arbre CART, il est en revanche admis (cf. [Tuf17]) qu'il n'apporte pas grand chose de plus en termes de performances sur un classifieur robuste tel qu'une régression logistique. C'est précisément ce qui est observé dans nos données : les moyennes des AUC sur les deux échantillons hors construction ne sont pas meilleures que celles observées sur le modèle de départ (pour rappel : 0,64 en validation et 0,63 en hors-du-temps, cf. page 73).
Ces résultats semblent confirmer un peu plus une de nos hypothèses de départ : l'apport de variables prédictives prime sur la technique de modélisation.

Application du boosting Nous avons effectué un **boosting de modèles logistiques**. Sa mise en œuvre est inspirée de la fonction **BoostLog** décrite dans le chapitre 14 de l'ouvrage [Tuf17] [13]. La sélection des variables est faite selon la méthode **forward** couplée à l'optimisation du critère **BIC** [14].
Nous rappelons que la méthode consiste à itérer un algorithme de classement ou de régression à partir d'un échantillon d'apprentissage initial qui est modifié à chaque pas afin de prendre en compte les erreurs de classement ou de prédiction de l'étape qui précède.
Les observations mal prédites sont ainsi surpondérées afin d'avoir plus de chance d'être sélectionnées dans le rééchantillonnage de l'étape suivante.

13. La plupart des packages R disponibles (lors de notre étude) et permettant d'effectuer des *boostings* travaillent en majorité avec des classifieurs de type **arbres de décision**.
14. Nous préférons le critère BIC au critère AIC car la pénalisation est croissante avec le nombre de variables **et d'individus**, ce qui peut être intéressant lorsque l'on souhaite obtenir un modèle prédictif.

L'apprentissage se focalise donc, au fur et à mesure, sur les données les plus atypiques. En parallèle, afin de contrôler le surapprentissage, chaque modèle est testé sur un échantillon de validation et un échantillon hors-du-temps.

Afin d'éviter d'avoir, d'une part un temps de calcul trop important (en partant de la base entière), d'autre part des résultats peu robustes (avec une seule base), nous avons décidé d'effectuer 10 cas de *boostings* (avec 3 000 itérations chacun) en choisissant aléatoirement 10 bases parmi les 1 000 bases de construction ayant servi à notre application du *bagging*.

En plus du modèle sans paramètre de pénalisation ($\lambda = 1$), quatre autres valeurs du paramètre ont été testées sur chaque échantillon. La synthèse des résultats est présentée ci-dessous.

Le tableau 5.5 synthétise, à l'aide de leur moyenne, l'ensemble des indicateurs retenus pour les 10 échantilons boostés (il s'agit en quelque sorte d'un *bagging* de *boosting*).

Shrink λ	**AUC maxi Val**	**AUC maxi Oot**	**AUC Val**	**AUC Oot**
0,0001	0,6298	0,6313	**0,6245**	**0,6271**
0,001	0,6307	0,6292	0,6238	0,6266
0,01	0,6302	0,6311	0,6200	0,6262
0,1	**0,6314**	0,6317	0,6190	0,6220
1	0,6309	**0,6350**	0,5917	0,5754

TABLE 5.5 – Moyennes des AUC (boosting).

Nous avons, dans un premier temps, considéré (pour chaque valeur λ ou shrink) les AUC maximales sur chacun des deux échantillons de test : validation et hors-du temps. Ces valeurs maximales sont atteintes pour un nombre d'itérations variable selon les échantillons.
Dans le tableau E.5, détaillé en annexe page 138, nous pouvons reconstituer, pour chacun des 10 échantillons, l'itération à partir de laquelle cette aire est maximale.
Un exemple graphique, assez représentatif de la convergence, est proposé page 137 afin de juger de la stabilité des AUC en fonction des itérations. En général (sauf pour le modèle sans pénalisation), l'aire maximale est atteinte assez rapidement, puis les performances restent stables au fil des itérations.

Le premier constat qui s'impose est que les performances mesurées sur l'échantillon hors-du-temps sont légèrement meilleures que celles mesurées sur l'échantillon de validation (sauf pour $\lambda = 0,001$).
Les AUC semblent maximisées pour les valeurs de shrink assez faibles ($\lambda = 1$ ou $0,1$).
Le second constat est que les AUC observées dans ces *boosting* sont meilleures que celles observées précédemment lors des *baggings* (pour rappel, dans la répartition 50-50 : 0,6248 en val et 0,6214 en oot).

Dans un second temps (colonnes bleues), nous nous sommes intéressés aux AUC moyens, sur l'ensemble des 3 000 itérations effectuées.
Ici encore, les AUC sont meilleures sur l'échantillon hors-du-temps (sauf pour le modèle sans pénalisation).
L'AUC présentée est l'aire finale mesurée sur l'échantillon boosté suite aux 3 000 itérations, ce qui explique que le niveau est inférieur à celui mesuré sur les deux précédentes colonnes du tableau. Nous observons, cependant, que l'AUC moyenne sur validation (0,6245) est quasiment du même niveau que le meilleur mesuré lors des *baggings* (0,6248). Il est meilleur, toutes méthodes confondues, sur l'échantillons hors-du-temps (0,6271 contre 0,6214).

Nous pouvons ainsi raisonnablement penser que le *boosting* permet de réduire le biais de la prédiction des modèles. Mais que pouvons-nous dire de sa variance ? Nous allons tenter d'y répondre en commentant le tableau 5.6, qui présente les écarts-type des AUC obtenues précédemment.

Bien que le nombre d'échantillons ne soit pas comparable avec celui des *baggings* (10 contre 1 000), les écarts-types sont inférieurs à ceux observés en test lors des différents *baggings*, ce qui laisse penser que le *boosting* permet bien de réduire la variance des prédictions.

Shrink λ	**AUC maxi Val**	**AUC maxi Oot**	**AUC Val**	**AUC Oot**
0,0001	0,0118	0,0108	**0,0126**	0,0149
0,001	0,0138	0,0140	0,0136	0,0144
0,01	0,0111	0,0142	0,0127	0,0127
0,1	0,0103	0,0118	0,0195	**0,0119**
1	**0,0099**	**0,0097**	0,0324	0,0365

TABLE 5.6 – Ecarts-type des AUC (*boosting*).

5.4 Élasticité du taux de résiliation à la prime

5.4.1 Généralités

Définition générale

En (micro)économie, l'**élasticité** de la demande d'un bien ou d'un service par rapport à son prix [15] est calculée afin de mesurer la sensibilité de la demande aux variations du prix. L'objectif est ainsi de chiffrer l'ampleur d'une variation de la demande suite à une variation du prix. Nous pouvons donc la définir ainsi :

$$\text{Élasticité-prix de la demande} = \frac{\text{taux de variation de la demande}}{\text{taux de variation du prix}}$$

Formellement, nous notons Q la quantité et P le prix du bien sous-jacent, l'élasticité E est par définition :

$$E = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} = \frac{\Delta Q}{\Delta P} \times \frac{P}{Q} \tag{5.1}$$

L'élasticité-prix [16] est en général négative, car la demande ou la consommation d'un bien diminue dès lors que son prix augmente, et à l'inverse augmente lorsque son prix baisse.
Il est également admis que la demande est d'autant plus élastique que l'horizon temporel est lointain.

Interprétation

- Si l'élasticité est **nulle**, cela signifie que la variation du prix (à la hausse comme à la baisse) n'a aucun impact sur la demande du bien. C'est le cas pour les biens de première nécessité (comme les médicaments, le pain, l'eau, etc.). Ces biens sont appelés **biens de Giffen** [17].
- Nous avons vu qu'en général l'élasticité-prix était **négative**. Dans ce cas, nous allons nous intéresser à son degré d'élasticité par rapport à la valeur absolue 1 :
  - Si l'élasticité est comprise entre -1 et 0 : le bien est **faiblement élastique** car la quantité demandée diminue relativement moins vite que le prix (comme la consommation de carburant par exemple).
  - Si l'élasticité est égale à -1 : l'élasticité est dite **unitaire** car la quantité demandée varie proportionnellement au prix du bien (comme la plupart des biens de consommation courante).
  - Si l'élasticité est inférieure strictement à -1 : le bien est **fortement élastique** car la quantité demandée diminue plus vite que le prix (comme les séjours touristiques par exemple).
- Si l'élasticité est **positive**, cela signifie que la demande augmente lorsque le prix augmente. Ce phénomène paradoxal est observé surtout sur les biens de luxe, appelés **biens de Veblen** [18].

15. Également appelée l'**élasticité-prix de la demande**.
16. D'autres types d'élasticités existent : élasticité-prix de l'offre, élasticité-revenu, élasticité-prix croisée, etc.
17. Du nom de l'économiste écossais Robert Giffen (1837-1910).
18. Du nom de l'économiste et sociologue américain Thorstein Veblen (1857-1929).

5.4.2 Application aux contrats d'assurance

La notion d'élasticité est potentiellement pertinente dans le cadre d'une étude sur les résiliations de contrats d'assurance. Les primes sont généralement revues tous les ans [19] afin de réajuster les conditions tarifaires sur les périmètres suivants :

- Les **affaires nouvelles** : nouveaux contrats du portefeuille.
- Le **stock** de contrats en cours.
- Les contrats **sous surveillance** : contrats en cours multi-sinistrés par exemple.

Pour les dépenses dites « pré-engagées » telles que les contrats d'assurance, il est admis que l'élasticité est quasi-nulle à court terme, mais pas à long terme. **Nous allons vérifier la propriété attendue à court terme [20] à partir des données MRH de notre portefeuille.**

Notons que l'élasticité peut être envisagée à partir de la mesure de deux types de prix :

- Le montant de la prime émise (annualisée).
- La variation de la prime entre deux dates : $N-1$ et N par exemple.

Attention, contrairement à la demande ou à la rétention de contrats, lorsque nous étudions le taux de résiliation, nous nous attendons à observer une élasticité **positive**. En effet, nous supposons implicitement qu'une augmentation de la prime [21] va entraîner une hausse des résiliations.

Application à nos données

Pour le modèle logistique de l'élasticité à la prime, présenté en équation (E.1) à partir de la page 139, nous avons considéré les stocks entre 2014 et 2017. Pour celui relatif à la variation de prime (également présenté en annexe, page 140), nous avons considéré les stocks entre 2015 et 2017. La distribution de la prime ou de sa variation, sur chacune les deux modalités cibles est décrite dans les graphiques en FIGURE 5.5.

Nous observons ainsi que le montant de prime est en moyenne moins élevé sur les contrats résiliés (213 €) que sur les contrats non résiliés (247 €). De la même façon, la variation de la prime est un peu moins élevée sur les contrats résiliés ($+3,37\%$) que sur les contrats non résiliés ($+3,48\%$).
A priori, une prime élevée ou une variation de prime plus forte semble permettre de conserver les clients.

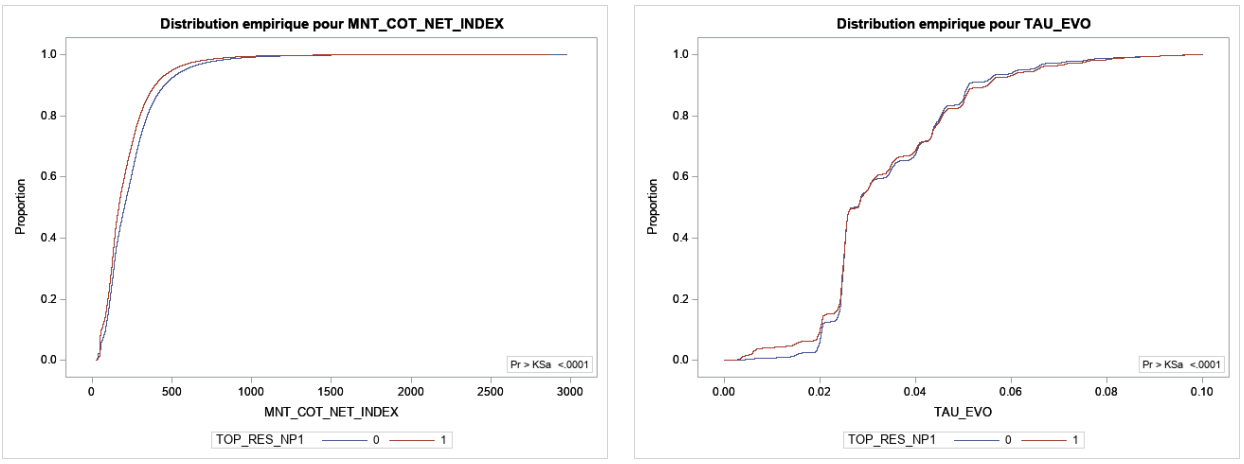


FIGURE 5.5 – Distribution des variables Primes et Variation

19. La stratégie de fixation des nouvelles primes revient au comité de tarification. L'objectif est essentiellement de préserver l'équilibre du portefeuille dans la durée longue.
20. Horizon annuel.
21. Ou une forte variation entre $N-1$ et N.

Un modèle logistique a ensuite été ajusté à partir d'un bonne partie des variables conservées dans le modèle sur le stock. Les résultats de la significativité des variables sont synthétisés en annexe page 140 dans le tableau E.6, pour le modèle **Prime**.

Ces deux modèles (Prime et Variation) ont donc permis de calculer une élasticité pour chaque contrat (ainsi qu'une probabilité de résilier dans l'année à venir). Nous sommes ainsi en mesure de proposer deux distributions de l'élasticité moyenne du portefeuille MRH (cf. FIGURE 5.6).

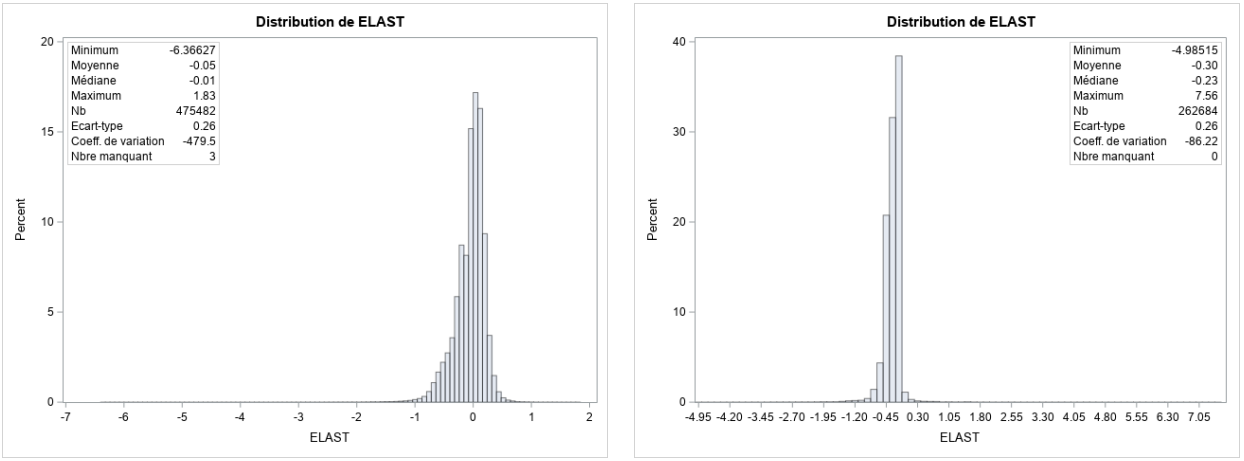


FIGURE 5.6 – Distributions de la variable Elasticité sur les modèles Prime et Variation

A ce stade, deux remarques peuvent être faites :

- L'elasticité est en moyenne légèrement négative : une augmentation de la prime entrainerait donc une diminution des résiliations. Ceci est en contradiction avec le dogme de départ : « la résiliation augmente avec le montant de prime ». Cependant, ce n'est pas en contradiction avec l'étude bivariée pour la variable cotisation : la résiliation diminue avec la hausse de son montant.
- L'élasticité est très proche de zéro au niveau de la médiane : une augmentation de la prime n'aurait donc pas d'effet significatif sur les résiliations. Ce qui confirme une des hypothèses de départ : « l'élasticité est quasi-nulle à court terme ».

Le résultat qui va suivre nous interpelle. Pour la variable « qualité juridique », l'élasticité moyenne par modalité est la suivante :

Variable d'analyse : ELASTICITE		
Qualité juridique	**Nombre**	**Moyenne**
Locataire	72 481	-0,2753
PNO	166 108	-0,0045
Propriétaire	236 890	-0,0212

TABLE 5.7 – Elasticité (Prime) en fonction de la qualité juridique.

Nous observons que l'élasticité moyenne est beaucoup plus faible pour les locataires, ce qui peut sembler étrange car nous nous attendions plutôt à un résultat inverse. L'explication est relativement simple et doit inciter l'actuaire à prendre garde aux **biais cognitifs** [22] et en particulier au **biais de confirmation** [23]. L'élasticité estimée semble d'autant plus faible que la différence entre la prime moyenne des contrats (résiliés ou pas) est élevée en faveur des non résiliés. L'écart relatif est de l'ordre de 16% pour les locataires, 0% pour les PNO et 3% pour les propriétaires.

22. Les biais cognitifs conduisent le sujet à accorder des importances différentes à des faits de même nature et peuvent être repérés lorsque des paradoxes ou des erreurs apparaissent dans un raisonnement.
23. Ce biais consiste à privilégier les informations confirmant ses idées préconçues et/ou à accorder moins de poids aux informations jouant en défaveur de ses conceptions.

Ainsi, et pour rejoindre une des remarques de [JR12], **les modèles d'élasticité proposés ne semblent guère réalistes dans notre cas d'étude**.

Il convient toutefois de nuancer ces résultats car notre étude comporte de nombreuses limites. D'une part, le risque est uniquement mesuré sur les contrats MRH du marché des particuliers. D'autre part, l'environnement concurrentiel n'est pas analysé. Il n'est pas absurde de penser que ces deux aspects soient liés. La mutuelle, à laquelle sont rattachés les contrats étudiés, applique globalement des évolutions tarifaires mesurées et en règle générale en dessous de celles pratiquées sur le marché. Dans ce contexte, le client ne va pas forcément trouver un meilleur tarif ailleurs. Le niveau des garanties et de flexibilité du contrat, assez bon dans la compagnie, peut également expliquer les résultats obtenus.

L'étude de la résiliation à horizon annuel, à partir d'un stock de contrats en cours, va ainsi se finaliser dans le prochain chapitre. Nous allons présenter la problématique des résiliations dans le contexte réglementaire « Solvabilité 2 ».

Chapitre 6

La résiliation dans Solvabilité 2

L'estimation du **taux de chute** [1] est un enjeu important pour l'entreprise lors du calcul annuel de la Meilleure Estimation des provisions pour primes (pilier 1) [2]. Sans rentrer dans les détails techniques du calcul, les limites suivantes émergent :

- Un seul taux global est appliqué sur l'ensemble du portefeuille pour l'année N.
- Des contraintes dans le processus de production des flux Solvabilité 2 imposent que ce taux soit calculé, au plus tard, en tout début d'année $N+1$ (avant $J+3$, soit moins de trois jours ouvrés après le 31/12).

Nous proposons donc deux axes d'améliorations possibles :

- Distinguer un taux pour chaque assiette de primes définie par la **frontière des contrats**.
- Calculer ces taux dans l'année N, ce qui implique le passage d'une logique de calcul a posteriori à une logique a priori et la nécessité de s'assurer que les ordres de grandeur des résultats des deux logiques ne soient pas trop éloignés les uns des autres.

6.1 Rappels sur Solvabilité 2

La directive européenne dont découle la réglementation dite « Solvabilité 2 » invite les assureurs et réassureurs, via la valorisation du bilan prudentiel, à :

- **Estimer les engagements futurs attendus** qu'ils soient potentiels ou avérés, qu'elle a vis-à-vis de ses assurés et qui sont matérialisés par les **provisions techniques**.
- **Estimer les engagements futurs inattendus** ou exceptionnels qu'elle pourrait avoir vis-à-vis de ses assurés à horizon d'un an [3] avec une probabilité (très) faible et qui sont matérialisés par le SCR ($\mathbb{P}[\text{Faillite}] < 0,5\%$) et sa vision moins restrictive qu'est le MCR ($\mathbb{P}[\text{Faillite}] < 15\%$). Nous précisons que ces deux indicateurs de solvabilité dépendent des provisions techniques calculées en amont.

La valeur de ces engagements est notamment fonction des volumes de primes des contrats en cours au moment de l'inventaire. Ce périmètre va être précisé dans les points suivants.

6.1.1 La frontière des contrats

La décomposition des contrats d'assurance peut être envisagée sous deux axes complémentaires :

- La segmentation des contrats, et en particulier des garanties associées.
- La notion de **frontière des contrats** [4].

Tout d'abord, le **principe de prééminence de la substance sur la forme** implique de classer les contrats sur des critères de risque intrinsèque et non plus sur des critères juridiques [5]. Cela a pour consé-

1. En référence au contexte Solvabilité 2, nous pourrons parler de taux de « chute » à la place de « résiliation ».
2. Les Articles 26 et 30 du Règlement Délégué 2015/35 de la commission du 10 octobre 2014 précisent que le comportement attendu des assurés doit être pris en compte dans l'évaluation de la Meilleure Estimation des provisions.
3. Fréquence de la supervision.
4. Conformément à l'article 18 du Règlement Délégué 2015/35 de la commission du 10 octobre 2014.
5. Conformément à l'article 55 du Règlement Délégué 2015/35 de la commission du 10 octobre 2014.

quence d'homogénéiser les résultats des modèles de provisionnement. Opérationnellement, la segmentation des types d'engagement doit se faire a minima par ligne d'activité [6].

Ensuite, la notion de **frontière** à l'intérieur de laquelle doivent être situés les contrats va permettre de quantifier le volume de primes à agréger. Cela implique de définir le périmètre temporel de ce qui doit être valorisé en termes d'engagements. Nous distinguons ainsi les **primes existantes** et les **primes futures**.

La prise en considération des **primes existantes** est comparable à ce qui est fait en normes comptables statutaires dites « S1 » pour le calcul des PPNA [7] et des PREC [8] (en non-vie). En revanche, la principale distinction entre normes S1 et S2 se fait au niveau des **primes futures**, car ces dernières ne sont pas intégrées dans le premier cas. La norme S2 intègre donc plus de primes dans son périmètre.
Bien que ces calculs quantitatifs soient effectués dans le cadre du pilier 1, et donc dans une logique de **run-off** (i.e. extinction de l'activité), les engagements liés aux **primes futures** rattachées aux contrats existants pour lesquels l'assureur ne peut agir unilatéralement doivent bien être intégrés au bilan S2.

La frontière des contrats va ainsi permettre de définir l'appartenance ou non d'une obligation d'assurance au périmètre de calcul de la **meilleure estimation des provisions**. Cette importante notion est abordée dans la section suivante.

6.1.2 Provisions techniques

Les **Provisions Techniques** sont la somme de la **meilleure estimation des provisions** [9] et de la **marge pour risque** [10]. La meilleure estimation des provisions distingue une partie relative aux risques avérés [11] (sinistres en cours à la date d'évaluation) et une partie relative aux sinistres potentiels [12] des contrats. La marge pour risque peut être considérée comme le montant, calculé dans une logique de coût d'immobilisation du capital, qu'un autre assureur demanderait pour reprendre et honorer les engagements d'assurance. Une insuffisance des provisions impacterait négativement l'image de l'entreprise. En effet, le respect de ses engagements à l'égard de ses assurés serait mis à mal et l'obligerait à puiser dans ses fonds propres (ce qui mettrait le management en position délicate vis-à-vis de son conseil d'administration ou de ses actionnaires). En résumé, sa bonne santé financière, mesurée par sa solvabilité, serait compromise.

6.2 Exemple illustratif

Plaçons-nous dans la situation d'une compagnie telle que Le Finistère Assurance, qui commercialise des contrats MRH à tacite reconduction avec un délai de préavis de résiliation de deux mois.
Nous rappelons que nous considérons les contrats en stock dans le portefeuille au 31/12 d'une année N. Ces contrats peuvent être à **échéance annuelle** ou à **échéances fractionnées** (mensuelles, trimestrielles ou semestrielles).

Aucune ambiguïté sur la frontière des contrats n'est à déplorer pour le BE de sinistres, car l'assureur est déjà engagé. En effet, une indemnisation est déjà en cours (ou devrait l'être dans le cas des sinistres déclarés tardivement).
Nous nous intéressons donc au **BE de primes** qui est, par définition, l'évaluation au plus juste (sans marge de prudence) des engagements liés aux **primes existantes** et aux **primes futures** associées aux contrats pour lesquels l'assureur n'a pas le droit unilatéral de résilier le contrat, de refuser le versement d'une prime ou de modifier librement la prime ou la garantie du contrat [13].
La compagnie est donc engagée au titre des primes futures, au $31/12/N$, sur les contrats des deux mois qui suivent (soit janvier $N+1$ et février $N+1$).

