

Mémoire présenté devant l'ENSAE Paris  
pour l'obtention du diplôme de la filière Actuariat  
et l'admission à l'Institut des Actuaires  
le 16/03/2022

Par : Boris Dilane NOUMEDEM DJIEUZEM

Titre : De la communauté à l'individu : Quantifier pour comprendre  
l'impact des pratiques actuarielles sur la mutualisation

Confidentialité :  NON  OUI (Durée :  1 an  2 ans)

*Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus*

*Membres présents du jury de la filière*

Caroline HILLAIRET

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

*Entreprise :*

 BNP PARIBAS CARDIF

*Signature :*



*Directeur du mémoire en entreprise :*

Clément SAILLARD

*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)**

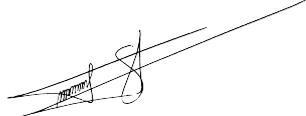
Signature du responsable entreprise



Secrétariat :

Bibliothèque :

Signature du candidat





## REMERCIEMENTS

La concrétisation de ce mémoire a été possible grâce au concours de plusieurs personnes envers qui je souhaite témoigner ma gratitude.

- ☞ En premier lieu, je remercie Pauline LECLERC-GLORIEUX de m'avoir accordé ce stage au sein de BNP Paribas Cardif et également Renaud DUMORA qui est à l'initiative du sujet que nous avons traité ;
- ☞ Je remercie tout particulièrement mon tuteur professionnel, Clément SAILLARD, de m'avoir accueilli et intégré dans son équipe, pour s'être si souvent rendu disponible, pour nos échanges et brainstormings, pour la qualité de son encadrement, sa sollicitude à mon endroit et ses précieux conseils. Ces quelques mois m'aideront bien au-delà de ce stage ;
- ☞ Également, je remercie tous mes collaborateurs de l'équipe *Pricing P&C* pour leur accueil, leurs enseignements et le cadre de travail chaleureux qui a été le nôtre durant ces 6 mois ; un merci particulier à Emmanuelle BRAIVE et à Pierre MEUNIER pour leur encadrement quant au produit d'extension de garantie et leur expertise sur la tarification dudit produit ;
- ☞ Mes remerciements vont également à l'endroit du corps enseignants de l'ENSAE, pour la richesse et la qualité de leurs enseignements. Spécialement à Caroline HILLAIRET, mon référent pédagogique au cours de ce stage, j'adresse mes remerciements pour la validation et le suivi du sujet ;
- ☞ À vous mes parents qui m'avez donné la vie, éduqué, et recouvert d'amour depuis mon enfance, je vous adresse mes infinis remerciements. Ce travail c'est à vous que je le dédie ;
- ☞ Je ne saurais terminer sans remercier mes amis Reine TEUFO, Arielle TSOFAK, Line MONKAM, Christine AGBOTSE et Maradonna CLINTON pour leur soutien émotionnel pendant ces années passées à l'ENSAE ou pour les relectures de ce mémoire.

Enfin, j'aimerais remercier toutes les personnes m'ayant apporté leurs aides et remarques afin que ce travail soit finalisé.

# TABLE DES MATIÈRES

<b>Remerciements</b> .....		<b>i</b>
<b>Résumé</b> .....		<b>v</b>
<b>Abstract</b> .....		<b>vi</b>
<b>Note de Synthèse</b> .....		<b>vi</b>
<b>Introduction</b> .....		<b>1</b>
<b>1 Mutualisation et démutualisation : enjeux et propositions de formalisation</b> .....		<b>3</b>
1.1 Nécessité de définir et de mesurer la mutualisation.....		3
1.1.1 Qu'est ce que la mutualisation des risques?.....		3
1.1.2 L'individualisation des risques et les conséquences sociales.....		6
1.2 Partage du risque entre assureur et assurés.....		8
1.2.1 Partage du risque et mutualisation (Denuit et Charpentier, 2004).....		9
1.2.2 Partage du risque en présence de chargements assureurs.....		14
1.3 Proposition de définition de la démutualisation.....		16
<b>2 Cadre théorique et propositions d'indicateurs de mutualisation</b> .....		<b>19</b>
2.1 Introduction à l'entropie sous plusieurs formes.....		20
2.1.1 Entropie en thermodynamique.....		20
2.1.2 Entropie en théorie de l'information.....		25
2.1.3 Entropie en statistique et en machine learning.....		26
2.1.4 Limite de l'entropie comme mesure d'hétérogénéité : <i>coarse graining</i> .....		27
2.1.5 Comparaison entre entropie et variance.....		27
2.2 Indices de mutualisation simples basés sur la prime et ses composantes.....		30
2.2.1 Construction de la prime pure : GLM et LightGBM.....		30
2.2.2 Indices de mutualisation calculés sur la prime et ses composantes.....		42
2.2.3 Indices de mutualisation calculés sur l'OL.....		47
2.2.4 Vers les indicateurs exotiques.....		52
2.3 Introduction à la complexité de Kolmogorov.....		52
2.3.1 Définition et formalisation de la complexité de Kolmogorov.....		53
2.3.2 Lien entre complexité de Kolmogorov, hétérogénéité et mesure de l'information.....		57
2.3.3 Méthodes d'approximation de la complexité de Kolmogorov.....		59
2.4 Indices de mutualisation exotiques basés sur les caractéristiques des assurés.....		63
2.4.1 Indice de démutualisation basé sur les contributions au risque des caractéristiques.....		63
2.4.2 Indice de mutualisation basé sur la complexité de Kolmogorov.....		69

2.4.3	Portefeuille de marché, un socle commun pour la comparaison de la mutualisation entre assureurs . . . . .	77
<b>3</b>	<b>Description de l'extension de garantie automobile et étude du portefeuille.</b>	<b>80</b>
3.1	L'assurance extension de garantie . . . . .	80
3.1.1	Les produits d'assurance affinitaires . . . . .	80
3.1.2	Le principe de l'assurance extension de garantie automobile . . . . .	82
3.2	Préparation des données : de la base brute à la base finale . . . . .	84
3.2.1	Présentation des données brutes . . . . .	84
3.2.2	Principaux traitements préliminaires . . . . .	85
3.2.3	Présentation des données finales . . . . .	89
3.3	Analyses descriptives du portefeuille d'extension de garantie automobile de BNP Paribas Cardif . . . . .	91
3.3.1	Étude de la dynamique de la sinistralité historique . . . . .	91
3.3.2	Facteurs de risque et structure du portefeuille . . . . .	92
3.3.3	Base d'apprentissage, base de test et base de comparaison des indicateurs de mutualisation . . . . .	95
<b>4</b>	<b>Évaluation et analyse de la mutualisation à partir des indicateurs simples .</b>	<b>98</b>
4.1	Calibration d'un modèle de calcul de la prime pure . . . . .	98
4.1.1	Modélisation de la prime pure à partir d'un GLM . . . . .	99
4.1.2	Modélisation de la prime pure à partir d'un LightGBM . . . . .	101
4.1.3	Comparaison des modèles de construction du Burning Cost . . . . .	102
4.1.4	Interprétation des modèles de calcul des profils de risque individualisés . . . . .	104
4.2	Applications numériques des indices de mutualisation simples . . . . .	108
4.2.1	Construction de la prime technique tarifée . . . . .	108
4.2.2	Évaluation et interprétation du niveau de démutualisation sur le portefeuille EG de BNP Paribas Cardif en 2020 . . . . .	109
4.2.3	Analyses numériques de propriétés des indicateurs de démutualisation vanilles . . . . .	115
4.3	Utilisations pratiques des indices de mutualisation basés sur la prime . . . . .	117
4.3.1	Indices de mutualisation pour guider le choix du niveau de segmentation tarifaire . . . . .	117
4.3.2	Comparaison du niveau de mutualisation entre assureurs et analyse des effets de la démutualisation . . . . .	119
<b>5</b>	<b>Évaluation et analyse de la mutualisation à partir des indicateurs exotiques</b>	<b>133</b>
5.1	Mesure de la démutualisation à partir de l'inertie des contributions au risque des variables tarifaires . . . . .	133
5.1.1	Calcul et visualisation des contributions au risque . . . . .	133
5.1.2	Part de risque de structure généralisée comme mesure de démutualisation . . . . .	135
5.2	Mesure de la mutualisation à partir de l'entropie algorithmique : le ratio de mutualisation généralisé . . . . .	140
5.2.1	Construction du <i>splitter</i> à partir d'un arbre CART . . . . .	140
5.2.2	Description et encodage du portefeuille à partir du <i>splitter</i> . . . . .	142

5.2.3	Approximation de la complexité de Kolmogorov pour quantifier la mutualisation connue . . . . .	146
5.2.4	Évaluation de la mutualisation résiduelle pour une description partielle par l'entropie de Shannon des micro-états . . . . .	149
5.2.5	Mesure de la mutualisation par le ratio de mutualisation généralisé . . . . .	151
<b>6</b>	<b>Synthèses, limites et perspectives pour la mesure de la mutualisation . . . . .</b>	<b>158</b>
6.1	Synthèses des indicateurs de mutualisation proposés. . . . .	158
6.1.1	Les usages pratiques de la mesure de la mutualisation . . . . .	158
6.1.2	Souhaits et attentes d'une métrique de mutualisation . . . . .	159
6.1.3	Les indicateurs proposés répondent-ils aux attentes ? . . . . .	160
6.2	Limites des indices de mutualisation. . . . .	163
6.2.1	Nécessité de disposer des informations sur le marché pour assurer la comparabilité externe. . . . .	163
6.2.2	La justice et l'équité des pratiques actuarielles pas directement prises en compte. . . . .	164
6.3	Perspectives des indicateurs de mutualisation. . . . .	166
6.3.1	Améliorer la conception et l'interprétation des indicateurs . . . . .	166
6.3.2	Perspectives chez BNP Paribas Cardif. . . . .	166
	<b>Conclusion . . . . .</b>	<b>168</b>
	<b>Glossaire et Acronyme . . . . .</b>	<b>170</b>
	<b>Liste des symboles. . . . .</b>	<b>172</b>
	<b>Executive summary. . . . .</b>	<b>172</b>
	<b>Bibliographie . . . . .</b>	<b>182</b>
	<b>Table des annexes . . . . .</b>	<b>187</b>
A	Maximisation de la variance et de l'entropie. . . . .	188
A.1	Maximisation de la variance . . . . .	188
A.2	Maximisation d'entropie. . . . .	189
B	Analyses descriptives complémentaires du portefeuille d'extension de garantie automobile . . . . .	190
C	Analyses complémentaires pour la modélisation du Burning Cost à partir d'un modèle de régression . . . . .	194
D	Approximation de la complexité de Kolmogorov de la description du portefeuille par la BDM et la Lossless compression . . . . .	199
E	Approximation de la complexité de Kolmogorov par l'entropie : robustesse à la taille de la décomposition . . . . .	201

## RÉSUMÉ

L'assurance repose sur la mutualisation des risques qui consiste à regrouper des assurés faisant face à des risques semblables au sein d'un même portefeuille. Pour se protéger de l'anti-selection, l'assureur peut segmenter son tarif sur la base des facteurs de risque connus. La tarification actuarielle fait appel à des modèles statistiques tels que le GLM et des algorithmes de machine learning. Avec l'accroissement du volume de données sur les assurés et sur leurs habitudes de vie, la volonté des assurés de payer le *juste prix* de leur risque, la segmentation des primes est de plus en plus poussée vers l'individualisation des tarifs. Cette personnalisation des risques revient dans les débats actuariels sous la désignation de démutualisation dont les dangers majeurs sont : pour la communauté, l'exclusion des *mauvais risques* qui se verraient proposer des contrats qu'à des prix trop élevés ; pour l'assureur, il pourrait resserrer son offre d'assurance sur une petite niche débouchant sur plus de volatilité et un risque de perte d'expertise sur les autres profils de risques. Bien que la (dé)mutualisation soit fondamentale en assurance et à l'ordre du jour dans les congrès d'actuaire, sa définition, sa mesure et son impact sur l'assurance demeurent très peu abordés dans la littérature, au-delà des intuitions solidement ancrées autour de la loi des grands nombres.

Pour mutualiser ou non son portefeuille, l'assureur peut s'appuyer sur deux leviers : un levier technique, portant sur les choix de segmentation lors de la construction de la prime technique ; un levier commercial, concernant les profils de marges que l'assureur décide d'appliquer arbitrairement par assuré. La démutualisation a alors été définie comme la désolidarisation des assurés d'un portefeuille ; le cas extrême étant une prestation individuelle certaine avec un niveau de marge constant par assuré, sans aucun subsidé croisé. Des indicateurs de mesure de mutualisation ont été élaborés en s'appuyant sur la variance, l'entropie et la complexité de Kolmogorov.

L'application des indicateurs dits simples sur un portefeuille réel révèle que ces indices sont cohérents et peuvent être utilisés comme outils d'aide à la décision pour le choix des variables tarifaires ou comme outils de communication externe. Parmi ces indicateurs, le niveau d'individualisation et la prime de solidarité se démarquent notamment pour leur simplicité de calcul et leur interprétabilité. Une seconde classe d'indicateurs plus complexe a été introduite afin de réduire le besoin en information sur le portefeuille des indicateurs simples et de mieux comprendre le rôle des facteurs de risque dans la démutualisation. Les implémentations sous Python des indicateurs proposés sont disponibles en accès libre sur GitHub.

**Mots clés** : Mutualisation, démutualisation, segmentation, solidarité, entropie, complexité de Kolmogorov, GLM, LightGBM.

## ABSTRACT

Insurance relies on risk mutualization, which consists of pooling policyholders facing similar risks within the same portfolio. In order to prevent anti-selection, the insurer can segment the risk premium according to the observed risk factors. Actuarial pricing uses statistical models such as GLM and machine learning algorithms to achieve this goal.

Given the growing data availability about policyholders and their usage, the social willingness of policyholders to pay the "*right price*", premium segmentation is becoming more and more granular, the extreme case being premium individualization. This personalization of risks by premium and his drawback is debated in actuarial committees as demutualization. One of the major dangers of demutualization is policyholders exclusion from the coverage due to unaffordable premium. For the insurer, it could tighten its insurance offer on a small market niche leading to more volatility results and a loss of expertise on the other risk profiles. Although (de)mutualization is fundamental in insurance and is on the agenda of many actuarial conferences, its definition, its measurement and its impact on insurance remain very poorly addressed in the literature, beyond the strongly held intuitions around the law of large numbers. This study aims to formalise these notions and to develop indicators for measuring the level of mutualisation of an insurance portfolio.

To mutualise or no risks in its portfolio, the insurer can rely on two mechanisms: a technical lever, related to segmentation choices made when estimating the technical premium; and a commercial lever, which concerns the margin profiles that the insurer decides to apply arbitrarily per insured. Demutualization was then defined as the desolidarisation of policyholders from a portfolio; the extreme case being a certain individual benefit with a constant margin level per policyholder, without any cross-subsidisation. Indicators for measuring mutualization were developed based on variance, entropy and Kolmogorov complexity.

The numerical application of a first class of so-called simple indicators on a real portfolio reveals that the mutualization indices are consistent and can be used as decision support tools for pricing choices. They can also be used as external communication tools. Among these indicators, the level of individualisation and the solidarity premium stand out for their simplicity of calculation and their interpretability. Some more elaborate indicators have been introduced to reduce the required information's on the portfolio, to better understand mutualization by highlighting the role of risk factors in demutualization. This class requires more complex calculations and seems more difficult to industrialise. Python implementations of the proposed indicators are available in open access on GitHub.

**Key words** mutualization, demutualization, segmentation, solidarity, entropy, Kolmogorov complexity, GLM, LightGBM.

# Note de synthèse

## 1 Introduction et contexte

Le mécanisme de l'assurance repose sur la **mutualisation** des risques qui consiste à regrouper des assurés faisant face à des risques semblables au sein d'un même portefeuille. Cette mutualisation est une déclinaison actuarielle de la loi des grands nombres (Baker et Simon, 2001; Charpentier, 2011), qui suppose alors l'indépendance et l'homogénéité des risques. Or, les risques auxquels sont soumis les assurés ne sont pas homogènes en général; ce qui pousse les assureurs à segmenter les primes en fonction des facteurs de risques captés à travers les variables tarifaires.

Avec l'accroissement du volume de données sur les assurés et sur leurs habitudes de vie, la volonté des assurés de payer le *juste prix* de leur risque, la segmentation des primes est de plus en plus poussée, le cas extrême étant une individualisation des tarifs. Les opportunités associées à ces tendances sont à l'origine de l'émergence des *insurtechs* et des produits d'assurance très personnalisés comme les produits basés sur les données d'usage (Usage Based Insurance). Cette personnalisation des risques et les inquiétudes qu'elle soulève reviennent dans les débats actuariels sous la désignation de **démutualisation** dont l'un des dangers majeurs pour la communauté est l'exclusion d'une catégorie de la population (L'Argus, 2021; Les Echos, 2019; L'Argus, 2016), notamment celle des *mauvais risques* (puisque'ils se verraient proposer des primes très - trop - élevées), faisant alors perdre à l'assurance sa raison d'être : protéger les plus à risque.

Bien que la mutualisation soit fondamentale en assurance et à l'ordre du jour dans les congrès d'actuariers, sa définition, sa mesure et les conséquences de la démutualisation demeurent très peu abordées dans la littérature, au-delà des intuitions solidement ancrées autour de la loi des grands nombres. Cette étude se propose d'apporter une formalisation de ces notions et de formuler des indicateurs calculables pour évaluer le niveau de mutualisation d'un portefeuille réel. L'étude s'intéresse également à l'analyse empirique et quantitative des effets de la démutualisation sur l'assureur et l'assurance.

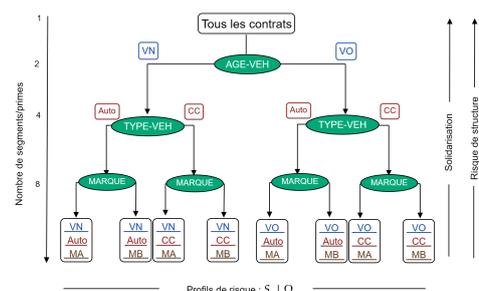
### Données d'application

Les données du produit d'extension de garantie automobile de BNP Paribas Cardif sur le périmètre de la France ont servi pour le cadre pratique de ce mémoire. Ce produit revêt un intérêt pour cette étude au vu de la profondeur de l'historique disponible, mais surtout de la finesse possible de segmentation qui peut être atteinte à partir des variables tarifaires; en considérant le modèle de véhicule dans la tarification par exemple.

## 2 Mutualisation et démutualisation : enjeux et formalisations

### 2.1 Mutualisation et segmentation

Pour se couvrir de la sélection adverse et construire un portefeuille d'assurance, il peut être nécessaire de procéder à une segmentation des assurés à partir des variables tarifaires. L'actuaire constitue des *classes* de risques homogènes au sein desquelles il mutualise les risques d'un pool d'assurés. Au sein de chaque classe, il subsiste une hétérogénéité résiduelle; due d'une part à l'incomplétude des facteurs de risque observés/utilisés et d'autre part au *hasard pur*. Un montant de prime par classe de risque est alors déterminé à partir d'une estimation empirique de la sinistralité attendue de la classe (Charpentier *et al.*, 2015).



**Fig. 1.** Segmentation des primes et solidarité au sein du portefeuille : extension de garantie automobile avec 3 facteurs de risque

Dans un scénario où les assurés intra-classes seraient suffisamment nombreux, l'on pourrait s'attendre à observer une compensation entre les primes perçues et les sinistres payés d'une classe, la loi des grands nombres s'appliquant dans chaque classe de risques homogènes. En pratique, l'homogénéité à l'intérieur d'une classe s'améliore avec le nombre de facteurs de risque utilisés, mais inversement le nombre d'assurés par classe diminue.

À mesure que la segmentation s'affine, les dangers de démutualisation comme les **exclusions** (Keller *et al.*, 2018), les **injustices actuarielles** (Frezal et Barry, 2020), les **instabilités des primes** (Hay, 2015) et le **positionnement sur des niches de risques** (Charpentier *et al.*, 2015) se multiplient. La solidarité entre assurés permet de privilégier l'équilibre du portefeuille directement : la prime payée par un assuré peut être transférée vers d'autres classes de risque.

## 2.2 Formalisation de la mutualisation

### Tarification technique comme levier de mutualisation

Soit  $S$  la charge de sinistre totale d'un assuré,  $\mathbf{X} = (X_1, X_2, \dots, X_p)$  le vecteur aléatoire des facteurs de risques observés par l'assureur et  $\mathbf{\Omega} = (\Omega_1, \Omega_2, \dots)$  le vecteur aléatoire (théorique) de l'ensemble des caractéristiques qui influent sur le risque. En étudiant trois scénarios de partage de risque sur la base de la prime pure entre assureur et assurés (Denuit et Charpentier, 2004), la relation entre les choix de segmentation et le partage de variance se précise :

**Tab. 1.** Partage du risque entre assureur et la communauté des assurés selon le niveau de segmentation

	Aucune segmentation (a)		Segmentation imparfaite (b)		Segmentation parfaite (c)	
Acteur	Assurés	Assureur	Assurés	Assureur	Assurés	Assureur
Dépense	$\mathbb{E}(S)$	$S - \mathbb{E}(S)$	$\mathbb{E}(S/\mathbf{X})$	$S - \mathbb{E}(S/\mathbf{X})$	$\mathbb{E}(S/\mathbf{\Omega})$	$S - \mathbb{E}(S/\mathbf{\Omega})$
Dépense moyenne	$\mathbb{E}(S)$	0	$\mathbb{E}(S)$	0	$\mathbb{E}(S)$	0
Variance	0	$\mathbb{V}(S)$	$\mathbb{V}(\mathbb{E}(S/\mathbf{X}))$	$\mathbb{E}(\mathbb{V}(S/\mathbf{X}))$	$\mathbb{V}(\mathbb{E}(S/\mathbf{\Omega}))$	$\mathbb{E}(\mathbb{V}(S/\mathbf{\Omega}))$

- Le scénario (a) correspond à une absence de segmentation, l'entièreté du risque de structure est portée par l'assureur. Dans ce scénario, chaque assuré cotise la même prime indépendamment de ses propres facteurs de risque ;
- Le scénario (c) est un cas théorique dans lequel l'assureur dispose de l'ensemble des facteurs de risque  $\mathbf{\Omega}$  et le risque de structure porté par l'assureur serait alors minimale. La prime payée par chaque assuré est conditionnelle à ses propres caractéristiques et dès lors aléatoire ;
- Le scénario (b) est un scénario intermédiaire entre (a) et (c) ; en fonction de la disponibilité des facteurs de risque et de ses considérations commerciales, l'assureur procède à une segmentation imparfaite des assurés créant conjointement la mutualisation et la solidarité entre les assurés.

### Les chargement assureurs comme levier de mutualisation

Commissions	Optimisation Layer (OL)	Prime commerciale
Frais d'acquisition		
Frais administratifs		
Marge assureur		
Prime pure / Burning cost (BC)	Profil de risque	

**Fig. 2.** Décomposition de la prime commerciale en optimisation layer et en prime pure.

En pratique, au montant de prime pure, il faut ajouter des chargements (positif ou négatif), désignés par optimisation layer (OL) pour obtenir la prime commerciale. En appliquant des profils de marge différents sur certains périmètres du portefeuille, l'assureur crée de la solidarité entre ses assurés. Sur des risques différents, l'assureur peut décider de jouer sur l'OL pour faire payer la même prime commerciale à tous les assurés aboutissant ainsi à un portefeuille solidaire. De même, l'assureur peut utiliser l'OL pour financer l'acquisition et la rétention des contrats, notamment les remises à la souscriptions (« **price-walking** » (FCA, 2020)).

## 2.3 Proposition de définition pour la démutualisation

Nous proposons de définir la **démutualisation** d'un portefeuille comme la désolidarisation des assurés de la communauté, le cas extrême étant une prestation individuelle certaine avec un niveau de marge constant par assuré, sans aucun subside croisé. Ainsi, un portefeuille sera d'autant plus démutualisé que chaque assuré cotise pour son risque indépendamment des autres assurés. Inversement, un portefeuille sera dit mutualisé si au sein de celui-ci l'assureur fait jouer la solidarité entre les assurés en plus de la mutualisation entre les assurés ayant le même profil de risque lorsque tous les facteurs de risque disponibles sont pris en considération, la mutualisation indivisible.

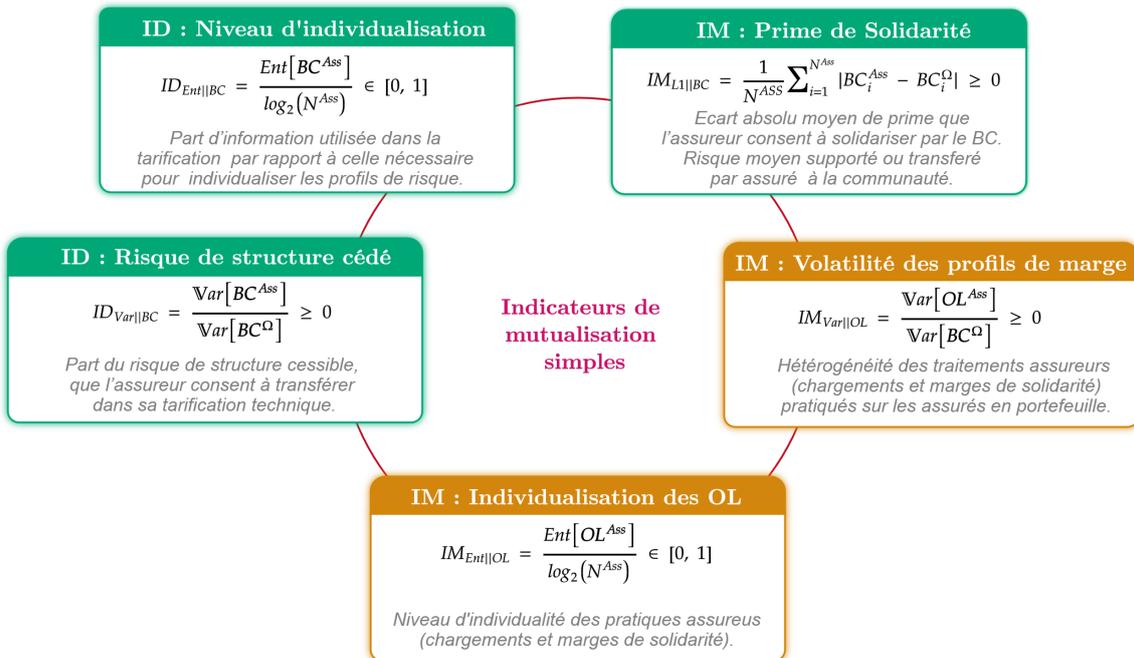
## 3 Construction des indicateurs de mutualisation

### 3.1 Les indicateurs simples basés sur la prime et ses composantes

Les pratiques de désolidarisation que sont l'hyper-segmentation des primes techniques et l'hétérogénéité des marges assureurs par assuré peuvent être quantifiées à partir d'une évaluation de la dispersion des primes techniques et des OL respectivement. Deux mesures d'incertitude ont été utilisées pour construire les indicateurs simples, pour une variable aléatoire  $X$  prenant les valeurs  $x_1, x_2, \dots, x_n$  avec probabilités  $p_1, p_2, \dots, p_n$  :

- **Variance**  $\mathbb{V}[X] = \sum_{i=1}^n p_i(x_i - \bar{x})^2$ , avec  $\bar{x} = \frac{1}{n} \sum_{i=1}^n p_i x_i$  : elle mesure la volatilité de la variable autour de sa valeur moyenne ; son utilisation découle de la formalisation du transfert de risque de structure à travers la segmentation des primes ;
- **Entropie**  $Ent(X) = - \sum_{i=1}^n p_i \log_2(p_i)$  : elle mesure l'hétérogénéité au sein d'une population ou d'un système ; en théorie de l'information, elle mesure la quantité d'information moyenne contenue dans un message généré par une source d'information.

L'idée charnière des indicateurs simples est d'évaluer l'écart entre la tarification technique adoptée par l'assureur ( $BC^{Ass}$ ) et celle qu'il pourrait appliquer s'il utilisait toute l'information à sa disposition dans son tarif ( $BC^\Omega$ ). Les indicateurs simples et leurs interprétations sont donnés par la figure 3.



**Fig. 3.** Indicateurs simples de mesure de mutualisation d'un portefeuille basés sur les profils de risque ■ ou sur les OL ■. IM : Indice de Mutualisation (croît avec la solidarité entre assurés), ID : Indice de Démutualisation (croît avec la désolidarisation de la communauté).

### 3.2 Les indicateurs exotiques basés sur les facteurs de risque

Pour que la mesure de la mutualisation puisse être communément adoptée par les assureurs, des généralisations des indicateurs simples sont introduits. Les indicateurs exotiques cherchent à s'affranchir de la dépendance des indicateurs simples au modèle de calcul des profils de risque individualisés, à alléger le besoin en informations sur le portefeuille et à mieux interpréter la mutualisation en précisant le rôle des facteurs de risque.

#### Inertie des contributions au risque

La contribution du facteur de risque  $X_j$  pour l'assuré  $i$  (notée  $c_{i,j}$ ) est calculée comme la sensibilité de la sinistralité prédite par le modèle pour cet assuré suite à une variation du facteur  $X_j$ . Lorsqu'un modèle de machine learning est utilisé pour prédire le burning cost,  $c_{i,j}$  peut être approché à l'aide des valeurs de SHAP (Lundberg et Lee, 2017), lorsque le GLM est utilisé pour modéliser la sinistralité attendue,  $c_{i,j}$  peut être déterminé comme suit :

$$c_{i,j} := \frac{\partial \ln(\hat{\mathbb{E}}[S_i | \mathbf{X}_i = \mathbf{x}_i])}{\partial X_j} = \frac{\partial}{\partial X_j} \left( \hat{\beta}_0 + \sum_{j=1}^p \sum_{l=1}^{L_j} \mathbb{I}_{\{x_{i,j} \in I_{j,l}\}} \hat{\beta}_{j,l} \right) = \hat{\beta}_{j,l}$$

où  $l$  est le rang de l'intervalle discrétisé de  $X_j$  dans lequel tombe l'observation  $x_{i,j}$ . Pour tous les assurés et tous les facteurs de risque, un nuage des contributions est obtenu  $\mathcal{C}_{N^{Ass}} = \{c_i \in \mathbb{R}^p, i = 1, \dots, N^{Ass}\}$ .

La dispersion de ce nuage rend compte du transfert de risque de structure consenti par l'assureur. Des contributions semblables et peu disparates traduisent une faible différenciation des profils de risque par la prime technique. Inversement, un nuage de contribution très éparse traduit une segmentation tarifaire très fine. L'utilisation de l'**inertie** pour évaluer la dispersion dans ce nuage de points conduit à l'indicateur suivant :

#### ID : Part de risque de structure cédée généralisée

$$ID_{Int|\partial\mu_i} = \frac{I(\mathcal{C}_N^{Ass})}{I(\mathcal{C}_N^\Omega)} = \frac{\sum_{i=1}^{N^{Ass}} \|c_i^{Ass} - \bar{c}^{Ass}\|_2^2}{\sum_{i=1}^{N^{Ass}} \|c_i^\Omega - \bar{c}^\Omega\|_2^2}$$

$c_i^{(a)}$  : contributions de l'assuré  $i$  pour le modèle  $(a)$  et  $\bar{c}^{(a)}$  : contributions moyennes par variable.

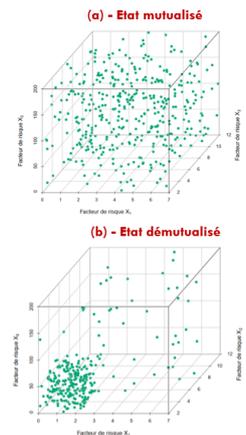
#### Entropie algorithmique d'une description du portefeuille

L'**entropie algorithmique** (Zurek, 1989) mesure l'hétérogénéité dans un système physique pour une description partielle (notée  $x_d$ ) de ce système. Elle se décompose comme la somme  $S_{\Gamma_{x_d}} = K(x_d) + H(\Gamma_{x_d})$  de :

La **complexité de Kolmogorov** (Kolmogorov, 1968; Li et Vitányi, 2019) de la description du système qui mesure l'hétérogénéité entre les particules décrite par  $x_d$ ;  $K(x_d) := \min_p \{|p| : U(p) = x\}$ , la *longueur du plus petit programme qui générerait la description  $x_d$  lorsqu'il est exécuté sur une machine de Turing universelle  $U(\cdot)$*  (Turing, 1937). Elle n'est pas calculable mais peut-être approchée.

L'**entropie de Shannon** du macro-état  $\Gamma_{x_d}$  quantifie l'hétérogénéité résiduelle c'est-à-dire l'incertitude sur les configurations microscopiques du système lorsque seule la description macroscopique  $x_d$ .

Pour mesurer la mutualisation, le système physique est remplacé par l'espace engendré par les facteurs de risque et les particules de gaz par les assurés. La description  $x_d$  du système physique est remplacée par une synthèse itérative de plus en plus précise de la dispersion des assurés suivant les facteurs de risque triés par ordre d'importance pour la modélisation du risque (notée  $x_d^{freq}$ ).



**Fig. 4.** Assurés dans l'espace des facteurs de risque : (a) assurés dans tout l'espace, (b) assurés recroquevillés

### Exemple de description du portefeuille

- $x_1^{freq} = \ll 40 \% \text{ des véhicules du portefeuille sont de marques TOYOTA ou FORD } \gg$
- $x_2^{freq} = x_1^{freq}$  et  $\ll 20 \% \text{ des véhicules TOYOTA ou FORD ont moins de 6 mois à la souscription } \gg$
- $x_3^{freq} = x_2^{freq}$  et ...

Une description  $x_d^{freq}$  est obtenue à l'aide d'un arbre de régression CART calibré en régressant les profils de risque individualisés sur l'ensemble des facteurs de risque sur la *distribution de marché*. Il en découle alors une séquence de couples (variable & seuil) qui partitionnent itérativement l'espace en nœuds ou sous-régions. Cet arbre appelé *splitter* encode la structure des risques du marché.

Si les OL sont connues,  $x_d^{freq}$  est complétée par la description de l'importance des marges de solidarité collectées par l'assureur pour le cross-financement entre les assurés de son portefeuille (notée  $x_d^{OL}$ ).  $x_d^{OL} := \ll \alpha_1^{OL} | \alpha_2^{OL} | \dots | \alpha_d^{OL} \gg$  où  $\alpha_d^{OL}$  mesure la solidarité créée entre les assurés du nœuds  $\mathcal{N}_d$  par le gain de stabilité d'OL consécutif partition du nœud  $\mathcal{N}_d$  en  $\mathcal{N}_{d,<}$  et  $\mathcal{N}_{d,>}$  calculé comme suit :

$$\alpha_d^{OL} := \frac{\text{Var}_{\mathcal{N}_d}(\mathbf{OL}) - \left( \frac{|\mathcal{N}_{d,<}|}{|\mathcal{N}_d|} \text{Var}_{\mathcal{N}_{d,<}}(\mathbf{OL}) + \frac{|\mathcal{N}_{d,>}|}{|\mathcal{N}_d|} \text{Var}_{\mathcal{N}_{d,>}}(\mathbf{OL}) \right)}{\text{Var}_{\mathcal{N}_0}(\mathbf{OL})}$$

où  $\text{Var}_{\mathcal{N}_{d,\cdot}}(\mathbf{OL})$  est la variabilité des OL des assurés du nœud  $\mathcal{N}_{d,\cdot}$  et  $\text{Var}_{\mathcal{N}_0}(\mathbf{OL})$  est la variabilité des OL dans tout le portefeuille. En normalisant l'entropie algorithmique par la dispersion maximale des assurés dans l'espace, on obtient un indicateur de mutualisation :

### IM : Ratio de mutualisation généralisé

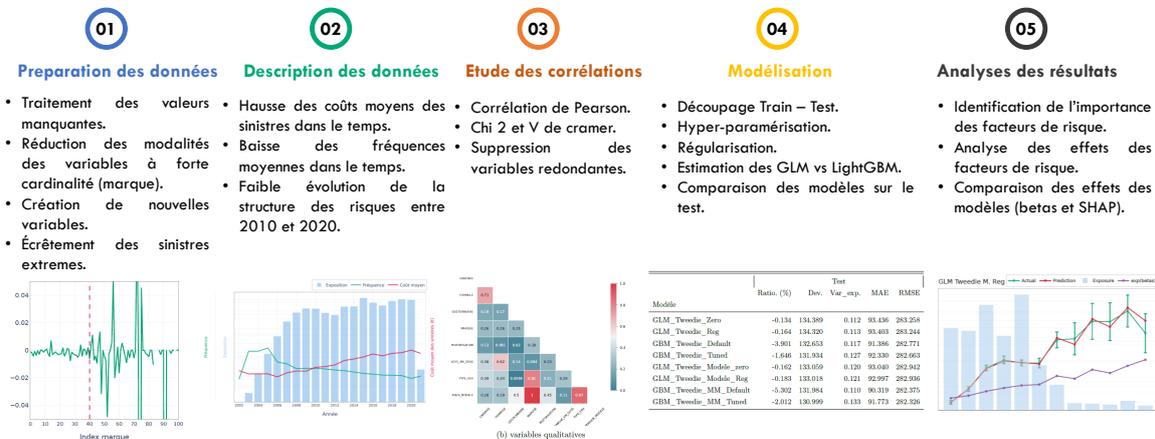
$$IM_{CK||Ent} = \frac{\min_d \{K(x_d) + H(\Gamma x_d)\}}{N^{Ass} \log_2(L_{0,0})}$$

où  $x_d$  : description (fréq ou OL) du portefeuille et  $L_{0,0}$  : nombre de feuilles terminales du *splitter*.

## 4 Cadre pratique des indicateurs de mutualisation

### 4.1 Estimation des profils de risque les plus individualisés

La plage de données 2010 - 2019 est utilisée pour calibrer les modèles de calcul des profils de risque (figure 5) et l'exercice de 2020 est consacré à l'analyse de la mutualisation. Deux familles de modèles sont mises en concurrence pour prédire au mieux la sinistralité attendue  $BC_i^\Omega$  pour chaque assuré  $i$  : le GLM et le LightGBM.



**Fig. 5.** Les modèles basés sur le GLM sont meilleurs pour reconstituer la charge de sinistre totale; notamment le meilleur modèle GLM prédit une charge globale de sinistres avec une erreur absolue de l'ordre de 0,2 % contre 2 % pour le meilleur LightGBM. Inversement, les modèles LightGBM ont des meilleurs scores sur les métriques qui privilégient les adéquations individuelles (MAE, RMSE, déviance). Les modèles s'accordent sur les variables les plus importantes.

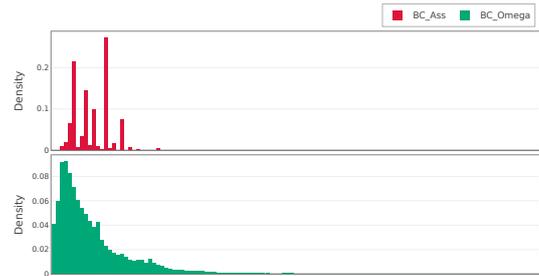
Le modèle de calcul des primes techniques pratiquées  $BC^{Ass}$  est calibré à partir d'un GLM en utilisant les variables formule, âge du véhicule, type de véhicule et puissance, chacune ayant 3 modalités.

## 4.2 Niveau de mutualisation du portefeuille extension de garantie en 2020

La stratégie tarifaire de BNP Paribas Cardif/Icare conduit à une agrégation importante des profils de risque différenciables ; il s'agit d'une création de solidarité entre les assurés du portefeuille (figure 6). Les indicateurs de mutualisation (tableau 2) viennent corroborer cette observation graphique en quantifiant la mutualisation induite par cette stratégie. La démutualisation du portefeuille est assez faible puisque cette stratégie n'utilise que 20,5 % de l'information nécessaire pour individualiser les primes techniques soit 5 profils de risque différents en moyenne par montant de prime technique. La prime de solidarité est également assez importante et s'élève à 34 €.

Indicateur	Valeur	
Part du risque de structure cédée	$ID_{Var  BC}$	0,175
Niveau d'individualisation	$ID_{Ent  BC}$	0,205
Prime de Solidarité	$IM_{L1  BC}$	34,215 €
Volatilité des profils de marge	$IM_{Var  OL}$	0,388
Niveau d'individualisation des OL	$IM_{Ent  OL}$	0,782
Part du risque de struct. cédée généralisée	$ID_{Int  \partial\mu}$	0,184

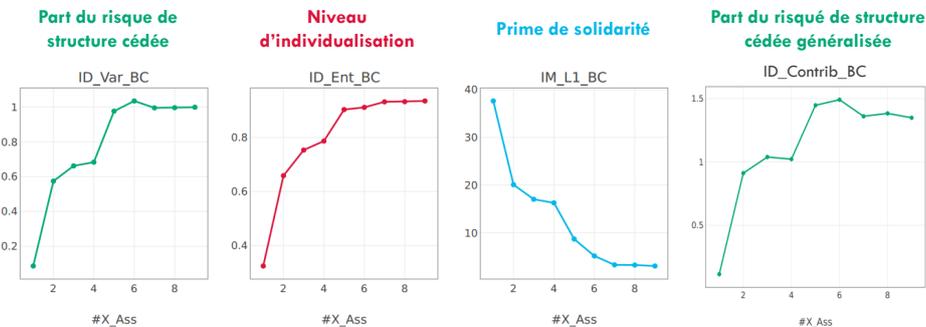
**Tab. 2.** Mesures de mutualisation sur les véhicules couverts en 2020. Les ID sont assez faible alors que les IM sont assez élevés.



**Fig. 6.** Distributions des profils de risque les plus individualisés ( $BC^{\Omega}$ ) et tarifés ( $BC^{Ass}$ ).

## 4.3 Indices de mutualisation pour guider la stratégie tarifaire

Les indices de mutualisation peuvent être utilisés pour guider les décisions de segmentation tarifaires lors de la construction de la prime technique. En effet, l'assureur pourrait ajouter une contrainte sur la solidarité minimale souhaitée dans son portefeuille (figure 7).



**Fig. 7.** La démutualisation (resp. la mutualisation) du portefeuille évaluée sur les primes techniques à partir des indicateurs proposés augmentent (resp. diminue) lorsque les variables tarifaires deviennent de plus en plus exhaustives ( $X_{Ass} \rightarrow X_{\Omega}$ ).

## 4.4 Hyper-segmentation, mutualisation et concurrence : quels enseignements ?

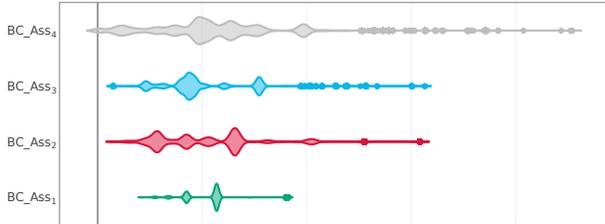
Les indicateurs de mutualisation peuvent également être utilisés à des fins de communication externe, notamment pour la comparaison entre plusieurs assureurs. Nous envisageons 4 assureurs qui effectuent des choix tarifaires différents.

**Tab. 3.** Stratégies tarifaires utilisées par les assureurs pour la construction de la prime technique.

Assureur	Variabes
$Ass_1$ ■	FORMULE, TYPE_VEH, AGE_VEH_SOUS
$Ass_2$ ■	FORMULE, TYPE_VEH, AGE_VEH_SOUS, PUISSANCE
$Ass_3$ ■	FORMULE, TYPE_VEH, AGE_VEH_SOUS, 'TRANCHE_KM_SOUS'
$Ass_4$ ■	FORMULE, TYPE_VEH, AGE_VEH_SOUS, MARQUE

#### 4.4.1 Comparaison des niveaux de mutualisation des assureurs en monopole

En situation de **monopole** chacun de ces assureurs commercialiserait les primes techniques dont les distributions sont données par la figure 8. L'assureur 4 en utilisant la variable MARQUE segmente très précisément son risque aboutissant à des primes très étalées contrairement à l'assureur 1 qui n'utilise que très peu d'information dans son modèle de prime technique. Les évaluations des niveaux de mutualisation correspondant à ces stratégies sont données par le tableau 4. L'assureur le plus démutualisé est l'assureur 4 avec un niveau d'individualisation de 36 %. À l'inverse, la stratégie de l'assureur 1 est très inclusive, avec une individualisation des primes à seulement 13 %.



**Fig. 8.** Primes techniques commercialisées par les assureurs en situation de monopole. Les primes de Ass 4 ■ sont très segmentées alors que les primes de Ass 1 ■ sont très agrégées.

Indicateur	Ass <sub>1</sub> ■	Ass <sub>2</sub> ■	Ass <sub>3</sub> ■	Ass <sub>4</sub> ■
ID_Ent_BC	0.131	0.199	0.216	0.360
IM_L1_BC	37.153	33.170	34.421	28.627
ID_Var_BC	0.047	0.175	0.147	0.344
IM_Var_OL	0.943	0.827	0.866	0.634
IM_Ent_OL	0.908	0.909	0.908	0.906
IM_L1_BC_Shift	35.065	29.703	31.473	26.101
Prime	4659776	4529484	4603268	5050225
S/P	0.975	1.003	0.987	0.899

**Tab. 4.** Indices de mutualisation simple de plusieurs assureurs en situation de monopole. Par ordre de démutualisation : Ass 4 ■ > Ass 2 ■ > Ass 3 ■ > Ass 1 ■

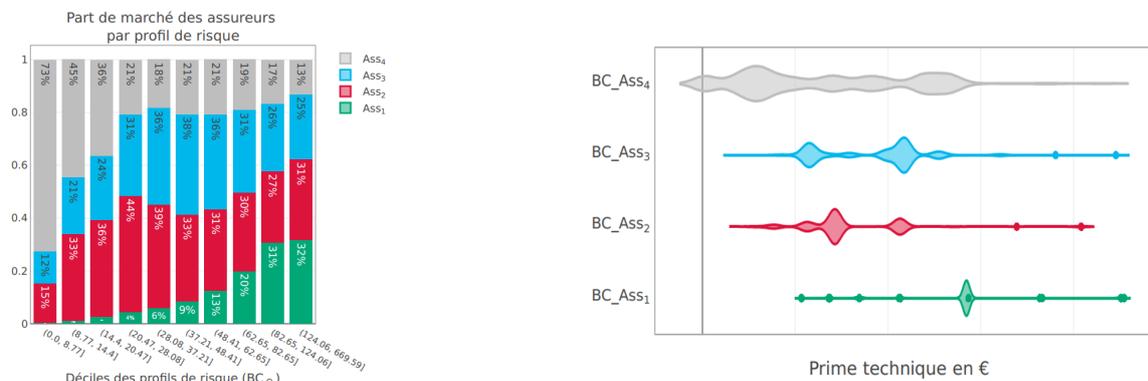
Les ratios S/P montrent que ces assureurs dégagent des résultats techniques positifs en situation de monopole avec un résultat bien plus favorable pour l'assureur le plus segmenté. Ce résultat est dû à l'amélioration de la sinistralité (incluant un effet Covid-19). La sur-performance de l'assureur 4 découle de l'augmentation du poids des mauvais risques sur le marché qui bénéficient également d'une plus grande amélioration de sinistralité. Ainsi, la démutualisation permet à l'assureur de mieux maîtriser son risque et de se couvrir contre une déviation de la structure des risques du marché.

#### 4.4.2 Comparaison des niveaux de mutualisation des assureurs en concurrence

Les quatre assureurs précédents sont placés sur un marché concurrentiel sur lequel chaque assuré a une connaissance parfaite de tous les contrats proposés et souscrit à la police la moins chère. Il s'en suit alors, une sélection par les prix ; en souscrivant uniquement les profils sur lesquels il est compétitif, chaque assureur va récupérer une fraction plus ou moins hétérogène des assurés. D'autre part, elle induit un effort de mutualisation sur des groupes de risque plus restreints. Aboutissant *in fine* à une démutualisation globale du marché (tableau 5). Au niveau de chaque assureur, la mutualisation diminue également par rapport à la situation de monopole en particulier la prime solidarité (corrigée de la valeur absolue du résultat technique attendu par tête) diminue d'environ 22 € par assureur.

**Tab. 5.** Indices de mutualisation simple de plusieurs assureurs sur un marché concurrentiel.

Indicateur	Ass <sub>1</sub> ■	Ass <sub>2</sub> ■	Ass <sub>3</sub> ■	Ass <sub>4</sub> ■	Marché
ID_Ent_BC	0.086	0.147	0.171	0.384	0.304
IM_L1_BC	52.769	28.484	30.457	16.337	28.469
ID_Var_BC	0.030	0.047	0.043	0.195	0.087
IM_Var_OL	0.922	0.747	0.849	0.594	0.724
IM_L1_BC_Shift	7.926	7.723	10.896	8.148	8.759
Part_Marche (%)	11.780	31.880	28.100	28.241	100.000
Primes	587156	893783	911934	666104	3058979
S/P	1.775	1.345	1.493	1.404	1.485



(a) Parts de marché des assureurs suivant les profils de risque en concurrence. (b) Distribution des primes commercialisées par assureurs en concurrence.

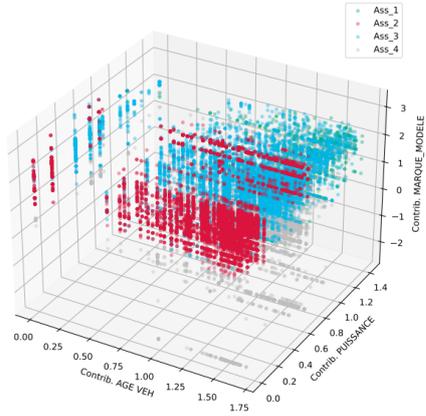
**Fig. 9.** Ass 4 ■ se positionne majoritairement sur les très bons risques auxquels il fait payer des primes très faibles, Ass 1 ■ est victime d'anti-sélection, il souscrit les mauvais risques à des tarifs sous-évalués, Ass 2 ■ et Ass 3 ■ sont présents sur tous les risques.

Cette mise en concurrence permet de mettre en évidence empiriquement les effets de l'hyper-segmentation sur les résultats des assureurs et sur le marché :

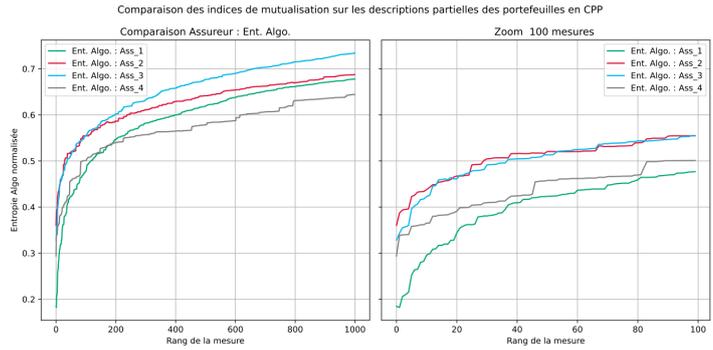
<p><b>La segmentation des prix conduit à de meilleurs résultats attendus</b></p>	<p><b>L'hyper-segmentation n'est pas une garantie de performance</b></p>	<p><b>L'hyper-segmentation présente également des risques d'exclusion</b></p>	<p><b>La spirale de la segmentation peut être néfaste pour le marché</b></p>
<p><b>La segmentation permet de se couvrir contre le risque de structure</b></p> <ul style="list-style-type: none"> <li>Proposer des tarifs individualisés permet à l'assureur d'ajuster le niveau global de prime aux sinistres attendus tout en se couvrant contre l'aléa-moral.</li> <li>Une segmentation plus élevée permet de mieux comprendre le risque et de capter les meilleurs risques du marché.</li> </ul>	<p><b>Une segmentation plus élevée peut conduire à de moins bons résultats dans certains cas</b></p> <ul style="list-style-type: none"> <li>En particulier, une segmentation élevée conduit à percevoir des primes plus faibles sur les bons risques qui rendent plus difficile une mutualisation avec d'autres risques, surtout lorsque ces risques sont volatils.</li> </ul>	<p><b>Une segmentation plus élevée peut conduire à des exclusions</b></p> <ul style="list-style-type: none"> <li>Baisser le prix sur les bons risques entraîne une augmentation mécanique des prix des mauvais risques, rendant certaines catégories « non assurables ».</li> <li>Financièrement, cela signifie également qu'un assureur hyper-segmenté peut « manquer » des opportunités d'amélioration des risques sur les « mauvais risques ».</li> </ul>	<p><b>L'hyper-segmentation pousse le marché vers la démutualisation</b></p> <ul style="list-style-type: none"> <li>Une segmentation plus élevée pousse l'ensemble du marché vers plus de démutualisation, ce qui rend l'assurance moins accessible pour certains risques.</li> <li>L'arrivée d'un nouvel acteur ayant une segmentation plus fine détruit plus de valeur pour le marché que toute autre concurrence.</li> </ul>
<p><b>Dans un environnement concurrentiel, la segmentation est incontournable.</b></p>	<p><b>Il n'y a pas de justification à sursegmenter trop finement par rapport aux concurrents.</b></p>	<p><b>Les choix tarifaires doivent être pris en tenant compte des concurrents et des risques d'exclusion.</b></p>	<p><b>La proposition de valeur d'un nouvel acteur ne saurait se résumer à des primes plus segmentées.</b></p>

#### 4.5 Les indicateurs exotiques précisent le rôle des facteurs dans la démutualisation

Contrairement aux indicateurs simples, les indicateurs exotiques fournissent des représentations permettant de mieux comprendre la mutualisation et la structure des risques en portefeuille. La figure 10 précise l'apport des indicateurs exotiques pour décrire la mutualisation des 4 assureurs précédents lorsqu'ils sont en concurrence.



(a) Positionnement des assurés dans l'espace des contributions des facteurs au profil de risque individualisé par assureur. Les assurés de Ass 1 ■ ont des contributions au risque suivant la puissance élevée et sont également des marques risquées. Ass 4 ■ se positionne sur les marques ayant une sinistralité moindre. Ass 2 ■ se positionne sur tous les risques par marque, mais effectue une sélection suivant la puissance. Les véhicules de Ass 3 ■ ont des contributions suivant la marque faibles ou élevées, mais ont des classes de puissance plus risquée que pour Ass 2 ■.



(b) Évolution du RMG - Fréquence par assureur suivant la description. Les portefeuilles de Ass 2 ■ et de Ass 3 ■ recouvrent mieux la structure du marché que ceux de Ass 1 ■ et de Ass 4 ■. Ass 1 ■ se positionne sur des risques très spécifiques puisque dès les premières mesures, le RMG - Freq est très faible (et donc la mutualisation résiduelle aussi); Ass 2 ■ et Ass 3 ■ se positionnent sur plusieurs profils différents et capturent un nombre important d'assurés sur chacun de ces segments de risque ( $K(x_d)$  et  $H(\Gamma_{x_d})$  élevés). Ass 4 ■ se positionne également sur plusieurs segments de risque différents, mais ne souscrit qu'un nombre limité d'assurés sur chacun de ces segments ( $K(x_d)$  faible et  $H(\Gamma_{x_d})$  élevé).

**Fig. 10.** Comparaison et analyse de la mutualisation entre assureurs en concurrence à partir des indicateurs de mutualisation exotiques.

## 5 Conclusion, limites et perspectives

La mesure de la mutualisation génère des souhaits et des attentes majeures auxquels les indicateurs proposés répondent assez favorablement. La prime de solidarité et le niveau d'individualisation constituent nos principales propositions pour évaluer la mutualisation d'un portefeuille : ils sont complémentaires, relativement faciles à calculer et à interpréter. Le RMG est assez pertinent lorsqu'il est question d'étudier comment les interactions entre assureurs en concurrence modifient la structure du portefeuille.

Les limites principales des indices sont la non-prise en compte directe de l'équité actuarielle dans la formulation des indicateurs et la dépendance des indicateurs à la modélisation du profil de risque le plus fin. Il faut en effet compléter les indicateurs avec d'autres mesures de valeur pour vérifier que la solidarité créée est socialement acceptable.

Cette étude est conduite afin d'inciter les assureurs à évaluer leur niveau de mutualisation sur divers produits comme l'assurance des emprunteurs. Ce qui permettra à terme de mieux circonscrire les dangers liés à la démutualisation. Dans ce sillage, l'implémentation sous Python pour calculer les indicateurs est mise à disposition sur <https://github.com/B-Noumedem/Indices-de-mutualisation.git>. Des axes d'améliorations envisagés pour cette étude sont : l'analyse de l'influence de la modélisation du profil de risque individuel sur les indices ; une intégration directe de l'équité dans la formulation des indicateurs ; l'étude plus approfondie sur les liens entre la mutualisation et l'inclusion.

## INTRODUCTION

Le mécanisme de l'assurance repose sur la mutualisation des risques qui consiste à regrouper un grand nombre d'assurés faisant face à des risques semblables au sein d'un même portefeuille. Cette mutualisation s'opère à condition que les risques mis ensemble soient suffisamment nombreux et indépendants. La mutualisation est alors une déclinaison actuarielle de la loi des grands nombres, qui suppose une indépendance et une homogénéité des risques. Or, bien que dans de nombreux cas il est possible d'adopter l'hypothèse d'indépendance des risques, l'homogénéité de ceux-ci n'est pas garantie en général. Cette hétérogénéité des risques pousse les assureurs à procéder à une segmentation des assurés au regard des facteurs de risque identifiés à partir des variables tarifaires dans l'optique de constituer des pools de risque homogènes.

Outre cette nécessité actuarielle d'assurer l'homogénéité, les tendances sociales marquées par la volonté des assurés de payer le *juste prix* de leurs risques, fait émerger des produits d'assurance fondés sur des données très fines telles que les données d'utilisation ; aboutissant ainsi à une individualisation des risques. Sur le marché de l'automobile par exemple, selon Global Markets Insight, les polices d'assurances basées sur les données d'usage représentent un marché de plus de 30 milliards de dollars et une croissance de plus 20 % serait attendue entre 2021 et 2027<sup>1</sup>. Cette hyper-segmentation est épaulée par les avancées technologiques qui permettent de collecter des données plus personnalisées et en plus grand volume sur les assurés. Il s'agit notamment de la démocratisation des objets connectés, des véhicules connectés ou avec GPS intégrés, etc. L'exploitation de ces données volumineuses se fait plus aisément avec la puissance de calcul dont disposent actuellement les assureurs et les insurtechs, mais surtout grâce aux méthodes d'intelligence artificielle qui font leur arrivée progressive sur toute la chaîne du parcours client et en particulier en tarification (La Fabrique d'Assurance, 2019) .

Cet effort croissant d'assurer l'homogénéité des risques, voire à les individualiser n'est pas sans risque. Le questionnement autour de l'individualisation, de l'hyper-segmentation des primes et des risques revient au sein de la communauté actuarielle et dans des articles depuis quelques années sous la dénomination de *démunicipalisation* (L'Argus, 2021 ; Les Echos, 2019 ; L'Argus, 2016 ; etc.). Cette question s'avère fondamentale puisqu'en absence de segmentation, l'assureur s'expose à une sélection adverse alors qu'en individualisant les tarifs d'assurance, les assurés les plus à risque (et donc qui auraient le plus besoin d'assurance) pourraient être découragés par les coûts d'entrée ou tout simplement exclus des couvertures d'assurance<sup>2</sup>. Malgré son importance apparente, aucune définition faisant le consensus, ni méthode de calcul du niveau de mutualisation ou de démutualisation d'un portefeuille n'ont été retrouvées dans la

---

1. Rapport annuel sur l'assurance des produits connectés publié en mai 2021, disponible ici

2. Le lecteur pourra penser au cas de l'assurance maladie où l'exclusion des personnes prédisposées à développer des maladies héréditaires pourraient faire l'objet d'exclusions discutables sur les plans éthique et sociale.

littérature, sans doute du fait de sa complexité et de son caractère protéiforme.

L'objectif de ce mémoire est de proposer une méthode d'évaluation du niveau de mutualisation d'un portefeuille assurantiel. Plus spécifiquement, il sera question dans ce mémoire d'apporter :

- Des éléments de réponse aux questions techniques et actuarielles, que pose la segmentation des primes en apportant une formalisation de cette notion omniprésente dans les conférences actuarielles professionnelles ;
- Des perspectives aux multiples articles « grand public » qui promeuvent la justesse des prix individuels et donc la démutualisation. S'il semble aisé de mettre en avant la démutualisation d'un portefeuille en assurance auto sur la base des données de conduite par exemple, la communication devient plus sensible quand il s'agit d'assurance santé où faire jouer la solidarité entre les assurés est indéniablement socialement nécessaire ;
- Des indicateurs de mutualisation concrets et calculables à insérer dans le suivi du portefeuille de BNP Paribas Cardif. Les indicateurs proposés n'apportant pas directement une mesure de l'équité ou de la justesse d'une pratique de mutualisation, ceux-ci devront être accompagnés d'autres indicateurs (découlant des métriques de démutualisation) qui rendent mieux compte de l'équité individuelle et sociale de la politique tarifaire de l'assureur. Les questions éthiques que pose la démutualisation sont en effet trop importantes pour ne pas être abordées dans cette étude ;
- Un cadre pour analyser et apprécier les effets que pourrait avoir la course à l'hypersegmentation sur la performance d'un assureur en particulier et sur la valeur de marché en général.

Le présent mémoire est structuré en six chapitres. Le premier rappelle le rôle clé de la mutualisation en assurance ainsi que les tendances à l'individualisation des risques avec leurs conséquences ; il s'achève sur une proposition de définition de la démutualisation. Le second chapitre est consacré à la construction théorique des indicateurs de mutualisation proposés. Il s'articule autour des outils techniques utilisés à cet effet, la formalisation des indicateurs et leurs interprétations. Dans le troisième chapitre le produit d'extension de garantie automobile ainsi que la base de données historiques sont étudiés. Il s'agit d'un cas d'usage utilisé pour les implémentations numériques des indicateurs de mutualisation tels qu'abordés aux chapitres quatre et cinq. Le chapitre quatre est consacré aux indicateurs simples et à l'analyse des conséquences de la démutualisation alors que le chapitre cinq implémente les indicateurs exotiques. Enfin, le dernier chapitre traite de la synthèse des indicateurs, de leurs utilisations, de leurs limites respectives et suggère quelques perspectives.

# MUTUALISATION ET DÉMUTUALISATION : ENJEUX ET PROPOSITIONS DE FORMALISATION

## Sommaire

<b>1.1</b>	<b>Nécessité de définir et de mesurer la mutualisation</b>	<b>3</b>
<b>1.2</b>	<b>Partage du risque entre assureur et assurés</b>	<b>8</b>
<b>1.3</b>	<b>Proposition de définition de la démutualisation</b>	<b>16</b>

Ce chapitre articulé en trois points délimite le cadre de l'étude et pose les définitions du vocabulaire commun autour de la mutualisation. Il revient sur les fondamentaux de l'assurance que sont la mutualisation et le partage du risque, ce qui permet de dégager des intuitions et de proposer une définition de la démutualisation.

## 1.1 Nécessité de définir et de mesurer la mutualisation

### 1.1.1 Qu'est ce que la mutualisation des risques ?

La peur de l'aléa et la recherche de la sécurité poussent les hommes à se prémunir des risques auxquels ils sont exposés. Face à ces besoins, les assureurs proposent de mettre en commun les risques des assurés afin de supprimer ou de réduire le risque individuel en le faisant porter par la communauté (Baker et Simon, 2001; Lehtonen et Liukko, 2015). Ce principe fondamental de l'assurance est connu comme la mutualisation des risques. Il consiste à regrouper un ensemble d'assurés faisant face à des risques semblables au sein d'un même portefeuille. En pratique, chaque assuré paie une prime contre une promesse d'indemnisation de l'assureur en cas de réalisation (ou de non-réalisation) d'un événement aléatoire. L'assureur utilise les primes collectées auprès de l'ensemble de ses assurés (généralement nombreux) pour verser des prestations à ceux qui ont subi un sinistre (généralement peu nombreux). Cette mutualisation actuarielle repose sur la loi des grands nombres qui suppose alors une indépendance et une homogénéité au sein d'un pool de risques (Charpentier, 2011).

### Indépendance entre les assurés en portefeuille

L'hypothèse d'indépendance entre les risques mis en commun revient à admettre que la survenance et l'ampleur du risque pour un assuré ne modifie pas les probabilités de survenance ni les montants moyens escomptés pour les autres assurés du portefeuille. Cette condition est vérifiée pour certains produits d'assurance (comme l'assurance auto, l'extension de garantie

constructeur, etc.), mais pour des risques tels que les catastrophes naturelles affectant l'ensemble des assurés d'une zone géographique donnée, cette condition n'est pas vérifiée et donc ce risque ne se prête pas à de la mutualisation à l'échelle d'un assureur.

### **Homogénéité construite et admise entre les assurés en portefeuille**

---

La seconde hypothèse pour appliquer la loi des grands nombres impose l'homogénéité des risques. Mais cette condition n'étant pas vérifiée en général, il importe aux assureurs de *construire* cette homogénéité. Le risque auquel est soumis l'assuré est (partiellement) la résultante d'un ensemble de facteurs qui lui sont propres comme son âge, sa zone de résidence, ses habitudes de conduite dans le cas de l'assurance auto par exemple ; ou qui sont spécifiques au bien assuré comme la marque du véhicule, la motorisation dans le cas d'un produit d'extension de garantie automobile. Cette hétérogénéité des risques pousse les assureurs à procéder à une segmentation des risques en fonction des facteurs captés à travers les variables tarifaires conduisant à la création des classes de risques équivalents au sein desquelles les assurés sont supposés avoir des comportements et des risques *identiques*. Chaque classe tarifaire correspond à une certaine combinaison de catégories de facteurs de risque ou d'intervalles dans le cas de facteurs de risque continus. Pour chaque classe tarifaire, les actuaires analysent les données historiques sur les sinistres pour estimer avec précision la sinistralité moyenne de la classe (Paefgen *et al.*, 2013; Desrosieres, 2008, p. 139). L'actuaire généralise la sinistralité moyenne au sein d'une classe de risque par le profil de risque de la classe, c'est-à-dire la distribution de la sinistralité étant données les caractéristiques qui définissent la classe. Cette généralisation fait appel à des modèles de régression qui imposent ou non a priori une loi sur la sinistralité à l'intérieur de la classe. Les paramètres (de position, de dispersion ou de forme) de cette loi sont dépendants des caractéristiques et sont estimés à l'aide d'un modèle statistique comme le modèle linéaire généralisé. Un montant de prime pure est alors proposé pour chaque classe, non plus nécessairement comme la moyenne historique de la sinistralité observée dans la classe, mais comme une estimation de l'espérance de loi décrivant la sinistralité de la classe.

Proposer des montants de prime pure différents d'une classe de risque à l'autre permet à l'assureur de se protéger de l'anti-sélection et de rester compétitif sur le marché pour les meilleures classes de risque. En effet, comme Charpentier *et al.* l'illustrent sur un marché concurrentiel (voir figure 1.1), un assureur qui proposerait un prix unique (égal à l'espérance de la sinistralité) alors que ses concurrents proposeraient des prix segmentés, se retrouverait en déséquilibre financier avec un ratio de sinistralité largement supérieur à 100 % et irait droit à la faillite. Les « *bons* » risques trouveraient des primes plus basses chez l'assureur qui segmente son tarif et les « *mauvais* » risques auraient des primes plus faibles chez l'assureur qui ne segmente pas. Les « *mauvais* » risques iraient alors s'assurer chez l'assureur qui ne discrimine pas par les prix ; mais ce dernier n'ayant pas pu souscrire les « *bons* » risques pour assurer un équilibre financier se retrouvera en perte (Akerlof, 1970). Inversement, les auteurs soulèvent une volatilité

importante du ratio sinistres/primes chez l'assureur qui segmente le plus ses tarifs. Bien que celui-ci soit le seul en équilibre, il est sur une très petite niche d'assurés (les plus *bons* risques), induisant alors une volatilité importante. Cette société a en effet près d'une chance sur cinq de se retrouver avec un ratio plus mauvais que celui du marché.

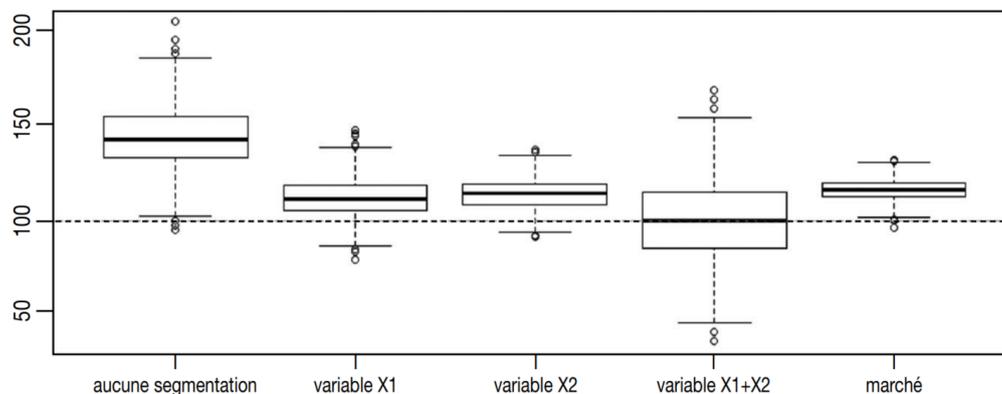


FIGURE 1.1 – Scénarios du ratio S/P sur un marché concurrentiel d'assurance auto.

**Note de lecture :** Distribution du ratio sinistres/primes en présence de quatre sociétés sur le marché de l'assurance automobile. une compagnie qui ne segmente pas, une qui segmente en utilisant l'âge (variable  $X_1$ ), une qui utilise le lieu d'habitation (variable  $X_2$ ) et une qui utilise les deux variables. La situation globale du marché est présentée à droite (Charpentier *et al.*, 2015).

## La segmentation accentuée par le marché et la technologie

Outre cette nécessité d'assurer l'homogénéité, les assureurs segmentent également les profils de risque sous l'impulsion des tendances commerciales et sociales. Les assurés manifestent de plus en plus la volonté de payer le *juste prix* de leur risque individuel. En effet, un assuré dont la conduite est très précautionneuse se demande pourquoi il va payer pour le dédommagement de son voisin dont le comportement est bien plus risqué. Cette tendance a fait naître les produits d'assurance basés sur les données comportementales et d'usages tels que pay-as-you-drive (PAYD), pay-how-you-drive (PAHD), pay-how-you-live (PAYL)... Ces produits sont portés essentiellement par les insurtechs à l'instar de la start-up « Root Insurance » basée aux États-Unis dont le slogan est « *drop the score* ». Aux États-Unis, le score de crédit, c'est-à-dire la probabilité de défaut d'un emprunteur, est utilisé comme facteur de risque dans les modèles de tarification pour l'assurance auto. Or ce dernier est construit à partir des variables qui ne causent pas nécessairement le risque lié à la conduite, mais qui en constituent des proxy. Root Insurance propose une assurance dont la prime est déterminée non pas sur des facteurs *a priori* tels que le score de crédit, mais sur le comportement de conduite déduit des données télématiques c'est-à-dire des données de géolocalisation du véhicule.

L'émergence et la généralisation de ces produits et concepts dans le monde des assurances ont été rendues possibles par les avancées technologiques et la popularisation des objets connectés. Il s'agit des dispositifs électroniques (tels que les GPS, les montres, les balances, les thermomètres, ...) capables de collecter en temps réel, de stocker et de transférer les données sur l'utilisateur

et son environnement. Cette disponibilité croissante et plus fine de la donnée apporte une meilleure compréhension des risques aussi bien au niveau individuel que global, débouchant sur des segmentations des risques encore plus fines. L'analyse prédictive portée par la communauté des *Data Scientists* et des chercheurs apporte de nouveaux modèles et algorithmes permettant de prédire la survenance des risques non plus au niveau d'une classe de risque, mais au niveau individuel.

Ainsi, la mutualisation est au cœur de la démarche actuarielle. Pour la mener à bien, l'assureur est amené à construire des classes de risque homogènes à partir des variables tarifaires. Les tendances sociales et les avancées technologiques lui confèrent aujourd'hui des possibilités de compréhension et de gestion des risques à une échelle très fine voire individuelle. Cependant, d'importantes conséquences sociales en découlent ainsi qu'un risque de démutualisation.

### 1.1.2 L'individualisation des risques et les conséquences sociales

Si les perspectives de personnalisation de gestion des risques en assurance portées par les avancées technologiques et méthodologiques telles que mises en avant par les *insurtechs* semblent prometteuses, elles ne font pas l'unanimité auprès de la communauté des assureurs. Ces perspectives soulèvent en effet des inquiétudes et des conséquences sur plusieurs plans :

#### Sur le plan social : il y a un risque d'exclusion

L'hyper-segmentation des risques et des primes pourrait conduire à des prix très élevés pour les individus les plus à risque au point de les pousser à renoncer à la souscription à une couverture d'assurance comme le précisent Keller *et al.* (2018). Ce risque d'exclusion est d'autant plus préoccupant que le risque couvert est corrélé avec un faible revenu ou une faible richesse comme c'est le cas en assurance santé. Bien que cette préoccupation ne soit pas nouvelle, elle est susceptible de s'accroître avec la société numérique. Ce qui impliquerait en particulier que les plus risqués qui ont le plus besoin d'assurance ne seraient pas couverts alors que les bons risques paieraient des montants de primes *négligeables*. Inversement, Cevolini et Esposito (2020) évoquent un effet paradoxal : les individus les moins risqués qui auraient souscrit à une police pourraient modifier leurs habitudes (en étant moins prudents par exemple) puisqu'une prime individualisée faible agirait comme un signal leur révélant qu'ils ont très peu à craindre du risque couvert. L'hyper-segmentation pourrait alors conduire indirectement l'assureur à promouvoir l'aléa moral.

#### Sur le plan de l'équité : il y a un risque d'« injustice actuarielle »

La controverse de l'équité sociale en assurance qui découle de la tarification basée sur les facteurs de risque trouve ses racines dans l'utilisation ou non des facteurs causaux du risque. Frezal et Barry (2020) discutent du rôle de l'assurance dans la justice individuelle et sociale à travers la segmentation des prix. S'il semble que l'hyper-segmentation tend à renforcer l'équité

individuelle telle que réclamée par les assurés, au niveau de la société, la justesse de cette pratique est très controversée. Si l'équité sociale est perçue comme « *un traitement équitable des assurés soumis aux mêmes risques* » (Frezal et Barry, 2020), la personnalisation ne contribuerait à renforcer l'équité sociale que si et seulement si elle est basée sur les facteurs qui causent effectivement le risque et non des facteurs seulement corrélés au risque comme c'est souvent le cas. Hyper-segmenter à travers l'ensemble des facteurs et données disponibles pourrait donc conduire à des discriminations injustes entre les assurés.

### **Sur le plan technique : changement de paradigme de l'assurance**

---

En passant de la modélisation prédictive d'un niveau groupe à celui de l'assuré, ou de très peu d'assurés par classe, les primes obtenues pourraient être hautement volatiles et difficiles à communiquer et in fine à mettre en pratique. En effet, Hay (2015) précise qu'il n'est pas « facile » pour les compagnies d'assurance de passer d'une dizaine d'informations contractuelles mises à jour chaque année aux flux de données comportementales et contextuelles des capteurs entourant les risques. Bien que les techniques et outils d'analyse quantitative se développent rapidement, la solidité et la crédibilité des prix en assurance reposent sur des moyennes dont la stabilité est fonction du nombre d'assurés sous-jacents. Dit autrement, l'application de la loi des grands nombres suppose qu'il y ait un grand nombre d'assurés qui peuplent une classe de risque donnée. Cette inquiétude rejoint les préoccupations de Barry et Charpentier (2020) quant au changement de perspective de l'assurance qui découlerait de l'arrivée du *big data* en assurance.

### **Sur le plan politique : remise en cause de la place de l'assurance**

---

D'un point de vue politique, la raison d'être sociale d'une protection d'assurance est d'apporter une sécurité et de renforcer la confiance des personnes, ceci en les protégeant contre des aléas qui les menacent tant sur le plan de leur intégrité physique qu'au niveau de leurs biens exposés à des risques. Ces aléas ne sont en général pas enrayables à l'échelle individuelle, mais à l'échelle d'une communauté d'assurés, ils le deviennent. Ce rôle est porté principalement par l'État et par les entreprises d'assurance grâce à la mutualisation des cotisations. Ainsi, une hyper-segmentation excluant une partie de la population d'une couverture aurait pour effet d'écarter certains risques pour les reporter sur le rôle social de l'État (et son modèle économique fondé sur l'impôt). L'assurabilité est donc un enjeu de place pour les assureurs comme le précise (Hay, 2015).

Le risque de démutualisation est encore contenu aujourd'hui grâce aux régulateurs qui imposent des mesures dans l'optique de maintenir un niveau acceptable d'équité et de solidarité au sein d'un portefeuille. Un exemple d'une telle mesure est l'arrêt « *Test-Achat* » de la Cour Européenne de Justice entré en vigueur au sein de l'Union européenne en décembre 2012 interdisant aux assureurs de proposer des produits dont la tarification ou les prestations diffèrent

en fonction du sexe des assurés. Dans le même sillage, la loi de la Financial Conduct Authority (FCA)<sup>1</sup> sur la gouvernance de la tarification des produits d'assurance, qui entrera en vigueur en 2022, peut être évoquée. Elle proscrit aux assureurs de proposer des prix différents pour les assurés partageant les mêmes facteurs de risque et donc a fortiori seraient soumis aux mêmes risques (FCA, 2020). Cette loi vise à enrayer la tendance croissante des prix de contrats observée lors du renouvellement (automatique) des polices d'assurance : le « price walking ». En effet, selon le rapport de l'étude *General insurance pricing practices market study* réalisée en 2018 par la FCA, les assureurs utilisent des modèles pour identifier les assurés qui auraient des probabilités de renouvellement très élevées et augmenteraient ensuite le prix de leur couverture en fin d'exercice. En particulier, ils ont observé sur le périmètre de l'assurance auto que les assurés avec une ancienneté de plus de 5 ans paient plus de 370 £ contre 285 £ pour les nouveaux contrats soit 30 % de plus. Cette pratique, qui se généralise à bien plus que l'assurance auto, s'apparente à de l'individualisation des primes dans la mesure où des personnes ayant des profils de risque similaires sont traitées différemment. Toutefois, cela ne s'oppose pas à la solidarité, puisque les assurés fidèles financent les nouvelles acquisitions, ce qui permet *in fine* l'équilibre du portefeuille. L'équité de cette pratique étant discutable, la solidarité au sein d'un portefeuille n'est pas nécessairement alignée avec l'équité collective, c'est pourquoi une métrique de mutualisation doit s'accompagner d'autres mesures de justesse.

Outre les restrictions des régulateurs, la rigidité du secteur de l'assurance à la technologie et les questions de confidentialité encadrées par le RGPD limitent encore la quantité de données collectées par les assureurs et le niveau de personnalisation des risques atteignable. Le scénario extrême d'hyper-personnalisation où chaque assuré aurait un profil de risque unique et connu de l'assureur bien que théoriquement intéressant pour traiter la question de solidarité au sein d'un portefeuille, n'est donc pas (encore) envisageable en pratique.

Ce débat autour de la segmentation du portefeuille et ses implications bien qu'omniprésent dans la littérature depuis plus d'une décennie n'a pas donné lieu à des indicateurs permettant de quantifier le niveau de mutualisation ou de solidarité d'un portefeuille. L'objectif du présent mémoire est d'apporter une contribution à cette réflexion à partir d'éléments pour la compréhension et pour l'évaluation du niveau de mutualisation ou de démutualisation d'un portefeuille réel. Ne pouvant séparer la justesse des prix de la notion de démutualisation, ce mémoire apporte également des éléments pour apprécier l'équité résultante d'une pratique de démutualisation. Une mise en application de la démutualisation et de l'analyse de l'équité sera effectuée sur un portefeuille réel de BNP Paribas Cardif dans les chapitres 4, 5 et 6.

## 1.2 Partage du risque entre assureur et assurés

Pour construire et maintenir aussi bien techniquement que commercialement un portefeuille d'assurance, il peut être nécessaire de procéder à une segmentation des assurés en classes à

---

1. le régulateur anglais des services et marchés financiers

partir des variables tarifaires. Au sein de chaque classe, il subsiste une hétérogénéité résiduelle. D'une part les facteurs de risque ne sont que partiellement connus ou observés et d'autre part, les modèles utilisés pour construire les classes homogènes embarquent des erreurs d'estimation. Une autre source d'hétérogénéité résiduelle est le *hasard pur* qui ne pourrait être prédit même dans le cas où l'assureur disposerait de tous les facteurs associés à un risque donné. Dans un scénario où les effectifs intra-classes seraient suffisamment grands, l'on pourrait s'attendre à observer une compensation entre les primes perçues et les sinistres payés d'une classe ; la loi des grands nombres s'appliquant dans chaque classe. En pratique, l'homogénéité à l'intérieur d'une classe s'améliore avec le nombre de facteurs de risque utilisés ; mais inversement, le nombre d'assurés par classe diminue. Il en résulte alors une instabilité des primes et des déséquilibres dans les classes. La solidarité joue au niveau du portefeuille ; elle consiste à privilégier l'équilibre du portefeuille en lieu et place de l'équilibre des classes de risque homogènes. Les primes payées par les assurés ne servent plus qu'à indemniser les assurés de la même classe de risque, mais peuvent être transférées vers d'autres classes de risque.

La solidarité et la mutualisation se rapportent directement à un portefeuille, il est alors nécessaire d'apporter une précision à la notion de portefeuille en assurance. Selon IFRS 17, le portefeuille de produits désigne les contrats soumis au même type de risque et gérés de la même manière. Par exemple, les contrats de la même ligne de produits chez un assureur donné (comme l'assurance automobile) sont censés appartenir au même portefeuille. IFRS 17 précise également qu'il est nécessaire de regrouper les contrats suivant le degré de rentabilité, lors de la comptabilisation initiale en utilisant les critères suivants : un groupe de contrats déficitaires dès leur comptabilisation initiale ; un groupe de contrats qui n'ont pas, lors de leur comptabilisation initiale, de possibilité significative de devenir déficitaire et le groupe des contrats restants. Dans le cadre de cette étude, un portefeuille se rapporte à un groupe de contrats couverts contre un risque donné, sur un exercice donné, administrés conjointement, sur lesquels les mêmes facteurs de risque sont observés et les montants de prime commerciale (éventuellement différents) sont reportés. Il n'y a aucune prise en considération du regroupement des contrats suivant leur rentabilité observée ou anticipée.

### 1.2.1 Partage du risque et mutualisation (Denuit et Charpentier, 2004)

Dans cette partie, la pratique de segmentation est formalisée, le sens de la mutualisation et celui de la solidarité sont précisés. Soit un assuré choisi aléatoirement dans un portefeuille d'assurance ; il est soumis à un risque dont l'ampleur est désigné par la variable aléatoire  $S$  définie sur  $\mathbb{R}^+$ . L'ensemble des facteurs de risque utilisés pour construire la prime pure sont consignés dans le vecteur aléatoire  $\mathbf{X} = (X_1, X_2, \dots, X_p)$ . Le vecteur aléatoire  $\mathbf{\Omega} = (\Omega_1, \Omega_2, \dots)$  représente l'ensemble des caractéristiques de l'assuré, du sous-jacent couvert et du contrat qui influent sur la survenance et la sévérité du risque, en général l'assureur ne l'observe qu'imparfaitement.

En étudiant trois scénarios de partage de risque entre assureur et assurés, une mesure de démutualisation d'un portefeuille induite par le choix de segmentation tarifaire de l'assureur se déduit des travaux de Denuit et Charpentier (2004, chap. 3) dans un cas simple où l'assuré ne paie que la prime pure ou la prime pure majorée d'un montant fixe de chargement.

### **En absence de segmentation la solidarité est maximale**

---

En absence de segmentation des primes, chaque assuré cotise le même montant déterminé par la sinistralité moyenne anticipée du portefeuille, indépendamment de ses facteurs de risque propres (tableau 1.1 ci-contre). La solidarité au sein du portefeuille est alors maximale dans la mesure où indépendamment des spécificités individuelles, les cotisations de tous les assurés sont mises en commun pour assurer le dédommagement des personnes sinistrées. Toute la variance des sinistres est portée par l'assureur, que celle-ci soit liée à la structure du portefeuille ou à la variabilité intrinsèque des montants de sinistres. L'équilibre financier de l'assureur à la fin de l'exercice est tributaire d'un alignement entre la structure du portefeuille (c'est-à-dire la distribution jointe des facteurs de risque) constitué pendant l'année et celle du portefeuille utilisé pour évaluer  $E[S]$ , soit la structure du *portefeuille de marché*. Dans le cas d'une déviation majeure de la structure du portefeuille assuré par rapport au portefeuille de marché, l'équilibre financier serait rompu. L'absence de segmentation correspond alors à un schéma où la solidarité est maximale au sein du portefeuille et où l'entièreté du risque de structure du portefeuille est portée par l'assureur.

TABLEAU 1.1 – Partage du risque entre assurés et assureur en l'absence de segmentation.

<b>Acteur</b>	<b>Assurés</b>	<b>Assureur</b>
Dépense	$\mathbb{E}[S]$	$S - \mathbb{E}[S]$
Dépense moyenne	$\mathbb{E}[S]$	0
Variance	0	$\mathbb{V}[S]$

### **En segmentation parfaite il n'y a pas de solidarité inter-classes de risque**

---

Dans le cas (théorique) où l'assureur observerait tous les facteurs de risque  $\Omega$  et les incorporerait dans son processus de tarification, le partage du risque entre assureur et assurés s'opèrerait comme indiqué dans le tableau 1.2 ci-contre :

TABLEAU 1.2 – Partage du risque entre assurés et assureur en segmentation parfaite.

<b>Acteur</b>	<b>Assurés</b>	<b>Assureur</b>
Dépense	$\mathbb{E}[S \Omega]$	$S - \mathbb{E}[S \Omega]$
Dépense moyenne	$\mathbb{E}[S]$	0
Variance	$\mathbb{V}[\mathbb{E}[S \Omega]]$	$\mathbb{E}[\mathbb{V}[S \Omega]]$

Contrairement au cas précédent, les primes payées par les assurés ne sont plus constantes, mais dépendent des caractéristiques  $\Omega$  propres à chaque assuré et dès lors la prime  $\mathbb{E}[S|\Omega]$  est aléatoire d'un assuré à l'autre au regard de la distribution des facteurs de risque  $\Omega$ . Celle-ci en moyenne est égale à l'espérance de sinistre du portefeuille :  $\mathbb{E}[\mathbb{E}[S|\Omega]] = \mathbb{E}[S]$ . En segmentant les risques, l'assureur désolidarise le portefeuille puisque chaque assuré paie une prime proportionnelle à son risque. La variance des primes  $\mathbb{V}[\mathbb{E}[S|\Omega]]$  correspond à un transfert du risque de structure de la part de l'assureur vers les assurés. L'assureur devient insensible aux déviations éventuelles de la structure de son portefeuille par rapport au portefeuille de marché. Le portefeuille de marché est assimilable à une distribution de la sinistralité de l'ensemble des assurés, de tous les assureurs d'un marché donné, suivant l'ensemble des facteurs de risque  $S|\Omega$ . La connaissance de cette distribution est nécessaire pour calculer  $\mathbb{V}(\mathbb{E}(S|\Omega))$  afin de transférer l'intégralité des déviations de structure du portefeuille assuré dans la prime pure. En alignant sa tarification à la distribution du marché, l'équilibre financier d'un assureur est la résultante de l'équilibre au sein de chaque classe de risque. Si le portefeuille de l'assureur n'est pas représentatif du portefeuille de marché, cela n'a pas une grande influence sur l'équilibre des risques entrés en portefeuille (à condition que ceux-ci soient suffisamment nombreux au sein de toutes les classes de risque).

La variance des montants de sinistres portée par l'assureur se détermine comme suit :

$$\begin{aligned}\mathbb{V}[S - \mathbb{E}[S|\Omega]] &= \mathbb{E} [(S - \mathbb{E}[S|\Omega])^2] \\ &= \mathbb{E} [\mathbb{E} [(S - \mathbb{E}[S|\Omega])^2|\Omega]] \\ &= \mathbb{E}[\mathbb{V}[S|\Omega]]\end{aligned}$$

Au sein d'une classe de risque donnée, l'assureur n'intervient que pour mutualiser les risques résultant du hasard pur, c'est-à-dire les risques imprévisibles malgré la connaissance de  $\Omega$ . En effet,  $\mathbb{V}[S|\Omega]$  représente les fluctuations de  $S$  dues au seul hasard lié aux des facteurs de risque constants. Dans cette situation, l'assureur mutualise le risque intra-classe et il n'y a aucune solidarité induite entre les assurés du portefeuille : chacun paie en fonction de son propre risque.

Il s'en suit un partage de variance entre assureur et assurés :  $\mathbb{V}[S] = \underbrace{\mathbb{E}[\mathbb{V}[S|\Omega]]}_{\rightarrow \text{assureur}} + \underbrace{\mathbb{V}[\mathbb{E}[S|\Omega]]}_{\rightarrow \text{assurés}}$ .

L'assureur prenant en charge la volatilité due au *hasard pur* et les assurés celle résultant de la structure du portefeuille.

### **En segmentation imparfaite, l'assureur combine solidarité et mutualisation**

---

Les deux scénarios illustrés plus haut ne sont généralement pas observés sur le marché. En effet, l'absence de segmentation conduit à de l'anti-sélection sur un marché concurrentiel (sur lequel les concurrents segmentent) ou alors sur des produits d'assurance optionnels. L'accès à l'information parfaite sur les facteurs de risque se heurte encore aujourd'hui à des limitations



d'assurance est présenté à la section 3.1). Dans cet exemple illustratif, les seuls facteurs de risque (prenant chacun deux modalités) suivant leur pertinence sont :

- AGE-VEH (âge du véhicule) : véhicule neuf (VN) ou en véhicule d'occasion (VO) ;
- TYPE-VEH (type de véhicule) : Camping-Car (CC) ou Automobile (Auto) ;
- MARQUE (marque du véhicule) : Marque A (MA) ou Marque B (MB).

L'évolution du niveau de mutualisation ou de solidarisation du portefeuille en fonction des facteurs de risque utilisés dans la tarification est représentée sur la figure 1.2. Au sommet de l'arbre, l'assureur propose un prix unique, i.e.  $\mathbb{E}[S]$ , la solidarité est alors maximale. En incorporant l'âge du véhicule dans sa tarification, il propose désormais deux prix différents à savoir  $\mathbb{E}[S|VN]$  et  $\mathbb{E}[S|VO]$  ; une désolidarisation du portefeuille suivant l'âge du véhicule est introduite. Néanmoins, il subsiste une solidarité suivant le type de véhicule et la marque. À la base de l'arbre, l'assureur constitue autant de segments qu'il y a de profils de risque et chaque assuré cotise un montant représentatif de son propre risque. La mutualisation se fait entre les contrats partageant les mêmes caractéristiques et il n'y a aucune forme de solidarité entre les profils de risque différents.

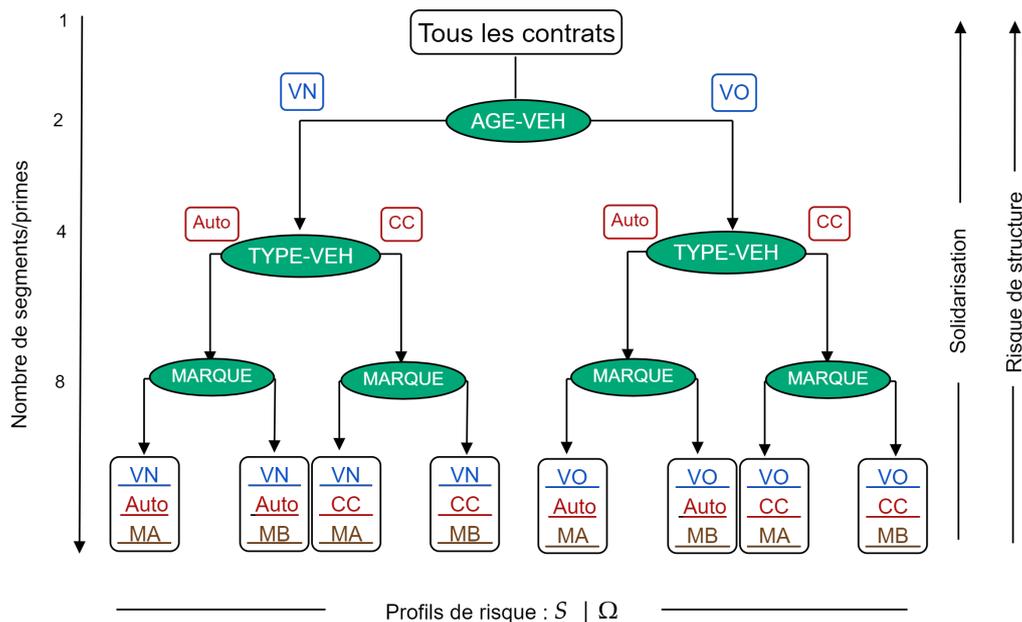


FIGURE 1.2 – Segmentation des primes au regard des facteurs de risque et solidarisation au sein du portefeuille : cas de l'extension de garantie automobile avec 3 facteurs de risque.

Cette formalisation du partage du risque et le lien avec la solidarité et la mutualisation bien que naturelle présente quelques limites qu'il convient de soulever :

- Le partage de risque se fait exclusivement suivant la prime pure en omettant la prime commerciale alors que les choix tarifaires majeurs de l'assureur s'opèrent lors du passage de la prime pure à la prime commerciale ;

- La seule mesure de risque utilisée est la variance, d'autres mesures de dispersion pourraient être envisagées comme l'entropie qui met plus en avant la diversité que les écarts extrêmes contrairement à la variance ;
- Ce partage du risque correspond à un transfert du risque de structure entre assureur et assurés ce qui nécessite de connaître la structure du marché auquel le produit s'adresse pour s'en prémunir ou en suivre l'évolution ;
- En pratique, l'estimation de la loi conditionnelle de  $S$  sachant  $\mathbf{X}$  ( $S|\mathbf{X}$ ) pour évaluer  $\mathbb{E}[S|\mathbf{X}]$  et  $\mathbb{V}[S|\mathbf{X}]$  passe par la calibration des modèles statistiques. Évaluer le niveau de démutualisation sur les primes est donc fortement dépendant des modèles et des choix de l'actuaire qui les a élaborés.

### 1.2.2 Partage du risque en présence de chargements assureurs

Les chargements assureurs jouent un rôle majeur dans la stratégie d'acquisition et de mutualisation de l'assureur. Il convient alors d'analyser comment évolue le partage du risque précédent en prenant en compte les chargements assureurs.

#### De la prime pure à la prime commerciale

Jusqu'ici, la prime évoquée ne représentait que la prime pure (ou *burning cost*) qu'il convient de distinguer de la prime effectivement versée par l'assuré : la prime commerciale. En effet au montant de prime pure, déterminé en tarification à l'aide des facteurs de risque (dont le principe est présenté dans la section 2.2.1), il faut ajouter en pratique des chargements (positifs ou négatifs) qui seront désignés par *optimisation layer* (OL) tout au long de ce mémoire. Cette couche d'OL (figure 1.3) comporte :

- La marge assureur : elle comprend la marge de sécurité pour couvrir les écarts entre la sinistralité et les encaissements ainsi que les bénéfices escomptés pour rémunérer les actionnaires ;
- Les commissions : elles sont versées par l'assureur aux partenaires et courtiers. Elles se calculent comme une fraction de la prime pure ou sont déterminées en fonction de la stratégie de l'assureur ;
- Les frais généraux : comprenant les frais administratifs et les frais d'acquisition, ils garantissent à l'assureur le paiement de ses charges.

Commissions	Optimisation Layer (OL)	Prime commerciale
Frais d'acquisition		
Frais administratifs		
Marge assureur		
Prime pure / Burning cost (BC)	Profil de risque	

FIGURE 1.3 – Décomposition de la prime commerciale en optimisation layer (OL) et en prime pure.

L'optimisation layer apparaît comme un levier sur lequel l'assureur peut s'appuyer pour ajuster le niveau de mutualisation de son portefeuille. En décidant d'appliquer des taux de marge plus importants ou plus faibles sur certains périmètres de portefeuille, l'assureur crée de la solidarité additionnelle même si celle-ci est sujette à discussion (voir réglementation sur le « **price-walking** » évoquée plus haut). De façon plus pragmatique, pour des niveaux de risque différents au regard de la prime pure, l'assureur peut décider de jouer sur l'OL pour faire payer la même prime commerciale à tous les assurés aboutissant ainsi à un portefeuille solidaire suivant la prime commerciale. De même, l'assureur peut utiliser l'OL pour financer l'acquisition et la rétention des contrats, en appliquant notamment des remises à l'entrée ou aux contrats ayant une probabilité de résiliation élevée. Ces remises sont alors financées par les autres assurés en portefeuilles.

### Mutualisation des risques en présence d'optimisation layer

En s'appuyant sur l'optimisation layer, l'assureur peut réduire ou accroître le niveau de démutualisation de son portefeuille selon qu'il corrèle ou non l'OL avec la prime pure ou plus généralement les facteurs de risque. L'OL peut se formaliser comme une fonction des caractéristiques des assurés à valeur dans  $\mathbb{R}$ . Elle sera notée à cet effet  $OL(\mathbf{X})$ ; toutefois cela n'exclut pas le scénario où l'OL serait indépendant des facteurs de risque. Le partage du risque en segmentation imparfaite, la prime pure étant majorée par l'OL se présente comme dans le tableau 1.4.

L'équilibre financier de l'assureur ne correspond plus à une dépense moyenne nulle, mais équivaut à l'opposé du résultat technique de l'opération d'assurance. Ce résultat technique est financé par les assurés et les scénarios favorables pour l'assureur où la sinistralité observée serait moins importante que celle attendue.

La variance des primes portée par les assurés augmente (algébriquement) de  $\mathbb{V}[OL(\mathbf{X})] +$

$2Cov[\mathbb{E}[S|\mathbf{X}], OL(\mathbf{X})]$ , donc plus les chargements assureurs sont différents d'un assuré à l'autre plus la solidarité entre assurés est importante. La corrélation entre les primes pures et les OL modifie également la part de risque structurelle portée par les assurés. En effet si  $Cov[\mathbb{E}[S|\mathbf{X}], OL(\mathbf{X})]$  est positif (i.e. les marges assureurs sont d'autant plus importantes que l'assuré est un mauvais risque), la désolidarisation augmente et inversement si cette corrélation est négative (l'assureur tend à proposer un tarif unique en fixant des marges importantes sur les bons risques et des marges faibles voir négatives sur les mauvais risques). Un cas particulier est celui où l'optimisation layer est indépendante ou du moins décorrélée de la sinistralité. Sous cette condition, la variance des montants à charge des assurés se décompose comme une somme des variances des deux couches qui constituent la prime commerciale.

TABLEAU 1.4 – Partage du risque entre assurés et assureur selon le niveau de segmentation avec prise en compte des chargements assureurs.

Acteur	Assurés	Assureur
Dépense	$\mathbb{E}[S \mathbf{X}] + OL(\mathbf{X})$	$S - \mathbb{E}[S \mathbf{X}] - OL(\mathbf{X})$
Dépense moyenne	$\mathbb{E}[S] + \mathbb{E}[OL(\mathbf{X})]$	$-\mathbb{E}[OL(\mathbf{X})]$
Variance	$\mathbb{V}[\mathbb{E}[S \mathbf{X}]] + \mathbb{V}[OL(\mathbf{X})] + 2Cov[\mathbb{E}[S \mathbf{X}], OL(\mathbf{X})]$	$\mathbb{E}[\mathbb{V}[S \mathbf{X}]] + \mathbb{V}[OL(\mathbf{X})] - 2Cov[S, OL(\mathbf{X})] + 2Cov[\mathbb{E}[S \mathbf{X}], OL(\mathbf{X})]$

### 1.3 Proposition de définition de la démutualisation

Dans le cadre de ces travaux, nous proposons de définir la démutualisation d'un portefeuille comme la désolidarisation des assurés de l'ensemble de la communauté des assurés, le cas extrême étant une prestation individuelle certaine avec un niveau de marge constant par assuré, sans aucun subside croisé. Dans ce cas extrême, l'équilibre financier d'un portefeuille, donc de la communauté, est inchangé (en proportion) si n'importe quel assuré décide de le quitter ou à l'inverse d'y entrer. Ainsi, un portefeuille sera d'autant plus démutualisé que chaque assuré cotise pour son risque indépendamment des autres assurés. Inversement, un portefeuille sera dit *mutualisé* si au sein de celui-ci l'assureur fait jouer la solidarité entre les assurés en plus de la mutualisation entre les assurés ayant le même profil de risque lorsque tous les facteurs de risque sont pris en considération.

La figure 1.4 fournit des illustrations dans l'optique de mieux appréhender cette proposition de définition. Le portefeuille considéré est celui introduit dans la section 1.2.1 sur l'extension de garantie sur les pannes mécaniques. Les pools de contrats sont triés par profil de risque croissant : VN|Auto|MA, VN|Auto|MB, VN|CC|MA, VN|CC|MB, VO|Auto|MA, VO|Auto|MB, VO|CC|MA et VO|CC|MB. À chacun de ces segments de risque, est associé une prime pure, notée respectivement  $BC_1, BC_2, \dots, BC_8$  et des OL, notées respectivement  $OL_1, OL_2, \dots, OL_8$ .

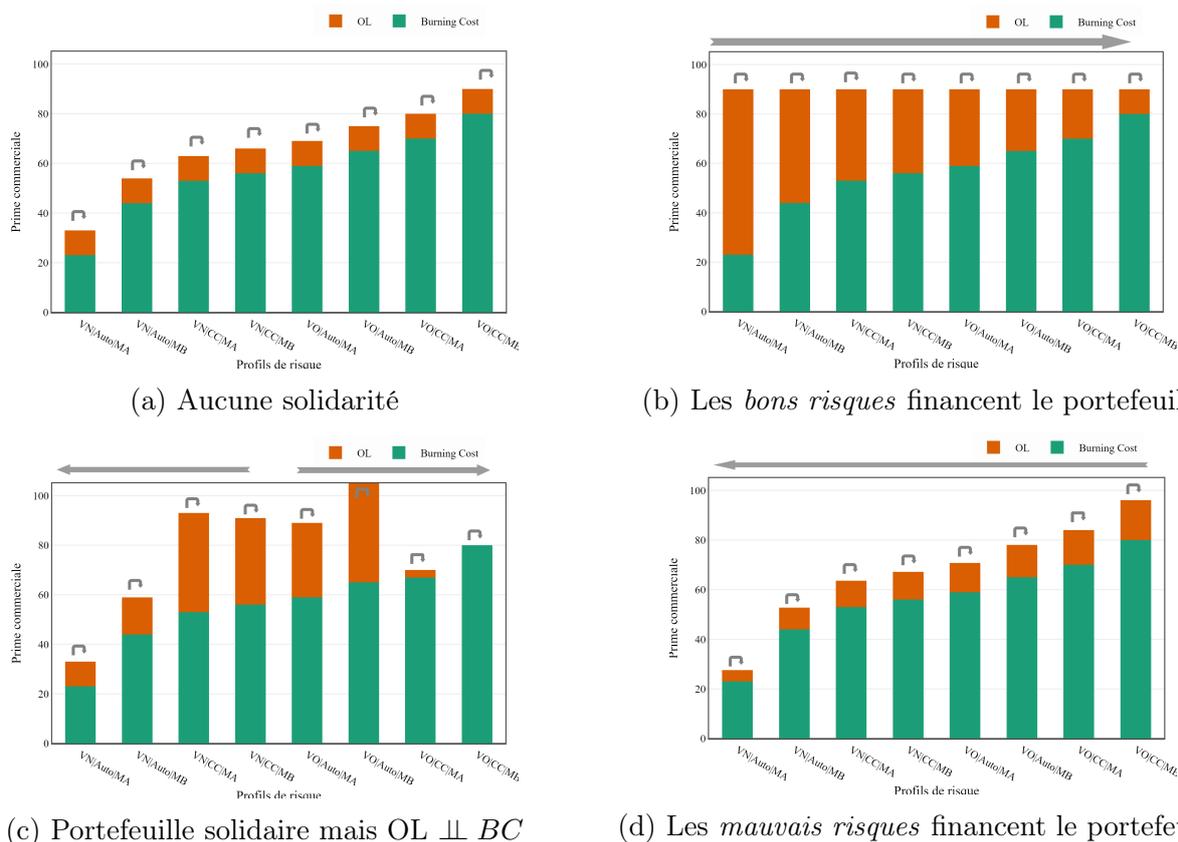


FIGURE 1.4 – Illustration de la démutualisation dans un portefeuille en présence d'OL.

On peut distinguer quatre schémas avec des niveaux de démutualisation différents illustrés sur cette figure.

- **a. Pas de solidarité au sein du portefeuille** : dans le cas 1.4a, chaque assuré paie pour son risque majoré d'un même montant d'OL pour tous les assurés. L'équilibre du portefeuille est alors la résultante des équilibres individuels par profil de risque et donc la perte d'une fraction (éventuellement importante) du portefeuille ne devrait pas induire un déséquilibre du portefeuille (proportionnellement). Le risque de structure porté par l'assureur se limite à un risque de modèle. Un indice de mutualisation sur de tels portefeuilles devrait prendre des valeurs faibles puisque la seule source de solidarité ici est la mutualisation intra-classe de risque ( $\mathcal{Q}$ ). Inversement, en proposant de définir la démutualisation comme l'absence de solidarité dans le portefeuille, un indice de démutualisation sur de tels portefeuilles devrait prendre des valeurs élevées ;
- **b. Les moins risqués financent les plus risqués** : dans le cas 1.4b, il y a une forte solidarité dans le portefeuille car tous les assurés paient le même montant de prime. En particulier, les *bons risques* (VN|Auto|MA, VN|Auto|MB, VN|CC|MA) se voient appliquer des OL plus importantes, ce qui pourrait permettre de réduire les chargements sur les mauvais risques (VO|CC|MA, VO|CC|MB) afin de les inciter à souscrire à l'assurance. L'équilibre du portefeuille est la résultante d'une compensation des flux entre les assurés, c'est-à-dire que la perte d'une fraction (éventuellement faible) du portefeuille (les *bons*

*risques* par exemple) pourrait induire son déséquilibre (proportionnellement à la fraction concernée), le risque de structure est important dans ce cas. L'indice de mutualisation proposé devrait prendre des grandes valeurs pour ce type de portefeuille, puisque l'assureur s'appuie sur deux formes de mutualisation : la mutualisation intra-classe (illustrée par  $\odot$ ) et la solidarité inter-classe (illustrée par  $\implies$ ). Inversement, un indice de démutualisation mesuré sur un tel portefeuille devrait atteindre sa valeur théorique minimale ;

- **c. un cas de solidarité où l'OL est indépendante du risque** : dans le cas 4, l'assureur introduit une solidarité dans le portefeuille qui n'est pas liée au risque. Des profils dont le niveau de risque est modéré comme VN|CC|MA, VN|CC|MB, VO|Auto|MA et VO|Auto|MB paient des chargements plus importants, ce qui permet notamment de financer aussi bien les bons que les mauvais risques (illustré par  $\Leftarrow$  et  $\Rightarrow$ ). Il n'est d'ailleurs pas exclu que les mauvais risques voient leur prix commercial final tomber en dessous de leur prime pure. Il s'agit notamment du cas où l'assureur accroît les marges sur certains profils pour financer l'acquisition d'autres profils, à qui il commercialise son produit à perte. Une telle pratique peut s'inscrire dans la stratégie de long terme de l'assureur ou être contrainte par les tendances de marché ;
- **d. Les plus risqués financent le portefeuille** : dans le cas 1.4d, l'assureur applique des marges proportionnelles au burning cost ce qui a pour conséquence d'accroître davantage la prime commerciale payée pour les contrats les plus à risque. Ils paient donc plus que le prix de leurs risques, ce qui pourrait permettre de financer les bons risques sur lesquels les marges appliquées sont moins importantes. Il s'agit là encore d'une forme de solidarité au sein du portefeuille et donc le comportement des indicateurs de mutualisation et de démutualisation devrait s'apparenter à celui du cas (b). Toutefois, il convient de souligner qu'ici l'équité individuelle n'est pas sacrifiée au profit de l'équité sociale dans la mesure où la prime finale payée par chaque assuré est proportionnelle à son niveau de risque.

## Conclusion du chapitre

Ce chapitre a permis de définir les concepts fondamentaux en assurance que sont la mutualisation et la solidarité. La segmentation en tarification apparaît comme étant d'une importance majeure pour se protéger de l'anti-sélection et du risque de structure. Toutefois, pousser la segmentation à l'extrême comporte des risques sociaux tels que l'exclusion des assurés ayant le plus besoin d'assurance. Partant du partage du risque entre assureur et assurés, nous avons proposé de définir la démutualisation comme l'absence de solidarité entre les classes de risque constituées à partir de l'ensemble des facteurs de risque dont disposent les assureurs pour un produit donné.

Cette définition étant posée, il convient à présent de proposer des indicateurs quantitatifs du niveau de mutualisation d'un portefeuille qui s'inscrivent en ligne avec ces intuitions actuarielles.

## CADRE THÉORIQUE ET PROPOSITIONS D'INDICATEURS DE MUTUALISATION

### Sommaire

<b>2.1</b>	<b>Introduction à l'entropie sous plusieurs formes.</b>	<b>20</b>
<b>2.2</b>	<b>Indices de mutualisation simples basés sur la prime et ses composantes</b>	<b>30</b>
<b>2.3</b>	<b>Introduction à la complexité de Kolmogorov</b>	<b>52</b>
<b>2.4</b>	<b>Indices de mutualisation exotiques basés sur les caractéristiques des assurés.</b>	<b>63</b>

Ce chapitre est consacré à la construction théorique des indicateurs de démutualisation. Les intuitions et l'état de l'art autour de la mutualisation ont permis d'identifier deux leviers de mutualisation à disposition de l'assureur : La différenciation des profils de risque par segmentation tarifaire et le traitement différentiel des assurés sur la base des chargements assureurs. Quantifier la mutualisation d'un portefeuille semble s'accorder avec une mesure de la dispersion et de l'hétérogénéité des profils de risque en portefeuille et des traitements commerciaux par assurés. En distinguant deux grandes familles de mesures de dispersion que sont la variance et l'entropie et deux classes de grandeurs sur lesquelles apprécier cette dispersion, deux familles d'indicateurs de mutualisation sont proposées dans ce chapitre.

La première famille d'indicateurs est dite simple ou vanille de par leur formulation intuitive et leur méthode de calcul relativement simple. Ces indicateurs résultent d'une application des mesures de dispersion sur la prime et ses composantes (prime pure et Optimisation Layer, voir figure 1.3) qui résument le profil de risque et/ou les chargements assureurs. Afin de justifier l'utilisation de l'entropie comme mesure de dispersion et de familiariser le lecteur avec cette notion, ce chapitre s'ouvre sur une introduction à l'entropie sous plusieurs formes. Un rappel de la démarche d'estimation de la prime pure est fait par la suite dans l'optique de légitimer pourquoi et comment la prime pure synthétise le profil de risque. La construction des indicateurs simples est ultérieurement présentée dans la section 2 du présent chapitre.

Les indicateurs exotiques tendent à généraliser les indicateurs simples en restreignant les hypothèses de calcul. Pour y parvenir, ils sont calculés sur les caractéristiques plutôt que sur des grandeurs agrégées. Ils nécessitent en conséquence des généralisations de l'entropie et de la variance qui s'appliqueraient sur des vecteurs. La troisième section de ce chapitre introduit la complexité algorithmique ou complexité de Kolmogorov qui est une brique essentielle au calcul de l'entropie algorithmique, qui se présente comme une généralisation de l'entropie. La dernière

section de ce chapitre est consacrée à la construction des indicateurs exotiques qui mesurent la démutualisation à partir de l'entropie algorithmique et l'inertie d'un nuage de points. La figure 2.1 ci-contre résume les outils et les grandeurs utilisés pour construire les indicateurs de démutualisation d'un portefeuille assurantiel.

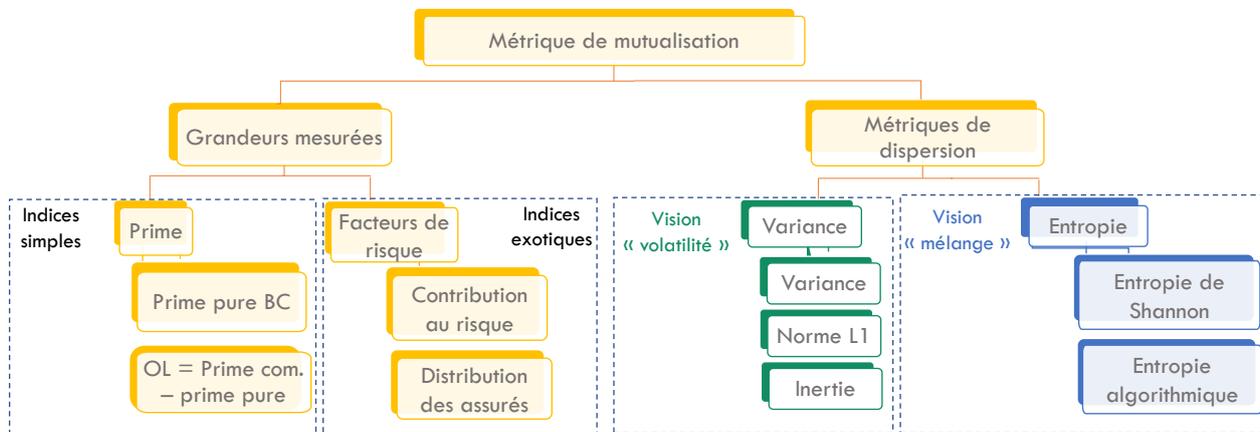


FIGURE 2.1 – Métriques de dispersions et grandeurs utilisées pour construire les indicateurs de démutualisation.

## 2.1 Introduction à l'entropie sous plusieurs formes

Le terme entropie a été introduit par le physicien Rudolf Clausius en 1865 à partir du mot grec « *entropé* » qui signifie « transformation ». Il mesure le niveau de désordre, d'imprédictibilité du contenu en information d'un système physique. Il a été popularisé par sa formulation probabiliste introduite par Ludwig Boltzman et Willard Gibbs (Jaynes, 1965), mais aussi par ses applications en théorie de l'information mises en avant par Shannon (1948). L'entropie introduite en thermodynamique s'est généralisée à plusieurs domaines comme : l'astrophysique, la théorie de l'information, la statistique etc. Le socle commun qui revient dans ces différentes applications de l'entropie est son emploi comme mesure d'hétérogénéité. C'est dans cette optique que l'entropie sera utilisée dans ce mémoire, d'abord pour mesurer la dispersion des primes pures/OL dans un portefeuille et ensuite pour mesurer l'hétérogénéité entre les assurés en portefeuille.

### 2.1.1 Entropie en thermodynamique

La thermodynamique est la branche de la physique qui étudie les échanges d'énergie entre les systèmes ou entre les systèmes et le milieu extérieur, lors des transformations de la matière. Un système physique est une collection d'objets dont l'état est déterminé par un ensemble de grandeurs physiques caractéristiques mesurées sur ces objets ; il s'agit des variables d'état. La pression, la température, la vitesse et la masse sont des exemples courants de variables d'état. Un système se distingue de son milieu extérieur, par une frontière qui peut être tangible ou

non. Lors d'une transformation, un système peut échanger de l'énergie avec son environnement s'il n'est pas isolé. Une transformation désigne la succession d'états qu'un système parcourt entre un état initial et un état final. Considérons à titre d'exemple une casserole contenant de l'eau et posée sur une plaque chauffante. L'eau constitue le système étudié et tout le reste de l'univers y compris la casserole correspond à l'environnement. Les grandeurs d'état comme la température ou la pression de l'eau peuvent être mesurées. Lorsque la plaque est mise en marche, la température de l'eau augmente, le milieu extérieur transmet de la chaleur au système. S'agissant de la description d'un système, deux prismes peuvent être considérés : une description macroscopique ou microscopique.

### L'entropie d'un système en vision macroscopique

L'entropie d'un système physique est une fonction d'état extensive c'est-à-dire dépendante du volume de matière présent dans ce système. Clausius a montré dans ses travaux relatifs au deuxième principe de la thermodynamique que la variation du rapport  $\frac{Q}{T}$  de la quantité de chaleur  $Q$  reçue par le système sur sa température  $T$ , lors d'une transformation restait inférieure ou égale à la variation d'une fonction d'état qu'il appela entropie :  $\Delta S \geq \int \frac{dQ}{T}$ ,  $dQ$  est la quantité de chaleur que le système échange à un point de la frontière avec le milieu extérieur,  $T$  est la température absolue en ce point, l'intégrale est sur l'ensemble des points du système. L'égalité n'est atteinte que si la transformation est réversible, c'est-à-dire qu'elle s'effectue sans frottement, sans dissipation de chaleur. Dans le cas contraire, l'inégalité est stricte et la différence représente la variation d'entropie du milieu extérieur.

Afin de mieux appréhender la relation entre l'entropie ainsi définie et la quantification du désordre au sein d'un système, les deux premiers principes de la thermodynamique sont rappelés à travers un exemple simple, tiré des notes de cours du lauréat du prix Nobel Feynman dans ses lectures (Feynman, 1963). Soit un système formé de deux solides  $S_1$  et  $S_2$  identiques, mais portés à des températures initiales respectives de  $75^\circ C$  et  $25^\circ C$  et placés dans un milieu adiabatique<sup>1</sup>. Ces solides initialement séparés par une paroi isolante, sont mis en contact et les évolutions des températures  $T_1$  et  $T_2$  de ces corps sont présentées sur la figure 2.2.

Le **premier principe de la thermodynamique** qui est un principe de conservation de l'énergie du système postule que : pour tout système thermodynamique, il existe une fonction des variables d'état, appelée énergie totale (notée  $U$ ), dont la variation lors de toute transformation est égale à l'énergie reçue par le système en provenance du milieu extérieur, sous forme de travail ( $W$ ) et de chaleur ( $Q$ ) :

$$\Delta U = W + Q$$

L'application de ce principe à l'exemple ci-dessus implique que l'énergie interne du système global ne varie pas, puisqu'il n'y a aucun échange avec le milieu extérieur. Si  $M$  est la masse de chacun des solides,  $c$  leur capacité calorifique, alors l'énergie interne du système est donnée

---

1. milieu qui empêche tout échange de chaleur avec l'environnement

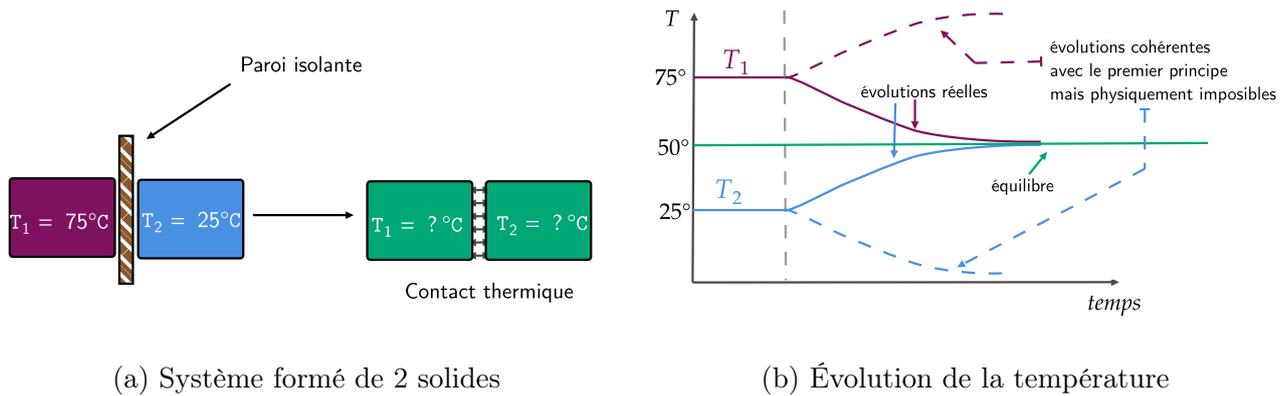


FIGURE 2.2 – Illustration des principes 1 et 2 de la thermodynamique à partir d'un système formé de deux solides ayant des températures initiales différentes.

par  $U = McT_1 + McT_2 = Mc(T_1 + T_2)$ . La constance de  $U$  impose que la somme  $T_1 + T_2$  reste constante, mais rien n'exclut *a priori* que la température du solide  $S_1$  augmente et passe de  $75^\circ\text{C}$  à  $100^\circ\text{C}$  alors que celle du solide  $S_2$  passe de  $25^\circ\text{C}$  à  $0^\circ\text{C}$ . Une telle transformation est contraire à l'expérience. En pratique les températures évoluent toujours en se rapprochant l'une de l'autre :  $T_1$  va diminuer pendant que  $T_2$  va augmenter jusqu'à l'obtention d'une température finale homogène entre les deux solides et égale à  $50^\circ\text{C}$ . Une fois cet état atteint, il n'y a plus d'évolution spontanée du système, le système est dit en état d'équilibre. Le second principe de la thermodynamique rend compte du sens d'évolution d'un système lors de sa transformation. Il prédit en particulier l'évolution du système vers son état d'équilibre correspondant à un état de désordre maximal. L'état d'équilibre d'un système thermodynamique est atteint lorsque les trois conditions suivantes sont satisfaites :

- Les vitesses des points matériels sont nulles et ses variables d'état sont constantes dans le temps ;
- La température du système est homogène (dans l'espace) ;
- Les actions extérieures qui assurent les valeurs constantes des variables d'état sont uniques.

Le **second principe de la thermodynamique** est un principe d'évolution qui introduit la notion d'irréversibilité des phénomènes physiques. Cette irréversibilité est formalisée par la fonction entropie  $S$ . Ce principe d'évolution postule que toute transformation d'un système doit s'effectuer dans le sens d'un bilan entropique global du système et de son milieu extérieur positif, autrement dit d'une création d'entropie. À tout système fermé, peut être associée une fonction d'état extensive  $S$ , appelée entropie, dont la variation lors de toute transformation élémentaire du système est la somme de deux contributions  $d_e S$  et  $d_i S$  :

$$dS = d_e S + d_i S$$

où

- $d_e S$  est due aux échanges de chaleur avec l'extérieur (est donc nulle si la transformation

est adiabatique) ;

- $d_i S$  est toujours positive ou nulle : strictement positive si la transformation est irréversible, nulle si la transformation est réversible.

Dans l'exemple précédent, la transformation est irréversible. Pendant un laps de temps  $dt$  consécutif à la mise en contact des deux solides, le corps chaud  $S_1$  va céder au corps froid  $S_2$  la quantité de chaleur  $\delta Q_{1 \rightarrow 2}$  dont le signe est inconnu *a priori*. La température de  $S_i$  va passer de  $T_i$  à  $T_i + dT_i$ . Le système étant supposé fermé et la transformation irréversible,  $d_e S = 0$  et  $d_i S \geq 0$  donc  $dS > 0$ .

Si le corps  $S_1$  était considéré comme un système à lui tout seul et le corps  $S_2$  le milieu extérieur, alors la variation de température de  $S_1$  pourrait s'apparenter à une transformation réversible dont la variation d'entropie serait  $dS_1 = d_e S_1 = -\delta Q_{1 \rightarrow 2}/T_1$ . De même si le corps  $S_2$  était considéré comme un système, son changement de température pourrait être assimilé à une transformation réversible dont la variation d'entropie avec le solide  $S_1$  (milieu extérieur) serait  $dS_2 = d_e S_2 = \delta Q_{1 \rightarrow 2}/T_2$ .

L'entropie étant une fonction extensive, l'entropie du système formé des deux solides est la somme des entropies de chacun des solides :

$$dS = dS_1 + dS_2 = \delta Q_{1 \rightarrow 2} \left( \frac{1}{T_2} - \frac{1}{T_1} \right) > 0$$

Comme  $T_1 > T_2$ , alors  $\delta Q_{1 \rightarrow 2}$  doit être positif pour que le bilan entropique de la transformation soit bien positif : la chaleur est transférée du corps chaud  $S_1$  vers le corps froid  $S_2$ .

L'évolution se fait donc, d'un état initial ségrégué où les molécules chaudes et les molécules froides sont bien séparées, à un état final dispersé où les molécules chaudes sont passées de  $S_1$  à  $S_2$  et inversement pour les molécules froides. Plus précisément, à l'état initial, la densité d'énergie est plus forte dans le corps chaud que dans le corps froid. En cours d'évolution, il y a transfert d'énergie du corps chaud vers le corps froid, ce qui se traduit par une homogénéisation des températures et une meilleure dispersion de l'énergie thermique. L'entropie d'un système mesure donc le degré de dispersion de l'énergie (sous toutes ses formes : thermique, chimique, électrique) à l'intérieur d'un système. Le second principe stipule que, dans un système isolé, l'énergie a tendance à se disperser le plus possible amenant le système vers un état d'uniformité.

### L'entropie d'un système en vision microscopique

La description précédente du système se fait sous un prisme macroscopique, soit à partir des mesures des agrégats, le *macro-état* est précisé. Toutefois, les valeurs des agrégats qui décrivent ce macro-état peuvent être obtenues sur différentes configurations si le système est observé à un niveau microscopique, c'est-à-dire au niveau des unités élémentaires qui le constituent comme la molécule. Un macro-état est en effet compatible avec un ensemble de micro-états possibles.

Partant de ce paradigme, Boltzmann formula vers 1870 une version statistique de l'entropie qui rend compte de l'incertitude sur la configuration du *micro-état* dans lequel se trouverait un système pour un macro-état donné.

L'exemple du gaz de Boltzmann est repris pour illustrer l'entropie d'un macro-état. Il s'agit d'un système thermodynamique formé d'un volume de gaz dans un cube isolé. Le gaz est formé d'un grand nombre de molécules en perpétuel mouvement et en collision. La somme des énergies de chacune des molécules du gaz définit l'énergie interne  $U$  du système. En plus de son énergie interne, le macro-état de ce gaz est caractérisé par son volume  $V$ . Pour un volume  $V$  et une énergie  $U$  fixés, les particules peuvent être disposées dans le volume d'une pléthore de façons différentes. De même l'énergie interne peut être répartie entre les particules suivant une multitude de configurations. Le nombre de particules dans un système macroscopique, étant immensément grand (de l'ordre de  $10^{23}$ ) le nombre  $\Omega(U, V)$  de ces configurations microscopiques est lui-même immensément grand. L'entropie de Boltzmann de ce gaz est définie par :

$$S = k_B \cdot \log(\Omega(U, V)), \text{ avec la constante de Boltzmann, } k_B = 1,381 \cdot 10^{-23} \text{ J.K}^{-1}$$

où  $J$  signifie Joule et  $K$  Kelvin. Cette formulation peut être réécrite sous un sens probabiliste, en considérant que les micro-états peuvent se réaliser de façon équiprobable. Sous cette hypothèse, une micro configuration  $\omega$  se réalise avec une probabilité de  $\frac{1}{\Omega(U, V)}$  et l'entropie de Boltzmann devient :

$$S = k_B \cdot \log(\Omega(U, V)) = -k_B \sum_{\omega \in \Omega} \frac{1}{\Omega(U, V)} \log\left(\frac{1}{\Omega(U, V)}\right)$$

L'entropie de Gibbs généralise cette formulation au cas où les micro-états ne se réalisent pas de façon équiprobable. Si au micro-état  $\omega$ , est associé la probabilité  $p_\omega$ , l'entropie du macro-état s'écrit alors :

$$S = -k_B \sum_{\omega \in \Omega} p_\omega \log(p_\omega)$$

Pour illustrer le lien entre cette formulation de l'entropie et la mesure de désordre, considérons à nouveau un gaz contenu dans un volume  $V$ , constitué d'environ  $10^{23}$  molécules. Parmi toutes les configurations possibles des molécules dans le volume, il y a un certain nombre qui laisse une moitié du volume vide mais, un nombre extrêmement plus grand de configurations pour lesquelles les molécules sont uniformément réparties dans tout le volume. Comme toutes ces configurations microscopiques sont équiprobables, les répartitions uniformes sont réalisées immensément plus souvent que toute autre situation particulière, au point qu'elles apparaissent macroscopiquement comme un équilibre stationnaire ; et ceci simplement parce que le nombre de configurations microscopiques, et donc l'entropie, qui lui correspondent ont atteint leur valeur maximale.

## 2.1.2 Entropie en théorie de l'information

L'entropie est un concept essentiel en informatique, publié par Shannon en 1948 dans *A mathematical theory of communication*. Cette notion a permis l'essor des communications modernes de la téléphonie jusqu'aux transmissions des données massives par internet. En théorie de l'information, l'entropie est vue comme une mesure de l'incertitude sur la réalisation d'un événement en fonction de la connaissance dont on dispose sur cet événement.

### Formulation intuitive de l'entropie de Shannon

Considérons une source qui émet une information sous forme binaire (séquence de 0 et de 1) et un destinataire qui formule des requêtes à cette source. Toute information transmise par la source doit être codée sous forme de bits et un seul bit à la fois ; un bit étant la plus petite unité d'information échangée. Considérons les situations suivantes :

- Si la source émet toujours la même information ( $x_1$ ), la quantité d'information apportée par une émission supplémentaire est nulle. Le destinataire n'a pas besoin de faire de requête pour lever l'incertitude sur la prochaine information transmise par la source ;
- Si la source émet deux informations ( $x_1$  ou  $x_2$ ) de façon équiprobable, une unique requête permet de lever l'incertitude sur l'information transmise par la source ;
- Si la source émet quatre informations ( $x_1, x_2, x_3$  et  $x_4$ ) de façon équiprobable, il faudra formuler entre une et quatre requêtes pour lever l'incertitude. Mais en adoptant un codage binaire des éventualités ( $x_1 \rightarrow 00, x_2 \rightarrow 01, x_3 \rightarrow 10$  et  $x_4 \rightarrow 11$ ), l'incertitude peut-être levée en 2 requêtes, une par bit suivi d'un décryptage une fois la séquence complète transmise ;
- De même que précédemment, la source émet quatre informations mais avec des probabilités différentes. À titre d'exemple, admettons que l'information  $x_1$  est émise avec une probabilité de  $1/2$ ,  $x_2$  avec une probabilité de  $1/4$ ,  $x_3$  et  $x_4$  avec une probabilité de  $1/8$  chacune. Le codage précédent n'est plus optimal pour minimiser le nombre de requêtes. Pour tirer profit des occurrences plus fréquentes de  $x_1$ , la première requête devrait identifier s'il s'agit de l'information  $x_1$  ou non. La deuxième requête (si besoin) viserait à savoir s'il s'agit de  $x_2$  ou non, et ainsi de suite. Le codage adopté prendrait la forme :  $x_1 \rightarrow 1, x_2 \rightarrow 01, x_3 \rightarrow 001$  et  $x_4 \rightarrow 000$ . Dans 50 % des cas, une requête suffit pour lever l'incertitude, dans 25 % des cas 2 requêtes sont suffisantes et dans 25 % des cas il faut 3 requêtes. Il faut en moyenne,  $1 \times 50\% + 2 \times 25\% + 3 \times 25\% = 1,75 < 2$ .

Dans chacun des cas précédents, la valeur correspondant au nombre de requêtes qu'il faut pour lever l'incertitude est l'entropie de Shannon, notée  $H$ . Elle s'interprète également comme le nombre de bits qu'il faut pour transmettre une information ou la quantité d'informations contenues ou délivrées par une source d'information.

## Formalisation de l'entropie de Shannon

Pour une source émettrice, qui est une variable aléatoire discrète  $X$  comportant  $n$  symboles, chaque symbole  $x_i$  ayant une probabilité  $p_i$  d'apparaître, l'entropie  $H$  de la source  $X$  est définie comme :

$$H(X) = -\mathbb{E} [\log_2 p(X)] = \sum_{i=1}^n p_i \log_2 \left( \frac{1}{p_i} \right) = - \sum_{i=1}^n p_i \log_2 p_i$$

Dans ce cas,  $H(X)$  peut être interprétée comme le nombre moyen de requêtes à réponse binaire que doit formuler le récepteur à la source, ou la quantité d'informations en bits que la source doit fournir au récepteur pour que ce dernier puisse déterminer sans ambiguïté la valeur prise par  $X$ .

### 2.1.3 Entropie en statistique et en machine learning

La statistique et plus généralement l'apprentissage statistique reprend la formulation de l'entropie de Shannon et l'utilise comme une mesure d'hétérogénéité. Lors de la construction d'un arbre de décision (section 2.2.1.2), l'entropie est utilisée pour mesurer l'impureté ou l'hétérogénéité des feuilles et pour évaluer le gain de pureté associé à un découpage donné. L'entropie est d'autant plus élevée qu'une feuille est composée d'un mélange de classes. Dans le cas de la classification binaire, l'évolution de l'entropie d'une population hétérogène formée de deux sous-groupes, « positif » et « négatif », est présentée sur la figure 2.3.

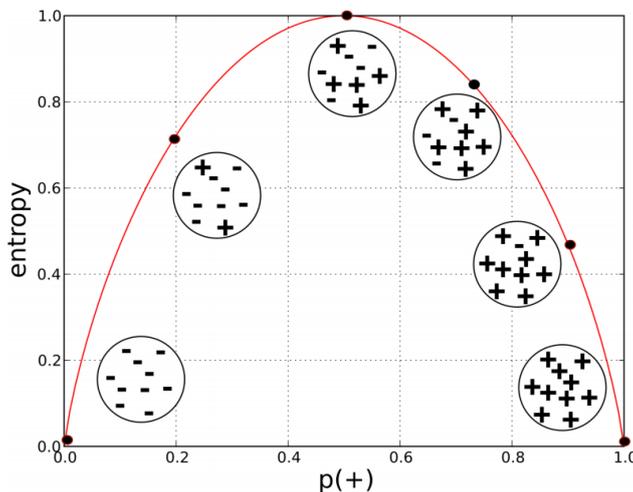


FIGURE 2.3 – Évolution de l'entropie en fonction de la proportion de « positifs » dans une population constituée de deux groupes (Provost et Fawcett, 2013, page 52).

L'entropie est également utilisée en statistique pour quantifier la dissimilarité entre deux distributions de probabilité ; elle est désignée par entropie croisée dans ce cas. Pour deux distributions de probabilités  $p$  et  $q$  définies sur un même espace probabilisé, l'entropie croisée est

donnée par :

$$H(p, q) = - \sum_x p(x) \log_2(q(x)) = H(p) + D_{KL}(p||q), \text{ avec } D_{KL}(p||q) := - \sum_x p(x) \log_2 \left( \frac{p(x)}{q(x)} \right)$$

$D_{KL}$  est la divergence de Kulback-Leibler. L'entropie croisée est utilisée en apprentissage statistique comme une fonction objective à minimiser afin que la distribution de la réponse prédite soit proche de sa distribution observée.

### 2.1.4 Limite de l'entropie comme mesure d'hétérogénéité : *coarse graining*

Il ressort des définitions et des emplois de l'entropie que cette notion est bien associée avec une mesure de désordre, d'informations non observées ou encore d'hétérogénéité. Mais elle n'en n'est pas pour autant une notion dépourvue de limites. Le calcul de l'entropie impose de retenir une granularité d'observation du système qui influence la valeur finale par l'entropie. Dans le cas du gaz de Boltzmann, pour connaître la distribution des molécules de gaz dans un volume  $V$  (disons un cube), un découpage de celui-ci en micro-cubes est nécessaire. Au sein de chaque micro-cube, les molécules qui s'y trouvent sont dénombrées. Plus il y a de micro-cubes occupés par les molécules plus le gaz est dispersé au sein du cube. Toutefois, la granularité du micro-cube est difficile à définir précisément. Cette nécessité de définir l'unité élémentaire pour réaliser des mesures sur un système est connu en physique sous le nom d'agrainissement ou *coarse graining*. Le prix Nobel de physique Penrose (2007) à la page 670 de son ouvrage *À la découverte des lois de l'univers*, rappelle cette limite de l'entropie en ces termes : « [...] pour que la formule  $S = k * \log(V)$  ci-dessus incarne quelque chose de précis pour la physique, encore faudrait-il que nous ayons une prescription claire et nette pour définir la granulation que nos "boîtes" sont censées représenter. Cette division en boîtes présente indubitablement un caractère arbitraire. La définition semble dépendre du degré de précision choisi pour étudier le système. [...] De plus, la position exacte de la frontière entre deux boîtes est elle aussi très arbitraire ». Il conclut alors que l'entropie « n'est pas une notion absolue des théories physiques contemporaines, mais une notion très commode ».

Cette limite de l'entropie bien connue des physiciens, se répercute lors de son emploi dans ce mémoire pour évaluer le niveau de mutualisation d'un portefeuille assurantiel.

### 2.1.5 Comparaison entre entropie et variance

Le chapitre 1 a posé les intuitions nécessaires pour comprendre et mesurer la démutualisation sur un portefeuille. Il en est ressorti en particulier, qu'une dispersion importante des primes pures correspond à un transfert du risque de structure et *in fine* à une démutualisation. S'il est immédiat de penser à la variance pour quantifier cette dispersion, elle n'en est pas l'unique

mesure. L'entropie s'est également imposée être une mesure commode pour mesurer le désordre au sein d'un système. En considérant un portefeuille comme un système, l'entropie pourrait permettre de mesurer l'hétérogénéité des primes pures ou des OL qui est assimilable au désordre. Soit  $X$  une variable aléatoire discrète pouvant prendre les valeurs réelles  $x_1, x_2, \dots, x_n$  avec les probabilités  $p_1, p_2, \dots, p_n$ . La variance et l'entropie de cette variable sont données par :

$$\text{Variance : } \mathbb{V}[X] = \sum_{i=1}^n p_i (x_i - \bar{x})^2, \text{ avec } \bar{x} = \frac{1}{n} \sum_{i=1}^n p_i x_i$$

$$\text{Entropie : } H(X) = - \sum_{i=1}^n p_i \log_2(p_i)$$

Les différences suivantes entre ces deux mesures d'hétérogénéité interpellent dans le cadre de la quantification de la mutualisation :

### La variance est sensible à un facteur d'échelle

À l'inverse de la variance, l'entropie ne dépend pas des valeurs  $x_1, x_2, \dots, x_n$  prises par la variable, mais seulement des probabilités  $p_1, p_2, \dots, p_n$ . De plus, en multipliant la variable  $X$  par un scalaire  $\alpha$ , la variance est multipliée par  $\alpha^2$ . Concernant la mesure de dispersion des primes, un assureur qui multiplie toutes ses primes pures par un même facteur ne verrait pas son niveau de démutualisation varier si la mesure est basée sur l'entropie, alors que celui-ci augmenterait pour un indice basé sur la variance. Pour un indicateur de démutualisation, il est souhaitable que consécutivement à cette pratique, sa valeur reste inchangée puisqu'il n'y a pas eu de traitement discriminant d'une classe d'assurés sur les autres. Inversement, pour mesurer l'équité sur le portefeuille, il est souhaitable que la valeur de la métrique utilisée change dans ce scénario.

### L'entropie privilégie une dispersion étalée uniformément

L'entropie est maximale lorsque toutes les probabilités  $p_1, p_2, \dots, p_n$  sont égales à  $\frac{1}{n}$ , donc en situation d'équiprobabilité. Le cas de variance maximale est observé lorsque  $p_1 = \frac{1}{2}$  et  $p_n = \frac{1}{2}$ , et tous les autres  $p_i$  pour  $i = 1 \dots n - 1$  sont nuls. Autrement dit, lorsque seules les valeurs extrêmes sont observées (voir figure 2.4 ci-dessous). Des démonstrations détaillées des cas de variance et d'entropie maximale sont données en annexe A.

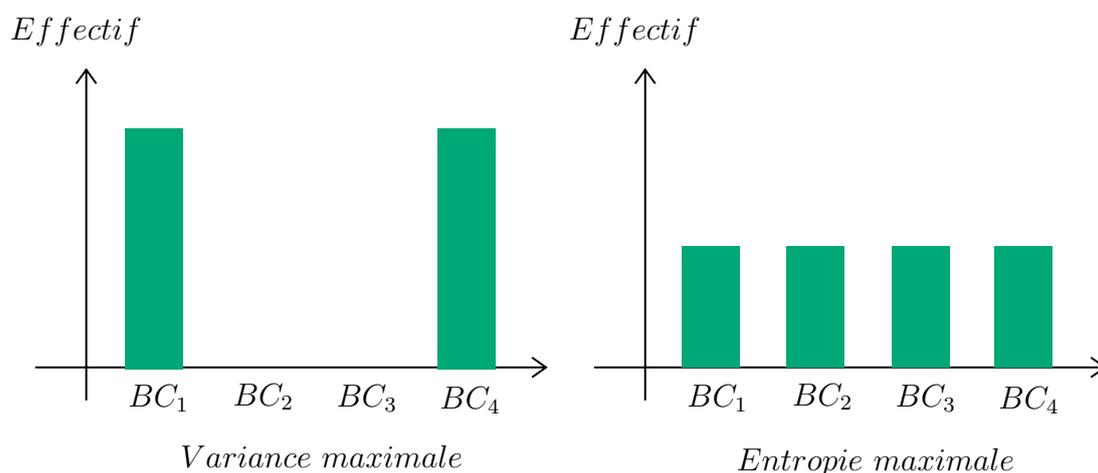


FIGURE 2.4 – Comparaison variance entropie comme mesure de dispersion.

Appliqué à la quantification de la mutualisation, un portefeuille sera d'autant plus mutualisé au sens de la variance que les burning costs du meilleur profil (min des burning cost) et du pire profil de risque (max des burning cost) sont distants l'un de l'autre et qu'entre ces deux extrêmes il n'y a pas ou alors très peu d'assurés. Autrement dit, si l'assureur réussit à mettre ensemble dans les mêmes proportions les très bons et les très mauvais risques, son niveau de mutualisation évalué par la variance sera élevé. Cette configuration n'est pas observée en pratique et ne semble pas refléter correctement la mutualisation telle que comprise par la communauté actuarielle. Au sens de l'entropie, un portefeuille sera d'autant plus mutualisé que les profils de risque élémentaires en portefeuille sont nombreux et équiprobables. C'est-à-dire un assureur qui réussit à mettre ensemble un grand nombre de profils différents et les rend solidaires les uns aux autres, mutualise beaucoup au sens d'une métrique fondée sur l'entropie. Cette configuration semble mieux s'aligner avec l'appréhension actuarielle de la mutualisation.

## 2.2 Indices de mutualisation simples basés sur la prime et ses composantes

La prime et particulièrement la prime pure sont des outils dont dispose l'assureur pour discriminer les assurés. Le calcul de la prime pure fait appel à des modèles statistiques et plus récemment à des algorithmes de machine learning qu'il convient de présenter dans ce cadre.

### 2.2.1 Construction de la prime pure : GLM et LightGBM

Le cadre de calcul de la prime pure en tarification actuarielle est le suivant : l'assureur dispose des bases historiques de sinistres (nombre de sinistres et montants individuels des sinistres) accompagnés des facteurs de risques (caractéristiques ou variables tarifaires) observés sur un ensemble d'assurés. Les notations suivantes sont adoptées :

- $\mathbf{X} = (X_1, X_2, \dots, X_p)$  : le vecteur aléatoire des facteurs de risque dont l'observation sur l'assuré  $i$  est noté  $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$  ;
- $N$  : la variable aléatoire entière qui donne pour chaque assuré le nombre de sinistres observés sur un exercice ou sur la période d'intérêt ;
- $S^{(k)}$  : variable aléatoire strictement positive qui indique la sévérité ou le montant du  $k$ -ième sinistre indemnisé pour un assuré donné ;
- $S = \sum_{k=1}^N S^{(k)}$  : la variable aléatoire positive qui donne la charge totale de sinistres d'un assuré, avec la convention  $S = 0$  si aucun sinistre n'est observé.

Le but de l'assureur est d'approcher  $S$  à partir des variables connues *a priori*  $\mathbf{X}$ . Plus précisément, l'assureur cherche à construire une fonction  $\Phi$  définie par :  $\Phi(\mathbf{x}) := \mathbb{E}[S \mid \mathbf{x}]$ .  $\Phi$  est la fonction de tarification qui donne pour chaque assuré (défini par ses caractéristiques), le montant de prime pure. Ce montant reflète la charge totale de sinistres attendue sur l'assuré. Dans la suite de ce mémoire, une telle fonction sera désignée par le *tarifneur*.

Pour calibrer un tarifneur à partir des données historiques, deux approches sont couramment utilisées en actuariat : l'approche fréquence/sévérité et l'approche Tweedie.

#### Approche fréquence/sévérité ou Tweedie pour la tarification

Dans le modèle de fréquence/sévérité, l'actuaire part d'une hypothèse d'indépendance entre les variables  $N$  et  $S^{(k)}$  pour tout  $k$ . Le tarifneur devient alors :

$$\Phi(\mathbf{x}) = \mathbb{E}[S \mid \mathbf{x}] = \underbrace{\mathbb{E}[N \mid \mathbf{x}]}_{\Phi_{freq}(\mathbf{x})} \underbrace{\mathbb{E}[S^{(1)} \mid x]}_{\Phi_{sev}(\mathbf{x})} = \Phi_{freq}(\mathbf{x})\Phi_{sev}(\mathbf{x})$$

Construire le tarifneur  $\Phi$  revient à calibrer à partir de modèles paramétriques ou non la fonction de fréquence moyenne  $\Phi_{freq}$  et la fonction de coût moyen  $\Phi_{sev}$ . Ensuite ces fonctions sont

multipliées. Il n'y a pas de restriction particulière sur les familles de lois qui peuvent être imposées sur la fréquence et sur la sévérité.

L'approche Tweedie est un cas particulier de l'approche fréquence/sévérité où la charge de sinistre totale  $S$  d'un assuré est supposée suivre une loi composée poisson-gamma. Soit le nombre de sinistres  $N \sim \mathcal{P}(\lambda)$  et le coût d'un sinistre  $S^{(k)} \sim \Gamma(\alpha, \beta)$ , les premiers moments de cette loi sont donnés par :

$$\begin{aligned} \mathbb{E}[S] &= \lambda \frac{\alpha}{\beta} =: \mu \\ \text{Var}[S] &= \lambda \frac{\alpha(1 + \alpha)}{\beta^2} =: \sigma^2 \mu^p \end{aligned}$$

La reparamétrisation à partir de  $\mu, \sigma^2$  et  $p$  permet de représenter cette loi comme appartenant à la famille des lois exponentielles où  $p$  est le paramètre de puissance. Dès lors, il est possible à partir d'un échantillon observé de déterminer les paramètres qui maximisent la vraisemblance de ce modèle, de construire le tarifeur  $\Phi$  directement.

Les primes pures sont calculées dans ce mémoire à partir de l'approche Tweedie puisque cette approche permet de limiter les accumulations des erreurs des modèles et assure de disposer clairement de la loi des profils de risque pour chaque assuré directement. Les formalismes des modèles GLM et LightGBM utilisés au chapitre 4 pour construire le tarifeur  $\Phi$  sont présentés ici.

### 2.2.1.1 Un modèle usuel en actuariat : GLM

Le modèle linéaire généralisé ou *Generalized Linear Model*, en anglais et en abrégé GLM est le modèle de calcul de prime pure le plus utilisé en pratique pour sa transparence et sa simplicité d'interprétation. Le but est de prédire une variable aléatoire  $Y$  (le coût total de sinistre, ou la fréquence, ou la sévérité d'un sinistre par exemple) à partir d'un vecteur de caractéristiques  $\mathbf{X}$  supposé exogène.

Le modèle GLM à estimer s'écrit :

$$\begin{aligned} h(\mathbb{E}[Y_i | \mathbf{X}_i = \mathbf{x}]) &= \eta_i(\mathbf{x}) \\ \eta_i(\mathbf{x}) &= \beta_0 + x_{i,1}\beta_1 + \dots + x_{i,p}\beta_p =: \langle \mathbf{x}, \boldsymbol{\beta} \rangle \\ \mu_i(\mathbf{x}) &= h^{-1}(\eta_i(\mathbf{x})) \end{aligned}$$

Le GLM nécessite des hypothèses explicites sur la réponse, la forme de la relation entre les variables explicatives et la réponse. Ces hypothèses ainsi que les procédures d'estimation du modèle sont examinées.

#### i. La composante aléatoire : la famille exponentielle

Dans un GLM la réponse  $Y$  est modélisée comme une variable aléatoire dont la distribution de probabilité appartient à la famille exponentielle. Cette famille fournit un cadre paramétrique

uniformisé et une formalisation générale pour l'estimation des GLM pour un grand nombre de distributions usuelles telles que : la distribution de poisson, la binomiale, la binomiale négative (utile pour modéliser la fréquence), la distribution gamma, exponentielle, log-normale (utile pour modéliser la sévérité)... Cette famille comprend également la distribution de Tweedie. La sélection de la loi se fait *a priori* au regard du support de  $Y$ , de sa distribution empirique et les spécificités du phénomène étudié. Le caractère aléatoire de la réponse pour une police d'assurance  $i$  se formalise dans un cas simple à deux paramètres comme suit :

$$Y_i \sim ED(\mu_i, \sigma^2)$$

Le paramètre  $\mu_i$  est la moyenne de la distribution de  $Y_i \mid \mathbf{X}_i$ . L'estimation de ce paramètre pour chaque police est donnée par la prédiction du modèle.  $\sigma^2$  est le paramètre de dispersion (lié à la variance mais pas nécessairement la variance). D'autres paramètres existent dans le cas des lois plus avancées comme la Tweedie.

Pour modéliser la charge totale par police d'assurance ( $S_i$ ), la distribution de Tweedie est particulièrement adaptée puisqu'elle comporte une masse de probabilité en zéro permettant de capter les polices non sinistrées et la partie de sa distribution continue sur  $\mathbb{R}_+$  est nécessaire pour modéliser les montants des sinistres lorsqu'ils surviennent comme illustré sur la figure 2.5. Lorsque le paramètre de puissance  $p \rightarrow 1$ , la loi s'approche d'une poisson et lorsque  $p \rightarrow 2$ , elle s'approche d'une loi gamma.

La charge de sinistre totale par assuré est modélisée par une loi de Tweedie :

$$S_i \sim Tw(\mu_i, \sigma^2, p) \text{ dont la densité est } f_{S_i}(y) = \sum_{n=1}^{\infty} e^{-\lambda_i} \frac{\lambda_i^n}{n!} \frac{\beta_i^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha-1} e^{-y\beta_i} \text{ et } f_{S_i}(0) = e^{-\lambda_i}$$

où les paramètres  $\lambda_i, \beta_i$  et  $\alpha_i$  sont reliés aux paramètres de la Tweedie

$$\lambda_i = \frac{\mu_i^{2-p}}{\sigma^2(2-p)}, \quad \alpha = \frac{2-p}{p-1} \quad \beta_i = \frac{1}{\sigma^2(p-1)\mu_i^{p-1}} \text{ avec } 1 < p < 2.$$

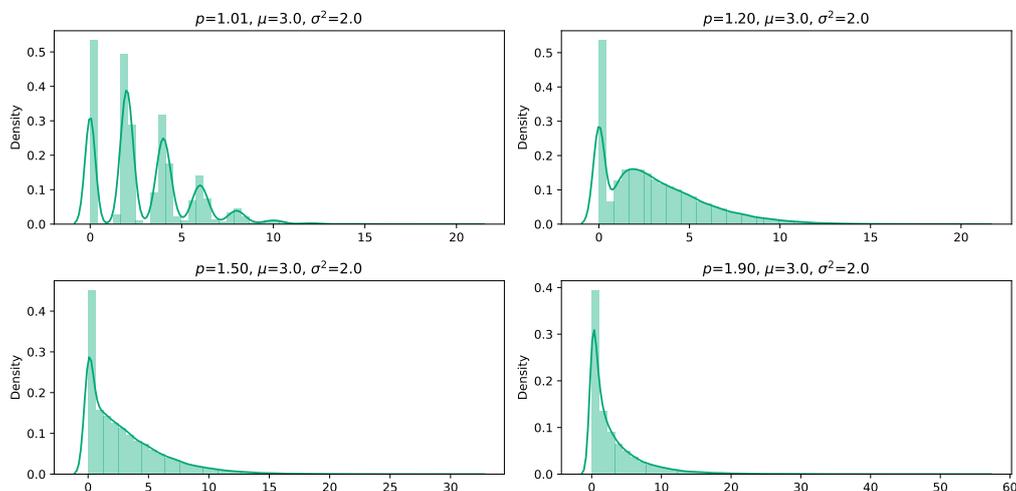


FIGURE 2.5 – Distributions Tweedie pour différentes valeurs du paramètre de puissance  $p$ .