6. Ou LoB (Line of Business). Ces lignes d'activité sont à mettre en relation avec les *catégories ministérielles* associées aux différentes garanties des contrats, mais attention car il n'existe pas forcément de bijection entre ces deux notions.
7. Provisions pour Primes Non Acquises.
8. Provisions pour Risques En Cours.
9. Ou *Best Estimate* (BE).
10. Ou *Risk Margin* (RM).
11. Appelée meilleure estimation des provisions pour sinistres ou BE de sinistres.
12. Appelée meilleure estimation des provisions pour primes ou BE de primes.
13. Conformément à l'article 17 du Règlement Délégué 2015/35 de la commission du 10 octobre 2014.

Pour les montants de fractionnement restant à émettre [14] deux points de vue sont possibles : primes futures car non émises à la date d'intérêt ou primes existantes car le contrat est bien en cours à la date d'intérêt. Nous considérons que ces derniers font partie des contrats existants au même titre que les PNA [15].

Le point important à comprendre, dans le cadre de ce mémoire, est le suivant : à ces montants de primes est en général appliqué un **taux de chute** permettant de déduire, du volume de primes, les résiliations probables dans l'année à venir.
Le taux de chute peut varier selon l'assiette de primes considérée : contrats existants ou futurs.
La sous-section suivante va nous préciser la méthode de calcul appliquée jusqu'à présent.

6.3 Méthode actuelle d'estimation du taux de chute

La méthode appliquée actuellement par la mutuelle permet de calculer un taux de chute **global**, appliqué à l'ensemble du portefeuille (particuliers et professionnels). Il s'agit d'un taux calculé a posteriori, c'est-à-dire calculé à partir d'informations observées à la date de l'inventaire, dont nous allons décrire les étapes de calculs.

Au niveau métier, nous pouvons distinguer trois manières de résilier (cf. chapitre 1) :

- **Résiliation à l'échéance** : avant l'application de la loi Hamon, il s'agissait de la principale façon de résilier pour les souscripteurs. Nous distinguons d'une part la résiliation effectuée en tenant compte du délai de préavis (en général deux mois), d'autre part la résiliation n'en tenant pas compte.
  - Avec délai de préavis : le cas classique avant la loi Hamon (type 1).
  - Hors délai de préavis : suite à un contentieux par exemple (type 2).
- **Résiliation hors échéance** : depuis la loi Hamon, l'assuré peut résilier à tout moment son contrat au bout d'un an (type 3).

Ces trois types de résiliations donnent lieu à des calculs de taux. En considérant, pour chaque année d'exercice N, le ratio suivant :

$$\text{Taux de résiliation}_N^{\text{type i}} = \frac{\text{Nb de résiliations}_N^{\text{type i}}}{\text{Nb de polices}_{N-1}}, \qquad \forall i \in \{1, 2, 3\}$$

Nous avons vu précédemment que la frontière des contrats nous incite à distinguer deux assiettes de primes :

- **Contrats existants** : matérialisés par **les PNA et les échéances fractionnées**.
- **Contrats futurs** : matérialisés par **les échéances de janvier et février** $N + 1$.

Par définition, il a été considéré que :

- **Taux Résiliation (Existants)** = Taux de résiliation$^{\text{type 3}}$
- **Taux Résiliation (Futurs)** = Taux de résiliation$^{\text{type 2 et 3}}$

En effet, le taux de résiliation relatif aux primes **existantes** est assimilé au taux de résiliation hors échéance. Cela se justifie par la durée de l'engagement restante : si l'assuré dénonce son contrat avant la fin de l'engagement, alors il ne peut le faire qu'avant l'échéance anniversaire de la police [16].
Le même raisonnement est fait pour les primes **futures**, à la différence que le client peut résilier dans les deux mois qui précèdent l'échéance : nous tenons donc compte, en plus, du taux de résiliation à l'échéance (mais hors délai de préavis, car il reste moins de deux mois).

14. Jusqu'à la date d'échéance principale, pour les contrats à échéances fractionnées.
15. Primes Non Acquises.
16. S'il le fait à l'échéance, cela concerne des engagements n'appartenant pas à la frontière des contrats.

6.4 Evolution possible de la méthode actuelle

Les taux appliqués sont calculés dans une logique de calcul a posteriori : il faut attendre la fin de l'année pour pouvoir les calculer. Nous avons donc décidé de nous inspirer du score sur stock afin d'estimer ces deux taux a priori [17].
Conformément à ce qui est précisé en section 6.3, nous avons décomposé la variable cible initiale afin de la transformer en une nouvelle variable cible afin de savoir si la résiliation est effectuée à l'échéance ou hors échéance.
Nous avons ensuite ajusté un modèle de régression logistique **multinomial à réponses polytomiques** à quatre modalités :

- Résiliation à horizon 12 mois hors échéance.
- Résiliation à horizon 12 mois à l'échéance hors délai de préavis.
- Résiliation à horizon 12 mois autres cas.
- Pas de résiliation à horizon 12 mois.

Le modèle multinomial est présenté en annexe F, page 141.

Application à nos données Nous sommes partis de la même base de construction que pour le modèle sur stock du chapitre 5, afin d'estimer les différents taux de résiliation à horizon annuel.
Un premier modèle polytomique a tout d'abord été ajusté, suivi de trois modèles binaires (un pour chacun des types de résiliation définis).
Après avoir vérifié sur l'échantillon de validation la relative robustesse des modèles construits à partir des données d'apprentissage, nous les avons appliqués sur les données en stock au 31/12/2017.
Nous ne présenterons pas les différents paramètres en sortie de l'outil afin de ne pas perdre le lecteur dans des détails considérés comme secondaires ici.
La qualité des modèles sera appréciée de manière globale, en comparant le taux prédit en cible avec le taux réellement observé.

		Apprentissage		**Validation**		**Échantillon 2017**	
		Moy mod	**Taux obs**	**Moy mod**	**Taux obs**	**Moy mod**	**Taux obs**
Hors échéance	**polyto**	6,62%	6,48%	6,60%	6,53%	6,53%	7,18%
	binaire	6,85%		6,83%		6,78%	
Éch. hors délai	**polyto**	2,21%	2,23%	2,21%	2,18%	2,14%	1,65%
	binaire	2,32%		2,31%		2,25%	
Éch. avec délai	**polyto**	1,66%	1,59%	1,66%	1,54%	1,65%	0,78%
	binaire	1,56%		1,56%		1,56%	
Total résil.	**polyto**	**10,49%**	**10,30%**	**10,47%**	**10,25%**	**10,32%**	**9,61%**
	binaire	**10,73%**		**10,70%**		**10,59%**	

Table 6.1 – Résultats des modèles polytomiques et binaires.

Nous observons que les taux moyens obtenus en validation sont cohérents avec les taux réels, quoique légèrement plus élevés (pour chacun des modèles).
Les résultats obtenus sur l'échantillon 2017 sont mitigés. Globalement, les modèles surestiment la résiliation. Si nous regardons de plus près, nous constatons que les résiliations hors échéance sont largement sous-estimées tandis que les résiliation à l'échéance sont largement surestimées. Cela peut s'expliquer par la dynamique des taux, observées en Figure F.1 (page 141), qui montre une forte augmentation des résiliations hors échéance au détriment des résiliations à l'échéance. Cette dynamique s'explique par l'application de la loi Hamon (cf. section 1.2).
Ainsi, les deux modalités « A l'échéance avec/sans délai » sont estimées avec relativement peu d'observations et, comme la tendance est à la baisse, nous pouvons nous attendre à une mauvaise prédictivité pour

17. À partir du stock N-1.

les années à venir.

D'un point de vue prudentiel, surestimer la résiliation revient à diminuer le volume de primes dans la frontière des contrats et donc le niveau de provisions du BE de primes (et donc des provisions techniques). Dans notre exemple d'application :

- Le taux de résiliation sur primes existantes (hors échéance) serait ainsi sous-estimé (6,53% ou 6,78% vs 7,18% observé)
- Le taux de résiliation sur primes futures (hors échéance + à l'échéance hors délai) serait quant à lui :
  - Sous-estimé pour le modèle polytomique (8,67% vs 8,83% observé)
  - Surestimé pour les modèles binaires (9,03% vs 8,83% observé)

Ainsi, les taux obtenus sont cohérents avec l'expérience du portefeuille et semblent d'un niveau prédictif acceptable. La prudence en termes de provisions est correctement respectée. L'application d'un modèle explicatif a priori d'estimation des taux de chute permettrait [18] ainsi d'anticiper considérablement les calculs annuels réglementaires. Ces types de modèles permettraient également de distinguer les volumes de primes à prendre en compte dans la frontière des contrats.

18. Bien que plus complexe à mettre en œuvre que la méthode actuelle.

Conclusion

L'observation du phénomène de résiliation nous a ainsi amené à explorer diverses pistes afin d'en comprendre les causes pour être en mesure de prédire au mieux leur évolution à plus ou moins court terme.
Nous nous sommes concentrés, dans ce mémoire, à étudier le comportement du portefeuille de contrats multirisques habitation de la compagnie. Nous ne pouvons donc pas prétendre généraliser les résultats obtenus à l'ensemble du marché.

La résiliation peut être prise en compte dès la souscription du contrat. Le chapitre 4 décrit ainsi une méthodologie de préparation des données (variables et échantillons) dans le cadre d'un score d'attrition à la souscription. Ce score est construit de façon supervisée à l'aide d'un modèle de régression logistique binaire. Ce type de modèle présente au moins deux avantages essentiels : robustesse et interprétabilité. Son principal inconvénient est sa maintenance : nous avons en effet vu que la stabilité des variables doit être mesurée dans le temps. Nous aurions pu rajouter un corolaire à cela : son pouvoir discriminant doit aussi être suivi.
A moyen terme, le pouvoir discriminant du modèle s'est avéré acceptable. La discussion menée par la suite sur l'horizon du risque a mis en évidence une corrélation entre moyen et court terme. La discussion a également permis de montrer que le choix de l'échantillon peut nuancer le jugement que l'on peut avoir des résultats (d'acceptable à excellent, dans notre étude).

Pour confronter les méthodes, nous avons aussi souhaité étudier les résiliations sous l'axe complémentaire des durées de vie, dans le chapitre 3.
Nous avons d'abord montré qu'une approche discrète ou continue du temps ne semblait pas affecter significativement notre étude, a minima pour l'estimation de la fonction de survie [19].
Nous avons ensuite introduit un aspect explicatif dans nos modèles afin de tenir compte de l'hétérogénéité du portefeuille. Les conditions d'utilisation des modèles à hasard proportionnel et accéléré ont ainsi été détaillées pour justifier de leur application. Bien que la stabilité temporelle des modèles ait été jugée correcte, les résultats obtenus sur un stock de contrats à horizon annuel se sont avérés décevants. Cette limite semble plaider pour l'introduction, dans les modèles, d'une dynamique temporelle capable de tenir compte de la baisse tendancielle observée du taux de résiliation. Cette piste n'a pas été explorée dans ce mémoire.

Pour tenter d'améliorer la prédictivité des modèles temporels, les deux derniers chapitres du mémoire (5 et 6) ont été consacrés à l'étude des résiliations sur le même type d'échantillon, à savoir un stock de contrats en cours, observé pendant un an. De nombreuses méthodes de segmentation supervisées ont ainsi été testées. Le chapitre 5 décrit tout d'abord l'application d'un modèle logistique à notre stock et montre un apport non négligeable au pouvoir prédictif des modèles de durée. Cependant, les performances attendues ne sont pas au niveau de celles observées dans le modèle du chapitre 4.

Nous avons ainsi été amenés à approfondir l'étude en complétant notre approche par des techniques d'utilisation plus récente telles que le *bagging* et le *boosting*.
Le temps très important consacré aux traitements des données ne semble pas justifié au regard, d'une part du gain de performance attendu, d'autre part de l'interprétabilité des modèles.
Cependant, grâce aux évolutions technologiques le temps de calcul sera vraisemblablement beaucoup plus réduit dans le futur afin de généraliser l'utilisation des différentes techniques d'apprentissage automatique dans les modèles prédictifs.
Notre étude s'est limitée à l'étude de deux approches populaires bien qu'il en existe beaucoup d'autres (ainsi que d'autres variantes) comme par exemple le *stacking* qui peut combiner plusieurs types de modèles.

19. Sans prise en compte de variable explicative.

Nous avons conclu le chapitre 5 par l'étude de l'élasticité du taux de résiliation au montant de la cotisation d'assurance. Nous n'avons pas mis en évidence de hausse du taux lorsque la prime augmente. En revanche, nous avons observé une élasticité presque nulle à court terme. Ce résultat, bien qu'attendu, nous a permis de nous interroger sur la notion d'erreur de modèle ainsi que les biais cognitifs que ces derniers peuvent induire.
De plus, le contexte concurrentiel n'a été qu'ébauché. Il est probable que les apporteurs d'affaires joue un rôle plus important que celui mesuré dans notre étude. Pour une partie d'entre eux, la politique de la société mandante va impacter leur apport en affaires nouvelles voire en résiliations.

Enfin, dans le chapitre 6 le mémoire s'est interessé aux résiliations dans le cadre prudentiel de Solvabilité 2 (pilier 1). Comme pour les chapitres 3 et 5, nous sommes partis d'un stock observé sur un an, modélisé par régression logistique. Mais le périmètre de la frontière des contrats nous a incité à modéliser une variable cible à quatre modalités.
Les résultats nous sont apparus suffisamment robustes et prudents pour permettre l'utilisation d'un tel modèle. Le principal intérêt est un gain de temps très important lors de la phase de production des calculs de provisions car l'estimation des taux de chute pourrait être faite de nombreux mois avant la date de l'inventaire.

Nous nous sommes limités au cadre du premier pilier, mais il aurait été intéressant de se placer dans le contexte de l'*ORSA*[20] (pilier 2). L'*ORSA* nous demande d'avoir une vision plus prospective du risque d'assurance. Il faudrait donc mener une réflexion complémentaire sur la modélisation de l'évolution du taux de résiliation dans le temps (sur un horizon de trois à cinq ans). Une des questions à se poser serait, par exemple, la suivante : quelle est la dynamique d'évolution du taux et quel pourrait être son impact sur le *Best Estimate* ?

Ainsi, les principaux enseignements de ce travail sont les suivants :

- Il est difficile de construire des modèles stables dans le temps.
- La prédictivité des modèles proposés est correcte mais pas excellente.
- Les méthodes complexes ne sont pas forcément les meilleures.
- Mieux vaut privilégier la qualité voire l'ajout de nouvelles données pour améliorer les modèles.

20. *Own Risk and Solvency Assessment* ou Évaluation interne des risques et de la solvabilité

Bibliographie

[AJB13] Sabine Abravanel-Jolly and Bernard Beignier. *Droit des assurances.* Ellipses Marketing, Paris, July 2013. 20

[BMTW12] Bernd Bischl, O Mersmann, Heike Trautmann, and Claus Weihs. Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary computation*, 20 :249–75, 02 2012. 64, 143

[Chi11] Ossama Chihi. *Sensibilité du taux de résiliation au prix en assurance MRH occupant et simulation du portefeuille.* Mémoire d'actuariat, ISUP, Université Pierre et Marie Curie, 2011. 139

[ET93] Bradley Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall/CRC, New York, January 1993. 70

[FHT00] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression : A statistical view of boosting. *The Annals of Statistics*, 28 :337–407, 04 2000. 77, 143

[Fon11] Léonard Fontaine. *Modélisation de la valeur contrat par la refonte du modèle de résiliation.* Mémoire d'actuariat, ISFA, Université de Lyon 1, 2011. 139

[HLP09] François Husson, Sébastien Lê, and Jérôme Pagès. *Analyse de données avec R.* PU Rennes, Rennes, September 2009. 54, 61

[JR12] Laurent Jacques and Etienne Rain. *Du modèle GLM à une apporche darwinienne : Nouvelle génération de concepts et d'indicateurs pour l'optimisation du renouvellement Auto.* Mémoire d'actuariat, CEA, 2012. 84

[LG03] Jean-Marie Le Goff. Modélisation des événements du parcours de vie. Une introduction. Technical report, Mimeo, Universités de Lausanne et de Genève, 2003. 35

[LMP00] Ludovic Lebart, Alain Morineau, and Marie Piron. *Statistique exploratoire multidimensionnelle.* Paris : Dunod, 2000. 61

[MN89] Peter McCullagh and John A. Nelder. *Generalized Linear Models, second edition.* Chapman and Hall, 1989. 107

[NC03] Jean-Pierre Nakache and Josiane Confais. *Statistique explicative appliquée : analyse discriminante, modèle logistique, segmentation par arbre.* Editions TECHNIP, 2003. Google-Books-ID : j3ciFgIyVisC. 52

[NC05] Jean-Pierre Nakache and Josiane Confais. *Approche pragmatique de la classification.* Paris : Technip, 2005. 54

[PT00] Frédéric Planchet and Pierre Thérond. *Modélisation statistique des phénomènes de durée - Applications actuarielles.* Economica, Paris, October 2000. 35, 38, 39

[PTJ11] Frédéric Planchet, Pierre Thérond, and Marc Juliard. *Modéles financiers en assurance, analyses de risque dynamiques.* Paris : Economica, 2011. 62

[SF12] Robert E. Schapire and Yoav Freund. *Boosting : Foundations and Algorithms.* The MIT Press, 2012. 77, 143

[SW93] Judith D. Singer and John B. Willett. It's about Time : Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics*, 18(2) :155–195, 1993. 35

[Tas04] Philippe Tassi. *Méthodes statistiques.* Economica, Paris, 3e édition edition, October 2004. 117

[Ten07] Michel Tenenhaus. *Statistique, méthodes pour décrire, expliquer et prévoir.* Paris : Dunod, 2007. 17, 65, 79, 108, 117

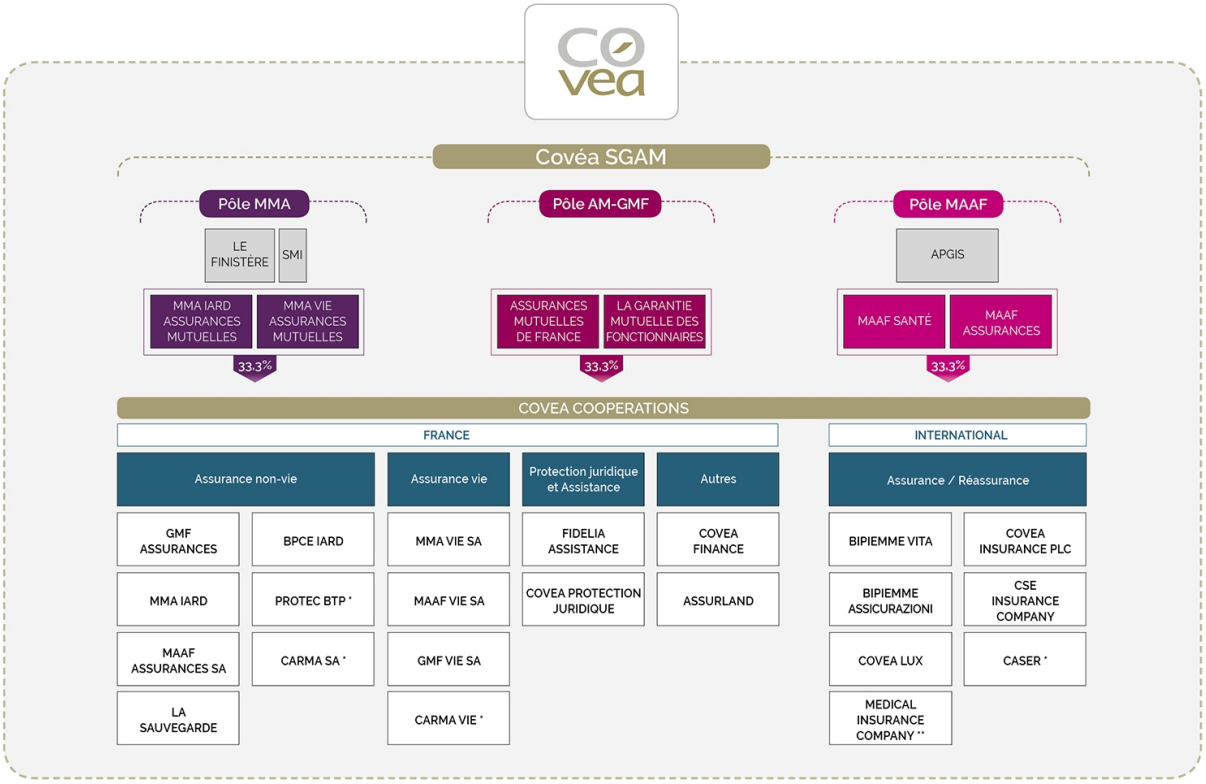
[TT17] Philippe Trainar and Patrick Thourot. *Gestion de l'entreprise d'assurance - 2e éd.* Dunod, 2e édition edition, July 2017. 20

[Tuf05] Stéphane Tuffery. Améliorer les performances d'un modèle prédictif : perspectives et réalité. In *Data Mining et Apprentissage Statistique : application en assurance, banque et marketing, Niort, France, 12-13 Mai 2005*, pages 45–72, 2005. 24

[Tuf17] Stephane Tuffery. *Modlisation Predictive et Apprentissage Statistique avec R.* Technip, 2e édition revue et augmentée edition, July 2017. 51, 57, 58, 79

Annexe A

Annexe - Généralités



** Méthode de mise en équivalence*
*** Filiale de MMA IARD Assurances Mutuelles*

FIGURE A.1 – Organigramme simplifié du groupe Covéa au 01/01/2019

Annexe B

Annexe - Présentation des données

Zoom sur le risque Habitation

			Comptages		
Echantillon	**Variable**	**Modalité**	**Global**	**Résiliés**	**Censurés**
PTF	**RISQUE**	**Habitation**	**383 344**	**259 114**	**124 230**
MRH	Fréquence	Annuel	305 695	221 228	84 467
MRH	Fréquence	Mensuel	71 765	34 396	37 369
MRH	Fréquence	Autre	5 884	3 490	2 394
MRH	Nb pièces	2 pièces ou -	81 587	67 980	13 607
MRH	Nb pièces	3 ou 4 pièces	105 942	76 670	29 272
MRH	Nb pièces	5 ou 6 pièces	96 193	61 893	34 300
MRH	Nb pièces	7 pièces ou +	58 400	32 691	25 709
MRH	Qualité Juridique	Locataire	132 542	114 483	18 059
MRH	Qualité Juridique	PNO	89 867	44 406	45 461
MRH	Qualité Juridique	Propriétaire	151 489	90 780	60 709
MRH	Qualité Juridique	Inconnue	9 446	9 445	1
MRH	Type Bien	Appartement	113 603	86 569	27 034
MRH	Type Bien	Maison	213 547	125 926	87 621
MRH	Type Bien	Autre	15 140	5 565	9 575
MRH	Zone géo.	Zone 1	39 215	22 626	16 589
MRH	Zone géo.	Zone 2	168 261	113 148	55 113
MRH	Zone géo.	Zone 3	175 868	123 340	52 528

Table B.1 – Comptages sur le portefeuille MRH

Statistiques univariées des variables explicatives MRH

Variables quantitatives

Nous avons dans un premier temps présenté les variables relatives au bien assuré (nombre de pièces, surface, valeur du mobilier, SMP), et dans un second temps celles relatives à la cotisation (franchise, montant et taux de réduction).

Données As-if Afin de pouvoir comparer des montants relatifs à des polices « anciennes » avec des montants plus récents, il convient de transformer les montants anciens à l'aide d'un indice considéré comme étant représentatif de l'évolution du montant considéré.
Le calcul attendu est assez simple :

$$\text{Mnt}^{\text{vu en année n}}_{\text{police année k}} = \text{Mnt}_{\text{police année k}} \cdot \frac{\text{Indice}_{\text{année n}}}{\text{Indice}_{\text{année k}}}$$

Ainsi, une indexation des variables suivantes est proposée : montant de cotisation, valeur du mobilier garanti, SMP. L'indice de référence est l'indice FFB [1].

Nombre de pièces

Les principes de comptage du nombre de **pièces principales** est généralement spécifié dans les Conditions Générales du contrat MRH. En 2017, la définition d'une pièce principale était la suivante :

- Toute pièce d'une surface minimale de 9 m^2 autre que : cuisine, arrière-cuisine, salle de bain et salle d'eau, toilettes, hall, couloir.
- L'ensemble des caves, garages, greniers utilisables et dépendances est à compter pour une pièce principale quelle que soit sa superficie développée, si cette dernière est comprise entre 40 et 300 m^2.
- Toute pièce à usage de bureau professionnel comprise entre 9 et 40 m^2.
- Toute pièce principale dont la superficie dépasse 40 m^2 est comptée pour autant de pièces que de tranche (ou fraction de tranche) de 40 m^2.

La Figure B.1 nous montre la distribution de la variable.