**ii. La composante déterministe : le prédicteur linéaire et la fonction de lien**

Le paramètre  $\mu_i$  est supposé dépendre des caractéristiques  $x_i$  par la relation :

$$h(\mu_i) = \beta_0 + x_{i,1}\beta_1 + \dots + x_{i,p}\beta_p$$

Cette relation révèle qu'une transformation spécifiée de la valeur attendue  $\mu_i$ , est égale à une combinaison linéaire des caractéristiques avec les coefficients  $\beta_1, \beta_2, \dots, \beta_p$  augmentée d'une constante  $\beta_0$ . Ces paramètres consignés dans le vecteur  $\beta$  sont estimés lors de la calibration d'un GLM.  $h$  est la fonction de lien, à spécifier lors de l'estimation ; son inverse confère au prédicteur linéaire  $\eta_i$  plus de flexibilité pour approcher la valeur de  $\mu_i$  et mieux coller à la réalité. La seule contrainte sur la fonction de lien est son inversibilité. Pour la régression Tweedie, les fonctions de lien *log* et *puissance* sont généralement utilisées. Notre choix est porté sur le lien *log*, puisque son inverse *exp* génère la prédiction finale comme une multiplication des contributions de facteurs de risque.

TABLEAU 2.1 – Fonctions de lien usuelles pour la régression Tweedie.

fonction	$h(x)$	$h^{-1}(x)$	$\mu_i$
<b>logarithme</b>	$\ln(x)$	$\exp(x)$	$e^{\beta_0} \times e^{\beta_1 x_{i,1}} \times \dots \times e^{\beta_p x_{i,p}}$
<b>puissance</b>	$x^q$	$x^{1/q}$	$(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p})^{1/q}$

**iii. Estimation du modèle : maximisation de la vraisemblance**

L'objectif est d'inférer  $\beta$  à partir d'un ensemble de données,  $\mathcal{D}_n = \{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$ . L'estimateur de  $\beta$  est déterminé par maximisation de la vraisemblance du modèle. Sous l'hypothèse que  $y_i$  est distribué suivant une Tweedie, la log-vraisemblance est donnée par (Dutang, 2020) :

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^n \frac{y_i \frac{\mu_i^{1-p}}{1-p} - \frac{\mu_i^{2-p}}{2-p}}{\sigma^2} + \sum_{i=1}^n c(y_i, \sigma^2, p) \tag{2.1}$$

où  $c(y_i, \sigma^2, p)$  est un terme indépendant de  $\mu_i$

La dérivée partielle par rapport au  $k$ -ième paramètre vaut :

$$\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta_k} = \frac{1}{\sigma^2} \sum_{i=1}^n \left( y_i x_{i,k} h'(\eta_i) h(\eta_i)^{-p} - x_{i,k} h'(\eta_i) h(\eta_i)^{1-p} \right) = \frac{1}{\sigma^2} \sum_{i=1}^n \frac{x_{i,k}}{\mu_i^p} h'(\eta_i) (y_i - \mu_i)$$

L'objectif revient à trouver  $\hat{\beta}_{MLE}$  qui maximise la log-vraisemblance, c'est-à-dire  $\hat{\beta}_{MLE}$  solution de l'équation donnée par la condition du premier ordre  $\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta_k} = 0$ . L'optimisation numérique se fait à partir des algorithmes de type Newton-Raphson ou de descente du gradient.

**iv. Sur-apprentissage et régularisation : la méthode smurf pour généraliser le lasso**

En pratique pour construire le prédicteur de prime pure  $\Phi$ , l'assureur peut disposer d'un grand nombre de caractéristiques sur les polices d'assurance. Outre cette multitude de caractéristiques, la forte cardinalité des variables qualitatives et des variables continues discrétisées<sup>2</sup> augmente également la taille de  $\beta$ . Dans ce cas, le tarifeur obtenu est instable et ne se généralise pas correctement sur de nouvelles polices ; il s'agit du sur-apprentissage du modèle. La régularisation permet de faire face à ce problème. Elle contraint les paramètres à demeurer faibles (pénalisation ridge) voire nuls (pénalisation lasso). Cette technique est largement utilisée en statistique et en machine learning pour créer des prédicteurs *sparses*. Devriendt *et al.* (2021) introduisent d'autres formes de régularisation dans GLM qui permettent de regrouper les modalités d'une variable catégorielle ordinaire ou nominale ; la régularisation *smurf* (pour Sparse Multi-Type Regularized Feature Modeling). Ils réécrivent la fonction à maximiser sous forme de problème de minimisation en rajoutant un terme de pénalisation :

$$\mathcal{O}(\beta; D) = -\mathcal{L}(\beta; D) + \lambda \cdot \sum_{j=1}^p g_j(\beta_j)$$

$g_j(\cdot)$  est une fonction convexe de pénalisation du sous ensemble de paramètres  $\beta_j$  (pour une variable  $j \in \{1, 2, \dots, p\}$ ,  $\beta_j$  comporte les coefficients de toutes les modalités de la variable discrétisée  $\beta_i = (\beta_{j,1}, \beta_{j,2}, \dots, \beta_{j,p_j})$ ). La somme commence à 1 car la constante  $\beta_0$  n'est pas régularisée. Le paramètre  $\lambda$  contrôle l'importance de la pénalité. Une valeur élevée de  $\lambda$  augmente le poids de la régularisation dans l'objectif  $\mathcal{O}(\cdot)$  et augmente la sparsité du modèle estimé. Plusieurs formes de pénalisation sont introduites par les auteurs :

- *Lasso* :  $g_{\text{Lasso}}(\beta_j) = \sum_{i=1}^{p_j} w_{j,i} |\beta_{j,i}|$  où  $w_{j,i}$  est le poids de la pénalisation sur le coefficient associé à la  $i$ -ème modalité de la variable  $X_j$ . une valeur élevée de  $w_{j,i}$  encourage les coefficients à être nuls ;
- *Groupe Lasso* :  $g_{\text{grpLasso}}(\beta_j) = w_j \sqrt{\sum_{i=1}^{p_j} \beta_{j,i}^2}$ , en introduisant une pénalité de type norme  $L_2$  sur l'ensemble des coefficients de  $X_j$ , cette fonction encourage l'ensemble des coefficients  $\beta_{j,i}$  à demeurer faibles ou nuls ;
- *Fused Lasso* :  $g_{\text{fLasso}}(\beta_j) = \sum_{i=2}^{p_j} w_{j,i-1} |\beta_{j,i} - \beta_{j,i-1}|$ , particulièrement utile pour regrouper les modalités consécutives d'une variable quantitative ordinaire ou continue discrétisée. Le Fused Lasso introduit une pénalité  $L_1$  sur la différence entre deux coefficients consécutifs ;
- *Generalized Fused Lasso* :  $g_{\text{gfLasso}}(\beta_j) = \sum_{(i,l) \in \mathcal{G}} w_{j,il} |\beta_{j,i} - \beta_{j,l}|$ , les paires de modalités d'une variable à regrouper ne sont plus nécessairement consécutives, mais déterminées

2. La discrétisation des variables continues dans un GLM permet de prendre en compte les effets non-linéaires du facteur de risque sur la réponse, mais au prix de l'augmentation des paramètres à estimer. Elle confère donc plus de flexibilité au modèle.

par un graphe  $\mathcal{G}$ . Cette régularisation est adaptée pour regrouper les modalités de la même variable catégorielle nominale ;

- Combinaison des pénalités :  $g_{s,grp,gfLasso}(\beta_j) = \lambda_1 g_{Lasso}(\beta_j) + \lambda_2 g_{grpLasso}(\beta_j) + g_{gfLasso}(\beta_j)$ , il est également possible de combiner les formes de pénalisation précédentes pour une même variable  $X_j$ .

#### v. Notion d'offset en tarification actuarielle

Il est courant que les polices d'assurance ne soient pas couvertes pour une durée fixée sur un exercice ; celles-ci peuvent aller d'un jour à une année entière. Les contrats n'ont pas tous la même durée d'**exposition** notée  $e_i$ . L'exposition d'un contrat sur un exercice est calculée par :

$$e_i = \frac{\text{Nombre de jours où le contrat est à risque}}{365.25}$$

Il s'en suit que les prédictions des  $\mu_i$  produites par le modèle ne sont pas comparables. Il faut alors ramener les prédictions à la même « échelle » pour les rendre comparables. Au lieu d'estimer  $\mu_i$ , la quantité à estimer est plutôt  $\frac{\mu_i}{e_i}$ . L'équation du modèle avec une fonction de lien *log*, peut alors se réécrire sous la forme :

$$\log(\mu_i) = \beta_0 + x_{i,1}\beta_1 + \dots + x_{i,p}\beta_p + \log(e_i)$$

S'assurer que le modèle prédise bien un burning cost annuel par contrat revient alors à ajouter une variable explicative égale au logarithme de l'exposition dont le coefficient est fixé à 1. Cette variable est connue dans la littérature des GLM sous le nom de *offset*.

#### 2.2.1.2 Un exemple de modèle de Machine Learning : LightGBM

---

Avec la généralisation des techniques de machine learning, de plus en plus d'études sont réalisées pour étudier l'opportunité de mettre à profit ces modèles dans le contexte de l'actuariat et précisément celui de la tarification. Au panel des méthodes les plus utilisées en machine learning, se retrouve les méthodes dites de boosting. Il s'agit de modèles supervisés, non paramétriques, ensemblistes, dont la prédiction finale s'écrit sous forme d'une somme des prédictions de plusieurs modèles *faibles* construits de façon itérative. Lorsque ces modèles sont appliqués à la régression Tweedie pour la tarification actuarielle, la prédiction finale est obtenue comme le logarithme de la combinaison linéaire des outputs des modèles faibles généralement les arbres CART. Plus spécifiquement, la régression Tweedie à partir des algorithmes de boosting des arbres se formalise comme suit :

Le boosting au service de la régression Tweedie :

$$\begin{aligned}
 h(\mathbb{E}[Y_i | \mathbf{X}_i = \mathbf{x}_i]) &= F(\mathbf{x}_i) \\
 F(\mathbf{x}_i) &= \sum_{k=1}^K f_k(\mathbf{x}_i), f_k \in \mathcal{F}, k = 1, 2, \dots, K \\
 \mu_i(\mathbf{x}_i) &= h^{-1}(F(\mathbf{x}_i))
 \end{aligned} \tag{2.2}$$

avec  $\mathcal{F} := \{f(\mathbf{x}) = w_{q(x)}, q : \mathbb{R}^m \rightarrow \{0, 1, \dots, T\}, w \in \mathbb{R}^T\}$  l'espace des arbres CART, où  $q$  représente la structure de chaque arbre qui associe à chaque individu un index d'une feuille de l'arbre,  $T$  est le nombre de feuilles de l'arbre,  $w$  est le vecteur des prédictions de l'arbre.

Comme dans le cas du GLM, les différentes composantes de ce modèle sont présentées afin de justifier son utilisation pour le calcul de la prime pure.

### i. La réponse $Y_i$ et sa loi implicite

La réponse  $Y_i$  du modèle est déterminée en divisant la charge totale de sinistres  $S_i$  du contrat par sa durée d'exposition  $e_i$  pour obtenir le burning cost. De même que pour le GLM, conditionnellement aux caractéristiques, cette réponse est supposée suivre un loi de Tweedie dont le paramètre de moyenne  $\mu_i$  est lié aux caractéristiques  $\mathbf{x}_i$  par une fonction  $F$  qui n'est pas nécessairement une combinaison linéaire des variables explicatives mais une somme d'arbres de décision.  $F$  est donc *a priori* plus flexible que le prédicateur linéaire  $\eta$ . L'hypothèse sur la loi de  $Y_i$  permet de calculer la log-vraisemblance dont l'opposé servira de fonction objectif pour la détermination des paramètres du modèle.

### ii. L'espace des arbres CART

La brique élémentaire pour construire la fonction  $F$  dans les modèles de boosting est un arbre CART (*Classification And Regression Tree*). Un arbre CART cherche à partitionner l'espace des caractéristiques  $\mathbf{X}$  en groupes homogènes (représentés par des feuilles) au regard de la réponse  $y_i$ . La construction d'un arbre est un partitionnement itératif de l'espace engendré par les caractéristiques. À chaque étape, le couple (variable & seuil) qui sépare le mieux les individus en groupes homogènes est identifié. Les sous-populations obtenues sont appelées nœuds. Puis la même opération est réitérée sur chaque nouveau nœud obtenu jusqu'à ce que la séparation des individus ne soit plus possible (tous les individus d'un nœud ont les mêmes caractéristiques) ou, ne soit plus souhaitable (au vu du faible nombre d'individus dans la feuille et/ou du faible gain de pureté qui résulterait d'un découpage). Les nœuds terminaux sont les feuilles. La prédiction d'un arbre est alors obtenue comme la moyenne des réponses (respectivement la classe majoritaire) calculée à partir des données d'entraînement des individus appartenant à la même feuille que le nouvel individu à prédire pour une régression (respectivement pour une classification).

À l'étape  $t$ , le choix optimal de la variable  $X_{j(t)}$  ainsi que du seuil  $a_{j(t)}$  pour le découpage du

noeud  $\mathcal{N}_t$  en deux noeuds  $\mathcal{N}_{t,<}$  et  $\mathcal{N}_{t,>}$  se base sur la décroissance d'une fonction d'impureté  $I(\cdot)$  qui est la variance de la réponse du noeud (resp. l'entropie ou l'indice de Gini) dans le cas d'un arbre de régression (resp. de classification). Plus spécifiquement, le couple  $(X_{j(t)}, a_{j(t)})$  découle de la minimisation suivante :

$$(j(t), a_{j(t)}) = \arg \max_{\substack{j \in \{1, 2, \dots, p\} \\ a_j \in \Omega(X_j)}} \left\{ I(\mathcal{N}_t) - \left( \frac{|\mathcal{N}_{t,<}|}{|\mathcal{N}_t|} I(\mathcal{N}_{t,<}) + \frac{|\mathcal{N}_{t,>}|}{|\mathcal{N}_t|} I(\mathcal{N}_{t,>}) \right) \right\}$$

où  $|\cdot|$  désigne le cardinal d'un noeud et  $I(\cdot)$  est choisi à partir du tableau 2.2

TABLEAU 2.2 – Principales mesures d'impureté utilisées dans les arbres CART.

Critère	Formule	Utilisation
Variance	$\frac{1}{ \mathcal{M} } \sum_{i=1}^{ \mathcal{M} } (y_i - \bar{y})^2$	arbre de régression
Entropie	$p_{<} \log_2(p_{<}) + p_{>} \log_2(p_{>})^*$	arbre de classification
Indice de Gini	$1 - p_{<}^2 - p_{>}^2$	arbre de classification

\* avec  $p_{<} = \frac{|\mathcal{N}_{<}|}{|\mathcal{M}|}$  et  $p_{>} = 1 - p_{<}$

La structure des arbres leur permet de capter les liens non-linéaires et les interactions entre les variables. D'autres avantages des arbres sont leur interprétabilité et leur visualisation qui découlent directement de leur construction. Toutefois, cette méthode manque de robustesse. En effet, un léger changement dans l'échantillon d'apprentissage peut conduire à un arbre final très différent puisque chaque découpage est conditionnel aux noeuds issus du découpage précédent. Pour stabiliser les arbres, les méthodes ensemblistes comme le gradient boosting combinent plusieurs arbres CART pour produire leur prédiction finale.

### iii. Gradient Boosting des arbres CART

La méthode du gradient boosting des arbres consiste en la construction itérative d'arbres, de sorte que l'arbre courant à construire corrige les erreurs observées sur l'arbre précédent (principe du boosting), et l'optimisation itérée pour réduire l'erreur est réalisée par descente de gradient. Les arbres utilisés dans le processus sont qualifiés de modèles « faibles » en raison de leur faible profondeur. Dans le modèle 2.2, il est possible de réécrire la fonction  $F(\cdot)$  en introduisant une paramétrisation des arbres faibles comme suit :

$$F_K(\mathbf{x}) := \sum_{k=1}^K \gamma_k f_k(\mathbf{x}; \boldsymbol{\theta}_k) = \sum_{k=1}^{K-1} \gamma_k f_k(\mathbf{x}; \boldsymbol{\theta}_k) + \gamma_K f_K(\mathbf{x}; \boldsymbol{\theta}_K) \tag{2.3}$$

$$= F_{K-1}(\mathbf{x}) + \gamma_K f_K(\mathbf{x}; \boldsymbol{\theta}_K)$$

où  $\gamma_k, k = 1, 2, \dots, K$  sont des paramètres de dilatation qui guident l'importance de chacun

des arbres dans la prédiction de  $\log(\mu)$ ;  $\boldsymbol{\theta}_k = \left( (j_{(1)}^k, a_{j_{(1)}}^k), (j_{(2)}^k, a_{j_{(2)}}^k), \dots, (j_{(T)}^k, a_{j_{(T)}}^k) \right)$  est le vecteur des paramètres (variables et seuils de décision) à déterminer lors de la construction du  $k$ -ième arbre CART. Estimer le modèle de régression Tweedie par le boosting revient à déterminer les paramètres  $\{\boldsymbol{\Gamma} = (\gamma_1, \gamma_2, \dots, \gamma_K), \boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K)\}$ .

Le cadre final de l'estimation d'une régression Tweedie par un modèle de boosting est le suivant : à partir d'un échantillon  $\mathcal{D}_n = \{(s_i, \mathbf{x}_i, e_i) \text{ pour } i = 1, 2, \dots, n\}$ , l'objectif est d'estimer la fonction  $F^*(\cdot)$  (plus précisément ses paramètres  $\gamma^*$  et  $\Theta^*$ ) qui minimise l'opposé de la log-vraisemblance du modèle 2.2.

$$\begin{aligned}
 F^*(\mathbf{x}) &= \arg \min_F \left\{ -\ell(F(\cdot), \sigma^2, p \mid \mathcal{D}_n) \right\} \\
 &= \arg \max_F \sum_{i=1}^n d(y_i, F(\mathbf{x}_i) \mid p) \\
 \text{où } \ell(F(\cdot), \sigma^2, p \mid \mathcal{D}_n) &= \sum_{i=1}^n \frac{e_i}{\sigma^2} \left( y_i \frac{\mu_i^{1-p}}{1-p} - \frac{\mu_i^{2-p}}{2-p} \right) + \sum_{i=1}^n c(y_i, \sigma^2, p) \text{ avec } y_i = \frac{s_i}{e_i} \\
 d(y_i, F(\mathbf{x}_i) \mid p) &= e_i \left\{ -\frac{y_i \exp[(1-p)F(\mathbf{x}_i)]}{1-p} + \frac{\exp[(2-p)F(\mathbf{x}_i)]}{2-p} \right\}
 \end{aligned}$$

La forme générale du problème à optimiser est semblable à celle de la régression GLM Tweedie exception faite du prédicteur linéaire qui est remplacé par  $F(\cdot)$ . La décomposition de  $F(\cdot)$  introduite par l'équation 2.3 suggère de déterminer l'estimation de  $(\boldsymbol{\Gamma}^*, \boldsymbol{\Theta}^*)$  de façon itérative telle qu'implémentée par l'algorithme *Gradient Tree Boosting* détaillé ci-après.

---

**Algorithm 1** Gradient Tree Boosting pour la régression Tweedie.

---

1. Initialisation :  $\hat{F}_0(\mathbf{x}) := \arg \min_{\eta \in \mathbf{R}} \sum_{i=1}^n d(y_i, \eta \mid p) = \log \left( \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \right)$  ▷ Initialisation du prédicteur par l'estimateur du maximum de vraisemblance du modèle sans les covariables
  2. Pour  $k = 1, \dots, K$  répéter les étapes 2(a) à 2(d)
    2. (a). Calcul du pseudo-résidu  $r_{i,k} := - \left[ \frac{\partial d(y_i, F(\mathbf{x}_i) \mid p)}{\partial F(\mathbf{x}_i)} \right]_{F=\hat{F}_{k-1}}$  pour  $i = 1, \dots, n$   
 par  $r_{i,k} = e_i \left\{ -y_i \exp \left[ (1-p)\hat{F}_{k-1}(\mathbf{x}_i) \right] + \exp \left[ (2-p)\hat{F}_{k-1}(\mathbf{x}_i) \right] \right\}$
    2. (b). Estimation  $\hat{\boldsymbol{\theta}}_k$  du modèle faible  $f_k(\mathbf{x}, \boldsymbol{\theta}_k)$  en calibrant un arbre CART pour prédire les  $r_{i,k}$  à partir de :  $\{(r_{i,k}, \mathbf{x}_i), i = 1, \dots, n\}$
    2. (c). Estimation de  $\hat{\gamma}$  comme solution du programme :  
 $\hat{\gamma}_k = \arg \min_{\gamma \in \mathbf{R}} \sum_{i=1}^n d(y_i, \hat{F}_{k-1}(\mathbf{x}_i) + \gamma \mid p)$
    2. (d). Mise à jour de  $\hat{F}_k(\mathbf{x}) = \hat{F}_{k-1}(\mathbf{x}) + \gamma_k f_k(\mathbf{x}, \hat{\boldsymbol{\theta}}_k)$
  3. Retourner  $\hat{F}_K(\mathbf{x})$  comme le prédicteur final.
- 

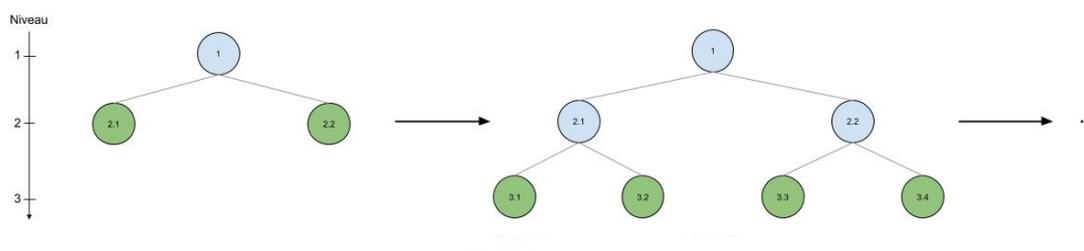
**iv. Spécificités de l'algorithme LightGBM**

Le LightGBM est un algorithme de machine learning qui appartient à la famille des méthodes

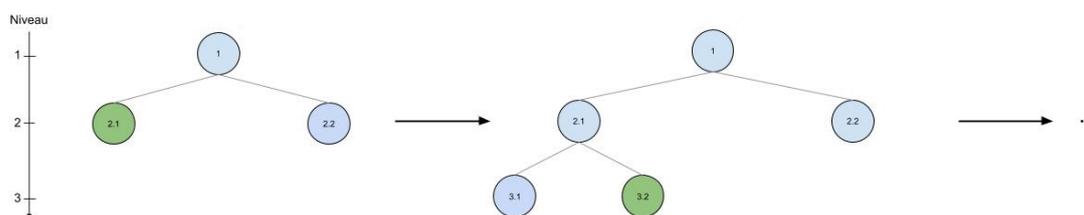
de boosting par gradient des arbres de décision. Ses auteurs Ke *et al.* (2017), proposent une architecture qui lui confère une efficacité et une vitesse d'exécution pouvant être jusqu'à 20 fois supérieures à celles du XGBoost (Chen et Guestrin, 2016), tout en gardant des performances prédictives comparables.

Les arbres dans les algorithmes de Gradient Boosting sont majoritairement construits en divisant les nœuds par étage<sup>3</sup>, c'est-à-dire que d'une itération à l'autre, l'ensemble des nœuds courants sont découpés pour produire 2 fois plus de nœuds à l'étage inférieur (cas de classification binaire et de la régression). Cette architecture est connue dans la littérature sous le nom de « *level-wise tree growth* » pour croissance de l'arbre par niveau. Sur la figure ci-contre, à la deuxième itération, les nœuds 2.1 et 2.2 sont sélectionnés et découpés en 2 nœuds chacun. À l'étape suivante, l'ensemble des nœuds de l'étage 3 à savoir 3.1, 3.2, 3.3 et 3.4 sont découpés en sous nœuds et ainsi de suite.

LightGBM construit ses arbres par nœuds, par opposition à d'autres algorithmes de gradient boosting qui s'étendent par étage. À chaque itération, le nœud terminal avec la perte maximale (l'impureté la plus élevée) est retenu et découpé pour le développement de l'arbre ; la terminologie retrouvée dans la littérature est « *leaf-wise tree growth* ». En guise d'illustration, à l'étape 2 seul le nœud 2.1 est retenu et découpé, puis à l'étape suivante le nœud 3.2 sera découpé. Mais il convient de souligner qu'à l'étape 3 ou ultérieurement, la feuille 2.2 peut faire l'objet d'un découpage.



(a) Croissance de l'arbre par étage (XGboost)



(b) Croissance de l'arbre par nœud (LightGBM)

FIGURE 2.6 – Comparaison des architectures de croissance des arbres CART dans les algorithmes de Gradient Boosting.

Le LightGBM permet également de prendre les variables catégorielles sous forme textuelle sans pré-traitement en créant notamment des variables indicatrices ; ce qui a pour conséquence

3. un étage regroupe l'ensemble des nœuds ayant la même profondeur

de réduire la qualité prédictive du modèle. De plus, conserver les variables catégorielles sous leur forme naturelle confère une interprétation directe de l'arbre.

### 2.2.1.3 Critères d'évaluation d'un modèle de calcul de prime pure

---

Une fois un modèle ajusté, il est convenable de disposer d'informations sur sa qualité d'adéquation et d'apprécier sa généralisation. Pour y parvenir, il convient de retenir les indicateurs pertinents pour évaluer et comparer les modèles considérés. Rappelons que les réponses (notées  $\hat{\mu}_i$ ) prédites par les différents modèles sont les burning costs annuels ( $SinistreTotal/Exposition$ ) dont la valeur observée est  $y_i = s_i/e_i$ . Les métriques utilisées dans cette étude sont héritées à la fois de l'actuariat (la déviance, la part de variance expliquée, l'équilibre financier du modèle) et du machine learning (RSME, MAE). La RSME et la MAE appréhendent la qualité du modèle en mesurant les écarts entre le  $\hat{\mu}_i$  et  $y_i$  au niveau individuel avant de les agréer au niveau global. Les métriques actuarielles quant-à-elles considèrent le portefeuille comme un pool sans passer au niveau individuel. Les métriques de machine learning implicitement favorisent alors la personnalisation des risques alors que celles actuarielles sont plus adaptées sous le prisme de la mutualisation. Les formules et les interprétations de ces métriques sont les suivantes :

- **Ratio\_Est\_Obs** =  $\frac{\sum_{i=1}^n e_i \hat{\mu}_i}{\sum_{i=1}^n e_i y_i} - 1$  : cet indicateur mesure la capacité du modèle à estimer correctement la charge totale de sinistre, c'est à dire l'équilibre financier au niveau du portefeuille. Un modèle qui prédit la charge totale de sinistres à plus de 2 % est considéré comme invalide ;
- **Déviance** =  $D_{\hat{\beta}} = -2 \left( \ln \mathcal{L}_{\hat{\beta}} - \ln \mathcal{L}_s \right) = 2 \sum_{i=1}^n \left( \frac{\max(y_i, 0)^{2-p}}{(1-p)(2-p)} - \frac{y_i \hat{\mu}_i^{1-p}}{1-p} + \frac{\hat{\mu}_i^{2-p}}{2-p} \right)$  : Cet indicateur mesure la qualité d'ajustement du modèle estimé aux données. Il est calculé comme l'écart entre la log-vraisemblance du modèle estimé ( $\ln \mathcal{L}_{\hat{\beta}}$ ) et la log-vraisemblance du modèle saturé ( $\ln \mathcal{L}_s$ ), c'est-à-dire un modèle avec autant de paramètres que d'observations ; de sorte que les données sont ajustées exactement. Plus cet indicateur est faible (en restant positif) plus le modèle s'ajuste bien aux données.
- **Déviance expliquée** =  $1 - \frac{D_{\hat{\beta}}}{D_{\hat{\beta}_0}}$  : Cet indicateur est une généralisation du coefficient de détermination  $R^2$ , il mesure la part de déviance expliquée par le modèle.  $D_{\hat{\beta}_0}$  est la déviance du modèle constant, sans variable explicative. Si l'ajustement est parfait la part de déviance expliquée est égale à 1 ; inversement elle vaut zéro si la prédiction est constante.
- **RMSE** =  $\sqrt{\frac{\sum_{i=1}^n (\hat{\mu}_i - y_i)^2}{n}}$  : c'est un indicateur de qualité global qui agrège les adéquations individuelles. Il quantifie à quel point une prédiction individuelle  $\hat{\mu}_i$  est proche de sa valeur observée  $y_i$  en moyenne ;
- **MAE** =  $\frac{1}{n} \sum_{i=1}^n |\hat{\mu}_i - y_i|$  : la philosophie de cette métrique est proche de celle de la RSME,



## 2.2.2 Indices de mutualisation calculés sur la prime et ses composantes

Les notations et définitions suivantes (et celles introduites à la section 2.2.1) sont nécessaires pour la formalisation des indicateurs de démutualisation proposés.

### Notations relatives aux facteurs de risque

- $\Omega$  : Ensemble des facteurs de risque qui influent sur la survenance et la sévérité du risque étudié, en général cet ensemble n'est pas connu de l'assureur ;
- $\mathbf{X}_\Omega \subseteq \Omega$  : Meilleure approximation de  $\Omega$  observée par l'assureur. Il s'agit de l'ensemble des variables tarifaires dont dispose l'assureur ;
- $\mathbf{X}_c^{Ass} \subseteq \mathbf{X}_\Omega$  : En pratique, l'assureur n'utilise pas explicitement toutes les variables observées  $\mathbf{X}_\Omega$  dans sa grille de prime pure, du fait des contraintes réglementaires, des contraintes technologiques, de la maille de segmentation choisie, etc. Il utilise plutôt un sous-ensemble  $\mathbf{X}_c^{Ass}$  et procède à une agrégation des autres variables. *Ass* désigne un assureur, puisque le choix de  $\mathbf{X}_c^{Ass}$  dépend de la stratégie commerciale de chaque assureur. Par contre,  $\mathbf{X}_\Omega$  est supposé homogène entre assureurs, du moins sur les facteurs de risque les plus importants<sup>4</sup> ;
- $J_c^{Ass} = |\mathbf{X}_c^{Ass}|$  est le nombre de variables tarifaires utilisées par l'assureur *Ass*. De même  $J_{\mathbf{X}_\Omega} = |\mathbf{X}_\Omega|$ , où  $|\cdot|$  représente le cardinal.

### Notations relatives au portefeuille

Un portefeuille est défini comme un ensemble d'assurés couverts pour un risque donné sur le même exercice, observés sur un ensemble de facteurs de risques communs et qui payent des montants de prime commerciale  $P_{c,i}$ . Il est formalisé par  $\mathcal{P}f^{Ass} = \{(\mathbf{x}_{c,i}^{Ass}, P_{c,i}^{Ass}), \text{ pour } i = 1, \dots, N^{Ass}\}$ . Cette définition est étendue au cas où l'assureur utilise explicitement  $\mathbf{X}_\Omega$  dans sa grille tarifaire :  $\mathcal{P}f_\Omega^{Ass} = \{(\mathbf{X}_{\Omega_i}^{Ass}, P_{c,i}^{Ass}), \text{ pour } i = 1, \dots, N^{Ass}\}$ .  $N^{Ass}$  est la taille du portefeuille, le nombre de polices couvertes sur l'exercice.  $\mathcal{P}$  désigne l'espace de tous les portefeuilles pour un risque donné.

### Notations relatives au tarifeur et aux primes

Le tarifeur  $\Phi$  a été introduit comme une fonction définie de l'espace des caractéristiques vers  $\mathbb{R}_+$ , il peut être étendu à un portefeuille en l'appliquant contrat par contrat :

$$\begin{aligned} \Phi: \mathcal{P} &\rightarrow \mathbb{R}_+^N \\ \mathcal{P}f^{Ass} &\mapsto \Phi(\mathcal{P}f^{Ass}) = (BC_1, \dots, BC_{N^{Ass}}) \end{aligned}$$

$\mathbf{BC}^{Ass} = (BC_1, \dots, BC_{N^{Ass}})$  est le vecteur des primes techniques de l'assureur calculé par son tarifeur commercial. Selon le choix des variables, leurs pré-traitements et la modélisation mise

---

4. Cette hypothèse *a priori* restrictive est nécessaire pour la comparaison du niveau de démutualisation entre assureurs. Elle peut être allégée, en supposant que la segmentation la plus fine pouvant être atteinte par les assureurs est relativement comparable.

en oeuvre, le tarifeur change d'un assureur à l'autre, il est par conséquent noté  $\Phi^{Ass}$ .

**Remarque : Prime technique et prime pure individualisée**

Le vecteur des primes pures (techniques)  $BC^{Ass}$  calculé à partir du tarifeur de l'assureur est en général différent du vecteur des primes pures (individuelles)  $BC^\Omega$  qui serait obtenu si le tarifeur utilisait l'ensemble des variables de  $X_\Omega$  comme inputs.

Du vecteur des primes commerciales et du vecteur des burning cost, le vecteur des OL est déduit :

$$OL^{Ass} = P_c^{Ass} - BC^\Omega$$

$$= \underbrace{\text{Chargements}}_{Commercial} + \underbrace{BC^{Ass} - BC^\Omega}_{technique}$$

Le choix de cette formulation de l'OL (et non  $OL^{Ass} = P_c^{Ass} - BC^{Ass}$ ) permet de prendre en compte les deux leviers de mutualisation de l'assureur : le levier technique à travers la segmentation plus ou moins fine de la prime tarifée ; le levier commercial à travers les chargements assureurs différenciés d'un profil de risque à l'autre. Toutefois, il est nécessaire de modéliser le profil de risque  $BC^\Omega$  pour calculer l'OL

Exhaustivité des facteurs de risques

	$X_c^{Ass}$	$X_\Omega$	$\Omega$
Information disponible	Facteurs de risque que l'assureur inclut explicitement dans sa grille tarifaire	Ensemble des facteurs de risque observés par l'assureur : approximation de $\Omega$	Ensemble des caractéristiques de l'assuré, du contrat qui influent la sinistralité
Burning cost	$\Phi^{Ass}(X_c^{Ass}) = BC^{Ass}$ Prime technique tarifaire calculée sur les facteurs $X_c^{Ass}$	$\Phi^\Omega(X_\Omega) = BC^\Omega$ Profil de risque le plus fin connu calculé sur les facteurs $X_\Omega$	— Profil de risque individuel obtenu si $\Omega$ était observé
Ex. Facteurs de risques EG, auto	Formule : 3 couvertures Age véh. à la souscription : 3 modalités Type de véhicule : camping car/voiture Puissance : 3 modalités	Formule : couvertures détaillées Age véh. à la souscription : continue sur R+ Type de véhicule : camping car/voiture Puissance : entière et positive  Motorisation km à la souscription Marque de véhicule Modèle de véhicule	Formule : couvertures détaillées Age véh. à la souscription : continue sur R+ Type de véhicule : camping car/voiture Puissance : entière et positive  Motorisation km à la souscription Marque de véhicule Modèle de véhicule  Historique des réparations et entretiens Prix de vente du véhicule et des pièces Détails sur les composants du véhicule ...
	$ X_c^{Ass}  = 4$ variables	$ X_\Omega  = 8$ variables	—

FIGURE 2.8 – Notations utilisées pour la formalisation des indicateurs de mutualisation.

**2.2.2.1 Variance des primes pures comme mesure de démutualisation**

La relation entre partage du risque et démutualisation présentée à la section 1.2.1 suggère que plus un assureur segmente sa prime technique, plus il transfère le risque de structure dans le tarif, (en vision prime technique) *in fine* plus il désolidarise son portefeuille. Dit autrement, à mesure que  $X_c$  se rapproche de  $X_\Omega$ , l'assureur accroît la démutualisation de son portefeuille. Il en découle alors un indice de démutualisation calculé comme la part du risque de structure transférée aux assurés lors de la tarification.

### ID : Part du risque de structure cédée

Variance des primes pures tarifées comme indice de démutualisation

$$ID_{\text{Var}||BC}(\mathcal{P}f^{Ass}) = \frac{\text{Var}[\Phi^{Ass}(\mathcal{P}f^{Ass})]}{\text{Var}[\Phi^{\Omega}(\mathcal{P}f_{\Omega}^{Ass})]} = \frac{\text{Var}[\mathbf{BC}^{Ass}]}{\text{Var}[\mathbf{BC}^{\Omega}]}$$

#### Étapes de calcul de $ID_{\text{Var}||BC}$

- (i) Estimation du tarifeur  $\Phi^{\Omega}$  à partir d'une base historique  $\mathcal{D}_n^{\Omega} = \{(s_i, \mathbf{x}_{\Omega,i}, e_i) \text{ pour } i = 1 \dots n\}$ . Le modèle utilisé peut être aussi bien un GLM qu'un modèle de machine learning ;
- (ii) Estimation du tarifeur  $\Phi^{Ass}$  suivant l'une ou l'autre des approches suivantes :
  - Agrégation des coefficients de modalités et/ou des variables du modèle  $\Phi^{\Omega}$  suivant des hypothèses de structure fixées par l'assureur. Cette agrégation est facilitée dans le cas où  $\Phi^{\Omega}$  est un GLM. C'est la méthode la plus utilisée en pratique puisqu'elle est rapide à mettre en œuvre (une fois  $\Phi^{\Omega}$  estimé) et permet à l'assureur d'inclure des *a priori* sur la structure du marché futur ;
  - Estimation du tarifeur  $\Phi^{Ass}$  à partir d'une base historique de sinistralité  $\mathcal{D}_n = \{(s_i, \mathbf{x}_{c,i}^{Ass}, e_i) \text{ pour } i = 1 \dots n\}$ . Cette approche est équivalente à la précédente si les hypothèses de structure fixées sont les mêmes que les distributions marginales historiques observées dans  $\mathcal{D}_n^{\Omega}$ .
- (iii) Calcul des vecteurs de primes pures  $\mathbf{BC}^{\Omega}$  et  $\mathbf{BC}^{Ass}$  par application des tarifeurs  $\Phi^{\Omega}$  et  $\Phi^{Ass}$  au portefeuille ;
- (iv) Calcul des variances empiriques des 2 vecteurs des primes pures par :  $\frac{1}{N^{Ass}} \sum_{i=1}^{N^{Ass}} (BC_i^a - \bar{BC}^a)^2$  pour  $a \in Ass, \Omega$  ;
- (v) Déduction de  $ID_{\text{Var}||BC}$  en divisant les variances de l'étape (iv).

#### Interprétation de $ID_{\text{Var}||BC}$

Le dénominateur représente la dispersion maximale des primes pures qui peut être atteinte au regard des facteurs de risque observés  $\mathbf{X}_{\Omega}$  en utilisant le tarifeur  $\Phi^{\Omega}$ . Il s'agit donc du scénario dans lequel l'assureur cède tout le risque de structure cessible dans son tarif. Le numérateur représente la variabilité des primes pures effectivement pratiquées par l'assureur, donc le risque de structure que l'assureur consent à céder dans son tarif. Cet indicateur mesure la part de risque de structure cessible, que l'assureur consent à transférer dans son tarif.

#### 2.2.2.2 Entropie des primes pures comme mesure de démutualisation

Alors que la variance des primes pures est maximale quand le portefeuille est formé de deux profils de risque qui s'opposent de part et d'autre du risque moyen, l'entropie privilégie une distribution plus étalée des profils de risque. Sa valeur maximale égale à  $\log_2(N^{Ass})$  est

atteinte lorsque chaque assuré a un profil de risque différent de celui de tous les autres assurés et vaut ( $BC_i \neq BC_j$  pour tout assuré  $i, j$  tel que  $i \neq j$ ). Un tel scénario d'individualisation des profils de risque serait atteignable si l'assureur disposait du vecteur  $\Omega$  de l'ensemble des facteurs de risque. La mesure de l'hétérogénéité des profils de risque en portefeuille à partir de l'entropie conduit à la formulation d'un nouvel indice de démutualisation basé sur les primes pures tarifées.

### ID : Niveau d'individualisation

Entropie des primes pures tarifées comme indice de démutualisation

$$ID_{Ent||BC}(\mathcal{P}f^{Ass}) = \frac{Ent[\Phi^{Ass}(\mathcal{P}f^{Ass})]}{\log_2(N^{Ass})} = \frac{Ent[BC^{Ass}]}{\log_2(N^{Ass})}$$

### Étapes de calcul de $ID_{Ent||BC}$

- (i) Identique au cas précédent si nécessaire pour calculer  $\Phi^{Ass}$  sinon inutile de calculer  $\Phi^\Omega$  ;
- (ii) Identique au cas précédent ;
- (iii) Calcul du vecteur de primes pures  $BC^{Ass}$  par application du tarifeur  $\Phi^{Ass}$  au portefeuille ;
- (iv) Regroupement des  $N^{Ass}$  assurés en  $M$  groupes de profils de risque uniques déterminés par le tarifeur  $\Phi^{Ass}$  et la précision (en terme de décimale) de la prime pure souhaitée. En effet, si  $\Phi^{Ass}$  est construit à partir d'un GLM, il partitionne l'espace des caractéristiques en  $\prod_{j=1}^{J_c^{Ass}} |\beta_j|$  sous-espaces correspondant à des croisements entre les variables, où  $|\beta_j|$  est le nombre de modalités de la variable  $X_j$ . En décidant d'arrondir la prime pure (à l'euro près par exemple), certaines valeurs de burning cost prédites par le modèle  $\Phi^{Ass}$  coïncident, il en résulte alors un nombre de classes de risque unique  $M \leq \prod_{j=1}^{J_c^{Ass}} |\beta_j|$ . Si  $m_i$  désigne le nombre de polices dont le burning cost est égal à la  $i$ -ème plus petite valeur des burning costs distincts et  $\sum_{i=1}^M m_i = N^{Ass}$  ;
- (v) Le numérateur est calculé comme l'entropie des burning costs uniques avec les fréquences empiriques ( $m_i/N^{Ass}, i = 1 \dots M$ ) comme distribution de probabilité :  $Ent[\Phi^{Ass}(\mathcal{P}f^{Ass})] = -\sum_{i=1}^M \frac{m_i}{N^{Ass}} \log\left(\frac{m_i}{N^{Ass}}\right)$
- (vi) L'indicateur  $ID_{Ent||BC}$  est déduit en divisant le résultat de l'étape (v) par  $\log(N^{Ass})$  pour le ramener sur  $[0, 1]$ .

### Interprétation de $ID_{Ent||BC}$

Le numérateur représente la quantité d'information en bits nécessaire pour encoder le vecteur des prix pratiqués par l'assureur ; elle est nulle lorsque l'assureur pratique un prix unique. Plus l'assureur segmente son risque, plus sa grille de tarif est détaillée, plus la prime payée par chaque assuré est informative ou révélatrice de son profil de risque et plus l'entropie du

vecteur des primes est grande. L'entropie du vecteur des primes d'un portefeuille de  $N^{Ass}$  assurés est maximale et vaut  $\log_2(N^{Ass})$  lorsque chaque assuré paie une prime individuelle. Cet indicateur normalisé entre 0 et 1 quantifie donc la part d'information utilisée par l'assureur dans sa tarification par rapport à l'information qu'il faudrait pour individualiser les profils de risque. De façon plus schématique, l'inverse de cet indicateur est assimilable au nombre de profils de risque (supposés individualisables) que l'assureur mutualise pour un tarif quelconque de sa grille de prix technique. Une construction plus étoffée de cette interprétation est donnée dans l'application numérique de cet indicateur (voir partie 4.2.2).

### 2.2.2.3 Distance $L_1$ entre les primes pures comme mesure de mutualisation

L'indicateur de variance de démutualisation calculé comme la variance des Burning costs mesure indirectement la distance entre les vecteurs  $BC^{Ass}$  et  $BC^\Omega$ . Au lieu de comparer un agrégat de ces 2 vecteurs (en occurrence la variance ici), il est possible de comparer directement ces vecteurs pour apprécier à quel point ils sont proches l'un de l'autre. Pour ce faire, la norme  $L_1$  du vecteur des écarts entre les primes pures tarifées et les primes pures les plus fines atteignables est évaluée pour déterminer la prime de solidarité.

#### IM : Prime de solidarité

Distance  $L_1$  entre les primes pures comme mesure de mutualisation

$$IM_{L_1||BC}(\mathcal{P}f^{Ass}) = \frac{1}{N^{Ass}} \|BC^{Ass} - BC^\Omega\|_1 = \frac{1}{N^{Ass}} \sum_{i=1}^{N^{Ass}} |BC_i^{Ass} - BC_i^\Omega|$$

Le calcul de cet indicateur est immédiat dès que les vecteurs  $BC^{Ass}$  et  $BC^\Omega$  sont déterminés.

#### Interprétation de $IM_{L_1||BC}$

Cet indicateur correspond à l'écart absolu moyen de prime que l'assureur consent à solidariser entre assurés ; autrement dit, il quantifie le risque moyen que chaque assuré supporte ou transfère à la communauté des assurés. Plus cette valeur est grande, plus  $BC^{Ass}$  s'éloigne de  $BC^\Omega$  et donc plus l'assureur mutualise son portefeuille. Il convient de préciser quelques remarques en lien avec cet indicateur :

- Si le résultat technique attendu, évalué comme l'écart algébrique entre le montant total de primes techniques perçues et la sinistralité totale attendue ( $\sum_{i=1}^{N^{Ass}} BC_i^{Ass} - \sum_{i=1}^{N^{Ass}} BC_i^\Omega$ ), n'est pas nul, cet indicateur ne capte plus un transfert de risque entre assurés. En guise d'exemple, considérons un portefeuille dans lequel un assureur fait payer aux mauvais risques des primes moins coûteuses que leur sinistralité attendue (i.e.  $BC_i^{Ass} < BC_i^\Omega$ , si l'assuré  $i$  est un mauvais risque), dans l'optique de financer le déficit sur ces derniers par un excédent sur les bons risques. Mais si à la souscription, cet assureur est victime de sélection adverse, et ne se retrouve qu'avec les mauvais, les écarts de primes ( $BC_i^{Ass} - BC_i^\Omega$ )

ne seront pas financées par d'autres assurés mais par d'autres moyens comme les fonds propres de l'assureur. Dans ce cas particulier, il est facile de comprendre que cet indicateur ne capte plus le transfert de risque entre assurés tel qu'escompté. Le résultat attendu peut également être non nul du fait des erreurs des modèles ou des choix tarifaires techniques. Dans tous les cas il convient de corriger l'indicateur en soustrayant de la valeur obtenue le résultat technique attendu par tête en valeur absolue. Cela revient à décaler l'indicateur vers zéro comme l'indique la formule ci-contre :

$$IM_{L_1||BC_{Shift}}(\mathcal{P}f^{Ass}) = IM_{L_1||BC}(\mathcal{P}f^{Ass}) - \left| \frac{1}{N^{Ass}} \sum_{i=1}^{N^{Ass}} (BC_i^{Ass} - BC_i^{\Omega}) \right|$$

- Cet indicateur mesure un transfert de risque en valeur absolue entre assurés. Il pourrait s'accompagner d'un indicateur de synthèse du burning cost comme le risque moyen par police pour lui apporter du contexte et éventuellement simplifier sa comparaison d'un produit à un autre. Il est même envisageable de considérer le ratio entre cet indicateur et le coût moyen des sinistres pour obtenir une grandeur sans dimension et *a priori* plus universelle. Dans le cadre de ces travaux, cet indicateur est conservé sous la forme absolue et son interprétation s'accompagnera du coût moyen de sinistre. Ce choix est guidé par la volonté que l'indicateur reste concret et interprétable pour le grand public (incluant notamment les assurés).

### 2.2.3 Indices de mutualisation calculés sur l'OL

Outre la prime pure, l'assureur peut s'appuyer sur les chargements et autres commissions pour accroître la solidarité entre assurés de son portefeuille; en appliquant des chargements constants quelque soit le profil de risque, la désolidarisation du portefeuille résulte de la proximité de  $\mathbf{X}_c$  à  $\mathbf{X}_\Omega$ , c'est-à-dire est liée à la quantité d'information incluse dans la grille tarifaire. Inversement, plus les marges sont variables d'un profil de risque à un autre, plus forte est la solidarité entre les assurés. L'OL calculée comme l'écart entre la prime payée et le profil de risque individuel de l'assuré (connu sous  $\mathbf{X}_\Omega$ ) permet de capter ces deux leviers de mutualisation du portefeuille. Une OL hétérogène entre les assurés correspond soit à des chargements assureurs variables d'un profil de risque à l'autre, soit à des primes techniques tarifées différentes des profils de risque individuels connus des assurés. La dispersion des OL est dès lors utilisée pour quantifier la mutualisation d'un portefeuille. Selon que la dispersion soit mesurée par la variance ou par l'entropie, deux indicateurs de mutualisation sont obtenus. Dans les cas extrêmes, il est souhaitable que ces indicateurs vérifient les deux propriétés suivantes :

$$\begin{aligned} \mathbf{P}_c^{Ass} = cte \times (1, 1, \dots, 1) &\implies IM_{...||OL} \text{ maximal} \\ \mathbf{OL}^{Ass} = cte \times (1, 1, \dots, 1) &\implies IM_{...||OL} \text{ minimal} \end{aligned}$$

### 2.2.3.1 Variance des OL comme mesure de mutualisation

La variance des OL  $\text{Var} [\mathbf{OL}^{Ass}]$  satisfait la deuxième condition en atteignant la valeur minimale de la variance zéro, lorsque l'OL est constante. Lorsque la prime commerciale  $\mathbf{P}_c^{Ass}$  est constante, la variance des OL est égale à la variance des primes pures  $\mathbf{BC}^\Omega$ . Une approche de normalisation envisageable serait donc de diviser la variance des OL par la variance des primes pures. Toutefois, deux limites immédiates sont à relever : l'assureur peut rendre arbitrairement l'OL aussi variable qu'il le souhaite, en particulier plus variable que le burning cost, conduisant à un indicateur supérieur à 1. Ensuite, l'ordre de grandeur des OL est en général plus faible que celui des primes pures, (en admettant notamment que les chargements assureurs sont marginaux devant le risque). Cette normalisation est tout de même conservée puisqu'en pratique la variabilité des OL est moins importante que celle des primes pures. D'autre part, des montants d'OL faibles comparativement à ceux des primes pures, pourraient s'interpréter comme une utilisation limitée des chargements pour mutualiser et des primes techniques tarifées proches des profils de risque individualisés. Soit au total, une faible solidarité créée entre les assurés.

#### IM : Variabilité des OL

Variance des OL comme indice de mutualisation

$$IM_{\text{Var}|\text{OL}} (\mathcal{P}f^{Ass}) = \frac{\text{Var} [\mathbf{OL}^{Ass}]}{\text{Var} [\Phi^\Omega(\mathcal{P}f_\Omega^{Ass})]} = \frac{\text{Var} [\mathbf{OL}^{Ass}]}{\text{Var} [\mathbf{BC}^\Omega]}$$

#### Étapes de calcul de $IM_{\text{Var}|\text{OL}}$

- (i) Estimation du tarifeur  $\Phi^\Omega$  à partir d'une base historique  $\mathcal{D}_n^\Omega = \{(s_i, \mathbf{x}_{\Omega,i}, e_i) \text{ pour } i = 1 \dots n\}$ . Le modèle utilisé peut être aussi bien un GLM qu'un modèle de machine learning ;
- (ii) Calcul du vecteur de primes pures  $\mathbf{BC}^\Omega$  en appliquant le tarifeur  $\Phi^\Omega$  au portefeuille ;
- (iii) Calcul du vecteur des OL par différence entre les primes commerciales et les primes pures obtenues en (ii) :  $\mathbf{OL}^{Ass} = \mathbf{P}_c^{Ass} - \mathbf{BC}^\Omega$
- (iv) Calcul des variances empiriques des vecteurs d'OL et de primes pures ;
- (v) Déduction de  $IM_{\text{Var}|\text{OL}}$  en divisant les variances de l'étape (iv).

#### Interprétation de $IM_{\text{Var}|\text{OL}}$

Le numérateur de cet indicateur quantifie l'hétérogénéité des traitements assureurs (charge-ments et marges de solidarité) pratiqués sur les assurés en portefeuille. Un indicateur proche de 0 correspond à une OL constante, soit une absence de solidarité entre les assurés en portefeuille (prime technique proche du profil de risque le plus fin) et des marges constantes. À mesure que la solidarité est créée en portefeuille, la valeur prise par l'indicateur augmente.

### 2.2.3.2 Entropie des OL comme mesure de mutualisation

Lorsque l'entropie est utilisée pour mesurer la disparité des OL, la normalisation se fait naturellement par  $\log(N^{Ass})$ .

#### IM : Niveau d'individualisation des OL

Entropie des OL comme indice de mutualisation

$$IM_{Ent||OL}(\mathcal{P}f^{Ass}) = \frac{Ent[\mathbf{OL}^{Ass}]}{\log_2(N^{Ass})}$$

#### Étapes de calcul de $IM_{Ent||OL}$

- (i) Estimation du tarifeur  $\Phi^\Omega$  à partir d'une base historique  $\mathcal{D}_n^\Omega = \{(s_i, \mathbf{x}_{\Omega,i}, e_i) \text{ pour } i = 1 \dots n\}$ .  
Le modèle utilisé peut être aussi bien un GLM qu'un modèle de machine learning ;
- (ii) Calcul du vecteur de primes pures  $\mathbf{BC}^\Omega$  en appliquant le tarifeur  $\Phi^\Omega$  au portefeuille ;
- (iii) Calcul du vecteur  $\mathbf{OL}^{Ass}$  des OL par différence entre les primes commerciales et les primes pures obtenues en (ii) :  $\mathbf{OL}^{Ass} = \mathbf{P}_c^{Ass} - \mathbf{BC}^\Omega$
- (iv) Regroupement des  $N^{Ass}$  assurés en  $M$  groupes suivant les valeurs uniques du vecteur  $\mathbf{OL}^{Ass}$ . La précision retenue influence le nombre de groupes  $M$  constitué ;
- (v) Le numérateur est calculé comme l'entropie des OL uniques avec les fréquences empiriques  $(m_i/N^{Ass}, i = 1 \dots M)$  comme distribution de probabilité :  
$$Ent[\mathbf{OL}^{Ass}] = - \sum_{i=1}^M \frac{m_i}{N^{Ass}} \log_2 \left( \frac{m_i}{N^{Ass}} \right)$$
- (vi) Déduction de  $IM_{Ent||OL}$  par normalisation du résultat de (v).

#### Interprétation de $IM_{Ent||OL}$

L'entropie du vecteur des OL mesure l'information contenue dans les OL. Elle est d'autant plus élevée que les pratiques tarifaires sont différentes d'un assuré à l'autre. Le cas maximal  $\log_2(N^{Ass})$  est celui où chaque assuré a une OL différente de celle de tous les autres. Cet indicateur s'interprète donc comme le niveau d'individualisation des OL.

Les difficultés du tableau 2.3 sont anticipées pour le calcul pratique des indicateurs vanilles. Par ailleurs, le tableau 2.4 apporte une synthèse de ces indicateurs en soulignant leurs interprétations, leurs propriétés et comment les difficultés sus-mentionnées les impactent.

TABLEAU 2.3 – Description des difficultés des indicateurs vanilles.

Difficulté	Description
$d_1$ - Effet modèle	La modélisation statistique passe par des choix qui ne font pas nécessairement l'unanimité d'un actuaire à un autre. Ainsi, pour une même base historique $\mathcal{D}_n$ le modèle de tarification commerciale $\Phi^\Omega$ obtenu ne produira pas nécessairement, les mêmes primes pures tarifées et par suite le même indice de démutualisation.
$d_2$ - Comparaison externe	Pour calculer un indice de démutualisation sur un portefeuille, il est indispensable de disposer d'un minimum d'informations sur ledit portefeuille. Mais un indice qui nécessite des informations trop fines sur le portefeuille de la concurrence serait difficile à mettre en oeuvre en pratique
$d_3$ - Portefeuille de marché	Certains indicateurs proposés reposent sur l'hypothèse selon laquelle les assureurs ont accès à un modèle de tarification $\Phi^\Omega$ qui donnerait pour chaque assuré le profil de risque le plus individualisé au regard des variables à leur disposition. Pour que ce modèle puisse être bien calibré, il importe de connaître la structure du marché, c'est à dire la distribution (et l'importance) des facteurs de risque $\mathbf{X}^\Omega$ sur le marché auquel s'adresse le produit.

TABLEAU 2.4 – Synthèses, propriétés et difficultés des indicateurs simples.

Indicateur	$ID_{Var  BC}$	$ID_{Ent  BC}$	$IM_{Var  OL}$	$IM_{Ent  OL}$
Inputs	Modèle de calcul du profil de risque le plus fin Modèle de pricing de la prime technique Portefeuille		Modèle de calcul du profil de risque le plus fin Prime commerciale par assuré Portefeuille	
Ens. Image : [Min, Max]	[0, 1]	[0, 1]	$\mathbf{R}_+$	[0, 1]
Interprétation	Part du risque de structure transférée dans la prime technique tarifée	Ratio d'information entre le modèle de primes techniques et le modèle d'individualisation des profils de risque.	Volatilité des OL relative à la volatilité des profils de risque	Proximité des OL à un scénario d'OL individualisées
	Croît avec la désolidarisation du portefeuille Évolution concave avec l'ajout des facteurs de risque		Décroît avec la désolidarisation du portefeuille >1 : Portefeuille très (trop) solidaire 1 : Prime commerciale unique sur des profils différents 0 : Chaque profil de risque paie pour son risque	
Unité	Sans dimension			
Difficulté de l'indicateur	-		-	
$d_1$ - Effet modèle	Oui : Nécessité de connaître le modèle de pricing commercial $\Phi^{Ass}$		Non : Pas besoin de connaître le modèle de pricing de l'assureur	
$d_2$ - Comparaison externe	Oui : Impose de connaître tout le portefeuille pour le calcul de l'indicateur		Oui : Impose de connaître le portefeuille et les primes commerciales	
$d_3$ - Portefeuille de marché	Oui : Il est nécessaire de connaître le portefeuille de marché pour estimer $\Phi^\Omega$		Oui : Il est nécessaire de connaître le portefeuille de marché pour estimer $\Phi^\Omega$	

## 2.2.4 Vers les indicateurs exotiques

Deux familles d'indicateurs de mesure de mutualisation ont été proposées dans la section précédente. La première appréhende la démutualisation comme l'écart entre le choix de prime technique pratiquée par l'assureur et la prime pure la plus individualisée que l'assureur pourrait appliquer sur son portefeuille. En enrichissant cet écart par les considérations commerciales véhiculées dans le portefeuille à travers les chargements assureurs, une seconde classe de mesure de démutualisation basée sur l'OL a été proposée. Toutefois, le calcul de ces indicateurs dits simples, nécessite de connaître suffisamment le portefeuille pour en évaluer précisément le niveau de mutualisation. En effet, les indicateurs supposent connus sur le portefeuille au moins les primes techniques, les pratiques commerciales ou les primes pures individualisées. Si en interne pour un assureur, l'accès à ces informations est assez immédiat, y avoir accès pour un concurrent (afin de se positionner par rapport à son niveau de mutualisation) est bien plus difficile, voire impossible. En effet, il sera au plus réalisable d'avoir accès à la grille des prix commerciaux pratiqués par l'assureur et donc indirectement aux primes techniques pratiquées (notamment en admettant des chargements proportionnels ou constants). À partir de cette grille tarifaire, un assureur pourrait évaluer le niveau de mutualisation de son concurrent si les assurés de son portefeuille avaient souscrit chez ce concurrent. Outre la connaissance de la donnée sur le portefeuille, il est également nécessaire de disposer d'un modèle commun d'évaluation du profil de risque individualisé  $\Phi^\Omega$ . Enfin, ces indicateurs renseignent de façon globale sur le niveau de démutualisation du portefeuille sans apporter une explication sur les facteurs de risque qui porte cette démutualisation, ils ne permettent pas de comprendre directement la démutualisation.

Pour conférer à la quantification de la mutualisation une portée générale qui puisse être communément adoptée par la communauté des assureurs, il semble nécessaire d'alléger le besoin en informations pour son calcul. Nous proposons alors de mesurer la mutualisation non plus sur la prime et ses composantes, mais sur les caractéristiques des assurés ou à partir d'une connaissance partielle et agrégée du portefeuille. La généralisation de la mesure de la mutualisation aux caractéristiques des assurés permettra également d'apporter des représentations graphiques pour expliquer le rôle des facteurs dans la démutualisation du portefeuille.

## 2.3 Introduction à la complexité de Kolmogorov

La section 2.4 introduit deux nouvelles métriques de mutualisation qui généralisent les indicateurs simples. Mais en prélude à la présentation de ces nouveaux indicateurs, il est nécessaire de s'attarder sur un outil majeur de leur calcul : la complexité de Kolmogorov (CK). Cette section apporte les formalismes physiques et informatiques nécessaires pour comprendre les indicateurs exotiques, présentés dans la suite de ce chapitre et mis en application au chapitre 5. En revanche, la lecture des chapitres 3 et 4 portant respectivement sur l'étude du portefeuille et l'application des indices simples, sont indépendants de cette section.

La complexité de Kolmogorov est utilisée dans ce mémoire pour évaluer la mutualisation pratiquée par l'assureur sur la base de la description connue du portefeuille. Elle est complétée par l'entropie de Shannon qui donne une quantification de la mutualisation résiduelle non-perceptible du fait d'une connaissance partielle du portefeuille.

### 2.3.1 Définition et formalisation de la complexité de Kolmogorov

La notion de « complexité » aboutit à deux formalisations selon le contexte : l'une s'appliquant à un *calcul* et l'autre à un *objet*. En informatique, la complexité d'un calcul est le temps et/ou l'espace nécessaire pour effectuer ce calcul. Lorsqu'il s'agit d'évaluer la complexité d'un objet, une approche naturelle est de faire appel à la description de l'objet. Quantifier la complexité d'un objet s'apparente alors à quantifier la complexité de sa description. Très naturellement, un objet sera d'autant plus complexe que sa description ne sera pas « simple ». Introduite indépendamment et simultanément par Kolmogorov (1968), Solomonoff (1964) et Chaitin (1966), la complexité de Kolmogorov mesure la complexité d'un objet comme la longueur du plus petit programme qui générerait sa description. Dans ce mémoire, seules la définition et les propriétés essentielles de la CK sont présentées, pour une description plus exhaustive et ses applications, le lecteur pourrait se référer à Li et Vitányi (2019).

L'objet dont la « simplicité » est calculée au moyen de la complexité de Kolmogorov dans ce mémoire est le portefeuille. De façon schématique, un portefeuille est d'autant plus simple que les assurés qui le composent sont recroquevillés dans un sous-espace restreint des facteurs de risque ; retranscrivant ainsi une situation dans laquelle le portefeuille n'est pas composé de profils très diversifiés. La description du portefeuille concerne alors la caractérisation de la dispersion des assurés. La description d'un portefeuille pourrait notamment adopter la forme d'une chaîne binaire reconstruisant la distribution jointe du portefeuille au regard des variables tarifaires. De façon plus générale, le cadre considéré pour la description d'un objet est celui des chaînes de caractères binaires révélant une information sur l'objet ; il peut s'agir notamment d'une séquence de réponses à des questions binaires qui, mises de bout en bout permettent de reconstruire l'objet.

#### Formalisation de la complexité de Kolmogorov

---

Soit  $\Sigma^* = \{0, 1\}^*$  l'ensemble des chaînes binaires finies, la chaîne vide est notée  $\varepsilon$ ,  $\Sigma^n$  désigne l'ensemble de toutes les chaînes de longueur  $n$  et  $|x|$  désigne la longueur de la chaîne  $x$ . La définition de la complexité de Kolmogorov se formalise alors comme suit :

Soit  $U$  une machine de Turing universelle, soient deux descriptions  $x, y \in \Sigma^*$ , la complexité de Kolmogorov de  $x$  sachant  $y$  est :

$$K(x|y) = \min_p \{|p| : U(p, y) = x\}$$

où  $U(p, y)$  est la sortie du programme  $p$  utilisant les arguments  $y$  quand il est exécuté sur la machine  $U$ ,  $p$  est un programme écrit en binaire dont la longueur est  $|p|$ .

Lorsque la génération de  $x$  se fait en absence d'arguments,  $y$  prend la valeur  $\varepsilon$  et la notation  $K(x)$  est adoptée en lieu et place de  $K(x|\varepsilon)$ .

Le sens global de cette définition est le suivant : un objet qui peut être généré par un programme court est un objet simple à décrire puisqu'il suffit de quelques lignes de code pour le représenter complètement, l'information contenue dans cet objet est donc faible. À l'inverse, si un objet nécessite un long programme pour le générer, il est plus complexe à décrire et donc moins compressible.

Pour se familiariser avec cette définition, considérons les séquences binaires de  $n$  caractères suivant dont les 20 premiers sont affichés :

$$x_1 = 11111111111111111111 \dots$$

$$x_2 = 101010101010101010 \dots$$

$$x_3 = 10100010011101010110 \dots$$

La chaîne  $x_1$  peut être générée par le programme Python 2.1, elle est assez simple et se décrit simplement par « afficher '1'  $n$  fois ». La séquence  $x_2$  correspond à une répétition de '10'  $n/2$  fois. Mais dans le cas où  $n$  est impair, la séquence  $x_2$  se termine par '1' et non '10'. La séquence  $x_3$  est générée par le programme 4.1. Ce programme est plus long que le premier programme, en ce sens qu'il nécessite plus de caractères que le premier.

Code Listing 2.1 – Programme Python : afficher '1'  $n$  fois.

```
for i in range(n):
    print('1')
```

Code Listing 2.2 – Programme Python : afficher '10'  $n/2$  fois.

```
for i in range(n // 2):
    print('10')
print('1') if !(n %% 2) else pass
```

Pour la séquence  $x_3$ , aucune régularité ne semble se dégager et par suite aucune description simple de la séquence contrairement aux séquences précédentes. Il est donc plus difficile de préciser le programme qui pourrait générer cette séquence. Il est même envisageable que le programme le plus court qui génère  $x_3$  soit celui qui affiche directement  $x_3$ . La complexité de Kolmogorov calculée à taille  $n$  fixée sur ces trois chaînes, révélerait que  $x_3$  est plus complexe que  $x_2$  qui à son tour est plus complexe que  $x_1$ . Dit autrement, la description  $x_1$  est plus simple que  $x_2$ , qui est plus simple que  $x_3$ . De façon plus générale, pour toute chaîne  $x$ ,  $K(x) \leq |x| + O(1)$ , puisque pour générer une séquence, il est toujours possible d'utiliser un programme

contenant exactement cette chaîne. Le terme  $O(1)$  est la constance de la machine de Turing, il représente les instructions du programme incompressibles comme *for, print, range, if, ...* dans le cas d'un langage évolué comme Python.

Dans cet exemple, le langage de programmation utilisé est Python ; il aurait pu être le langage Java, Perl, C, Fortran etc. Le choix du langage modifierait alors inéluctablement la valeur de la complexité de Kolmogorov du fait des différences syntaxiques. Néanmoins, le *théorème d'invariance* stipule qu'à une constante près (de l'ordre de  $O(1)$ ), la complexité obtenue est équivalente quel que soit le langage de programmation utilisé pour générer la chaîne. Ce résultat érige la complexité de Kolmogorov au rang d'outils de caractérisation universel et assure en particulier qu'en passant d'un langage de programmation à un autre, l'ordonnement des chaînes suivant leur complexité reste inchangé. Afin d'uniformiser la définition de la complexité de Kolmogorov, les programmes sont considérés dans la littérature sous forme d'algorithmes et la machine de Turing utilisée comme compilateur universel.

### Machine de Turing

Dans la définition précédente,  $U(\cdot, \cdot)$  est une machine de Turing dite universelle. Elle prend en paramètre un programme  $p$  (écrit en binaire) et un ensemble d'arguments  $y$  et produit le résultat de la compilation du programme à partir des paramètres. Son caractère universel tient à sa capacité à simuler n'importe quel programme ou machine simple de Turing qui aurait pour objectif de réaliser une tâche spécifique. La machine de Turing est en fait un modèle de machine de calcul abstrait qui se concrétisera plus-tard sous la forme d'ordinateur. Elle a été introduite par le chercheur anglais Turing (1937) dans son article fondateur *On Computable Numbers, With An Application To The Entscheidungsproblem*. Turing propose de répondre à la question de décidabilité : « existe-t-il un algorithme qui décide si une proposition énoncée dans un système logique est valide ou non ? ». Combinée aux travaux de Alonzo Church sur les nombres et les fonctions calculables, cette machine va conduire à la formulation de la thèse de Church-Turing. Elle postule que pour tout problème pour lequel il existe un algorithme (procédure automatique) le résolvant, il existe une machine de Turing qui résout ce problème. La thèse de Church peut s'interpréter aujourd'hui comme suit : tout traitement systématique d'information peut être réalisé par un ordinateur correctement programmé et suffisamment puissant ; ou encore, tout nombre ou fonction calculable peut être généré par une Machine de Turing.

Alan Turing considérait sa machine comme une version minimaliste de la machine à écrire qui pourrait être contrôlée à l'aide d'un programme. Une machine à écrire peut écrire ou effacer un symbole sur la feuille, avancer ou reculer d'un ou plusieurs caractères. La machine de Turing se contente d'une bande infinie à gauche et à droite et segmentée en cellules pouvant contenir chacune un symbole. Une tête de lecture/écriture remplace les barres à caractère physique et agit sur la bande en observant les règles du programme. Par commodité, les cases de cette

bande sont souvent indexées avec l'ensemble des entiers relatifs  $\mathbb{Z}$ .

La bande assure à la fois le rôle de périphérique d'entrée et de sortie. Initialement vide, les données à traiter par le programme (les entrées,  $y$ ),  $y$  sont inscrites. Celles-ci sont codifiées en utilisant les symboles de l'alphabet  $\Sigma$  arbitraire, mais fini ( $\Sigma = \{0,1\}$  dans le cas des descriptions des objets par des chaînes binaires). Par convention, les mots sont séparés par des cellules vides et le premier symbole du premier mot est placé dans la cellule d'indice 0 où se trouve la tête de lecture/écriture. Pendant l'exécution du programme, la tête de lecture/écriture peut se déplacer à gauche ou à droite, lire le contenu de la cellule sur laquelle elle se trouve, l'effacer ou y écrire un symbole de  $\Sigma$ . À la fin de l'exécution du programme, les mots sur la bande constituent le résultat ou la sortie du calcul (la description  $x$ ).

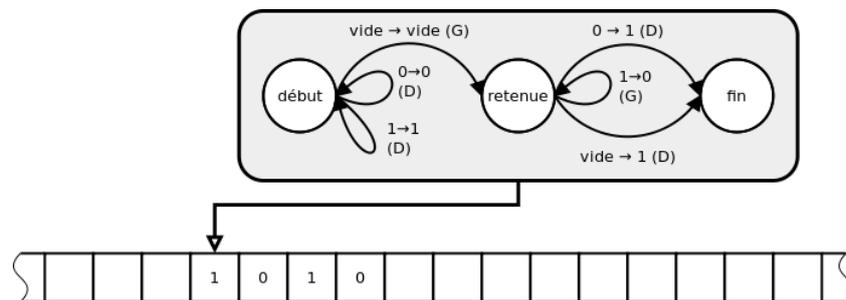


FIGURE 2.9 – Exemple de machine de Turing : programme de calcul de « +1 » en binaire.

Le fonctionnement de la machine est discret et son action à chaque étape ne dépend que de deux paramètres : le symbole lu, c'est-à-dire celui qui est situé dans la cellule sur laquelle est placée la tête de lecture/écriture, et l'état de la machine déterminé par le programme. Pour contrôler la machine, le programmeur dispose d'un ensemble fini d'états  $Q$  de son choix. Il fixe également l'état initial. C'est le programme qui va déterminer ce que la machine va faire en fonction du couple  $(p,s)$  constitué par l'état courant de la machine et le symbole lu. Sur la figure 2.9 ci contre, le ruban de la machine de Turing est représenté avec l'entrée  $y = 1010$ . Le programme qui contrôle la tête de lecture est donnée sous forme d'une suite d'instructions ; ici il s'agit d'un programme à trois états dont l'objectif est d'ajouter de « +1 » en binaire à l'entrée  $y$ .

Un programme est une suite d'instructions ou de règles destinée à résoudre un problème. Les règles peuvent être représentées sous la forme :  $p, s : o, q$  où  $(p, s)$  est le couple *condition* et  $(o, q)$  est le couple *opération*. Cette règle s'interprète par : « Si la machine est dans l'état  $p$  et que le symbole lu est  $s$  (i.e. la case sous la tête de lecture/écriture contient le symbole  $s$ ), alors la machine exécute l'opération  $o$  et passe dans l'état  $q$  ». Les opérations possibles sont :

- Écrire un symbole de l'alphabet  $\Sigma$  ;
- Effacer le contenu actuel de la cellule ;
- Déplacer la tête de lecture vers la droite ou vers la gauche.

### 2.3.2 Lien entre complexité de Kolmogorov, hétérogénéité et mesure de l'information

#### Complexité de Kolmogorov et chaîne aléatoire

Pour une taille de chaîne de caractère  $n$  fixée, l'espace  $\Sigma^n$  peut être muni d'une mesure de probabilité discrète en associant à chaque chaîne une probabilité d'être générée. Une mesure de probabilité triviale sur cet espace est la mesure uniforme qui à toute chaîne associe la même probabilité  $\frac{1}{\text{Card}(\Sigma^n)} = \frac{1}{2^n}$  d'être tirée. Pour  $n=20$ , les chaînes  $x_1$ ,  $x_2$  et  $x_3$  suivantes auront la même probabilité d'être générée :

$$\text{Chaîne régulière : } P(x_1) = P(11111111111111111111) = 2^{-20}$$

$$\text{Chaîne régulière : } P(x_2) = P(10101010101010101010) = 2^{-20}$$

$$\text{Chaîne aléatoire : } P(x_3) = P(10100010011101010110) = 2^{-20}$$

Mais en regardant ces 3 descriptions,  $x_1$  apparaît comme plus « rare » et moins probable que la chaîne  $x_3$ . L'impression que  $x_3$  soit plus probable que  $x_1$  est liée à la « régularité » apparente de  $x_1$  quand la chaîne est prise globalement. Les chaînes qui semblent plus aléatoires ou dénuées de structure à l'instar de  $x_3$ , sont plus nombreuses et apparaissent plus fréquemment que des chaînes particulières. Bien qu'elles soient rigoureusement distinctes les unes des autres, une perception globale de la chaîne générée donne l'impression que la chaîne la plus aléatoire apparaît plus fréquemment.

La complexité de Kolmogorov apporte des précisions sur le sens cet « aléa ». Si pour chaque chaîne  $x \in \Sigma^n$  sa complexité  $K(x)$  est supposée connue, être aléatoire au sens de Kolmogorov est synonyme d'avoir une complexité proche de sa longueur. Autrement dit, est aléatoire ce qui est difficile à décrire. Une chaîne facile à décrire peut être fortement compressée ( $x_1$  : '1'  $n$  fois,  $x_2$  : '10'  $n/2$  fois) contrairement à une chaîne aléatoire comme  $x_3$  qui ne peut pas être compressée. La taille de la séquence comprimée représente l'information effective contenue dans la séquence. Cette vision de l'aléa telle que précisée par la complexité de Kolmogorov et exploitée dans les algorithmes de compression, est liée au désordre de la chaîne de caractère. Une chaîne est d'autant plus désordonnée qu'elle est complexe au sens de Kolmogorov.

#### Complexité de Kolmogorov, information et entropie

La complexité de Kolmogorov et l'entropie bien que définies sur des espaces et sur des principes différents, se rapprochent en ce sens qu'elles mesurent le désordre de la description d'un objet. En effet, munissons  $\Sigma$  d'une mesure de probabilité  $P$  qui donne pour chaque chaîne sa probabilité d'être tirée.  $P$  est supposée numériquement calculable et d'entropie finie. Soit  $X$  la variable aléatoire dont les valeurs possibles sont les chaînes  $x$  distribuées suivant  $p(x)$ . Alors, la relation suivante entre l'entropie et la complexité de Kolmogorov peut être démontrée (voir

Teixeira *et al.* (2011) pour démonstration) :

$$0 \leq \sum_{x \in \Sigma^n} p(x)K(x) - H(X) \leq K(P) + O(1)$$

Cette relation indique qu'à une constante additive près, ne dépendant que de la distribution de probabilité  $P^5$ , l'entropie de Shannon de cette distribution est égale à l'espérance de la complexité de Kolmogorov des chaînes associées aux réalisations de cette distribution. Ceci implique en particulier que la complexité de Kolmogorov mesure la quantité d'information présente dans une chaîne donnée.

Une autre façon plus pratique d'étayer le lien entre la complexité de Kolmogorov et l'entropie de Shannon est donnée par Li et Vitányi (2019, chap. 2, page 190). Soit une chaîne binaire  $x = y_1y_2y_3 \dots y_{n/r}$  finie de taille  $n$  qui se décompose en  $m = n/r$  sous-chaînes de longueur  $r$ . Chaque  $y_i$  peut prendre jusqu'à  $2^r$  valeurs indicées par  $k = 1, 2, \dots, 2^r$  suivant l'ordre lexicographique. En désignant par  $p_k = |\{i : y_i = k\}|/m$  la fréquence empirique d'occurrence de la  $k$ -ième chaîne parmi les  $y_i$ , la relation suivante est démontrée :

$$K(x) \leq m(H + \varepsilon(m))$$

où  $H = -\sum_{k=1}^{2^r} p_k \log_2(p_k)$  l'entropie des blocs de  $x$  et  $\varepsilon(m) = 2^{r+1} \frac{\log_2(m)}{m}$ ,  $\varepsilon(m) \rightarrow 0$  quand  $m \rightarrow \infty$  à  $r$  fixé. Cette relation indique qu'à une constante additive près ne dépendant que du nombre de blocs  $m$  et de leur taille  $r$ , la CK d'une séquence  $x$  est égale à l'entropie de la distribution empirique des blocs multipliée par leur nombre  $m$ . Ce résultat est utilisé à la section 5.2.3.3 pour évaluer la mutualisation dans un portefeuille assurantiel.

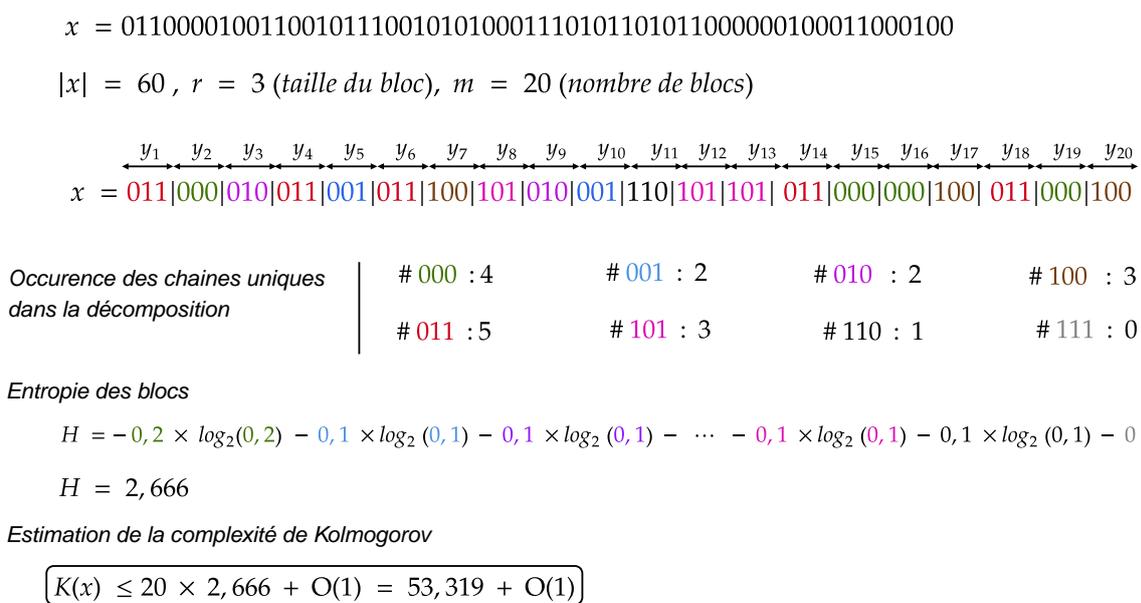


FIGURE 2.10 – Relation entre la complexité de Kolmogorov d'une chaîne binaire et l'entropie de ses sous-chaînes.

5.  $K(P)$  est la longueur du plus petit programme nécessaire pour calculer la distribution de probabilité  $P$

Si la distribution d'un portefeuille d'assurance est représentée de façon adéquate par une chaîne de nombres binaires, la complexité de Kolmogorov révèle alors l'information contenue dans cette distribution. En particulier, si le portefeuille est hétérogène c'est-à-dire disparate au regard des caractéristiques, la description obtenue devrait être difficilement compressible et donc avoir une complexité élevée. Par contre, si les assurés sont concentrés dans une sous région spécifique de l'espace des caractéristiques, la description de cette configuration nécessiterait moins d'informations et donc aboutirait à une chaîne plus simple et *in fine* à une complexité faible. La complexité de Kolmogorov est utilisée dans le cadre de ce mémoire pour mesurer l'hétérogénéité des assurés en portefeuille au regard de leurs caractéristiques.

### 2.3.3 Méthodes d'approximation de la complexité de Kolmogorov

#### La complexité de Kolmogorov n'est pas calculable

---

Il n'est pas possible pour toute chaîne  $x$  de trouver le programme optimal qui génère la chaîne  $x$  en un temps fini, ce qui implique immédiatement que la complexité de Kolmogorov n'est pas calculable. Le paradoxe « du plus petit entier naturel qui ne peut être décrit en français en moins de vingt mots » illustre bien le fait que  $K(\cdot)$  n'est pas calculable. En effet si  $K(\cdot)$  est supposé calculable, alors la fonction  $f(n) = \text{« première chaîne } x \text{ dans l'ordre lexicographique tel que } K(x) > n \text{ »}$  l'est également (cette fonction est bien définie car le nombre de chaîne binaire de longueur inférieure ou égale à  $n$  est au plus de :  $\sum_{i=1}^n 2^i = 2(2^n - 1) < 2^{n+1}$ ). Par définition,  $K(f(n)) > n$  pour tout  $n \in \mathbb{N}$ . Il existe également une constante  $c > 0$  telle que  $K(f(n)) \leq K(n) + c$ <sup>6</sup> Par ailleurs  $K(n) \leq |n| + O(1) = \log_2(n) + O(1)$ . Il vient alors qu'il existe  $d > 0$  tel que, pour tout  $n \in \mathbb{N}$ ,  $n < \log_2(n) + d$ . Ce qui est absurde.

De façon plus intuitive, pour que la complexité de Kolmogorov soit calculable, il faudrait pour toute chaîne  $x$ , faire tourner tous les programmes de longueur  $1, 2, 3, 4, 5, \dots, k, \dots$  bits et identifier le premier qui produit la chaîne  $x$ . Or, pour une longueur de  $k$  bits fixée, il peut y exister des programmes qui ne s'arrêtent pas en un temps fini (non pas à cause d'une syntaxe comme une boucle infinie, mais à cause des ressources en temps d'exécution et en puissance de calcul extrêmement grandes). Ces programmes qui ne terminent pas (c'est-à-dire qui ne s'arrêtent pas en un temps fini) ne permettent pas de lever l'ambiguïté sur le fait qu'ils génèrent ou non  $x$ . Toutefois,  $K(\cdot)$  peut être calculée pour certaines chaînes et approchée pour toute chaîne. Pour tout  $x \in \Sigma^*$ ,  $K(x)$  peut être approchée par une suite d'algorithmes  $K_1, K_2, \dots, K_i, \dots$  tel que la suite d'entiers  $K_1(x), K_2(x), \dots, K_i(x), \dots$  soit décroissante et converge vers  $K(x)$ . L'existence de cette suite d'algorithmes calculables découle de la majoration supérieure de  $K(x)$  par longueur de  $x$ .

---

6. Plus généralement, pour toute fonction calculable  $f : \Sigma^* \rightarrow \Sigma^*$ , il existe une constante  $c_f$  telle que, pour tout  $x \in \Sigma^*$ ,  $K(f(x)) \leq K(x) + c_f$ . Cette relation signifie que la complexité d'une transformation connue d'une chaîne est au plus égale à la complexité de la chaîne initiale augmentée d'une constante qui ne dépend que de la fonction. Pour générer  $f(x)$  il suffit en effet de générer  $x$  à l'aide de son programme le plus court et d'en calculer la transformée par  $f(\cdot)$ .

## Approximation de CK par les algorithmes de compression

---

Une classe de méthodes connue à travers la littérature pour approcher la complexité de Kolmogorov est l'utilisation des algorithmes de compression sans perte d'information. Si une chaîne  $x$  peut être compressée en une chaîne  $x'$  alors  $K(x)$  est approchée par  $|x'|$ . Les techniques de compression sans perte telles que l'algorithme de LempelZiv-Welch (LZW) fournissent clairement des estimations intuitives de la complexité de Kolmogorov d'un objet. Les chaînes les plus aléatoires sont difficilement compressibles contrairement aux chaînes ayant des régularités. Le taux de compression de la chaîne, calculé comme le rapport  $\frac{|x'|}{|x|}$  donne une vision normalisée de l'information apportée par la description.

## Approximation de CK par la CTM

---

La probabilité d'observer aléatoirement une chaîne de  $s$  bits de longueur  $n$  parmi les  $2^n$  chaînes possibles est de  $P(s) = 1/2^n$ . Le concept de probabilité algorithmique (également connu sous le nom de semi-mesure de Levin) remplace la production aléatoire d'une chaîne par sa génération par un programme tiré au hasard parmi une famille de programmes qui s'arrêtent. La probabilité algorithmique d'une chaîne de caractères  $s$  est donc une mesure qui estime la probabilité qu'un programme aléatoire  $p$  produise la chaîne  $s$  lorsqu'il est exécuté sur un ordinateur universel. Elle est définie par :

$$m(s) = \sum_{p:U(p)=s} \frac{1}{2^{|p|}}$$

Cette semi-mesure est liée à la complexité de chaîne  $s$  puisque le programme qui contribue le plus à  $m(s)$  est celui dont la longueur  $p$  est la plus courte. Intuitivement si la complexité d'une chaîne est faible,  $m(s)$  devrait être élevée et inversement. Le théorème de codification (Coding Theorem) formalisé par Levin (1974) précise cette relation comme suit :

$$|-\log_2(m(s)) - K(s)| < c \implies K(s) = -\log_2(m(s)) + O(1)$$

La Coding Theorem Method (CTM) proposée par Soler-Toscano *et al.* (2014) est une application directe de ce théorème. La probabilité algorithmique  $m(s)$  n'étant pas calculable pour les mêmes raisons que  $K(s)$ , les auteurs proposent de l'estimer par la fréquence empirique définie par :

$$D(t, k)(s) = \frac{|\{T \in (t, k) : T \text{ génère } s\}|}{|\{T \in (t, k) : T \text{ s'arrête}\}|}$$

où  $(t, k)$  représente l'ensemble des machines de Turing à  $t$  états internes pour un alphabet à  $k$  symboles ( $k = 2$  pour des chaînes binaires). Pour des valeurs faibles de  $t$  et des séquences binaires d'au plus 12 bits, les auteurs évaluent  $K(s)$  en remplaçant  $m(s)$  par son estimation empirique.

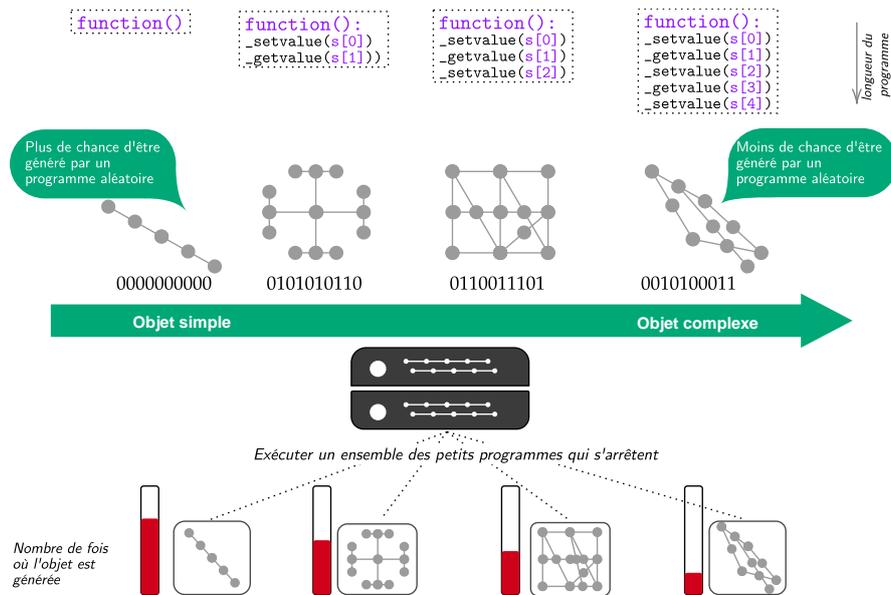


FIGURE 2.11 – Approximation de la complexité de Kolmogorov par la CTM (Coding Theorem Method).

### Approximation de CK par la BDM

Zenil *et al.* (2018) ont introduit une généralisation de la CTM qui permet de calculer la CK pour de longues descriptions sur lesquelles la CTM ne peut plus être appliquée : la Bloc Decomposition Method (BDM). Il s'agit de la méthode d'approximation de la CK la plus récente retrouvée dans la littérature. Lorsque la chaîne  $s$  devient longue, il est nécessaire d'utiliser des machines de Turing avec beaucoup plus d'états interne, mais il devient difficile d'approcher  $m(s)$  puisque de moins en moins de programmes s'arrêtent en un temps raisonnable.

Pour remédier à cette limitation, la BDM étend la CTM via une règle d'agrégation conçue pour reconstruire la complexité de Kolmogorov d'un grand objet à partir de ses composantes plus petites. Pour une séquence  $s$ , une décomposition en sous-chaînes consécutives  $s_i$  appelées blocs, de longueur constante  $r$  est effectuée :  $\mathcal{S}_s = \{s_1, s_2, \dots, s_{\lfloor \frac{|s|}{r} \rfloor}\}$ . La longueur d'un bloc  $r$  est choisie assez faible pour que la CK d'une sous-chaîne  $s_i$  puisse être approchée par la CTM. Les sous-chaînes  $s_i$  ne sont pas nécessairement uniques, elles se répètent d'autant plus que la chaîne  $s$  présente des régularités et que  $r$  est petit. Les auteurs introduisent alors l'ensemble des sous-chaînes uniques  $\mathcal{U}_s$  de  $\mathcal{S}_s$ . Au final, la complexité de la chaîne  $s$  est calculée comme suit :

$$CK(s) = BDM_r(s) = \sum_{u \in \mathcal{U}_s} CTM_r(u) + \log_2(c_u)$$

où  $c_u$  est la multiplicité de la sous-chaîne  $u$  dans  $\mathcal{S}_s$ . Cette formule s'entend comme suit : Pour générer la séquence  $s$ , il suffit d'identifier les programmes qui génèrent les séquences uniques  $u$  ; répéter l'appel de chacun de ces programmes autant de fois que chaque chaîne unique se répète dans  $s$ . La complexité additionnelle induite par les appels multiples d'un programme donné est

spécifiée par le terme  $\log_2(c_u)$  alors que  $CTM_r(u)$  approche la CK du programme optimal.

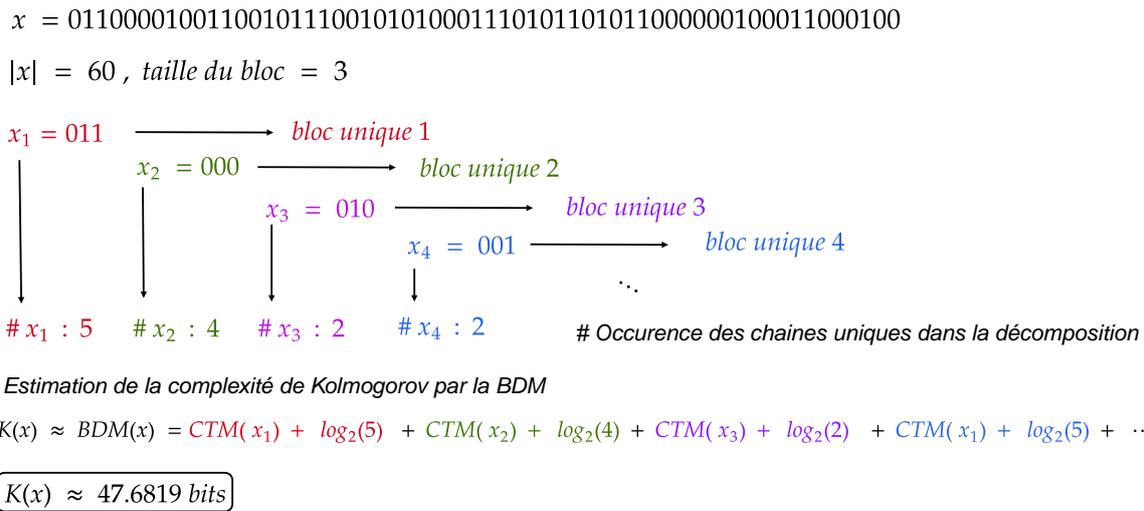


FIGURE 2.12 – Approximation de la complexité de Kolmogorov par la BDM (Bloc Decomposition Method).

Lorsque la longueur  $|s|$  de la chaîne n'est pas un multiple de  $r$ , il est nécessaire de gérer les conditions de bords. Trois approches sont présentées dans l'article originel et implémentées dans `pyBDM`<sup>7</sup> :

- *Ignoré* : décomposer la séquence  $s$  en  $\lfloor \frac{|s|}{r} \rfloor$  sous-séquences disjointes et ignorer les bits au delà de la dernière sous-chaîne ;
- *Corrélé* : considérer la décomposition précédente mais évaluer aussi la complexité de la sous-chaîne de bord ignorée ( $s \left[ \lfloor \frac{|s|}{r} \rfloor + 1, |s| \right]$ ) à partir de la CTM de l'ordre inférieur correspondant ;
- *Récuratif* : décomposer la séquence  $s$  en considérant à chaque fois un glissement de 1 bit. Deux séquences consécutives coïncident alors sur  $r - 1$  bits. Cette approche donne des résultats plus précis mais est très coûteuse en temps de calcul.

Les auteurs montrent toutefois sur des exemples empiriques que ces différentes approches sont asymptotiquement équivalentes.

7. module de calcul de la CK par la BDM implémentée sous Python et disponible <https://github.com/sztal/pybdm>

## 2.4 Indices de mutualisation exotiques basés sur les caractéristiques des assurés

La démutualisation est comprise dans cette étude comme étant la désolidarisation des assurés en portefeuille. Pour désolidariser ou non les assurés, l'assureur s'appuie sur les facteurs de risque qu'il observe. Plus la segmentation tarifaire est fine, plus la désolidarisation est prononcée. Or jusqu'à présent les indicateurs proposés ne prennent en compte qu'indirectement les variables tarifaires. Elles sont utilisées comme inputs pour le calcul des primes pures par les modèles de tarification, et seuls les burning costs générés par ces modèles sont utilisés par les indicateurs proposés à la section 2.2.2.

Deux orientations pour la généralisation de ces indicateurs sont considérées dans cette section. La première mesure la dispersion des profils de risque non plus par la variance de la prime pure qui est une agrégation, mais par la dispersion des contributions au risque des différents facteurs utilisés dans les modèles de tarification. La seconde généralisation est axée sur la dispersion des caractéristiques des assurés en portefeuille ; l'accent est porté sur la répartition des assurés dans l'espace des facteurs de risque avec des importances variables selon la pertinence du facteur de risque.

### 2.4.1 Indice de démutualisation basé sur les contributions au risque des caractéristiques

À travers cet indice, l'indice de démutualisation fondé sur le transfert de risque de structure élaboré à partir des travaux de Denuit et Charpentier, est généralisé. Résumer la sinistralité par le burning cost s'accompagne d'une perte d'information sur la dispersion des facteurs de risque dans le portefeuille. Il est en effet possible que les contributions des facteurs de risque se neutralisent de sorte que deux assurés ayant des profils très différents se retrouvent au final avec la même prime pure. Une mesure de la dispersion des risques en portefeuille pourrait être calculée en amont de la synthèse du risque en burning cost pour limiter ces effets d'agrégation.

Afin d'illustrer cette généralisation, considérons le cas simple suivant. À partir de 2 facteurs de risque ayant chacun deux modalités - marque (M1, M2) et motorisation (Essence, Diesel) -, un assureur souhaite construire la prime pure pour un contrat d'extension de garantie. À partir d'un modèle GLM dont la fonction de lien est de type exponentielle, les coefficients suivants ont été obtenus pour chaque modalité. Les différentes valeurs de burning costs à une constante  $\lambda_0$  (représentant la constante du modèle) près sont alors consignées dans le tableau suivant.

TABLEAU 2.5 – Agrégation des profils à la suite du calcul du burning cost à partir des contributions au risque.

Burning cost		Motorisation	
		Essence ( $e^\beta = 1$ )	Diesel ( $e^\beta = 0,91$ )
Marque	M1 ( $e^\beta = 1$ )	$\lambda_0$	$0,91\lambda_0$
	M2 ( $e^\beta = 1,1$ )	$1,1\lambda_0$	$\lambda_0$

Au regard des burning costs, les profils de risque  $M1 - Essence$  et  $M2 - Diesel$  sont confondus puisqu'ils cotiseraient le même burning cost, conduisant au final à trois profils de risque différents pour 4 groupes d'assurés potentiellement différents. Or en pratique, une déviation structurelle ou de la tendance de la sinistralité peut affecter le groupe  $M1 - Essence$  sans pour autant toucher le groupe  $M2 - Diesel$ . Il subsiste alors un risque de structure que l'assureur prend à sa charge en pratiquant trois tarifs. Porter un regard directement sur le burning cost ne permet pas de faire la distinction entre 2 assureurs qui :

- Pour le premier, pratiquerait 3 prix directement notamment à travers une grille tarifaire ayant 3 cases à cocher ;
- Pour le second, pratiquerait 4 prix dont 2 seraient confondus mais, qui aurait bien en entrée 4 cases à cocher sur la grille.

Le premier assureur mutualise davantage que le second assureur alors que la part du risque de structure transférée, mesurée à partir de la variance des primes pures tarifées reste inchangée entre ces deux assureurs.

La mesure de la démutualisation à travers les contributions aux risques permet ainsi de généraliser la mesure du risque de structure transféré par la segmentation tarifaire.

#### 2.4.1.1 Calcul des contributions des facteurs à la prime pure

Afin de mesurer la dispersion des profils de risque en amont de l'agrégation en prime pure, un vecteur de contributions des différents facteurs à son risque est calculé pour chaque assuré. Le choix de la méthode de calcul des contributions dépend fortement du modèle estimé pour déterminer la prime pure. Dans le cas d'un modèle GLM, la contribution des facteurs est immédiatement déterminée par les coefficients. Lorsqu'un modèle de machine learning moins explicable est utilisé, le calcul des contributions nécessite de faire appel à des outils supplémentaires d'analyse de sensibilité. Dans ce mémoire, les valeurs de SHAP Lundberg et Lee (2017) sont utilisées pour approcher ces contributions.

#### Cas d'un modèle GLM - les $\beta$

Si le burning cost est calculé à partir d'un GLM Tweedie comme présenté en section 2.2.1, alors sa prédiction (notée en utilisant le chapeau) pour un assuré  $i$  dont les caractéristiques  $x_i$

sont connues s'écrit :

$$\hat{\mathbb{E}}[S_i | \mathbf{X}_i = \mathbf{x}_i] = \hat{\mu}_i = \exp(\hat{\beta}_0 + x_{i,1}\hat{\beta}_1 + \dots + x_{i,p}\hat{\beta}_p)$$

Habituellement, les variables  $X_j$  sont discrétisées en variables indicatrices ce qui revient à créer pour chaque variable autant de variables que de modalités retenues. De façon précise, le burning cost prédit s'écrit alors :

$$\hat{\mathbb{E}}[S_i | \mathbf{X}_i = \mathbf{x}] = \hat{\mu}_i = \exp\left(\hat{\beta}_0 + \sum_{j=1}^p \sum_{l=1}^{L_j} \mathbb{I}_{\{x_{i,j} \in I_{j,l}\}} \hat{\beta}_{j,l}\right)$$

où la variable  $X_j$  est discrétisée en  $L_j$  intervalles  $I_{j,l}$  et  $\mathbb{I}_{\{\cdot\}}$  est la fonction indicatrice. Notons par  $c_{i,j}$  la contribution de la variable  $X_j$  au risque pour l'assuré  $i$  alors :

$$c_{i,j} := \frac{\partial \hat{\mathbb{E}}[S_i | \mathbf{X}_i = \mathbf{x}]}{\partial X_j} = \hat{\beta}_{j,l} \hat{\mu}_i$$

où  $l$  est le rang de l'intervalle discrétisé de  $X_j$  dans lequel tombe l'observation  $x_{i,j}$ . Il est également possible d'évaluer la contribution de chaque facteur au logarithme du burning cost aboutissant alors à :

$$c_{i,j} := \frac{\partial \ln(\hat{\mathbb{E}}[S_i | \mathbf{X}_i = \mathbf{x}])}{\partial X_j} = \hat{\beta}_{j,l}$$

En répliquant ce calcul pour chaque assuré et pour chaque facteur de risque, une matrice  $n \times p$  des contributions est obtenue  $\mathcal{M} = [c_{i,j}]_{\substack{i \in \{1,2,\dots,n\} \\ j \in \{1,2,\dots,p\}}}$

### Cas d'un modèle LightGBM - valeurs de SHAP

---

Lorsque le burning cost est calculé à partir d'un modèle de type boîte noire comme le LightGBM (voir section 2.2.1), il est plus difficile d'exprimer la prime de risque à partir d'une formule explicite et d'en déduire la contribution des facteurs de risque comme précédemment. L'obtention de la matrice  $\mathcal{M}$  est faite dans ce cas à partir des valeurs de SHAP, présentées succinctement dans cette section. Les lecteurs pourront se référer à l'article original (Lundberg et Lee, 2017) pour plus de détails.

La méthode SHAP pour SHapley Additive exPlanations (ou explication incrémentale par les valeurs de Shapley) est une méthode qui quantifie le rôle de chaque variable dans la prédiction finale du modèle. Les valeurs de SHAP décomposent en effet la sortie du modèle comme la somme des effets des différentes variables conduisant à une interprétation semblable à celle d'une régression linéaire multiple. Cette décomposition repose sur les valeurs de Shapley issues de la théorie des jeux coopératifs.

En théorie des jeux coopératifs,  $n$  joueurs collaborent pour l'obtention d'un gain  $G$ . L'intérêt porte sur la répartition *équitable* du gain entre ces joueurs. Par *équitable* il faut comprendre : «

en prenant en compte la contribution des joueurs dans l'obtention du gain ». En particulier, un joueur ne reçoit pas uniquement ce qu'il est capable d'obtenir comme gain s'il était seul, mais également pour sa contribution au groupe, lorsqu'il interagit avec les autres joueurs. Notons  $\mathcal{N} := \{1, 2, \dots, n\}$  l'ensemble des  $n$  joueurs et  $\mathbb{P}(\mathcal{N})$  l'ensemble des parties de  $\mathcal{N}$ . Une coalition de joueurs est un élément de  $\mathbb{P}(\mathcal{N})$ . On suppose qu'on dispose d'une fonction caractérisant le jeu  $c : \mathbb{P}(\mathcal{N}) \rightarrow \mathbb{R}$  qui à toute coalition de joueurs associe le gain maximal qu'ils peuvent obtenir en collaborant. En particulier  $c(\emptyset) = 0$  et  $c(\mathcal{N}) = G$ . L'allocation du gain est assimilable à une fonction  $\phi : \mathcal{N} \rightarrow \mathbb{R}$  avec  $\phi(i) = \phi_i(c)$ . Shapley (1953) introduisit quatre axiomes pour l'existence et l'unicité d'une telle fonction qui assurerait un partage efficace.

- **Efficacité** :  $\sum_{i \in \mathcal{N}} \phi_i(c) = c(\mathcal{N})$ , la somme des parts de chaque joueur doit recouvrir intégralement le gain total ;
- **Symétrie** : pour tout  $Z \in \mathbb{P}(\mathcal{N})$ ,  $c(Z \cup i) = c(Z \cup j) \implies \phi_i(c) = \phi_j(c)$ , si deux joueurs contribuent de la même façon dans toutes les coalitions dans lesquelles ils apparaissent, leurs parts doivent être égales ;
- **Joueur nul** : pour tout  $Z \in \mathbb{P}(\mathcal{N})$ ,  $c(Z \cup i) = c(Z) \implies \phi_i(c) = 0$ , si toutes les coalitions dans lesquelles un joueur est présent rapportent le même gain avec et sans lui alors la part de ce joueur doit être nulle ;
- **Additivité** :  $\phi_i(c_1 + c_2) = \phi_i(c_1) + \phi_i(c_2)$  et  $\phi_i(ac_1) = a\phi_i(c_1)$ , les valeurs de Shapley sont additives par rapport à la fonction caractéristique, ce qui signifie que distribuer successivement ou simultanément deux gains acquis l'un à la suite de l'autre aboutit aux mêmes parts.

À partir de ces axiomes, l'unique fonction qui assure un partage équitable du gain est :

$$\phi_i(c) = \frac{1}{n} \sum_{k=0}^{n-1} \binom{n-1}{k}^{-1} \sum_{\substack{Z \in \mathbb{P}(\mathcal{N} \setminus \{i\}) \\ |Z|=k}} [c(Z \cup i) - c(Z)]$$

La part d'un joueur  $\phi_i$  est obtenue comme la moyenne de tous les gains additionnels (algébriques) consécutifs  $(c(Z \cup i) - c(Z))$  à sa participation à une coalition. Il y a  $\binom{n-1}{k}$  coalitions de  $k$  joueurs et des coalitions de 1 à  $n-1$  joueurs peuvent être considérées.

Lundberg et Lee font le parallèle suivant avec la théorie des jeux où on cherche à répartir le gain  $G$ , entre les joueurs  $\mathcal{N}$  connaissant la fonction caractéristique  $c$ . En machine learning, on cherche à expliquer les contributions des observations  $x$  à la prédiction  $f(x)$  produite par le modèle  $f$ . Pour exploiter les valeurs de Shapley, la méthode SHAP remplace :

- Les joueurs par les valeurs  $x_i$  prises par  $X$  sur chaque variable explicative ;
- Le gain  $G$  à répartir par la différence  $f(x) - \hat{\mathbb{E}}[f(X)]$ , où  $\hat{\mathbb{E}}[f(X)]$  est la moyenne des prédictions ;
- La fonction caractéristique du jeu  $c$  par  $c(u) = \hat{\mathbb{E}}[f(X)|X_u = x_u]$ , définie pour toute coalition ou sous-ensemble  $u$  de valeurs  $x_u$  de  $x$ .

La déduction des contributions des différentes variables se fait en remplaçant  $c(u)$  par sa valeur dans la formule précédente. Les axiomes des valeurs de Shapley assurent en particulier que :

$$f(x) = \hat{\mathbb{E}}[f(X)] + \sum_{i=1}^p \phi_i$$

La matrice des contributions au risque des  $\mathcal{M} = [c_{i,j}]_{\substack{i \in \{1,2,\dots,n\} \\ j \in \{1,2,\dots,p\}}}$  pour un modèle de machine learning est obtenue à partir de la matrice des valeurs de SHAP. Le calcul pratique se fait à travers des implémentations de cette méthode en Python. En particulier, Lundberg *et al.* (2019) ont développé l'algorithme *Tree Explainer* pour optimiser le calcul des valeurs de SHAP dans le cas des modèles basés sur des arbres CART comme le LigthGBM.

### 2.4.1.2 Inertie des contributions comme mesure de démutualisation

Nous disposons à présent des contributions au risque pour chaque individu sous forme de matrice. La dispersion de cette matrice rend compte du transfert de risque de structure décidé par l'assureur. Des contributions semblables et peu disparates traduisent un choix de faible différenciation des profils de risque par l'assureur. Inversement, une matrice de contribution très éparse traduit un choix de démutualisation important de l'assureur, puisque ce dernier aurait inclut dans son tarif un grand nombre de variables à des mailles très fines. Cette appréhension de la démutualisation sous le prisme de la matrice de contributions au risque suggère de mesurer la démutualisation à partir de la dispersion de cette matrice. En observant qu'on peut considérer une matrice  $n \times p$  comme un nuage de  $n$  points de  $\mathbb{R}^p$ , quantifier sa dispersion peut se faire à travers le calcul de l'inertie de ce nuage de points.

L'inertie d'un nuage de points  $\mathcal{C}_n = \{c_i \in \mathbb{R}^p, i = 1, \dots, n\}$  par rapport à un point  $g \in \mathbb{R}^p$  est définie par :

$$I(\mathcal{C}_n) = \frac{1}{n} \sum_{i=1}^n d^2(c_i, g), \text{ où } d(\cdot, \cdot) \text{ est une distance}$$

Nous prenons  $g = \frac{1}{n} \sum_{i=1}^n c_i$  le vecteur des contributions moyennes des variables et  $d(u, v) = \|u - v\|_2$  la distance euclidienne. L'assureur considère toujours deux modèles de pricing, le tarifeur  $\Phi^\Omega$  qui génère les profils de risque les plus fins qu'il peut pratiquer et le tarifeur  $\Phi^{Ass}$  moins fin qui donne les primes pures en fonction des variables explicitement utilisées par l'assureur dans son tarif. Pour chacun de ces modèles, la matrice de contribution est déterminée à partir de la méthode adéquate. Elles sont notées respectivement  $\mathcal{M}^\Omega = [c_{i,j}^\Omega]_{\substack{i \in \{1,2,\dots,n\} \\ j \in \{1,2,\dots,J_{\mathbf{X}^\Omega}\}}}$  et  $\mathcal{M}^{Ass} = [c_{i,j}^{Ass}]_{\substack{i \in \{1,2,\dots,n\} \\ j \in \{1,2,\dots,J_c^{Ass}\}}}$ . La formule suivante pour l'indice de démutualisation basé sur les contributions au facteur de risque est déduite :

**ID : Part de risque de structure cédée généralisée**

Inertie des contributions comme mesure de démutualisation

$$ID_{Int||\partial\mu_i}(\mathcal{P}f^{Ass}) = \frac{I(\mathcal{C}_n^{Ass})}{I(\mathcal{C}_n^\Omega)} = \frac{\sum_{i=1}^n \|c_i^{Ass} - \bar{c}^{Ass}\|_2^2}{\sum_{i=1}^n \|c_i^\Omega - \bar{c}^\Omega\|_2^2}$$

où  $c_i^{(a)}$  est le vecteur des contributions pour l'individu  $i$  pour le modèle  $(a)$  et  $\bar{c}^{(a)}$  est le vecteur des contributions moyennes par variable.

**Étapes de calcul de  $ID_{Int||\partial\mu_i}$**

- (i) Estimation du tarifeur  $\Phi^\Omega$  à partir d'une base historique. Le modèle utilisé peut être aussi bien un GLM qu'un modèle de machine learning ;
- (ii) Estimation du tarifeur  $\Phi^{Ass}$  par un GLM, un modèle de machine learning ou l'agrégation des coefficients de  $\Phi^\Omega$  ;
- (iii) Calcul des matrices de contributions  $\mathcal{M}^\Omega$  et  $\mathcal{M}^{Ass}$  en utilisant les coefficients du modèle où les valeurs de SHAP selon que les étapes (i) et (ii) aient été réalisées avec un GLM ou un modèle de machine learning ;
- (iv) Calcul des inerties des matrices de l'étape (iii) en remarquant notamment que  $I(\mathcal{C}_n^{(a)}) = n \sum_{j=1}^{J(a)} \text{Var}(c_j)$  ;
- (v) Déduction de  $ID_{Int||\partial\mu_i}$  en divisant les inerties de l'étape (iv).

**Interprétation de  $ID_{Int||\partial\mu_i}$**

Cet indicateur généralise la mesure du transfert de risque de structure par la tarification pratiquée par l'assureur. Plus il est grand plus le modèle tarifaire est segmenté et plus éparses sont les contributions au risque des assurés. La décomposition de l'inertie du nuage des contributions comme la somme des variances des facteurs de risque apporte une interprétation plus précise de cet indicateur. L'indicateur peut en effet se réécrire comme suit :

$$ID_{Int||\partial\mu_i}(\mathcal{P}f^{Ass}) = \sum_{j=1}^{J_{Ass}} \frac{\text{Var}(C_j^{Ass})}{\sum_{l=1}^{J_{X^\Omega}} \text{Var}(C_l^\Omega)}$$

Le terme,  $\frac{\text{Var}(C_j^{Ass})}{\sum_{l=1}^{J_{X^\Omega}} \text{Var}(C_l^\Omega)}$  s'interprète comme une contribution à la démutualisation du  $j$ -ième facteur tarifaire utilisé dans la grille de tarifs. Plus il est élevé, plus la segmentation tarifaire induite par ce facteur de risque est importante.

## 2.4.2 Indice de mutualisation basé sur la complexité de Kolmogorov

### Intuitions sur la construction de l'indicateur

Cet indicateur est élaboré à partir d'une extension de l'idée charnière des indicateurs simples, c'est-à-dire la mesure de la démutualisation à partir de la dispersion des profils de risque en portefeuille ou de la concentration des OL. Une généralisation est introduite pour passer d'une dispersion mesurée sur les primes et ses composantes à une dispersion calculée sur les facteurs de risque en s'affranchissant du modèle tarifaire de l'assureur. Schématiquement, en considérant chaque facteur de risques comme un axe d'un espace vectoriel à  $|\mathbf{X}_\Omega|$  dimensions. Chaque assuré en portefeuille est représenté par un point dans cet hyper-cube. Dans cet hyper-cube, tous les assurés recroquevillés au même endroit traduisent un profil de risque remarquablement similaire. Alors que les assurés répartis partout dans l'hyper-espace laissent penser à des profils de risque très différents. Dans la configuration des assurés recroquevillés, le portefeuille de l'assureur n'est constitué que d'une fraction minime des risques présents sur le marché. Une telle configuration traduit alors un état démutualisé du portefeuille puisque l'assureur se concentre sur des niches spécifiques du marché. Inversement, si l'assureur réussit à mettre ensemble des assurés avec des profils différents, alors une représentation des assurés dans l'hyper-espace des facteurs de risque donnerait lieu à des points dispersés, traduisant un portefeuille mutualisé.

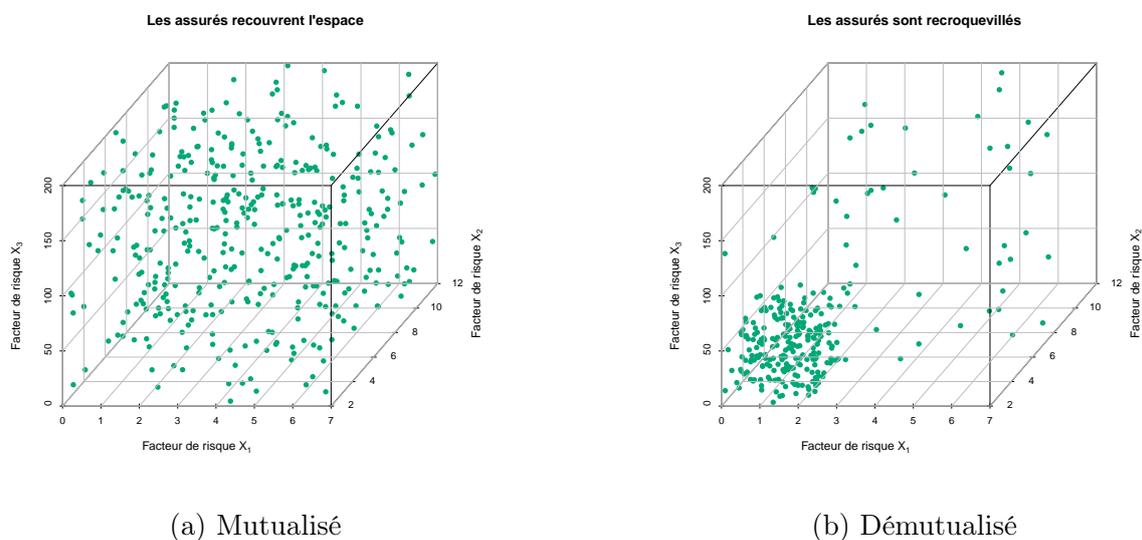


FIGURE 2.13 – Nuages des points des assurés dans l'espace des facteurs de risque : portefeuille mutualisé à gauche et portefeuille démutualisé à droite.

Ainsi mesurer la dispersion des assurés dans cet espace permettrait de quantifier l'hétérogénéité des profils de risque en portefeuille. Toutefois, les facteurs de risques ne possèdent pas la même importance, il est de ce fait nécessaire de choisir judicieusement les facteurs suivant lesquels la mesure de la dispersion sera priorisée. Pour y parvenir, l'approche proposée dans ce mémoire consiste à observer les profils de risque individualisés dans cet hyper-espace, afin d'identifier les régions de cet espace où le profil de risque est homogène mais suffisamment

distant de celui des autres régions. Soit un partitionnement supervisé de l'espace des caractéristiques. La méthode des arbres CART apparaît alors comme une méthode de prédilection pour réaliser un tel partitionnement.

Par ailleurs, la dispersion des assurés dans l'hyper-espace seule ne suffit pas pour rendre compte de la mutualisation ou non de l'assureur. La répartition géographique des assurés rend compte des disparités des profils collectés, mais elle n'indique pas comment l'assureur crée ou non la solidarité entre ces derniers. Si en situation concurrentielle, l'hétérogénéité des profils rend indirectement compte de la capacité de l'assureur à se positionner sur tous les risques du marché et à les mutualiser, l'incertitude sur le niveau véritable de mutualisation est plus pesante en situation de monopole. En effet, considérons deux assureurs qui seraient mutualisés à l'extrême pour le premier (prime commerciale unique) et démutualisés à l'extrême pour le second (pratique des prix individualisés et des chargements constants). En situation de monopole, les deux posséderaient exactement la même représentation du portefeuille dans l'espace des caractéristiques bien qu'il soit clair que les niveaux de mutualisation des portefeuilles sont différents.

Afin de lever cette ambiguïté, il est possible d'observer également la dispersion de l'OL dans les différentes régions de l'espace des assurés. Des régions de l'espace avec des OL qui varient peu au sein de la région, mais beaucoup par rapport à d'autres régions de l'espace traduisent une création de solidarité à travers l'OL entre ces groupes d'assurés par l'assureur. En effet, si l'OL est constante quelle que soit la région de l'espace, alors la prime commerciale est une simple majoration du profil de risque  $BC^\Omega$  et le portefeuille est démutualisé à l'extrême. Par contre, plus la variabilité des OL dans l'espace des caractéristiques est importante, plus l'assureur mutualise son portefeuille. Dans le cas extrême d'une prime commerciale unique, la variation de l'OL s'aligne exactement avec la variation des profils de risque.

### **Gaz de Boltzmann et démutualisation d'un portefeuille d'assurance**

---

Une solution technique à ce problème a été proposée par la communauté des physiciens dans le cas d'un gaz contenu dans un hyper-cube. En effet, cette représentation des assurés dans un hyper-espace, peut faire penser au gaz de Boltzmann introduit à la section 2.1.1 sur l'entropie. Les assurés peuvent être assimilés à des molécules de gaz en interaction (solidarité par un transfert de prime) dans un hyper-cube (espace des caractéristiques). Partant de ce parallèle, la dispersion au sein de cet espace peut être évaluée, à partir d'une mesure d'entropie, d'un point de vue micro-état et/ou macro-état selon la granularité d'observation des assurés retenue.

Un gaz de Boltzmann peut être décrit par la mesure d'une ou de plusieurs grandeurs d'état, conduisant alors à une description macroscopique notée  $\Gamma_x$  du gaz (le système). « La fraction des molécules dans une moitié de l'hypercube » ou « l'écart entre les vitesses moyennes des particules dans les deux moitiés de l'hyper-cube » sont des exemples de macro description  $\Gamma_x$ . À cette description macroscopique, peut correspondre plusieurs configurations microscopiques

plus ou moins hétérogènes, avec une éparsité des molécules plus ou moins importante. Il s'agit de l'hétérogénéité résiduelle qui subsiste même après la mesure  $\Gamma_x$ . À titre d'exemple, si la description macroscopique  $\Gamma_x$  spécifie que 20 % des molécules sont dans une moitié donnée de l'hyper-cube, aucune précision n'est apportée sur la répartition de ces 20 % dans tout ce volume. Ces molécules peuvent être toutes recroquevillées en un point de cette moitié de l'hyper-cube ou distribuées uniformément. Cette incertitude qui demeure malgré la connaissance partielle  $\Gamma_x$  correspond à l'hétérogénéité non observée. L'hétérogénéité totale correspond à la somme de l'hétérogénéité captée à travers la mesure  $\Gamma_x$  et l'hétérogénéité résiduelle.

### Entropie Algorithmique

Il existe une notion en physique pour mesurer la dispersion dans un système, introduite par Zurek (1989) qui recouvre bien cette situation : l'entropie algorithmique. Il s'agit de la somme de l'information contenue dans la description  $\Gamma_x$  mesurée à partir de la complexité de Kolmogorov et d'une approximation de l'information non observée mesurée à partir de l'entropie de Shannon. Si la description du macro-état  $\Gamma_x$  est encodée sous forme d'une chaîne binaire  $x$ , (ex :  $x = 100011001011\dots$ ), et que  $H(\Gamma_x)$  désigne l'entropie de Shannon correspondant à ce macro-état, alors l'entropie algorithmique (notée  $\mathcal{S}_{\Gamma_x}$ ) est définie par :

$$\mathcal{S}_{\Gamma_x} = K(x) + H(\Gamma_x)$$

où  $H(\Gamma_x) := - \sum_k p_{k|\Gamma_x} \log_2(p_{k|\Gamma_x})$  avec  $p_{k|\Gamma_x}$  la probabilité que le système se trouve dans le micro-état  $k$  conditionnellement à la connaissance du macro-état  $\Gamma_x$ . Lorsque les descriptions sont réalisées de façon itérative et apportent une connaissance de plus en plus riche sur le système étudié, ( $\Gamma_{x_1} \supseteq \Gamma_{x_2} \supseteq \Gamma_{x_3} \supseteq \dots \supseteq \Gamma_{x_i} \supseteq \dots$ ), la complexité de Kolmogorov de la description croît puisque l'état du système est de plus en plus connu. Inversement, à chaque nouvelle description, l'information résiduelle est précisée, ce qui permet d'affiner les probabilités conditionnelles  $p_{k|\Gamma_{x_i}}$ , et en particulier d'écartier certains micro-états qui deviennent incompatibles avec le micro-état. Par suite,  $H(\Gamma_{x_i})$  décroît avec l'enrichissement de la connaissance du système. Dans le cas d'un portefeuille assurantiel, les descriptions envisagées apportent une information sur la dispersion des assurés dans l'espace des facteurs de risque et/ou la concentration des OL. Plus les mesures réalisées sur le portefeuille sont nombreuses, meilleure est la connaissance sur la dispersion des assurés, et *in fine*, meilleure est l'appréhension du recouvrement de la structure du marché par le portefeuille étudié.

Pour un gaz, l'entropie de Boltzmann peut se noter  $\mathcal{S}_{\Gamma_x}$  où  $x$  désigne l'encodage de la description sous forme d'une suite de 0 et de 1 ( $\Gamma_x$  est compatible avec un ensemble de micro-états dont la description macroscopique binaire serait  $x$ ). Typiquement, le macro-état  $\Gamma_x$  correspond à des informations mesurées de haut niveau : le nombre de particules ou leur vitesse dans une région spécifique du cube par exemple. Plus des mesures sont réalisées, plus l'état

du système se précise (la longueur de  $x$  augmente), plus le nombre de micro-états compatibles diminue.

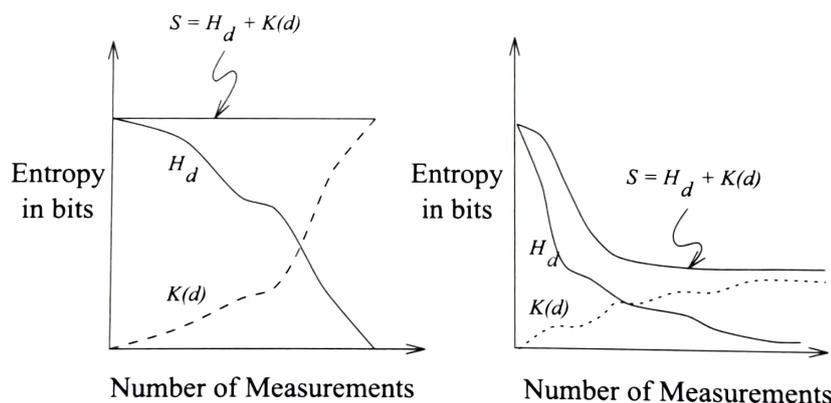


FIGURE 2.14 – Entropie algorithmique : à gauche système en équilibre, à droite système dans un état simple particulier, simple à décrire.

**Note de lecture :**  $H_d$  est l'entropie de Shannon calculée après  $d$  mesures. Elle décroît avec le nombre de mesures puisque l'incertitude est atténuée progressivement.  $K(d)$  est la complexité de Kolmogorov sur la description obtenue au bout de  $d$  mesures. Elle croît avec la connaissance sur le système. Dans le cas d'un système en équilibre, c'est-à-dire en situation de dispersion maximale, la décroissance de  $K(d)$  est exactement égale à l'accroissement de  $H_d$  et conduit à une hétérogénéité  $S$  du système constant (Li et Vitányi, 2019, chap. 8.7).

Afin d'apprécier à quel point les assurés sont dispersés dans l'espace des caractéristiques, il faut préalablement définir une stratégie de mesures itératives du portefeuille dans l'espace des facteurs de risque. Pour quantifier la dispersion des assurés en portefeuille par l'entropie algorithmique, il est nécessaire de procéder à un encodage du portefeuille sous forme de chaînes binaires et de spécifier une approche pour définir et estimer les probabilités  $p_{k|\Gamma_{x_i}}$ .

#### 2.4.2.1 Encodage du portefeuille : mesures itératives de la dispersion

##### Encodage du portefeuille par un arbre CART

On souhaite mesurer l'hétérogénéité des profils de risque et la dispersion des OL pratiquées par l'assureur dans l'espace engendré par les facteurs de risque. Pour y parvenir, l'espace des caractéristiques est partitionné en régions homogènes au regard du risque couvert. Un arbre CART avec une croissance par feuille, dont le principe est présenté à la section 2.2.1.2, est utilisé à cet effet. Un arbre CART est construit pour expliquer le profil de risque individualisé par les facteurs de risque. La lecture se fait du haut vers le bas :

- Initialement le couple (variable & seuil) le plus important est déterminé et l'espace des facteurs de risque est partagé en deux régions. Dans chacune de ces régions, deux mesures sont réalisées : la fraction d'assurés présente dans le sous-espace et la variance de l'OL des assurés du sous-espaces. Ces deux valeurs sont ensuite converties en chaînes binaires et conduisent à la première description du portefeuille  $x_1$  soit un macro-état  $\Gamma_{x_1}$  ;

- Ensuite, la feuille avec le plus d'hétérogénéité suivant le risque est scindée en deux feuilles filles au regard du nouveau couple (variable & seuil) le plus pertinent, correspondant à un partitionnement d'un des deux sous-espaces en deux nouveaux sous-espaces (soit trois sous-espaces au total). Dans les deux nouveaux sous-espaces précisés, la fraction des assurés et la variabilité de l'OL sont calculées. Une concaténation de ces nouvelles informations sous forme binaire avec la description  $x_1$  permet d'obtenir  $x_2$  dont le macro-état correspondant est  $\Gamma_{x_2}$  ;
- De façon plus générique, après  $d$  mesures, l'espace des facteurs a été partitionné en  $d + 1$  régions ou feuilles. À chaque région  $i$  sont associées la fraction  $f_{d,i}$  des assurés qui s'y trouvent et la variance des OL  $V(OL)_{d,i}$  calculée sur ces derniers. Les représentations binaires  $({}^b f_{d,i}, {}^b V(OL)_{d,i})$  de ces grandeurs sont empilées entre elles pour définir le macro-état  $\Gamma_{x_d}$  dont les descriptions sont  $x_d^{freq} = \bigcup_{i=1}^{d+1} {}^b f_{d,i}$  et  $x_d^{OL} = \bigcup_{i=1}^{d+1} {}^b V(OL)_{d,i}$ . Pour passer à l'étape suivante, la région ayant le plus de disparité en terme de profils de risque est scindée au regard du couple (variable & seuil) le plus impactant, en deux sous régions. Une nouvelle fraction est ajoutée  $x_d^{freq}$  pour obtenir  $x_{d+1}^{freq}$ . De même  $x_d^{OL}$  est mis à jour pour obtenir  $x_{d+1}^{OL}$  ;
- L'étape précédente est réitérée jusqu'à ce que les subdivisions des feuilles ne soient plus possibles. Soit parce qu'il ne reste plus qu'un assuré unique dans cette région ; soit parce que tous les assurés de la région ont rigoureusement les mêmes caractéristiques ; soit parce que les différences de caractéristiques entre les assurés des régions terminales ne permettent plus de séparer la sinistralité.

Dans le cas d'un portefeuille d'extension de garantie automobile, à titre d'exemple, les descriptions macroscopiques des dispersions de fréquence du portefeuille peuvent prendre les formes suivantes :

$\Gamma_{x_1}^{freq} = \ll 40 \% \text{ des véhicules du portefeuille sont de marques TOYOTA ou FORD} \gg$

$\Gamma_{x_2}^{freq} = \Gamma_{x_1}^{freq}$  et  $\ll 20 \% \text{ des véhicules TOYOTA ou FORD ont moins de 6 mois à la souscription} \gg$

$\Gamma_{x_3}^{freq} = \Gamma_{x_2}^{freq}$  et  $\ll 75 \% \text{ des véhicules qui ne sont ni des TOYOTA ni des FORD ont une puissance fiscale de plus de 10 chevaux fiscaux} \gg$

$\Gamma_{x_4}^{freq} = \Gamma_{x_3}^{freq}$  et ...

**Remarque : un *splitter* pré-entraîné sur le portefeuille de marché**

Les descriptions itératives du portefeuille présentées plus haut supposent connu un arbre CART qui indiquerait les différents couples (variable & seuil) ainsi que leur ordre d'importance pour encoder le portefeuille. Une présentation plus détaillée de ce « *splitter* » ainsi que celle de sa construction sont données en fin de chapitre 2.4.3.

### Mesure de l'hétérogénéité résiduelle avec l'entropie

---

À l'étape de déroulement  $d$  du *splitter*, les micro-états compatibles avec l'état  $\Gamma_{x_d}$  dont la description est  $x_d$ , sont toutes les répartitions possibles des assurés dans les feuilles terminales de l'arbre de sorte que les fréquences  $f_{d,i}$  et les variances  $V(OL)_{d,i}$  mesurées soient obtenues si tous les assurés sous les  $d + 1$  nœuds sont regroupés dans les nœuds supérieurs. De façon plus pratique, il s'agit de considérer pour chacun des nœuds  $\mathcal{N}_{d,i}$  contenant une fraction  $f_{d,i}$  du portefeuille, les répartitions possibles des assurés dans les feuilles terminales de l'arbre qui sont sous ce nœud. Si les feuilles sous le nœud  $\mathcal{N}_{d,i}$  sont indexées par  $l = 1, 2, \dots, L_{d,i}$  et  $p_{l|d,i}$  la probabilité qu'un assuré du nœud intermédiaire  $\mathcal{N}_{d,i}$  se retrouve dans la feuille terminale indexée  $l$ . L'entropie pour la région correspondant au nœud  $\mathcal{N}_{d,i}$  est calculée par :

$$H(\mathcal{N}_{d,i}) = -|\mathcal{N}_{d,i}| \sum_{l=1}^{L_{d,i}} p_{l|d,i} \log_2(p_{l|d,i})$$

Cette formule s'entend comme suit : à chaque assuré du nœud  $\mathcal{N}_{d,i}$ , correspond  $L_{d,i}$  feuilles terminales possibles (micro-état) dans lesquelles pourraient se trouver l'assuré si la description était complète. L'ampleur de l'incertitude sur la position exacte de l'assuré est donnée par l'entropie individuelle :  $\sum_{l=1}^{L_{d,i}} p_{l|d,i} \log_2(p_{l|d,i})$ . L'incertitude sur la position exacte de chaque assuré étant exactement la même, il suffit de multiplier l'entropie d'un assuré par le nombre d'assurés du nœud :  $|\mathcal{N}_{d,i}|$ .

L'entropie étant une grandeur extensive, l'entropie microscopique correspondant au macro-état  $\Gamma_{x_d}$  est la somme des entropies microscopiques des  $d + 1$  régions connues à cette étape de l'encodage :

$$H(\Gamma_{x_d}) = \sum_{d=1}^{d+1} H(\mathcal{N}_{d,i})$$

Les probabilités  $p_{l|d,i}$  sont déterminées à partir des fractions d'assurés du marché qui se retrouveraient dans les feuilles terminales si le *splitter* était appliqué à l'encodage du *portefeuille de marché* c'est-à-dire à tous les assurés du marché observés sur les facteurs de risque  $\mathbf{X}_\Omega$ . Un cas particulier est celui de la distribution uniforme des assurés dans les feuilles sous un nœud donné. Cela revient à poser :  $p_{l|d,i} = \frac{1}{L_{d,i}}$ . L'entropie microscopique du nœud  $\mathcal{N}_{d,i}$  se simplifie et devient l'entropie de Boltzmann (à une constante près) :  $H(\mathcal{N}_{d,i}) = |\mathcal{N}_{d,i}| \log_2(L_{d,i})$  et l'entropie du macro-état se déduit comme suit :

$$H(\Gamma_{x_d}) = \sum_{d=1}^{d+1} |\mathcal{N}_{d,i}| \log_2(L_{d,i}) = \log_2 \left( \prod_{d=1}^{d+1} L_{d,i}^{|\mathcal{N}_{d,i}|} \right)$$

En pratique, les probabilités  $p_{l|d,i}$  seront déterminées à partir du *portefeuille de marché* et utilisées sur les portefeuilles pour lesquels la mesure du niveau de démutualisation est envisagée. En absence d'informations sur la distribution du portefeuille de marché, l'hypothèse de répartition uniforme sera adoptée.

### 2.4.2.2 Entropie algorithmique comme mesure de mutualisation

L'indice de mutualisation fondé sur l'entropie algorithmique se formalise finalement comme suit. L'assureur dispose de :

- Un arbre de régression appelé *splitter*, entraîné sur le portefeuille de marché afin d'expliquer le profil de risque individualisé par les facteurs connus par les assureurs ;
- La distribution du portefeuille (fréquence et/ou variance des OL) dans les différents nœuds de l'arbre. Notons qu'une connaissance même partielle de la distribution dans les feuilles permet de calculer une estimation du niveau de mutualisation. Il n'est donc plus nécessaire de connaître tout le portefeuille ou toute la stratégie commerciale de l'assureur.

Pour une description macroscopique du portefeuille  $\Gamma_{x_d}$  (en terme de fréquences des assurés dans les nœuds du *splitter* ou de variabilité des OL dans ces nœuds), deux composantes de l'entropie algorithmique sont évaluées : la dispersion des assurés connue, évaluée comme la complexité de la description  $\Gamma_{x_d}$  à travers  $K(x_d)$  ; la dispersion résiduelle des assurés quantifiée à partir de l'entropie de Shannon  $H(\Gamma_{x_d})$  des micro-états compatibles avec  $\Gamma_{x_d}$ . La somme de ces deux composantes indique à quel point les assurés sont dispersés dans l'espace des facteurs de risques. Plus le nombre de mesures  $d$  augmente, plus la description  $\Gamma_{x_d}$  est riche et plus précise est la connaissance sur le recouvrement de l'espace par le portefeuille. La valeur retenue de cet indice de mutualisation est la dernière valeur de l'entropie algorithmique qui peut être calculée à partir de la plus longue description  $\Gamma_{x_d}$  connue du portefeuille. Si l'entièreté du portefeuille est connue, l'arbre d'encodage est déroulé jusqu'aux feuilles terminales. L'entropie algorithmique étant décroissante, sa valeur terminale calculable correspond nécessairement à sa valeur minimale.

Une forme normalisée de cet indicateur est introduite pour en faire une mesure de mutualisation comprise entre 0 et 1. Le facteur de normalisation vaut :  $N^{Ass} \log_2(L_{0,0})$  avec  $L_{0,0}$  le nombre de feuilles terminales du *splitter*. Ce terme correspond à l'hétérogénéité maximale qui puisse être observée sur un portefeuille de  $N^{Ass}$  assurés.

#### IM : Ratio de mutualisation généralisé

Entropie algorithmique comme mesure de mutualisation : le ratio de mutualisation généralisé

$$IM_{CK||Ent} = \frac{\min_d \{K(x_d) + H(\Gamma_{x_d})\}}{N^{Ass} \log_2(L_{0,0})}$$

avec  $x_d$  l'encodage binaire du portefeuille au bout de  $d$  mesures,  $H(\Gamma_{x_d})$  l'entropie microscopique du portefeuille lorsque  $x_d$  est connu et  $L_{0,0}$  le nombre de feuilles terminales du *splitter*.

### Étapes de calcul de $IM_{CK||Ent}$

- (i) La fréquence et/ou la variance des OL des assurés dans les  $d_{max}$  premiers nœuds du *splitter* sont utilisés comme inputs du calcul. Si cette description n'est pas connue mais que le portefeuille l'est, il suffit de la déterminer en :
  - (a) Calculant l'OL à partir de la prime commerciale et du profil de risque  $BC^\Omega$  ;
  - (b) Appliquant le splitter sur le portefeuille pour distribuer les assurés dans les nœuds et d'en déduire les descriptions de fréquences et de variations d'OL.
- (ii) Pour  $d \in 1, 2, \dots, d_{max}$  :
  - (a) Déterminer l'encodage du portefeuille  $x_d$  ( $x_d^{freq}$  et/ou  $x_d^{OL}$ ) et estimer la complexité de l'encodage  $K(x_d)$  ;
  - (b) Calculer l'entropie  $H(\Gamma_{x_d})$  à partir des probabilités obtenues sur le portefeuille de marché ;
  - (c) Calculer l'entropie algorithmique  $\mathcal{S}_{\Gamma_{x_d}} = K(x_d) + H(\Gamma_{x_d})$ .
- (iii) Normaliser les  $d_{max}$  valeurs de  $\mathcal{S}_{\Gamma_{x_d}}$  en les divisant par  $N^{Ass} \log_2(L_{0,0})$  ;
- (iv) Visualiser l'évolution du ratio de mutualisation généralisé en fonction du nombre de mesures  $d$  ;
- (v)  $IM_{CK||Ent}$  est la plus petite valeur obtenue.

### Interprétation de $IM_{CK||Ent}$

Lorsque l'indicateur est calculé à partir des descriptions de fréquences, il mesure le niveau de recouvrement de la structure de marché résumée dans le *splitter* . Plus le portefeuille est constitué de profils de risque différents plus l'indicateur est élevé ; il s'agit donc d'un indicateur de mutualisation, perçu comme la capacité de l'assureur à réunir des profils de risque différents dans le même portefeuille. Si la complexité de Kolmogorov de la description est élevée, alors les assurés sont très dispersés dans l'espace des caractéristiques. De même une entropie de Shannon du macro-état précisé par sa description signifie que les assurés sont dans plusieurs sous-espaces de l'espace des caractéristiques. Un décrochage de la courbe de l'entropie de Shannon est le résultat d'une concentration des assurés dans une sous-région de l'espace.

L'interprétation dans le cas d'une description des OL est similaire à celle des fréquences. Ainsi, l'indicateur capte non seulement la mutualisation liée l'hétérogénéité des profils en portefeuilles, mais également aux pratiques tarifaires favorisant la solidarité en portefeuille. Une complexité de la description élevée signifie que les OL sont très variables dans le portefeuille ou que les profils de risque sont hétérogènes.

### 2.4.3 Portefeuille de marché, un socle commun pour la comparaison de la mutualisation entre assureurs

#### Sens et intérêt du portefeuille de marché

Les indicateurs de mutualisation (aussi bien simples qu'exotiques) proposés peuvent intervenir pour guider le choix du niveau de segmentation des primes et des risques de l'assureur. Mais l'usage majeur escompté dans ce mémoire est de mettre à disposition de la communauté des assureurs, des outils leur permettant de calculer et de comparer leur niveau de mutualisation sur un produit donné. Cette nécessité de comparer suppose alors l'existence d'un portefeuille de référence sur lequel sont construits les outils de calcul des métriques de mutualisation que sont les modèles de calcul de la prime pure et l'arbre descriptif appelé « *splitter* ». Pour un produit fixé<sup>8</sup>, ce portefeuille générique est désigné dans ce mémoire par *portefeuille de marché*. Dans l'absolu, le *portefeuille de marché* est formé de tous les assurés ayant souscrit à un contrat d'assurance quel que soit l'assureur. Mais en pratique, il peut s'agir plus simplement d'une description de la distribution des facteurs de risque pour un produit donné.

#### Le profil de risque individuel ( $BC_\Omega$ ) commun entre assureurs

Pour la calibration des modèles de calcul de la prime pure, bien que les facteurs de risque utilisés puissent être quelque peu différents d'un assureur à l'autre, les facteurs les plus importants sont généralement les mêmes puisque les assureurs mènent en permanence une veille concurrentielle pour ajuster leurs outils de tarification. Au regard des contraintes de transparence imposées par les régulateurs, les méthodes de modélisations utilisées en pratique par les assureurs ont des complexités limitées et sont alors susceptibles de produire des profils de risque individualisés semblables. Dans ce sillage, considérer pendant le calcul des mesures de mutualisation les primes pures déterminées à partir d'un modèle interne et non avec un modèle calibré sur le *portefeuille de marché*, ne devrait pas être une source d'erreurs importantes pour quantifier la mutualisation.

#### Le *splitter* doit être représentatif de tous les risques du marché

Par contre, il est essentiel que le « *splitter* » soit représentatif du portefeuille de marché. En effet, si le portefeuille dont l'assureur souhaite évaluer le niveau de mutualisation est utilisé à la fois pour l'entraînement du « *splitter* » et son encodage, le niveau de mutualisation résultant ne pourra être que maximal. Lors de la construction du « *splitter* », le choix des seuils et des variables se fait dans l'optique de répartir des assurés en plusieurs sous-groupes, aussi disparates que possible au regard du profil de risque. Encoder un portefeuille avec un arbre construit sur le même portefeuille, ne permettrait pas d'identifier les régions de l'hyper-espace (du marché)

---

8. Il importe également de supposer que pour un même produit, les assureurs commercialisent des garanties semblables (ou au moins comparables)

non peuplées par les assurés dudit portefeuille. Ceux-ci apparaîtront comme dispersés dans l'espace selon cet arbre, mais cette dispersion ne se limiterait qu'au sous-espace délimité par les caractéristiques des assurés en portefeuille. À titre d'exemple, si un portefeuille constitué exclusivement des véhicules de moins de 2 ans à la souscription est utilisé pour entraîner le *splitter*, ce dernier ne pourra pas apprécier la dispersion des véhicules de plus de 2 ans à la souscription lors de la mesure de la mutualisation sur un autre portefeuille.

Dit autrement, considérer les assurés comme recroquevillés dans une région de l'espace des assurés suppose avoir une vision d'ensemble de cet espace. Au cas contraire, un simple *zoom* dans une région spécifique de l'espace induirait en erreur en laissant croire que les assurés sont bien dispersés dans toutes les régions de l'espace comme l'illustre la figure suivante.

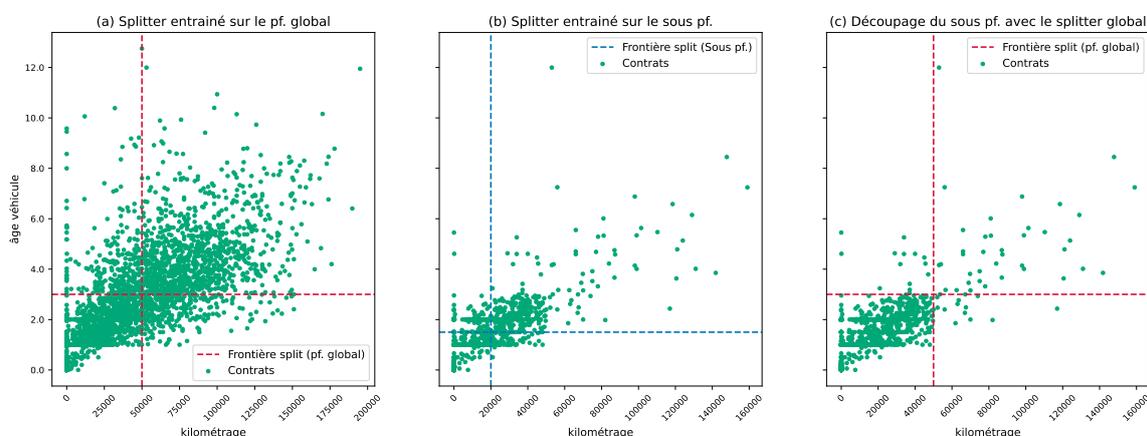


FIGURE 2.15 – Choix entre le portefeuille de marché ou le sous portefeuille pour entraîner le splitter et encoder le sous-portefeuille.

Il est donc nécessaire de disposer d'un portefeuille générique, de référence dont la distribution rendrait compte de la distribution des assurés de tout le marché. En pratique, constituer un tel portefeuille requiert de mettre en commun les informations (agrégées) des bases techniques des assureurs en concurrence sur un même marché. La constitution d'un tel portefeuille impose une certaine coopération entre les acteurs de ce marché, ce qui historiquement a été observé entre les assureurs français (par exemple dans le cadre des travaux de la Mission Risques Naturels). Par ailleurs, le régulateur dispose d'une vision plus globale du marché ; à partir de cette connaissance, il pourrait élaborer un « *splitter* » et le mettre à disposition des assureurs. Dès lors, la nécessité de connaître la distribution du marché peut être approchée au prix d'un effort collectif et coopératif entre assureurs.

Pour les indicateurs de mutualisation nécessitant le recours à un portefeuille de marché, en première approximation et en l'absence d'une base de données globale du marché, la présente étude considère le portefeuille de BNP Paribas Cardif, comme représentatif de la structure du marché de l'assurance extension de garantie pannes mécaniques en France. Des sous-portefeuilles seront extraits de ce portefeuille dit de marché, et leur niveau de mutualisation sera évalué à partir des différents indicateurs proposés.

## Conclusion du chapitre

Ce chapitre méthodologique a permis de présenter les outils statistiques, physiques et informatiques utilisés pour la construction des indicateurs de démutualisation. Ces outils sont utilisés pour mesurer la dispersion des profils de risque (burning cost le plus fin ou burning cost tarifé) en portefeuille ou l'hétérogénéité des traitements que l'assureur opère à travers l'optimisation layer. Ces outils sont la variance et l'entropie. Les indicateurs basés sur la prime sont généralisés aux caractéristiques qui déterminent ces primes à partir d'une généralisation de la variance (respectivement de l'entropie) par l'inertie (respectivement par l'entropie algorithmique). Les indices proposés se regroupent donc en indices vanilles (simples et dont le sens est plus ou moins immédiat) et en indices exotiques (dont l'interprétation et le calcul sont moins immédiats). Les indices simples imposent de connaître les primes pratiquées et le portefeuille alors que les indices exotiques permettent d'avoir une estimation du niveau de démutualisation même si la distribution du portefeuille n'est que partiellement connue et même sans la connaissance des modèles de risque (à partir du moment où un *splitter* est donné, c'est-à-dire les variables les plus importantes pour déterminer le risque).

L'implémentation de ces indicateurs et leur calcul sur des portefeuilles réels se feront aux chapitres 4 et 5. Mais en prélude à ces implémentations, le chapitre 3 présente le produit d'assurance retenu pour les illustrations ainsi que l'étude de sa base de données historique.

**DESCRIPTION DE L'EXTENSION DE GARANTIE  
AUTOMOBILE ET ÉTUDE DU PORTEFEUILLE**

**Sommaire**

---

<b>3.1</b>	<b>L'assurance extension de garantie . . . . .</b>	<b>80</b>
<b>3.2</b>	<b>Préparation des données : de la base brute à la base finale . . . . .</b>	<b>84</b>
<b>3.3</b>	<b>Analyses descriptives du portefeuille d'extension de garantie au- tomobile de BNP Paribas Cardif. . . . .</b>	<b>91</b>

---

Ce chapitre présente le produit d'extension de garantie panne mécanique qui a été retenu comme cas d'application des indicateurs proposés. Après une présentation du principe du produit ainsi que de son marché actuel, une analyse descriptive du portefeuille de BNP Paribas Cardif sur ce produit est réalisée. La fin du chapitre est consacrée à l'extraction des sous-portefeuilles pour comparer leur niveau de démutualisation.

### **3.1 L'assurance extension de garantie**

Pour les applications numériques de ce mémoire, le choix est porté sur l'extension de garantie automobile. Il s'agit d'un produit sur lequel BNP Paribas Cardif, organisme d'accueil du présent stage, dispose d'un historique suffisamment long (produit lancé depuis 2002) et des facteurs de risques susceptibles de bien expliquer la sinistralité à un niveau suffisamment fin. Le produit est commercialisé depuis 2016 par sa filiale Icare. À terme, la structure envisage le déploiement de la quantification de la mutualisation à tous ses produits de protection en particulier à l'assurance des emprunteurs.

Les couvertures du produit d'« extension de garanties » appartiennent à la branche de l'assurance non-vie et se retrouvent dans la classe de l'assurance affinitaire, ce qui leur confère un schéma particulier pour la distribution et la gestion des sinistres.

#### **3.1.1 Les produits d'assurance affinitaires**

Selon la Fédération des Garanties et Assurances affinitaires (FG2A), « est considérée comme affinitaire, toute garantie d'assurance, d'assistance ou service accessoire en lien avec l'univers d'un produit ou service présenté par un distributeur non-assureur et qui n'est pas le motif principal d'achat du client ». Ce sont des garanties proposées alors quasi-systématiquement lors de l'achat d'un produit ou d'un service. Les champs d'intervention sont très larges, ils englobent

notamment : les produits nomades (téléphonie mobile, tablettes, objets connectés, ...), les moyens de paiement, le voyage (assurance temporaire), l'extension de garantie (automobiles, produits électroménagers, ...), ... . Les garanties proposées dépendent du sous-jacent, mais elles prennent généralement une ou plusieurs des formes suivantes : l'extension de garantie constructeur, le contrat d'assistance, l'assurance-annulation, la garantie casse, vol ou encore les garanties facture. La notion de « partenaires » déjà au centre de la stratégie de BNP Paribas Cardif est très prépondérante dans le schéma de commercialisation des produits affinitaires. En effet, les partenariats sont dits en « B2B2C », Business To Business To Customer. Dans ce schéma, le client final c'est-à-dire l'acheteur du bien ne se retrouve pas à la souscription en lien direct avec l'assureur, mais seulement avec le partenaire qui lui fait la proposition d'extension de garantie. Dans ce circuit, il est possible de rajouter un TPA (third party administrator) ou « tiers administrateur » dont le rôle est de traiter les sinistres liés aux contrats d'assurance sans pour autant porter le risque qui incombe à l'assureur.

La pluralité des acteurs dans la distribution et dans la gestion des sinistres impose à l'assureur des contraintes additionnelles dans la gestion des flux de données, mais surtout dans la proposition de prix commerciaux. En fonction du modèle économique et du circuit de distribution du partenaire, l'assureur pourrait proposer une gamme plus ou moins large de prix correspondant à des couvertures différentes et à des niveaux de solidarité différents d'un portefeuille à l'autre. Notamment, si la vente a lieu sur un site de e-commerce comme c'est le cas pour certains produits nomades et que la technologie du distributeur le permet, l'assureur pourrait proposer un large panel de prix et en fonction des informations sur la vente (produit, acheteur, etc.) un système d'aide à la décision suggérerait une ou plusieurs couvertures à l'acheteur. Inversement, si la vente se fait en magasin, avec une technologie plus limitée, la grille de prix et de garantie doit être facilement accessible pour que le commercial puisse faire une proposition de garantie assez rapide et facilement. Plus généralement, l'assureur et le distributeur peuvent opter pour un produit simple avec peu de prix pour simplifier le parcours client. Cette spécificité liée au circuit de distribution pourrait donc induire un niveau de mutualisation plus ou moins important résultant non pas directement des choix de l'assureur, mais des limitations technologiques du distributeur ou de ses choix de simplicité.

Cette structure de distribution répond à une demande émanant initialement d'un constructeur et/ou captive automobile. En effet, ce partenaire souhaite proposer à ses clients une nouvelle offre, afin qu'ils puissent se prémunir contre un risque de panne non encore couvert. Pour cela, il va diffuser auprès des assureurs un appel d'offre, auquel ces derniers répondent dans un délai imparti. Les assureurs doivent alors faire une offre respectant les contraintes du partenaire, proposer le meilleur produit, des primes compétitives sur le marché, le meilleur suivi des sinistres et souscriptions, et aussi les meilleures conditions financières pour le distributeur tout en respectant un niveau convenable de valeur pour les clients.