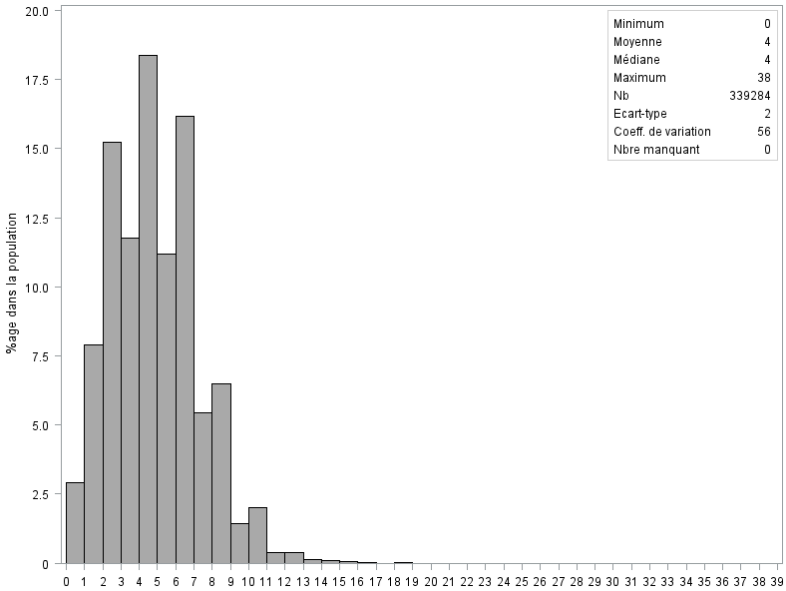


Figure B.1 – Distribution de la variable Nombre de pièces

La très grande majorité des contrats souscrits concerne des logements avec moins de huit pièces principales. Notons que les souscriptions autorisées dans les conditions tarifaires (par défaut) ne doivent pas dépasser dix pièces principales. Pour les logements dont le nombre de pièces est compris entre 11 et 15, le souscripteur doit déroger pour tarifer le contrat. Enfin, au-delà de 15 pièces, la tarification est censée se faire en fonction de la superficie précise du bien.
Nous avons ainsi choisi d'exclure les 13 individus dont le nombre de pièces déclaré nous semblait trop élevé (plus de 40 pièces).

Surface du bien

La **superficie développée** est obtenue, tenant compte de l'épaisseur des murs extérieurs, en additionnant la surface totale des rez-de-chaussée ainsi que de chacun des niveaux du bien. Les caves, garages et greniers utilisables comptent pour moitié de leur surface réelle.

1. Cet indice trimestriel (base 1 au 01/01/1941) du coût de la construction, basé sur le prix de revient d'un immeuble de rapport de type courant à Paris, est calculé par la Fédération Française du Bâtiment.

La Figure B.2 nous montre la distribution de la variable.

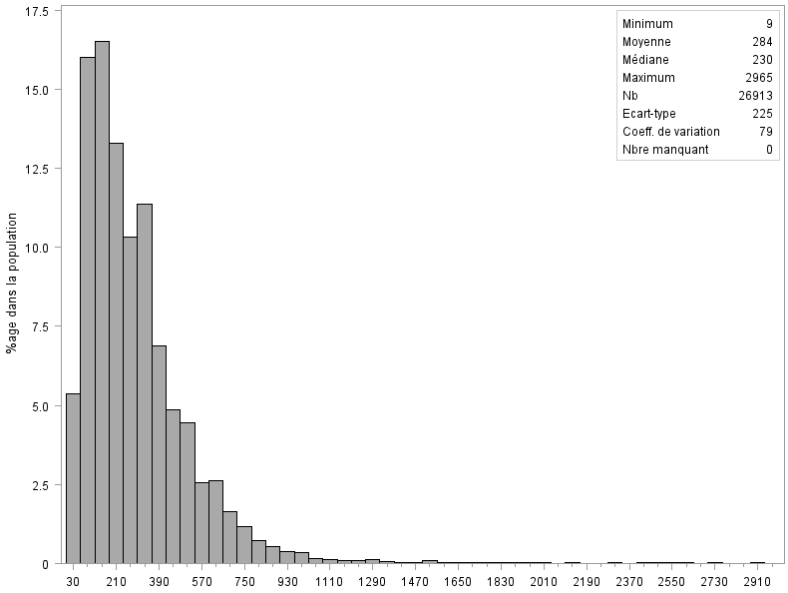


Figure B.2 – Distribution de la variable Surface du bien

Comme précisé pour la variable « nombre de pièces », la tarification se fait à la superficie lorsque le client déclare plus de 15 pièces principales. Cela explique la relativement faible volumétrie du nombre de contrats au regard du portefeuille (car si la superficie est saisie, alors le souscripteur ne renseigne pas le nombre de pièces [2]). Notons que les souscriptions autorisées dans les conditions tarifaires (par défaut) ne doivent pas dépasser 750 m^2.
Nous avons choisi d'exclure les 12 contrats dont la surface déclarée était supérieure à 3 000 m^2.

Valeur du mobilier garanti

La valeur du mobilier garanti est laissée au libre choix de l'assuré, qui doit choisir entre plusieurs tranches de montants en fonction de ses besoins. La Figure B.3 nous montre la distribution de la variable.

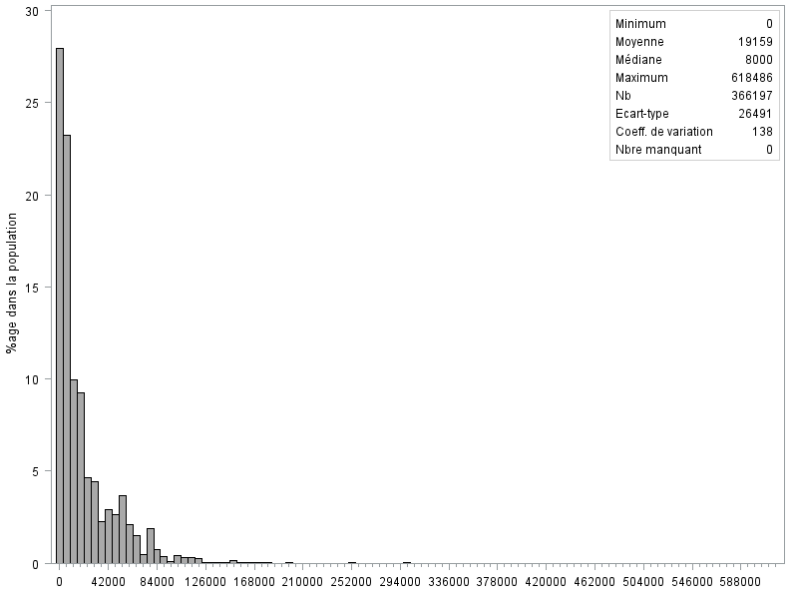


Figure B.3 – Distribution de la variable Valeur du mobilier garanti

2. Seuls 8 contrats, relativement anciens, possèdent les deux informations.

Nous observons un montant médian relativement faible (8 000 €) pour une moyenne de l'ordre de 20 000 €. Notons que les souscriptions autorisées dans les conditions tarifaires de base ne doivent pas dépasser 122 000 €.
Nous avons ainsi choisi d'exclure les 6 individus dont le montant de garantie était supérieur à 500 000 €.

Valeur du SMP

La Figure B.4 nous montre la distribution de la variable relative au SMP [3]. Le SMP correspond au montant des dommages matériels le plus important pouvant résulter d'un événement garanti par le contrat d'assurance. Il s'agit donc du pire scénario entraînant l'épuisement complet de la garantie. Son calcul est automatisé et dépend, pour les contrats MRH, du nombre de pièces principales et de la valeur du mobilier garanti.

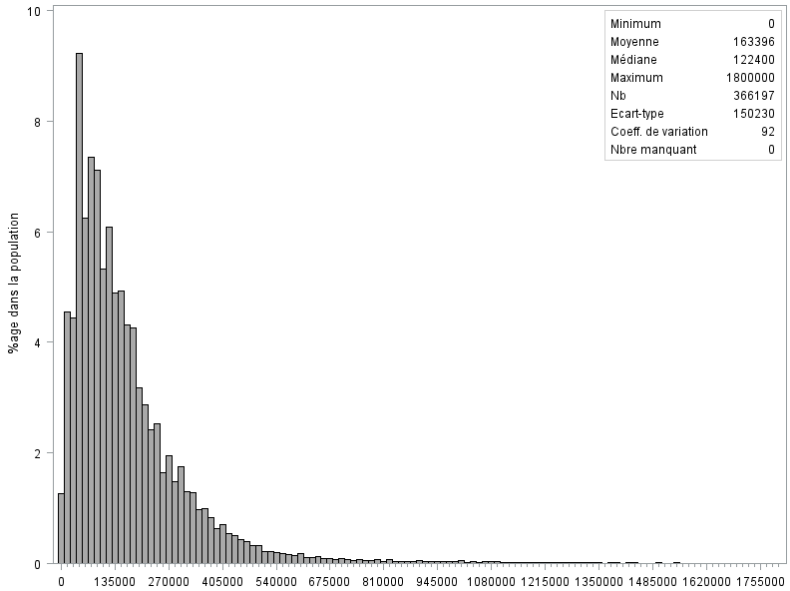


Figure B.4 – Distribution de la variable Valeur du SMP

Nous avons choisi d'exclure un contrat très atypique et dont le montant du SMP était supérieur à 10 000 000 €.

Montant de la cotisation nette

Il s'agit du montant de cotisation nette [4] annuelle. Nous rappelons que ce montant est indexé à l'indice FFB afin de pouvoir comparer les montants de primes des contrats anciens et des contrats plus récents.
La cotisation dépend du risque inhérent à chaque contrat, estimé à partir des données saisies lors de la souscription. L'objectif est d'avoir une prime qui couvre l'ensemble des sinistres du portefeuille (principe de la mutualisation) sans pour autant être trop élevée (car sinon les clients peu risqués sont susceptibles de résilier leur contrat).

La Figure B.5 nous montre la distribution de la variable.

3. Sinistre Maximum Possible.
4. Hors frais de chargement et hors taxes.

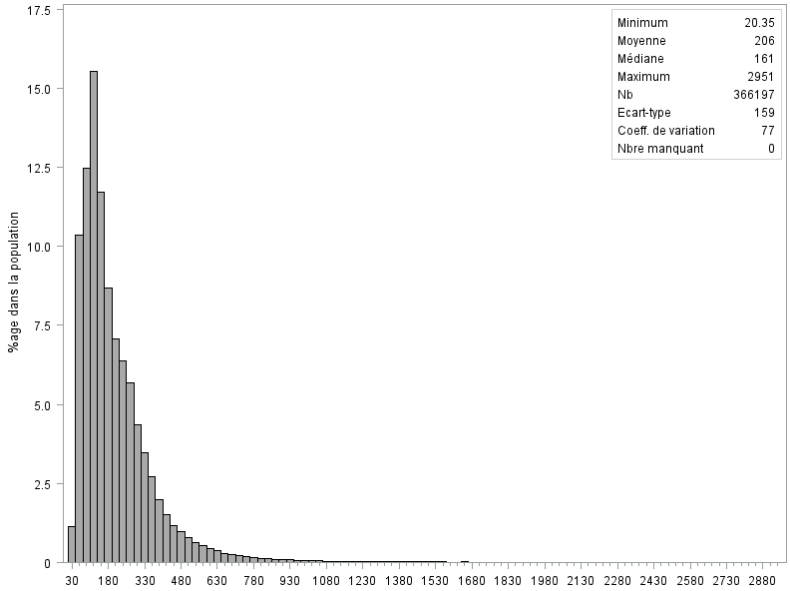


FIGURE B.5 – Distribution de la variable Montant de cotisation

Pour des raisons de cohérence, nous avons choisi d'exclure les 65 individus dont le montant de la cotisation était soit trop faible ($\leq$ 20 €) soit trop élevé ($\geq$ 3000 €). En effet, après vérification, il s'est avéré que ces anomalies relevaient d'erreurs opérationnelles de saisie. Notons également qu'un minimum de cotisation est appliqué (en 2017 il était fixé à 65 € TTC [5]).

Variables qualitatives

Nous avons présenté les variables relatives au bien assuré (département du risque, qualité juridique, type de bien), celles relatives à la cotisation (produit, nature des garanties, fréquence de paiement), et enfin celles relatives à l'apporteur d'affaire (réseau, type d'agent).

Département du risque assuré

La FIGURE B.6 nous montrent la distribution géographique des biens assurés. Sans surprise, les quatre départements bretons (Finistère, Morbihan, Île-et-Vilaine et Côte d'Armor) ainsi que la Loire Atlantique sont les plus représentés. Cela est logique car le développement commercial dans les autres départements s'est fait progressivement :

- Maine-et-Loire (49) et Mayenne (53) à partir de 1990.
- La Vendée (85) et le sud de la Manche (50) en 1997.
- La Sarthe (72), l'Orne (61) et les Deux-Sèvres (79) en 1998.
- Le Calvados (14) et l'ensemble de la Manche en 1999.
- Enfin la Charente-Maritime (17), l'Indre-et-Loire (37) et la Vienne (86) en 2006.

La cartographie (à droite) met en évidence les 16 départements de prospection du Finistère Assurance.

5. Toute Taxe Comprise.

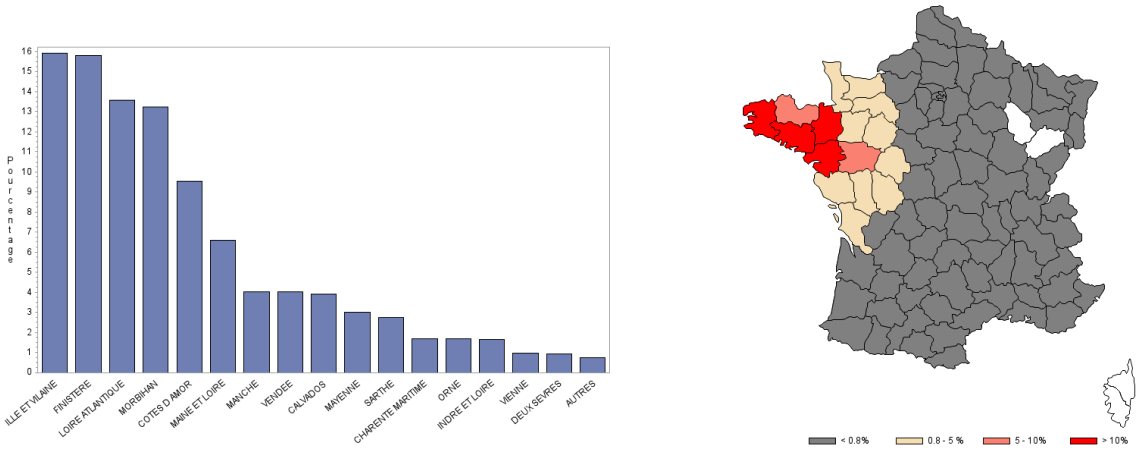


FIGURE B.6 – Distribution de la variable Département

Notons, pour l'anecdote, que quelques rares risques peuvent apparaître dans d'autres départements (en France métropolitaine).

Qualité juridique

La qualité juridique décrit le statut du titulaire du contrat et permet de savoir notamment s'il occupe ou pas le logement assuré et s'il en est le propriétaire ou le locataire. La FIGURE B.7 nous montrent la distribution de la variable. Le graphique de droite met en évidence l'agrégation « métier » choisie parmi les différentes modalités possibles (visibles sur le graphe de gauche).

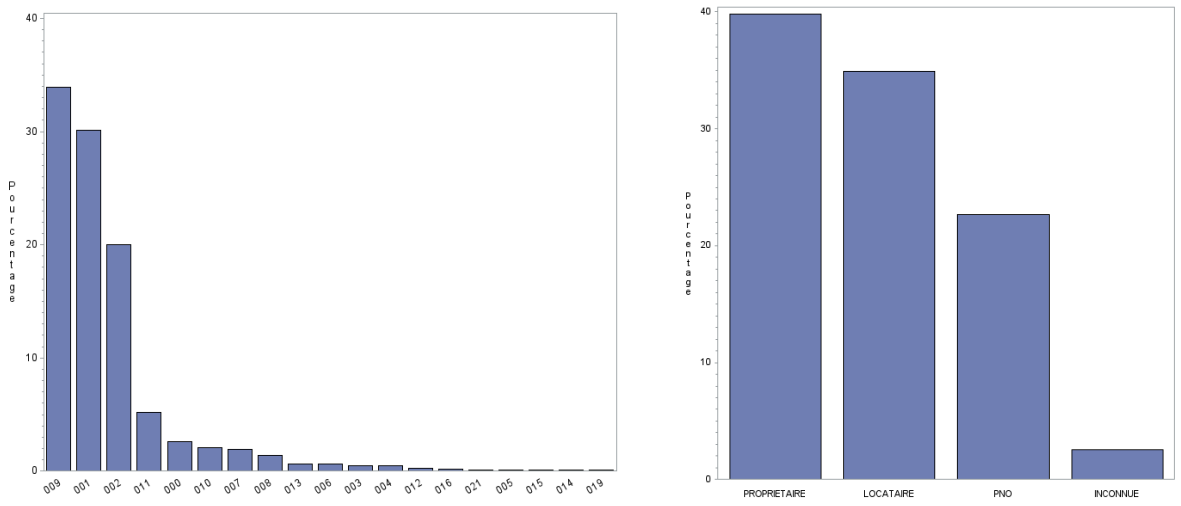


FIGURE B.7 – Distribution de la variable Qualité juridique

Nous observons une majorité de propriétaires (occupant ou pas le logement). Il sera intéressant de voir par la suite, la dynamique temporelle des trois principales modalités de la variable agrégée. Notons que nous ne disposons pas de cette information pour quelques contrats (modalité INCONNUE).

Type de bien assuré

La FIGURE B.8 nous montre la distribution de la variable. Comme pour la variable précédente, les différentes modalités possibles ont été agrégées selon une logique métier pour n'en conserver que quatre. La modalité AUTRE peut concerner des immeubles, garages ou d'autres types de dépendances.

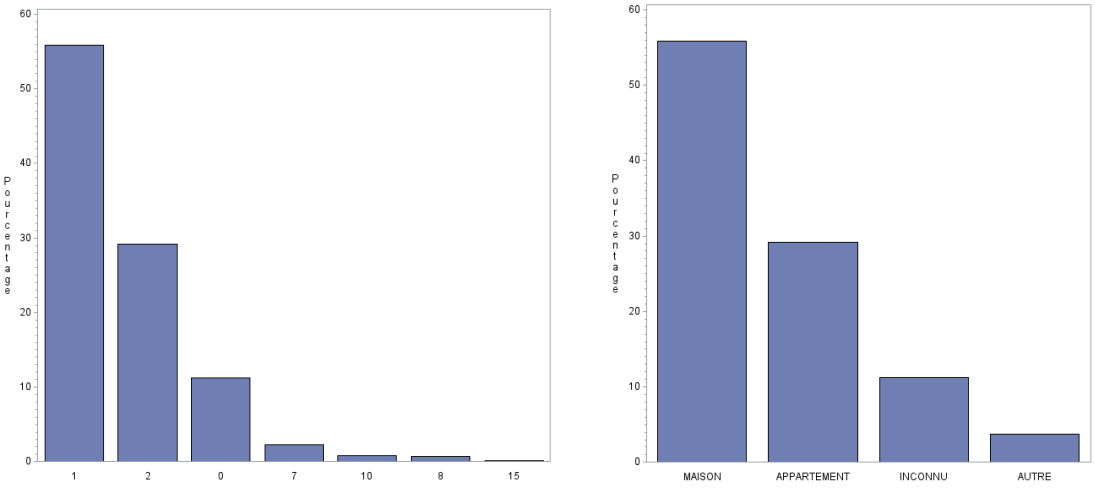


FIGURE B.8 – Distribution de la variable Type de bien

La modalité INCONNU représente environ 10% des contrats de la base.

Type de produit

Le Finistère Assurance commercialise une gamme de produits MRH dont voici les principales caractéristiques :

- Evolution : MRH complète et haut de gamme (Tout Risque Sauf).
- Globale : MRH complète.
- Menhir : MRH complète pour les sociétaires privilégiant le meilleur rapport qualité-prix.
- Promotion : MRH complète pour les sociétaires privilégiant la performance tarifaire.
- Etudiant : MRH adaptée aux étudiants (moins de deux pièces, faible valeur de mobilier garantie).

La FIGURE B.9 nous montre la distribution de cette variable.

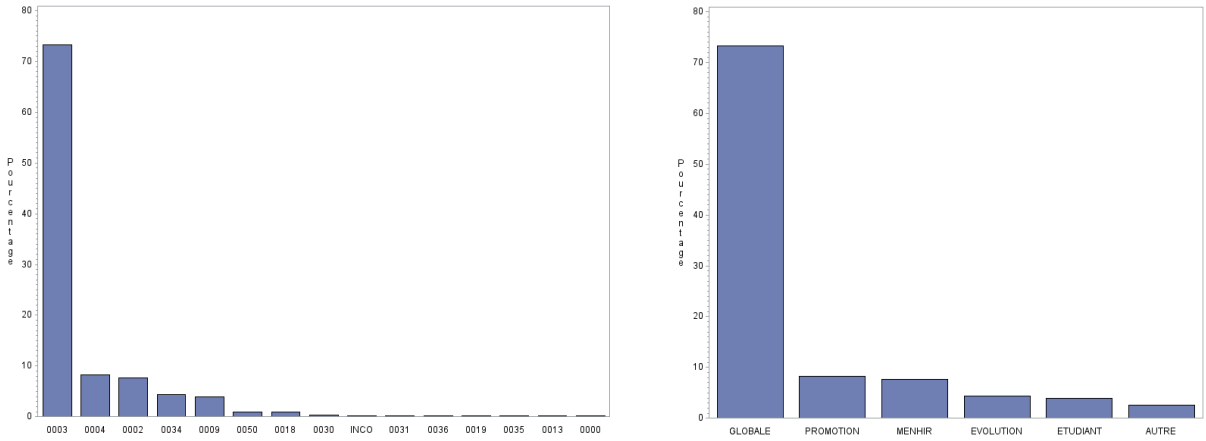


FIGURE B.9 – Distribution de la variable Type de produit

Le produit phare de la société d'assurance mutuelle est sans conteste le produit **Globale**. Il reflète l'appétence des sociétaires à une assurance complète et avec des garanties de qualité.

Type de formule

La formule est en quelque sorte le barême tarifaire appliqué au contrat. Il résulte donc d'un paramétrage des caractéristiques du contrat afin d'en faciliter sa tarification. La FIGURE B.10 nous montre la distribution de cette variable.

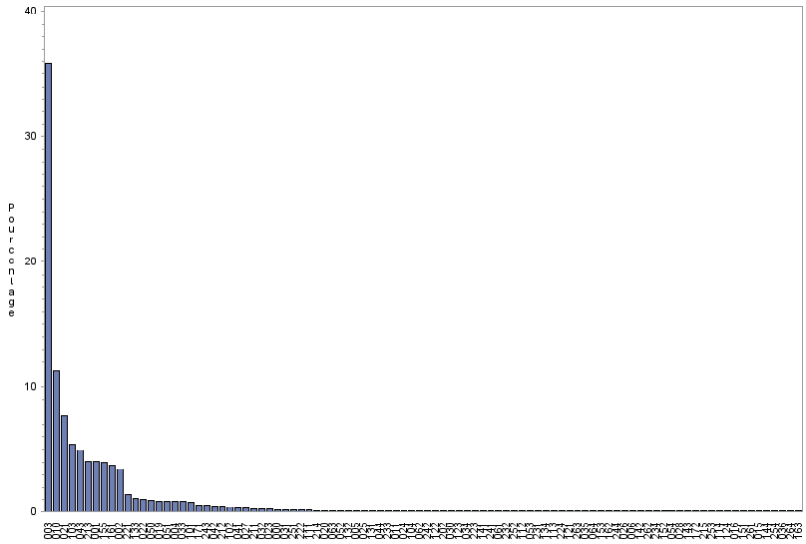


FIGURE B.10 – Distribution de la variable Type de formule

Nous observons un nombre très important de formules, ce qui peut sembler normal a priori car leur paramétrage croise potentiellement beaucoup de variables tarifaires. Cependant, environ une douzaine d'entre elles concentrent la majorité des contrats du portefeuille (trois formules concernent plus de la moitié du portefeuille).

Nature des garanties

La nature des garanties présente trois segments :

- Bien d'habitation (BH) : bâtiment et dépendances associées.
- Mobilier d'habitation (MH) : meubles et objets à usage domestique appartenant à l'assuré.
- Bien d'habitation et mobilier d'habitation (BHMH) : bâtiment et contenu d'habitation.

La FIGURE B.11 nous montre la distribution de la variable. Elle comporte quatre modalités (dont une concerne les données manquantes).