### 3.1.2 Le principe de l'assurance extension de garantie automobile

#### Positionnement de l'extension de garantie

---

Au moment de la vente d'un véhicule neuf, deux formes de garanties peuvent être mises en avant ; les garanties légales et les garanties contractuelles. Ces couvertures sont illustrées sur la figure 3.1.

- La garantie constructeur (GC) qui est une garantie légale, d'une durée minimale de deux ans sur les véhicules neufs, est imposée par la loi pour protéger les consommateurs. Celle-ci peut s'étendre jusqu'à 7 ans selon les constructeurs. Cette garantie assure à l'acheteur que tous les frais de pièces de rechange et de main-d'œuvre seront à la charge du constructeur si un défaut de fabrication ou une panne mécanique devait se manifester sur le véhicule pour une durée de validité fixée par le constructeur. Il s'agit d'une garantie non-contractuelle ne faisant l'objet d'aucun contrat ni échange monétaire entre l'acheteur et le vendeur. L'entièreté du risque est portée par le constructeur et non par un assureur ;
- Les garanties contractuelles ou commerciales sont des garanties facultatives dont la couverture et les horizons de couverture sont définis librement par le vendeur. Selon le Code de la consommation, articles L217-15 et L217-16, cette forme de garantie : « s'entend de tout engagement contractuel d'un professionnel à l'égard du consommateur en vue du remboursement du prix d'achat, du remplacement ou de la réparation du bien ou de la prestation de tout autre service en relation avec le bien, en sus de ses obligations légales visant à garantir la conformité du bien. ». Il s'agit alors d'extension de garantie (EG) ou *Extended Warranty (EW)* en anglais. Cette garantie donne lieu à un versement de prime et à la rédaction d'un contrat d'assurance qui encadre les garanties, leurs prestations ainsi que les exclusions.

Sur le marché des véhicules d'occasion, le dealer c'est-à-dire le vendeur du véhicule peut proposer (ou non) une garantie pour se démarquer des autres vendeurs en signalant ainsi la qualité de son véhicule (et de sa préparation par le garage). Cette garantie optionnelle est généralement appelée Garantie Dealer (GD) ; elle se place entre la GC et l'EG (figure 3.1 b). La GD peut être proposée alors que le véhicule est encore sous garantie constructeur. Dans ce cas, les trois formes de garanties se succèdent les unes après les autres sur la vie du contrat (figure 3.1 c).

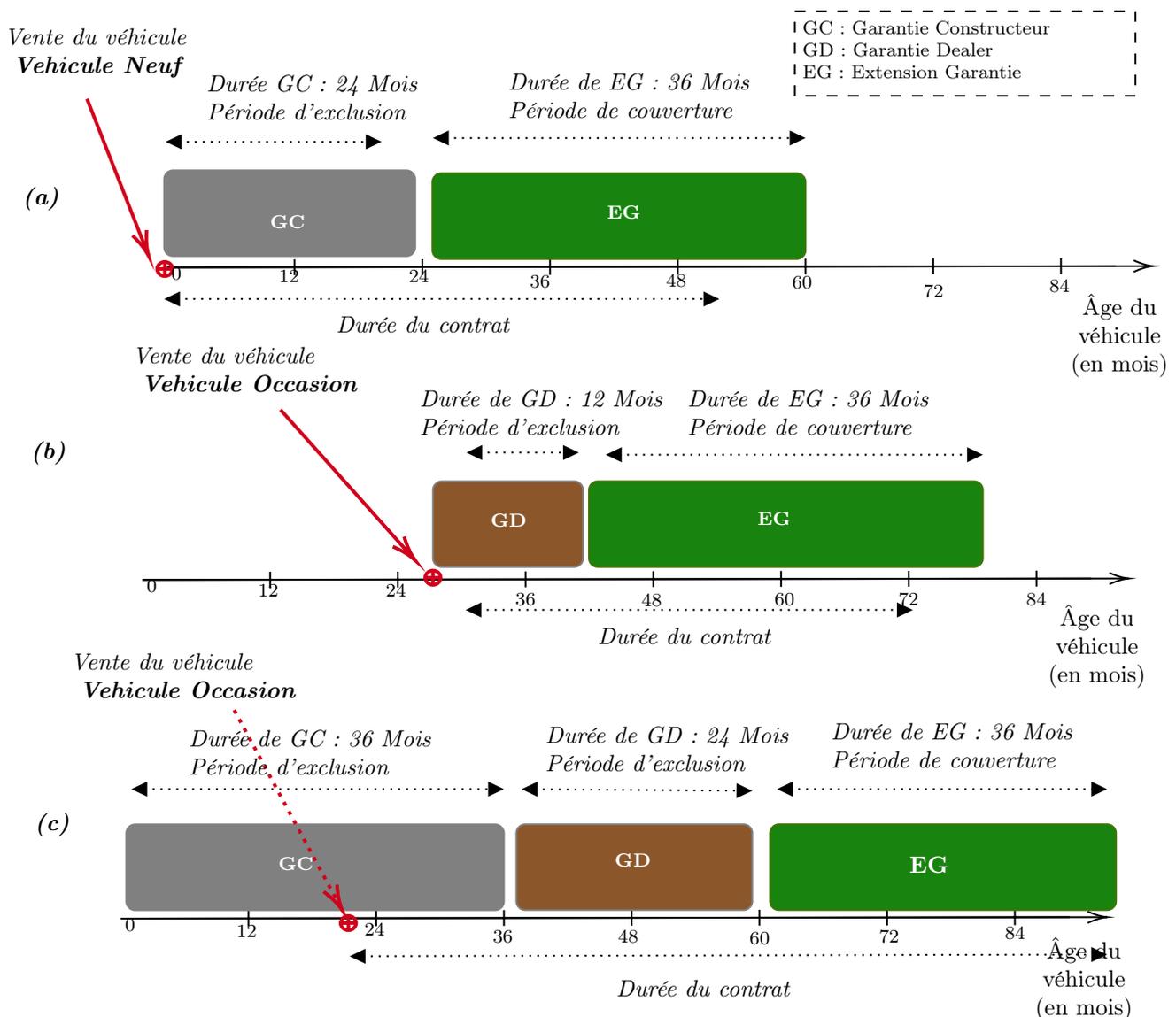


FIGURE 3.1 – Extension de garantie constructeur, en présence de garantie constructeur et/ou de garantie dealer.

À la date de souscription, la durée de la couverture est fixée contractuellement entre l'assureur et l'assuré, même si celle-ci ne débutera que plusieurs mois après la date de souscription. Il s'agit de la période de validité de l'EG. La durée qui s'écoule entre la date de souscription et le début de la couverture est une période d'exclusion appelée carence. La période de carence englobe alors la garantie constructeur (résiduelle) et éventuellement la garantie dealer. Le montant de prime pure est fixé au moment de la souscription pour toute la durée de la couverture. Son versement est généralement échelonné sur toute la vie du contrat.

### Les couvertures de la garantie pannes mécaniques

Les garanties pannes mécaniques permettent aux bénéficiaires de se couvrir contre les frais de pièces de rechange et les coûts de la main d'œuvre résultant de la remise en état de son véhicule, suite à une panne couverte par le contrat. Cependant, tous les incidents mécaniques ne sont pas pris en charge, d'une part à cause des exclusions des mauvaises pratiques pouvant

entraîner des pannes. Ces exclusions des mauvaises pratiques justifient la non prise en compte des caractéristiques de l'acheteur du véhicule comme facteurs de risque, mais exclusivement ceux du véhicule et du contrat. D'autre part, puisque les pièces couvertes par le contrat dépendent de la couverture souscrite par le client. Il existe en effet trois niveaux de couverture proposés chez BNP Paribas Cardif :

- La couverture « *Classic* » : elle se compose d'une liste restrictive de pièces, se limitant généralement aux seuls éléments essentiels du véhicule que sont le moteur, la boîte de vitesses, le différentiel et pont, la transmission, d'où son autre appellation « Moteur Boîte Pont ». Cette couverture est proposée aux véhicules d'occasion les plus âgés ou les plus kilométrés ;
- La couverture « *Expert* » : elle est privilégiée pour les véhicules d'occasion plutôt récents avec un faible kilométrage au compteur, elle couvre une liste complète d'organes comprenant l'ensemble des pièces incluses dans la couverture « Moteur Boîte Pont », ainsi que le système de direction, le circuit de refroidissement, le dispositif d'embrayage, les suspensions, et les éléments électriques et électroniques tels que la climatisation, le circuit d'alimentation, etc. ;
- La couverture « *Tout Sauf* » : elle permet de prendre en charge l'ensemble des organes du véhicule, sauf une liste définie et stipulée dans le contrat. En général, cette liste inclut entre autres la sellerie, la carrosserie, la vitrerie, les consommables et les accessoires. Elle est principalement proposée à la souscription pour les véhicules neufs ou très récents, compte tenu de l'étendue des pièces couvertes. Dans le cas où les garanties correspondraient exactement aux pièces couvertes par la garantie constructeur, celle-ci est désignée par la formule « *Bumper to Bumper* ». Dans la suite des travaux, ces deux formules sont regroupées sous la couverture « *Neuf* ».

En pratique, il existe des exceptions à ces formes génériques. À la demande du distributeur notamment, les couvertures Classic et Expert peuvent être étendues à plus de pièces que celles considérées initialement. Il convient donc de se rapporter aux contrats pour s'assurer des pièces effectivement couvertes.

## 3.2 Préparation des données : de la base brute à la base finale

### 3.2.1 Présentation des données brutes

Les données utilisées dans le cadre de ce mémoire proviennent de deux gestionnaires de sinistres différents, Gras Savoye et Icare. Les données fournies par Gras Savoye vont de janvier 2002 (au lancement du produit chez Cardif) au 30 avril 2021. Les contrats souscrits avant juin 2016 sont gérés par Gras Savoye et ceux souscrits depuis juin 2016 sont gérés par Icare.

Chacun de ces gestionnaires a mis à disposition pour l'étude deux bases de données :

- Une base de souscription : cette base fournit les informations sur les caractéristiques de chaque contrat souscrit ainsi que les informations concernant le véhicule assuré. Le format des caractéristiques n'est pas nécessairement le même d'un gestionnaire à l'autre. La variable kilométrage à la souscription par exemple est discrétisée chez Icare alors qu'elle est continue chez Gras Savoye. De plus, les deux gestionnaires n'ont pas nécessairement les mêmes variables. La base de souscription indique également le montant de la prime commerciale hors taxes et la fréquence de paiement ;
- Une base de sinistres : cette base renseigne ligne par ligne les pannes survenues sur les contrats de la base de souscription. Les sinistres sont décrits par leur date de survenance, la pièce concernée, leur statut et le montant pris en charge par l'assureur si la panne répond bien aux modalités de couverture du contrat. L'identifiant du contrat auquel est rattaché le sinistre permet de faire la jointure avec la base de souscription.

### 3.2.2 Principaux traitements préliminaires

#### i. Traitement des valeurs manquantes

Les variables *marque* et *modèle* présentent des valeurs manquantes dans les bases des deux gestionnaires qu'il a fallu imputer avant toute analyse. Chez Icare, 10,33 % des contrats n'avaient pas de marque renseignée et 35,31 % n'avaient pas de modèle. Chez Gras savoye, ces pourcentages sont respectivement de 1, 01 % et de 4,69 %. Pour chaque gestionnaire, les marques manquantes chez un distributeur donné ont été remplacées par la marque la plus souscrite par ce distributeur (la marque modale par distributeur). S'agissant du modèle, la même stratégie a été retenue en remplaçant le distributeur par la marque (imputation par le modèle modal pour une marque donnée).

#### ii. Harmonisation des modalités entre gestionnaires de sinistres

À partir de la désignation du produit, la couverture (FORMULE) est créée et harmonisée entre les deux gestionnaires. Celle-ci comporte finalement trois modalités : NEUF, EXPERT et CLASSIC. La variable type de véhicule est créée dans la base Icare à partir de la *désignation technique formule*. Celle-ci prend deux modalités à savoir camping Car et voiture. Les motos sont très peu représentées (moins de 0,01 % des contrats) et sont donc écartées de l'analyse. De même, pour la motorisation, les véhicules électriques et à GPL sont écartés et seuls les véhicules à moteur essence et diesel sont conservés. Puisque le kilométrage à la souscription chez Icare est renseigné en tranche, cette variable a été discrétisée à la même granularité dans la base de Gras Savoye.

#### iii. Traitement des variables catégorielles à forte cardinalité

Au regard de la multitude des marques présentes dans la base (plus de 140) et de leur inégale répartition au sein du portefeuille (figure 5a en annexe), seules certaines marques ont été traitées explicitement et les autres marques étant très faiblement représentées (ou provenant des erreurs de saisie qui n'ont pas pu faire l'objet d'une correction) ont été regroupées dans une modalité AUTRES-MARQUES. Le seuil d'identification des marques les plus représentées a été déterminé sur la base de la stabilité de la fréquence de sinistre par marque. Spécifiquement, les marques ont été triées suivant leur importance dans le portefeuille. Puis, pour chaque marque, les variations de fréquence de sinistre d'une année sur l'autre ont été calculées entre 2010 et 2020. Enfin, un accroissement moyen par marque a été calculé sur cette période. Il en ressort que les marques les plus représentées ont des fréquences de pannes annuelles relativement stables. Au-delà de la 40<sup>ème</sup> marque la plus importante en portefeuille, la volatilité de la fréquence des sinistres d'une année sur l'autre explose (figure 5b en annexe). Ces 40 marques ont au moins 250 contrats chacune et représentent 97,3 % des contrats en 2020.

Au même titre que les marques, les modèles de véhicules dans la base sont très nombreux (plus de 7 218), or pour la question de la démutualisation, il semble pertinent d'utiliser cette variable pour la construction du tarif; son exclusion systématique n'est donc pas une option envisageable. Le traitement appliqué à cette variable pour la rendre exploitable est le suivant : pour la marque  $marque_i$  prise parmi les marques conservées explicitement, les modèles qui représentent 80 % des contrats de cette marque sont retenus et recodés  $marque_{i,mod_1}, marque_{i,mod_2}, \dots$ . Les modèles restants sont regroupés sous la modalité  $marque_{i,mod_{autre}}$ . Cette approche permet de créer une nouvelle variable MARQUE  $\times$  MODÈLE dont l'apport en informations pour appréhender la sinistralité sera testé lors de la phase de modélisation de la prime pure.

#### iv. Calcul de nouvelles variables

- L'exposition annuelle des contrats actifs a été calculée à partir des dates de début et de fin du risque effectives sur une année donnée ;
- L'âge du véhicule au moment de la souscription est calculé comme la différence entre la date de souscription et la date de première mise en circulation ; toutefois, cette variable ne représente qu'imparfaitement l'âge du véhicule au début du risque du fait de l'existence éventuelle d'une période de carence (voir figure 3.2). La variable « âge du véhicule au début du risque » est donc créée pour tenir compte de cette spécificité des garanties affinitaires. Elle est calculée comme l'écart entre la date de mise en circulation du véhicule et la date de début du risque sur l'exercice. Il convient de souligner que de même que l'âge à la souscription, le kilométrage à la souscription est biaisé par l'existence de la carence, mais aucun traitement particulier n'est apporté ici à cette variable puisque le kilométrage parcouru par le véhicule entre la souscription et la date d'entrée en risque n'est pas connu ;

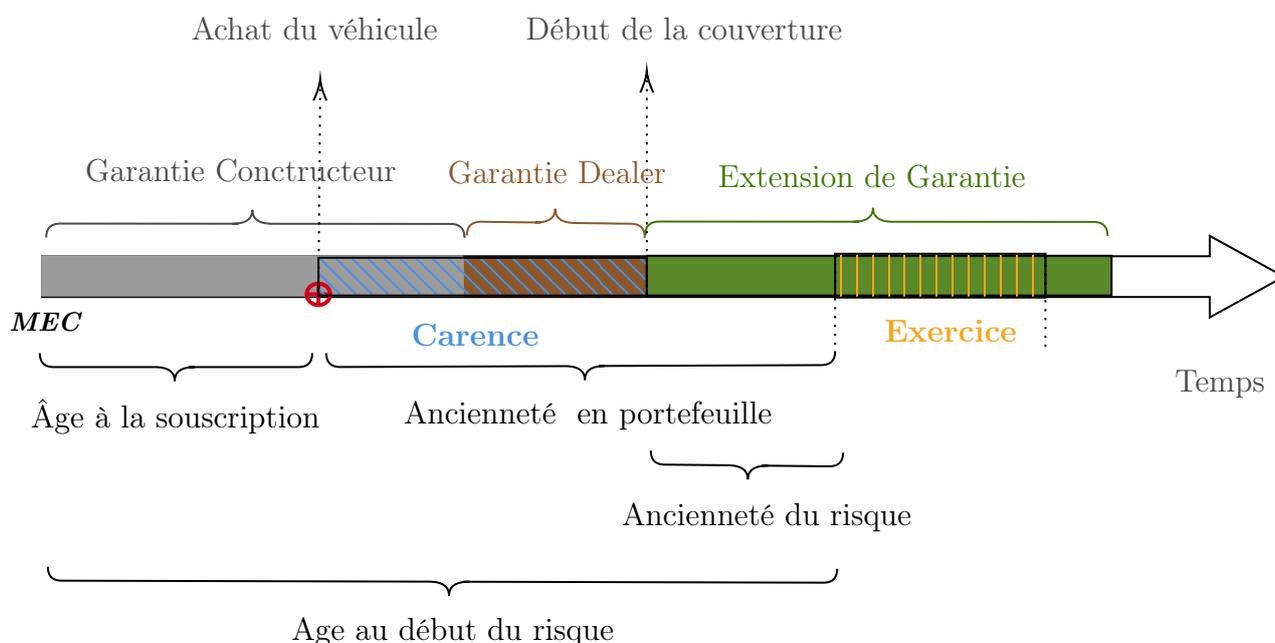


FIGURE 3.2 – Présence d'une période de carence en assurance EC et création de nouvelles variables.

- La variable « ancienneté du contrat en portefeuille » au début de chaque exercice ou au début d'un nouveau risque a été calculée comme l'écart entre la date de souscription et la date de début du risque sur l'exercice. Cette variable donne une indication sur l'usure du véhicule au début de chaque exercice.

#### v. Actualisation des montants pris en charge

Les bases de données recouvrent un long historique, ce qui permet d'observer des fréquences stables sur des véhicules à la maille marque  $\times$  modèle, mais cela impose de prendre en compte l'inflation sur le prix des pièces et sur le coût de la main d'œuvre à la réparation. En effet, sur les 11 dernières années, le prix des pièces détachées a augmenté de 4 % en moyenne chaque année et les coûts de la main d'œuvre de 3 % sur la même période (tableau 3.1).

TABLEAU 3.1 – Évolution du prix des pièces détachées et du coût de la main d'œuvre entre 2009 et 2020 (base 100 = 2009).

Année (après Q4)		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Pièces détachées	Accrois.	3,9%	4,7%	2,2%	0,4%	1,7%	2,0%	3,0%	3,5%	6,4%	6,5%	8,1%
	Prix	103,9	108,8	111,2	111,6	113,5	115,8	119,3	123,4	131,3	139,9	151,2
Main d'œuvre	Accrois.	3,2%	4,1%	3,7%	2,6%	3,2%	2,7%	2,5%	2,7%	2,5%	2,5%	3,2%
	Coût	103,2	107,4	111,4	114,3	117,9	121,1	124,2	127,5	130,7	133,9	138,2

Source : Site de l'association Sécurité et Réparation Automobiles (SRA), consulté en août 2021.

L'actualisation des coûts des sinistres peut se faire de plusieurs manières :

- Ne pas actualiser les sinistres en amont des modélisations de la prime pure, mais introduire une variable « année génération » ou « année d'exercice » dans les modèles, afin de capter

l'effet du temps sur le coût et la fréquence des sinistres. Cette méthode a l'avantage d'être facile à mettre en œuvre, mais elle rajoute une variable explicative pour la construction de la prime pure ;

- Fixer une année de référence et calculer tous les coûts des sinistres de la base en *as-if*, c'est-à-dire actualiser les montants des sinistres en amont de la construction de la prime pure. L'hypothèse implicite ici est que d'une année à l'autre, les variations du coût moyen des sinistres sont imputables à l'inflation. Cette approche impose de disposer d'un historique de l'évolution des coûts de prise en charge des sinistres sur le marché de l'automobile sur le périmètre géographique de l'étude, la France dans ce cas. Il est également nécessaire de faire des projections de cette évolution sur les années à venir afin de construire un tarif pour les contrats qui seront actifs ultérieurement.

Dans le cadre de cette étude, l'approche retenue est la première. En effet, cette approche permet de prendre en compte la structure du portefeuille (éventuellement différente de celle du marché) et ses changements dans le temps lors de l'actualisation des montants. De plus, l'inflation sur les pièces n'étant pas constante d'une marque à l'autre, considérer la deuxième approche alourdirait les pré-traitements puisqu'il faudrait récupérer les facteurs de corrections, marque par marque avant de les appliquer à notre base. Enfin, il est observé sur le marché que la fréquence des pannes a une tendance baissière au regard du temps, car les constructeurs dans les générations des modèles tendent à les améliorer. Il semble alors judicieux de prendre en compte l'année lors de la modélisation de la fréquence, ce qui n'est pas le cas avec la seconde approche.

#### **vi. Traitement des sinistres extrêmes par écrêtement**

Une analyse de la distribution des coûts des sinistres (figure 3.4a) fait ressortir des queues lourdes, il s'agit des sinistres graves ou extrêmes. Il est nécessaire de traiter particulièrement ces sinistres puisqu'ils ne sont pas suffisamment nombreux, mais sont de grandes ampleurs. Dit autrement, leur faible fréquence empêche d'estimer correctement les queues de distribution, mais leur forte sévérité impose de ne pas les écarter de l'analyse. Ils sont alors traités par écrêtement. La technique d'écrêtement des sinistres extrêmes consiste à ramener les coûts de ces sinistres graves au seuil d'identification des sinistres graves déterminé et, à répartir les coûts excédentaires sur l'ensemble des sinistres du portefeuille. Une telle pratique est assimilable à une forme de mutualisation puisque les coûts (historiques) des sinistres les moins graves sont légèrement surévalués pour réduire les coûts des sinistres les plus graves.

Le problème majeur pour procéder à l'écrêtement des sinistres extrêmes est la détermination du seuil d'écrêtement des sinistres graves. La théorie des valeurs extrêmes fournit un cadre propice pour le calcul de ce seuil (Robert, 2020). De façon simplifiée, le seuil d'identification des valeurs extrêmes d'une variable aléatoire peut se faire au moyen de la *mean excess plot* (ou courbe de dépassement moyen). Cette démarche repose sur le résultat suivant : si une variable

aléatoire  $X \sim GPG(\beta, \xi)$  c'est-à-dire une loi de Pareto généralisée de paramètres d'échelle  $\beta > 0$  et de forme  $\xi$ , alors,

$$X - u | X > u \sim GPG(\beta + \xi u, \xi) \text{ et } \mathbb{E}(X - u | X > u) = \frac{\beta + \xi u}{1 - \xi}$$

Ce résultat suggère que si la variable aléatoire  $X$  suit une loi GPD, la fonction de dépassement moyen, devrait être linéaire en fonction du seuil de dépassement  $u$ . En pratique, la démarche consiste à évaluer l'estimation empirique de la fonction de dépassement moyen sur les coûts des sinistres au regard de plusieurs seuils candidats croissants  $u_i$ . Il faut ensuite identifier le seuil au-delà duquel la linéarité est rompue. Au-delà de ce seuil, il n'y a plus assez de sinistres pour estimer avec robustesse cette espérance et l'évolution de la *mean excess plot* devient erratique.

### 3.2.3 Présentation des données finales

Un travail de restructuration de la base de données a été réalisé afin d'obtenir une base en « vision exercice ». La base sous sa forme finale renseigne alors pour une année donnée, l'ensemble des contrats effectivement couverts sur cet exercice. Les contrats sont décrits par des caractéristiques calculées ou obtenues grâce aux questionnaires des sinistres. Les variables constituant la base finale sont présentées dans le tableau 3.2. Toutes ces variables ne seront pas nécessairement utilisées pour la construction de la prime pure puisque certaines sont fortement corrélées entre elles (figures 3.7 et 3.7a).

TABLEAU 3.2 – Récapitulatif des variables de l'étude après prétraitement.

Code variable	Type	Description	Valeurs/Modalités
ID_CONTRAT	object	Identifiant du contrat	-
ANNEE	int64	Année d'exercice	2010, ..., 2020
AGE_VEH_SOUS	float64	Âge du véhicule à la souscription (en année)	$\mathbf{R}_+$
CARENCE	float64	Durée de la carence	0, 6M, 12M, ..., 36M
FORMULE	object	Formule du contrat	NEUF, EXPERT, CLASSIC
GESTIONNAIRE	object	Gestionnaire des sinistres	Icare, Gras Savoye
MOTORISATION	object	Motorisation du véhicule	ESSENCE, DIESEL
PUISSANCE	float64	Puissance fiscale du véhicule	$\mathbf{N}^*$
TRANCHE_KM_SOUS	object	Tranche de kilométrage du véhicule à la souscription	Intervalle de 10k km
TYPE_VEH	object	Type de véhicule	Camping-Car, Auto
AGE_VEH_RISK	float64	Âge du véhicule au début du risque de l'exercice (en année)	$\mathbf{R}_+$
ANC_VEH_PF	float64	Ancienneté du véhicule en portefeuille (en année)	$\mathbf{R}_+$
ANC_RISK	float64	Temps écoulé depuis le début de la période de couverture du contrat (en année)	$\mathbf{R}_+$
MARQUE	object	Constructeur du véhicule	RENAULT, VOLVO, FIAT, ...
MARQUE_MODELE	object	Constructeur et modèle du véhicule	RENAULT_LAGUNA, RENAULT_CLIO, ...
EXPO	float64	Exposition du contrat sur l'exercice	$]0, 1]$
NB_SINISTRE	int64	Nombre de sinistre pris en charge sur l'exercice	$\mathbf{N}$
TOT_MNT_SINISTRE	float64	Montant total des sinistres pris en charge sur le contrat sur l'exercice	$\mathbf{R}_+$
LIST_MNT_SINISTRE	object	Liste des montants pris en charge sur le contrat	$[S_1, S_2, S_2, \dots]$

### Remarque : Suppression des valeurs sur les axes

Pour lever la confidentialité sur ce mémoire afin d'en permettre un plus large accès, les échelles de certains graphiques ainsi que certains chiffres confidentiels mais non indispensables sont délibérément omis. Toutefois, cela n'impacte pas la compréhension des analyses ni leurs interprétations.

### 3.3 Analyses descriptives du portefeuille d'extension de garantie automobile de BNP Paribas Cardif

#### 3.3.1 Étude de la dynamique de la sinistralité historique

Le suivi de la sinistralité du portefeuille d'extension de garantie constructeur automobile se fait à travers la fréquence et la sévérité des sinistres. Les évolutions des fréquences et des coûts moyens pris en charge suivant les exercices sont présentées sur la figure 3.3. Sur les premières années du produit, l'exposition totale des contrats croît du fait de l'existence de la période de carence et du développement du produit, puis elle se stabilise à partir de 2008. La fréquence assez importante en début d'exercice, atteint sa plus grande valeur en 2005. Ensuite, la fréquence baisse continuellement sur le reste de la période à l'exception de l'année 2021, mais celle-ci est incomplète. Cette baisse des fréquences s'explique par les efforts des constructeurs à améliorer leurs véhicules d'une génération à l'autre et des progrès technologiques qui assurent une meilleure qualité de construction. La dynamique inverse est observée pour le coût moyen qui est à la hausse depuis 2007. Cette hausse est essentiellement imputable à l'inflation sur les pièces détachées et les coûts de la main d'œuvre.

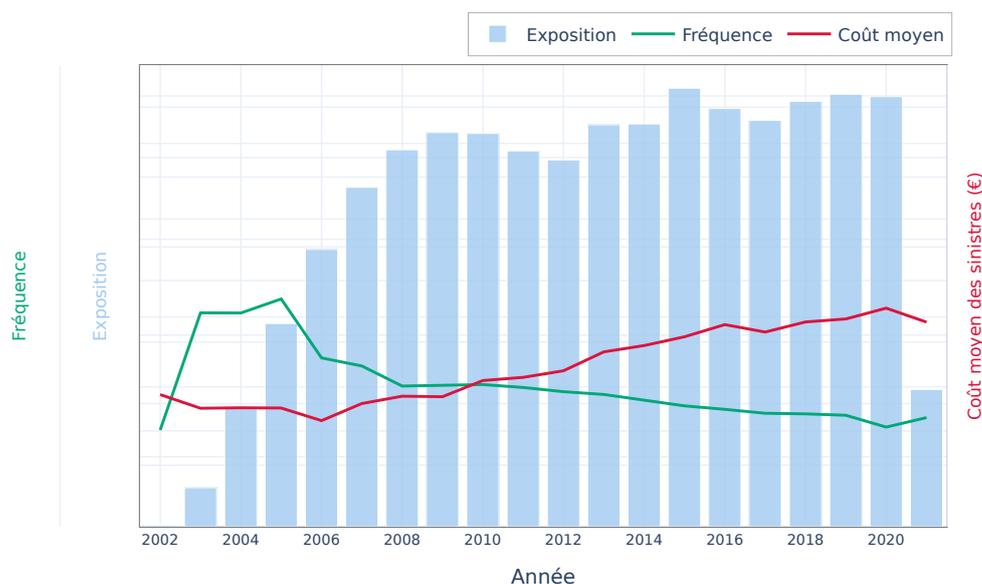


FIGURE 3.3 – Évolution de la fréquence et du coût moyen des sinistres du portefeuille global entre 2002 et 2021 (vision au 30 avril 2021).

Cette dynamique du montant moyen pris en charge cache une volatilité importante des montants des sinistres au sein d'une année. En effet, l'analyse de la distribution des coûts sur l'exercice 2019 révèle une queue de distribution très étalée à droite témoignant alors de l'existence de quelques sinistres dont les coûts sont bien plus élevés que ceux des sinistres de la base (figure 3.4a). La fréquence de survenance des sinistres graves reste très faible, mais du fait de leur forte sévérité, un traitement particulier doit être adopté. Celui adopté dans le

cadre de ce mémoire est l'écrêtement tel que présenté précédemment. Pour l'année 2019, le seuil d'écrêtement tiré de l'analyse de la figure 3.4b est de 3 800 € correspondant au quantile d'ordre 98,7 % des coûts de sinistres. Concrètement, ces sinistres sont ramenés à 3 800 € et la charge de sinistre totale restante est distribuée uniformément sur l'ensemble des sinistres de la base. Au regard de la tendance haussière des coûts PEC, des seuils d'écrêtement sont déterminés de même que pour 2019, année après année depuis 2010 (figure 10 en annexes). Les seuils identifiés ont effectivement une tendance à la hausse et leur ordre de quantile associé est au voisinage de 99 % (tableau 4 en annexes).

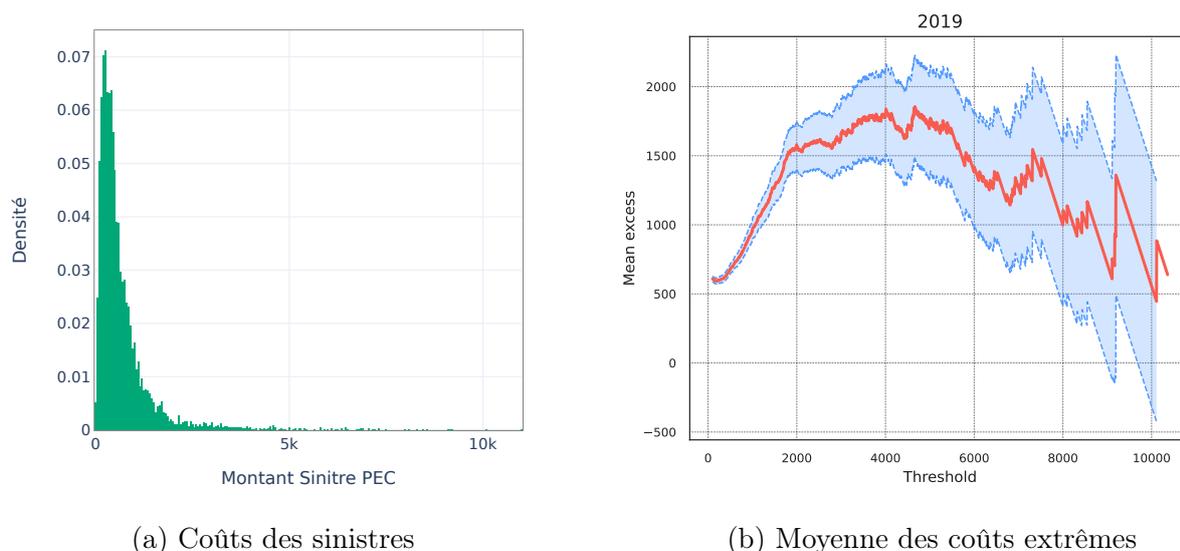


FIGURE 3.4 – Distribution des coûts de sinistre en 2019 et identification du seuil de sinistre extrême par *mean excess plot*.

### 3.3.2 Facteurs de risque et structure du portefeuille

La description des caractéristiques se fait suivant plusieurs axes. Un intérêt est d'abord porté sur la distribution marginale des facteurs de risque et leur évolution sur l'historique de la base. Ceci dans l'optique de repérer d'éventuels changements structurels importants du portefeuille qui viendraient biaiser les modèles de calcul de prime pure. Une attention particulière est accordée à l'année 2016 correspondant au changement de gestionnaires des sinistres. Ensuite, les fréquences empiriques et la distribution des coûts des sinistres sont comparées entre différentes modalités ou niveaux de facteurs de risque. Seule l'année 2019 est utilisée pour cette description des facteurs de risques vis-à-vis de la fréquence et de la sévérité des sinistres.

#### Évolution de la structure du portefeuille

Au regard de la distribution du portefeuille suivant la formule souscrite (figure 3.5), il en ressort que la formule la plus souvent souscrite par les assurés est la formule Expert avec près de 2/3 de l'exposition annuelle. La formule NEUF se situe à la deuxième place, avec

environ le quart des risques actifs sur un exercice. Toutefois entre 2013 et 2020, la couverture CLASSIC a gagné en importance dans le portefeuille passant de 5 % en 2013 des polices à 10 % en 2020. Cet accroissement s'aligne avec le vieillissement du parc automobile en France et donc l'accroissement des véhicules d'occasion qui souscrivent majoritairement à la couverture CLASSIC.

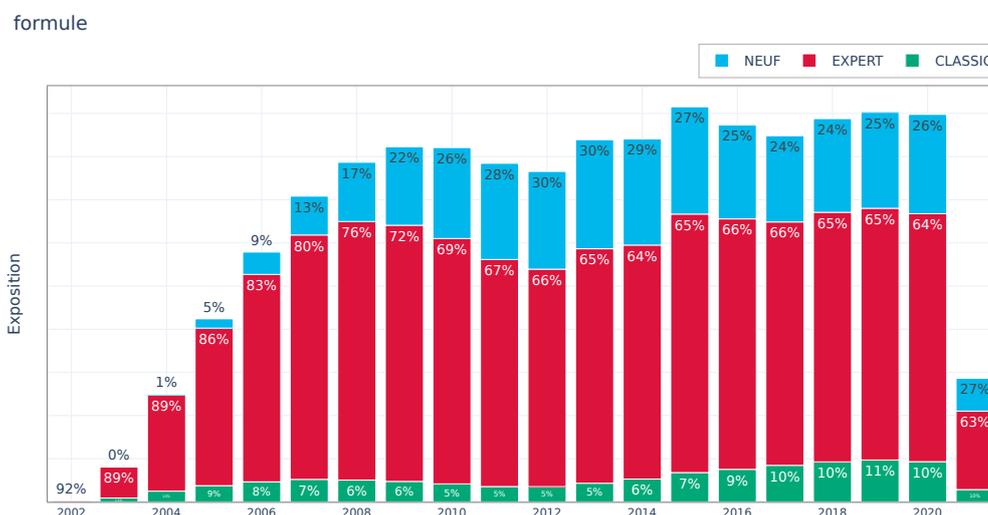


FIGURE 3.5 – Évolution de la distribution du portefeuille suivant la formule entre 2002 et 2021.

De façon plus sommaire, la structure du portefeuille et sa dynamique sont synthétisées comme suit en se référant aux figures 6 à 9 en annexe B :

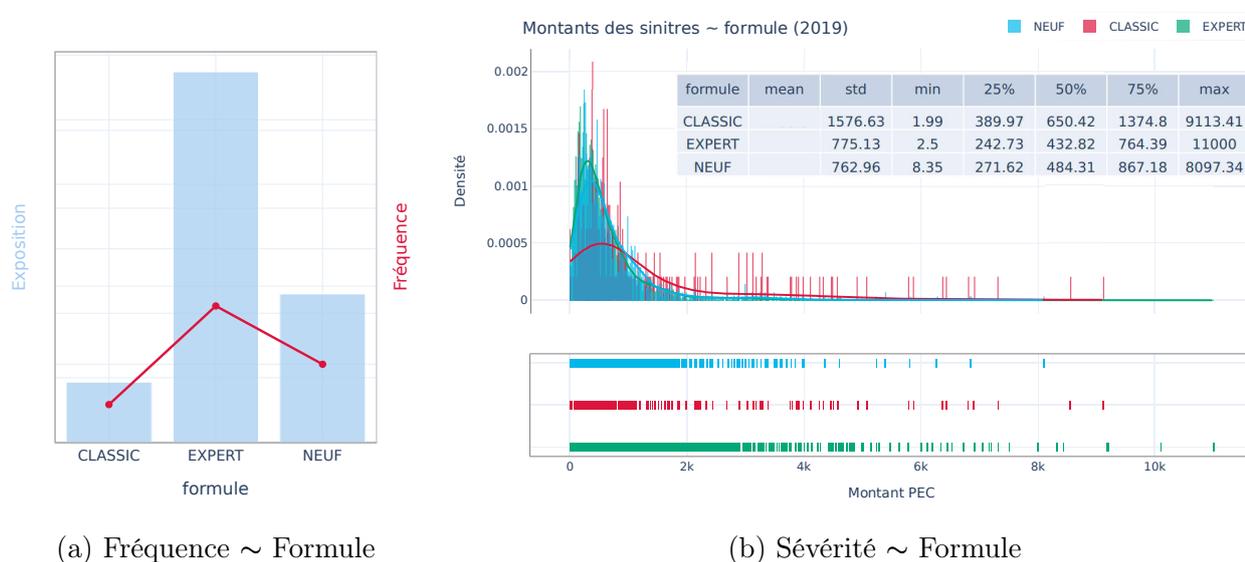
- S'agissant de la motorisation, le portefeuille est composé d'environ 80 % de véhicules à moteur Diesel et de 20 % à Essence. Une augmentation lente du poids des véhicules à ESSENCE est observée à partir de 2016 ;
- La distribution marginale des marques de véhicules les plus prépondérantes dans le portefeuille révèle de légers changements sur l'historique. Certaines marques gagnent de l'importance au détriment des autres. La diversification du portefeuille à d'autres marques peut également être soulignée, puisque la part des marques regroupées sous « autres marques » augmente sur toute la période de l'étude ;
- La part des véhicules couverts sans délai de carence augmente entre 2011 et 2020, passant de 29 % à 38 %. Ceci témoigne d'un vieillissement du parc automobile en portefeuille puisque les véhicules sans carence sont essentiellement des véhicules d'occasion. La durée moyenne de carence (sur les véhicules avec carence) reste stable et se situe autour de 1 an et 5 mois chaque année ;
- La part annuelle des camping-cars dans le portefeuille augmente rapidement, passant de 11 % en 2010 à 28 % en 2020.

Ces analyses de dynamique révèlent quelques changements dans la structure du portefeuille, mais leur importance reste modérée. L'inclusion de l'année d'exercice comme variable dans la

partie modélisation devrait aider à prendre en considération ces évolutions de structure. Sur l'historique disponible, la plage 2010 - 2020 est retenue pour la construction de la prime pure et le calcul des indicateurs de mutualisation.

### Fréquence et sévérité des sinistres suivant les facteurs de risque

La figure 3.6 ci contre, au même titre que les figures 11, 12 et 13 en annexes permettent d'apprécier la sinistralité sur les contrats couverts en 2019 au regard des caractéristiques. Une synthèse des observations est donnée ici. Au regard de la formule, la fréquence la plus élevée est enregistrée pour les contrats sous la couverture « EXPERT » ; viennent ensuite les assurés ayant souscrit à la couverture « NEUF » avec deux fois moins de fréquence, puis la couverture « CLASSIC ». Inversement, les coûts sont plus élevés pour la couverture « CLASSIC » qui a le moins de fréquence. Ils restent très comparables pour les deux autres couvertures.



(a) Fréquence ~ Formule

(b) Sévérité ~ Formule

FIGURE 3.6 – Fréquence et sévérité des sinistres sur le portefeuille en 2019 suivant la formule.

Les fréquences observées sur les véhicules à essence sont bien moins importantes que pour les diesels, mais l'analyse des coûts par classe ne révèle pas de différences majeures entre ces deux catégories de contrats.

Aussi bien en fréquence qu'en coût de sinistre, le type de véhicule ne semble pas présenter de différence notable.

Les fréquences et les coûts de sinistres sont très volatils d'une marque à l'autre. Ceci résulte en partie des différences entre les niveaux d'exposition des marques, mais ces différences ne traduisent pas exclusivement cet effet de taille, puisqu'un nombre minimal de contrats a été imposé pour traiter une marque distinctement.

Ces descriptions croisées apportent les premières appréhensions sur les facteurs de risque pertinents pour construire la prime pure. Seulement, il serait préférable de déterminer la pertinence de ces facteurs en se plaçant dans un contexte *toutes choses égales par ailleurs*. Ceci

nécessite de faire appel à un modèle qui analyserait la pertinence d'un facteur de risque en prenant en compte les autres facteurs. En prélude à cette modélisation, un regard est porté sur les corrélations entre les variables explicatives pour identifier les paires de variables qui ne peuvent pas être incluses simultanément dans le modèle.

### **Analyse des corrélations entre les facteurs de risque**

---

Les liaisons entre les variables quantitatives (réelles et entières) sont étudiées à partir du coefficient de corrélation linéaire de Pearson. Ce coefficient noté  $r$  compris entre  $-1$  et  $+1$  indique s'il existe une relation linéaire forte à mesure que  $|r|$  se rapproche de 1. La relation est positive si  $r > 0$  et négative si  $r < 0$ . Calculé sur les variables de la base, (figure 3.7a) ce coefficient met en exergue des relations fortes entre les paires de variables liées aux âges du véhicule calculés à différents instants. La variable âge à la souscription est écartée de la modélisation car en plus d'être biaisée comme facteur de risque, elle est corrélée à la carence et à sa version corrigée de l'âge du véhicule au début du risque de l'exercice. Par construction des variables ancienneté du risque et ancienneté en portefeuille, il n'est pas possible de les inclure toutes les deux dans la modélisation en plus de la carence<sup>1</sup>. De plus, l'ancienneté du risque est corrélée à l'âge du véhicule au début du risque tout comme à l'ancienneté en portefeuille. Cette variable est donc exclue *a priori* de la construction de la prime pure.

Pour les variables qualitatives, l'intensité de la liaison est mesurée à partir du V de Cramer qui est construit sur la statistique du chi-2. La valeur minimale de zéro s'interprète comme une absence de relation alors que la valeur maximale de 1 traduit une dépendance exacte entre les variables en présence. La figure 3.7b permet de repérer une dépendance forte entre le type de véhicule et la marque, celle-ci est d'autant plus prononcée que la marque est remplacée par la variable `marque_modèle`. L'intérêt à préserver ces deux variables sera étudié lors de la modélisation. Une dépendance moins forte, mais tout de même notable, est observée entre la formule et la carence. Ceci découle de la présence systématique de carence sur les véhicules neufs auxquels la couverture « NEUF » est proposée. La sélection éventuelle entre ces deux variables se fera lors de la modélisation de façon itérative.

### **3.3.3 Base d'apprentissage, base de test et base de comparaison des indicateurs de mutualisation**

En prélude à la modélisation de la prime pure et à la construction des indicateurs, il convient de construire des échantillons pour construire, tester et comparer les métriques de démutualisation.

Les indicateurs de mutualisation proposés au chapitre 2 supposent connus et estimés les modèles de calculs de primes pures incluant plus ou moins des facteurs de risque. En effet, le modèle de calcul du burning cost est calibré et mis en production pour quelques années avant

---

1.  $ANC\_VEH\_PF = CARENCE + ANC\_RISK$

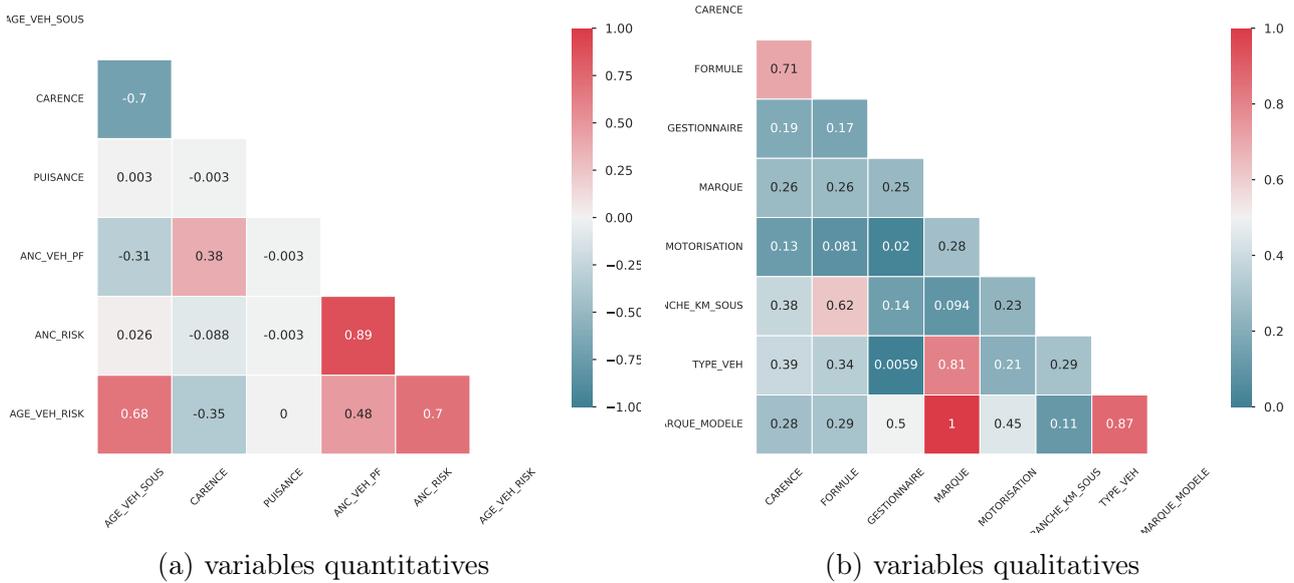


FIGURE 3.7 – Corrélations de Pearson (variables quantitatives) et V de cramer (variables qualitatives) entre les variables tarifaires.

de subir des ajustements. Dans ce sillage, l'historique entre 2010 et 2019 servira au calibrage des modèles. L'année 2020 servira quant à elle pour le calcul des indicateurs et leur comparaison sur des sous-portefeuilles extraits sur la même année.

Par ailleurs, pour un groupe de facteurs de risque donné, construire un prédicteur du burning cost passe par l'entraînement et la comparaison des performances de plusieurs modèles ou familles de modèles différents. Il est donc nécessaire de disposer d'une base test différente de la base d'entraînement. Précisément, 80 % des données entre 2010 et 2019 serviront à l'entraînement et les 20 % restants à la comparaison des modèles. Le choix des hyper-paramètres se fera par validation croisée sur l'échantillon d'entraînement.

La figure 3.8 ci-contre présente le découpage de la base de données pour l'entraînement, le test et la comparaison.

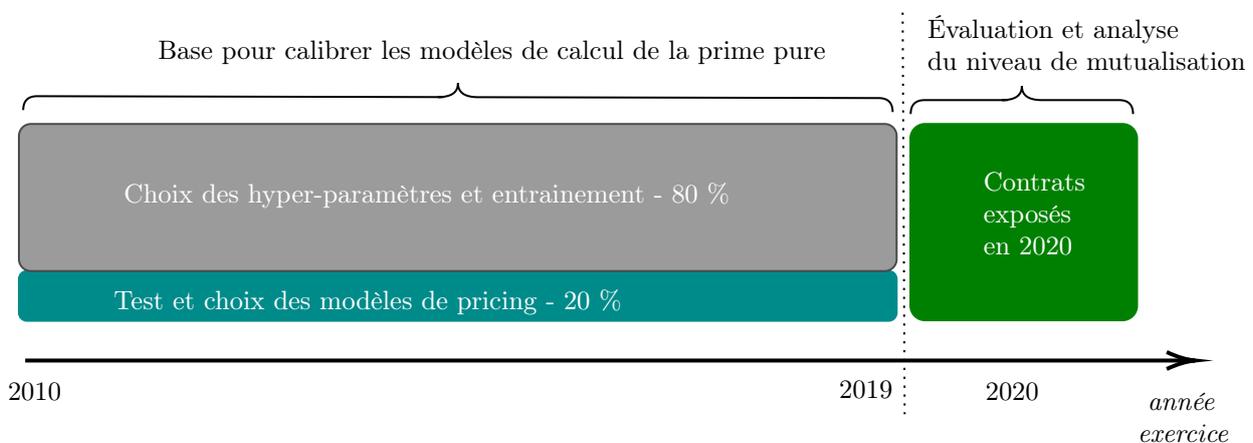


FIGURE 3.8 – Base d'apprentissage, base test et base de calcul/comparaison des indicateurs de mutualisation.

## Conclusion du chapitre

Il était question dans ce chapitre de présenter le produit d'extension de garantie automobile. Ce produit revêt un intérêt pour les illustrations de cette étude au vu de la profondeur de l'historique disponible, mais surtout de l'hyper-segmentation atteignable à partir des variables tarifaires ; en considérant des primes pures pour des contrats à la maille *marque x modèle x génération* par exemple. La constitution de la base et la description du portefeuille ont également été présentées. La dynamique de la sinistralité révèle une tendance haussière (resp. baissière) du coût moyen (resp. de la fréquence) du portefeuille, conduisant alors à inclure l'année d'exercice comme facteur de risque. La structure du portefeuille n'a pas connu une déviation majeure sur l'historique, en particulier entre 2010 et 2020. Cette plage retenue pour l'étude a servi à constituer une base d'apprentissage, une base test et des portefeuilles de comparaison des indicateurs de mutualisation.

La suite de ce mémoire est consacrée à la modélisation du profil de risque ainsi qu'à la mesure et l'analyse des effets de la démutualisation sur un portefeuille d'assurance.

## ÉVALUATION ET ANALYSE DE LA MUTUALISATION À PARTIR DES INDICATEURS SIMPLES

### Sommaire

---

4.1 Calibration d'un modèle de calcul de la prime pure .....	98
4.2 Applications numériques des indices de mutualisation simples ....	108
4.3 Utilisations pratiques des indices de mutualisation basés sur la prime .....	117

---

Ce chapitre est consacré à l'implémentation et la comparaison des indicateurs de mutualisation simples dont l'approche générique est de mesurer la mutualisation d'un portefeuille à travers la dispersion des primes tarifées ou la concentration des Optimisation Layers (OL). Mais en prélude au calcul de ces indicateurs, la construction des modèles de tarification qui permettront de calculer le burning cost est présentée. Ce chapitre s'achève sur la comparaison de ces indicateurs en considérant plusieurs assureurs qui appliqueraient des stratégies de tarification différentes. Cette comparaison entre assureurs permet également d'analyser les effets de la course à l'hyper-segmentation sur le marché et les assureurs.

### 4.1 Calibration d'un modèle de calcul de la prime pure

Déterminer la prime pure ou le profil de risque par police revient à prédire la charge de sinistres attendue par contrat pour une période donnée, en occurrence un exercice. L'approche retenue est la régression Tweedie implémentée à travers le GLM et l'algorithme de machine learning LightGBM sur Python. L'utilisation d'un algorithme de machine learning pour modéliser le burning cost permet d'apprécier les performances et de jauger les effets des variables résultants de la méthode la plus utilisée en pratique qu'est le GLM. Outre les algorithmes mis en comparaison, l'intérêt de la variable *modèle de véhicule* pour la détermination du burning cost est également étudié. Pour chaque modèle étudié, un modèle dit *par défaut* est estimé (il s'agit du modèle utilisant les paramètres incorporés directement dans le module d'estimation utilisé), ensuite les hyper-paramètres optimaux sont déterminés avant une nouvelle estimation du modèle, celui-ci est dit *tuné*. Le tableau suivant résume l'ensemble des modèles estimés. Les variables communes à tous les modèles sont : ANNEE, AGE\_VEH\_RISK, CARENCE, FORMULE, MOTORISATION, PUISSANCE, TRANCHE\_KM\_SOUS, TYPE\_VEH, ANC\_VEH\_PF.

TABLEAU 4.1 – Liste des modèles implémentés pour le calcul du profil de risque individuel.

	Modèle			
	GLM		LightGBM	
	<i>Défaut</i>	<i>Tuné</i>	<i>Défaut</i>	<i>Tuné</i>
<b>Variables + MARQUE</b>	GLM_Tweedie_zero	GLM_Tweedie_reg	GBM_Tweedie_default	GBM_Tweedie_tuned
<b>Variables + MODÈLE</b>	GLM_Tweedie_Mod_zero	GLM_Tweedie_Mod_reg	GBM_Tweedie_Mod_default	GBM_Tweedie_Mod_tuned

La base de données est découpée en base apprentissage et en base test à raison d'une partition de 80-20.

#### 4.1.1 Modélisation de la prime pure à partir d'un GLM

Pour calibrer les paramètres d'une régression GLM Tweedie, il faut au préalable déterminer l'hyper-paramètre de puissance  $p$  de la loi de Tweedie qui est partagé par tous les profils de risque. Sa détermination est faite de façon itérative par maximisation de la vraisemblance. Pour un ensemble de 50 valeurs candidates de  $p$  prises sur l'intervalle  $]1, 2[$  avec un pas de 0,02, des modèles GLM sont estimés et leurs AIC sont calculés sur l'échantillon d'apprentissage. La valeur optimale de  $p = 1,17$  est retenue puisque c'est pour cette valeur que l'AIC atteint sa valeur minimale comme le montre la figure 4.1 ci-contre.

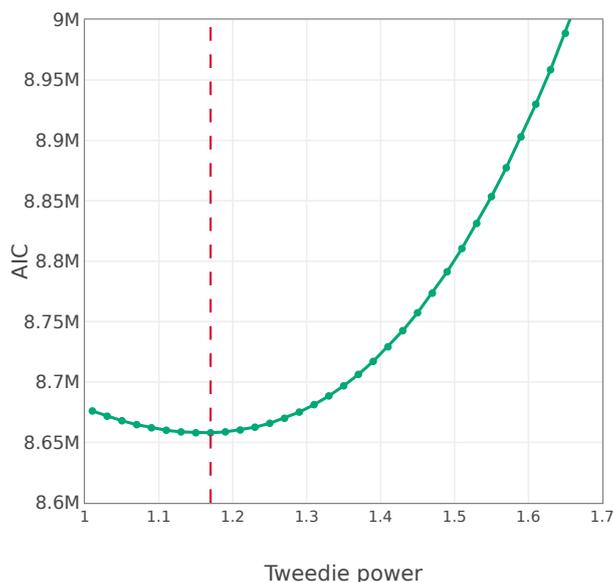


FIGURE 4.1 – Recherche du paramètre de puissance de Tweedie par minimisation de l'AIC.

Le modèle GLM\_Tweedie\_zero par défaut utilisant la variable MARQUE (respectivement GLM\_Tweedie\_Mod\_zero utilisant le croisement MARQUE x MODÈLE) est calibré en utilisant la valeur du paramètre de puissance précédemment déterminée. À l'exception de la variable ANNÉE, les variables quantitatives sont discrétisées. Les performances globales de ces modèles, mesurées par les indicateurs présentés au chapitre 2, sont calculés aussi bien en apprentissage qu'en test. Le rapprochement des valeurs des indicateurs entre ces deux échantillons permet

d’apprécier le sur-apprentissage éventuel du modèle alors que les indicateurs en test permettent de comparer les modèles et donnent une idée des performances si le modèle était déployé. Les valeurs obtenues des métriques sont consignées dans le tableau de comparaison des modèles 4.2.

Un regard porté sur les coefficients estimés montre qu’un grand nombre d’entre eux sont non-significatifs. De plus, des effets erratiques sont observés sur certaines variables comme l’âge du véhicule au début du risque, pour laquelle une évolution en dents de scie pour les intervalles au delà de 5,6 ans est observée. Ces observations suggèrent de procéder à une régularisation de ces modèles pour *contraindre* les coefficients peu faibles à être nuls et lisser les coefficients qui fluctuent beaucoup en regroupant les modalités dont la différenciation n’apporte pas assez de pouvoir explicatif au modèle. Cet objectif est atteint à partir de l’implémentation de la régulation *smurf* présentée dans la section 2.2.1.1.

La régularisation passe par le choix du paramètre de pénalisation  $\lambda$  qui guide son importance. Il est déterminé par maximisation d’un critère d’ajustement en validation croisée sur l’échantillon d’apprentissage. Le critère de maximisation retenu est la part de variance de Tweedie expliquée en validation. Pour le modèle utilisant la variable MARQUE, la valeur optimale du paramètre est de  $\lambda = 0,184$  ; le modèle correspondant ayant une part de variance expliquée de 10 %. Le coefficient de régularisation est bien plus faible pour le modèle utilisant la variable MODÈLE de véhicule :  $\lambda = 0,005$ . La régularisation est donc moins importante dans ce cas alors que la part de variance expliquée en validation monte à 10,5 %.

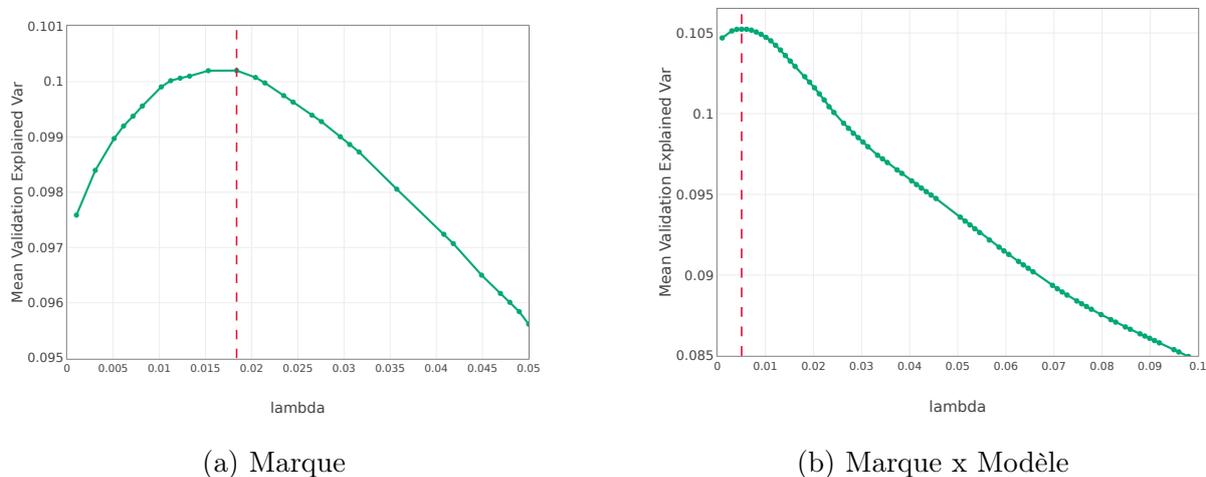


FIGURE 4.2 – Recherche du paramètre de régularisation par validation croisée pour les modèles GLM.

L’estimation du modèle GLM régularisé en utilisant la marque (respectivement le modèle de véhicule) conduit au modèle GLM\_Tweedie\_reg (respectivement GLM\_Tweedie\_Mod\_reg) dont les performances en apprentissage et en test sont consignées dans le tableau 4.2.

### 4.1.2 Modélisation de la prime pure à partir d'un LightGBM

Comme indiqué au chapitre 2, le LightGBM est envisagé comme alternative au modèle GLM pour le calcul du Burning Cost. La principale force du LightGBM réside dans sa vitesse d'entraînement mais également dans ses nombreux hyper-paramètres qui permettent d'ajuster un modèle le plus finement possible. D'entrée de jeu, des modèles LightGBM (GBM\_Tweedie\_default et GBM\_Tweedie\_Mod\_default) sont entraînés en considérant les paramètres par défaut intégrés dans le module de cette méthode. La spécification du modèle est celle d'une régression Tweedie avec fonction de lien exponentielle. Les performances de ces deux modèles sont consignées dans le tableau 4.2. Ensuite, les hyper-paramètres optimaux sont déterminés à partir d'une grille de recherche et d'un optimiseur bayésien, *optuna* (Akiba *et al.*, 2019). Le principe est d'orienter les paramètres à tester de façon itérative dans la direction qui maximise la part de variance expliquée sur la fraction de l'échantillon de validation dans une validation croisée à 5 sous-échantillons ou folds. Les hyper-paramètres optimisés sont :

- `tweedie_variance_power` : les valeurs testées sont tirées uniformément entre 1,001 et 1,5. Il s'agit du paramètre de puissance de la loi de Tweedie, retenue comme loi du burning cost. Ce paramètre doit être compris entre 1 et 2 pour que la régression Tweedie modélise bien une loi composée poisson-gamma ;
- `num_leaves` : les valeurs testées sont des entiers tirés uniformément entre 32 et 515. Ce paramètre correspond au nombre maximal de feuilles terminales par arbre. Plus ce paramètre est grand, plus important est le risque de sur-apprentissage. La croissance des arbres CART lors de la construction du LightGBM étant par feuille, le nombre de feuilles doit être suffisamment élevé pour donner un peu de profondeur aux arbres. Une valeur de 32 correspond à un arbre dont la profondeur de la branche la plus longue est au moins de 4 ( $= \log_2(32)$ ) ;
- `min_child_weight` : les valeurs testées sont des entiers tirés uniformément entre 1 et 16. Il s'agit de la somme minimale de poids d'instance nécessaire dans un noeud fils. Plus ce paramètre est grand, plus l'algorithme est conservateur c'est-à-dire robuste au sur-apprentissage ;
- `min_child_sample` : les valeurs testées sont des entiers tirés uniformément entre 1 et 50. Il s'agit du nombre minimal d'assurés dans un noeud pour procéder à son découpage. Le caractère conservateur des arbres s'améliore avec ce nombre, mais une valeur trop grande limite l'ajustement du modèle ;
- `bagging_fraction` : les valeurs testées sont tirées uniformément entre 0,7 et 1. Il s'agit de la fraction des assurés échantillonnés pour la construction de chaque nouvel arbre. Tiré aléatoirement, un sous échantillon des données d'apprentissage conduit à des arbres légèrement différents les uns des autres et au final à un modèle *ensembliste* plus robuste ;
- `feature_fraction` : les valeurs testées sont tirées uniformément entre 0,7 et 1. Ce paramètre

donne la fraction des facteurs de risque échantillonnés à chaque nouvel arbre. Il joue le même rôle que `bagging_fraction` ;

- `lambda_l1` (et `lambda_l2`) : les logarithmes des valeurs testées sont tirés uniformément entre  $10^{-8}$  et 1. Il s'agit des paramètres de régularisation de type L1 et L2. Plus ces paramètres sont grands, plus l'algorithme est conservateur ;
- `max_depth` : les valeurs testées sont des entiers tirés uniformément entre 3 et 16. Ce paramètre contrôle la profondeur maximale des arbres élémentaires. Plus il est faible, plus l'algorithme est conservateur.

Le graphique suivant présente l'évolution de la part de variance expliquée sur les 100 itérations utilisées pour déterminer les hyper-paramètres optimaux. La part de variance en validation croise pour se stabiliser au bout de 40 itérations, après quoi les augmentations sont très faibles. Un grand nombre de combinaisons permet d'atteindre une part de variance expliquée de l'ordre de 11 % pour le modèle utilisant la marque du véhicule et de 12 % pour le modèle utilisant le croisement `marque × modèle de véhicule`. De façon plus détaillée, les figures 4.3 (et 16 en annexe) présentent l'exploration de l'espace des hyper-paramètres et leurs effets sur la part de variance expliquée. Le paramètre de puissance optimal vaut 1,2, restant très comparable à celui déterminé dans le cas des modèles GLM.

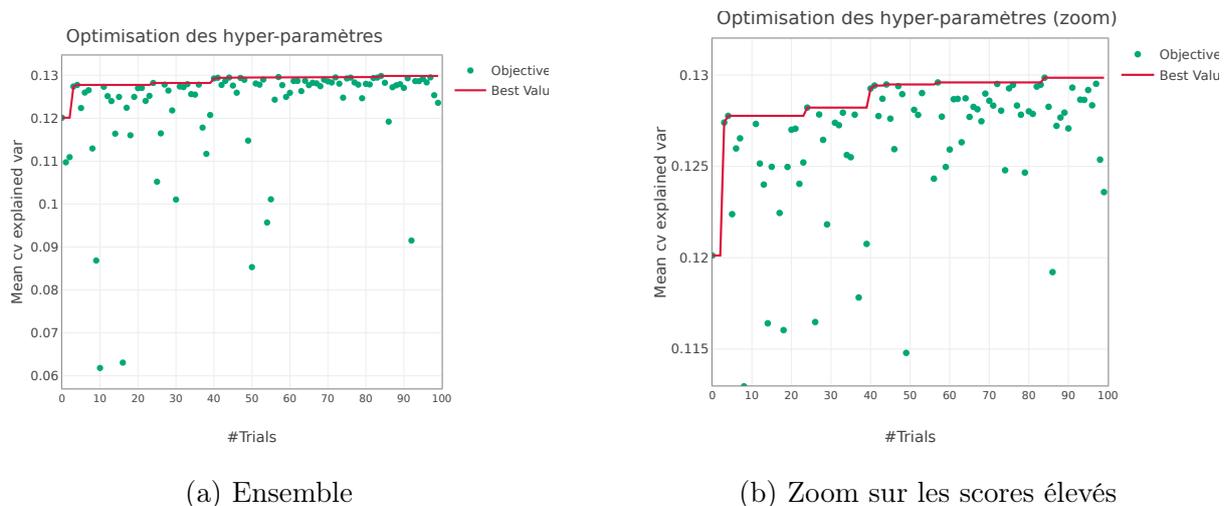


FIGURE 4.3 – Évolution de la part de variance expliquée lors de l'optimisation des hyper-paramètres du modèle LightGBM.

À la suite de cette optimisation des hyper-paramètres, les modèles `GBM_Tweedie_tuned` et `GBM_Tweedie_Mod_tuned` sont calibrés et utilisés pour prédire les polices du test. Leur évaluation est consignée dans le tableau 4.2.

### 4.1.3 Comparaison des modèles de construction du Burning Cost

Le tableau 4.2 ci-contre récapitule les performances des différents modèles de prédiction du Burning Cost estimé. Les résultats globaux suivants peuvent être observés :

- Tous les modèles estimés tendent à sous-estimer légèrement (de 0,13% et 5,30%) la charge totale de sinistres en test ;
- Le choix des hyper-paramètres dans les cas du LigthGBM permet d'améliorer la qualité d'ajustement des modèles aussi bien sur l'échantillon d'apprentissage que sur le test set et ceci quelque soit la métrique utilisée. Le choix des hyper-paramètres bien que coûteux en temps, est donc bien rentabilisé ;
- La régularisation des modèles GLM réduit l'écart de performance observé entre le train et le test, tout en maintenant des scores proches de ceux des modèles non régularisés. La régularisation des modèles GLM par la méthode *smurf* permet donc d'atteindre des capacités de généralisation très comparables à ceux en apprentissage ;
- Aussi bien pour le GLM que pour le LightGBM, l'utilisation de la variable marque x modèle de véhicule améliore les performances du modèle tout en évitant le sur apprentissage à travers la régularisation induite par le choix des hyper-paramètres ;
- Les modèles GLM sont meilleurs pour reconstituer la charge globale de sinistres (indicateur du ratio des sinistres prédits sur les sinistres observés) que les modèles LigthGBM. Inversement, les modèles de machine learning ont des meilleures performances sur les métriques qui privilégient les prédictions individuelles. Ceci est lié à l'approche théorique de ces modèles : collectif pour le GLM et individualiste pour le LightGBM. Un lien immédiat avec la démutualisation s'en suit puisqu'elle s'apparente à l'individualisation de la gestion des risques (favorisée par les modèles de machine learning) et s'oppose au traitement en pool des risques (tel que modélisé par le GLM).

TABLEAU 4.2 – Récapitulatif des performances des modèles de calcul du burning cost en test et en apprentissage.

Modèle estimé	Test					Train				
	Ratio. (%)	Dev.	Dev_exp.	MAE	RMSE	Ratio. (%)	Dev.	Dev_exp.	MAE	RMSE
GLM_Tweedie_Zero	-0,134	134,389	0,112	93,436	283,258	0,022	133,957	0,115	93,243	287,517
GLM_Tweedie_Reg	-0,164	134,320	0,113	93,403	283,244	0,022	133,862	0,115	93,225	287,488
GBM_Tweedie_Default	-3,901	132,653	0,117	91,386	282,771	-3,691	128,111	0,158	90,992	286,254
GBM_Tweedie_Tuned	-1,646	131,934	0,127	92,330	282,663	-1,430	127,818	0,153	91,876	285,618
GLM_Tweedie_Modele_zero	-0,162	133,059	0,120	93,040	282,942	0,030	132,231	0,125	92,789	287,027
GLM_Tweedie_Modele_Reg	-0,183	133,018	0,121	92,997	282,936	0,032	132,038	0,127	92,752	286,978
GBM_Tweedie_MM_Default	-5,302	131,984	0,110	90,319	282,375	-5,070	124,305	0,183	89,690	285,138
GBM_Tweedie_MM_Tuned	-2,012	130,999	0,133	91,773	282,326	-1,751	124,987	0,172	91,104	284,583

Dans le cadre de cette étude, deux modèles de construction du burning cost le plus fin  $\Phi^\Omega$  sont retenus. Le *meilleur* modèle GLM (au sens de la part de variance expliquée, de la déviance, de la MAE et de la RSME) GLM\_Tweedie\_Mod\_reg permettra d'évaluer le burning cost le plus fin que l'assureur puisse commercialiser. Il s'agit du modèle d'évaluation du profil de risque le plus individualisé que l'assureur puisse atteindre avec les informations à sa disposition. Le *meilleur* modèle LightGBM (au sens des mêmes critères que pour le choix du GLM) est le

modèle GBM\_Tweedie\_Mod\_tuned. Il permet de challenger le *meilleur* modèle GLM ainsi qu'à apprécier le niveau de démutualisation si le profil de risque était déterminé à partir de ce dernier.

#### 4.1.4 Interprétation des modèles de calcul des profils de risque individualisés

Deux modèles de calcul de primes pures les plus fines ont été calibrés et évalués numériquement. Il convient à présent de les interpréter pour déterminer les variables importantes ainsi que leurs effets sur le risque. L'interprétation du GLM est immédiate de part sa formulation relativement simpliste. Il s'agit globalement d'analyser les coefficients estimés. Le modèle LightGBM par contre nécessite de faire appel aux valeurs de SHAP pour en faire l'interprétation puisqu'il est considéré comme un modèle *boite noire*. Une exploration plus détaillée des méthodes d'interprétation des modèles de machine learning utilisés en actuariat est présentée par Delcaillau (2019)

##### Interprétation du modèle GLM Tweedie marque x modèle régularisé

La charge totale de sinistres estimée est de 0,18 % inférieure à la charge observée sur l'échantillon test. Cette valeur proche de zéro indique de façon globale la capacité du modèle à être en équilibre financier. La courbe lift répond au même objectif mais suivant les niveaux de burning cost. La figure 4.4 indique en apprentissage et en test l'adéquation des burning cost prédits et des burning cost observés suivant leurs montants. Que ce soit en apprentissage ou en test, les montants de sinistres observés et prédits s'alignent correctement, traduisant un modèle auto-calibré c'est-à-dire qui se généralise bien quelque soit le niveau de risque des assurés.

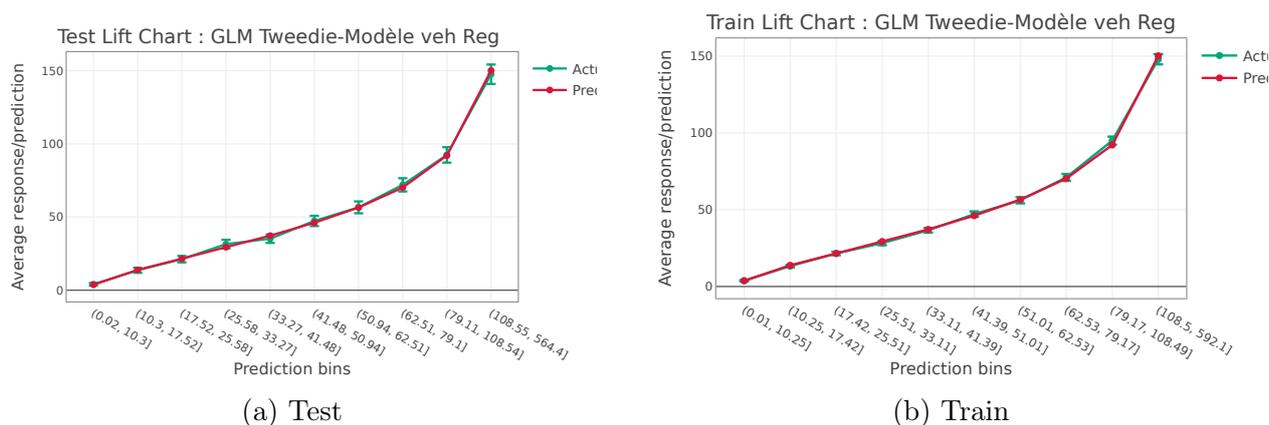


FIGURE 4.4 – Évaluation du modèle GLM par la courbe lift en apprentissage et en test.

La figure 4.5 suivante donne pour le modèle GLM optimal l'importance des variables. Une variable est d'autant plus importante que la moyenne des coefficients associés à ses différentes modalités est grande. Les variables étant toutes discrétisées, il n'y a pas d'effet de taille et les coefficients indiquent directement comment évolue le burning cost lorsque l'assuré présente

ou non la modalité. Par ordre d'importance, les variables les plus importantes sont l'âge du véhicule au début du risque, le croisement marque x modèle de véhicule, la formule, la puissance, l'ancienneté du véhicule en portefeuille et dans une bien moindre mesure la motorisation, la tranche de kilométrage, la carence, l'année d'exercice et le type de véhicule.

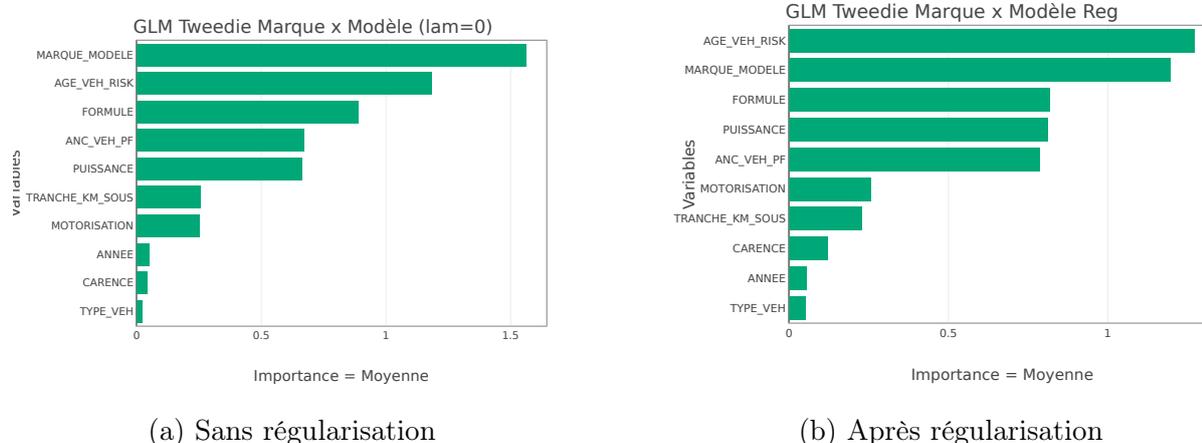


FIGURE 4.5 – Importances des variables du modèle GLM déterminées par l'effet moyen de la variable.