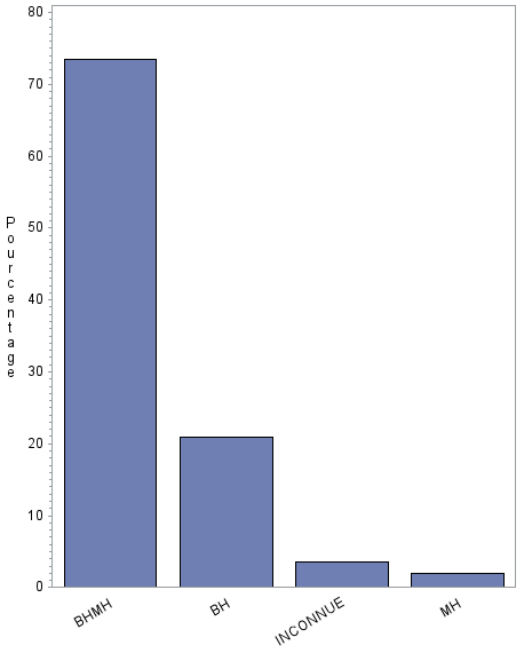


FIGURE B.11 – Distribution de la variable Nature des garanties

La très grande majorité des contrats comporte des garanties pour le bien et le mobilier (+70 %).

Fréquence de paiement

La fréquence de paiement est, dans la grande majorité des cas, annuelle (cf. FIGURE B.12). Le paiement semestriel est admis pour les cotisations annuelles supérieures à 300 € TTC. Les paiements semestriels (inférieurs à 300 €), trimestriels ou mensuels ne sont autorisés que par prélèvement.

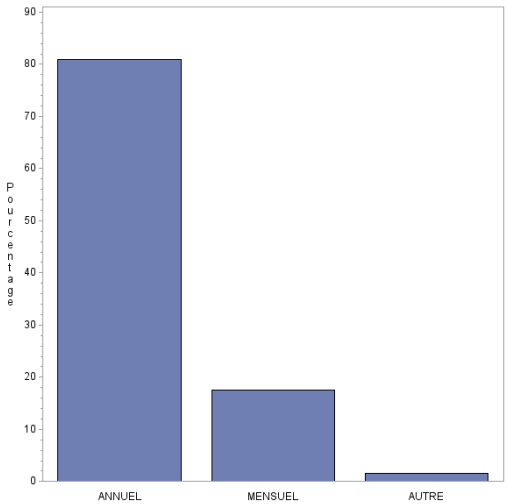


FIGURE B.12 – Distribution de la variable Fréquence de paiement

Notons que la dynamique actuelle tend à montrer que la part des paiements par prélèvement mensuel augmente continûment (cf. le cinquième graphe en FIGURE D.3, en annexe page 127).

Réseau

Le Finistère Assurance est une mutuelle qui fonctionne en **B to B to C**, c'est-à-dire qu'elle passe par un intermédiaire avant de faire une offre à un éventuel assuré. Cet apporteur d'affaire, commissionné, peut

dépendre d'un réseau d'agences (Allianz, Axa, etc.) ou pas (courtier pur). La FIGURE B.13 nous montre la distribution de cette variable.

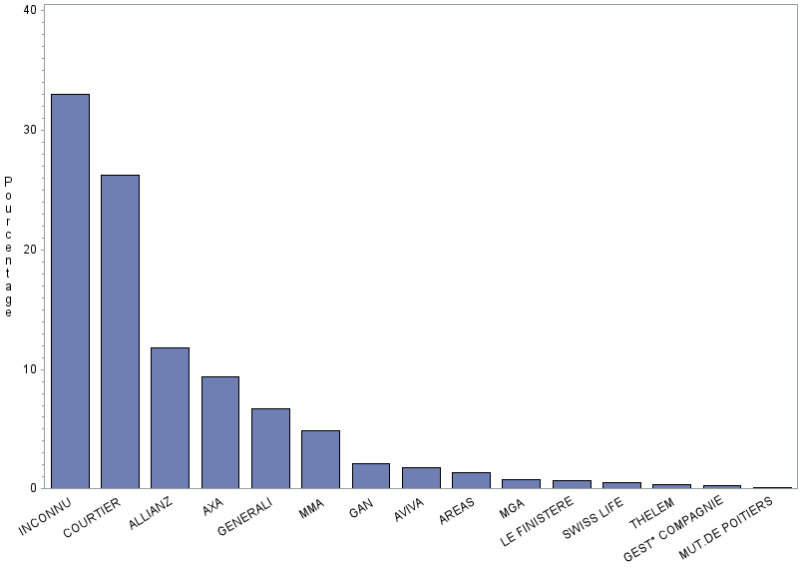


FIGURE B.13 – Distribution de la variable Réseau

Nous observons tout d'abord un nombre important de valeurs manquantes qui résulte d'une évolution du SID : la donnée n'a commencé à être historisée correctement qu'à partir de la fin des années 2000.
Nous constatons ensuite que la majorité des contrats provient des courtiers purs. Allianz, Axa et Generali sont, historiquement, les principaux contributeurs à la production de l'entreprise.

Type d'agent

En complément du réseau, nous avons l'information du type d'agent apporteur. La FIGURE B.14 nous en montre la distribution.

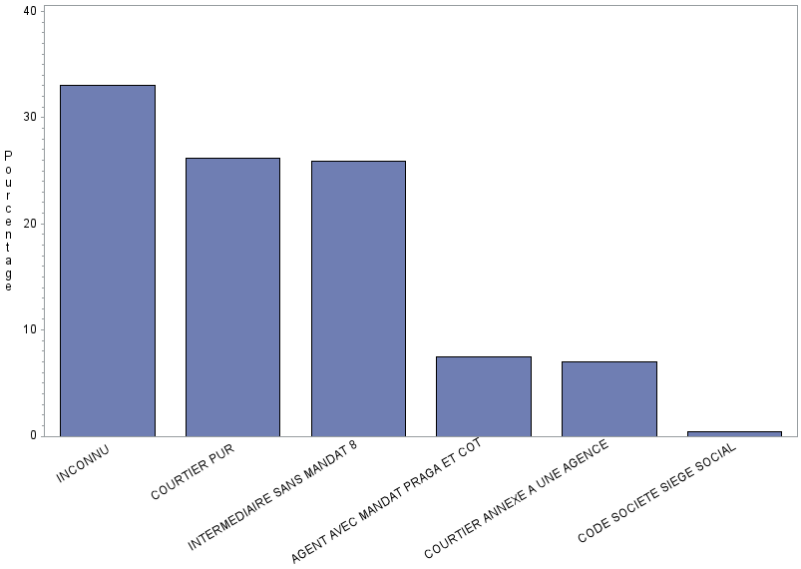


FIGURE B.14 – Distribution de la variable Type d'agent

La même observation et la même remarque est faite concernant les valeurs manquantes que pour la variable précédente (Réseau). Cette variable peut également être vue comme un regroupement métier de la variable réseau.

Annexe C

Annexe - Etude de la durée de vie des contrats MRH

Rappels théoriques

Modèles de durée

Formalisation de la survie Nous notons X la durée de vie dont nous cherchons à estimer la loi. Pour toute la suite, on se donne un réel $t \in \mathbb{R}^+$. Notons $f_X(t)$ la densité de X.

La **fonction de répartition** de X, par construction croissante et continue à droite, s'écrit formellement :

$$F_X(t) = \mathbb{P}[X \leq t] = \int_0^t f_X(u)du$$

Par complémentarité, la **fonction de survie** est :

$$S_X(t) = 1 - F_X(t) = \int_t^{+\infty} f_X(u)du$$

Cette fonction mesure la probabilité de ne pas avoir résilié à une maturité t donnée. Nous observons également que la densité de survie est l'opposée de la dérivée par rapport au temps de la fonction de survie. Nous définissons également la **fonction de hasard** h_X, qui peut s'interpréter comme le taux de résiliation « instantanée » (à t) du portefeuille :

$$h_X(t) = \frac{f_X(t)}{S_X(t)}$$

Sa version cumulative nous donne la **fonction de hasard cumulée** :

$$H_X(t) = \int_0^t h_X(u)du$$

Nous pouvons montrer que la fonction de survie et la fonction de hasard sont liées par la bijection suivante :

$$S_X(t) = \exp[-H_X(t)] \tag{C.1}$$

Nous définissons enfin la **durée de vie résiduelle**, qui mesure le nombre d'années restant à vivre à l'instant t (sachant que la police n'est pas résiliée à t) :

$$r_X(t) = \mathbb{E}[X - t|X > t] = \frac{\int_t^{+\infty} S_X(u)du}{S_X(t)} \tag{C.2}$$

Rappelons que ces mesures caractérisent la loi de X et que la connaissance de l'une implique la connaissance des autres. L'objectif est de trouver le meilleur estimateur de la fonction de survie S_X.
Nous donnerons par la suite, si besoin, des versions discrétisées de certaines des mesures présentées ci-dessus.

Phénomène de censure Il est fondamental de tenir compte de ce phénomène pour limiter les biais d'estimation de la fonction de survie. La censure est définie comme le fait de ne pas observer la survenance de l'événement d'intérêt durant la phase d'observation. Elle peut arriver avant (censure à gauche) ou après (censure à droite) voire avant et après.

La censure peut être considérée comme une variable temporelle aléatoire ou pas. Nous distinguons ainsi les censures aléatoires (de type II ou III), des censures déterministes (de type I).

Notre échantillon d'observations est composé de données censurées à droite. Cette censure est de type I car la date de fin d'observation s'arrête le 31/12/2017 pour toutes les polices.
Dans ce cas, nous ne savons pas quand la résiliation aura lieu. Nous savons juste qu'à cette date la police est toujours dans le portefeuille.
Nous allons décrire formellement ci-dessous le modèle avec censure.

Nous considérons un n-échantillon $(X_i)_{1\leq i\leq n}$ de durées de survie et nous fixons[1] $C_i > 0$. La censure à droite est décrite de la façon suivante :

$$\begin{cases} T_i = X_i \wedge C_i \\ D_i = \mathbb{1}_{[T_i = X_i]} \end{cases}$$

En effet, nous n'observons pas X mais le couple (T, D) en réalité (à cause des censures).

Phénomène de troncature Par définition, nous parlons de troncature à gauche (ou à droite) quand la durée d'intérêt n'est pas observable : lorsqu'elle est inférieure (ou supérieure resp.) à un seuil $c > 0$ (ou $C > 0$ resp.). Nous ne sommes pas concernés, dans notre étude, par ce phénomène.

Estimateur de Kaplan-Meier Il s'agit d'un estimateur de la fonction de survie qui tient compte des données censurées. Il a été proposé en 1958 par les deux auteurs dont il porte le nom.
Formellement, plusieurs écritures sont proposées dans la littérature. Par exemple, en admettant la présence d'ex-aequo et en supposant que les observations non censurées précèdent toujours les observations censurées :

$$\hat{S}(t) = \prod_{T_{(i)} \leq t} \left[1 - \frac{d_i}{r_i}\right] \tag{C.3}$$

Où d_i représente le nombre de résiliations au temps t_i, et r_i le nombre de polices à risque juste avant le temps t_i.
Sans rentrer dans les détails, cette formulation découle des propriétés de la fonction de survie conditionnelle.

Cet estimateur, très populaire dans les applications, est considéré comme le meilleur estimateur de la fonction de survie car il possède les propriétés suivantes :

- Il est convergent (sous conditions).
- Il en est l'unique estimateur cohérent.
- Il est GMLE (estimateur du maximum de vraisemblance généralisé).
- Il est consistant.
- Il est asymptotiquement normal (sous conditions).

Son principal défaut est d'être **biaisé positivement** :

$$\text{Biais}\left(\hat{S}(t)\right) = \mathbb{E}\left[\hat{S}(t)\right] - S(t) \geq 0$$

Ainsi en moyenne, l'estimateur de la survie est plutôt supérieur à la valeur théorique. La survie est donc surestimée.

1. Les polices sont issues de cohortes différentes.

Variance et propriétés asymptotiques de l'estimateur de Kaplan-Meier Sa variance peut être mesurée à l'aide de l'estimateur de Greenwood :

$$\hat{\sigma}_{gw}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{T_{(i)} \leq t} \frac{d_i}{r_i(r_i - d_i)}$$

La normalité asymptotique de cet estimateur permet de calculer des intervalles de confiance point par point.

Intervalles de confiance KM ponctuels Nous faisons l'hypothèse que la censure qui intervient dans nos données est **non informative**, i.e. qu'elle est indépendante de l'événement de résiliation [2].
Nous faisons également l'hypothèse que la population étudiée est **homogène** [3].
Nous notons que sous ces deux hypothèses, l'estimateur de KM, noté $\hat{S}(t)$, converge asymptotiquement vers une gaussienne centrée sur $S(t)$.

L'estimateur de Greenwood, ainsi que la propriété de normalité asymptotique de l'estimateur de KM, autorisent à proposer des intervalles de confiance, maturité par maturité, sous la forme :

$$S(j) \in \left[\hat{S}_j \left[1 \pm u_{1-\frac{\alpha}{2}} \times \gamma(j)\right]\right] \text{ avec } \gamma(j) = \sqrt{\sum_{k=1}^{j} \frac{d_k}{r_k(r_k - d_k)}} \tag{C.4}$$

où $u_{1-\frac{\alpha}{2}}$ est le fractile de rang $(1 - \frac{\alpha}{2}) \times 100$ d'une loi gaussienne standardisée.

Remarque sur l'estimateur de Nelson Aalen Cet estimateur permet également d'estimer indirectement la fonction de survie à partir de la fonction de hasard cumulée. L'écriture suivante est proposée :

$$\hat{H}_{NA}(t) = \sum_{T_i \leq t} \frac{d_i}{n - i + 1}$$

Puis, par l'équation (C.1) on en déduit $\hat{S}_{NA}$, appelé estimateur de Fleming et Harrington.

Cependant, cet estimateur est biaisé et sous-estime généralement la fonction de hasard cumulée. Selon la stratégie voulue, nous pourrons choisir l'un ou l'autre.

Modèles linéaires généralisés

Il s'agit dans ce paragraphe de décrire quelques notions de base sur ce type de modèle. Le lecteur intéressé pourra se référer au livre de référence de [MN89] pour un exposé exhaustif du sujet.

L'un des modèles les plus populaires en statistique inférentielle est le modèle multilinéaire classique qui permet, sous diverses conditions, de modéliser une réponse gaussienne en fonction de covariables (continues ou pas) supposées indépendantes. Sa principale limite est donc d'imposer de travailler sur une variable d'intérêt de loi normale, et donc uniquement continue, symétrique et à valeurs dans $\mathbb{R}$.

Cette contrainte rend donc particulièrement ardue la tâche de modéliser des nombres de sinistres (à valeurs dans $\mathbb{N}$), des coûts de sinistres (à valeurs dans $\mathbb{R}^+$) ou encore des taux de résiliations (à valeurs dans $[0, 1]$).

L'introduction des modèles linéaires généralisés (GLM en anglais), par John Nelder et Robert Wedderburn, dans les années 70 a donc ainsi permis d'élargir le cadre d'étude en généralisant la modélisation aux variables appartenant à la famille de lois de probabilité possédant une **structure exponentielle**.

On dit que la variable aléatoire Y, discrète ou continue, appartient à cette famille si sa densité peut se formaliser de la façon suivante :

2. Ce qui est vrai car elle est uniquement liée à la date de vue de l'ensemble des données.
3. Ce mémoire montre que ce n'est pas le cas.

$$f_Y(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

où y prend ses valeurs dans un sous-ensemble de $\mathbb{N}$ ou $\mathbb{R}$. La fonction b est définie sur $\mathbb{R}$, deux fois dérivable et sa dérivée première est injective. La fonction c est définie sur $\mathbb{R}^2$. Cette famille de modèles a donc deux paramètres :

- Un paramètre **naturel** ou canonique : θ.
- Un paramètre d'**échelle** ou de dispersion : ϕ (qui peut être pondéré éventuellement).

On montre que si Y appartient à cette famille, alors elle possède les propriétés suivantes :

$$\mathbb{E}(Y) = b'(\theta) \text{ et } \mathbb{V}(Y) = b''(\theta)\cdot\phi$$

Nous notons que la variance de Y est une fonction de sa moyenne : elle évolue donc avec elle, l'hypothèse d'homoscédasticité n'est donc plus nécessaire dans le cadre des GLM.

Sans rentrer dans les détails, cette famille de fonctions possède des propriétés algébriques (distributions conjuguées) et statistiques (cf. le théorème de Darmois) remarquables, permettant notamment d'obtenir des formules fermées lors de la résolution de certains problèmes.

Formellement, un GLM est défini par la donnée :

i d'un vecteur aléatoire $Y = (y_i)_{1\leq i\leq n}$ de dimension n ayant des composantes indépendantes dont la loi est issue d'une même structure exponentielle.

ii d'une matrice $X = (x_i^t)_{1\leq i\leq n}$ d'ordre $n \times p$ connue, de rang $p < n$ appelée **matrice des variables explicatives** (ou du plan d'expérience).

iii d'un vecteur p-dimensionnel $\beta = (\beta_i)_{1\leq i\leq p}$ appelé **vecteur des paramètres**.

iv d'une **fonction de lien** g, monotone, différentiable, telle que $g(\mu_i) = x_i^t\beta$ où $\mu_i = \mathbb{E}(y_i)$.

Le paramètre $\eta = X\beta$ est appelé **prédicteur linéaire**.

L'estimation du vecteur des paramètres β se fait par maximisation de la vraisemblance, méthode qui est implémentée dans la plupart des logiciels de traitement de données disponibles sur le marché.

Le cas de la régression logistique binaire

Pour un exposé complet et pédagogique sur la régression logistique, se référer à [Ten07] par exemple.

On montre que la loi binomiale de paramètres (n,p) appartient à la famille exponentielle si l'on pose :

- $\theta = \ln(\frac{p}{1-p})$ son paramètre naturel, appelé **fonction logit** (d'où le nom du type de régression).
- $b(.) = n\ln(1 + exp(.))$
- $\phi = 1$, son paramètre de dispersion.
- $c(y,\phi) = \ln(C_n^y)$

Nous rappelons que la moyenne d'une v.a. Y de loi $\mathcal{B}(n,p)$ est $\mathbb{E}[Y] = np$ et que sa variance est $\mathbb{V}[Y] = np(1-p)$. Nous rappelons également qu'une loi $\mathcal{B}(n,p)$ peut être considérée comme une somme de n lois de Bernoulli indépendantes, de paramètre p.

Notons que le paramètre n, représentant l'effectif, est supposé ici fixe (car sinon nous ne sommes plus dans le cadre GLM). Le paramètre d'intérêt est donc la proportion p dont le logit est supposé être linéaire en fonction de variables explicatives à préciser.

Intervalles de confiance des paramètres d'un modèle logistique

Les propriétés asymptotiques des paramètres $\beta = (\beta_j)_j$ d'un modèle logistique, estimés par maximisation de la vraisemblance, nous permettent de déduire un intervalle de confiance avec un niveau d'erreur α :

$$\beta_j \in \left[\hat{\beta}_j \pm u_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{\beta}_j)}\right] \tag{C.5}$$

Modèle en temps continu ou en temps discret ?

Précisions techniques sur la construction des données

Dans le modèle discret, les données sont préalablement mises sous forme longitudinale :

- Avec autant de lignes par individu que d'années observables : de la maturité 0 à min(date résiliation, 31/12/2017).
- Pour la dernière ligne associée à chaque police : soit il s'agit d'une résiliation, soit d'une censure (si non résiliée au 31/12/2017).
- La variable cible est donc $Y = 1$ si événement (résiliée) ; $Y = 0$ si non événement (non résiliée i.e. censure).
- Les variables explicatives relatives à la maturité sont alors créées : X_{ij} afin d'indexer l'individu i à la maturité j (vaut 1 si la police i est observée à la maturité j, 0 sinon).
- Un modèle de régression logistique (de type Bernoulli ou Binomial, selon que l'on modélise au niveau des individus ou des maturités) est alors ajusté afin d'estimer les paramètres (β) relatifs aux maturités.
- Le modèle tient donc bien compte de la censure par construction car ces données sont considérées comme des non événements ($Y = 0$).

Param	**DDL**	**Estimation**	**Erreur type**	**Khi-2 de Wald**	**Pr > Khi-2**	**h**	**S**
α_0	1	-3,0158	0,0077	155 281	<.0001	4,67%	**95,33%**
α_1	1	-1,5193	0,0044	117 421	<.0001	17,96%	**78,21%**
α_2	1	-1,6078	0,0052	96 455	<.0001	16,69%	**65,16%**
α_3	1	-1,7838	0,0062	82 346	<.0001	14,38%	**55,79%**
α_4	1	-1,9154	0,0073	69 311	<.0001	12,84%	**48,62%**
α_5	1	-1,9892	0,0083	56 925	<.0001	12,03%	**42,77%**
α_6	1	-2,0093	0,0093	46 385	<.0001	11,82%	**37,72%**
α_7	1	-2,0760	0,0106	38 150	<.0001	11,15%	**33,51%**
α_8	1	-2,1134	0,0119	31 601	<.0001	10,78%	**29,90%**
α_9	1	-2,1348	0,0131	26 661	<.0001	10,58%	**26,74%**
α_{10}	1	-2,1828	0,0145	22 700	<.0001	10,13%	**24,03%**
α_{11}	1	-2,2110	0,0159	19 217	<.0001	9,88%	**21,66%**
α_{12}	1	-2,2639	0,0178	16 193	<.0001	9,42%	**19,62%**
α_{13}	1	-2,2321	0,0192	13 538	<.0001	9,69%	**17,72%**
α_{14}	1	-2,2629	0,0214	11 192	<.0001	9,42%	**16,05%**
α_{15}	1	-2,2462	0,0233	9 285	<.0001	9,57%	**14,51%**
α_{16}	1	-2,2603	0,0257	7 717	<.0001	9,45%	**13,14%**
α_{17}	1	-2,3053	0,0288	6 393	<.0001	9,07%	**11,95%**
α_{18}	1	-2,3177	0,0320	5 252	<.0001	8,97%	**10,88%**
α_{19}	1	-2,3404	0,0360	4 227	<.0001	8,78%	**9,92%**
α_{20}	1	-2,3105	0,0395	3 419	<.0001	9,03%	**9,03%**
α_{21}	1	-2,3956	0,0458	2 740	<.0001	8,35%	**8,27%**
α_{22}	1	-2,4528	0,0526	2 177	<.0001	7,92%	**7,62%**
α_{23}	1	-2,4208	0,0582	1 733	<.0001	8,16%	**7,00%**
α_{24}	1	-2,3830	0,0642	1 378	<.0001	8,45%	**6,40%**
α_{25}	1	-2,2870	0,0688	1 106	<.0001	9,22%	**5,81%**
α_{26}	1	-2,4296	0,0822	873	<.0001	8,09%	**5,34%**
α_{27}	1	-2,2520	0,0864	679	<.0001	9,52%	**4,83%**
α_{28}	1	-2,2599	0,0980	532	<.0001	9,45%	**4,38%**
α_{29}	1	-2,2691	0,1120	411	<.0001	9,37%	**3,97%**
α_{30}	1	-2,3994	0,1337	322	<.0001	8,32%	**3,64%**
α_{31}	1	-2,1700	0,1386	245	<.0001	10,25%	**3,26%**
α_{32}	1	-2,3311	0,1699	188	<.0001	8,86%	**2,98%**
α_{33}	1	-2,5745	0,2118	148	<.0001	7,08%	**2,76%**
α_{34}	1	-2,6210	0,2589	102	<.0001	6,78%	**2,58%**
α_{35}	1	-1,9459	0,2390	66	<.0001	12,50%	**2,26%**
α_{36}	1	-1,8269	0,2880	40	<.0001	13,86%	**1,94%**
α_{37}	1	-2,1595	0,4312	25	<.0001	10,34%	**1,74%**
α_{38}	1	-1,8245	0,4819	14	0.0002	13,89%	**1,50%**
α_{39}	1	-2,4849	0,7360	11	0.0007	7,69%	**1,38%**
α_{40}	1	-2,3514	0,7400	10	0.0015	8,70%	**1,26%**
α_{41}	1	-2,1972	0,7454	9	0.0032	10,00%	**1,14%**
α_{42}	1	-2,6391	1,0351	7	0.0108	6,67%	**1,06%**
α_{43}	1	-2,3979	1,0445	5	0.0217	8,33%	**0,97%**

TABLE C.1 – Paramètres estimés et fonction de survie pour le modèle Bernoulli.