Une attention particulière est portée sur la variable la plus impactante à savoir l'âge du véhicule au début du risque. Cette variable présente après estimation et régularisation 53 coefficients (donc 53 intervalles au final après discrétisation et agrégation). L'effet de l'âge du véhicule sur le risque est croissant et concave pour les véhicules de moins de 5 ans et décroissant au delà de cette valeur (voir figure 4.6). Cette décroissance *a priori* contradictoire de l'effet de l'âge sur le risque de panne mécanique trouve une explication dans la politique de paiement des sinistres pratiquée en extension de garantie automobile. Au delà d'une certaine ancienneté du véhicule, le montant pris en charge par l'assureur est un pourcentage (qui décroît avec l'ancienneté) du montant de la réparation facturée par le garagiste. La fraction résultante est assimilée à de l'usure normale du véhicule.

L'analyse des coefficients suivant le croisement marque x modèle de véhicule indique un regroupement d'une vingtaine de modèles de voiture et un traitement différencié pour plus de 130 modèles au final. Les différences de contribution sont assez importantes d'une marque à l'autre. Suivant la formule, en considérant la couverture « classic » comme référence, la couverture « NEUF » coûte plus cher que la « EXPERT » (voir figure 14 en annexe).

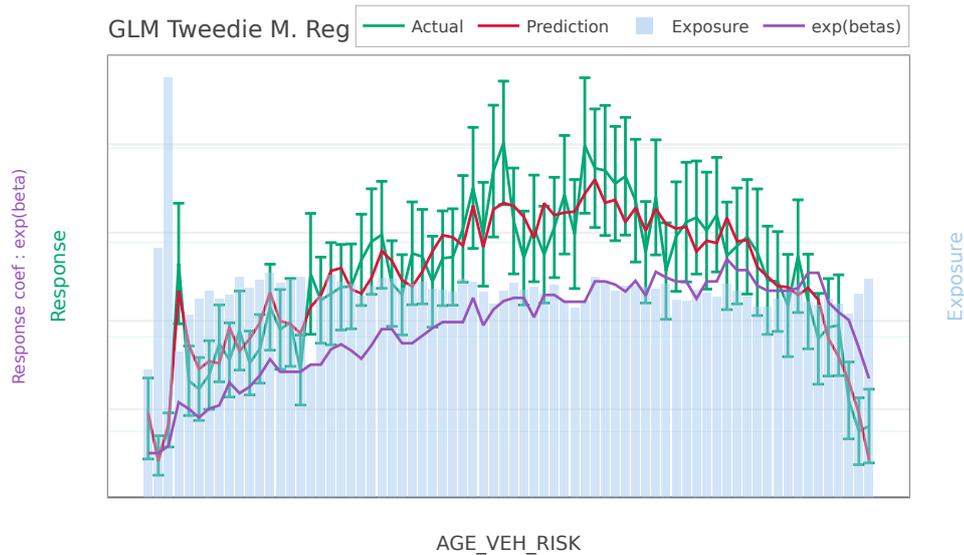


FIGURE 4.6 – Effet estimé par le modèle GLM de l’âge du véhicule au début du risque sur la sinistralité.

Bien que le GLM privilégie le traitement des risques sous forme de pool de par son principe, l’analyse des coefficients du modèle révèle que ceux-ci deviennent assez nombreux en incluant des discrétisations fines des variables continues ou des variables avec un grand nombre de modalités comme les modèles de véhicule. Le GLM estimé a en effet plus de 200 coefficients et génère près de 200 000 profils de risque différents sur le test soit près de 88 % de la taille du test. Ce modèle a donc déjà tendance à personnaliser les risques par contrats couverts. En pratique un tel modèle est difficilement utilisable. Les assureurs choisissent des découpages moins fins des variables continues et des regroupements des modalités pour obtenir des modèles parcimonieux. Mais ici, l’objet est de prédire aussi finement que possible les profils de risque individuels ; des scénarios où l’assureur réduit volontairement l’information à sa disposition sont envisagés ultérieurement (voir partie 4.2.1).

### Interprétation du modèle GBM Tweedie marque x modèle paramétré

De même que pour le GLM précédent, la courbe lift est utilisée pour évaluer la capacité du modèle `GBM_Tweedie_Mod_tuned` à prédire efficacement la charge de sinistres attendue (4.7 ci contre). La charge de sinistres globale estimée est de 2 % inférieure à la charge totale observée. L’alignement des courbes sur l’échantillon test est satisfaisant bien que ce modèle sous-estime grandement la charge de sinistres sur les mauvais risques lorsqu’il est évalué en restitution sur l’échantillon d’apprentissage.

La figure 4.8 permet d’apprécier l’importance des variables à partir de la contribution absolue moyenne de la variable au burning cost. Le modèle `LightGBM` accorde une grande importance à la variable `marque x modèle de véhicule`. Viennent ensuite avec 3 fois moins d’impor-

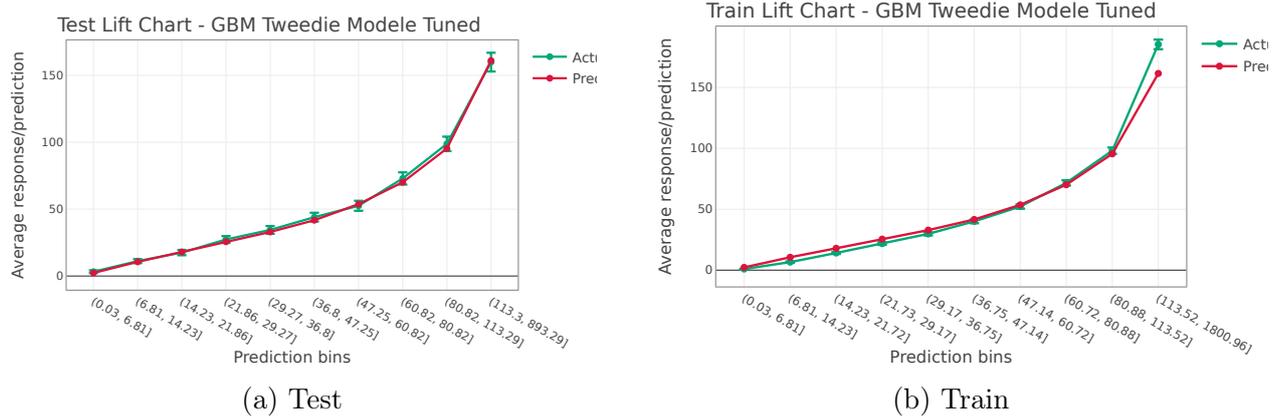


FIGURE 4.7 – Évaluation du modèle LightGBM par la courbe lift en apprentissage et en test.

tance : la puissance, l'âge du véhicule, l'ancienneté du contrat, la motorisation, le kilométrage et la formule. L'année, la carence et le type de véhicule ont une importance marginale.

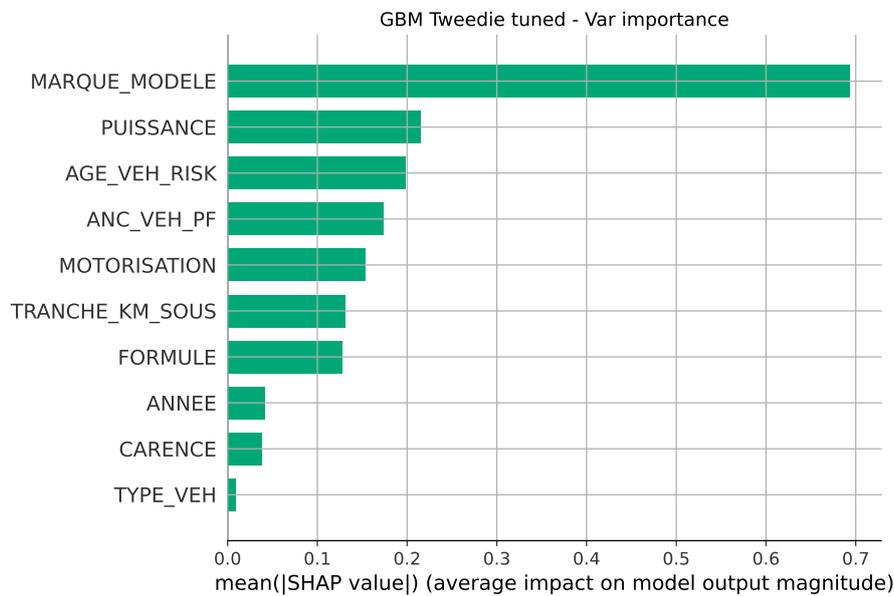


FIGURE 4.8 – Importances des variables du LightGBM mesurées à partir des valeurs de SHAP.

Le graphe de dépendance de SHAP apporte une visualisation de l'évolution des valeurs de SHAP en fonction des valeurs prises sur un facteur de risque donné. Il permet alors d'apprécier l'effet marginal d'un facteur de risque donné sur la sinistralité. Pour les variables âge du véhicule au début du risque et puissance, ces effets sont représentés sur la figure 4.9. Pour les véhicules de moins de 5 ans au début de la couverture, le risque croit assez rapidement. Au delà de cette valeur, l'effet de l'âge sur le burning cost est négatif. Cet effet est similaire à celui observé avec le GLM, excepté l'importance moindre que le GLM accorde aux véhicules de plus de 9 ans en les regroupant tous et leur attribuant un coefficient de risque unique. La puissance qui était la quatrième variable la plus importante pour le GLM avec un effet positif et linéaire sur la sinistralité (voir 15 en annexe) apparaît pour le LightGBM comme la deuxième variable la plus importante. Son effet reste positif mais présente deux phases : un effet qui croit plus rapidement

pour les véhicules dont les chevaux fiscaux sont inférieurs à 15 et un effet relativement stable au delà de cette valeur. Ici encore, la différence observée entre le GLM et le LightGBM correspond à un regroupement des véhicules ayant plus de 20 chevaux fiscaux par le GLM.

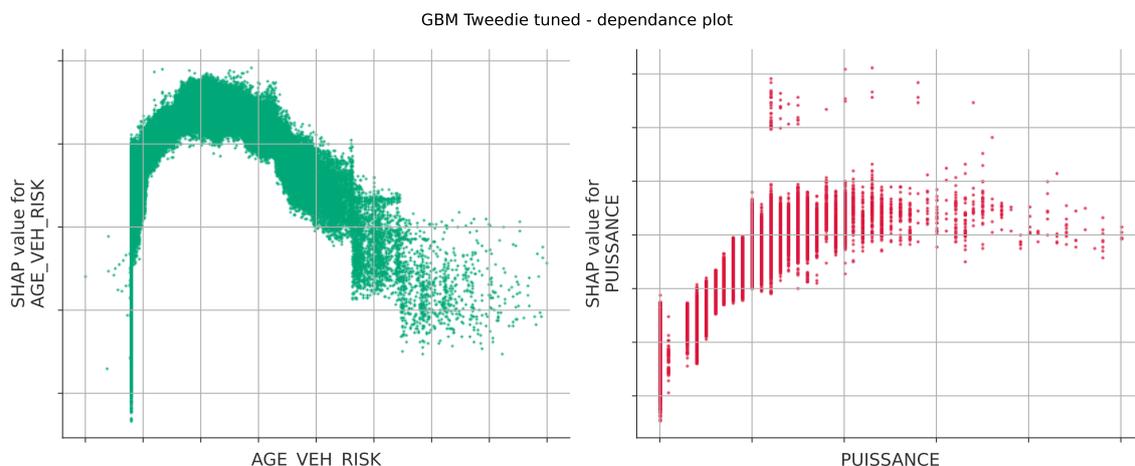


FIGURE 4.9 – Effets des variables âge au début du risque et puissance du modèle LightGBM mesurés à partir des valeurs de SHAP.

De façon générale il se dégage que les variables les plus importantes et leurs effets sur la sinistralité identifiés par le LightGBM sont assez comparables à ceux identifiés avec le GLM. Le modèle GLM\_Tweedie\_Mod\_reg semble donc assez robuste et performant pour modéliser de façon très fine le profil de risque.

## 4.2 Applications numériques des indices de mutualisation simples

Le modèle GLM\_Tweedie\_Mod\_reg est ressorti des estimations précédentes comme moyen de calcul des profils de risque le plus fins que BNP Paribas Cardif puisse atteindre. Il s'agit du tarifeur noté  $\Phi^\Omega$  dans la partie théorique de ce mémoire. Pour évaluer le niveau de démutualisation, il faut également connaître le modèle de tarification ( $\Phi^{Ass}$ ) de la prime technique que l'assureur choisit de déployer effectivement dans sa grille tarifaire. De plus, la prime commerciale (ou une description de la stratégie qui permet à l'assureur de passer de la prime technique à la prime commerciale) est nécessaire pour déterminer l'OL.

### 4.2.1 Construction de la prime technique tarifée

En pratique, pour passer de  $\Phi^\Omega$  à  $\Phi^{Ass}$ , l'assureur peut soit agréger les coefficients en introduisant un *a priori* relatif aux hypothèses de structure des modalités qu'il agrège, soit estimer un nouveau modèle en se restreignant à certaines variables et en les discrétisant manuellement (moins finement que la méthode *smurf* utilisée pour le modèle  $\Phi^\Omega$ ). Afin d'illustrer le calcul

des indicateurs, une stratégie de tarification semblable à celle utilisée par BNP Paribas Cardif est implémentée. Explicitement, les variables du tableau 4.3 sont utilisées pour la construction de la prime technique tarifée. Le choix des variables et des modalités est non seulement guidé par les interprétations du modèle  $\Phi^\Omega$  mais également par les pratiques commerciales observées chez Cardif et sur le marché de l'assurance extension de garantie.

TABLEAU 4.3 – Exemple de variables et discrétisation pour un modèle pratique de tarification.

Variable	Modalités
Formule	Neuf, Expert, Classic
Âge à la souscription	<5 ans, 5 - 7 ans, plus de 7 ans
Type de véhicule	Auto, Camping Car
Puissance	<6 cv, 6 - 9 cv, plus de 9 cv

Avec ces variables, l'assureur aura au plus  $3^3 \times 2 = 54$  segments de burning cost tarifés différents. En suivant la même démarche que précédemment, le modèle  $\Phi^{Ass}$  est estimé et testé sur la base des contrats historiques précédents. Il permet également d'atteindre l'équilibre financier global en test, puisque la sinistralité prédite en test est juste de 0,14 % inférieure à celle observée effectivement. Par contre, la part de variance expliquée chute à 3 % (contre 12 % avec  $\Phi^\Omega$ ). La déviance, la MAE et la RSME se détériorent également en augmentant comparativement au modèle  $\Phi^\Omega$  (voir tableau 7 en annexe). Ainsi, simplifier le modèle à déployer (pour répondre aux exigences du distributeur ou être plus *transparent* ou être plus mutualiste, ...) s'accompagne d'un risque de structure additionnel que l'assureur accepte de porter en mutualisant ou en regroupant certains profils de risque qu'il pourrait différencier.

## 4.2.2 Évaluation et interprétation du niveau de démutualisation sur le portefeuille EG de BNP Paribas Cardif en 2020

À partir des modèles  $\Phi^\Omega$  (GLM\_Tweedie\_Mod\_reg),  $\Phi^{Ass}$  (GLM avec les variables du tableau 4.3) et du vecteur des primes commerciales pratiquées sur les contrats couverts en 2020, les 3 vecteurs indispensables pour le calcul des indicateurs simples sont déterminés comme suit :

Code Listing 4.1 – Calcul des vecteurs nécessaires à l'évaluation des indicateurs simples.

```
BC_Omega = GLM_Tweedie_BC_Omega.predict(base_indic)*np.exp(beta_annee_2020)
BC_Ass = GLM_Tweedie_BC_Ass.predict(base_indic_ass)
OL_Ass = base_indic['PRIME_COM'] - BC_Omega
```

Le coefficient `beta_annee_2020` corrige les prédictions pour l'année 2020 puisque le modèle  $\Phi^\Omega$  a été entraîné sur un historique n'incluant pas 2020. Ce coefficient est obtenu en projetant la fréquence moyenne annuelle et le coût moyen annuel. Les projections se font à partir des

régressions linéaires simples de la fréquence et du coût moyen sur l'année d'exercice. Les résultats des droites de projection sont consignés en annexes dans les tableaux 5 et 6. Ce coefficient reflète la meilleure anticipation de la dynamique de la sinistralité calculée par un actuaire en fin d'exercice 2019. Il ne prend donc pas en compte les effets liés à la crise sanitaire Covid-19 qui ne sont correctement quantifiables qu'en fin d'exercice 2020. Le choix de ce coefficient (plutôt que d'un coefficient calculé rétrospectivement, qui prendrait alors en compte l'effet de la crise sanitaire) est guidé par le timing de mise à jour des modèles de tarification. La mise à jour de la grille de tarif est en effet construite ou mise à jour en amont de l'exercice. Il est donc plus naturel d'utiliser une anticipation de la dynamique des sinistres pour corriger les prédictions de 2020, plutôt qu'une réalisation de cette dynamique.

Une représentation graphique des distributions des primes est donnée par la figure 4.10. Celle-ci illustre en particulier un étalement très prononcé vers la droite des profils de risque à calculer à la maille la plus fine. Utiliser le tarifeur  $\Phi^{Ass}$  crée un important regroupement des 115 000 assurés exposés en 2020 avec notamment 48 valeurs uniques de prime pure contre plus de 60 000 pour les profils les plus individualisés (avec une précision au centime près). La stratégie de tarification implémentée chez BNP Paribas Cardif semble donc induire une mutualisation technique importante dans le portefeuille.

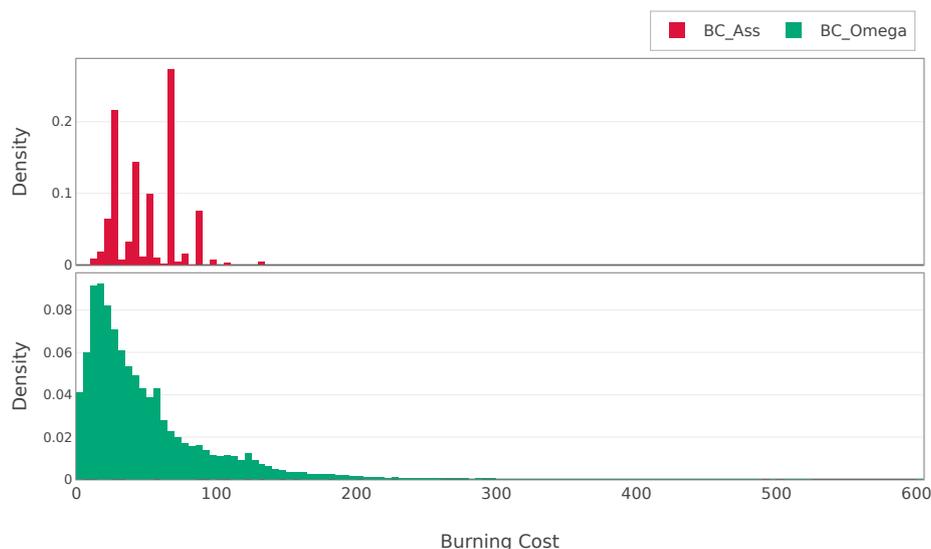


FIGURE 4.10 – Distributions des burning costs les plus fins (■ BC\_Omega) et tarifés (■ BC\_Ass).

Pour visualiser les OL, 200 profils de risque (générés par  $\Phi^\Omega$ ) sont tirés aléatoirement et triés par ordre croissant. Les OL correspondantes sont ensuite superposées au dessus de ces profils de risque, aboutissant à la prime commerciale payée par l'assuré. Cette représentation montre que les moins risqués financent le portefeuille puisque la hauteur des barres d'OL décroît avec la dégradation du profil de risque, c'est-à-dire à mesure que le burning cost augmente. Très peu de montants uniques de primes commerciales sont observés en fin de compte. Ces primes commerciales stables entre assurés reflètent une mutualisation importante aussi bien technique

que commerciale sur ce portefeuille.

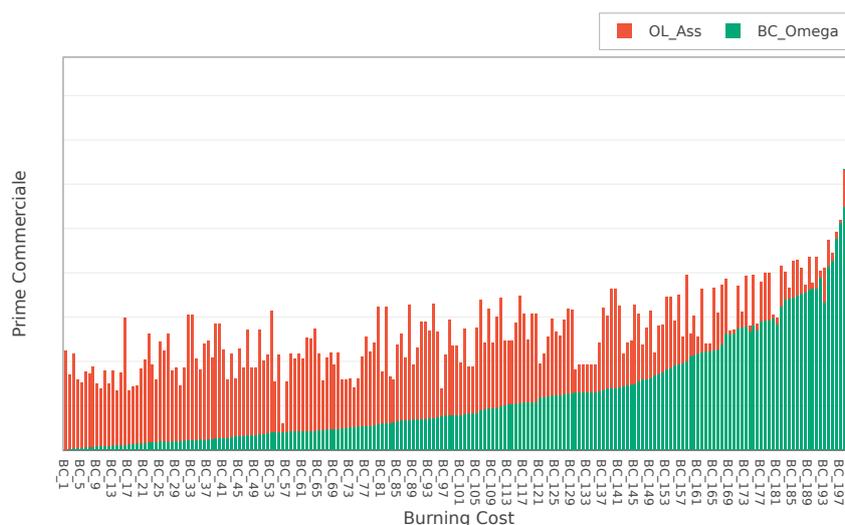


FIGURE 4.11 – Décomposition de la prime commerciale suivant l’OL ■ et le burning cost le plus fin ■ sur un échantillon de 200 profils de risque triés par ordre croissant.

Le tableau 4.4 ci-contre donne sur le portefeuille des contrats couverts en 2020, les valeurs calculées des indices simples de démutualisation en considérant la stratégie de tarification définie dans le tableau 4.3.

TABLEAU 4.4 – Mesures de démutualisation par les indices vanilles sur les véhicules couverts par BNP Paribas Cardif en 2020.

Indicateur	ID_Var_BC	ID_Ent_BC	IM_L1_BC (Shift)	IM_Var_OL	IM_Ent_OL
Valeur	0,175	0,205	34,215 (33,964)	0,388	0,782

Les indicateurs de démutualisation et de mutualisation s’accordent pour rendre compte d’une mutualisation importante dans ce portefeuille. Plus précisément, les interprétations suivantes en découlent :

#### ID\_Var\_BC : Transfert de 17,5 % du risque de structure cessible

La stratégie de tarification utilisée permet de transférer effectivement 17,5 % du risque de structure cessible dans le tarif technique. En renonçant à une tarification plus fine, Cardif consent à porter plus de risques de structure. Construire les primes techniques commerciales en se limitant aux 4 facteurs sus-mentionnés implique en effet que l’assureur supporte le risque d’un changement de distribution sur les facteurs non-pris en compte dans le tarif. Cet indicateur apporte une appréciation du gain du risque de structure que l’assureur pourrait encore céder s’il passait d’une tarification utilisant  $\mathbf{X}_{Ass}$  à une tarification basée sur  $\mathbf{X}_{\Omega}$ .

## ID\_Ent\_BC : Un tarif technique pour 5 profils de risque

L'entropie des primes techniques quantifie l'information contenue dans la grille des tarifs techniques, cet indicateur est alors un rapport de quantité d'information. En s'inspirant directement de la théorie de l'information, cet indicateur s'interprète ainsi : encoder les primes techniques tarifées ne nécessite que 20,5 % des bits qu'il aurait fallu utiliser si l'assureur pratiquait un tarif différent par assuré. Ce qui revient à dire que Cardif n'utilise dans sa stratégie tarifaire que 20,5 % de l'information nécessaire pour individualiser les profils de risque. L'information ici correspond aux facteurs de risques, leur discrétisation et leur pertinence. De façon plus détaillée, la figure 4.12 donne l'entropie des vecteurs de primes techniques qui résulterait de la segmentation agrégée (basée sur  $\mathbf{X}_{Ass}$ ), la plus fine (basée sur  $\mathbf{X}_{\Omega}$ ) ou individualisée (basée sur  $\Omega$  et supposée individualisable). Le tableau 4.5 quant-à lui se concentre sur l'encodage des primes dans le cas de la segmentation agrégée.

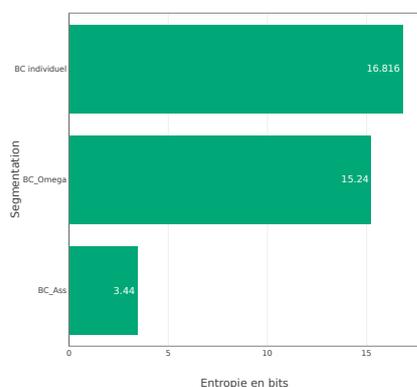


FIGURE 4.12 – Entropie des primes selon la segmentation assureur.

$BC_{Ass}$	Prop.	Prop. Cumulée	Codage Huffman	code
$BC_{u1}$	0,258	0,258	01	2
$BC_{u2}$	0,209	0,467	00	2
$BC_{u3}$	0,140	0,607	011	3
$BC_{u4}$	0,075	0,682	0111	4
...	...	...	...	...
$BC_{u45}$	0,000	1,000	1000100111011111	16
$BC_{u46}$	0,000	1,000	00010100111011111	17
$BC_{u47}$	0,000	1,000	11100100111011111	17
$BC_{u48}$	0,000	1,000	01100100111011111	17

TABLEAU 4.5 – Encodage en bits des primes techniques tarifées sur le portefeuille de Cardif.

Avec les variables tarifaires de Cardif, il suffit en moyenne de 3,44 bits pour encoder le profil de risque d'un assuré, contre 15,24 si la prime technique était modélisée à partir du modèle d'évaluation des profils le plus segmenté  $\Phi^{\Omega}$ . Si on admet que l'assureur dispose d'assez d'informations sur son risque pour associer à chaque assuré un profil de risque spécifique, alors l'entropie de ces primes techniques individualisées serait de 16,82 très proche de l'entropie des primes techniques les plus fines. Pour coder efficacement les profils en portefeuille, il suffit d'associer aux burning costs les plus fréquents des codes courts et les descriptions longues sont utilisées pour les profils rares. En segmentant plus finement le risque, chaque profil se raréfie et les longueurs des codes associés augmentent également, aboutissant alors à une entropie qui croît avec la finesse de la segmentation. Plus l'assureur utilise des facteurs de risques dans son tarif, plus il crée de la différenciation entre ses assurés.

Une interprétation plus schématique de cet indicateur et éventuellement plus parlante pour la communauté des assureurs, émerge lorsqu'on « passe » d'une vision information en bits à une vision nombre de profils de risque. Un code court est associé à un profil de risque fréquent alors qu'un codage long est utilisé pour les profils rares. La longueur du code est alors liée

négativement à l'occurrence du profil en portefeuille. Lorsqu'un codage optimal comme celui de Huffman est adopté pour représenter les profils de risque, l'entropie est égale à la longueur moyenne du codage, soit :  $Ent(\mathbf{BC}) = \sum_i p_i |code_i|$  où  $p_i$  est la proportion d'assurés ayant la prime tarifée codée par  $code_i$  dont la longueur est  $|code_i|$ . L'indice de la somme porte sur les profils de risque uniques identifiés par la segmentation utilisée. Sous cette réécriture de l'entropie,  $ID\_Ent\_BC$  devient :

$$\begin{aligned} ID_{Ent||BC}(\mathcal{P}f^{Ass}) &= \frac{\sum_i p_i |code_i|}{\log_2(N^{Ass})} \\ &= \sum_i p_i \frac{|code_i|}{\log_2(N^{Ass})} \end{aligned}$$

le terme  $\log_2(N^{Ass})$  est le nombre de bits nécessaire pour encoder un profil de risque pour une segmentation individualisée de la prime technique ; ou encore la longueur du codage associée à chaque assuré lorsqu'il est considéré comme un risque spécifique. Il s'agit de la valeur maximale de  $|code_i|$  pour tout macro profil de risque  $i$  issue d'une segmentation donnée. Le rapport  $\frac{|code_i|}{\log_2(N^{Ass})}$  est d'autant plus petit que le macro profil codé par  $code_i$  est fréquent, c'est-à-dire que le burning cost tarifé  $BC_i$  regroupe des profils individualisables sous  $\Omega$ . Son inverse  $\frac{\log_2(N^{Ass})}{|code_i|}$  est alors assimilable au nombre de profils de risque (supposés individualisés) regroupés sous la prime technique  $BC_i$ . En effet si :

- $\frac{\log_2(N^{Ass})}{|code_{i_1}|}$  est petit c'est à dire proche de 1, alors le tarif  $BC_{i_1}$  codé par  $code_{i_1}$  ne concerne qu'un très faible nombre d'assurés en portefeuille puisqu'il faut pratiquement  $\log_2(N^{Ass})$  bits pour représenter ce profil ;
- $\frac{\log_2(N^{Ass})}{|code_{i_2}|}$  est grand c'est à dire proche de  $\log_2(N^{Ass})$ , alors le tarif  $BC_{i_2}$  est commun à une grande fraction du portefeuille, puisqu'il est codé sur très peu de bits.

Il en découle que l'inverse de l'indicateur  $ID\_Ent\_BC$  est la moyenne harmonique pondérée du « nombre d'assurés regroupés sous un tarif technique pour une segmentation donnée » :

$$\frac{1}{ID_{Ent||BC}} = \frac{1}{\sum_i \frac{p_i}{\frac{\log_2(N^{Ass})}{|code_i|}}}$$

Une interprétation actuarielle de cet indicateur est alors déduite ; le ratio d'individualisation correspond à l'inverse du nombre moyen de profil de risque mutualisé sous un tarif technique donné. Dans le cas de l'exemple d'application, l'inverse de cet indicateur est égal à 4,88, dès lors Cardif choisit de considérer en moyenne une prime technique pour 5 profils de risque différents.

**IM\_L1\_BC : Un montant moyen de prime annuel solidarisé de 34,25 €**

Le montant moyen annuel de prime que Cardif choisit de solidariser entre assurés de son portefeuille est de 33,31 €. Ce montant traduit un transfert de risque entre assurés en portefeuille. En le normalisant par rapport à la prime de risque moyenne, il en résulte un transfert de risque de l'ordre de 43,67 % du risque moyen. Un assuré est considéré comme finançant lorsque sa prime technique est plus grande que son profil de risque connu -  $BC_i^{Ass} > BC_i^{\Omega}$ . À l'inverse, un assuré est financé par le portefeuille lorsque sa prime technique est inférieure à son profil de risque connu -  $BC_i^{Ass} < BC_i^{\Omega}$ . Il convient de préciser que ce montant solidarisé ne prend pas en compte le risque mutualisé à l'intérieur des classes de risque obtenues avec la segmentation la plus fine connue.

La figure 4.13 ci-contre donne la distribution des montants solidarisés par assuré dans le portefeuille. Compléter cet indicateur par ces distributions ou leurs synthèses numériques données par le tableau 4.6 permet d'apprécier comment se répartit le financement de cette solidarité. En particulier sur ce portefeuille, les montants perçus sont très étalés à gauche avec un montant minimal perçu de l'ordre de -562 € de prime contre un montant maximal financé par tête de 136 €. Ceci s'explique par l'anticipation des *très mauvais risques* pour lesquels les primes techniques sont considérablement abaissées, rendant ainsi leur inclusion à l'assurance possible. Par contre, le financement de ces profils est bien pris en charge par la communauté et non par quelques assurés particuliers puisque la distribution des montants financés n'a pas de queue épaisse à droite. Enfin, le coefficient de corrélation entre les montants algébriques solidarisés et le profil de risque  $BC_{\Omega}$  vaut -0,91. Cette valeur très proche de -1 confirme le sens du financement du portefeuille, des moins risqués vers les plus risqués, ce qui contribue à rendre ce produit plus inclusif.

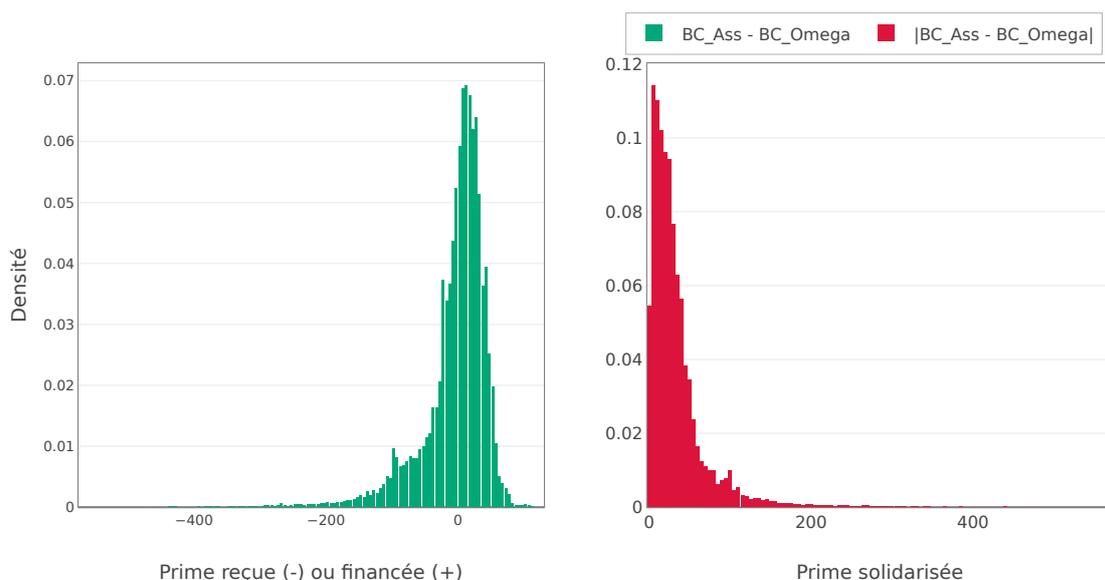


FIGURE 4.13 – Distribution des montants de primes annuelles solidarisées en valeurs algébriques (■ à gauche) et en valeurs absolues (■ à droite).

	count	mean	std	min	Q-25 %	median	Q-75 %	max
BC_Ass - BC_Omega	115359.0	-0.251	49.943	-562.216	-15.357	10.396	30.349	135.758
BC_Ass - BC_Omega	115359.0	34.215	36.383	0.001	12.082	24.908	43.529	562.216

TABLEAU 4.6 – Description numérique des distributions des montants de primes annuelles solidarisées.

Il convient de remarquer enfin que le résultat moyen escompté par assuré est pratiquement nul, donc le modèle d'identification des profils de risque  $\Phi^\Omega$  et de calcul des primes techniques  $\Phi^{Ass}$  s'accorde bien sur la charge globale de sinistres. La valeur de l'indicateur précédent est donc très proche de sa version corrigée (IM\_L1\_BC\_Shift) et s'interprète alors effectivement comme un montant solidarisé au sein de la communauté.

### IM\_Var\_OL : Les OL sont 2,5 fois moins volatiles que les profils de risque

Le rapport entre la variabilité des OL et celle des profils de risque calculés à partir de l'ensemble des facteurs de risque à disposition vaut 0,388, c'est-à-dire les OL sont 2,5 fois moins volatiles que les profils de risque en portefeuille. L'OL n'est donc pas constante sur ce produit et cet indicateur rend également compte d'un portefeuille mutualisé.

### IM\_Var\_OL : 78,2 % des assurés ont une OL personnalisée

L'assureur utilise l'OL pour regrouper les assurés à 78,2 %, conduisant ainsi à un portefeuille très solidaire. La mutualisation par l'OL concerne non seulement l'agrégation technique des profils de risque mais également l'application des chargements assureurs différents d'un profil de risque à l'autre. Cardif s'appuie donc sur les deux leviers pour mutualiser ce portefeuille.

## 4.2.3 Analyses numériques de propriétés des indicateurs de démutualisation vanilles

Les propriétés numériques étudiées ici fournissent une appréciation de la stabilité de ces indicateurs au regard de la précision souhaitée sur les montants de prime et de la taille du portefeuille. Les échelles des axes des ordonnées sont fixées entre 0 (la valeur minimale théorique) et 1 (valeur maximale) à l'exception de IM\_L1\_BC qui est fixé entre 0 et 60 €.

### Robustesse des indicateurs vanilles à la précision sur les primes

La précision ou nombre de chiffres après la virgule détermine l'ordre d'arrondi sur les prix que l'assureur considère pour la commercialisation de ses produits. Les tarifs sont généralement « ronds » et il n'est pas courant d'observer des primes avec une précision de 4 chiffres après la virgule. En effet, si la monnaie utilisée pour les transactions est l'euro par exemple, seuls les deux premiers chiffres après la virgule ont une signification économique. Ils correspondent

au centime et en deçà d'un centime, aucune transaction financière n'est possible. Les montants d'OL et de primes techniques tarifées sont arrondis aux ordres  $-1, 0, 1, 2, 3, 4, 5, 10$ . La précision  $p$  correspond à l'arrondi d'ordre  $10^{-p}$ . Notamment, la précision de  $-1$  correspond à des tarifs arrondis à la dizaine d'euro près,  $0$  à l'euro près et  $2$  au centime près. Bien qu'il n'y ait pas d'intérêt économique pratique (dans le cas des prix exprimés en euro) d'une précision de plus de 3 chiffres après la virgule, cela présente un intérêt académique pour étudier le comportement des indicateurs.

La figure 4.14 montre l'évolution des indicateurs en fonction de la précision. Il en ressort que les indicateurs basés sur la variance et la norme L1 sont assez robustes à l'ordre d'arrondi. ID\_Ent\_BC est légèrement sensible à la précision, en effet pour des prix arrondis à la dizaine d'euro, l'indice vaut  $0,166$  et croit ensuite pour se stabiliser à  $0,204$  pour une précision à la dizaine de centime. Cette évolution s'explique par la solidarité additionnelle créée en réduisant la précision sur les burning costs tarifés. Par contre, IM\_Ent\_OL est très sensible au regard de l'arrondi et évolue en sens inverse avec l'intuition précédente. Il passe de  $0,22$  pour une précision de  $-1$  à  $0,90$  pour une précision de 3 chiffres après la virgule. Cet effet de la précision sur les indicateurs basés sur l'entropie est la matérialisation du problème de *coarse graining* présent même en physique lors de l'utilisation de l'entropie (voir partie 2.1.1).

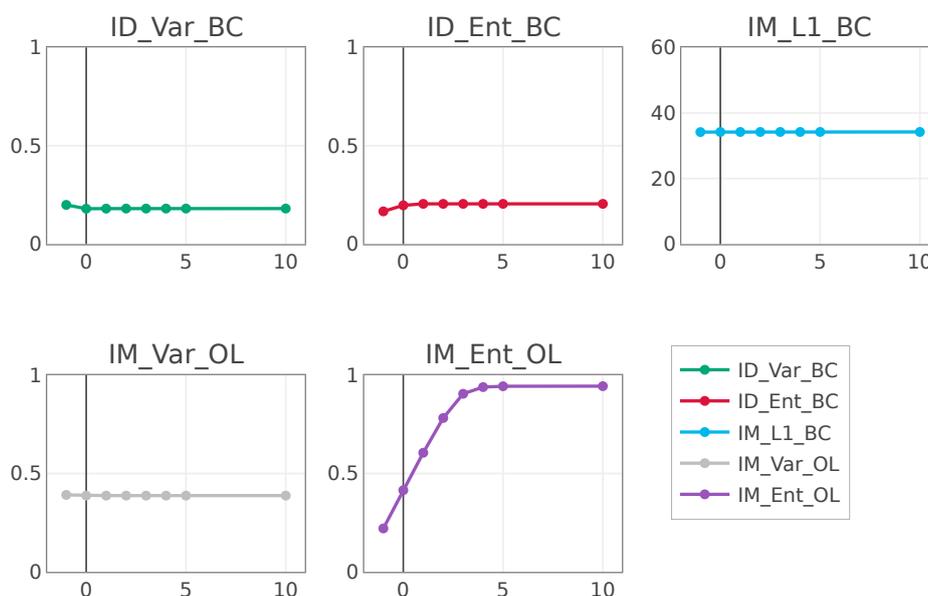


FIGURE 4.14 – Analyse de l'effet de la précision des montants (arrondi) sur les indices de mutualisation simples.

### Robustesse des indicateurs vanilles à la taille du portefeuille

Pour étudier l'effet de la taille du portefeuille sur ces indicateurs, des fractions de plus en plus importantes des contrats couverts en 2020 sont tirées aléatoirement et les indicateurs vanilles sont calculés sur les portefeuilles ainsi constitués. Les tailles de portefeuilles obtenus varient entre 5 000 à 110 000 polices. La figure 4.15 présente le résultat de cette expérience. Les indicateurs sont assez stables à l'exception de l'IM\_Ent\_OL qui décroît de  $0,944$  à  $0,782$ . L'autre

indice entropique  $ID\_Ent\_BC$  décroît très lentement, pour des portefeuilles peu matériels et se stabilise quand la taille du portefeuille est au delà de 20 000 polices. Avec l'accroissement des polices, les fréquences d'observation des macro profils en portefeuille se stabilisent de plus en plus.

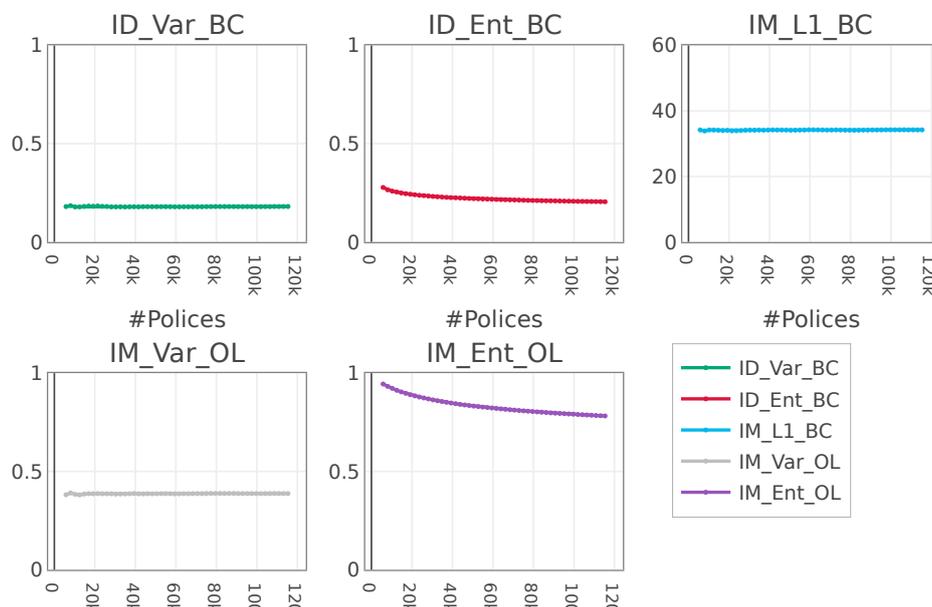


FIGURE 4.15 – Analyse de l'effet de la taille du portefeuille sur les indices de mutualisation simples.

### 4.3 Utilisations pratiques des indices de mutualisation basés sur la prime

Le calcul des indicateurs ainsi que leurs interprétations étant présentés, deux exemples d'utilisation de ces indicateurs sont explorés; en interne chez un assureur et à des fins de comparaison avec d'autres assureurs.

#### 4.3.1 Indices de mutualisation pour guider le choix du niveau de segmentation tarifaire

Pour répondre à la question « quelles sont les variables tarifaires à inclure explicitement dans le modèle de tarification  $\Phi^{Ass}$  », l'assureur se tourne en général vers des indicateurs qui mesurent la qualité de l'ajustement du modèle. Mais la question pourrait être formulée autrement : « Quelles sont les variables à inclure dans le tarif pour transférer au moins 80 % du risque de structure cessible dans la grille tarifaire ? » ou encore « quelles sont les variables qu'il ne faut pas agréger pour maintenir au moins deux profils de risque par tarif ? ».

Répondre à ces problématiques revient à décliner une stratégie de tarification qui respecte un niveau de mutualisation donné. Le régulateur pourrait également *imposer* un niveau de mutualisation minimale pour contrôler des dérives éventuelles sur le marché comme

l'hyper-segmentation, qui peut mener à des exclusions. L'analyse ci-après apporte des éléments pour éclaircir ces réflexions. L'expérience consiste à enrichir le modèle de tarification  $\Phi^{Ass}$  en lui rajoutant itérativement les variables (avec toutes leurs modalités ou discrétisations fines) du modèle d'évaluation du profil de risque individuel  $\Phi^\Omega$  suivant leur importance. L'intérêt est donc d'étudier l'évolution de la démutualisation du portefeuille lorsque  $X^{Ass} \rightarrow X^\Omega$  c'est-à-dire lorsque l'assureur tend à utiliser toute l'information à sa disposition. Les variables sont donc incorporées dans le modèle dans l'ordre : AGE\_VEH\_RISK > MARQUE\_MODELE > PUISSANCE > FORMULE > ANC\_VEH\_PF > MOTORISATION > TRANCHE\_KM\_SOUS > TYPE\_VEH > CARENCE.

L'évolution des indicateurs est représentée sur la figure 4.16 au regard de ces modèles de plus en plus segmentés. Les indicateurs reflètent bien l'intuition selon laquelle plus la segmentation du risque est fine plus le niveau de démutualisation est élevé. Si les évolutions des indicateurs ID\_Ent\_BC et IM\_L1\_BC sont en parfait accord avec cette tendance générale, l'évolution de l'indice basé sur la variance des burning cost est légèrement erratique. Le montant annuel de prime que l'assureur décide de solidariser passe de 37,62 € pour la segmentation du risque exclusivement par la variable AGE\_VEH\_RISK à 3,05 € pour le modèle le plus exhaustif. Une décroissance importante est observée lorsqu'on passe d'un modèle de 4 à 5 variables en ajoutant la variable ANC\_VEH\_PF. Après cette baisse importante, la décroissance reste relativement stable et atteint un plateau au bout de 7 variables. L'effet symétrique est observé sur l'indice de démutualisation ID\_Ent\_BC qui croît assez rapidement pour moins de 5 variables et se stabilise au bout d'un modèle à 5 facteurs de risques. De façon générale, l'accentuation de la démutualisation par les facteurs les plus impactants est très prononcée contrairement à la démutualisation induite par les facteurs qui n'ont que très peu d'effet dans la segmentation du risque.

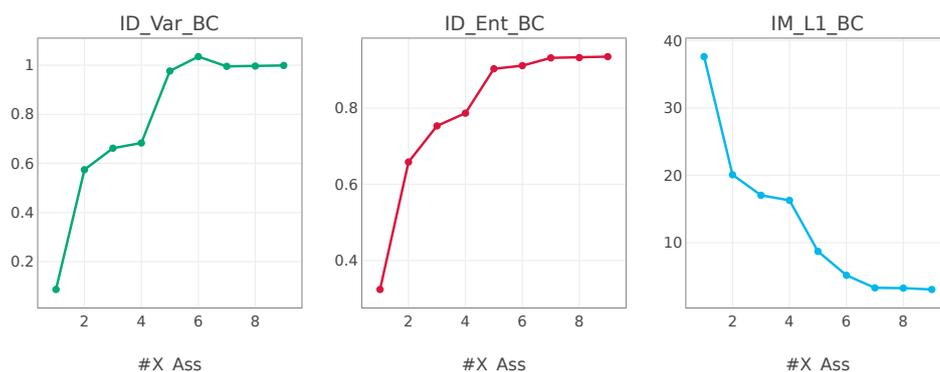


FIGURE 4.16 – Évolution du niveau de démutualisation lorsque :  $X_{Ass} \rightarrow X_\Omega$ .

En rapprochant ces indices de démutualisation aux mesures d'adéquation (figure 17 en annexe), il apparaît que le nombre optimal de variables à considérer est de 5 pour les métriques d'évaluation globale que sont la déviance, la part de variance expliquée et l'AIC. Par contre, pour les métriques issues de la communauté des Data Scientists et construites à partir des écarts entre

les prédictions individuelles et observations, l'ajout des variables au delà des 5 premières reste pertinent. Selon la vision de l'assureur : plutôt promoteur de la solidarité et de l'accessibilité de l'assurance ou promoteur du *juste prix individuel*, il pourra décider de se restreindre à au plus 5 variables dans sa stratégie tarifaire (également en fonction de la compétitivité sur le produit considéré, qui pourrait l'inciter à aller au delà pour se protéger du risque d'anti-sélection).

### 4.3.2 Comparaison du niveau de mutualisation entre assureurs et analyse des effets de la démutualisation

Si quantifier le niveau de mutualisation revêt un intérêt en interne pour l'assureur notamment quant au choix de sa stratégie tarifaire, l'appréciation et le pilotage de son risque de structure ; la communication sur la mutualisation du portefeuille et sur la stratégie tarifaire présente également un intérêt indéniable. D'une part, pour plus de transparence auprès de la communauté des assurés et d'autre part pour enrichir la discussion autour de la démutualisation sur le marché. La connaissance du niveau de mutualisation pratiqué par les assureurs sur un marché donné permettrait aux régulateurs d'éviter des dérives éventuelles ; quantifier la mutualisation représente donc un enjeu de place.

Cette partie est consacrée à la comparaison de la démutualisation induite par quatre assureurs possédant des stratégies tarifaires diverses. Les assureurs sont supposés disposer des mêmes variables tarifaires, mais ils n'utilisent pas exactement les mêmes variables pour évaluer leurs tarifs techniques (tableau 4.7). À ces quatre assureurs, sont rajoutés deux assureurs extrêmes (et fictifs) qui pratiquent respectivement une prime unique et la tarification la plus fine que les assureurs puissent atteindre. Les discrétisations des variables continues sont supposées communes à tous les assureurs et correspondent à celles du tableau 4.3. Le tableau suivant résume les variables utilisées par les assureurs sur ce marché.

TABLEAU 4.7 – Stratégies tarifaires utilisées par les assureurs pour la construction de la prime technique.

Assureur	Variables
$Ass_1$	FORMULE, TYPE_VEH, AGE_VEH_SOUS
$Ass_2$	FORMULE, TYPE_VEH, AGE_VEH_SOUS, PUISSANCE
$Ass_3$	FORMULE, TYPE_VEH, AGE_VEH_SOUS, 'TRANCHE_KM_SOUS'
$Ass_4$	FORMULE, TYPE_VEH, AGE_VEH_SOUS, MARQUE
$Ass_\emptyset$	$\emptyset$ - Aucune variable
$Ass_\Omega$	$X_\Omega$ - Toutes variables à disposition

Pour l'assureur n'utilisant pas de facteur de risque, le burning cost représente la moyenne historique pondérée par l'exposition des montants totaux des sinistres pris en charge par contrat.

Les modèles d'évaluation de la prime technique associée à ces assureurs sont calibrés sur un historique de sinistralité commun correspondant aux bases d'apprentissage et de test utilisées pour la calibration du modèle de calcul des profils de risque les plus fins. Le tableau 8 en annexe indique les performances de ces modèles sur les données d'apprentissage, les données test et sur les contrats couverts en 2020. Sur la base d'apprentissage ainsi que celle de test, la charge globale de sinistres prédite correspond effectivement à la charge totale observée, pour l'ensemble des quatre assureurs. Mais pour l'année 2020, tous les assureurs sur-estiment la charge de sinistres réelle. Ceci vient d'une amélioration de la sinistralité en 2020 due notamment à la crise sanitaire Covid-19 qui a limité les déplacements et *in fine* les pannes sur les véhicules.

#### 4.3.2.1 Comparaison du niveau de mutualisation en situation de monopole

Supposons dans un premier temps que chacun de ces assureurs soit en situation de monopole sur le marché et que tous les assurés sur le marché souscrivent bien au contrat qui leur est proposé. Dans ces conditions, les assurés paieraient les primes techniques distribuées comme le montre la figure 4.17 suivante en fonction de l'assureur.

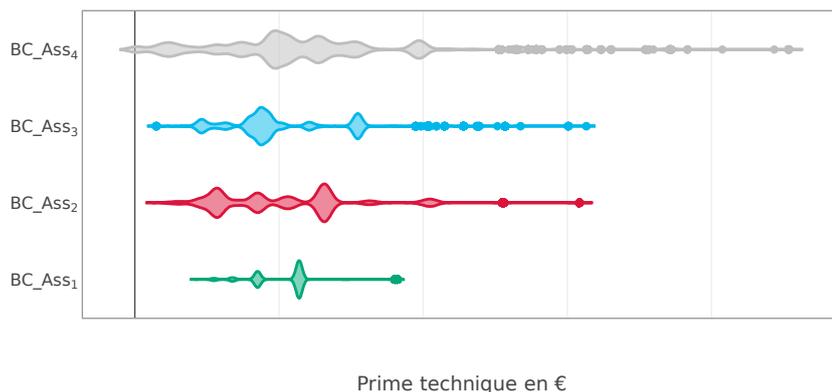


FIGURE 4.17 – Distributions des primes techniques commercialisées par les assureurs en situation de monopole.

**Note de lecture :** Les primes commercialisées par l'assureur 4 ■ sont tellement faibles que leur distribution représentée par un *violin plot* semble faire apparaître à tort des primes négatives.

#### Des structures de primes très différentes issues des stratégies de tarification

Les différences entre les distributions des primes commercialisées par ces assureurs sont assez prononcées. Alors que l'assureur 1 propose des primes parfaitement concentrées autour de la sinistralité moyenne, l'assureur 4 propose une gamme de prix plus étalée allant d'un montant minimal pratiquement nul à un montant maximal de plus de 200 €. Ces différences entre les vecteurs de primes laissent entrevoir une démutualisation prononcée pour l'assureur 4 contrairement à l'assureur 1. Calculés sur le portefeuille de 2020 pour chaque assureur, les indices de mutualisation consignés dans le tableau 4.8 viennent corroborer cette intuition graphique.

TABLEAU 4.8 – Indices de mutualisation vanilles de plusieurs assureurs sur le même portefeuille.

Indice	$Ass_1$	$Ass_2$	$Ass_3$	$Ass_4$	$Ass_\emptyset$	$Ass_\Omega$
ID_Ent_BC	0.131	0.199	0.216	0.360	0.000	0.942
IM_L1_BC	37.153	33.170	34.421	28.627	38.435	0.000
ID_Var_BC	0.047	0.175	0.147	0.344	0.000	1.000
IM_Var_OL	0.943	0.827	0.866	0.634	1.000	0.000
IM_Ent_OL	0.908	0.909	0.908	0.906	0.942	0.000
IM_L1_BC_Shift	35.065	29.703	31.473	26.101	38.184	0.000
Prime	4659776	4529484	4603268	5050225	4827475	4878049
S/P	0.975	1.003	0.987	0.899	0.941	0.931
Résultat	118245	-12047	61737	508694	285943	336518

### Des indicateurs cohérents avec les intuitions actuarielles et entre eux

Les indicateurs simples sont cohérents entre eux puisqu'ils conduisent globalement au même classement des assureurs suivant le niveau de démutualisation. L'assureur 4 est celui dont la stratégie tarifaire démutualise le plus la communauté des assurés en opposition à l'assureur 1 pour lequel les indices de mutualisation sont supérieurs à ceux des autres assureurs. En effet, le montant moyen de prime solidarisée de l'assureur 1 (35,07 €) est très voisin de celui de l'assureur qui pratiquerait la politique de prime unique (38,18 €). Il utilise seulement 13,1 % de l'information nécessaire pour individualiser les risques, pratiquant ainsi un tarif pour près de 8 profils de risque. Enfin, il ne transfère que 4,7 % du risque de structure qu'il pourrait céder dans la tarification. Inversement, l'assureur 4 en incluant la variable marque de véhicule dans sa stratégie, aboutit à des tarifs bien plus individualisés que tous les autres assureurs. Il commercialise un tarif pour moins de 3 profils de risque différents avec un montant de prime solidarisée de 26 €. Entre ces 2 assureurs, se trouve les assureurs 2 et 3 pour lesquels les indicateurs sont assez comparables, mais l'assureur 2 pratique une segmentation plus fine que l'assureur 3 (en dépit du nombre de modalités de la tranche de kilométrage plus important).

Sans perte de généralité, les chargements assureurs sont supposés constants et arbitrairement fixés et l'OL nécessaire au calcul des indicateurs IM\_Var\_OL et IM\_Ent\_OL est déduite immédiatement. L'indicateur IM\_Ent\_OL est inchangé quelque soit l'assureur, mais il est source d'erreurs. En effet, les valeurs uniques dans les OL quel que soit l'assureur sont très nombreuses puisque l'OL est calculé par différence entre la prime commerciale (relativement agrégée) et le profil de risque  $BC_\Omega$  (très différencié entre les assurés). Dès lors, il devient hautement improbable d'obtenir plusieurs valeurs d'OL égales, même pour les assurés ayant pratiquement le même profil de risque. Le calcul de l'entropie ne prenant pas en compte directement les valeurs, mais leur fréquence d'occurrence, IM\_Ent\_OL est très proche de sa valeur maximale, mais à tort. Cet indicateur ne capte pas efficacement les traitements différenciés pratiqués par l'assureur. Une correction possible de cet indicateur consisterait à le calculer sur les chargements

assureurs, pour évaluer la mutualisation commerciale du portefeuille. Mais cette formulation nécessiterait la connaissance de la stratégie commerciale de l'assureur.

Finalement, à l'exception de l'indicateur IM\_Ent\_OL, les indices de mutualisation simples sont cohérents et informatifs : cohérents puisqu'ils conduisent aux mêmes classements des assureurs au regard de la démutualisation ; informatifs puisqu'ils permettent d'évaluer à quel point une stratégie tarifaire mutualise plus qu'une autre.

### Démutualiser permet à l'assureur de se couvrir contre le risque de structure

Les trois dernières lignes du tableau 4.8 sont consacrées à l'analyse des résultats financiers des assureurs en situation de monopole. Elles révèlent des ratios de sinistralité S/P au plus de 100,3 %. En effet, les assureurs 1, 3 et 4 ont un S/P très inférieur à 100 % avec un résultat technique bien plus favorable pour l'assureur 4 (S/P de 89,9 %). Ce résultat positif quelle que soit la stratégie est dû à l'amélioration de la sinistralité (incluant probablement un effet Covid-19) ou à une rupture de tendance. Mais pour comprendre pourquoi le résultat technique est plus important pour l'assureur 4, il convient d'observer l'évolution de la structure du portefeuille. En effet entre 2019 et 2020, le poids de la marque « Marque\_AR »<sup>1</sup> a augmenté assez rapidement comme le montre la figure 18 en annexe. L'assureur 4 ayant utilisé cette variable dans sa grille tarifaire s'est ainsi couvert contre cette évolution de la structure contrairement aux autres assureurs. De plus, l'analyse des résultats financiers par marque de véhicule pour les assureurs 1 et 4 (figure 4.18 ci-contre) permet de constater un résultat très positif pour l'assureur 4 sur cette marque.

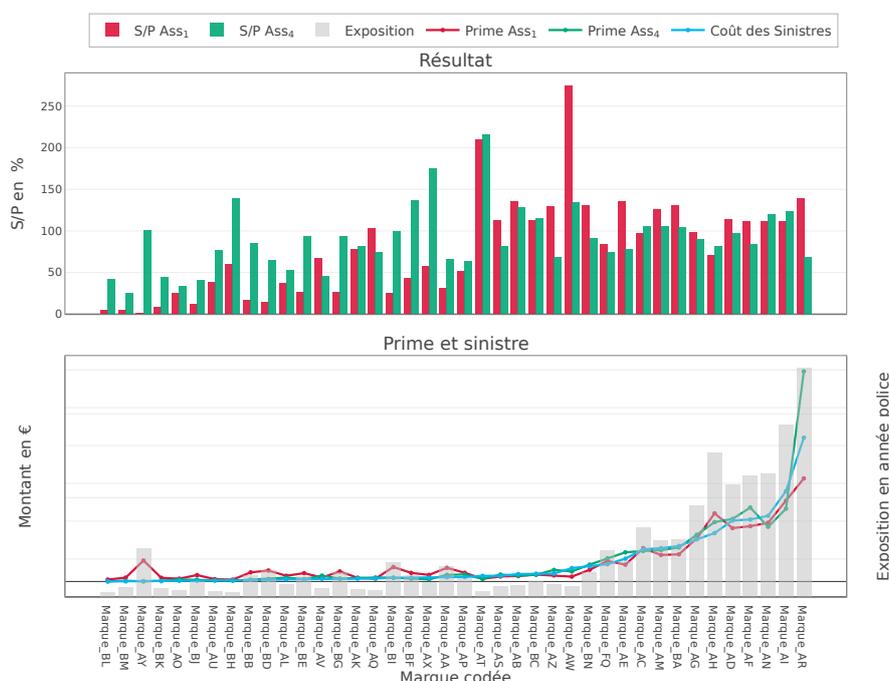


FIGURE 4.18 – Résultats financiers des assureurs 1 et 4 suivant les marques véhicules.

1. codage d'une marque utilisé pour lever la confidentialité

Il y a un effet croisé de rupture de tendance (amélioration de la sinistralité) et de changement de structure, dont l'assureur 4 tire profit en situation de monopole. Il fait payer plus cher (mais conformément à la sinistralité historique) les détenteurs de véhicules Marque\_AR qui ont été plus nombreux dans le portefeuille en 2020. Les autres assureurs bénéficient également de l'amélioration de la sinistralité puisque dans l'absolu, ils font face à une charge de sinistre moindre que celle escomptée. Mais cet excédent est limité par l'augmentation de ces véhicules de marque Marque\_AR qui sont en fait des mauvais risques (car la sinistralité attendue sous  $BC_{\Omega}$  et pour l'assureur 4 est élevée) auxquels les assureurs faisaient déjà payer moins cher grâce à la mutualisation. Dit autrement, en l'absence de l'amélioration de la sinistralité des mauvais risques Marque\_AR, à l'exception de l'assureur 4, tous les assureurs auraient subi des pertes techniques puisque leur équilibre suppose une fraction stable des marques Marque\_AR dans le portefeuille.

Il découle de cette analyse des résultats financiers que la démutualisation permet de se couvrir contre une déviation de la structure du marché : en cas d'augmentation des mauvais risques, elle limite les pertes techniques, chaque nouveau risque payant pour son *juste risque* ; en cas d'augmentation des bons risques, le résultat augmente également de façon proportionnelle.

#### **4.3.2.2 Comparaison du niveau de mutualisation en situation de concurrence**

---

La comparaison précédente se fait en situation de monopole et ne prend pas en compte les effets de la stratégie de tarification sur les contrats souscrits. Les indicateurs ont été calculés sur le même portefeuille alors qu'en pratique, les prix par le biais des mécanismes de la concurrence, peuvent influencer sur les souscriptions et ainsi modifier les structures des portefeuilles détenus par chaque assureur. Si pour certains produits, il est possible de considérer que la stratégie tarifaire n'aura qu'une influence modérée sur la structure souscrite (du fait d'un avantage concurrentiel lié à un canal captif de distribution par exemple), il n'en est rien pour l'essentiel des produits d'assurance.

Les produits d'assurance commercialisés sur les comparateurs font l'objet d'une concurrence importante et la structure des prix joue un rôle important à la souscription. Concrètement, les comparateurs vont trier les polices d'assurance suivant les prix croissants et un assureur qui propose des prix très élevés se retrouvera le plus souvent sur des pages secondaires et donc ne souscrira pas les bons risques. Pour se rapprocher de cette situation concurrentielle, les quatre assureurs précédents sont placés sur un marché concurrentiel sur lequel chaque assuré a une connaissance parfaite de tous les contrats proposés et souscrit à la police la moins chère pour lui. Il est alors admis implicitement que le prix est le seul élément susceptible d'influencer la décision d'un assuré, en particulier que les couvertures/exclusions sont toutes identiques ou que les différences sont suffisamment marginales pour ne pas jouer dans le choix de l'assuré. De même, d'autres éléments différenciants comme la qualité de la relation client, la fidélité à son l'assureur ou plus généralement l'existence des avantages concurrentiels ne sont pas pris en

compte.

### Exclusion des risques par la démutualisation en concurrence

En situation de concurrence, les distributions des primes techniques payées par les assurés aux différents assureurs sont données par la figure 4.19. L'assureur 4, c'est-à-dire le plus démutualisé, commercialise essentiellement des primes très basses avec une masse importante d'assurés qui paient des primes au voisinage de 10 €. Contrairement au prix en monopole, il ne souscrit aucun profil au-delà de 90 €. Sa stratégie très segmentée empêche les mauvais risques de choisir ses contrats à lui puisqu'ils sont très chers. Dit autrement, il choisit d'exclure par sa tarification les plus mauvais risques de son portefeuille. La proportion des assurés ainsi exclue est évaluée pour chacun des assureurs en déterminant le nombre d'assurés à qui l'assureur propose un contrat plus cher que la prime maximale commercialisée *in fine*<sup>2</sup>. Ces proportions d'exclusion sont consignées dans le tableau 4.9. Il en ressort que plus la segmentation est poussée, plus grande est la part des assurés exclus par les tarifs. L'assureur qui mutualise le plus (celui qui propose le prix maximal le plus faible) se retrouve à souscrire les profils exclus par les autres acteurs sur le marché. La course à la démutualisation pourrait alors conduire à une exclusion des profils de risque les plus à risque de l'assurance.

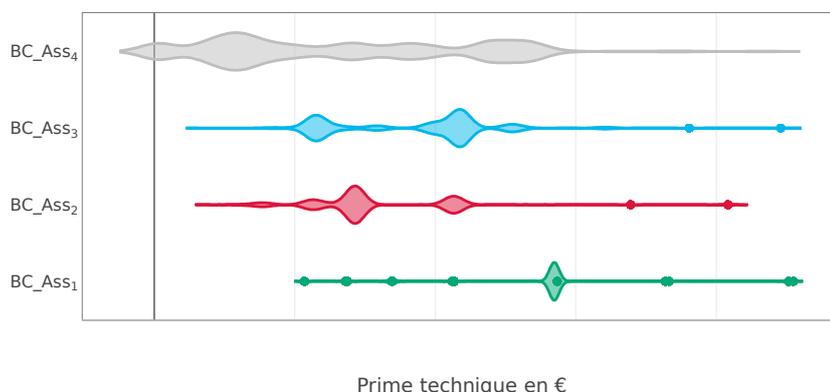


FIGURE 4.19 – Distributions des primes techniques commercialisées par les assureurs en situation de concurrence.

TABLEAU 4.9 – Proportion des assurés du marché exclus par la stratégie tarifaire segmentée.

Assureur	ASS_1	ASS_2	ASS_3	ASS_4
Exclusion (en %)	0,00	7,13	5,42	14,31
Prime max. souscrite (en €)	█	█	█	█
Prime max. proposée	1	1,89	1,76	2,61
Prime max. souscrite				

2. nombre d'assurés exclus par  $Ass_i = |\{\text{assuré}_j, BC_j^{Ass_i} > \text{Max}\{BC_k^{Ass_i}, \text{assuré}_k \text{ a souscrit chez } Ass_i\}\}|$  où  $|\cdot|$  désigne le cardinal

## Démutualisation globale du marché en situation de concurrence

---

Le tableau 4.10 apporte une quantification du niveau de démutualisation des assureurs en concurrence ainsi qu'une synthèse de la situation globale du marché. Les valeurs des  $ID\_Ent\_BC$  pour les différents assureurs n'évoluent pas drastiquement par rapport aux valeurs en situation de monopole. En effet, la mise en concurrence provoque un double effet :

- Une sélection par les prix : en souscrivant uniquement les profils sur lesquels il est compétitif, chaque assureur va récupérer une fraction plus ou moins hétérogène des assurés ;
- Une (dé)mutualisation sur un pool de risque restreint : sur les profils acquis par le mécanisme des prix, l'assureur fait jouer une mutualisation ou une démutualisation plus ou moins importante selon sa stratégie initiale.

Ainsi, il y a une baisse conjointe de l'hétérogénéité des profils de risque en portefeuille et des tarifs effectivement commercialisés aboutissant d'où l'évolution contenue de l' $ID\_Ent\_BC$ . Cet indicateur continue donc à rendre compte du niveau de démutualisation de la stratégie tarifaire, tout en prenant en compte les évolutions de la structure du portefeuille. Il indique en particulier qu'en situation de concurrence, l'assureur 4 pratique un tarif pour 2,6 assurés soit une démutualisation plus importante qu'en situation de monopole. Le même indicateur baisse chez l'assureur 1 puisqu'il est désormais à 0,086 (contre 0,131) traduisant ainsi que peu de prix de ce dernier sont attractifs.

Les pools de risque par portefeuille étant restreints, il devient plus difficile de transférer le risque de structure dans la tarification puisque les assurés en portefeuille tendent à se ressembler. La volatilité des primes diminue alors, aboutissant à une baisse mécanique de l'indice  $ID\_Var\_BC$ . La diminution est d'autant plus importante que les prix pratiqués sont solidaires puisque l'assureur 4 continue à être l'assureur le plus démutualisé au sens de cet indice alors que l'assureur 1 demeure l'assureur le plus mutualiste.

S'agissant du montant moyen risque transféré entre assurés ( $IM\_L1\_BC$ ), il est impérieux de considérer sa forme corrigée puisque le résultat technique attendu est déficitaire pour tous les assureurs (du fait notamment d'une déformation de structure - voir détails plus loin dans cette partie). Les montants de primes solidarisées baissent chez tous les assureurs comparativement à leurs valeurs en situation de monopole et se situent désormais autour de 8,76 €. La sélection des profils par les prix s'accompagne naturellement d'une baisse des montants cross financés entre assurés puisque les assurés payent des primes qui sont : soit proches de leur profil de risque individuel (comme chez l'assureur 4), soit distants de leur profil de risque en restant similaires (que des mauvais risques et donc pas d'assurés pour contre financer l'excédent de risque sous-évalué) - comme chez l'assureur 1.

De façon plus sommaire, la course à l'hyper-segmentation en situation de concurrence entraîne une démutualisation globale du marché. L'analyse des indicateurs de marché s'aligne bien avec cette aggravation de la démutualisation déjà entrevue chez les assureurs. Les assureurs proposent un tarif pour un peu plus de 3 profils de risque (très proche d'un marché monopolistique

avec l'assureur 4 comme unique acteur) pour un transfert de risque entre assurés de 8,76 € soit un montant moindre quelque soit la situation de monopole.

TABLEAU 4.10 – Indices de démutualisation vanilles de plusieurs assureurs sur un marché concurrentiel.

Indice	ASS_1	ASS_2	ASS_3	ASS_4	Marché
ID_Ent_BC	0.086	0.147	0.171	0.384	0.304
IM_L1_BC	52.769	28.484	30.457	16.337	28.469
ID_Var_BC	0.030	0.047	0.043	0.195	0.087
IM_Var_OL	0.922	0.747	0.849	0.594	0.724
IM_Ent_OL	0.992	0.900	0.978	0.913	0.943
IM_L1_BC_Shift	7.926	7.723	10.896	8.148	8.759
$\Delta$ IM_L1_BC_Shift monopole	-27.14	-21,98	-20,58	-17,95	-
Part_Marche (%)	11.780	31.880	28.100	28.241	100.000
Primes	587156	893783	911934	666104	3058979
Expected Loss Ratio (ELR = $\frac{BC_{\Omega}}{P}$ )	1.832	1.694	1.549	1.315	1.595
S/P	1.775	1.345	1.493	1.404	1.485
Résultat	-454980	-308791	-449367	-269412	-1482551

### La démutualisation engendre une destruction de la valeur pour le marché

L'analyse des résultats techniques de marché donnés par les dernières lignes du tableau 4.10 révèle un ratio de sinistralité de 149 % largement supérieur à ses valeurs pour les différents monopoles. La course à la démutualisation en situation de concurrence fait en conséquence perdre de la valeur à tout le marché (et aux différents assureurs). En effet, si la sinistralité globale reste la même, l'assiette de prime du marché baisse puisque chaque assuré va choisir la police la moins chère pour lui sans prendre en compte son apport pour le marché global. À titre d'exemple, les bons assurés qui *pourraient* financer le marché vont souscrire aux contrats les moins chers chez l'assureur hyper-segmenté, mais les contrats destinés aux mauvais risques chez le même assureur seront boycottés par ceux-ci qui trouveront moins chers chez la concurrence. Au final, les bons risques paient leur *juste prix* et les mauvais risques un montant sous-évalué, mais sans contrepartie pour pallier ce déficit.

Pour mieux comprendre la dégradation de la valeur imputable à chaque assureur, les assureurs sont mis en concurrence en triplet en excluant à chaque fois un des assureurs. Pour les marchés ainsi obtenus, les situations globales sont analysées et les S/P de marché sont représentés sur la figure ci-contre. Le constat flagrant est que l'arrivée d'un nouvel assureur hyper-segmenté provoque une diminution de la valeur plus importante que celle d'un assureur mutualisé. En effet, sans l'assureur 4, le S/P de marché est à 128 % (figure 4.20) soit une dégradation de la valeur client de 20 points de pourcentage consécutive à son arrivée sur le marché. À sa suite, se retrouve l'assureur 2 qui est le second le plus démutualisé avec une perte

de valeur de 16 points. À l'inverse l'arrivée de l'assureur 1 ne détériore pratiquement pas la valeur de marché puisque sans lui, le S/P ne réduit que de 4 points.

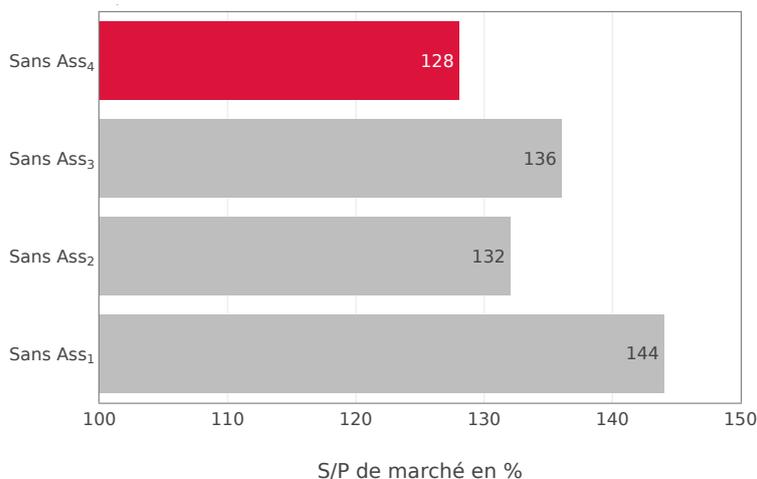


FIGURE 4.20 – Valeur de marché (S/P) en situation de concurrence entre trois assureurs.

**Remarque : Approche alternative - destruction de valeur**

Pour analyser la destruction de valeur pour le marché consécutive à l'arrivée d'un nouvel acteur sur le marché, on pourrait également considérer toutes les situations concurrentielles à 1, 2, 3 ou 4 assureurs dans lesquelles intervient cet assureur afin d'identifier sa contribution à la valeur de marché; une mesure plus robuste de la détérioration de la valeur en découlerait alors. Cette approche s'inspire du calcul des valeurs de Shapley présenté dans la partie 2.4.1.1. Pour éviter d'alourdir cette partie, cette méthode n'a pas été adoptée.

Les analyses précédentes soulignent des conséquences plutôt négatives de la course à la démutualisation sur le marché pour ce cas pratique. Ces bilans concernent l'exclusion des mauvais risques, la démutualisation globale du marché et la dégradation de la valeur client de marché. Ces conséquences sont d'autant plus graves que les assureurs démutualisent plus finement leurs tarifs. Le cadre d'analyse reste tout même assez simple puisque la concurrence exclusivement sur les prix ne reflète pas d'autres aspects bénéfiques pour les clients (qualité de service, pression à conserver des ratios de frais bas etc.).

Une analyse des performances financières par assureur révèle également des effets négatifs pour les assureurs. En effet, dans cet exemple, tous les assureurs obtiennent un résultat technique déficitaire comme le montre le tableau 4.10. Pour illustrer les causes de ces contreperformances financières, une compréhension des profils de risque (déterminés par  $BC_{\Omega}$ ) souscrits par chaque assureur (dont la distribution est donnée par la figure 4.21 ci-contre) ainsi qu'une analyse des montants de sinistres pris en charge par assureur (figure 19 en annexe) sont nécessaires.