Maturité	Nb non résiliées	Nb résiliées	Nb à risque	Nb censurées
0	365 457	17 910	383 367	20 071
1	283 290	62 000	345 290	14 832
2	223 569	44 785	268 354	13 327
3	179 924	30 228	210 152	11 005
4	147 165	21 675	168 840	11 180
5	119 545	16 355	135 900	9 286
6	97 177	13 030	110 207	7 745
7	79 424	9 962	89 386	5 825
8	65 631	7 930	73 561	3 751
9	55 314	6 542	61 856	2 963
10	47 028	5 301	52 329	2 839
11	39 801	4 362	44 163	2 736
12	33 555	3 488	37 043	2 489
13	28 041	3 009	31 050	2 428
14	23 192	2 413	25 605	1 917
15	19 234	2 035	21 269	1 569
16	15 990	1 668	17 658	1 392
17	13 266	1 323	14 589	1 282
18	10 904	1 074	11 978	1 266
19	8 786	846	9 632	983
20	7 096	704	7 800	853
21	5 718	521	6 239	751
22	4 567	393	4 960	619
23	3 624	322	3 946	486
24	2 872	265	3 137	343
25	2 294	233	2 527	305
26	1 828	161	1 989	271
27	1 407	148	1 555	190
28	1 102	115	1 217	161
29	851	88	939	118
30	672	61	733	106
31	508	58	566	77
32	391	38	429	52
33	315	24	339	79
34	220	16	236	60
35	140	20	160	39
36	87	14	101	29
37	52	6	58	16
38	31	5	36	5
39	24	2	26	1
40	21	2	23	1
41	18	2	20	3
42	14	1	15	2
43	11	1	12	11

Table C.2 – Table annuelle des polices résiliées - Vue au 31/12/2017 - Portefeuille MRH

Comparaison des étendues

Des intervalles de confiance à un niveau $1 - \alpha = 95\%$ sont proposés, afin d'avoir une idée de la précision des estimateurs obtenus par les deux méthodes. Les représentations graphiques sont présentées dans la Figure C.1 :

- Le graphique à gauche représente l'estimateur KM avec intervalles obtenus à l'aide de l'équation (C.4).

- Le graphique à droite représente l'estimateur en temps discret de type Bernoulli avec intervalles obtenus à l'aide de l'équation (C.5).

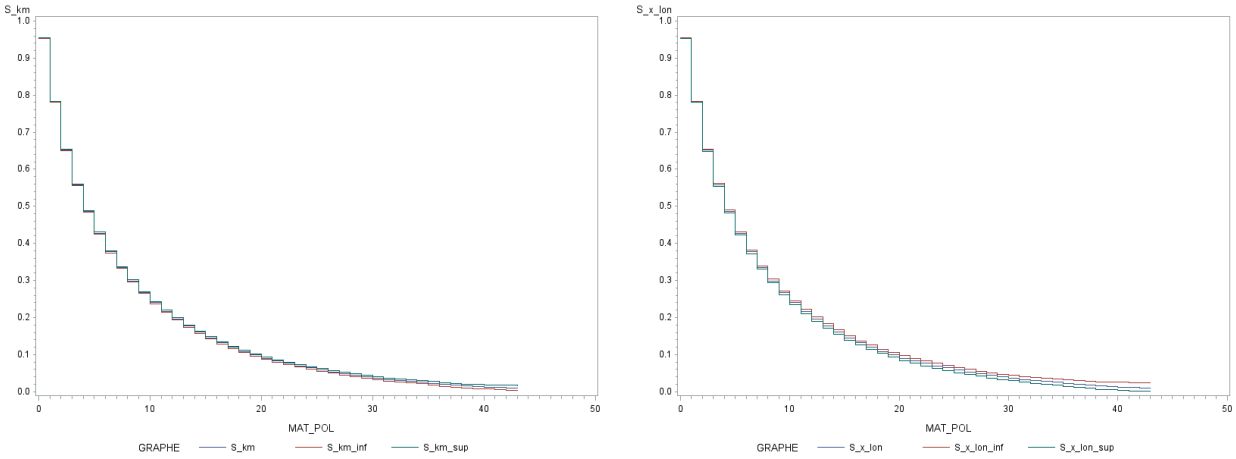


FIGURE C.1 – Comparaison des approches.

Nous observons que les intervalles de confiance sont plutôt fins (ce qui peut s'expliquer par la volumétrie d'événements assez élevée des vingt premières maturités). Nous resterons donc assez prudents et critiques par la suite concernant l'incertitude a priori, car n'oublions pas que l'objectif est d'appliquer une estimation de survie à horizon annuel.

La FIGURE C.2 nous montre, pour chacune des méthodes, l'écart par maturité entre la borne inférieure et la borne supérieure. Nous observons que :

- Les deux courbes sont croissantes, ce qui est logique car l'incertitude augmente lorsque diminue le nombre d'observations.
- La méthode KM (temps continu) est la plus optimiste au niveau de l'estimation moyenne (les intervalles de confiance sont plus fins).

Intégration de variables exogènes

Modèles AFT

Démonstration de la formule de départ

Formellement, la fonction de survie d'un modèle AFT s'écrit :

$$S(t|X;\theta) = S_0(\phi_i \cdot t) \tag{C.6}$$

où S_0 est la fonction de survie (à spécifier) d'un individu de référence (i.e. avec des valeurs nulles pour toutes les indicatrices associées aux variables explicatives) et $\phi_i = \mathrm{e}^{-(\theta_1 X_1 + \cdots + \theta_p X_p)}$.

Cette expression provient du fait que pour la variable T modélisant la survie, le modèle proposé est le suivant :

$$\log(T|X;\theta) = {}^tX\theta + \sigma\epsilon \tag{C.7}$$

où σ est appelé le *coefficient d'échelle* et ϵ est une v.a. centrée de paramètre d'échelle unitaire.

Si l'on pose ${}^tX\theta = \sum_{i=0}^{p} \theta_i X_i$, alors pour l'individu de référence décrit précédemment on a ${}^tX\theta = \theta_0 X_0 = \theta_0$ et donc :

$$\log(T_0|X;\theta) = \theta_0 + \sigma\epsilon \tag{C.8}$$

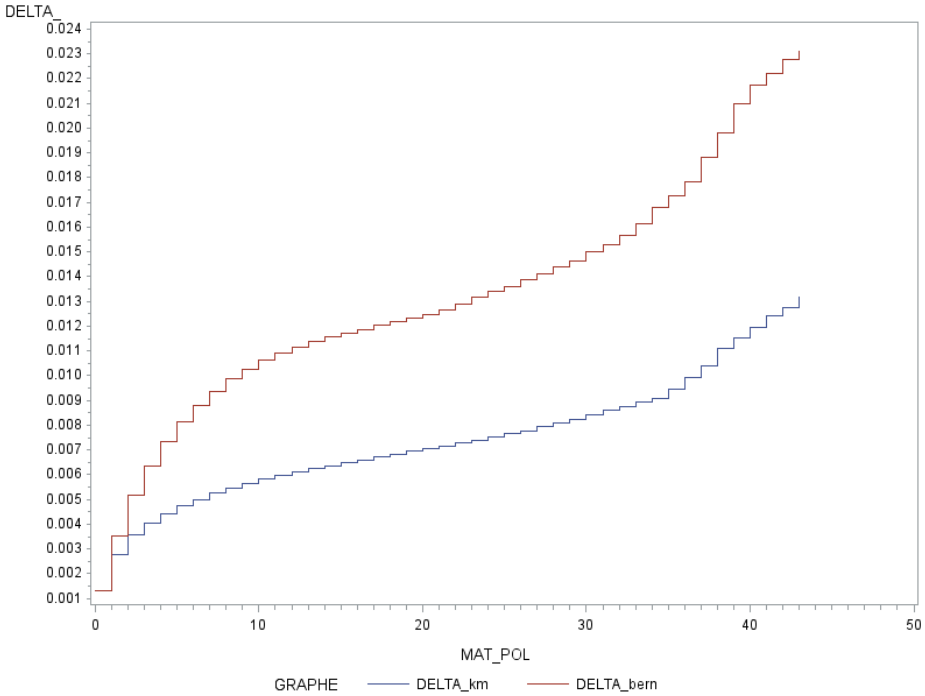


FIGURE C.2 – Comparaison des étendues entre les bornes inf et sup des deux approches.

Lorsque nous passons à la fonction de survie, cela donne (les notations sont allégées pour plus de clarté) :

$$\begin{aligned}
S_0(t) &= \mathbb{P}[T_0 > t] \\
&= \mathbb{P}[\log(T_0) > \log(t)] \text{ car la fonction } \log(.) \text{ est croissante} \\
&= \mathbb{P}[\theta_0 + \sigma\epsilon > \log(t)] \text{ d'après l'équation (C.8)} \\
&= \mathbb{P}\left[\epsilon > \frac{\log(t) - \theta_0}{\sigma}\right] \text{ car } \sigma \text{ est } > 0 \\
&= \mathbb{P}\left[\epsilon > \frac{\log[t\,\mathrm{e}^{-\theta_0}]}{\sigma}\right] \qquad \text{(C.9)}
\end{aligned}$$

Nous généralisons le calcul à T :

$$\begin{aligned}
S(t) &= \mathbb{P}[T > t] \\
&= \mathbb{P}[\log(T) > \log(t)] \\
&= \mathbb{P}[{}^tX\theta + \sigma\epsilon > \log(t)] \text{ d'après l'équation (C.7)} \\
&= \mathbb{P}\left[\epsilon > \frac{\log(t) - {}^tX\theta}{\sigma}\right] \\
&= \mathbb{P}\left[\epsilon > \frac{\log[t\,\mathrm{e}^{-{}^tX\theta}]}{\sigma}\right] \\
&= \mathbb{P}\left[\epsilon > \frac{\log[t\,\mathrm{e}^{-\theta_0}\,\mathrm{e}^{-(\theta_1 X_1 + \cdots + \theta_k X_k)}]}{\sigma}\right] \\
&= S_0\left(t \cdot \mathrm{e}^{-(\theta_1 X_1 + \cdots + \theta_k X_k)}\right) \text{ d'après l'équation (C.9)} \\
&= S_0\left(t \cdot \phi_i\right) \text{ et on retrouve bien la formule (C.6)}
\end{aligned}$$

Tests graphiques d'adéquation

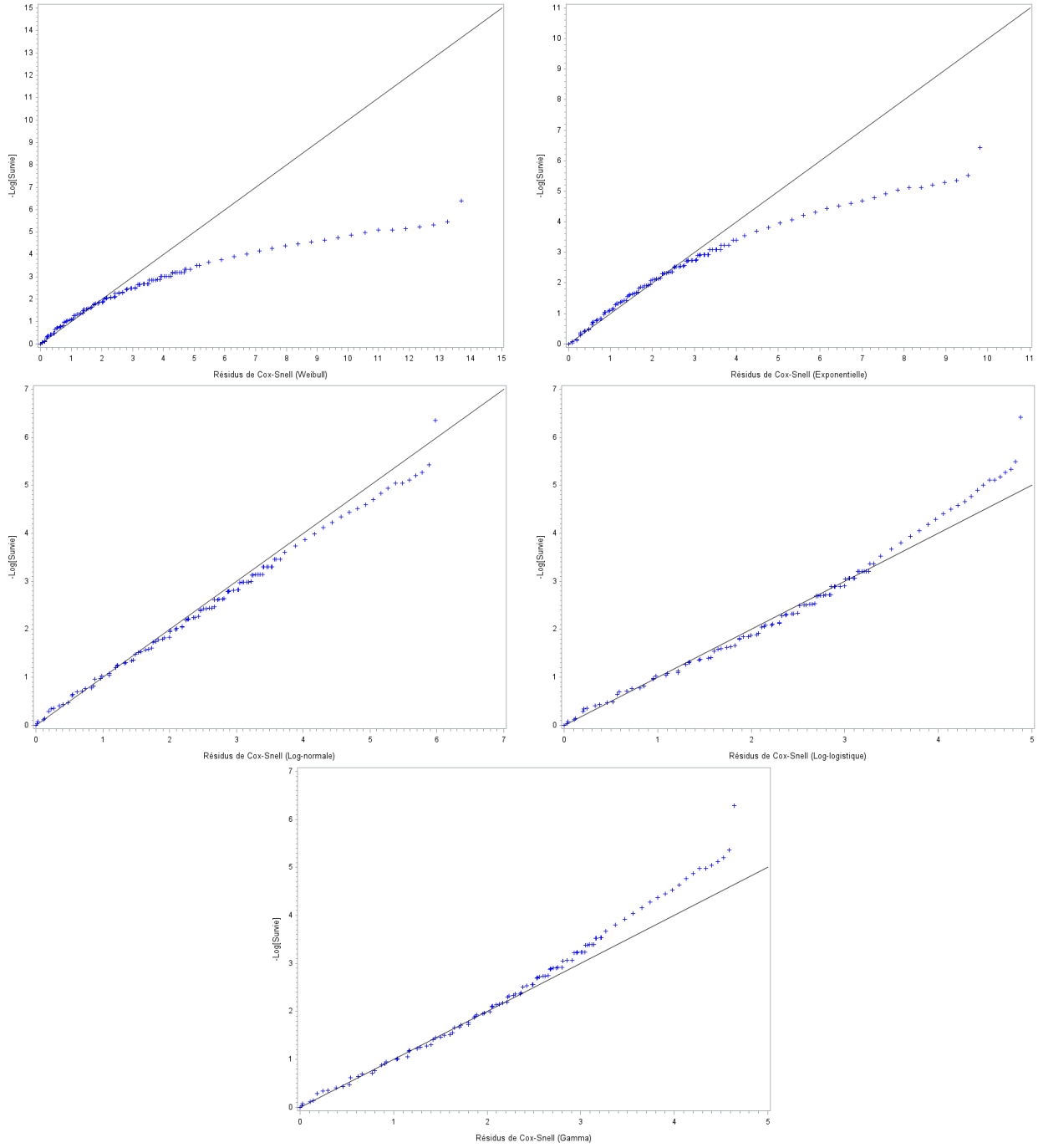


FIGURE C.3 – Résidus de Cox-Snell pour la variable Qualité juridique

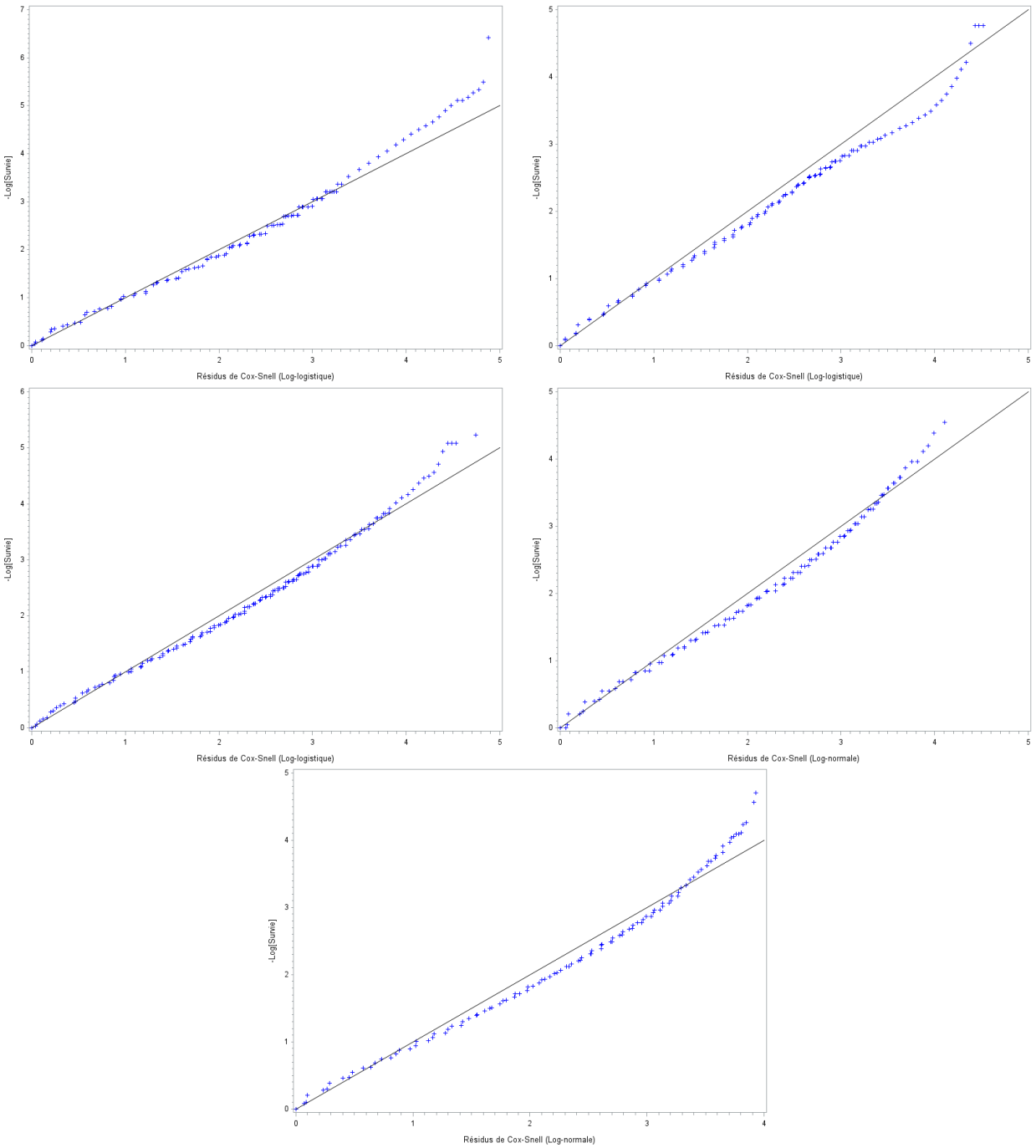


FIGURE C.4 – Résidus de Cox-Snell pour les cinq variables : QJ, TB, NP, FP et ZG (resp.)

Quelques précisions sur les distributions : log-logistique et log-normale

La distribution log-logistique On considère une variable aléatoire $T > 0$ et deux paramètres : $\alpha > 0$ (échelle) et $\beta > 0$ (forme). On dit que T suit une loi log-logistique $\mathcal{LL}(\alpha, \beta)$, si elle admet pour densité la fonction suivante :

$$f(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{[1 + (t/\alpha)^{\beta}]^2}$$

Cette expression permet de déduire la formalisation des fonctions de survie et de hasard :

$$S(t) = \frac{1}{1 + (t/\alpha)^{\beta}} \text{ et } h(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1 + (t/\alpha)^{\beta}}$$

Notons quelques propriétés :

- Le paramètre d'échelle α est égale à la médiane de la distribution.
- La distribution est unimodale lorsque $\beta > 1$.
- La dispersion de la distribution diminue lorsque β augmente.
- Le $k^{\text{ème}}$ moment existe lorsque $k < \beta$ et vaut $\mathbb{E}[T^k] = \alpha^k \frac{k\pi/\beta}{\sin(k\pi/\beta)}$.
- $Y = \log(T)$ suit une loi logistique avec $\log(\alpha)$ pour paramètre de position et $1/\beta$ pour paramètre d'échelle.

On note que, contrairement à la loi de Weibull, la loi log-logistique permet d'obtenir des fonctions de hasard croissantes puis décroissantes. Cette propriété est également partagée par les distributions log-normales, que nous allons présenter lors du prochain paragraphe.

La distribution log-normale On considère une variable aléatoire $T > 0$ et deux paramètres : $\mu \in \mathbb{R}$ et $\sigma > 0$. On dit que T suit une loi log-normale $\mathcal{LN}(\mu, \sigma)$, si $Y = \log(T)$ suit une loi normale d'espérance μ et de variance σ^2. T admet alors pour densité la fonction suivante :

$$f(t) = [t\sigma\sqrt{2\pi}]^{-1} \exp\left(-\frac{1}{2}\left[\frac{\log(t) - \mu}{\sigma}\right]^2\right)$$

Cette expression permet de déduire la formalisation des fonctions de survie et de hasard :

$$S(t) = 1 - \Phi\left[\frac{\log(t) - \mu}{\sigma}\right] \text{ et } h(t) = \frac{[t\sigma]^{-1}\Phi\left[\frac{\log(t)-\mu}{\sigma}\right]}{\Phi\left[-\frac{\log(t)-\mu}{\sigma}\right]}$$

Où Φ est la fonction de répartition d'une loi normale $\mathcal{N}(\mu, \sigma)$.

Notons également quelques propriétés :

- Sa médiane vaut e^{μ}.
- Le $k^{\text{ème}}$ moment existe $\forall k \geq 0$ et vaut $\mathbb{E}[T^k] = \mathrm{e}^{k\mu + \frac{1}{2}k^2\sigma^2}$.
- Il s'agit d'un cas particulier d'une Gamma généralisée (lorsque le paramètre de forme de la Gamma vaut zéro).

Applications des modèles AFT

Modèle	Variable	Modalité	Nb de polices à fin 2017	Nb résiliations prédites dans l'année	Taux de résiliation à un an
LOGLOG	Qualité jur.	Locataire	18 060	4 012	22%
		PNO	45 461	4 716	10%
		Propriétaire	60 709	5 577	9%
		TOTAL	124 230	14 305	11,52%
LOGLOG	Type de bien	Appartement	27 034	5 704	21%
		Maison	87 621	9 580	11%
		Autre	9 575	898	9%
		TOTAL	124 230	16 183	13,03%
LOGLOG	Nb de pièces	2 pièces ou -	13 607	2 805	21%
		3 ou 4 pièces	29 272	3 688	13%
		5 ou 6 pièces	34 300	2 339	7%
		7 pièces ou +	25 709	1 888	7%
		Surface	21 342	1 888	9%
		TOTAL	124 230	12 607	10,15%
LOGNOR	Frq de paiement	Annuel	84 467	10 849	13%
		Mensuel	37 369	4 798	13%
		Autre	2 394	238	10%
		TOTAL	124 230	15 885	12,79%
LOGNOR	Zone géo.	ZONE 1	16 589	2 265	14%
		ZONE 2	55 113	7 765	14%
		ZONE 3	52 528	6 197	12%
		TOTAL	124 230	16 227	13,06%

Table C.3 – Résultats des modèles AFT

Stabilité temporelle des données

Principe des tests d'hypothèses Se référer à [Tas04] pour un exposé complet sur le sujet, ou encore à [Ten07] pour un exposé plus synthétique.

L'objectif de l'inférence statistique est de généraliser les résultats obtenus sur un n-échantillon [4] $(x_1, \ldots, x_n)$ à l'ensemble d'une population d'intérêt $(X_1, \ldots, X_N)$. La question est donc de savoir si l'échantillon sous-jacent est bien représentatif de la population et si oui, quelle confiance accorder aux résultats obtenus.
Ainsi, nous estimons $\hat{\theta}$ à partir de cet échantillon censé généraliser une population de paramètre de référence θ_0. Autrement dit, nous allons vérifier que l'échantillon étudié possède une propriété intrinsèque à la population de référence [5]. Pour cela, nous allons nous aider d'une statistique $T = f(x_1, \ldots, x_n)$ mesurant un écart normalisé entre l'estimation $\hat{\theta}$ et la valeur théorique testée θ_0.
La loi de T est supposée être connue sous l'hypothèse **nulle** $[H_0 : \theta = \theta_0]$, et une règle de décision doit être définie afin de choisir entre l'hypothèse nulle et l'hypothèse **alternative** $[H_1 : \theta \in \Theta_1]$ où Θ_1 est l'ensemble des valeurs possibles de θ lorsque l'hypothèse nulle H_0 est rejetée.

Typiquement, la règle de décision est de la forme : « rejet de H_0 si $|T| > c$ ». Cette règle conduit généralement à deux types d'erreur :

- L'**erreur de première espèce** ou de type I, qui consiste à rejeter H_0 alors que H_0 est vraie.
- L'**erreur de seconde espèce** ou de type II, qui consiste à accepter H_0 alors que H_0 est fausse.

4. Ce n-échantillon est supposé être la réalisation de variables aléatoires indépendantes x_i de loi connue ou pas.
5. Par exemple : même moyenne, même loi, etc.

Ainsi nous notons α le risque de première espèce, qui représente le niveau d'erreur habituellement défini dans les intervalles de confiance [6], et β le risque de seconde espèce [7].
Selon le domaine métier (industrie, finance, médecine, etc.) dans lequel sont effectués les tests, l'importance accordée à ces deux indicateurs peut considérablement varier afin d'amener les décideurs à adopter la solution la plus appropriée.

6. Formellement : $\alpha = \mathbb{P}[|T| > c|H_0 \text{ vraie}]$ et $1 - \alpha$ est appelée la confiance du test.
7. Formellement : $\beta = \mathbb{P}[|T| \leq c|H_1 \text{ vraie}]$ et $1 - \beta$ est appelée la puissance du test.