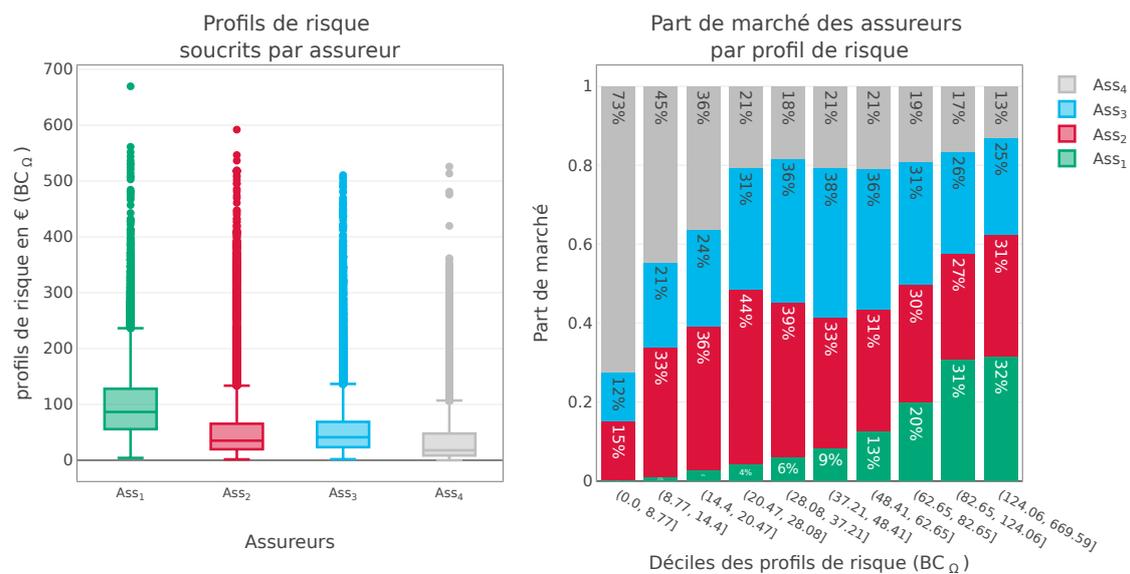


FIGURE 4.21 – Répartition des parts de marché par profils de risque souscrits par assureur en concurrence.

### Sur un marché concurrentiel, l'absence de segmentation n'est pas envisageable

L'assureur le plus mutualisé est celui qui a le résultat technique le plus déficitaire avec le ratio de sinistralité le plus élevé de tout le marché. Il est en effet victime de la sélection adverse bien connue et illustrée par plusieurs auteurs dans des travaux aussi bien empiriques que théoriques. En proposant des prix relativement constants, cet assureur n'est pas compétitif sur les bons risques comme le montre la figure 4.21. Au contraire, il ne souscrit que les mauvais risques à des tarifs sous-évalués. Il n'acquiert pas assez de prime pour prendre en charge les sinistres élevés observés sur ces derniers (figure 19 en annexe). Les pertes techniques sont dès lors inéluctables.

### L'hyper-segmentation ne garantit pas la sur-performance financière

Bien qu'en vision attendue l'assureur le plus démutualisé soit le plus performant financièrement, force est de constater qu'en fin d'exercice, son résultat est plus déficitaire qu'escompté et il n'a pas le meilleur S/P du marché. Les causes de ce résultat inattendu seraient liées aux mécanismes inhérents à une tarification hyper-segmentée :

- L'assureur le plus démutualisé souscrit l'essentiel des très bons risques du marché en leur faisant payer des primes extrêmement faibles. En effet il récupère 73 % des 10 % des assurés ayant la sinistralité attendue la plus faible. Les très bons risques constituent alors une part importante de son portefeuille et son assiette de prime résultante est relativement faible, elle représente 21,78 % des primes du marché contre des parts de marché de 28,24 %.
- En segmentant plus finement son tarif que les autres assureurs, il est capable de séparer

les « moins mauvais risques ». Il est notable en effet que cet assureur se positionne sur quelques mauvais risques également. Mais son modèle étant très segmenté, la volatilité des sinistres sur de telles classes est importante puisqu'il s'agit de petites niches qui ne sont que faiblement peuplées.

Ainsi, l'assureur 4 souscrit majoritairement les très bons risques à des tarifs faibles, ce qui certes lui permet de financer les sinistres sur cette classe, mais pas de dégager une marge technique (suffisante) pour cross financer la volatilité des sinistres sur les niches de bons parmi les mauvais risques souscrits également. Il serait ainsi nécessaire que l'assureur le plus démutualisé mette en œuvre des politiques de souscriptions restrictives pour limiter les mauvais risques captés, renforçant alors les exclusions des personnes ayant le plus *besoin* d'assurance.

### **Un niveau de mutualisation intermédiaire semble être le bon compromis**

---

L'assureur qui réalise le meilleur S/P dans cet exemple est l'assureur 2. Deux observations imbriquées sur son portefeuille permettent de comprendre cette sur-performance : une démutualisation contenue et l'amélioration de la sinistralité sur les mauvais souscrits. D'une part, les profils de risque souscrits par cet assureur recouvrent bien l'ensemble des profils de risque du marché avec une présence assez prononcée sur les risques intermédiaires. À la souscription, les prix pratiqués par cet assureur sont plus proches de ceux pratiqués par les assureurs les plus mutualistes 1 et 3 (comme le montre l'indicateur de la prime solidarisée avant correction du résultat technique attendu) mais avec des marges de solidarité près de deux fois supérieures à celle de l'assureur 4. Il est donc *a priori* bien couvert contre le risque de volatilité sur des niches de risque contrairement à l'assureur le plus démutualisé. D'autre part, l'analyse du ratio des sinistres observés sur les sinistres attendus à iso profil de risque met en évidence une amélioration de la sinistralité importante sur les très mauvais risques du portefeuille (figure 4.22 ci-contre). En 2020, les sinistres observés sur les très mauvais risques ne représentent que 86 % de ceux anticipés par la modélisation historique du profil de risque. L'assureur 4 en proposant des tarifs très segmentés, n'est pas compétitif sur ces contrats qui se rabattent sur des assureurs ayant des prix plus mutualisés en particulier chez les assureurs 1 et 2. Cette rupture de tendance est alors favorable à l'assureur 2 et constitue un manque à gagner pour l'assureur 4. Ce manque à gagner est amplifié par l'augmentation du poids des très mauvais risques sur le marché durant la même année.

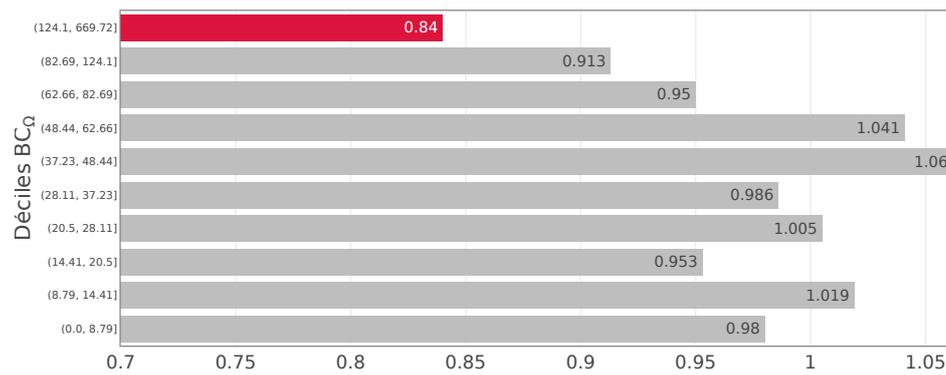


FIGURE 4.22 – Analyse de l’amélioration des sinistres suivant la classe de risque.

Dans le cas hypothétique d’une rupture de tendance défavorable, typiquement, une augmentation de la sinistralité sur une classe de risque spécifique, l’ensemble des assureurs ayant souscrits ces profils en souffriraient et l’équilibre global du portefeuille tiendrait à la capacité des autres classes de risque à remédier à ce déficit. Or en cas de démutualisation importante, il n’y a point ou pas assez de marge technique sur les autres classes puisque les assurés cotisent des montants proches de leur profil de risque individuel. Sur la base de cet exemple empirique, il ressort que la démutualisation à l’extrême rend le résultat plus sensible à une rupture de tendance (un manque à gagner en cas de rupture favorable et des pertes financières en cas de rupture défavorable). Cette sensibilité est accentuée lorsque la rupture de tendance s’accompagne d’une déviation de structure conjointement.

#### 4.3.2.3 Stabilité des indicateurs au regard des fluctuations de marché

Les résultats précédents ont été obtenus sur un échantillon de données, bien qu’ils s’emboîtent parfaitement avec les intuitions actuarielles (biens connues et moins apparentes), leur généralisation n’est pas immédiate comme pour toute étude empirique. En réalisant des tirages successifs de 50 % des contrats couverts en 2020, il en découle un portefeuille de marché qui fluctue d’un tirage à l’autre. Ainsi, sur la base de 100 tirages bootstraps, la figure 4.23 permet d’analyser la stabilité des indicateurs de démutualisation simples et celle du S/P au regard des fluctuations d’échantillons.

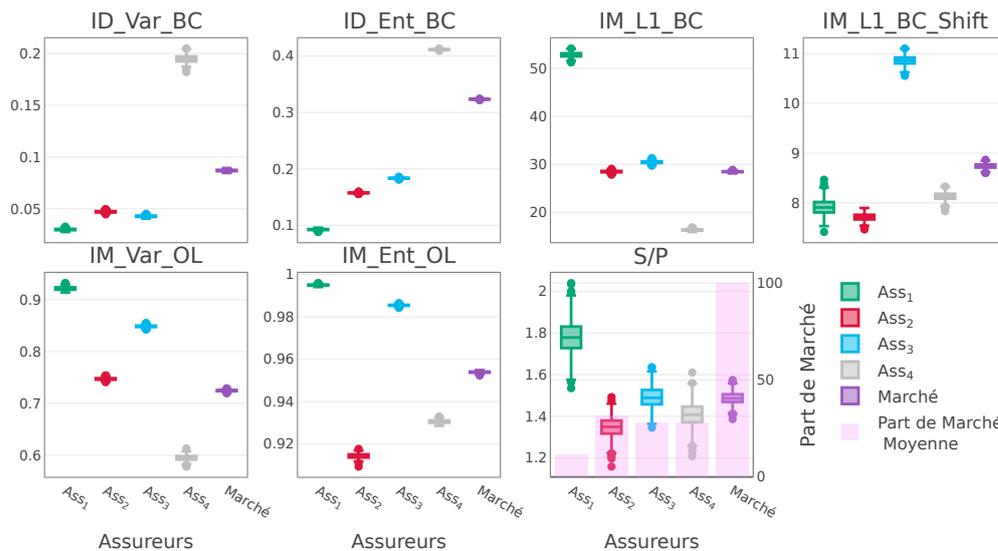


FIGURE 4.23 – Stabilité des indicateurs de démutualisation sur 500 échantillons bootstrap des contrats de 2020

Les indicateurs sont robustes aux fluctuations d'échantillonnage puisque leurs valeurs sur ces différents tirages sont extrêmement concentrées autour de leur valeur médiane. Si des dispersions plus marquées du S/P sont observées, celles-ci restent en ligne avec les résultats précédents à savoir que l'assureur très mutualisé court au suicide financier ; l'assureur très segmenté ne dispose pas de la garantie d'être le meilleur financièrement ; dans de nombreux cas, l'assureur disposant d'un niveau de mutualisation intermédiaire réalise un meilleur résultat.

#### 4.3.2.4 Discussions sur les évolutions possibles des assureurs et du marché

Le modèle de concurrence analysé dans cet exemple reste relativement simple. Il suppose une concurrence sur les primes techniques exclusivement sans modéliser les considérations commerciales. Il ne prend pas non plus en compte les spécificités de distributions qui pourraient conférer des avantages concurrentiels à certains assureurs. Enfin, l'analyse se fait sur une période unique. Tous les assureurs étant en perte technique ; quelques scénarios possibles d'évolutions des stratégies des assureurs, peuvent être envisagés :

- Une augmentation générale des tarifs techniques qui serait *in fine* plus défavorable pour toute la communauté des assurés qu'une situation de primes mutualisées ;
- Une exclusion par des pratiques commerciales à la souscription des mauvais risques par les assureurs les plus démutualisés, les contraignant à se replier sur les autres assureurs qui n'auraient d'autres choix que d'augmenter leur tarif puisqu'ils n'ont plus les bons risques pour financer leur portefeuille. Pouvant conduire de proche en proche à une exclusion des plus mauvais risque de l'assurance ;
- La mise en place des stratégies tarifaires plus segmentées par les assureurs les plus mutualisés qui accusent les pertes les plus importantes afin de capter également les bons risques.

De même que dans le scénario précédent, les mauvais risques pourraient se retrouver avec des primes trop élevées pour souscrire à des contrats d'assurance.

Sur la base de cet exemple empirique et ces scénarios dans un cadre simplifié, il ressort que segmenter permet à l'assureur de mieux comprendre le risque assuré et de se protéger contre la sélection adverse, mais les dangers associés à une démutualisation trop importante du portefeuille sont bien réels et préoccupants. Il serait donc judicieux de suivre, de comprendre et de mesurer la démutualisation des portefeuilles pour se prémunir des dérives et piloter ces risques de démutualisation.

## Conclusion du chapitre

Il était question dans ce chapitre d'élaborer le modèle de calcul de prime pure le plus fin que l'assureur puisse construire à partir des variables dont il dispose. Pour y parvenir, deux classes de modèles ont été mises en comparaison : des modèles basés sur le GLM et des modèles basés sur le boosting des arbres de régression. À la suite, les indicateurs vanilles de démutualisation ont révélé que le portefeuille d'extension de garantie automobile de BNP Paribas Cardif est assez mutualisé de par la stratégie tarifaire mise en œuvre. En particulier, le ratio d'individualisation est de 20,5 % et la prime de solidarité est de 34 €. Appliqués à différentes stratégies de tarifications, les indicateurs sont cohérents avec les intuitions actuarielles et conduisent aux rangs équivalents de mutualisation. Ils révèlent notamment que l'intégration de la variable marque démutualiserait près de 3 fois plus les tarifs techniques pratiqués.

Si en situation de monopole la démutualisation procure une couverture contre les risques de structure, en situation concurrentielle, il semblerait que le bon compromis soit une stratégie qui conduit à une mutualisation intermédiaire. Trop mutualiser expose à de la sélection adverse ; à l'inverse démutualiser à l'extrême rend sensible à la volatilité sur les niches puisqu'il devient impraticable de cross-financer cette volatilité sans marges de solidarité qui proviendraient des autres profils de risque. Le risque d'exclusion des mauvais risques de l'assurance est bien réel. L'assureur le plus démutualisé exclut mécaniquement les mauvais risques par sa tarification et pourrait également les exclure à la souscription pour améliorer son équilibre financier. De plus, cet assureur tire le marché vers une démutualisation globale et la course à la démutualisation détruit de la valeur pour les assurés.

Comme évoqué dans le chapitre 2, ces indices peuvent se généraliser en indices exotiques afin de limiter la nécessité de connaître tout le portefeuille pour évaluer le niveau de démutualisation. L'implémentation de ces indicateurs est l'objet du chapitre suivant.

## ÉVALUATION ET ANALYSE DE LA MUTUALISATION À PARTIR DES INDICATEURS EXOTIQUES

### Sommaire

<b>5.1</b> Mesure de la démutualisation à partir de l'inertie des contributions au risque des variables tarifaires . . . . .	<b>133</b>
<b>5.2</b> Mesure de la mutualisation à partir de l'entropie algorithmique : le ratio de mutualisation généralisé. . . . .	<b>140</b>

À la suite des indicateurs simples calculés au chapitre précédent, le niveau de mutualisation d'un portefeuille peut être évalué à partir des indicateurs exotiques proposés au chapitre 2. Ces derniers se présentent comme des généralisations de la mesure du risque de structure cédé ( $ID\_Var\_BC$ ) et du niveau d'individualisation ( $ID\_Ent\_BC$ ). Le présent chapitre est consacré au calcul numérique des indicateurs exotiques, il est de ce fait structuré en deux parties; chacune d'entre elle abordant l'un des indicateurs exotiques. Chacune de ces parties s'ouvre sur une description illustrée du calcul d'un indicateur exotique; elle s'attarde ensuite sur les applications et les interprétations de l'indicateur sur les quatre stratégies tarifaires introduites en fin de chapitre 4.

### 5.1 Mesure de la démutualisation à partir de l'inertie des contributions au risque des variables tarifaires

Comme décrit à la section 2.4.1, cet indice mesure la part de risque de structure cédée dans la prime technique à travers des choix tarifaires de l'actuaire. Il s'agit donc d'un indice de démutualisation qui devrait croître à mesure que l'actuaire utilise l'information à sa disposition dans le tarif technique.

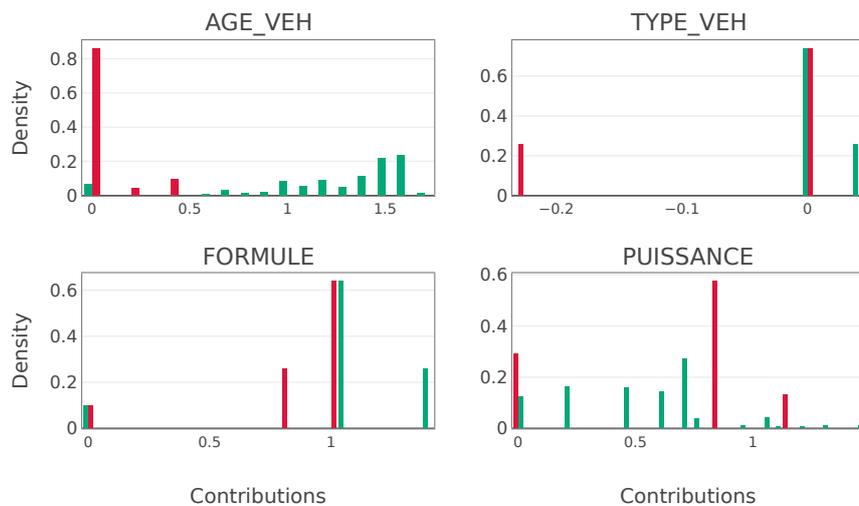
#### 5.1.1 Calcul et visualisation des contributions au risque

Les contributions au risque sont déterminées à partir des coefficients des modèles GLM estimés au chapitre 4. Celles-ci mesurent pour chaque assuré à quel point les modalités observées sur chaque variable contribuent à la prédiction linéaire du risque (le logarithme du burning cost). Lorsque la modalité réalisée pour un assuré est la modalité de référence (celle pour laquelle aucun coefficient n'est estimé), la contribution est arbitrairement fixée à zéro.

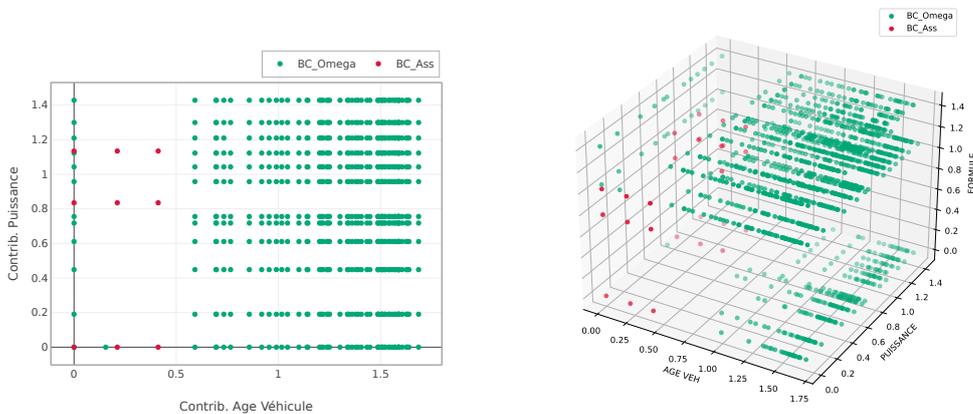
Les contributions au risque pour les contrats couverts en 2020 sur le portefeuille d'extension

de garantie automobile sont représentées sur la figure 5.1. Ces contributions sont calculées pour le modèle de prime technique ( $BC_{Ass}$  en rouge) et pour le modèle d'évaluation du profil de risque individualisé ( $BC_{\Omega}$  en vert).

La visualisation en une dimension, c'est-à-dire variable par variable (figure 5.1a) illustre comment considérer un nombre limité de modalités dans le modèle de tarification agrège les contributions au risque de la variable. En particulier pour les variables âge et puissance du véhicule, les contributions sont limitées à trois valeurs pour le modèle tarifaire (associées aux trois modalités retenues) contre des contributions bien plus disparates à l'échelle du profil de risque individualisé.



(a) Contributions 1D



(b) Contributions 2D

(c) Contributions 3D

FIGURE 5.1 – Illustrations des contributions des facteurs de risque sur le portefeuille de d'extension de garantie automobile

Les représentations multidimensionnelles des contributions quant à elles placent les assurés du portefeuille dans un espace où les axes sont les contributions des facteurs de risque. En particulier dans l'espace défini par les contributions des facteurs âge et puissance (figure 5.1b), un point rouge correspond à l'ensemble des assurés qui ont le même couple de contribution (âge, puissance) dans le modèle de calcul de prime technique ; de même, un point vert correspond à

l'ensemble des assurés qui ont le même couple de contribution (âge, puissance) dans le modèle d'évaluation du profil de risque individuel. Il est immédiat que dans le modèle de calcul de la prime technique, les assurés sont concentrés en quelques points contrairement au modèle utilisant toute l'information disponible. En rajoutant la formule comme nouvelle dimension, la visualisation des contributions en dimension 3 (figure 5.1c) reste en ligne avec cette agrégation des risques en passant du burning cost individualisé au burning cost commercial.

L'inertie est utilisée pour quantifier combien le nuage des assurés est disparate dans l'hyper-espace dont les dimensions sont les facteurs de risque utilisés par le modèle. Plus les contributions sont disparates, plus la modélisation du burning cost est fine et inversement.

## 5.1.2 Part de risque de structure généralisée comme mesure de démutualisation

### 5.1.2.1 Application au portefeuille BNP Paribas Cardif

Le tableau 5.1 donne une synthèse de l'hétérogénéité des primes issues des modèles de calcul des burning costs. La dernière colonne évalue la proximité entre la segmentation tarifaire adoptée et la plus fine accessible pour BNP Paribas Cardif. En effet, la stratégie tarifaire adoptée regroupe les 315 modalités de risque utilisées pour évaluer le profil de risque individuel en seulement 7 modalités. Le rapport de la variance des primes techniques sur la variance des profils de risque individualisé est de 0,175, soit un transfert de 17,5 % du risque de structure cessible. En remontant ce calcul au rapport des dispersions des contributions au risque, la part de risque de structure transférée généralisée vaut 18,4 %. Cette augmentation de près d'un point de pourcentage par rapport à l'indice calculé sur les burning costs est due aux compensations entre les facteurs des risques différents (partie 2.4.1). Néanmoins, les deux valeurs restent très comparables, ce qui conforte la quantification du transfert de risque associé à la finesse de la segmentation.

TABLEAU 5.1 – Mesure de la démutualisation sur le portefeuille à partir de l'inertie des contributions au risque

	$\Phi^{Ass}$ - Prime Technique	$\Phi^{\Omega}$ - Profil de risque	Rapport
Nombre de facteurs	4	10	0,4
Nombre de coefficients	7	315	0,022
Var. Burning cost	377	2150	0,175
Inertie des contributions	0,292	1,587	<b>0,184</b>

De même que pour sa contrepartie en indicateurs simples, la sensibilité de cet indicateur suivant la taille du portefeuille est observée numériquement. Quelle que soit la taille du portefeuille, la valeur obtenue pour cet indicateur ne varie que très peu comme le montre la figure

5.2a. Un regard est également porté sur la dynamique de l'indicateur lorsque les facteurs de risque tarifaires utilisés sont de plus en plus nombreux (figure 5.2b). Le comportement reste similaire à la part de risque calculée à partir des facteurs des burning costs directement : une démutualisation croissante avec la finesse de la segmentation, mais dont l'évolution est erratique. Il convient également de souligner que l'indicateur n'est pas nécessairement normalisé entre 0 et 1 comme escompté. Ceci pourrait s'expliquer par le fait que les inerties en présence sont calculées dans deux espaces différents et rien ne garantit *a priori* que l'inertie des contributions dans l'espace des variables utilisées pour la tarification restera inférieure à celle de l'espace de tous les facteurs de risque. Il serait souhaitable de ramener les deux nuages de contribution dans le même espace pour espérer observer une meilleure normalisation.

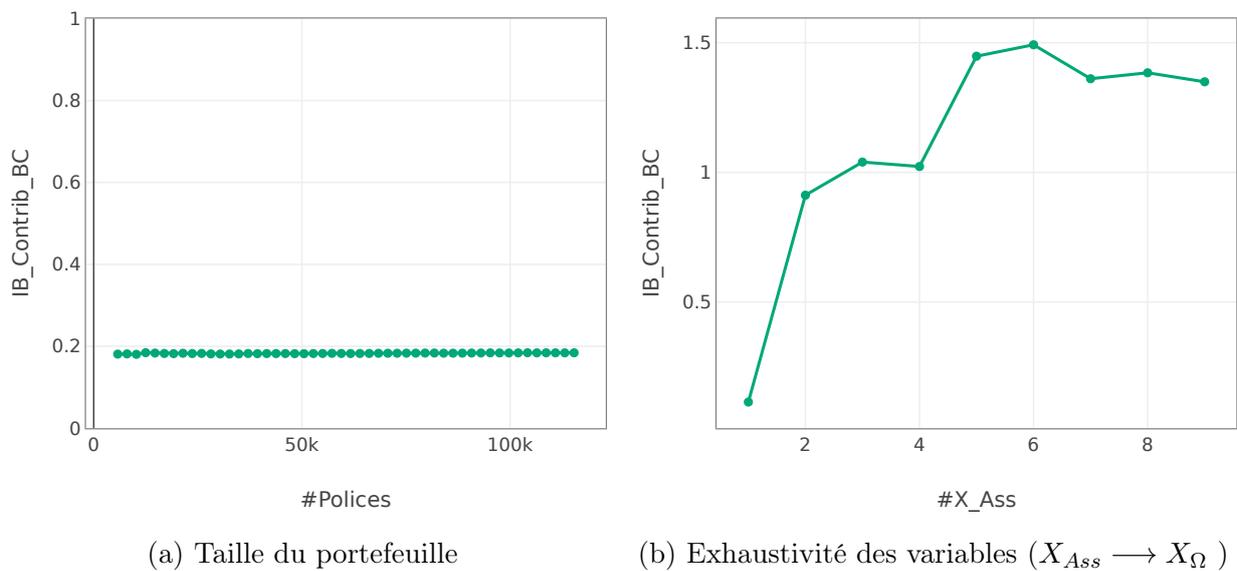


FIGURE 5.2 – Évolution de la part de risque de structure cédée généralisée lorsque la taille du portefeuille augmente (a) et lorsque  $X_{Ass}$  s'enrichit (b)

### 5.1.2.2 Comparaison des stratégies de tarification en monopole

Les quatre stratégies de tarification introduites au chapitre 4 sont reprises ici afin de les trier suivant le niveau de démutualisation qu'elles induiraient sur le portefeuille en situation de monopole. Les valeurs de la part de risque de structure cédée généralisée accompagnée du rang de démutualisation pour chaque stratégie sont consignées dans le tableaux 5.2. La forme simple de cet indicateur est également reportée dans le même tableau. Cet indicateur identifie correctement les deux stratégies les plus extrêmes : la stratégie tarifaire de l'assureur 4 (intégrant la marque du véhicule dans la grille tarifaire) est identifiée comme la plus démutualisée ; à l'opposé la stratégie de l'assureur 1 est celle qui conduirait à la plus grande mutualisation du portefeuille le plus en situation de monopole. Contrairement à l'indicateur simple, cet indicateur ne permet pas de classer efficacement l'assureur 2 par rapport à l'assureur 3. En effet, la stratégie de l'assureur 2 segmente davantage que celle de l'assureur 3, puisqu'elle utilise la puissance qui est un facteur de risque plus impactant que le kilométrage à la souscription.

TABLEAU 5.2 – Comparaison des stratégies de démutualisation en monopole par les mesures de transfert de risque de structure.

Assureur	ID_Inertie_Contrib.	Rang ID_Inerti_Contrib.	ID_Var_BC	Rang ID_Var_BC
Ass <sub>1</sub>	0,086	4	0,047	4
Ass <sub>2</sub>	0,185	3	0,175	2
Ass <sub>3</sub>	0,388	2	0,147	3
Ass <sub>4</sub>	0,996	1	0,344	1

Pour ces quatre stratégies, les valeurs de la forme généralisée de l'indicateur augmentent par rapport à sa forme simple. Cette augmentation pourrait s'expliquer par les différences importantes entre les assurés cachés derrière des risques équivalents au regard de la prime technique. Dans le même sillage, l'accroissement de la démutualisation induit par le passage de la stratégie de l'assureur 1 à celle de l'assureur 4, mesuré par le transfert de risque de structure, passe de 5,5 pour l'indicateur simple à 10,58 pour sa forme exotique.

### 5.1.2.3 Comparaison des stratégies de tarification en situation de concurrence

Les quatre assureurs précédents sont mis en concurrence comme au chapitre précédent. Sur les portefeuilles constitués à travers les mécanismes concurrentiels, le niveau de démutualisation des assureurs est évalué et consigné dans le tableau 5.3. Ces valeurs sont également accompagnées des valeurs et des rangs de  $ID_{Var||BC}$ . Les rangs des assureurs suivant l'indice de démutualisation basé sur l'inertie des contributions au risque sont conservés, mais les ordres de grandeurs des indicateurs pour les assureurs 3 et 4 sont modifiés. Alors que pour l'assureur 3, la valeur de l'indice baisse pour se rapprocher de celle de l'assureur 2 comme c'est le cas pour les indicateurs simples (comme la part du risque de structure cédée), la valeur pour l'assureur 4 augmente fortement pour atteindre 1,901.

TABLEAU 5.3 – Comparaison des stratégies de démutualisation en concurrence par les mesures de transfert de risque de structure.

Assureur	ID_Inertie_Contrib.	Rang ID_Inerti_Contrib.	ID_var_BC	Rang ID_Var_BC
Ass <sub>1</sub>	0,089	4	0,030	4
Ass <sub>2</sub>	0,197	3	0,047	2
Ass <sub>3</sub>	0,232	2	0,043	3
Ass <sub>4</sub>	1,901	1	0,195	1

Il ressort des applications précédentes que les propriétés de l'indice de démutualisation calculé comme le ratio des inerties des contributions au risque, partage les propriétés de la part du risque de structure cédée dans le tarif. Ils sont, en effet, semblables en terme d'interprétation et de dynamique lorsque la segmentation devient de plus en plus fine. De plus, ce nouvel indice apporte une meilleure compréhension de l'importance des facteurs dans la démutualisation ; il

permet en effet de mieux comprendre les effets sous-jacents de la mutualisation. En se basant sur les contributions au risque, en amont de leur agrégation en burning cost, il permet de retranscrire l'hétérogénéité des profils mutualisés. Toutefois, l'échec de sa normalisation rend difficilement comparable sa valeur entre plusieurs assureurs. Un axe d'exploration possible serait de recalculer cet indicateur en évaluant les contributions au burning cost directement (et non au logarithme du burning cost comme c'est le cas actuellement). Néanmoins, l'orientation privilégiée *a priori* consisterait à replacer les contributions du modèle de tarification adopté dans l'hyper-espace des contributions au profil de risque. Le ratio des inerties des nuages des contributions considérés dans le même hyper-espace correspondrait effectivement à une mesure du transfert de risque.

Aucune de ces deux approches n'est implémentée dans ce mémoire. En effet, cet indicateur est très similaire à la mesure de la part du risque de structure transférée dans le tarif. Même s'il devrait supprimer les compensations mécaniques liées à l'agrégation des contributions en burning cost qui donne l'impression de plus mutualiser, cet indicateur ne permet pas de s'affranchir de la nécessité de modéliser le profil de risque individualisé, ni celle de connaître l'entièreté du portefeuille. Le gain escompté à essayer d'améliorer cet indicateur ne semble donc pas justifier le temps nécessaire pour y parvenir. Toutefois, la visualisation des assurés dans l'espace des contributions au profil de risque individualisé en les distinguant par assureur permet de faire une transition pour le dernier indicateur proposé dans ce mémoire.

La figure 5.3 ci-contre montre comment les assurés du marché, en situation de concurrence, se répartissent entre les différents assureurs. Cette représentation illustre les hétérogénéités des profils constitués en portefeuille par les assureurs. Tous les 4 assureurs souscrivent des véhicules pour lesquels le risque imputable à l'âge est faible ou élevé. Néanmoins, des spécificités existent par assureur suivant la marque et la puissance du véhicule qui sont des facteurs utilisés seulement par les assureurs 4 et 2 respectivement. Les assurés captés par l'assureur 1 ont majoritairement une contribution au risque portée par la puissance élevée et sont également des marques risquées. L'assureur 4 se positionne essentiellement sur les marques avec une sinistralité moindre, c'est-à-dire les véhicules qui ont des contributions suivant la marque faibles. L'assureur 2 se positionne sur tous les risques par marque, mais effectue une sélection (pas très restrictive) par la puissance. Enfin, les véhicules qui souscrivent chez l'assureur 3 ont des contributions suivant la marque assez variables, mais ont des classes de puissance dont l'impact sur la sinistralité est plus élevé que ceux qui vont chez l'assureur 2.

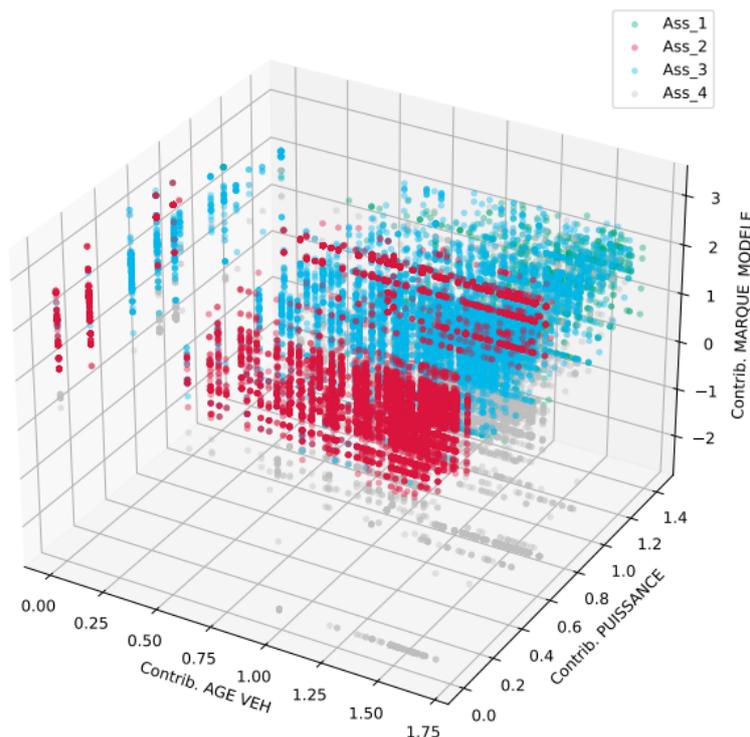


FIGURE 5.3 – Positionnement des assurés dans l’espace des contributions des facteurs au profil de risque individualisé selon l’assureur.

Replacer tous les assurés des portefeuilles d’assureur dans le même hyper-espace fournit une nouvelle perspective pour mesurer la mutualisation. En effet, un assureur qui a réussi à se positionner sur l’ensemble des risques du marché et à les faire vivre dans un même portefeuille aurait créé un portefeuille solidaire. Contrairement à un assureur qui se concentrerait sur des niches de risque spécifiques. Une approche intuitive serait d’évaluer à quel point le nuage des assurés d’un portefeuille donné recouvre le portefeuille de marché, en déterminant notamment le ratio d’inertie du nuage de l’assureur par rapport à celui du nuage du portefeuille de marché (c’est-à-dire de tous les assureurs). À noter que le centre de ces nuages devrait être partagé. Néanmoins, la mise en pratique d’un tel indicateur nécessiterait une connaissance excessivement détaillée (contributions au risque de toutes les modalités de tous les facteurs) d’un modèle commun d’évaluation du profil de risque individualisé ; tout comme une connaissance des portefeuilles des assureurs concurrents.

L’ultime indicateur dont la mise en œuvre pratique est réalisée dans la section suivante indique à quel point un portefeuille dont la description serait connue même partiellement recouvre les risques présents sur le marché, tout en réduisant les données nécessaires pour y parvenir.

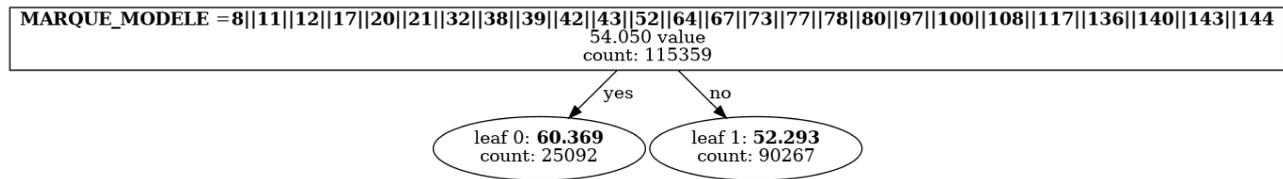
## 5.2 Mesure de la mutualisation à partir de l'entropie algorithmique : le ratio de mutualisation généralisé

Le ratio de mutualisation généralisé dont l'approche technique est présentée à la section 2.4.2, nécessite dans un premier temps d'élaborer un *splitter* de marché. Il s'agit d'une séquence de partitionnements des assurés du marché suivant les facteurs de risque les plus importants pour prédire le profil de risque. Dans ce mémoire, ces règles de partitionnement sont déterminées à partir d'un arbre de régression du profil de risque individualisé sur les facteurs de risque. Néanmoins, il n'est pas indispensable de réunir tous les portefeuilles des assureurs pour construire le *splitter*. Il peut également se décliner à partir d'une connaissance de la distribution des facteurs de risque et de leur impact. Plus le *splitter* est précis et riche, meilleur devrait être l'estimation du niveau de mutualisation produit par le ratio de mutualisation généralisé. Dans un cas d'absence de données pour prédire les profils de risque, il serait également envisageable de définir conjointement un *splitter* comme une référence normée (comme par convention), en listant par ordre d'importance les variables jugées les plus importantes pour mesurer la mutualisation.

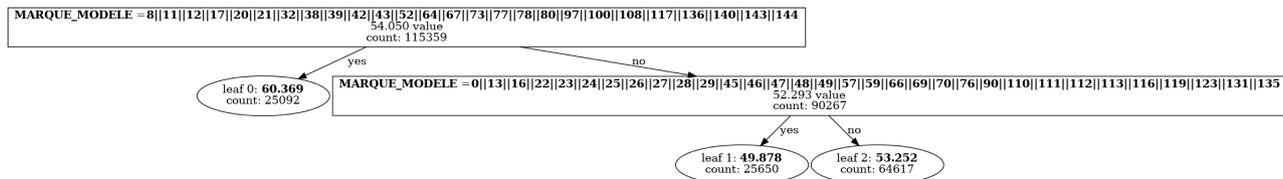
### 5.2.1 Construction du *splitter* à partir d'un arbre CART

La caractérisation de la structure des risques du marché est obtenue en régressant le profil de risque individuel  $BC_{\Omega}$  sur l'ensemble des facteurs de risque observés  $\mathbf{X}_{\Omega}$ . L'arbre est construit en croissance par nœud et non par étage (section 2.2.1.2). À chaque étape, le nœud contenant les risques les plus hétérogènes est partitionné à partir du facteur qui assure plus d'homogénéité dans les feuilles filles. Les feuilles terminales de l'arbre contiennent alors les risques identiques ou du moins indiscernables sous  $\mathbf{X}_{\Omega}$ . Sous  $\Omega$ , il devrait être possible d'atteindre des feuilles terminales ayant chacune un unique assuré.

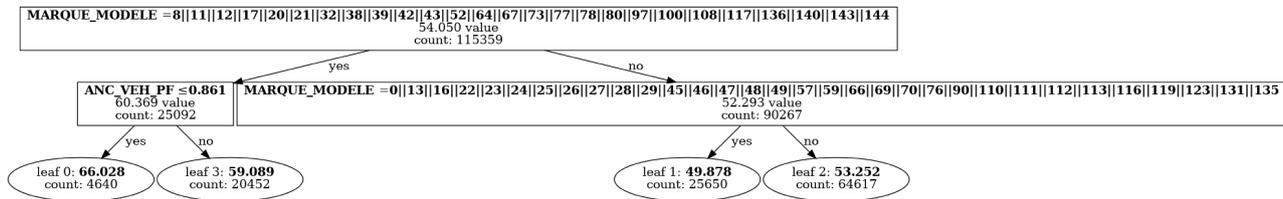
La figure 5.4 illustre la construction et la croissance du *splitter* calculé sur les contrats couverts en 2020. Le premier critère partitionne l'espace des facteurs en deux régions suivant l'appartenance ou non à une liste de modèles de véhicule. Le second critère précise un peu plus la région des véhicules n'appartenant pas à la liste spécifiée par le premier critère. Le troisième critère quant à lui partitionne la région des véhicules qui vérifient le premier critère suivant l'âge du véhicule au début du risque. L'extension de cet arbre à un peu plus de 4 000 critères permet de préciser complètement la distribution des facteurs ainsi que leurs impacts sur le risque à partir de 4 200 feuilles terminales ou régions de risque homogènes. La profondeur maximale de ce *splitter* est de 28 étapes.



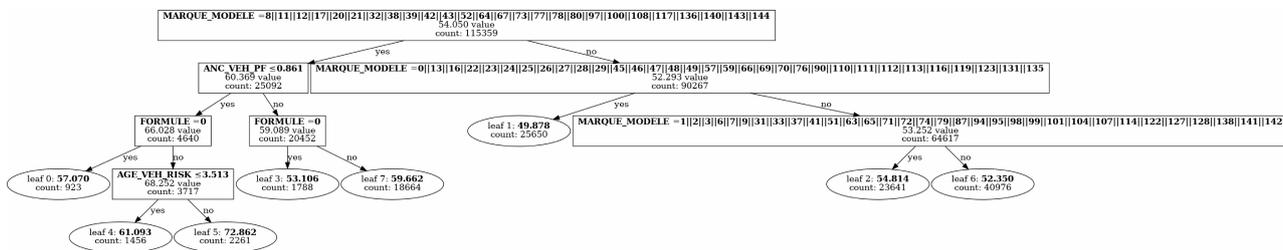
(a) Étape 1 : 1 critère et 2 régions



(b) Étape 2 : 2 critères et 3 régions



(c) Étape 3 : 3 critères et 4 régions



(d) Étape 7 : 7 critères et 8 régions

FIGURE 5.4 – Construction du *splitter* des risques sur les contrats couverts en 2020

De façon plus générale, le tableau 5.4 indique quelles sont les variables les plus utilisées dans les premiers critères. Il s'en dégage en particulier que les variables les plus importantes dans le modèle  $\Phi_\Omega$  sont celles qui sont les plus utilisées dans les premiers critères. Les variables âge au début du risque, marque  $\times$  modèle, puissance et ancienneté en portefeuille sont les plus utilisées dans les premiers critères. Inversement, les variables kilométrage à la souscription, motorisation, type de véhicule et carence ne sont pratiquement pas utilisées dans les premiers critères.

TABLEAU 5.4 – Fréquences relatives d'utilisation des facteurs de risque en fonction des critères cumulés

Critères	AGE_VEH_RISK	MARQUE_MODELE	PUISSANCE	ANC_VEH_PF	FORMULE	TRANCHE_KM,	MOTO,	TYPE_VEH	CARENCE
5,0	0,200	0,400	0,000	0,200	0,200	0,000	0,000	0,0	0,000
15,0	0,200	0,333	0,067	0,200	0,133	0,000	0,067	0,0	0,000
25,0	0,160	0,240	0,120	0,240	0,120	0,080	0,040	0,0	0,000
35,0	0,171	0,200	0,114	0,229	0,143	0,086	0,057	0,0	0,000
45,0	0,200	0,178	0,178	0,222	0,111	0,067	0,044	0,0	0,000
55,0	0,200	0,182	0,182	0,182	0,127	0,091	0,036	0,0	0,000
65,0	0,215	0,169	0,215	0,154	0,123	0,077	0,031	0,0	0,015
75,0	0,213	0,147	0,253	0,160	0,107	0,067	0,040	0,0	0,013
85,0	0,235	0,141	0,235	0,165	0,094	0,071	0,047	0,0	0,012
95,0	0,263	0,137	0,232	0,147	0,095	0,074	0,042	0,0	0,011

Le présent *splitter* est conservé et appliqué à l'ensemble des portefeuilles pour lesquels mesurer la mutualisation est envisagé à la suite.

## 5.2.2 Description et encodage du portefeuille à partir du *splitter*

Afin d'apprécier le niveau de mutualisation d'un portefeuille, des mesures de dispersion sont réalisées sur ce dernier en utilisant le *splitter* précédent. Plus précisément, deux séquences de mesures précisant de plus en plus la distribution du portefeuille et la concentration des OL sont réalisées à partir du *splitter*.

### 5.2.2.1 Les mesures de distribution du portefeuille

Une mesure élémentaire est calculée comme la fraction des assurés du portefeuille qui vérifie un critère du *splitter*. Typiquement, la première mesure est la proportion des véhicules du portefeuille dont la marque est contenue dans la liste du premier critère. Sur 115 359 véhicules du portefeuille, 25 092 ont ces modèles, soit 0,2175. La deuxième mesure est donnée par la fraction des assurés qui vérifie le deuxième critère, mais pas le premier, soit 0,2223. De façon plus générique, la  $d$ -ième mesure est le rapport entre le nombre d'assurés vérifiant le  $d$ -ième critère du *splitter* au nombre d'assurés en portefeuille. Réaliser des mesures de dispersion par la fréquence revient à dérouler le *splitter* et à compter progressivement les assurés qui tombent dans les nouveaux déroulés. Par symétrie, il n'est pas nécessaire de compter les assurés qui ne vérifient pas un critère donné. Une description de la distribution du portefeuille est donnée par la concaténation des mesures réalisées successivement. Le tableau 5.5 donne les premières descriptions du portefeuille avec une précision de 2 chiffres après la virgule. Les mesures précédentes sont effectuées au regard du nombre d'assurés en portefeuille. Elles sont dites de ce fait absolues, contrairement aux mesures relatives qui sont déterminées par la fraction des assurés du nœud qui vérifie son critère de subdivision nœuds fils. Le passage de l'une des mesures à l'autre se fait assez rapidement par de simples multiplications. Un portefeuille qui aurait une distribution uniforme suivant tous les facteurs de risque aurait sa description relative donnée par « 0,5|0,5|0,5|0,5|... » et sa description absolue serait « 0.5|0.25|0.25|0.125|0.0625|... ».

Inversement, un portefeuille dans lequel les assurés seraient recroquevillés dans des régions spécifiques aurait une description relative constituée essentiellement des fractions proches de 0 ou de 1 (avec la convention qu'un nœud n'ayant pas d'assurés, ne produira en dessous de lui que des mesures de 0). Pour le portefeuille des contrats couverts en 2020, la figure 5.5 donne l'évolution des mesures absolues de dispersion obtenues à partir du *splitter*. Les valeurs décroissent assez rapidement et au bout de 50 mesures, les valeurs observées sont très faibles.

TABLEAU 5.5 – Un exemple court pour illustrer la description et l'encodage (précision = 2)

#Mesure	Description Dec	Description bits
1	0.22	00010110
2	0.22 0.22	00010110 00010110
3	0.22 0.22 0.04	00010110 00010110 00000100
4	0.22 0.22 0.04 0.01	00010110 00010110 00000100 00000001
5	0.22 0.22 0.04 0.01 0.01	00010110 00010110 00000100 00000001 00000001
6	0.22 0.22 0.04 0.01 0.01 0.2	00010110 00010110 00000100 00000001 00000001 00010100
7	0.22 0.22 0.04 0.01 0.01 0.2 0.02	00010110 00010110 00000100 00000001 00000001 00010100 00000010
8	0.22 0.22 0.04 0.01 0.01 0.2 0.02 0.02	00010110 00010110 00000100 00000001 00000001 00010100 00000010 00000010
9	0.22 0.22 0.04 0.01 0.01 0.2 0.02 0.02 0.08	00010110 00010110 00000100 00000001 00000001 00010100 00000010 00000010 00001000

### 5.2.2.2 Les mesures de dispersion des OL en portefeuille

La description des assurés du portefeuille par des fréquences de dispersion indique à quel point les assurés recouvrent l'ensemble des profils sur le marché. Toutefois, ces descriptions ne rendent pas compte directement de la stratégie mise en place par l'assureur pour rassembler ce portefeuille. Des mesures sur l'OL dans les différentes régions permettraient de capter plus directement la mutualisation par les pratiques de l'assureur. Résumer la dispersion des OL dans une région de l'espace des assurés est moins immédiat que d'appréhender la distribution du portefeuille par des fréquences dans les sous-régions. La mesure proposée s'inspire directement de la formule du gain de stabilité de construction d'un arbre de régression (voir section 2.2.1.2) ; il s'agit en fait d'une forme normalisée du gain de stabilité d'OL associé à chaque nouvelle partition. Elle est obtenue en rapportant la diminution de variance d'OL à la variabilité totale de l'OL en portefeuille. Plus formellement, si le nœud  $\mathcal{N}_d$  est partagé par le  $d$ -ième critère en nœud  $\mathcal{N}_{d,<}$  et  $\mathcal{N}_{d,>}$ , alors la  $d$ -ième mesure de dispersion des OL est donnée par :

$$\alpha_d^{OL} = \frac{\text{Var}_{\mathcal{N}_d}(\mathbf{OL}) - \left( \frac{|\mathcal{N}_{d,<}|}{|\mathcal{N}_d|} \text{Var}_{\mathcal{N}_{d,<}}(\mathbf{OL}) + \frac{|\mathcal{N}_{d,>}|}{|\mathcal{N}_d|} \text{Var}_{\mathcal{N}_{d,>}}(\mathbf{OL}) \right)}{\text{Var}_{\mathcal{N}_0}(\mathbf{OL})}$$

où  $\text{Var}_{\mathcal{N}_{d,..}}(\mathbf{OL})$  est la variabilité des OL des assurés qui tombent dans le nœud  $\mathcal{N}_{d,..}$  et  $\text{Var}_{\mathcal{N}_0}(\mathbf{OL})$  est la variabilité des OL en portefeuille.

Une mesure relative est également introduite pour la variabilité d'OL, en divisant le gain d'OL associé à la partition du nœud par la variance du nœud  $\text{Var}_{\mathcal{N}_d}(\mathbf{OL})$  (et non par la variance totale d'OL du portefeuille  $\text{Var}_{\mathcal{N}_0}(\mathbf{OL})$ ).

Si un portefeuille a des OL constantes quelles que soient les caractéristiques des assurés (ce

qui implique en particulier que la prime technique est égale au profil de risque individualisé pour chaque assuré), les mesures sont fixées par convention à 0. Donc pour un portefeuille démutualisé à l'extrême, la description des OL est sous la forme : « 0,00|0,00|0,00|0,00|0,00|... ». Inversement, si la prime commerciale est constante et les profils de risque nombreux, alors les OL en portefeuille sont très variables et leur variabilité correspond à celle des profils de risque. Les mesures de variabilité des OL sont alors variables d'une mesure à l'autre et évoluent entre 0 et 1. Plus la mesure est grande, plus l'OL est disparate entre les risques et donc plus importante est la mutualisation. Pour le portefeuille des contrats couverts en 2020, la figure 5.5 donne l'évolution des mesures absolues de dispersion d'OL obtenues à partir du *splitter*.

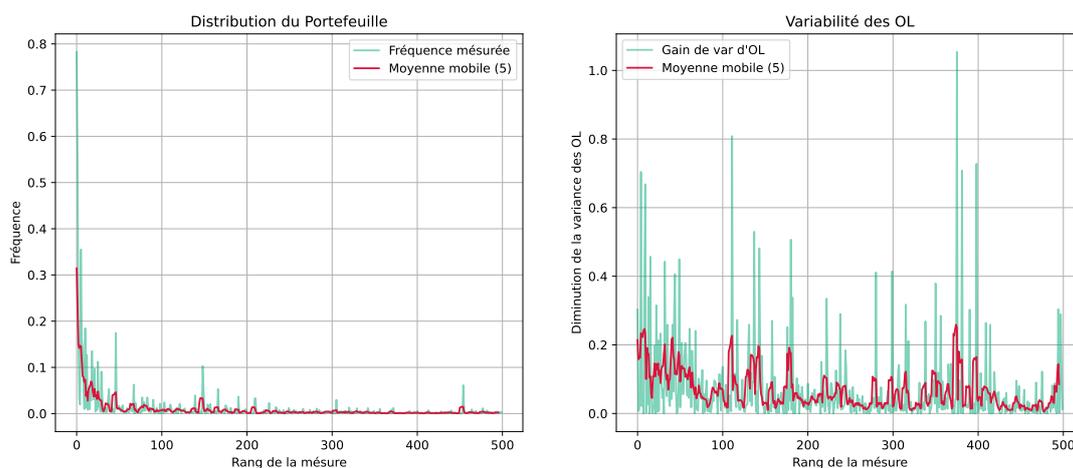


FIGURE 5.5 – Évolution des mesures (absolues) de fréquence et de dispersion d'OL réalisées sur le portefeuille

### 5.2.2.3 Encodage des descriptions sous forme de chaîne binaire

Les descriptions du portefeuille retranscrivent l'hétérogénéité des assurés en portefeuille. La quantification de cette hétérogénéité à partir d'une description du portefeuille se fait à partir de la complexité de Kolmogorov qui est une généralisation (dans un certain sens) de l'entropie qui s'est avérée prometteuse dans les indicateurs simples, mais qui permet ici de travailler en amont, sur la description du portefeuille, avant tout modèle de risque. La formalisation et l'approximation de la complexité de Kolmogorov supposent une représentation des descriptions d'objets sous forme de chaînes binaires. Dans ce sillage, il est nécessaire de convertir les descriptions précédentes sous forme de chaîne de binaire également. Une description (de fréquence ou d'OL) est en effet une séquence de réels compris entre 0 et 1 :  $desc_d = \alpha_1\alpha_2\alpha_3 \dots \alpha_d$  avec  $\alpha_i \in [0, 1]$ . La réécriture binaire de cette description consiste à coder chaque mesure  $\alpha_i$  par une représentation binaire. À cet effet, trois approches sont implémentées (voir les programmes 5.1), elles conduisent à des représentations binaires différentes des descriptions et *in fine* à des CK arbitrairement différents lorsque le calcul est basé sur l'une de ces approches :

- La première approche d'encodage est basée sur la conversion des nombres réels (en base 10) en nombre binaire (en base 2). La norme IEEE 754 permet de représenter un nombre

réel sous la forme « *signe (s) - exposant (e) - mantisse (m)* » :  $\alpha_i = s_i \times m_i \times 2^{e_i}$ . Chacune des composantes  $(s_i, m_i, e_i)$  est représentée sur un nombre de bits donné. Pour une représentation binaire sur 16 bits dite *demi précision*, le signe est codé sur 1 bit, la mantisse sur 10 bits et l'exposant sur 5 bits<sup>1</sup> ;

- Le codage systématique des mesures sur 16 bits précédent rend les descriptions du portefeuille arbitrairement longues. Pour y remédier, la représentation binaire de  $\alpha_i$  est obtenue comme la concaténation de la partie entière de  $\alpha_i$  et de la conversion en base 2 de la partie décimale de  $\alpha_i$  mais considérée comme un entier naturel. À titre d'exemple, pour  $\alpha_i = 0,36$ , la représentation binaire sera 0100100 obtenue en concaténant 0 (  $[0, 36] = 0$  ) et 100100 ( $(36)_{10} = (100100)_2$ ). Cette approche permet de limiter la taille de la description. En effet, à partir de la précision sur les mesures, entendue simplement comme le nombre de chiffres retenus après la virgule, le nombre de bits de la représentation binaire de la partie décimale d'une mesure est déterminé par la formule :  $\lceil \log_2(10^p - 1) \rceil$ , où  $10^p - 1$  est l'entier le plus grand de  $p$  chiffres. Pour une précision  $p = 2$ , les représentations binaires des mesures sont codées sur 8 bits (1 bit pour la partie entière et 7 bits pour l'entier de la partie décimale). Il peut être nécessaire de compléter la représentation de l'entier à droite de la virgule par des zéros pour remplir les bits restants. La représentation binaire de  $\alpha_i = 0,36$  avec une précision  $p = 2$ , est alors 00100100 avec cette approche ;
- Les propositions d'encodage précédentes n'accordent pas d'importance directe au sens de la mesure à représenter sous forme binaire. En effet,  $\alpha_i$  est la proportion des individus du nœud qui vérifient le critère de partition. Si  $\alpha_i$  est proche de 0,5 alors la dispersion des assurés dans le nœud est maximale. Inversement, si  $\alpha_i$  est proche de 0 ou de 1, alors la mesure indique que les assurés ne recouvrent pas uniformément les deux sous-espaces spécifiés par le critère du nœud. Dans le premier cas  $K(x_i)$  doit être grand et faible dans le second cas. Dans ce sillage  $K(x_i)$  fait penser à la complexité d'une loi binomiale dont les paramètres seraient : le nombre d'assurés dans le nœud ( $|\mathcal{N}_{d,i}|$ ) et la proportion des assurés qui vérifient le critère du nœud ( $\alpha_i$ ). Une représentation binaire de la mesure  $\alpha_i$  est donnée par une séquence de  $20 + \lceil \log_2(10^p - 1) \rceil$  bits dans laquelle la fraction des  $\alpha_i$  derniers bits est codée par des 1 et la fraction des  $1 - \alpha_i$  premiers bits est codée par des 0. Le nombre 20 est arbitrairement choisi. Plus la taille de la représentation est faible, plus grande est la perte d'information. En revanche, plus elle est grande plus il devient difficile de construire une description binaire du portefeuille.

Code Listing 5.1 – Proposition de méthodes d'encodage des mesures en bits

```
def frac2bit_half_precision(frac):
    return bin(np.float16(frac).view('H'))[2:].zfill(16)

def frac2bits_DecAsInt(frac, precision = 4, decimal_char = ',.')
```

1. voir <https://trekhele.dev/blog/2021/binary-floating-point/> pour une description plus détaillée

```

bit_rep = frac[0]
dec_part = frac.split(decimal_char, 1)[1][:precision]
dec_part = dec_part + '0'*(precision - len(dec_part))
dec_part_as_bit = "{0:b}".format(int(dec_part))
bitstring = '0' * (precision2bitsize(precision) - len(dec_part_as_bit))
                + dec_part_as_bit

return bit_rep + bitstring

def frac2bits_Simple(frac, precision = 4, decimal_char = '.'):
    size = 20 + precision2bitsize(precision)
    nb_ones = round(float(frac)* size)
    bit_rep = '0'* (size- nb_ones) + '1'*nb_ones
    return bit_rep

```

### 5.2.3 Approximation de la complexité de Kolmogorov pour quantifier la mutualisation connue

Quelle que soit l'approche d'encodage retenue, celle-ci revêt une dimension quelque peu arbitraire; soit par le choix de la longueur de la représentation binaire, soit par son principe même. C'est pourquoi une approche plus directe pour estimer  $K(x_d)$  est donnée en 5.2.3.3. Mais avant, une discussion rapide sur l'approximation de  $K(x_d)$  par les méthodes tirées de l'état de l'art est faite.

#### 5.2.3.1 BDM - Une forte dépendance à l'encodage et à la paramétrisation

La BDM présentée à la section 2.3.3 est une méthode d'approximation de la complexité de Kolmogorov qui s'appuie sur une décomposition d'une chaîne binaire en sous-chaînes sur lesquelles il est possible d'appliquer la CTM (section 2.3.3) pour en approximer la complexité. Elle nécessite donc en entrée que la description de la dispersion du portefeuille (ou la volatilité des OL) soit représentée sous forme de chaîne binaire. La spécification de cette méthode impose de préciser la taille des sous-chaînes ainsi que la gestion des conditions de bord (qui spécifie indirectement la façon de décomposer la description). La figure 20 en annexe donne la CK de la description du portefeuille (par la dispersion des assurés et le gain de volatilité des OL) en fonction de la taille des sous-chaînes et de la stratégie de gestion des conditions de bord. Sans perte de généralité; la deuxième stratégie d'encodage *DecAsInt* est utilisée pour représenter les mesures sous forme binaire. Les valeurs de la complexité obtenues pour une description donnée sont croissantes avec la longueur des sous-chaînes et plus importantes pour la condition de bord *PartitionCorrelated*. La tendance haussière de la CK en fonction de la description traduit bien une connaissance de plus en plus précise de la mutualisation du portefeuille. Toutefois, cette dynamique est fortement influencée par la paramétrisation du modèle de calcul PyBDM implémenté par les auteurs de la méthode.

Une longueur de bloc faible donne lieu à un grand nombre de sous-chaînes identiques ; ce qui est moins le cas lorsque la taille du bloc est plus élevée<sup>2</sup>. Le terme qui compte le nombre d'occurrences des sous-chaînes identiques étant en  $\log_2$ , il croit beaucoup moins vite que la complexité des sous-chaînes unitaires lorsque la taille des sous-chaînes augmente. Donc la complexité totale de la description dépend positivement de la taille de la décomposition. De plus, normaliser la CK calculée par la BDM (Zenil *et al.*, 2018) dans l'optique de s'affranchir de cette dépendance à la taille du bloc de décomposition n'est pas concluant. Puisque la dynamique de la complexité normalisée n'est pas en ligne avec la complexité initiale qui devrait être croissante avec l'enrichissement des mesures sur le portefeuille comme attendu (figure 21 en annexe).

Ainsi, l'approximation de la CK par la BDM dépend fortement de son hyper-paramétrisation, mais également du choix d'encodage du portefeuille sous forme binaire qui reste très arbitraire. Dès lors, la comparabilité des niveaux de mutualisation évalués en utilisant la BDM pour quantifier la mutualisation du portefeuille connue à travers la description ne peut se faire en l'absence des méta données sur les paramètres utilisés pour le calcul.

### **5.2.3.2 Compression sans perte : une forte dépendance au choix d'encodage**

---

L'emploi des algorithmes de compression sur des chaînes binaires est récurrent dans la littérature pour évaluer la complexité des objets. Dans ce sillage, après avoir encodé la description sous forme binaire en utilisant les trois approches précédentes, l'implémentation Python bz2<sup>3</sup> de l'algorithme de compression Bzip2 a été utilisé pour approcher la CK. La figure 22a donne l'évolution de la CK estimée en fonction de la richesse de la description pour différentes stratégies d'encodage. Il en ressort que la CK estimée est fortement dépendante du choix de l'encodage. La croissance est lente pour l'encodage simple alors qu'elle est pratiquement linéaire avec la taille de la description pour l'encodage demi-précision. En effet, l'encodage à demi-précision ne prend pas en compte la nature de la mesure et par suite ne retranscrit pas l'idée d'une information sur la distribution du portefeuille croissante avec les nouvelles mesures, mais à taux décroissant. La tentative de normalisation donnée par la figure 22b n'est pas en ligne non plus avec la dynamique attendue.

### **5.2.3.3 Approximation de CK par l'entropie**

---

Les tentatives d'évaluation de la CK par la BDM et les algorithmes de compression se sont révélées peu concluantes. Du fait de la relation entre l'entropie et la CK (partie 2.3.2), l'entropie est souvent utilisée pour mesurer la CK. Nous proposons ici une approche de calcul fondée sur l'entropie. Rappelons que  $x_d$  désigne la description du portefeuille obtenue après  $d$  mesures ;  $x_d = \alpha_1|\alpha_2| \dots |\alpha_i| \dots |\alpha_d|$ , il s'agit spécifiquement d'une liste de  $d$  fractions comprises entre 0 et 1. Commençons par proposer une stratégie d'encodage qui assurerait la conversion d'une

---

2. pour s'en convaincre, il suffit d'observer que la probabilité d'obtenir une séquence de 4 bits identiques 1111 est plus élevée que celle d'obtenir 5 bits identiques.

3. <https://docs.python.org/3/library/bz2.html>

mesure  $\alpha_i$  en bits sans perte d'information. Pour un nœud  $\mathcal{N}_d$  contenant  $|\mathcal{N}_d|$  assurés dont la proportion  $\alpha_d$  vérifie le critère de partition du nœud, un encodage sans perte d'information de  $\alpha_d$  est donné par :  ${}^b\alpha_d = y_1 y_2 \dots y_{|\mathcal{N}_d|}$  où chaque  $y_i$  est un bit égal à 1 si l'assuré  $i$  du nœud vérifie le critère et à 0 sinon. Il y a alors,  $\alpha_d \times |\mathcal{N}_d|$  bits égaux à 1 dans cette chaîne binaire. Il convient de préciser qu'en pratique construire cette représentation poserait deux contraintes majeures : l'ordonnancement des assurés du nœud et la longueur de la description totale du portefeuille. Toutefois, l'approximation de la CK par l'entropie que nous proposons permet d'évaluer la complexité de cette description sans avoir à la représenter explicitement.

Pour évaluer l'information contenue dans la description  $x_d$ , nous proposons une approche incrémentale de l'information :  $K(x_d) \approx \sum_{i=1}^d K({}^b\alpha_i)$ . La relation entre la CK et l'entropie donne un moyen d'évaluer l'information associée à la mesure  $\alpha_d$ . En effet, la partie 2.3.2 illustre que pour une chaîne  $x = y_1 y_2 \dots y_{n/r}$  de taille  $n$  qui se décompose en  $n/r$  sous chaînes de taille  $r$ , (à condition que  $m \rightarrow \infty$  avec  $r$  fixé) sa complexité peut être approchée par :  $K(x) = mH_r + O(1)$  où  $H_r$  est l'entropie empirique associée à la distribution des chaînes uniques parmi les  $m$  chaînes  $y_i$ . L'application de ce théorème au calcul de la complexité de la  $d$ -ième mesure  ${}^b\alpha_d$ , pour  $r = 1$  et  $m = |\mathcal{N}_d|$  est donnée par :

$$K({}^b\alpha_d) \approx |\mathcal{N}_d| H_1 = -|\mathcal{N}_d| ((1 - \alpha_d) \log_2(1 - \alpha_d) + \alpha_d \log_2(\alpha_d))$$

Lorsque la mesure  ${}^b\alpha_d$  est partitionnée en sous-chaînes d'un bit, les "1" apparaissent avec la probabilité  $\alpha_d$ , conduisant alors à l'entropie empirique de  $-(1 - \alpha_d) \log_2(1 - \alpha_d) - \alpha_d \log_2(\alpha_d)$  pour chacun des  $|\mathcal{N}_d|$  bits. De façon équivalente, il est possible de considérer des sous-chaînes de  $r = 2$  bits, la décomposition de  ${}^b\alpha_d$  donnerait  $\frac{|\mathcal{N}_d|}{2}$  sous-chaînes dont les quatre chaînes uniques suivantes seraient observées avec ces probabilités :

$$\begin{array}{ll|ll} 00 & \text{avec probabilité} & (1 - \alpha_d)^2 & 01 & \text{avec probabilité} & \alpha_d(1 - \alpha_d) \\ 10 & \text{avec probabilité} & \alpha_d(1 - \alpha_d) & 11 & \text{avec probabilité} & \alpha_d^2 \end{array}$$

L'entropie empirique  $H_2$  serait :

$$\begin{aligned} H_2 &= -2(1 - \alpha_d)^2 \log_2(1 - \alpha_d) \\ &\quad - 2\alpha_d(1 - \alpha_d) \log_2(\alpha_d(1 - \alpha_d)) \\ &\quad - 2\alpha_d^2 \log_2(\alpha_d) \\ &= -\log_2(1 - \alpha_d) (2(1 - \alpha_d)^2 + 2\alpha_d(1 - \alpha_d)) \\ &\quad - \log_2(\alpha_d) (2\alpha_d^2 + 2\alpha_d(1 - \alpha_d)) \\ &= -2((1 - \alpha_d) \log_2(1 - \alpha_d) + \alpha_d \log_2(\alpha_d)) \\ &= 2H_1 \end{aligned}$$

Par suite, la complexité de Kolmogorov de la mesure  ${}^b\alpha_d$  approchée par une décomposition en séquences de 2 bits conduirait au même résultat. De façon plus générale, un développement

combinatoire conduit à  $H_r = rH_1$  (annexe E), garantissant ainsi la robustesse de l'approximation de CK par l'entropie des sous-chaînes vis-à-vis de la taille du partitionnement. Cette robustesse résulte de la nature extensive de l'entropie qui entraîne en particulier que l'entropie d'un système est égale à la somme des entropies des sous-systèmes issus d'une décomposition quelconque.

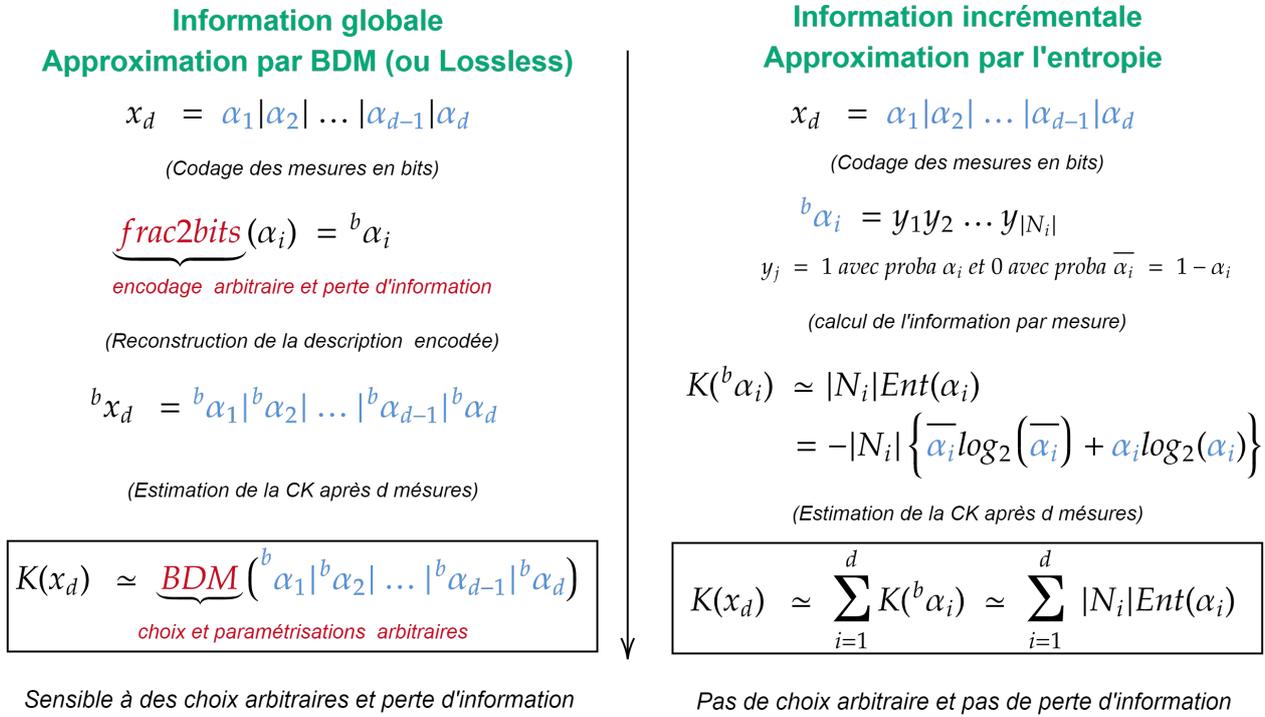


FIGURE 5.6 – Adaptation de la démarche de Zureck pour approcher CK par l'entropie

Au final, la complexité de la description  $x_d$  est donnée par :

$$K(x_d) = - \sum_{i=1}^d |N_i| \left( (1 - \alpha_i) \log_2(1 - \alpha_i) + \alpha_i \log_2(\alpha_i) \right)$$

où  $|N_i|$  est le nombre d'assurés qui tombent dans le nœud associé à la  $i$ -ième mesure et  $\alpha_i$  la fraction de ceux qui vérifient le critère de partition du nœud.

### 5.2.4 Évaluation de la mutualisation résiduelle pour une description partielle par l'entropie de Shannon des micro-états

Une description du portefeuille correspond à une spécification partielle de la distribution des assurés dans l'espace des facteurs de risque partitionné en sous-espaces. Chaque nouvelle mesure précise d'avantage la répartition des assurés dans cet espace en considérant des partitions plus fines de l'espace des caractéristiques. Lorsque le *splitter* est déroulé jusqu'aux feuilles terminales, la description du portefeuille est complète et le portefeuille est entièrement déterminé. En pratique, l'information n'est disponible que pour les premières variables (lors d'un

appel d'offre avec un partenaire qui partage certaines informations sur son portefeuille mais pas intégralement le portefeuille), permettant alors de construire une description incomplète. Pour  $d$  mesures réalisées sur le portefeuille, le partitionnement correspondant et la dispersion des assurés associée apportent une information partielle sur le recouvrement de l'espace par les assurés. Les partitions de l'espace correspondent à des nœuds intermédiaires du *splitter*. En dessous des nœuds révélés, l'incertitude sur la répartition des assurés (et donc sur la mutualisation suivant les autres critères) subsiste. En effet, un assuré dans un nœud peut se trouver dans n'importe laquelle des feuilles terminales sous celui-ci. Plus il y a des assurés dans un nœud et plus le nombre de feuilles terminales sous ce dernier est grand, plus l'ignorance sur leur mutualisation est importante.

À titre d'exemple, sur le portefeuille des polices de 2020, une première mesure indique que 22 % des véhicules ont l'un des modèles qui définissent la division. Ce chiffre représente la description connue après une mesure, mais aucune information n'est donnée sur la distribution en âge ou en puissance sur ces 22 % ni les 78 % restants. Après cette première description, Si tous les véhicules dans chacun des nœuds fils sont identiques, alors les profils en portefeuille ne diffèrent que par ce premier critère, pas de mutualisation ultérieure. À l'inverse, la dispersion résiduelle peut être assez importante suivant les critères non encore observés. Cette incertitude sur la dispersion des assurés au-delà de l'information connue (il s'agit des micro-états compatibles avec le macro-état précisé) est qualifiée de mutualisation résiduelle.

Dans l'article fondateur de Zurek, l'évaluation de l'ampleur de cette hétérogénéité résiduelle correspondant au micro-état sous-jacent se fait en admettant que la micro-configuration devrait être la plus probable : celle ayant l'entropie de Shannon la plus élevée. Le parallèle de cette définition originelle à la mesure de la mutualisation en assurance revient à supposer une dispersion des assurés dans les feuilles terminales sous leurs nœuds respectifs avec des probabilités de marchés.

La figure 5.7 indique comment évolue la mutualisation résiduelle lorsque l'information sur le portefeuille des contrats couverts en 2020 est de plus en plus exhaustive. La baisse de l'incertitude sur la mutualisation est très rapide pour les premières mesures, c'est-à-dire les plus importantes sur le risque. La connaissance des probabilités de marché, c'est-à-dire des proportions des assurés dans les feuilles terminales du *splitter* conduit à une mutualisation résiduelle moins importante que celle obtenue sous l'hypothèse d'une distribution uniforme. Néanmoins, l'écart entre les deux courbes reste très contenu, l'utilisation des probabilités uniformes dans les feuilles terminales est donc un bon compromis pour évaluer la mutualisation résiduelle en absence des probabilités de marché.

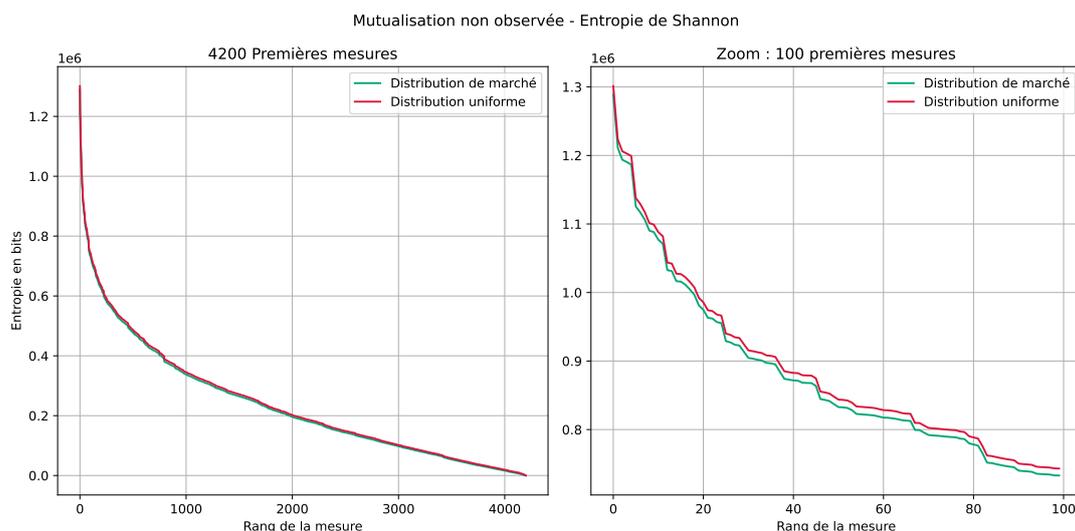


FIGURE 5.7 – Évolution de la mutualisation résiduelle pour une description partielle du portefeuille

### Remarque : Non prise en compte de l'hétérogénéité résiduelle des OL

Lorsqu'il s'agit d'évaluer la mutualisation par la dispersion des OL, la mesure de la mutualisation résiduelle par l'entropie des micro-états correspondants aux distributions des assurés dans les feuilles terminales n'est pas suffisante. Il faudrait également considérer l'incertitude pendante sur la volatilité des OL. Néanmoins, cette composante et la complexité qui en résulterait ne sont pas prises en compte dans ce mémoire ; l'estimation de la mutualisation non observée est alors sous estimée dans le cas d'une description par les OL.