Nom de la variable	Modalité	Hazard Ratio	Lower	Upper	Per	Test
FRQ_PAI	ANNUEL	2,031	1,945	2,121	2005	
FRQ_PAI	ANNUEL	1,307	1,277	1,337	2009	0
FRQ_PAI	ANNUEL	1,114	1,097	1,131	2013	0
FRQ_PAI	ANNUEL	1,083	1,071	1,096	2017	0
FRQ_PAI	AUTRE	1,501	1,380	1,632	2005	
FRQ_PAI	AUTRE	1,082	1,023	1,144	2009	0
FRQ_PAI	AUTRE	0,936	0,897	0,977	2013	0
FRQ_PAI	AUTRE	0,955	0,922	0,989	2017	1
LIB_NB_PIE	2 pièces ou -	2,771	2,690	2,853	2005	
LIB_NB_PIE	2 pièces ou -	2,634	2,570	2,698	2009	1
LIB_NB_PIE	2 pièces ou -	2,609	2,556	2,662	2013	1
LIB_NB_PIE	2 pièces ou -	2,640	2,596	2,685	2017	1
LIB_NB_PIE	3 ou 4 pièces	1,750	1,700	1,802	2005	
LIB_NB_PIE	3 ou 4 pièces	1,723	1,682	1,765	2009	1
LIB_NB_PIE	3 ou 4 pièces	1,714	1,680	1,749	2013	1
LIB_NB_PIE	3 ou 4 pièces	1,712	1,684	1,740	2017	1
LIB_NB_PIE	5 ou 6 pièces	1,325	1,286	1,366	2005	
LIB_NB_PIE	5 ou 6 pièces	1,312	1,280	1,345	2009	1
LIB_NB_PIE	5 ou 6 pièces	1,297	1,271	1,324	2013	1
LIB_NB_PIE	5 ou 6 pièces	1,308	1,286	1,330	2017	1
LIB_NB_PIE	7 pièces ou +	1,151	1,114	1,189	2005	
LIB_NB_PIE	7 pièces ou +	1,131	1,102	1,162	2009	1
LIB_NB_PIE	7 pièces ou +	1,100	1,075	1,124	2013	1
LIB_NB_PIE	7 pièces ou +	1,095	1,075	1,116	2017	1
QUA_JUR_2	LOCATAIRE	2,846	2,809	2,884	2005	
QUA_JUR_2	LOCATAIRE	2,734	2,703	2,765	2009	0
QUA_JUR_2	LOCATAIRE	2,773	2,745	2,801	2013	1
QUA_JUR_2	LOCATAIRE	2,801	2,776	2,826	2017	1
QUA_JUR_2	PNO	1,113	1,091	1,137	2005	
QUA_JUR_2	PNO	1,062	1,045	1,080	2009	0
QUA_JUR_2	PNO	1,035	1,021	1,049	2013	1
QUA_JUR_2	PNO	1,026	1,015	1,038	2017	1
TYP_BIE	APPARTEMENT	2,102	2,073	2,132	2005	
TYP_BIE	APPARTEMENT	1,976	1,954	1,999	2009	0
TYP_BIE	APPARTEMENT	1,953	1,934	1,973	2013	1
TYP_BIE	APPARTEMENT	1,879	1,863	1,896	2017	0
TYP_BIE	AUTRE	1,102	1,085	1,119	2005	
TYP_BIE	AUTRE	1,098	1,083	1,113	2009	1
TYP_BIE	AUTRE	1,121	1,107	1,135	2013	1
TYP_BIE	AUTRE	1,126	1,113	1,139	2017	1
ZON_GEO	Zone 1	1,202	1,169	1,235	2005	
ZON_GEO	Zone 1	1,181	1,156	1,207	2009	1
ZON_GEO	Zone 1	1,138	1,118	1,157	2013	1
ZON_GEO	Zone 1	1,086	1,071	1,102	2017	0
ZON_GEO	Zone 2	1,196	1,181	1,211	2005	
ZON_GEO	Zone 2	1,173	1,161	1,186	2009	1
ZON_GEO	Zone 2	1,158	1,147	1,169	2013	1
ZON_GEO	Zone 2	1,131	1,121	1,140	2017	0

Table C.4 – Résultats des tests pour les autres paramètres

Nom de la variable	Modalité	Estimate	Lower	Upper	Per	Test
FRQ_PAI	ANNUEL	-0,5001	-0,5333	-0,4668	2005	
FRQ_PAI	ANNUEL	-0,1825	-0,2027	-0,1623	2009	0
FRQ_PAI	ANNUEL	-0,0344	-0,0484	-0,0203	2013	0
FRQ_PAI	ANNUEL	-0,0281	-0,0388	-0,0174	2017	1
FRQ_PAI	AUTRE	-0,2544	-0,3242	-0,1847	2005	
FRQ_PAI	AUTRE	-0,0210	-0,0712	0,0292	2009	0
FRQ_PAI	AUTRE	0,1120	0,0724	0,1515	2013	0
FRQ_PAI	AUTRE	0,1060	0,0727	0,1392	2017	1
LIB_NB_PIE	2 pièces ou -	-1,0977	-1,1244	-1,0711	2005	
LIB_NB_PIE	2 pièces ou -	-1,0697	-1,0920	-1,0474	2009	1
LIB_NB_PIE	2 pièces ou -	-1,0567	-1,0757	-1,0376	2013	1
LIB_NB_PIE	2 pièces ou -	-1,0649	-1,0805	-1,0492	2017	1
LIB_NB_PIE	3 ou 4 pièces	-0,5718	-0,5983	-0,5454	2005	
LIB_NB_PIE	3 ou 4 pièces	-0,5728	-0,5948	-0,5507	2009	1
LIB_NB_PIE	3 ou 4 pièces	-0,5637	-0,5825	-0,5449	2013	1
LIB_NB_PIE	3 ou 4 pièces	-0,5512	-0,5666	-0,5358	2017	1
LIB_NB_PIE	5 ou 6 pièces	-0,2640	-0,2908	-0,2372	2005	
LIB_NB_PIE	5 ou 6 pièces	-0,2580	-0,2803	-0,2357	2009	1
LIB_NB_PIE	5 ou 6 pièces	-0,2385	-0,2575	-0,2194	2013	1
LIB_NB_PIE	5 ou 6 pièces	-0,2357	-0,2513	-0,2201	2017	1
LIB_NB_PIE	7 pièces ou +	-0,1109	-0,1396	-0,0823	2005	
LIB_NB_PIE	7 pièces ou +	-0,0924	-0,1163	-0,0684	2009	1
LIB_NB_PIE	7 pièces ou +	-0,0565	-0,0770	-0,0360	2013	1
LIB_NB_PIE	7 pièces ou +	-0,0393	-0,0562	-0,0224	2017	1
QUA_JUR_2	LOCATAIRE	-1,0436	-1,0545	-1,0327	2005	
QUA_JUR_2	LOCATAIRE	-1,0322	-1,0418	-1,0225	2009	1
QUA_JUR_2	LOCATAIRE	-1,0577	-1,0665	-1,0490	2013	0
QUA_JUR_2	LOCATAIRE	-1,0778	-1,0858	-1,0698	2017	0
QUA_JUR_2	PNO	-0,1313	-0,1481	-0,1144	2005	
QUA_JUR_2	PNO	-0,0946	-0,1084	-0,0808	2009	0
QUA_JUR_2	PNO	-0,0741	-0,0860	-0,0623	2013	1
QUA_JUR_2	PNO	-0,0644	-0,0744	-0,0544	2017	1
TYP_BIE	APPARTEMENT	-0,7391	-0,7514	-0,7268	2005	
TYP_BIE	APPARTEMENT	-0,7213	-0,7319	-0,7107	2009	1
TYP_BIE	APPARTEMENT	-0,7313	-0,7407	-0,7218	2013	1
TYP_BIE	APPARTEMENT	-0,7105	-0,7190	-0,7020	2017	0
TYP_BIE	AUTRE	0,1346	0,1208	0,1484	2005	
TYP_BIE	AUTRE	0,1311	0,1182	0,1441	2009	1
TYP_BIE	AUTRE	0,0909	0,0787	0,1032	2013	0
TYP_BIE	AUTRE	0,0558	0,0442	0,0673	2017	0
ZON_GEO	Zone 1	-0,1903	-0,2150	-0,1655	2005	
ZON_GEO	Zone 1	-0,1761	-0,1961	-0,1561	2009	1
ZON_GEO	Zone 1	-0,1567	-0,1730	-0,1405	2013	1
ZON_GEO	Zone 1	-0,1044	-0,1180	-0,0907	2017	0
ZON_GEO	Zone 2	-0,1772	-0,1886	-0,1658	2005	
ZON_GEO	Zone 2	-0,1665	-0,1765	-0,1565	2009	1
ZON_GEO	Zone 2	-0,1594	-0,1683	-0,1504	2013	1
ZON_GEO	Zone 2	-0,1378	-0,1459	-0,1296	2017	0

Table C.5 – Résultats des tests pour les modèles AFT
Paramètres relatifs aux variables

Le tableau C.6 nous présente les résultats obtenus pour les autres paramètres propres aux modèles, à savoir le paramètre lié à l'indicatrice et celui lié à l'échelle.

Nom de la variable	Modalité	Estimate	Lower	Upper	Per	Test
FRQ_PAI	Echelle	1,0304	1,0258	1,0350	2005	
FRQ_PAI	Echelle	1,0373	1,0334	1,0413	2009	1
FRQ_PAI	Echelle	1,0521	1,0486	1,0556	2013	0
FRQ_PAI	Echelle	1,0721	1,0689	1,0753	2017	0
FRQ_PAI	Intercept	1,9752	1,9423	2,0081	2005	
FRQ_PAI	Intercept	1,6466	1,6270	1,6662	2009	0
FRQ_PAI	Intercept	1,5163	1,5030	1,5296	2013	0
FRQ_PAI	Intercept	1,5518	1,5419	1,5617	2017	0
LIB_NB_PIE	Echelle	0,5784	0,5755	0,5813	2005	
LIB_NB_PIE	Echelle	0,5847	0,5822	0,5872	2009	0
LIB_NB_PIE	Echelle	0,5952	0,5930	0,5974	2013	0
LIB_NB_PIE	Echelle	0,6068	0,6048	0,6088	2017	0
LIB_NB_PIE	Intercept	1,9956	1,9711	2,0201	2005	
LIB_NB_PIE	Intercept	1,9606	1,9403	1,9808	2009	1
LIB_NB_PIE	Intercept	1,9435	1,9264	1,9606	2013	1
LIB_NB_PIE	Intercept	1,9618	1,9481	1,9754	2017	1
QUA_JUR_2	Echelle	0,5387	0,5360	0,5413	2005	
QUA_JUR_2	Echelle	0,5467	0,5444	0,5490	2009	0
QUA_JUR_2	Echelle	0,5559	0,5539	0,5580	2013	0
QUA_JUR_2	Echelle	0,5666	0,5647	0,5684	2017	0
QUA_JUR_2	Intercept	1,9366	1,9286	1,9446	2005	
QUA_JUR_2	Intercept	1,9028	1,8958	1,9098	2009	0
QUA_JUR_2	Intercept	1,9035	1,8971	1,9099	2013	1
QUA_JUR_2	Intercept	1,9218	1,9161	1,9276	2017	0
TYP_BIE	Echelle	0,5710	0,5682	0,5739	2005	
TYP_BIE	Echelle	0,5823	0,5798	0,5848	2009	0
TYP_BIE	Echelle	0,5947	0,5925	0,5969	2013	0
TYP_BIE	Echelle	0,6103	0,6083	0,6124	2017	0
TYP_BIE	Intercept	1,6672	1,6589	1,6755	2005	
TYP_BIE	Intercept	1,6543	1,6475	1,6612	2009	1
TYP_BIE	Intercept	1,6762	1,6701	1,6822	2013	0
TYP_BIE	Intercept	1,7158	1,7104	1,7211	2017	0
ZON_GEO	Echelle	1,0241	1,0195	1,0287	2005	
ZON_GEO	Echelle	1,0315	1,0276	1,0354	2009	1
ZON_GEO	Echelle	1,0469	1,0434	1,0503	2013	0
ZON_GEO	Echelle	1,0682	1,0650	1,0713	2017	0
ZON_GEO	Intercept	1,5741	1,5664	1,5818	2005	
ZON_GEO	Intercept	1,5572	1,5504	1,5639	2009	0
ZON_GEO	Intercept	1,5672	1,5610	1,5734	2013	1
ZON_GEO	Intercept	1,5992	1,5935	1,6049	2017	0

Table C.6 – Résultats des tests pour les modèles AFT
Autres paramètres relatifs aux modèlex

Nous observons ainsi que l'hypothèse H_0 est presque toujours rejetée pour le paramètre lié à l'échelle : ceci est assez logique car la variance décroit avec le temps puisque les données s'enrichissent constamment (avec des individus assez homogènes). Le paramètre lié à l'indicatrice n'est pas beaucoup plus stable (sauf pour la variable **Nombre de pièces**). En fait, ce paramètre reflète surtout l'évolution temporelle de la durée de vie des contrats au niveau **global** du portefeuille [8].

8. Il est précisé, dans la documentation de SAS, que ces paramètres correspondent à la moyenne pour celui lié à l'indicatrice et à l'écart-type pour celui lié à l'échelle.

Annexe D

Annexe - Segmentation supervisée à la souscription

Découpage et regroupement des variables

Synthèse pour les variables quantitatives

Variable	Regroup	Nombre polices	Borne Mini	Borne Maxi	Taux résiliation
Nombre de pièces	0	7 327	1	1	72,64
Nombre de pièces	1	9 922	2	2	58,32
Nombre de pièces	2	9 761	3	3	48,32
Nombre de pièces	3	15 241	4	4	38,42
Nombre de pièces	4	10 559	5	5	34,97
Nombre de pièces	5-6	18 559	6	7	27,01
Nombre de pièces	7	10 846	8	37	20,96
Superficie	0-1-2-3	3 690	9	173	33,79
Superficie	4-5-6-7-8-9	5 544	174	2 900	24,30
SMP	0	9 468	1	44 299	53,55
SMP	1	9 470	44 400	88 770	60,37
SMP	2	9 587	88 800	132 000	48,76
SMP	3-4	18 599	132 002	197 981	39,99
SMP	5	9 814	198 000	237 000	36,20
SMP	6	9 422	237 096	281 963	31,05
SMP	7	9 486	282 000	337 390	26,12
SMP	8	9 485	337 440	434 400	22,28
SMP	9	9 482	434 640	1 800 000	19,33
Valeur Mobilier Garanti	1	27 125	0	0	27,99
Valeur Mobilier Garanti	2	8 980	500	4 000	75,55
Valeur Mobilier Garanti	3	10 041	4 020	9 980	55,51
Valeur Mobilier Garanti	4	13 061	10 000	19 960	46,29
Valeur Mobilier Garanti	5	11 869	20 000	39 984	34,10
Valeur Mobilier Garanti	6	11 592	40 000	59 988	27,37
Valeur Mobilier Garanti	7	12 145	60 000	420 000	21,30
Mt cotisation	0	9 480	22	65	58,49
Mt cotisation	1-2	18 965	65	123	50,88
Mt cotisation	3-4-5	28 443	123	212	40,93
Mt cotisation	6	9 481	212	261	33,19
Mt cotisation	7	9 483	261	319	24,58
Mt cotisation	8	9 480	319	424	20,31
Mt cotisation	9	9 481	424	2 951	16,50

Table D.1 – Synthèse Regroupement des variables quantitatives

Regroupement des variables qualitatives

Variable	Cluster	Regroupement	Nombre	Taux résiliation
Départ.	CLUSTER 1	MORBIHAN, MAYENNE, MANCHE, COTES D AMOR, FINISTERE, DEUX SEVRES, SARTHE	40 914	35,49
Départ.	CLUSTER 2	ORNE, INDRE ET LOIRE, ILLE ET VILAINE, VENDEE, CHARENTE MARITIME, CALVADOS, LOIRE ATLANTIQUE, MAINE	51 002	38,95
Départ.	CLUSTER 3	VIENNE	2 183	45,49
Départ.	CLUSTER 4	AUTRES	714	59,52
Formule	CLUSTER 1	000, 254, 261,etc.	4 687	86,43
Formule	CLUSTER 2	062, 134, 244, etc.	44 634	45,17
Formule	CLUSTER 3	063, 242, 162, etc.	41 742	26,37
Formule	CLUSTER 4	015, 025, 054, etc.	3 750	15,55
Réseau	CLUSTER 1	AREAS, MMA, GENERALI, GEST. COMPAGNIE, ALLIANZ, MGA, AXA	48 691	30,85
Réseau	CLUSTER 2	COURTIER, AVIVA, LE FINISTERE, SWISS LIFE, GAN, MUT.DE POITIERS, THELEM	38 423	36,40
Réseau	CLUSTER 3	INCONNU	7 699	88,25
Type d'agent	CLUSTER 1	SIEGE SOCIAL, COURTIER ANNEXE A UNE AGENCE, INTERMEDIAIRE SANS MANDAT	47 055	30,45
Type d'agent	CLUSTER 2	AGENT AVEC MANDAT PRAGA, COURTIER PUR	40 059	36,64
Type d'agent	CLUSTER 3	INCONNU	7 699	88,25
Produit	CLUSTER 1	EVOLUTION, AUTRE	17 212	20,50
Produit	CLUSTER 2	MENHIR, GLOBALE	68 486	37,54
Produit	CLUSTER 3	PROMOTION	4 602	57,45
Produit	CLUSTER 4	ETUDIANT	4 513	86,77

Table D.2 – Synthèse Regroupement des autres variables qualitatives

Autres variables qualitatives

Ces variables sont regroupées selon une logique métier [1].

1. Elles sont présentées de gauche à droite et du haut vers le bas.

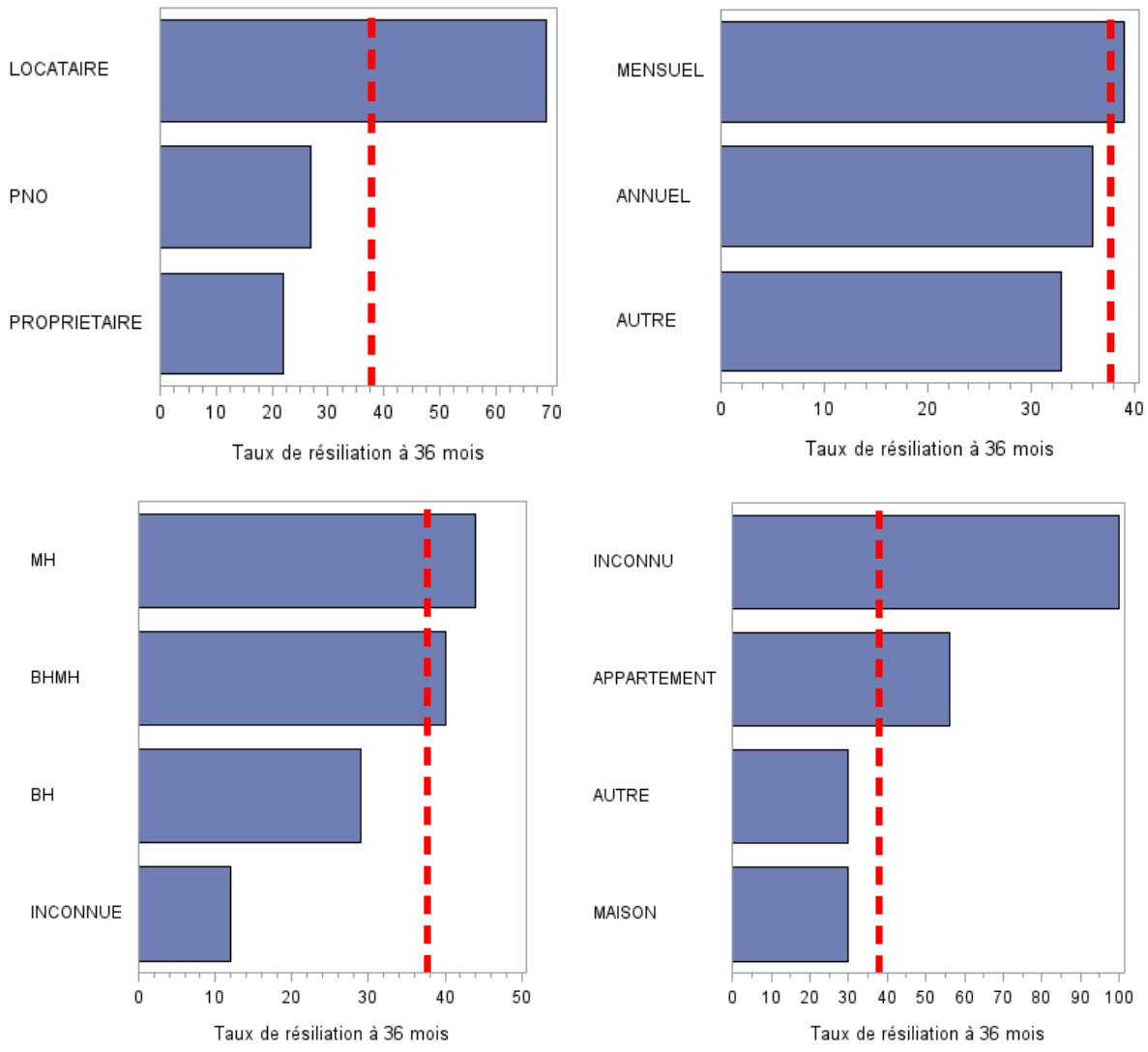


FIGURE D.1 – Variables : Qualité juridique, Fréquence de paiement, Nature des garanties, Type de bien.

Stabilité temporelle des variables explicatives

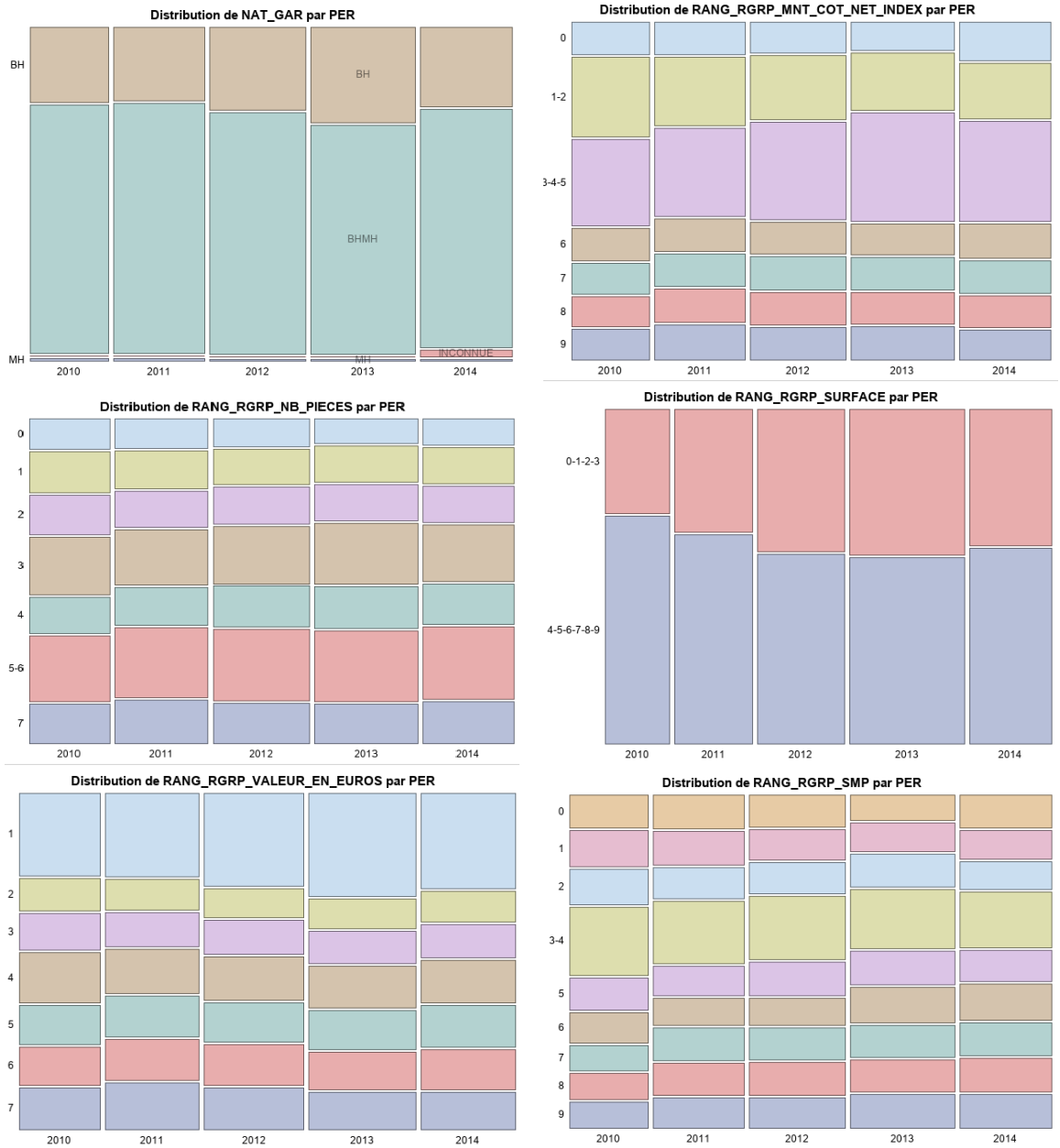


FIGURE D.2 – Synthèse stabilité temporelle des variables (1).

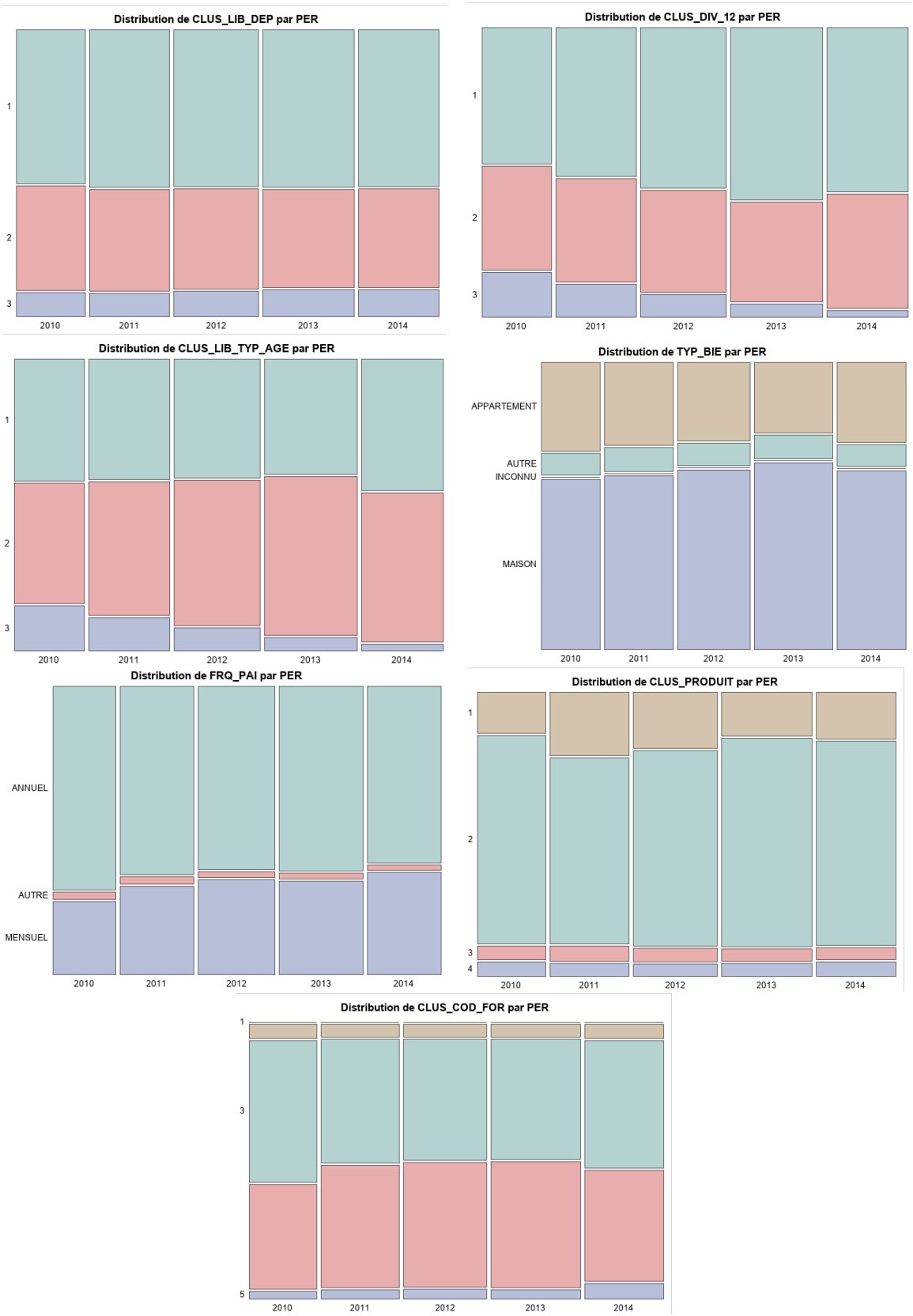


FIGURE D.3 – Synthèse stabilité temporelle des variables (2).

Période >	2010_2011					2011_2012					2012_2013					2013_2014					2010_2014				
Variable	V de Cramer	Khi²		Khi² MH		V de Cramer	Khi²		Khi² MH		V de Cramer	Khi²		Khi² MH		V de Cramer	Khi²		Khi² MH		V de Cramer	Khi²		Khi² MH	
		Valeur	*Proba*	Valeur	*Proba*		Valeur	*Proba*	Valeur	*Proba*		Valeur	*Proba*	Valeur	*Proba*		Valeur	*Proba*	Valeur	*Proba*		Valeur	*Proba*	Valeur	*Proba*
CLUS_COD_FOR	**0,10**	333,62	*0,00*	248,24	*0,00*	**0,04**	59,20	*0,00*	1,76	*0,19*	**0,07**	197,35	*0,00*	31,40	*0,00*	**0,10**	433,42	*0,00*	0,00	*0,98*	**0,07**	1771,21	*0,00*	21,24	*0,00*
CLUS_DIV_12	**0,06**	139,40	*0,00*	121,69	*0,00*	**0,06**	161,04	*0,00*	134,77	*0,00*	**0,07**	222,15	*0,00*	151,31	*0,00*	**0,07**	206,21	*0,00*	1,63	*0,20*	**0,12**	2879,47	*0,00*	1618,03	*0,00*
CLUS_LIB_DEP	**0,02**	9,60	*0,02*	6,94	*0,01*	**0,02**	11,62	*0,01*	0,56	*0,46*	**0,01**	9,21	*0,03*	0,60	*0,44*	**0,01**	3,16	*0,37*	0,47	*0,49*	**0,02**	70,20	*0,00*	0,10	*0,75*
CLUS_LIB_TYP_AGE	**0,07**	147,78	*0,00*	23,92	*0,00*	**0,06**	162,37	*0,00*	21,28	*0,00*	**0,08**	239,77	*0,00*	13,36	*0,00*	**0,07**	220,95	*0,00*	190,93	*0,00*	**0,12**	2952,06	*0,00*	667,17	*0,00*
CLUS_PRODUIT	**0,11**	421,95	*0,00*	216,39	*0,00*	**0,04**	64,69	*0,00*	46,08	*0,00*	**0,06**	139,56	*0,00*	74,25	*0,00*	**0,02**	15,88	*0,00*	0,87	*0,35*	**0,05**	813,61	*0,00*	68,31	*0,00*
NAT_GAR	**0,01**	3,07	*0,38*	0,55	*0,46*	**0,03**	41,63	*0,00*	40,54	*0,00*	**0,04**	75,27	*0,00*	71,85	*0,00*	**0,12**	552,41	*0,00*	193,35	*0,00*	**0,08**	1902,00	*0,00*	29,43	*0,00*
QUA_JUR_2	**0,04**	46,35	*0,00*	46,18	*0,00*	**0,03**	43,08	*0,00*	1,36	*0,24*	**0,05**	93,61	*0,00*	2,95	*0,09*	**0,03**	44,11	*0,00*	0,12	*0,73*	**0,05**	443,05	*0,00*	39,65	*0,00*
RGRP_MNT_COT_NET_INDEX	**0,05**	71,74	*0,00*	43,92	*0,00*	**0,04**	47,68	*0,00*	0,64	*0,42*	**0,04**	67,65	*0,00*	3,64	*0,06*	**0,06**	143,30	*0,00*	25,45	*0,00*	**0,04**	676,83	*0,00*	8,15	*0,00*
RGRP_NB_PIECES	**0,03**	33,38	*0,00*	24,58	*0,00*	**0,03**	21,59	*0,00*	0,81	*0,37*	**0,03**	24,01	*0,00*	1,01	*0,31*	**0,02**	20,61	*0,00*	0,82	*0,36*	**0,02**	159,85	*0,00*	56,73	*0,00*
RGRP_SMP	**0,07**	151,16	*0,00*	60,89	*0,00*	**0,03**	31,71	*0,00*	4,12	*0,04*	**0,06**	164,22	*0,00*	49,94	*0,00*	**0,06**	125,84	*0,00*	1,56	*0,21*	**0,04**	714,26	*0,00*	299,58	*0,00*
RGRP_SURFACE	**-0,06**	10,81	*0,00*	10,81	*0,00*	**-0,06**	12,33	*0,00*	12,33	*0,00*	**-0,01**	0,42	*0,52*	0,42	*0,52*	**0,03**	3,30	*0,07*	3,30	*0,07*	**0,09**	70,94	*0,00*	44,24	*0,00*
RGRP_VALEUR_EN_EUROS	**0,04**	48,64	*0,00*	12,78	*0,00*	**0,04**	53,32	*0,00*	36,20	*0,00*	**0,04**	64,67	*0,00*	51,69	*0,00*	**0,03**	31,94	*0,00*	16,35	*0,00*	**0,03**	399,22	*0,00*	115,90	*0,00*
FRQ_PAI	**0,06**	125,32	*0,00*	124,72	*0,00*	**0,02**	23,11	*0,00*	17,03	*0,00*	**0,00**	0,81	*0,67*	0,80	*0,37*	**0,03**	40,32	*0,00*	36,34	*0,00*	**0,05**	453,34	*0,00*	347,67	*0,00*
TYP_BIE	**0,02**	21,72	*0,00*	10,81	*0,00*	**0,02**	18,62	*0,00*	15,16	*0,00*	**0,03**	40,83	*0,00*	34,98	*0,00*	**0,04**	52,48	*0,00*	46,28	*0,00*	**0,03**	216,12	*0,00*	97,08	*0,00*

FIGURE D.4 – Synthèse stabilité temporelle des variables (3).

Modèle initial considéré

Résultats estimés des paramètres et intervalles de confiance de Wald										Estimations des OR et IC de vraisemblance de profil		
Libellé de la variable	**Libellé de la modalité**	**Paramètre**	**DDL**	**Estimation beta**	**Erreur type**	**Khi-2 de Wald**	**Pr > khi-2**	**IC à 95%**		**Estimation OR**	**IC à 95%**	
		Indicatrice	1	-0,38	0,09	18,79	<.0001	-0,55	-0,21			
Type d'agent	Type d'agent inconnu	CLUS_AGE_3	1	2,56	0,05	3011,90	<.0001	2,47	2,65	12,89	11,77	14,13
Formule	Formules cluster 2	CLUS_FOR_2	1	-2,27	0,08	732,80	<.0001	-2,44	-2,11	0,10	0,09	0,12
Qualité juridique	Qualité juridique locataire	LOCATAIRE	1	1,37	0,03	2800,00	<.0001	1,32	1,42	3,94	3,74	4,14
Formule	Formules cluster 4	CLUS_FOR_4	1	-1,02	0,08	149,24	<.0001	-1,18	-0,86	0,36	0,31	0,43
Croisement Mt cotisation et type de bien	Mt cotisation < 65 €	CR0	1	0,91	0,07	180,15	<.0001	0,78	1,05	2,49	2,18	2,85
Formule	Formules cluster 3	CLUS_FOR_3	1	-0,68	0,08	68,53	<.0001	-0,84	-0,52	0,51	0,43	0,60
Mt de cotisation	Mt cotisation dans (65 € - 123 €)	RGRP_COT_NET_1-2	1	0,66	0,03	481,66	<.0001	0,60	0,72	1,94	1,83	2,06
Croisement Mt cotisation et type de bien	Mt cotisation > 424 €	CR9	1	-0,62	0,04	195,15	<.0001	-0,71	-0,53	0,54	0,49	0,59
Croisement Surface et nb de pièces	Surface > 174 m²	CROIS4_SU_NP456789	1	0,49	0,05	111,71	<.0001	0,40	0,58	1,64	1,49	1,79
Croisement Mt cotisation et type de bien	Mt cotisation dans (319 € - 424 €)	CR8	1	-0,40	0,04	94,34	<.0001	-0,48	-0,32	0,67	0,62	0,73
Croisement Surface et nb de pièces	Surface < 174 m²	CROIS4_SU_NP0123	1	0,33	0,05	44,97	<.0001	0,23	0,43	1,39	1,26	1,53
Croisement Mt cotisation et type de bien	Mt cot dans (123 € -212 €) et type de bien maison	CR345_MAISON	1	0,31	0,03	108,98	<.0001	0,25	0,36	1,36	1,28	1,44
Croisement Mt cotisation et type de bien	Mt cotisation dans (261 € - 319 €)	CR7	1	-0,27	0,04	46,59	<.0001	-0,34	-0,19	0,77	0,71	0,83
Croisement Surface et nb de pièces	4 pièces	CROIS4_SU_NP3	1	-0,24	0,03	75,84	<.0001	-0,29	-0,18	0,79	0,75	0,83
Croisement Surface et nb de pièces	2 pièces	CROIS4_SU_NP1	1	-0,22	0,03	42,44	<.0001	-0,29	-0,16	0,80	0,75	0,85
Département	Départements cluster 1	CLUS_DEP_1	1	-0,13	0,02	43,21	<.0001	-0,17	-0,09	0,88	0,85	0,91

FIGURE D.5 – Résultats de la régression logistique - souscription à 36 mois.

Discussion sur l'horizon du risque mesuré

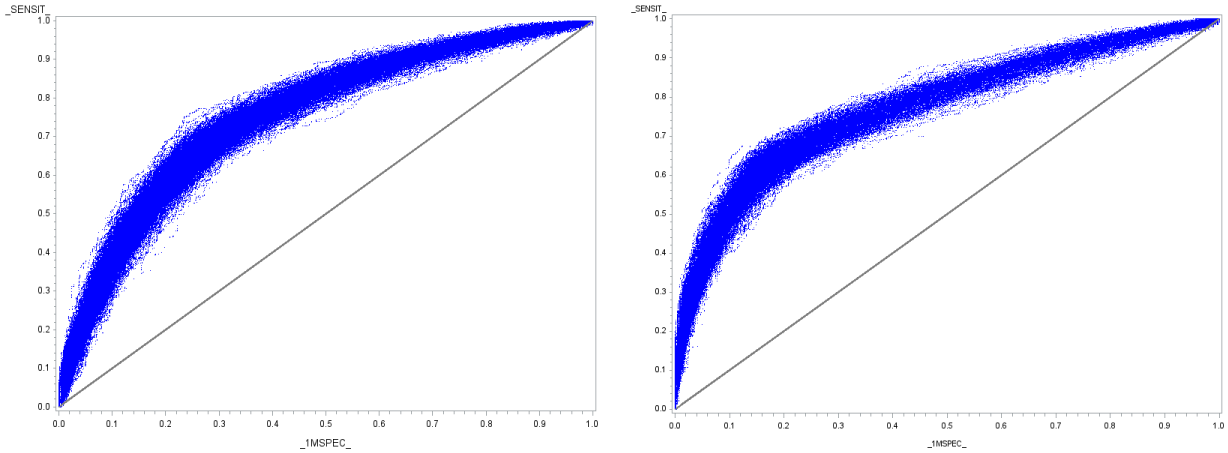


FIGURE D.6 – Courbes ROC des 1 000 échantillons (validation) à 12 et 36 mois.

Annexe E

Annexe - Segmentation supervisée sur stock

Découpage et regroupement des variables

Synthèse pour les variables quantitatives

Variable	Regr.	Nombre polices	Borne Mini	Borne Maxi	Taux résiliation
Nombre de pièces	0	54 120	1	2	18,92
Nombre de pièces	1	115 660	3	4	11,09
Nombre de pièces	2	52 735	5	5	10,15
Nombre de pièces	3	82 263	6	6	8,16
Nombre de pièces	4	30 280	7	7	8,70
Nombre de pièces	5	68 764	8	38	6,99
Superficie	1	424 104	0	0	10,31
Superficie	2	51 378	9	2 900	8,45
SMP	0	52 827	0	80 997	13,01
SMP	1	52 833	81 008	129 976	14,46
SMP	2-3	105 686	129 985	206 250	10,79
SMP	4	52 808	206 252	247 767	9,91
SMP	5-6	105 663	247 768	354 042	8,78
SMP	7-8	105 662	354 052	1 895 990	7,25
Valeur Mobilier Garanti	0	157 257	0	0	8,15
Valeur Mobilier Garanti	1-2	80 482	217	11 687	17,68
Valeur Mobilier Garanti	3	79 225	11 689	30 932	10,65
Valeur Mobilier Garanti	4	79 277	30 933	60 382	8,40
Valeur Mobilier Garanti	5	79 238	60 384	664 869	7,51
Mt cotisation	0	47 554	23	82	14,27
Mt cotisation	1	47 544	82	111	10,96
Mt cotisation	2-3-4	142 640	111	197	11,43
Mt cotisation	5	47 555	197	240	10,00
Mt cotisation	6	47 548	240	283	8,77
Mt cotisation	7-8	95 093	283	449	7,88
Mt cotisation	9	47 548	449	2 977	7,08
Maturité (an)	0	78 064	0	0	14,55
Maturité (an)	1	119 267	1	2	11,61
Maturité (an)	2	87 508	3	4	9,43
Maturité (an)	3	100 146	5	9	8,22
Maturité (an)	4	90 497	10	43	7,06

Table E.1 – Synthèse Regroupement des variables quantitatives (stock).

Regroupement des variables qualitatives

Variable	Cluster	Regroup	Nombre polices	Taux résiliation
Départ.	CLUSTER 1	MORBIHAN, MAYENNE, FINISTERE, SARTHE, ILLE ET VILAINE, MANCHE, COTES D AMOR	288 200	9,32
Départ.	CLUSTER 2	CHARENTE MARITIME, DEUX SEVRES, ORNE, VENDEE, MAINE ET LOIRE, INDRE ET LOIRE, LOIRE ATLANTIQUE, CALV	184 688	11,24
Départ.	CLUSTER 3	AUTRES	2 594	17,58
Formule	CLUSTER 1	114, 234, 006, etc.	26 897	23,68
Formule	CLUSTER 2	024, 153, 015, etc.	447 783	9,31
Formule	CLUSTER 3	060, 113, 121, etc.	802	2,12
Réseau	CLUSTER 1	SWISS LIFE, MMA, ALLIANZ, GENERALI, GAN, GEST. COMPAGNIE, INCONNU, COURTIER, AREAS, THELEM	378 664	10,5
Réseau	CLUSTER 2	AVIVA, MGA, AXA, LE FINISTERE	96 781	8,58
Type d'agent	CLUSTER 1	INTERMEDIAIRE SANS MANDAT, COURTIER ANNEXE A UNE AGENCE, AGENT AVEC MANDAT	298 037	9,64
Type d'agent	CLUSTER 2	INCONNU, COURTIER PUR	174 243	10,97
Type d'agent	CLUSTER 3	CODE SOCIETE SIEGE SOCIAL	3 202	7,59
Produit	CLUSTER 1	MENHIR, GLOBALE, EVOLUTION, AUTRE	450 156	9,36
Produit	CLUSTER 2	PROMOTION	17 531	15,27
Produit	CLUSTER 3	ETUDIANT	7 795	41,94

Table E.2 – Synthèse Regroupement des autres variables qualitatives (stock).

Autres méthodes de segmentation

Boosting

Algorithme Real Adaboost

1. **Initialiser** uniformément le poids des observations : $w_i = 1/n \ \forall i \in \{1, \ldots, n\}$
2. **Pour** $m = 1$ à M :
   (a) Estimer la probabilité $p_m(x) = \mathbb{P}_w [y = 1|x]$ sur l'échantillon d'apprentissage pondéré par les poids w_i.
   (b) Si $p_m(x) = 0$ ou 1, réinitialiser les poids $w_i = 1/n$.
   (c) Sinon :
   - Calculer $\mathrm{H}_m(x) = \frac{1}{2} \log \left[\frac{p_m(x)}{1-p_m(x)} \right]$.
   - Réajuster les poids $w_i = w_i \exp \left[-y_i \mathrm{H}_m(x_i) \right], \ \forall i \in \{1, \ldots, n\}$.
   - Normaliser les poids w_i pour que leur somme fasse 1.
3. **Sortie** : Le classifieur boosté est $\mathrm{H}(x) = \sum_{m=1}^{M} \mathrm{H}_m(x)$

Attention à ne pas prendre M trop grand, car risque de surapprentissage important (contrairement au bagging).

Applications du bagging

	BAGGING classique			BAGGING stepwise			BAGGING type RF (14)			BAGGING type RF (67)		
Type Ech	**APP**	**VAL**	**OOT**	**APP**	**VAL**	**OOT**	**APP**	**VAL**	**OOT**	**APP**	**VAL**	**OOT**
Coeff Variation	1,411	2,464	1,894	2,136	2,436	1,831	3,234	3,870	3,909	1,492	2,527	2,529
Ecart interquartile	0,013	0,021	0,016	0,017	0,020	0,015	0,027	0,030	0,030	0,014	0,022	0,022
Ecart-type	0,010	0,016	0,012	0,014	0,015	0,011	0,020	0,023	0,023	0,010	0,016	0,016
Intervalle	0,123	0,153	0,137	0,083	0,102	0,068	0,119	0,134	0,144	0,061	0,109	0,096
Moyenne	0,689	0,635	0,628	0,647	0,625	0,621	0,606	0,596	0,594	0,668	0,633	0,629
Nb Echantillons	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000
Min	**0,596**	**0,545**	**0,525**	**0,595**	**0,569**	**0,584**	**0,536**	**0,523**	**0,515**	**0,636**	**0,584**	**0,581**
P1	**0,667**	**0,596**	**0,601**	**0,605**	**0,587**	**0,594**	**0,555**	**0,539**	**0,536**	**0,645**	**0,597**	**0,591**
P2.5	**0,671**	**0,605**	**0,606**	**0,612**	**0,594**	**0,598**	**0,565**	**0,548**	**0,545**	**0,650**	**0,602**	**0,598**
P5	**0,673**	**0,609**	**0,610**	**0,622**	**0,599**	**0,602**	**0,571**	**0,556**	**0,555**	**0,652**	**0,607**	**0,602**
P10	**0,676**	**0,616**	**0,614**	**0,631**	**0,606**	**0,607**	**0,580**	**0,565**	**0,563**	**0,655**	**0,613**	**0,608**
P25	**0,682**	**0,625**	**0,620**	**0,640**	**0,616**	**0,614**	**0,593**	**0,581**	**0,579**	**0,661**	**0,621**	**0,618**
P50	**0,689**	**0,635**	**0,628**	**0,648**	**0,625**	**0,622**	**0,607**	**0,597**	**0,595**	**0,668**	**0,632**	**0,629**
P75	**0,695**	**0,646**	**0,636**	**0,656**	**0,636**	**0,629**	**0,620**	**0,612**	**0,609**	**0,675**	**0,644**	**0,640**
P90	**0,700**	**0,655**	**0,643**	**0,664**	**0,645**	**0,636**	**0,630**	**0,623**	**0,624**	**0,681**	**0,654**	**0,650**
P95	**0,704**	**0,661**	**0,647**	**0,668**	**0,650**	**0,639**	**0,636**	**0,633**	**0,631**	**0,684**	**0,660**	**0,656**
P97.5	**0,707**	**0,665**	**0,652**	**0,670**	**0,653**	**0,643**	**0,641**	**0,640**	**0,638**	**0,688**	**0,664**	**0,660**
P99	**0,710**	**0,669**	**0,656**	**0,674**	**0,657**	**0,648**	**0,647**	**0,649**	**0,642**	**0,692**	**0,671**	**0,666**
Max	**0,720**	**0,697**	**0,662**	**0,678**	**0,671**	**0,652**	**0,654**	**0,657**	**0,659**	**0,697**	**0,692**	**0,677**

TABLE E.3 – Synthèse des AUC obtenus par bagging (répartition 90-10).

	BAGGING classique			BAGGING stepwise			BAGGING type RF (14)			BAGGING type RF (67)		
Type Ech	**APP**	**VAL**	**OOT**	**APP**	**VAL**	**OOT**	**APP**	**VAL**	**OOT**	**APP**	**VAL**	**OOT**
Coeff Variation	1,596	3,265	2,320	2,635	3,071	2,204	3,384	4,233	4,321	1,799	3,088	3,197
Ecart interquartile	0,015	0,026	0,019	0,022	0,026	0,018	0,027	0,032	0,034	0,018	0,025	0,027
Ecart-type	0,011	0,020	0,014	0,017	0,019	0,014	0,021	0,025	0,026	0,012	0,019	0,020
Intervalle	0,070	0,133	0,114	0,108	0,121	0,096	0,130	0,158	0,182	0,077	0,134	0,122
Moyenne	0,711	0,623	0,618	0,649	0,618	0,614	0,611	0,595	0,592	0,683	0,625	0,621
Nb Echantillons	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000
Min	**0,674**	**0,554**	**0,563**	**0,587**	**0,565**	**0,570**	**0,542**	**0,501**	**0,486**	**0,647**	**0,560**	**0,562**
P1	**0,683**	**0,575**	**0,584**	**0,605**	**0,576**	**0,580**	**0,557**	**0,529**	**0,532**	**0,655**	**0,584**	**0,574**
P2.5	**0,688**	**0,581**	**0,589**	**0,611**	**0,581**	**0,588**	**0,568**	**0,538**	**0,539**	**0,659**	**0,590**	**0,580**
P5	**0,692**	**0,588**	**0,595**	**0,617**	**0,587**	**0,592**	**0,576**	**0,550**	**0,548**	**0,663**	**0,595**	**0,589**
P10	**0,696**	**0,597**	**0,600**	**0,626**	**0,594**	**0,597**	**0,583**	**0,562**	**0,559**	**0,667**	**0,601**	**0,596**
P25	**0,703**	**0,610**	**0,608**	**0,639**	**0,604**	**0,605**	**0,597**	**0,580**	**0,576**	**0,674**	**0,611**	**0,608**
P50	**0,711**	**0,623**	**0,617**	**0,649**	**0,618**	**0,614**	**0,612**	**0,596**	**0,594**	**0,684**	**0,624**	**0,621**
P75	**0,718**	**0,636**	**0,627**	**0,661**	**0,630**	**0,624**	**0,624**	**0,612**	**0,609**	**0,692**	**0,637**	**0,635**
P90	**0,726**	**0,648**	**0,635**	**0,670**	**0,641**	**0,632**	**0,636**	**0,627**	**0,624**	**0,699**	**0,649**	**0,646**
P95	**0,729**	**0,656**	**0,642**	**0,675**	**0,650**	**0,637**	**0,644**	**0,635**	**0,634**	**0,703**	**0,658**	**0,653**
P97.5	**0,733**	**0,663**	**0,645**	**0,680**	**0,656**	**0,639**	**0,649**	**0,640**	**0,642**	**0,706**	**0,667**	**0,660**
P99	**0,738**	**0,671**	**0,653**	**0,684**	**0,664**	**0,644**	**0,656**	**0,647**	**0,649**	**0,709**	**0,675**	**0,668**
Max	**0,744**	**0,687**	**0,677**	**0,695**	**0,687**	**0,666**	**0,672**	**0,659**	**0,668**	**0,724**	**0,694**	**0,684**

TABLE E.4 – Synthèse des AUC obtenus par bagging (répartition 50-50).

Applications du boosting

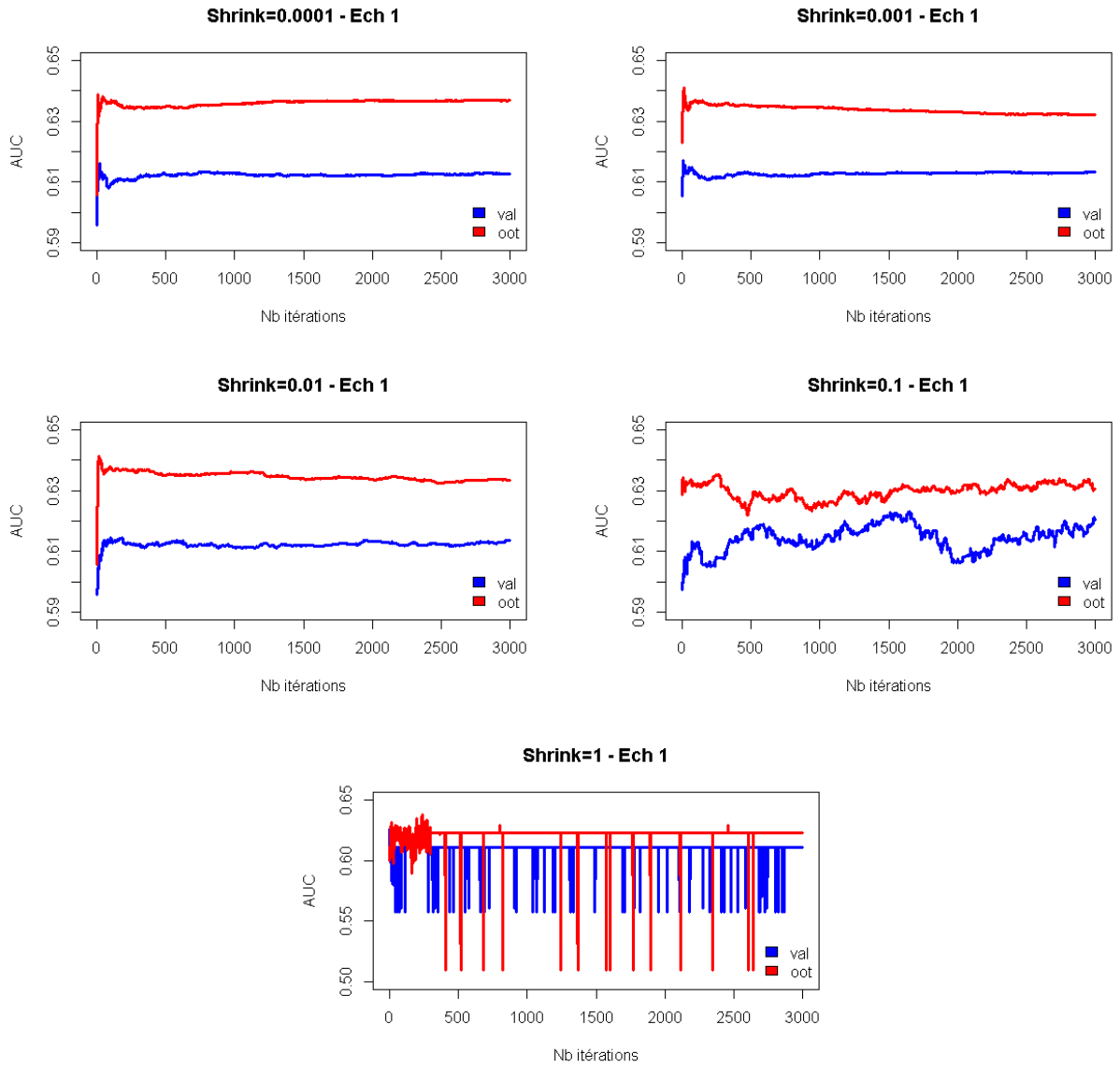


FIGURE E.1 – Boosting - Echantillon 1 (val et oot) - 3 000 itérations.

Ech	**Shrink λ**	**AUC maxi Val**	**AUC maxi Oot**	**It AUC maxi Val**	**It AUC maxi Oot**	**AUC Val**	**AUC Oot**
1	0,0001	0,6162	0,6387	18	6	0,6126	0,6368
2	0,0001	0,6284	0,6200	20	5	0,6227	0,6123
3	0,0001	0,6271	0,6488	319	8	0,6263	0,6470
4	0,0001	0,6592	0,6255	23	22	0,6552	0,6230
5	0,0001	0,6233	0,6269	6	42	0,6111	0,6255
6	0,0001	0,6220	0,6433	11	2440	0,6175	0,6428
7	0,0001	0,6384	0,6367	6	116	0,6293	0,6338
8	0,0001	0,6272	0,6261	8	5	0,6162	0,6196
9	0,0001	0,6275	0,6331	86	980	0,6256	0,6328
10	0,0001	0,6286	0,6140	2931	7	0,6284	0,5972
1	0,001	0,6170	0,6411	6	13	0,6132	0,6321
2	0,001	0,6369	0,6142	30	2	0,6221	0,6113
3	0,001	0,6301	0,6477	2	162	0,6280	0,6461
4	0,001	0,6596	0,6182	7	2972	0,6548	0,6181
5	0,001	0,6118	0,6285	143	54	0,6050	0,6267
6	0,001	0,6180	0,6455	36	253	0,6131	0,6445
7	0,001	0,6284	0,6350	872	259	0,6264	0,6333
8	0,001	0,6310	0,6208	3	33	0,6176	0,6206
9	0,001	0,6333	0,6345	6	92	0,6275	0,6329
10	0,001	0,6413	0,6063	16	23	0,6303	0,5999
1	0,01	0,6145	0,6414	185	16	0,6135	0,6332
2	0,01	0,6326	0,6195	27	7	0,6195	0,6105
3	0,01	0,6382	0,6466	6	380	0,6231	0,6411
4	0,01	0,6492	0,6211	23	841	0,6455	0,6200
5	0,01	0,6201	0,6303	25	8	0,5975	0,6260
6	0,01	0,6141	0,6470	193	78	0,6112	0,6423
7	0,01	0,6330	0,6357	2	38	0,6237	0,6308
8	0,01	0,6323	0,6246	2	1426	0,6130	0,6230
9	0,01	0,6305	0,6421	190	28	0,6246	0,6326
10	0,01	0,6370	0,6029	21	2701	0,6285	0,6022
1	0,1	0,6230	0,6353	1648	262	0,6205	0,6304
2	0,1	0,6290	0,6188	151	790	0,6232	0,6134
3	0,1	0,6311	0,6469	65	2	0,6214	0,6380
4	0,1	0,6570	0,6210	2921	318	0,6554	0,6120
5	0,1	0,6213	0,6316	3	664	0,5764	0,6271
6	0,1	0,6281	0,6483	10	111	0,6126	0,6307
7	0,1	0,6335	0,6387	10	79	0,6289	0,6321
8	0,1	0,6232	0,6294	31	315	0,6130	0,6247
9	0,1	0,6312	0,6348	48	63	0,6135	0,6073
10	0,1	0,6371	0,6121	2	6	0,6250	0,6037
1	1	0,6270	0,6381	6	235	0,6105	0,6231
2	1	0,6291	0,6278	514	62	0,5810	0,5565
3	1	0,6356	0,6420	8	400	0,6055	0,5339
4	1	0,6485	0,6304	4	2513	0,6088	0,6188
5	1	0,6229	0,6329	3	2	0,5909	0,5579
6	1	0,6175	0,6477	12	202	0,5741	0,6238
7	1	0,6243	0,6498	12	36	0,5199	0,5505
8	1	0,6300	0,6249	14	34	0,6115	0,5659
9	1	0,6279	0,6366	167	74	0,5744	0,5302
10	1	0,6460	0,6198	785	2381	0,6403	0,5935

Table E.5 – Détails des AUC obtenus par boosting.

Élasticité du taux de résiliation à la prime

Remarques

- L'élasticité est une quantité sans unité.
- L'élasticité est formellement calculée à la limite. L'utilisation des variations est une approximation.
- Exemple : si l'élasticité est égale à -3, cela veut dire qu'une augmentation de 1% du prix va diminuer la demande de 3%.

Intérêt pour une entreprise L'élasticité-prix de la demande est calculée par les entreprises pour fixer un prix de vente :

- Si le bien a un prix peu élastique, une hausse du prix peut être envisagée.
- Plus le bien a un prix élastique, plus il est délicat d'augmenter son prix de vente.

Modèle logistique de l'élasticité à la prime Cette partie s'inspire notamment des mémoires de [Chi11] et [Fon11].

Soit p la probabilité de résilier que nous cherchons à estimer avec $\hat{p}$. Nous rappelons que le modèle logistique de départ est le suivant :

$$\text{logit}(\hat{p}) = X\hat{\beta}$$

La première idée est de décomposer les variables explicatives en deux blocs :

- Un premier bloc, **indépendant de la prime**, mesurant des effets intrinsèques au bien assuré.
- Un second bloc, mesurant les **interactions** entre le niveau de prime P et certaines variables explicatives.

Le modèle logistique du taux de résiliation, tenant compte de l'élasticité à la prime, peut ainsi être formalisé comme tel :

$$\text{logit}(\hat{p}(P)) = X\hat{\beta} + P \cdot \Pi\hat{\pi} \tag{E.1}$$

Avec :

- X la matrice des variables explicatives indépendantes de la prime et $\hat{\beta}$ les coefficients associés.
- P le montant de la prime.
- Π la matrice des variables explicatives en interaction avec la prime et $\hat{\pi}$ les coefficients associés.

La seconde idée est de faire le lien entre ce dernier modèle logistique et la formule infinitésimale de l'élasticité, décrite ci-dessous et généralisant l'équation (5.1) (page 81) :

$$\hat{E}(P) = \frac{\frac{\partial \hat{p}(P)}{\hat{p}(P)}}{\frac{\partial P}{P}} = \frac{\partial \hat{p}(P)}{\partial P} \times \frac{P}{\hat{p}(P)}$$

Or d'après l'équation (E.1), nous avons :

$$\frac{\partial \hat{p}(P)}{\partial P} = -\frac{-\Pi\hat{\pi}\,\mathrm{e}^{-X\hat{\beta}-P\cdot\Pi\hat{\pi}}}{\left[1+\mathrm{e}^{-X\hat{\beta}-P\cdot\Pi\hat{\pi}}\right]^2}$$

Comme le **taux de renouvellement** du portefeuille, soit $1-\hat{p}(P)$, est égal à $\left[1+\mathrm{e}^{X\hat{\beta}+P\cdot\Pi\hat{\pi}}\right]^{-1}$, nous en déduisons que :

$$\hat{E}(P) = P \cdot \Pi\hat{\pi}[1 - \hat{p}(P)] \tag{E.2}$$

Ainsi, l'élasticité peut se déduire du montant de la prime, du taux de renouvellement et du vecteur des coefficients associés aux interactions de la prime avec certaines variables explicatives.
Si nous nous plaçons au niveau d'un contrat i, en supposant que la matrice Π possède $m+1$ colonnes et que $\Pi_0 = i_0 = 1$, alors nous avons :

$$\hat{E}(P)_i = P(i)[1 - \hat{p}(P)_i] \sum_{k=0}^{m} \pi_k \mathbb{1}_{[\Pi_k = i_k]} = P(i)[1 - \hat{p}(P)_i]\pi(i)$$

L'étude aux limites en 0 ou en $+\infty$ montre que l'élasticité tend vers zéro si nous admettons que $\pi(i)$ est positif.

Modèle logistique de l'élasticité à la variation de prime Le principe est exactement le même sauf que nous ne nous intéressons plus au montant de la prime annuelle P mais à sa variation annuelle[1] ΔP.

$$\Delta P = \frac{P_N - P_{N-1}}{P_{N-1}}$$

Application à nos données

Analyse des effets Type 3			
Effet	**DDL**	**Khi-2 de Wald**	**Pr > khi-2**
Qualité Juridique	2	1046,20	<.0001
Valeur de la garantie	4	39,55	<.0001
Nb_de_pièces_Surface	7	624,31	<.0001
Type de bien	2	67,36	<.0001
Maturité	4	1027,00	<.0001
Département	2	23,35	<.0001
Fréquence de paiement	2	18,88	<.0001
Type agent	2	15,30	0.0005
Mnt prime	1	57,06	<.0001
Mnt prime*Qualité Juridique	2	91,37	<.0001
Mnt prime*Valeur de la garantie	4	23,76	<.0001
Mnt prime*Nb_de_pièces_Surface	7	218,17	<.0001
Mnt prime*Type de bien	2	26,98	<.0001
Mnt prime*Maturité	4	307,41	<.0001
Mnt prime*Département	2	15,94	0.0003

Table E.6 – Résultat de l'ajustement logistique de l'élasticité à la prime

1. Entre les années $N-1$ et N.

Annexe F

Annexe - La résiliation dans Solvabilité 2

Méthode actuelle

Période observation	Nb résiliations à l'échéance avec délai	Nb résiliations à l'échéance hors délai	Nb résiliations hors échéance	Nb résiliations	Nb contrats stock N-1	Taux résiliation à l'échéance avec délai	Taux résiliation à l'échéance hors délai	Taux résiliation hors échéance	Taux de chute Existants	Taux de chute Futurs	Taux de résiliation
2012	2 752	3 830	5 400	11 982	101 221	2,72%	3,78%	5,33%	5,33%	9,12%	11,84%
2013	3 102	3 909	5 549	12 560	107 997	2,87%	3,62%	5,14%	5,14%	8,76%	11,63%
2014	2 996	3 886	5 960	12 842	114 922	2,61%	3,38%	5,19%	5,19%	8,57%	11,17%
2015	2 478	3 524	7 928	13 930	118 669	2,09%	2,97%	6,68%	6,68%	9,65%	11,74%
2016	1 361	2 680	9 344	13 385	120 998	1,12%	2,21%	7,72%	7,72%	9,94%	11,06%
2017	1 194	2 166	10 137	13 497	123 194	0,97%	1,76%	8,23%	8,23%	9,99%	10,96%
2018	1 037	2 256	10 469	13 762	126 332	0,82%	1,79%	8,29%	8,29%	10,07%	10,89%

FIGURE F.1 – Comptages résiliations

Evolution de la méthode actuelle

Modèle de régression logistique multinomial

Nous avons vu jusqu'à présent comment modéliser une réponse **binaire**. Le problème est désormais de généraliser le modèle à une réponse à plus de deux modalités.
En complément des modèles de régression logistiques binaires, nous avons choisi de nous pencher sur le cas d'un modèle **multinomial**. Nous en profiterons ainsi pour comparer les résultats des deux approches.
La variable cible d'un tel modèle peut être soit **polytomique**, soit **ordinale**. Dans notre étude, nous souhaitons distinguer les **résiliations hors échéance** des **résiliation à l'échéance hors délai de préavis**. Aucune structure d'ordre n'est donc supposée en sortie : nous nous trouvons ainsi dans le cadre d'un modèle polytomique.

Loi multinomiale

Cette loi généralise la loi binomiale [1]. Considérons $(N_i)_{1\leq i\leq k}$ variables aléatoires de loi binomiale $\mathcal{B}(n_i, p_i)$.
Nous imposons les deux contraintes suivantes : $\sum_{i=1}^{k} n_i = n$ et $\sum_{i=1}^{k} p_i = 1$.
La loi du vecteur multinomial se formalise ainsi :

$$\mathbb{P}\left[N_1 = n_1, \ldots, N_k = n_k\right] = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k}$$

Chacune des variables aléatoires du vecteur est supposée suivre une loi binomiale dont les propriétés d'espérance et de variance sont les suivantes $\forall i \in \{1, \ldots, k\}$:

1. Schématiquement, et dans le cas équiprobable, nous passons d'un lancer de pièce à un lancer de dé.

$$\mathbb{E}[N_i] = np_i \qquad \text{et} \qquad \mathbb{V}[N_i] = np_i(1-p_i)$$

Les covariances se déduisent :

$$\mathbb{C}[N_i, N_j] = np_ip_j$$

Ainsi cette loi peut être formalisée par un vecteur $N = (N_i)_{1\leq i\leq k}$ dont les composantes sont binomiales mais non indépendantes [2].

Formalisation du modèle

Soit Y la variable réponse à k modalités et X la matrice des variables explicatives. Nous noterons $p_i = P[Y = i|X], \forall i \in \{1, \ldots, k\}$. Le modèle polytomique est le suivant :

$$\log\left(\frac{p_i}{p_k}\right) = X\beta^i, \forall i \in \{1, \ldots, k-1\} \tag{F.1}$$

Nous appelons **logit généralisé** la quantité $\log(p_i/p_k)$. La dernière modalité de la variable cible Y est souvent celle prise comme référence [3].

Nous déduisons de l'équation (F.1) que $p_i = p_k\,\mathrm{e}^{X\beta^i}$. Comme par construction $\sum_{j=1}^{k} p_j = 1$, alors :

$$\begin{aligned} p_k &= 1 - \sum_{i=1}^{k-1} p_i \\ &= 1 - \sum_{i=1}^{k-1} p_k\,\mathrm{e}^{X\beta^i} \\ &= 1 - p_k \sum_{i=1}^{k-1} \mathrm{e}^{X\beta^i} \end{aligned}$$

Ce qui revient à écrire :

$$p_k\left[1 + \sum_{j=1}^{k-1} \mathrm{e}^{X\beta^j}\right] = 1$$

Au final, nous pouvons présenter une écriture plus directe des probabilités p_i :

$$\begin{cases} p_k = \frac{1}{1+\sum_{j=1}^{k-1} \mathrm{e}^{X\beta^j}} \\ \\ p_i = \frac{\mathrm{e}^{X\beta^i}}{1+\sum_{j=1}^{k-1} \mathrm{e}^{X\beta^j}} \end{cases}$$

La loi multinomiale appartient, comme la loi binomiale, à la famille exponentielle. Sa fonction de lien est appelée le **logit généralisé** [4].
Ses paramètres $\beta = \left(\beta^i\right)_{1\leq i\leq k-1}$ sont estimés par maximisation de la log-vraisemblance du modèle à l'aide de la méthode de Newton-Raphson.

2. La loi du vecteur est notée $N \sim \mathcal{M}ult(n, p_1, \ldots, p_k)$.
3. Nous avons donc $k-1$ modèles et le dernier (référence) est déduit des autres.
4. Attention à bien paramétrer, dans SAS, la procédure **logistic** avec l'option **link=glogit**.

Table des figures

1 Schéma des différents axes d'analyse du risque. 5
2 Objectif et méthodes proposées. 5
3 Évolution mensuelle du taux de résiliation. 6
4 Hétérogénéité des variables explicatives. 7
5 Diagram of the different risk analysis axes. 10
6 Objective and proposed methods. 10
7 Monthly evolution of the termination rate. 11
8 Heterogeneity of explanatory variables. 12

1.1 Répartition de l'encaissement par produit pour 2019 19

2.1 Évolution annuelle des affaires nouvelles et résiliations. 27
2.2 Évolution annuelle du stock et des résiliations. 27
2.3 Évolution mensuelle et annuelle du taux de résiliation. 28

3.1 Estimateur KM de la survie sur le portefeuille MRH 35
3.2 Durée de vie résiduelle moyenne sur le portefeuille MRH 35
3.3 Comparaison des fonctions de survie KM et Bernoulli. 37
3.4 Courbes LLS 41
3.5 Résidus de Schœnfeld 42
3.6 Résultats des tests pour les survies de référence 46
3.7 Evolution de la survie en fonction de la maturité. 49

4.1 Découpage de la variable Nombre de pièces 53
4.2 Variable Département du risque - Arbre 55
4.3 Variable Département du risque - R^2 55
4.4 Regroupements de la variable Département du risque 56
4.5 Liaison des variables explicatives avec la variable cible 58
4.6 Liaison des variables explicatives entre elles 59
4.7 Évolution temporelle de la variable Qualité juridique 60
4.8 Régression linéaire sur chacune des trois modalités 60
4.9 Analyse des Correspondances Multiples 62
4.10 Complexité et erreur (source :[BMTW12]) 64
4.11 Courbe ROC sur l'échantillon d'apprentissage. 67
4.12 Distribution des deux évènements modélisés. 68
4.13 Taux de risque en fonction de la classe de score. 68
4.14 Comparaison des coefficients 12M et 36M 69
4.15 Distribution bootstrap des AUC (validation) à 12 et 36 mois. 70

5.1 Courbes ROC - Apprentissage, Validation et Out-Of-Time 73
5.2 Principe du *boosting* (source : [SF12]). 77
5.3 Exemples de fonctions de pertes (source : [FHT00]). 77
5.4 Comparaison de la distribution des AUC pour 90-10 et 50-50. 79
5.5 Distribution des variables Primes et Variation 82
5.6 Distributions de la variable Elasticité sur les modèles Prime et Variation 83

A.1 Organigramme simplifié du groupe Covéa au 01/01/2019 94

B.1 Distribution de la variable Nombre de pièces 96
B.2 Distribution de la variable Surface du bien 97
B.3 Distribution de la variable Valeur du mobilier garanti 97
B.4 Distribution de la variable Valeur du SMP 98
B.5 Distribution de la variable Montant de cotisation 99
B.6 Distribution de la variable Département 100
B.7 Distribution de la variable Qualité juridique 100
B.8 Distribution de la variable Type de bien 101
B.9 Distribution de la variable Type de produit 101
B.10 Distribution de la variable Type de formule 102
B.11 Distribution de la variable Nature des garanties 103
B.12 Distribution de la variable Fréquence de paiement 103
B.13 Distribution de la variable Réseau 104
B.14 Distribution de la variable Type d'agent 104

C.1 Comparaison des approches. 112
C.2 Comparaison des étendues entre les bornes inf et sup des deux approches. 113
C.3 Résidus de Cox-Snell pour la variable Qualité juridique 114
C.4 Résidus de Cox-Snell pour les cinq variables : QJ, TB, NP, FP et ZG (resp.) 115

D.1 Variables : Qualité juridique, Fréquence de paiement, Nature des garanties, Type de bien. 125
D.2 Synthèse stabilité temporelle des variables (1). 126
D.3 Synthèse stabilité temporelle des variables (2). 127
D.4 Synthèse stabilité temporelle des variables (3). 128
D.5 Résultats de la régression logistique - souscription à 36 mois. 129
D.6 Courbes ROC des 1 000 échantillons (validation) à 12 et 36 mois. 130

E.1 Boosting - Echantillon 1 (val et oot) - 3 000 itérations. 137

F.1 Comptages résiliations 141