## 5.2.5 Mesure de la mutualisation par le ratio de mutualisation généralisé

Le ratio de mutualisation généralisé est défini comme le rapport entre l'entropie algorithmique d'une macro-description du portefeuille et l'hétérogénéité maximale atteignable à partir du nombre d'assurés en portefeuille. Plus schématiquement, l'entropie algorithmique quantifie à quel point les assurés du portefeuille dont la description est partiellement connue, recouvrent l'espace des facteurs de risque (par rapport à une configuration de marché).

### 5.2.5.1 Application au portefeuille de BNP Paribas Cardif

L'application du ratio de mutualisation généralisé pour mesurer le niveau de mutualisation du portefeuille de contrats d'extension de garantie couverts en 2020 à partir de la distribution des assurés et de la dispersion des OL conduit à la figure 5.8. Le niveau de mutualisation évalué à partir de l'hétérogénéité des assurés en portefeuille est voisin de 100 % comme anticipé. En effet, le *splitter* de marché est construit sur le portefeuille de Cardif, ce qui signifie en particulier que ce portefeuille est celui qui recouvre le plus l'espace des facteurs de risque encodé par le

*splitter* .

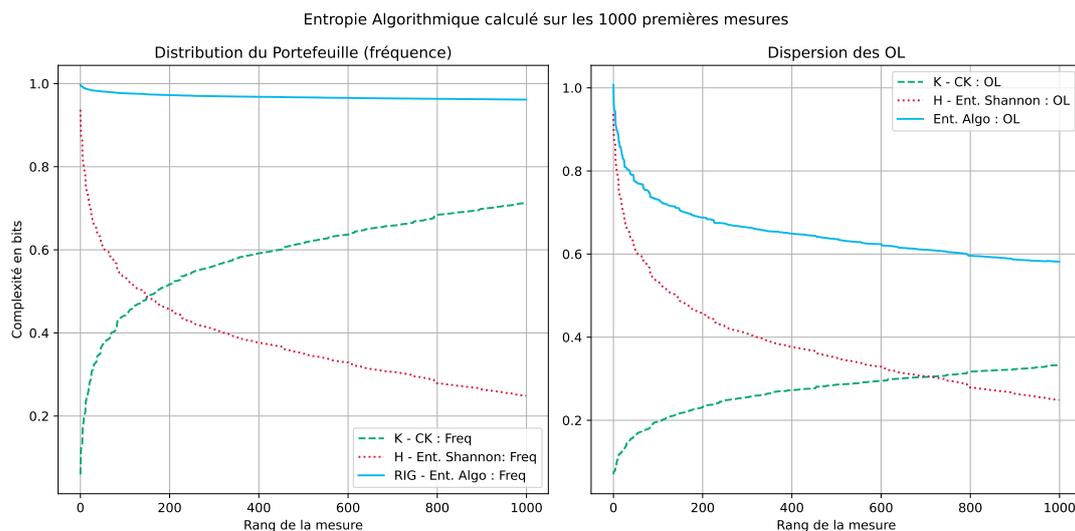


FIGURE 5.8 – Mesure de la démutualisation à partir du ratio de mutualisation généralisé (entropie algorithmique) sur le portefeuille de BNP Paribas Cardif.

À partir des descriptions de baisse de volatilité d'OL, le ratio de mutualisation généralisé mesure à quel point les OL sont différentes d'une région à l'autre (correspondant à des financements croisés entre assurés). Plus cet indice est grand, plus important est le financement croisé du portefeuille. Contrairement au cas de la description du portefeuille par des mesures d'effectifs dans les sous-régions, l'indicateur construit à partir des mesures sur l'OL n'est pas constant ; il baisse assez rapidement sur les premières mesures. En effet, la complexité de la description des OL, reste faible comparativement à la mesure de la mutualisation résiduelle. Cela traduit en particulier que même si les assurés sont répartis dans tout l'espace, les OL par assuré sont différentes (dans le cas contraire, la CK serait bien plus faible voir nulle).

Ainsi, calculé à partir des descriptions des fréquences, le ratio de mutualisation généralisé indique le pourcentage de recouvrement de la structure du marché par le portefeuille décrit. Cette mesure de la mutualisation sous-entend que plus les assurés en portefeuille ont des profils différents plus la mutualisation est importante. Calculé à partir des descriptions de volatilité d'OL, cet indicateur mesure la mutualisation induite par les choix actuariels (tarifaires et commerciaux). Ces deux formes du ratio de mutualisation généralisé sont donc complémentaires.

### 5.2.5.2 Comparaison des stratégies de tarification en situation de monopole

À partir des stratégies tarifaires correspondantes aux variables du tableau 4.7, quatre suites de descriptions de la Volatilité des OL ont été construites. Le *splitter* commun utilisé pour l'élaboration de ces descriptions est celui présenté précédemment. La figure 5.9 suivante présente l'évolution du niveau de mutualisation correspondant à chacune de ces stratégies.

L'assureur 4 dont la stratégie emploie la marque de véhicule dans la grille tarifaire a le niveau de mutualisation le plus faible, c'est-à-dire qu'il crée moins de solidarité dans le portefeuille

par les pratiques tarifaires. À partir d'une description de plus de 200 mesures, son ratio de mutualisation généralisé est de 70% soit près de 10 points de pourcentage inférieur à celui du deuxième assureur le plus démutualisé qui n'est autre que l'assureur 2. Au regard des premières mesures qui sont également les plus impactantes, les niveaux de mutualisation des assureurs 1 et 3 sont très comparables et sont les plus élevés des quatre stratégies. Pour une description de 200 mesures, le RMG vaut pratiquement 85 %. Au-delà de ce nombre de mesures, le RMG de l'assureur 3 baisse plus rapidement que celui de l'assureur 1. En effet, bien que le kilométrage à la souscription ne joue qu'un rôle marginal sur la segmentation du risque, l'assureur 3 en fait usage dans sa grille tarifaire contrairement à l'assureur 1. Il est donc naturel qu'avec beaucoup d'informations sur le portefeuille, l'écart entre les mutualisations qui découlent de ces deux stratégies se précise. Il convient de noter que ces stratégies étant appliquées au même portefeuille, les RMG - Fréquence qui seraient obtenus à partir des descriptions de fréquences coïncideraient exactement.

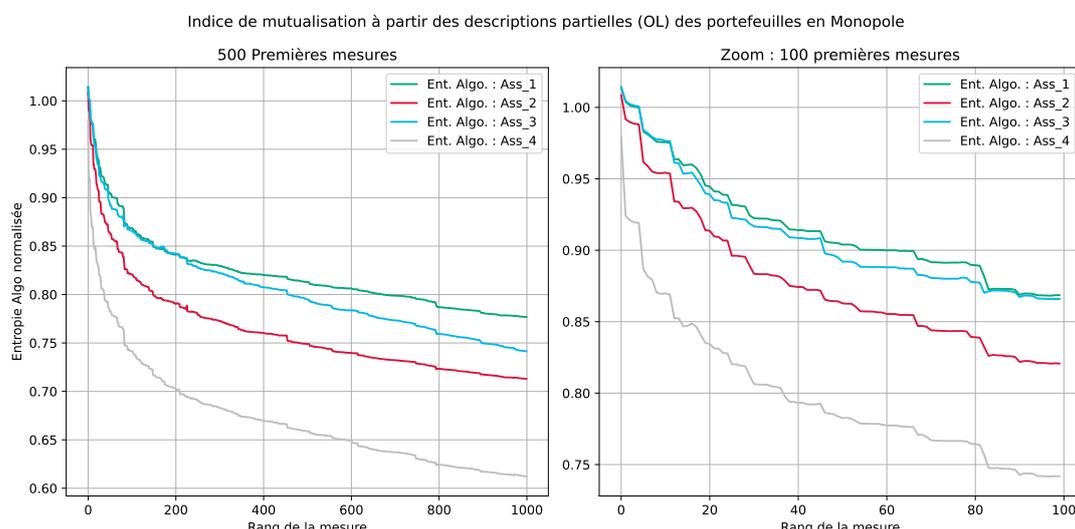


FIGURE 5.9 – Comparaison du niveau de mutualisation entre assureurs en monopole par le ratio de mutualisation généralisé

### 5.2.5.3 Comparaison des stratégies de tarification en situation de concurrence

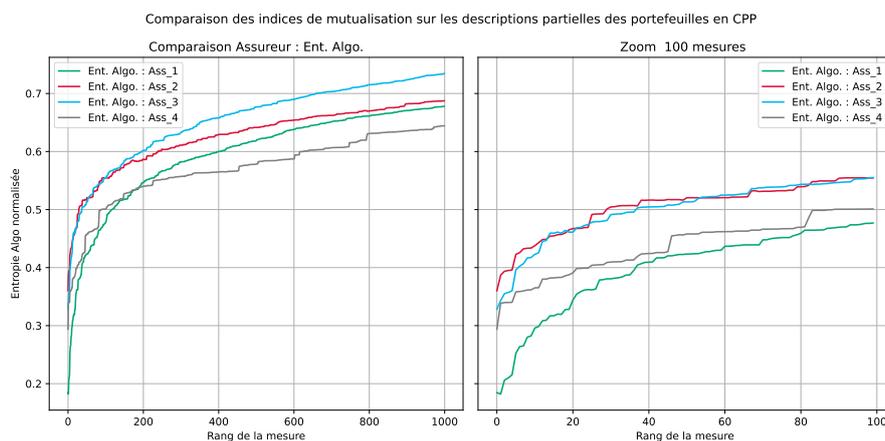
En situation concurrentielle, la structure du portefeuille de chaque assureur dépend fortement des prix qu'il pratique. Les portefeuilles des assurés recouvrent alors différentes régions de l'espace des facteurs de risque. Dès lors, le RMG calculé sur des descriptions de fréquences rend bien compte de la capacité de l'assureur à se positionner sur tous les segments de risques du marché. La déclinaison du RMG obtenue à partir des descriptions de volatilité d'OL traduit le niveau de solidarité créé entre les assurés du portefeuille.

### Comparaison du niveau de mutualisation en concurrence par le RMG - Fréquence

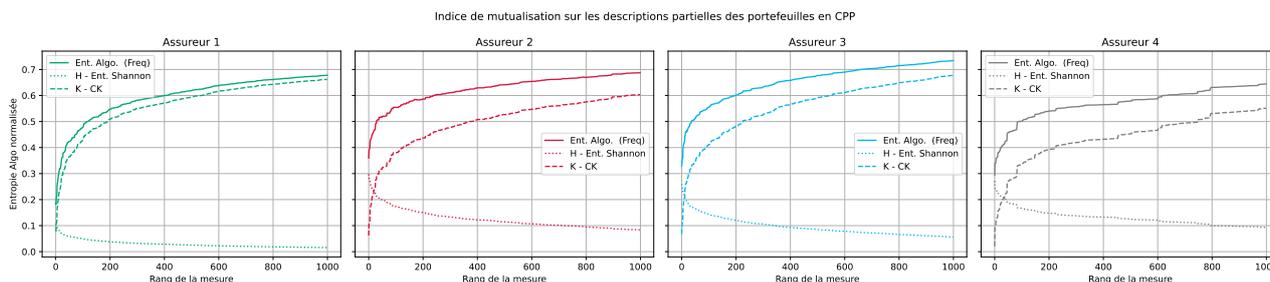
Les évolutions des RMG basés sur des descriptions de distribution des assurés sont représentées sur la figure 5.11. La figure 5.10a apporte une vision synthétique permettant de comparer

le niveau de mutualisation entre assureurs ; tandis que la figure 5.10b donne les détails sur le niveau de mutualisation de chaque assureur en le décomposant suivant la complexité de Kolmogorov (mutualisation connue) et l'entropie de Shannon (mutualisation résiduelle). Ces figures montrent que malgré la stratégie très inclusive de l'assureur 1, il ne se positionne que sur une faible portion des risques du marché, car il est victime de sélection adverse. Les portefeuilles des assureurs 2 et 3 recouvrent plus la structure du marché que ceux des assureurs 1 et 4. Les assureurs 2 et 3 sont donc ceux dont les stratégies permettent de mettre ensemble plus de risques hétérogènes. Plus précisément par assureur, il ressort que :

- **Assureur 1** : l'entropie de Shannon mesurée sur le portefeuille de cet assureur est très faible, dès la première mesure. Ceci traduit une part de marché relativement faible et un positionnement sur des segments de risques spécifiques. Il est, en effet, victime de sélection adverse et son portefeuille au final n'est constitué que de mauvais risques ;
- **Assureur 2 et Assureur 3** : l'entropie de Shannon, relativement élevée et sa décroissance lente traduisent des positionnements de ces assureurs sur plusieurs segments de risque. De plus, les courbes de complexité sont assez élevées et présentent des rythmes de croissance comparables (avant 200 mesures). Cette dynamique des CK traduit un nombre important d'assurés dans les régions dans lesquelles se positionnent ces assureurs. Pour des descriptions de plus de 200 mesures, l'accroissement de la complexité de la description pour l'assureur 2 est plus lente que celle de l'assureur 3 et les RMG - Fréquence de ces deux assureurs s'écartent à partir de cette description. La tranche de Kilométrage est de plus en plus utilisée et l'assureur 3 (contrairement à l'assureur 2) ayant segmenté son tarif suivant ce critère, dispose *in fine* des véhicules hétérogènes suivant le kilométrage à la souscription ;
- **Assureur 4** : comme les assureurs 2 et 3, l'assureur 4 est présent sur plusieurs segments de risque différents puisque l'entropie de Shannon des assurés est élevée et décroît très lentement. Par contre, la complexité de Kolmogorov calculée sur les descriptions est assez faible et croît très lentement. Elle indique ainsi que sur les différents risques où cet assureur est présent, il n'a souscrit qu'un nombre limité d'assurés. En proposant une segmentation très fine des primes, l'assureur 4 écrème chaque segment de risque du marché et n'y récupère que les meilleurs risques.



(a) Comparaison RMG - Fréquence entre assureurs



(b) Détails RMG - Fréquence par assureur

FIGURE 5.10 – Comparaison de la mutualisation entre assureurs en situation de concurrence à partir du RMG - Fréquence

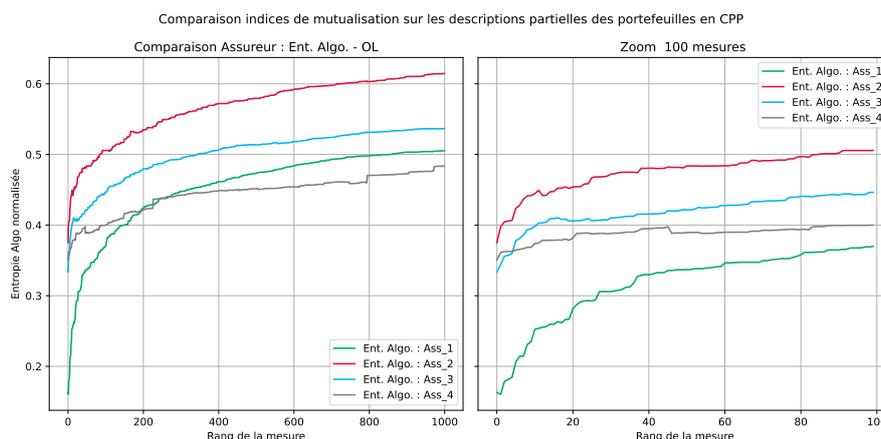
### Comparaison du niveau de mutualisation en concurrence par le RMG - OL

De même que pour les fréquences, les RMG calculés sur des descriptions de volatilité d'OL sont présentés en vision synthétique (figure 5.11a) et par assureur (figure 5.11b). En situation concurrentielle, l'assureur 2 est celui qui crée plus de solidarité dans son portefeuille suivi de l'assureur 3. Les interprétations suivantes par assureur viennent compléter les interprétations précédentes sur les positionnements sur des segments plus ou moins hétérogènes de risque :

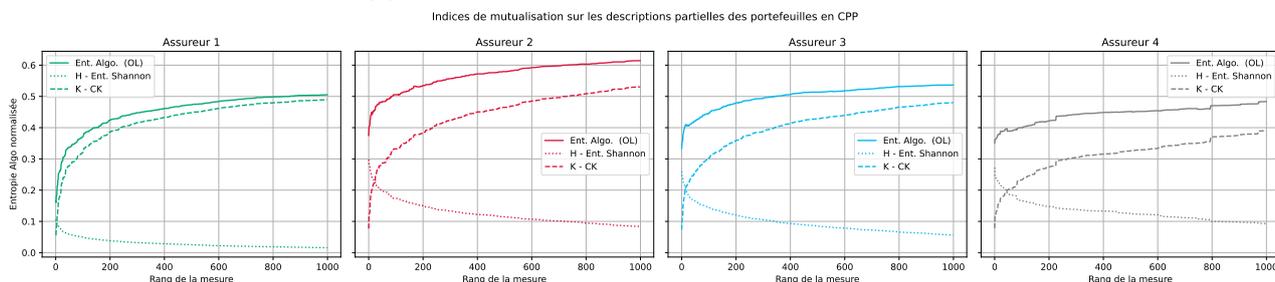
- **Assureur 1** : en se retrouvant sur des mauvais risques exclusivement et avec une faible part de marché, l'entropie de Shannon qui mesure la mutualisation comme l'hétérogénéité résiduelle des assurés en portefeuille est assez faible et au final le RMG - OL au même titre que le RMG - Fréquence est faible ;
- **Assureur 2** : le niveau de mutualisation de cet assureur est particulièrement élevé pour deux raisons : sa présence sur plusieurs segments de risque et la solidarité qu'il fait jouer en pratiquant des primes techniques qui laissent une marge de solidarité entre assurés. Cette seconde raison s'observe sur la figure 5.11b à partir de la courbe de CK de cet assureur. Elle est assez élevée, traduisant ainsi des variabilités d'OL fortes entre les profils de risque en portefeuille ;
- **Assureur 3** : l'observation pour l'assureur 3 est très semblable à celle de l'assureur 2 ; il présente le second RMG - OL le plus élevé. Ce niveau de mutualisation *in fine* plus

faible que celui de l'assureur 2 est lié conjointement à une CK et à une mutualisation résiduelle plus faibles chez l'assureur 3. Puisque les prix pratiqués par l'assureur 3 sont plus agrégés que ceux de l'assureur 2, cette faible mutualisation devrait être le fait de la similarité entre les profils souscrits par l'assureur 3 ;

- **Assureur 4** : la courbe de la CK est basse et croit très lentement, traduisant ainsi une désolidarisation des assurés du portefeuille. Les primes de cet assureur étant très segmentées, elles sont assez proches des profils de risque individualisés et par suite les OL sont faibles. L'absence de marge de cross-financement entraîne alors une faible variabilité des OL entre assurés, d'où cette dynamique de la CK observée.



(a) Comparaison RMG - OL entre assureurs



(b) Détails RMG - OL par assureur

FIGURE 5.11 – Comparaison de la mutualisation entre assureurs en situation de concurrence à partir du RMG - OL

## Conclusion du chapitre

Il était question dans ce chapitre d'implémenter les indicateurs exotiques sur des données réelles. Ces indicateurs sont proposés afin de s'affranchir des modèles de risque, d'évaluer la mutualisation avec une connaissance moindre des portefeuilles considérés, de tenir compte de la capacité à mettre ensemble des assurés différents et de mieux comprendre la mutualisation. À partir de l'inertie, la mesure de la démutualisation par la part de risque de structure transférée dans la tarification est généralisée. L'intérêt de cet indicateur est de fournir une compréhension des contributions variable par variable à la démutualisation notamment en représentant les

assurés dans un hyper-espace de contribution au risque. Le ratio de mutualisation généralisé quant à lui permet d'évaluer le niveau de mutualisation à partir d'une connaissance partielle du portefeuille. Selon que la description porte sur la distribution des assurés par les fréquences dans des sous-régions ou sur les écarts de volatilité d'OL, cet indicateur va mesurer à quel point la structure du portefeuille recouvre celle du marché ou à quel point l'assureur crée de la solidarité à partir de ses choix tarifaires.

Ces indicateurs permettent de mieux comprendre les facteurs de risque sur lesquels s'appuie l'assureur pour démutualiser son portefeuille. Par contre, les niveaux de mutualisation donnés par ces indices sont difficilement calculables comparativement aux indices simples; du fait notamment de l'impossibilité de calculer la complexité de Kolmogorov qui impose de concevoir des méthodes pour l'approximer. Si ces indicateurs sont difficilement explicables à un public non expert comme l'assuré ou le partenaire, ils apportent une interprétation et une représentation plus riches et détaillées de la mutualisation pour les assureurs.

**SYNTHÈSES, LIMITES ET PERSPECTIVES POUR LA  
MESURE DE LA MUTUALISATION**

**Sommaire**

---

<b>6.1 Synthèses des indicateurs de mutualisation proposés. . . . .</b>	<b>158</b>
<b>6.2 Limites des indices de mutualisation. . . . .</b>	<b>163</b>
<b>6.3 Perspectives des indicateurs de mutualisation . . . . .</b>	<b>166</b>

---

Ce chapitre structuré en trois points apporte une synthèse générale des indicateurs proposés dans ce mémoire. La synthèse des indicateurs passe par la confrontation des indicateurs aux souhaits attachés à une mesure de mutualisation, le choix des indicateurs les plus pertinents et leurs limites. S’en suivent alors les perspectives d’amélioration et d’utilisation de ces indicateurs.

**6.1 Synthèses des indicateurs de mutualisation proposés**

Au total, sept indicateurs de mutualisation (ou démutualisation) ont été proposés et implémentés dans cette étude. Ils présentent des propriétés et des approches différentes qu’il convient de mettre en perspective pour une prise de recul et un choix d’indicateurs les plus prometteurs.

**6.1.1 Les usages pratiques de la mesure de la mutualisation**

La nécessité de quantifier la mutualisation s’est précisée avec la montée des débats autour de l’hyper-segmentation notamment en assurance protection. La réflexion autour de ces indicateurs a permis d’identifier des usages pratiques d’une mesure de mutualisation. La figure suivante synthétise les principales utilisations des indicateurs de démutualisation.

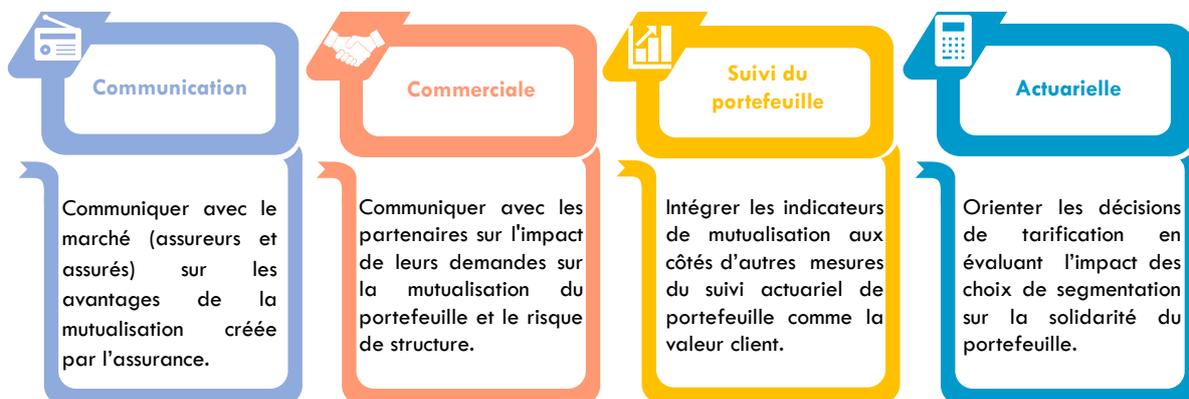


FIGURE 6.1 – Utilisations pratiques des indices de mutualisation par les assureurs

L'adoption d'un ou de plusieurs indices de mutualisation servirait à plusieurs fins en particulier :

- **Communication** : pour épauler les discussions autour de l'hyper-segmentation, en apportant une quantification à cette notion débattue, mais pas encore réellement définie ; des métriques de mutualisation fourniraient des outils aux assureurs pour défendre leurs orientations tarifaires ou de souscription auprès de la communauté des assureurs et des assurés ;
- **Commerciale** : dans les appels d'offres des partenaires, il est de coutume que les modalités de distribution soient restrictives pour assurer la simplicité du parcours client à la souscription par exemple, évaluer les effets de l'agrégation des prix sur la mutualisation et le risque de structure porté devrait simplifier les négociations avec le partenaire ;
- **Suivi du portefeuille** : pour surveiller les risques liés à la démutualisation, un suivi du niveau de mutualisation des portefeuilles peut être intégré dans le dashboard de suivi actuariel du risque. L'assureur peut en effet choisir de maintenir un niveau minimal de mutualisation ;
- **Actuarielle** : une contrainte additionnelle à prendre en compte lors de la construction de la prime technique tarifée pourrait être le maintien d'un niveau de mutualisation minimal convenu à l'avance par l'assureur en fonction de sa stratégie.

### 6.1.2 Souhaits et attentes d'une métrique de mutualisation

La littérature relative à la mutualisation en assurance, ainsi que les intuitions actuarielles précisées au chapitre 1 permettent de décliner les souhaits escomptés d'une mesure de mutualisation dans le tableau 6.1 ci-contre.

TABLEAU 6.1 – Liste des souhaits escomptés d'un indice de mutualisation

Souhait	Description
1. Dep $BC_{\Omega}$	La mesure de la mutualisation devrait dépendre des profils de risque individuel ( $BC_{\Omega}$ ) en portefeuille. Ex. un assureur qui se positionne sur un segment de risque très spécifique devrait avoir un indice de mutualisation faible
2. Dep $BC_{Ass}$ et $OL$	La mesure de la mutualisation devrait dépendre des choix tarifaires techniques ( $BC_{Ass}$ ) et commerciaux de l'assureur ( $OL$ ). Ex. un assureur qui pratique une prime commerciale unique sur tous les risques devrait avoir un niveau de mutualisation assez élevé.
3. Dep $X_{\Omega}$ et $\Omega$	La mesure de la mutualisation devrait dépendre de l'information disponible ( $X_{\Omega}$ ) qui est susceptible d'évoluer vers $\Omega$ . Ex. Sur des produits où très peu de facteurs de risque sont connus ou observés, le niveau de mutualisation devrait être <i>a priori</i> élevé.
4. Indep modèle $BC_{\Omega}$	Le profil de risque individuel $BC_{\Omega}$ lorsque $X_{\Omega}$ est fixé devrait être équivalent d'un assureur à l'autre pour garantir la comparabilité de la mutualisation entre assureurs. Ainsi, l'indice devrait être indépendant des choix de modèle $\Phi^{\Omega}$
5. Calc. pour concurrent	Pour des besoins de comparaison, il est souhaitable qu'un indice de mutualisation permette à l'assureur de se projeter un minimum sur le niveau de mutualisation de ses concurrents à partir de sa grille de prix publics par exemple.

Ces souhaits sont ceux jugés essentiels conformément aux usages des indicateurs de mutualisation indiqués sur la figure 6.1. En particulier, les souhaits 4 et 5 sont indispensables lorsque la mesure de mutualisation est utilisée à des fins de comparaison externes. D'autres souhaits secondaires peuvent être envisagés :

- L'indépendance de l'indicateur à l'historique de sinistralité : ce qui permettrait en particulier de calculer l'indicateur au lancement d'un produit ou à la signature d'un partenariat ;
- La dépendance (faible) à la taille du portefeuille : si un assureur accumule un très grand nombre d'assurés ayant des profils de risque très similaires, il crée une mutualisation sur son portefeuille bien qu'elle soit faible relativement aux risques présents sur le marché. Il y a en effet une baisse mécanique de la volatilité des sinistres qui accompagne l'accroissement du portefeuille ;
- La simplicité de calcul de l'indicateur : il est souhaitable qu'un indice de mutualisation soit simple à calculer et à expliquer pour un actuaire ;
- L'interprétabilité de l'indicateur : plus l'indicateur est interprétable, plus directe sera la communication autour de ce dernier.

Bien que certains indicateurs proposés vérifient également ces critères additionnels, ils ne sont pas utilisés pour jauger de la pertinence des indicateurs.

### 6.1.3 Les indicateurs proposés répondent-ils aux attentes ?

Le rapprochement entre les souhaits escomptés et les indicateurs proposés (tableau 6.2) montre que les attentes directement liées à la segmentation tarifaire, aux choix commerciaux et à l'information disponible sont vérifiées (au moins partiellement) par l'ensemble des indicateurs. Par contre, la majorité des indicateurs ne répond pas aux attentes 4 et 5, relatives respectivement à l'indépendance de l'indicateur aux choix de modélisation du profil de risque individuel et la possibilité de les calculer simplement pour la concurrence. En particulier, aucun des indicateurs n'est calculable pour la concurrence puisqu'ils nécessitent un minimum d'informations sur le portefeuille.

L'ambition d'obtenir un indicateur calculable pour la concurrence est en effet difficile (voire impossible) à atteindre complètement puisqu'il est nécessaire de disposer des informations sur le portefeuille et les choix de l'assureur pour évaluer son niveau de mutualisation. Toutefois, le niveau d'individualisation ne dépendant pas directement des profils de risque en portefeuille, un assureur peut évaluer la mutualisation de son portefeuille s'il avait appliqué les primes pratiquées par un assureur concurrent. De même, le ratio de mutualisation généralisé calculé sur des descriptions de fréquence peut-être évalué pour la concurrence pour peu que l'on dispose d'une description même agrégée de son portefeuille.

TABEAU 6.2 – Vérification des souhaits initiaux par les indicateurs de mutualisation

Indicateur de mutualisation/démutualisation	Dépend de $BC_{\Omega}$	Dépend de $BC_{Ass}$ ou OL	Dépend de $X_{\Omega}$ ou $\Omega$	Indépendant du modèle $BC_{\Omega}$	Calculable pour concurrent	Calculable Simplement	
Part de risque de structure cédé $ID_{Var  BC} = \frac{Var(BC^{Ass})}{Var(BC^{\Omega})}$	✔ OUI	✔ OUI	✔ OUI	✘ NON	✘ NON	✔ OUI	
Niveau d'individualisation $ID_{Ent  BC} = \frac{Ent(BC^{Ass})}{log(N^{Ass})}$	— PART.	✔ OUI	✔ OUI	✔ OUI	— PART.	✔ OUI	
Prime de solidarité $IM_{L1  BC} = \frac{1}{N^{Ass}} \sum_i  BC_i^{Ass} - BC_i^{\Omega} $	✔ OUI	✔ OUI	✔ OUI	✘ NON	✘ NON	✔ OUI	
Variabilité des OL $IM_{Var  OL} = \frac{Var(OL^{Ass})}{Var(BC^{\Omega})}$	✔ OUI	✔ OUI	✔ OUI	✘ NON	✘ NON	✔ OUI	
Individualisation des OL $IM_{Ent  BC} = \frac{Ent(OL^{Ass})}{log(N^{Ass})}$	✔ OUI	✔ OUI	✔ OUI	✘ NON	✘ NON	✔ OUI	
Part de risque de structure cédé généralisée $ID_{Int  \partial\mu_i} = \frac{\sum_{i=1}^n \ c_i^{Ass} - c^{\Omega}\ }{\sum_{i=1}^n \ c_i^{\Omega} - c^{\Omega}\ }$	✔ OUI	✔ OUI	✔ OUI	✘ NON	✘ NON	✘ NON	
Ratio de mutualisation généralisé $IM_{CK  Ent} = \frac{\min\{K(x_d) + H(T_{x_d})\}}{N^{Ass} \log_2(L_{0,0})}$	Fréquence	✔ OUI	— PART.	✔ OUI	✔ OUI	— PART.	✘ NON
	OL	✔ OUI	✔ OUI	✔ OUI	✘ NON	✘ NON	✘ NON

✔ OUI : attente vérifiée    
 — PART. : attente partiellement (indirectement) vérifiée    
 ✘ NON : attente non vérifiée

Au final, les propriétés et les interprétations des indicateurs se résument en ces termes :

- $IM_{Var||OL}$  et  $ID_{Int||\partial\mu_i}$  : les mesures du risque de structure cédé dans la segmentation sont assez pertinentes pour le monitoring du risque de structure, mais elles ne capturent pas la mutualisation telle qu'elle est perçue intuitivement, un mélange d'assurés hétérogènes. De plus, leur dépendance aux choix de modélisation du profil de risque individuel les rend difficilement comparables entre assureurs. Sous sa forme généralisée, cet indicateur permet d'identifier les facteurs qui permettent de mieux se couvrir contre le risque de structure ;
- $IM_{Ent||OL}$  : le niveau d'individualisation des OL s'est avéré insatisfaisant en pratique. Du fait, des différences (éventuellement très proches de zéro en valeur absolue) entre les profils de risque individualisés, l'entropie des OL est mécaniquement très proche de sa valeur maximale (traduisant des montants OL très différents) en portefeuille que les assurés paient des primes flattes ou très personnalisées ;
- $IM_{L1||BC}$  : la prime de solidarité représente le montant moyen de prime de risque que chaque assuré reçoit ou finance au portefeuille (en plus de sa contribution à son propre pool de risque). Sa définition, sa formule et son interprétabilité en font l'un des indicateurs les plus parlant pour la communauté des assureurs et même pour les assurés. Il dépend tout de même directement des valeurs des profils de risque modélisés par l'assureur, ce qui rend difficile de le comparer entre assureurs ou plus ardue encore de l'estimer pour la concurrence ;
- $IM_{Ent||BC}$  : le niveau d'individualisation est l'indicateur qui satisfait le plus d'attentes. Il

est facilement calculable pour peu que l'on dispose de la distribution des primes techniques du portefeuille, il ne dépend pas directement des profils de risque individualisés  $BC_{\Omega}$ . Il est donc un bon candidat pour la comparabilité externe. Son interprétation est moins immédiate que celle de la prime de solidarité mais elle reste accessible. Il s'interprète en effet comme le ratio de l'information (au sens de la multitude des facteurs de risque) contenue dans les profils de risque techniques tarifés sur l'information nécessaire pour obtenir une prime personnalisée par assuré. Son inverse est relié au nombre de profils de risque mutualisés sous un tarif technique ;

- $IM_{CK|Ent}$  : le ratio de mutualisation généralisé indique à quel point le portefeuille recouvre la structure de marché encodée dans un partitionnement de la distribution des assurés sur le marché. Sous réserve qu'une collaboration entre assureurs donne lieu à la construction d'un outil de partitionnement, cet indicateur serait très intéressant pour la comparaison externe. En effet, il dépend de la structure des risques en portefeuille (puisqu'elle est retranscrite dans les descriptions) ; il dépend partiellement des choix commerciaux et tarifaires lorsqu'il est calculé sur des descriptions de fréquence (en situation concurrentielle, les choix de l'assureur ont une incidence importante sur les profils sous-crits). Lorsque les descriptions des OL sont disponibles, l'indicateur est plus riche mais plus dépendant des choix de modélisation également. Néanmoins, le calcul du ratio de mutualisation généralisé et son interprétation sont relativement longs et complexes.

### **Notre proposition : le niveau d'individualisation et la prime de solidarité**

---

Nous suggérons *in fine* de retenir deux métriques de mutualisation complémentaires : **le niveau d'individualisation et la prime de solidarité**. Ils ont de bonnes intuitions et interprétations, ont été cohérents sur les applications empiriques tout au long de ce mémoire. Le niveau d'individualisation répond très bien aux attentes aussi bien principales que secondaires, mais il ne dépend pas directement des montants de primes. La prime de solidarité calculée en € vient combler cette limitation. Une mention particulière est également portée sur le ratio de mutualisation généralisé pour son calcul basé sur une description agrégée du portefeuille (soit moins d'informations sur le portefeuille), sa comparabilité entre assureurs du fait d'un socle de calcul commun et son interprétation qui éclaire sur le rôle des facteurs dans la mutualisation.

#### **6.1.3.1 Les effets de la démutualisation sur l'assureur et sur le marché**

---

Sur la base de ces indicateurs, les niveaux de mutualisation associés à différentes stratégies tarifaires ont été évalués. A partir de la mise en concurrence sur les prix de quelques stratégies tarifaires différentes, les effets suivants (figure 6.2) de la segmentation tarifaire sur les performances financières des assureurs et les effets sur le marché ont été observés.

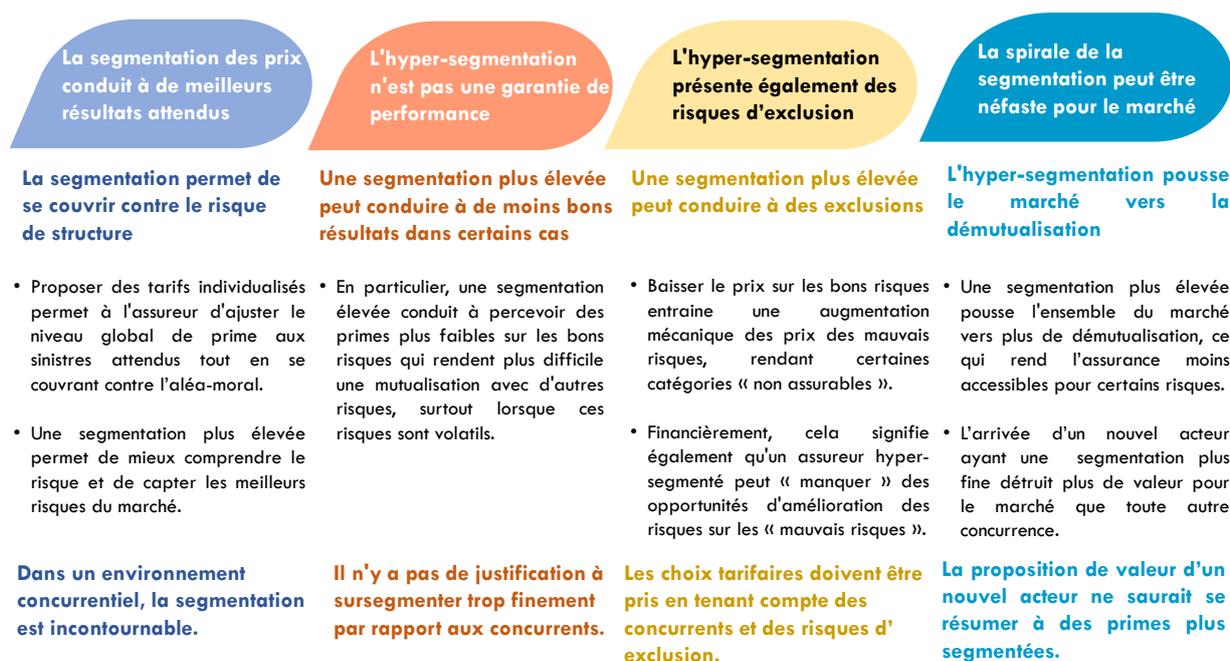


FIGURE 6.2 – Effets de la démutualisation sur les assureurs et sur le marché en situation de concurrence

## 6.2 Limites des indices de mutualisation

Bien que les indicateurs soient en accord avec les intuitions actuarielles, ils ne sont pas pour autant exempts de limites. Dans cette section, une discussion concernant les principales limites théoriques et pratiques identifiées sur les trois indicateurs les plus pertinents est menée.

### 6.2.1 Nécessité de disposer des informations sur le marché pour assurer la comparabilité externe

Pour assurer la comparabilité des indicateurs, il est nécessaire de disposer d'une référence commune à tous les assureurs actifs sur un marché donné. Ce socle commun concerne : la connaissance de la structure du portefeuille de marché et l'homogénéité des profils de risque individualisés entre assureurs. La bonne comparabilité des indicateurs suppose que les modèles de profils de risque individuels sont calculés sur des distributions comparables des facteurs de risque. En particulier, les assureurs disposeraient d'un même niveau d'information sur les assurés et plus encore adopteraient des modélisations qui conduiraient à des profils de risque individualisés semblables.

En pratique, ces deux conditions ne sont pas nécessairement vérifiées. Certains assureurs disposent d'un avantage technologique qui pourrait leur conférer à court ou moyen terme une exclusivité sur des variables tarifaires (comme les données télématiques qui se démocratisent lentement chez tous les assureurs), un assureur spécialisé sur des classes de risque spécifiques ne dispose pas nécessairement des informations sur tout le marché. Enfin, on observe empi-

riquement que sur la même base de données, pour déterminer le plus finement possible les profils de risque individuels, les actuaires et les data scientists emploient des méthodes diverses (GLM, GAM, boosting, réseau de neurones, etc.) et aboutissent à des profils de risque individuels différents (observation faite dans le cadre des compétitions de tarification sur des données d'assurance, *Pricing Games*<sup>1</sup>).

Le niveau d'individualisation est robuste à ces limites, mais ce n'est pas le cas ni pour la prime de solidarité ni, pour le ratio de mutualisation généralisé.

### 6.2.2 La justice et l'équité des pratiques actuarielles pas directement prises en compte

À ce stade de construction des indicateurs de mutualisation, la réflexion s'est concentrée sur les aspects techniques et actuariels des métriques sans apporter de jugement de valeur. De plus, dans le cadre de l'extension de garantie automobile, les questions d'équité actuarielle sont moins importantes ; toutefois, les questions de jugement de valeur peuvent vite gagner en importance en cas d'application des indicateurs à d'autres produits comme l'assurance des emprunteurs.

La « justice actuarielle » constitue encore aujourd'hui un sujet de discussion ouvert en assurance . En effet, lorsque la segmentation tarifaire n'est pas réalisée à partir des facteurs *causaux* ou à des facteurs *causaux* inacceptables socialement (exemple : sexe, état de santé etc.), elle donne lieu à des polémiques et/ou à des interdictions légitimes. Une polémique qui subsiste même lorsque seuls des facteurs *causaux* sont utilisés, est celle de la solidarité créée par l'assureur qui n'est pas toujours bien reçue ou acceptée par les assurés. Comme expliqué dans cette étude, modéliser le profil de risque s'apparente à la constitution des groupes d'assurés, à admettre au regard de l'information disponible qu'à l'intérieur de ces classes, les assurés sont homogènes et donc cotiseront les mêmes primes techniques. L'assureur crée une solidarité entre les assurés puisqu'il n'a qu'une information imparfaite ou par choix. D'aucuns pourraient alors envisager les prix individualisés ou les « justes prix » comme portés par l'analyse prédictive ; mais là encore il subsiste le problème de la causalité des facteurs utilisés pour l'hyper-segmentation et plus généralement les risques de démutualisation.

Les indicateurs proposés mesurent le niveau de mutualisation d'un portefeuille, mais ne traduisent pas nécessairement que les pratiques mises en place pour constituer ledit portefeuille sont justes et équitables :

- La prime de solidarité calcule la solidarité créée entre les assurés sans tenir compte du sens de financement. En effet, cet indicateur ne pénalise pas les situations où les mauvais risques financeraient les bons risques. Or la solidarité entre assurés évoque intuitivement que les bons risques sont ceux en position de financement et que les mauvais risques sont assurables au prix de cet effort collectif. La figure suivante illustre cette invariance de la

---

1. organisés par la chaire ACTINFO intitulée Valorisation et nouveaux usages actuariels de l'information (2016 - 2019), <https://actinfo.hypotheses.org/category/game>.

prime de solidarité au sens de financement.

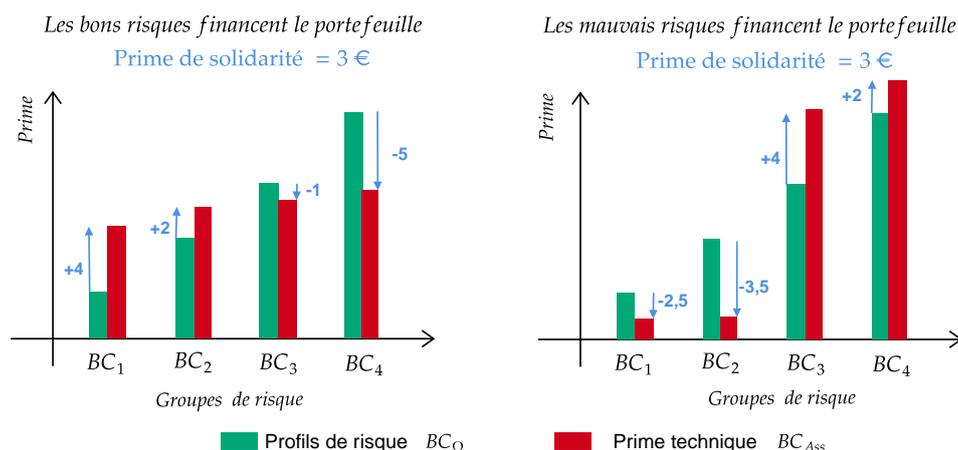


FIGURE 6.3 – Invariance de la prime de solidarité au sens de financement du portefeuille

Il convient de souligner tout de même qu'en pratique les régulateurs (par des restrictions tarifaires) et la présence des assureurs concurrents (par le jeu de la concurrence sur les prix) limitent grandement les schémas dans lesquels les mauvais risques financeraient les bons risques ;

- Le niveau d'individualisation par construction dépend de la distribution des primes techniques pratiquées, mais pas de leurs montants explicitement. Cela signifie en particulier que si un assureur décide de modifier les montants de primes tout en gardant les distributions constantes, le niveau d'individualisation demeurera constant. Il peut par exemple décider d'augmenter la prime technique des mauvais risques d'un montant équivalent prélevé sur les bons risques.

Ces exemples simples illustrent bien le fait que la mesure de la mutualisation seule ne rend pas compte de l'équité des pratiques tarifaires actuarielles. Ces indicateurs doivent donc être accompagnés d'autres mesures de valeur client. Le tableau suivant liste quelques mesures couramment utilisées par les actuaires ou suggérées sur la base des indicateurs de mutualisation.

TABLEAU 6.3 – Quelques exemples d'indicateurs susceptibles d'accompagner la mesure de la mutualisation

Grandeur	Métrique
<b>Niveau de mutualisation</b> mesure de la solidarité créée entre les assurés du portefeuille	Niveau d'individualisation : ID_Ent_BC, prime de solidarité : IM_LI_BC
<b>Valeur client</b> mesure la part de la prime retournée aux assurés sous forme d'indemnisation de sinistres	S/P du portefeuille Distribution des S/P à profils de risque équivalents
<b>Écart (signés) entre les primes techniques et les profils de risque</b> Permet de cerner le « sens » de la solidarité créée par l'assureur	Synthèse de distribution : Moyenne, et mesures de dispersions Corrélation avec les profils de risque les plus fins
<b>Hétérogénéité des traitements assureurs</b> mesure des différences des traitements commerciaux pratiqués par l'assureur (ex. price walking)	Volatilité du rapport OL/ Prime pure individualisée en portefeuille Distribution de la volatilité des OL à profil de risque équivalent
<b>Taille du portefeuille</b>	Nombre d'assurés en portefeuille
<b>Nombre de profils de risque</b>	Nombre de prime technique unique

## 6.3 Perspectives des indicateurs de mutualisation

### 6.3.1 Améliorer la conception et l'interprétation des indicateurs

Comme souligné à la section précédente, les indicateurs de mutualisation proposés ne sont pas exempts de limites ; remédier à ces limites constitue un axe privilégié d'amélioration des indicateurs. Plus précisément, des études ultérieures pourraient :

- Analyser l'influence théorique et empirique de la modélisation du profil de risque individuel sur les indices qui souffrent de ces effets de modèle que sont la prime de solidarité et le ratio de mutualisation généralisé ;
- S'intéresser à l'intégration de la mesure de l'équité dans les indicateurs de mutualisation directement ;
- Améliorer la formulation théorique de la part du risque de structure généralisée qui n'est pas correctement normalisée et donc difficilement comparable ;
- Prolonger les analyses sur les liens entre la mutualisation et l'inclusion à l'assurance, pour mieux préciser et contenir les risques d'exclusion associés à l'hyper-segmentation.

Certaines interrogations demeurent ouvertes à la fin de cette étude : elles concernent non seulement la conception des moyens de quantification de la mutualisation, mais également l'évaluation des conséquences de la démutualisation. Néanmoins, l'étude a permis de mieux cerner le sens de la mutualisation et de la démutualisation, d'apporter une appréciation quantitative des conséquences de la démutualisation sur des données réelles, de décliner des indicateurs certes perfectibles, mais satisfaisants pour évaluer la mutualisation d'un portefeuille.

### 6.3.2 Perspectives chez BNP Paribas Cardif

Une mesure de mutualisation n'aura véritablement de sens que si elle est adoptée par la communauté des assureurs. De ce fait ce mémoire souhaite promouvoir la mesure de la mutualisation pour en faire un enjeu de place. Dans ce sillage, le script Python implémenté pour le calcul des indicateurs ainsi qu'un exemple d'application numérique sont mis à la disposition des assureurs sous la forme de projet GitHub <https://github.com/B-Noumedem/Indices-de-mutualisation.git>.

Chez BNP Paribas Cardif, l'ambition quant-à la mesure de la mutualisation est double : l'intégrer dans les tableaux de bord du suivi des portefeuilles et continuer à améliorer les indicateurs pour corriger les limites sus mentionnées. BNP Paribas Cardif envisage très prochainement l'application des indicateurs proposés au cas de l'assurance des emprunteurs, produit pour lequel la mutualisation fait grandement débat auprès de la communauté aujourd'hui. Des travaux sur la prise en compte de l'équité dans la formulation des indicateurs et l'étude des liens entre mutualisation et assurabilité sont envisagés comme prolongement de ce stage. La figure suivante

présente un exemple d'intégration du niveau de mutualisation au tableau de bord de suivi du portefeuille d'extension de garantie automobile.

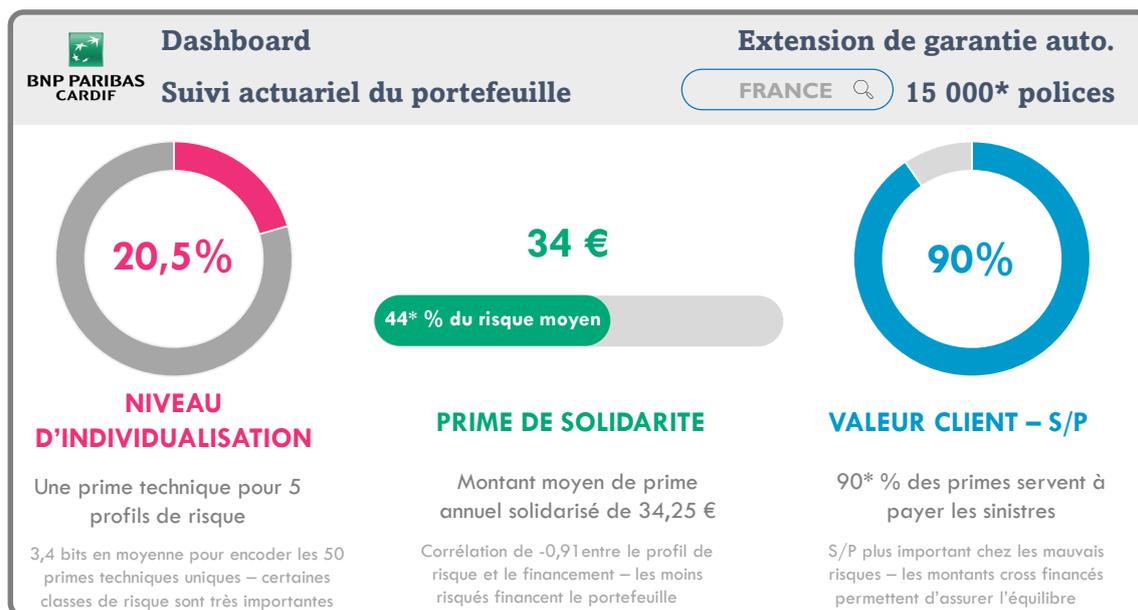


FIGURE 6.4 – Intégration des indicateurs de mutualisation à un dashboard de suivi de la conformité

## Conclusion du chapitre

La mesure de la mutualisation génère des souhaits et des attentes majeures auxquels les indicateurs proposés répondent assez favorablement. Les souhaits directement associés aux pratiques actuarielles qui conduisent à la démutualisation du portefeuille sont vérifiés par tous les indicateurs. À l'inverse des souhaits visant à assurer la comparabilité des mesures entre plusieurs assureurs, sont plus délicats et ne sont approchés que par deux des indicateurs.

La prime de solidarité et le niveau d'individualisation constituent nos principales propositions pour évaluer la mutualisation d'un portefeuille : ils sont complémentaires, relativement faciles à calculer et à interpréter. Une prise de recul sur les indicateurs permet d'en identifier quelques limites dont les plus importantes sont la non-prise en compte de l'équité et la dépendance des indicateurs à la modélisation du profil de risque le plus fin. Remédier à ces limites constitue l'axe d'amélioration privilégié de ces indicateurs.

---

## Conclusion

L'accès à des données plus granulaires sur les assurés et l'utilisation des algorithmes de machine learning offrent de nouvelles possibilités. Couplée aux pratiques actuarielles et commerciales qui tendent à une plus grande individualisation des tarifs en assurance, l'exploitation de ces opportunités fait émerger plusieurs questionnements inhérents à la démutualisation et à ses conséquences. Ce mémoire apporte une contribution à l'analyse des causes, des moyens et des effets de la (dé)mutualisation des portefeuilles. Cette contribution concerne spécifiquement :

- La proposition d'une formalisation de la (dé)mutualisation en proposant une définition inspirée de la littérature, des débats et des intuitions actuarielles ;
- La déclinaison des indicateurs calculables sur des portefeuilles afin d'évaluer leur niveau de mutualisation ou de démutualisation ;
- L'analyse empirique et quantitative des effets de la démutualisation sur l'assureur et l'assurance.

Pour formaliser la démutualisation, il a été nécessaire de revenir sur une notion fondamentale en assurance, la mutualisation. Son lien indéniable avec la segmentation des assurés et le transfert du risque de structure qui en découle ont été étudiés dans ce sillage. Lorsque l'assureur adopte des primes techniques très finement segmentées, il fait payer à chaque assuré un montant proche de son risque individuel : le portefeuille n'est pas cross-financé entre assurés, c'est-à-dire que chaque groupe de risque ne génère pas assez de marges pour financer les déficits éventuels des autres groupes de risque. Sur l'aspect commercial, lorsque l'assureur applique des chargements différents entre les assurés, il prélève chez les uns pour financer les autres, il crée donc de la solidarité entre les assurés. Nous avons alors proposé de définir la démutualisation comme « la désolidarisation des assurés de l'ensemble de la communauté d'un assureur, le cas extrême étant une prestation individuelle certaine avec un niveau de marge constant par assuré, sans aucun subside croisé ».

Les indicateurs de mutualisation ont été élaborés pour quantifier les effets des choix tarifaires sur la solidarité entre assurés en portefeuille. Ces indicateurs sont basés principalement sur deux métriques de dispersion : la variance et l'entropie. La variance mesure la volatilité d'une variable tandis que l'entropie mesure l'hétérogénéité au sein d'une population. Des indicateurs dits simples ont été obtenus en appliquant ces métriques à la prime et à ses composantes. L'idée charnière des cinq indicateurs simples proposés est d'évaluer l'écart entre la tarification technique adoptée par l'assureur (la prime technique pratiquée) et celle qu'il pourrait appliquer s'il utilise toute l'information à sa disposition dans son tarif (profil de risque individualisé). À l'exception du niveau d'individualisation défini comme le rapport de l'information utilisée dans la tarification sur l'information nécessaire pour individualiser les primes techniques, tous les indicateurs simples dépendent des choix de modélisation du profil de risque individualisé. Cette dépendance rend la comparabilité des niveaux de mutualisation entre assureurs assez

---

discutable.

Pour s'affranchir de cette dépendance au modèle, du besoin en informations détaillées sur le portefeuille et pour mieux interpréter la mutualisation, les indicateurs exotiques ont été proposés. Ils sont calculés sur les caractéristiques des assurés ou une synthèse de la distribution de ces caractéristiques. La variance et l'entropie ont été généralisées respectivement en inertie et en entropie algorithmique. L'inertie mesure la dispersion dans un nuage de points multidimensionnel ; elle est appliquée au nuage des contributions des facteurs à la prime pure. L'indicateur qui en découle est une généralisation de la mesure du risque de structure cédée par la segmentation tarifaire. L'entropie algorithmique quant à elle est une mesure de l'hétérogénéité dans un système physique pour une description partielle du système. Elle se décompose comme la somme de la complexité de Kolmogorov de la description du système (qui mesure l'hétérogénéité entre les particules, connue à partir de la description) et de l'entropie de Shannon (qui quantifie l'hétérogénéité résiduelle). Pour l'appliquer à la mesure de la mutualisation, le système physique est remplacé par l'espace engendré par les facteurs de risque, les particules par les assurés et la description du système par une synthèse de plus en plus précise de la dispersion des assurés du portefeuille.

Le portefeuille d'extension de garantie automobile de BNP Paribas Cardif a servi pour les analyses numériques de ces indicateurs. Après des pré-traitements et des analyses descriptives de ce jeu de données, des modèles de calcul des primes pures ont été calibrés à partir du GLM et du LightGBM. Ces modèles permettent d'évaluer la sinistralité globale attendue avec une erreur absolue de moins de 2 %, la part de déviance expliquée étant de l'ordre de 13 %. L'application des métriques de mutualisation sur ces primes conduit à des résultats très cohérents : les indicateurs sont en accord avec les intuitions actuarielles et la définition formulée tout en restant cohérents entre eux. Notamment, le niveau d'individualisation de ce portefeuille est de 20,5 % ce qui signifie que la stratégie tarifaire n'utilise qu'un cinquième de l'information nécessaire pour individualiser les prix ; ou plus schématiquement, un tarif technique regroupe en moyenne cinq profils de risque différents. La prime de solidarité, calculée comme le montant moyen cross-financé entre assurés, est de l'ordre de 34 €.

Les indicateurs de mutualisation peuvent être utilisés pour choisir les variables tarifaires qui permettent de respecter un certain niveau de mutualisation minimal ou pour des communications externes. Ils ont également été utilisés pour apprécier les effets de la démutualisation sur le marché et sur les assureurs. Si hyper-segmenter permet de se couvrir contre le risque de structure et de mieux comprendre le risque assuré, sur un exemple empirique de mise en concurrence, il a été observé que l'assureur le plus performant financièrement n'est pas nécessairement le plus démutualisé, mais peut-être celui qui a une segmentation intermédiaire. La démutualisation en situation de concurrence entraîne également une diminution de la valeur pour le marché et des exclusions qui s'accroissent avec la finesse de la segmentation. Les indicateurs proposés ne rendent pas compte de l'équité ou non d'une pratique actuarielle. La prise en compte de l'équité dans leur construction pourrait constituer un axe d'amélioration.

## GLOSSAIRE ET ACRONYME

**BDM** Bloc Decomposition Method.

**Bit** Unité élémentaire d'information pouvant prendre deux valeurs distinctes, 0 ou 1. Une chaîne binaire est une séquence de bits décrivant un objet.

**Bon/mauvais risques** Assurés présentant une sinistralité attendue faible/élevée.

**Boosting** Famille de modèles d'apprentissage automatique qui regroupe un ensemble de modèles « simples » de façon itérative.

**Burning cost/Prime pure** Sinistralité moyenne attendue pour un assuré en portefeuille sur un exercice. Le burning cost le plus fin est la prédiction la plus précise (par assuré) que l'assureur peut construire. Le burning cost tarifé ou prime technique est une estimation moins fine de la sinistralité individuelle utilisée directement dans la grille des tarifs.

**CART** Classification And Regression Tree.

**Complexité de Kolmogorov (CK)** Longueur du plus petit programme informatique (exprimée en bits) qui génère un objet donné.

**CTM** Coding Theorem Method.

**Démutualisation** Désolidarisation des assurés de l'ensemble de la communauté des assurés, le cas extrême étant une prestation individuelle certaine avec un niveau de marge constant par assuré, sans aucun subside croisé.

**Entropie** Fonction qui caractérise l'état de «

désordre » d'un système, qui mesure l'incertitude sur la nature d'un système ou d'un message.

**Entropie Algorithmique** Somme de la complexité de Kolmogorov d'une description d'un système et de l'entropie de Shannon des micro-états compatibles avec cette macro-description.

**facteur de risque** Variable tarifaire utilisée pour segmenter les assurés en groupes de risque homogènes.

**FCA** Financial Conduct Authority.

**GLM** General Linear Model ou Modèle linéaire généralisé.

**GMI** Global Market Insights.

**GPS** Global Positioning System ou Système de localisation par satellite.

**Juste prix/prix individuel** Prime d'un contrat d'assurance tarifé à partir d'un modèle de segmentation très fin.

**LightGBM** Light Gradient Boosting Machine.

**Macro-état/micro-état** Le macro-état d'un système est une description synthétique de l'état d'un système par une mesure des agrégats sur ce dernier. Un micro-état est une caractérisation plus précise d'un système, au niveau des unités élémentaires qui constituent le

---

ystème. À un macro-état donné, correspond un ou plusieurs micro-états.

**MAE** Mean Absolute Error ou Erreur absolue moyenne.

**Mutualisation** action par laquelle l'assureur met en commun plusieurs assurés faisant face à des risques homogènes et indépendants.

**OL** Optimisation Layer, écart entre la prime commerciale hors taxes et le profil de risque le plus fin estimé de l'assuré. Elle comporte les marges assureurs, les commissions, les frais et les marges de solidarité entre assurés.

**Portefeuille** Ensemble d'assurés couverts pour un risque donné et observés sur un ensemble de caractéristiques sur un exercice.

**Portefeuille de marché** Distribution (théorique) de l'ensemble des assurés ou prospects suivant les facteurs de risque pour une couverture d'assurance donnée.

**RGPD** Règlement Général sur la Protection des Données.

**Risque de structure** Risque de déviation

de la structure du portefeuille par rapport à la structure utilisée (ou anticipée) lors de la construction des primes du produit.

**RMG** Ratio de mutualisation Généralisé.

**RMSE** Root Mean Square Error ou Erreur quadratique moyenne.

**Segmentation** Pratique de différenciation de la prime de risque en fonction des caractéristiques de l'assuré.

**SHAP** SHapley Additive exPlanations.

**SMURF** Sparse Multi-Type Regularized Feature Modeling.

**Solidarité** compensation des flux financiers au sein d'un portefeuille d'assurés faisant face à des risques semblables ou non.

**Splitter** Partitionnement itératif de l'espace des facteurs de risques suivant leur importance, par un arbre de régression par exemple.

**Structure du portefeuille** Distribution empirique des assurés au regard des facteurs de risque pertinents au sein d'un portefeuille donné.

## LISTE DES SYMBOLES

$\mathbf{X} = (X_1, X_2, \dots, X_p)$ Vecteur aléatoire des facteurs de risque, variables tarifaires	$\mathbf{P}_c^{Ass}$ Vecteur des primes commerciales hors taxes
$\Omega$ Ensemble des facteurs de risque qui influent sur la survenance et la sévérité du risque étudié	$\mathbf{OL}^{Ass}$ Vecteur des optimisations layers
$\mathbf{X}_\Omega \subseteq \Omega$ Meilleure approximation de $\Omega$ connue par l'assureur	$\Gamma_{x_d}$ État macroscopique du portefeuille décrit par $x_d$
$\mathbf{X}_c^{Ass} \subseteq \mathbf{X}_\Omega$ Sous ensemble de $\mathbf{X}_\Omega$ utilisé par l'assureur dans la grille de prime technique	$x_d^{freq}$ Description du portefeuille au bout de $d$ mesures sur les fréquences
$\Phi$ Fonction de tarification qui associe à un vecteur de facteurs de risques une prime pure	$x_d^{OL}$ Description du portefeuille au bout de $d$ mesures sur les variations d'optimisation layer
$\mathcal{P}f^{Ass}$ Portefeuille des assurés	$\text{Var}(\cdot)$ Variance de la variable
$N^{Ass}$ Taille du portefeuille, nombre d'assurés	$\text{Ent}(\cdot)$ Entropie de la variable
$\mathbf{BC}^{Ass}$ Vecteur des primes techniques ou burning costs techniques	$IM$ Indice de mutualisation
	$ID$ Indice de démutualisation

# Executive summary

## 1 Introduction and background

The insurance mechanism relies on risks mutualization, which consists of pooling together insureds facing similar risks within the same portfolio. This mutualisation is an actuarial decline of the law of large numbers (Baker and Simon, 2001; Charpentier, 2011), which assumes the independence and homogeneity of risks. However, the risks faced by policyholders are rarely homogeneous, this leads insurers to segment premiums according to the risk factors captured through the pricing variables.

Given the growing data availability about policyholders and their life habits, and the increasing willingness of policyholders to pay the “*right price*” of their personal risk, premium segmentation is becoming more and more granular, the extreme case being the premium individualization. The opportunities arising from these trends lead to the emergence of *insurtechs* and highly personalised insurance products such as Usage Based Insurance. This penalization of risks and the fears that it raises come up in actuarial debates under the heading of **demutualization**. One of the major dangers for the community is the exclusion of a category of the population (L’Argus, 2021; Les Echos, 2019; L’Argus, 2016), in particular that of the *bad risks* (since they would be proposed very or unaffordable premiums). For the insurer, it could tighten its insurance offer on a small market niche leading to more volatility result and a loss of expertise on the other risk profiles.

Although mutualization is fundamental in insurance and is on the agenda of actuarial congresses, its definition, its measurement and the consequences of demutualization remain poorly addressed in the literature, beyond the intuitions around the law of large numbers. This study aims to formalise these notions and to provide calculable indicators to quantify the level of mutualization in a real portfolio. The study also examines the empirical and quantitative analysis of the effects of demutualization on the insurer and the insurance market.

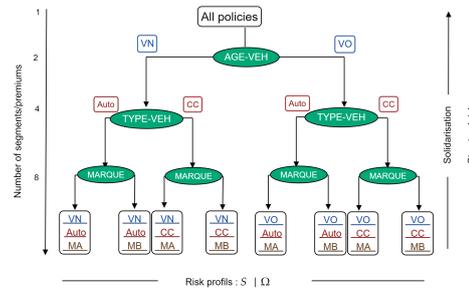
### Data

The data from BNP Paribas Cardif’s extended motor warranty distributed in France was used for the numerical application of this study. This product is interesting for this study due to the long period of data availability. More importantly, the possible thinness in segmentation that can be achieved with the pricing factors. For example by taking into account the vehicle model in the pricing model which has a very high cardinality.

## 2 Mutualization and demutualization: issues and formalizations

### 2.1 Mutualization and segmentation

In order to protect against adverse selection and to set up an insurance portfolio, it may be necessary to segment the policyholders according to pricing factors. The actuary establishes homogeneous risk classes within which he mutualises the risks of a group of insured persons. Within each cluster, there is still residual heterogeneity. This is due to the incompleteness of the observed or used risk factors and to the “unpredictable randomness”. A pure premium amount per risk class is then determined from an empirical estimate as the expected loss of the cluster (Charpentier et al., 2015).



**Fig. 1.** Premium segmentation and solidarity within the portfolio: extended motor warranty with 3 risk factors

In a scenario where there are enough policyholders within each cluster, the law of large numbers would be expected to apply within each cluster of homogeneous risks, so that premiums collected and claims paid within a cluster would balance out. In practice, homogeneity within a cluster increases with the number of risk factors used, but inversely the number of policyholders per cluster decreases.

As segmentation becomes more granular, the dangers of demutualization such as **exclusions** (Keller et al., 2018), **actuarial unfairness** (Frezal and Barry, 2020), **premium instability** (Hay, 2015), **focusing on market niches** (Charpentier et al., 2015) increase. Solidarity between insureds allows the portfolio to be balanced directly instead of each cluster : the premium paid by any insured can be transferred to other pools of risk.

## 2.2 Formalizing mutualization

### Technical pricing as a tool for mutualization

Let  $S$  be the total loss per policyholder,  $\mathbf{X} = (X_1, X_2, \dots, X_p)$  the random vector of observed risk factors and  $\Omega = (\Omega_1, \Omega_2, \dots)$  the theoretical random vector including all the characteristics which affect the covered risk. By studying three risk sharing scenarios based on the pure premium between insurer and insured introduced by Denuit and Charpentier (2004), the relationship between segmentation and risk sharing can be formalized:

**Tab. 1.** Risk sharing between insurer and policyholder community by level of segmentation.

	Without segmentation (a)		Imperfect segmentation (b)		Perfect segmentation (c)	
	Insured	Insurer	Insured	Insurer	Insured	Insurer
Expenditure	$\mathbb{E}(S)$	$S - \mathbb{E}(S)$	$\mathbb{E}(S/\mathbf{X})$	$S - \mathbb{E}(S/\mathbf{X})$	$\mathbb{E}(S/\Omega)$	$S - \mathbb{E}(S/\Omega)$
Average expenditure	$\mathbb{E}(S)$	0	$\mathbb{E}(S)$	0	$\mathbb{E}(S)$	0
Variance	0	$\mathbb{V}(S)$	$\mathbb{V}(\mathbb{E}(S/\mathbf{X}))$	$\mathbb{E}(\mathbb{V}(S/\mathbf{X}))$	$\mathbb{V}(\mathbb{E}(S/\Omega))$	$\mathbb{E}(\mathbb{V}(S/\Omega))$

- Scenario (a) corresponds to absence of segmentation, the entire structure risk is supported by the insurer. In this scenario, each policyholder pays the same premium regardless of their own risk factors;
- Scenario (c) is a theoretical case in which the insurer observe and use every risk factors  $\Omega$  and the structural risk supported by the insurer would then be minimal. The premium paid by each insured is conditional on his own characteristics and therefore random ;
- Scenario (b) is an intermediate scenario between (a) and (c); depending on the availability of risk factors and its commercial strategy, the insurer proceeds to a partial segmentation of policyholders. That creates both mutualization and solidarity among policyholders.

### Insurance loadings as a tool for mutualization

Commissions	Optimisation Layer (OL)	Commercial premium
Acquisition costs		
Administrative fees		
Insurer's margin		
Prime pure / Burning cost (BC)	Risk profile	

**Fig. 2.** Decomposition of commercial premium into optimisation Layer (OL) and pure premium (BC).

In practice, on the the pure premium, the insurer add positive or negative loadings, (designated by optimisation layer -OL -) to have the commercial premium paid by the insured. By choosing to apply different margin levels to specific parts of the portfolio, the insurer creates additional solidarity between its policyholders. In fact, on different risks, the insurer may decide to use the OL to make all policyholders pay the same commercial premium, thus creating a mutualise portfolio. Similarly, the insurer can use the OL to finance acquisition or retention of contracts, by applying underwriting discounts or to the contracts with a high probability of cancellation (see “**price-walking**” (FCA, 2020)).

## 2.3 Proposed definition for demutualization

We propose to define the demutualization of a portfolio as the desolidarisation of of policyholders from the community, the extreme case being a certain individual benefit with a constant level of margin per policyholder, without any cross subsidy. Thus, as each insured pays for his risk independently of the other policyholders, a portfolio will be all the more demutualized. On the other hand, a portfolio will be said to be mutualized if, within it, the insurer create solidarity between policyholders in addition to mutualization between policyholders with the same risk profile when all available risk factors are taken into consideration, i.e. indivisible mutualization.

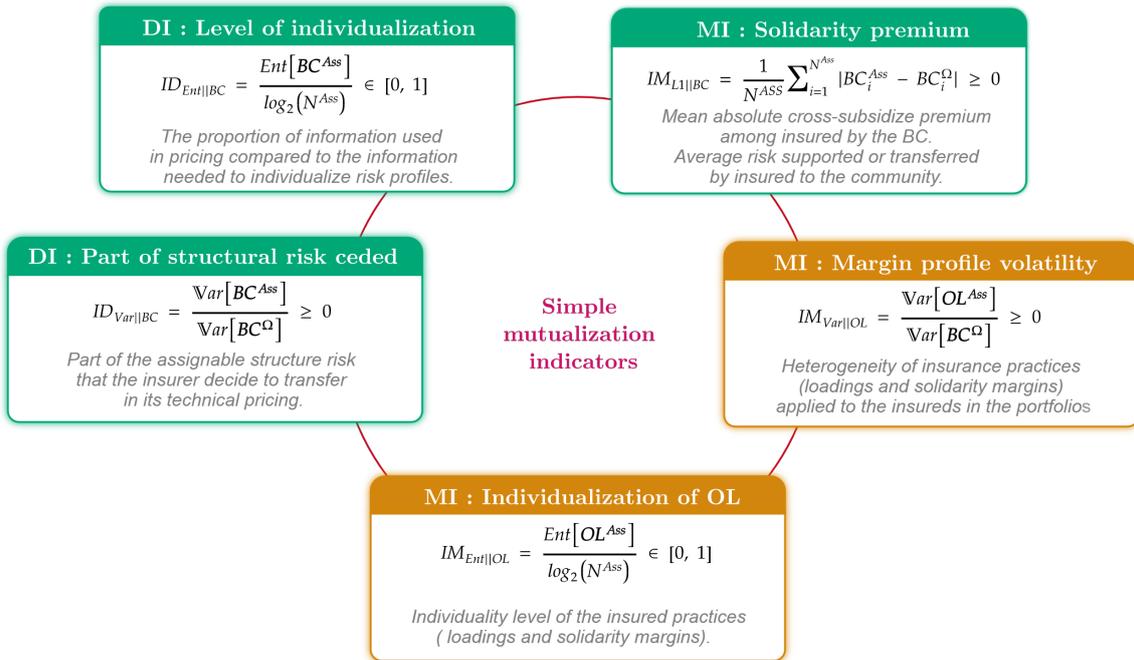
## 3 Theoretical proposals for mutualization indicators

### 3.1 Simple indicators based on the premium and its components

Demutualization practices such as hyper-segmentation of technical premiums and heterogeneity of margins per insured can be quantified by assessing the dispersion of technical premiums and OL respectively. Two measures of uncertainty were used to establish simple indicators. For a random variable  $X$  taking the values  $x_1, x_2, \dots, x_n$  with probabilities  $p_1, p_2, \dots, p_n$  :

- **Variance**  $\mathbb{V}[X] = \sum_{i=1}^n p_i(x_i - \bar{x})^2$ , avec  $\bar{x} = \frac{1}{n} \sum_{i=1}^n p_i x_i$ : it measures the volatility of the variable around its mean value; is as a result of the formalization of the transfer of structural risk through the segmentation of premiums ;
- **Entropy**  $Ent(X) = - \sum_{i=1}^n p_i \log_2(p_i)$  : it measures the heterogeneity within a population or a system; in information theory, it measures the average quantity of information contained in a message delivered by a random source.

The key idea of the simple indicators is to evaluate the difference between the technical premium adopted by the insurer ( $BC^{Ass}$ ) and the most personalized premium ( $BC^\Omega$ ) that could reach if the insurer used all the information available. The simple indicators and their interpretations are given by figure 3.



**Fig. 3.** Simple indicators for measuring the mutualization of a portfolio based on risk profiles ■ or on OL ■. MI: Mutualization Index (increases with the solidarity between policyholders), DI: Demutualization Index (increases with the desolidarisation of the community).

### 3.2 Exotic indicators of mutualization based on risk factors

In order for the measurement of mutualization to be commonly adopted by insurers, generalizations of simple indicators are introduced. The exotic indicators seek to overcome the dependence of simple indicators on the individualized risk profile modeling choices, to reduce the required information on a portfolio and to better characterize mutualization by identifying the role of risk factors.

#### Inertia of risk contributions to quantify mutualisation

The contribution of the risk factor  $X_j$  for insured  $i$  (denoted  $c_{i,j}$ ) is calculated as the sensitivity of the expected claims predicted by the model for this insured following a variation of the factor  $X_j$ . When a machine learning model is used to predict the burning cost,  $c_{i,j}$  can be approximated using SHAP values (Lundberg and Lee, 2017), when the GLM is used to model the expected loss experience,  $c_{i,j}$  can be computed as follows :

$$c_{i,j} := \frac{\partial \log \left( \hat{\mathbb{E}}[S_i | \mathbf{X}_i = \mathbf{x}_i] \right)}{\partial X_j} = \frac{\partial}{\partial X_j} \left( \hat{\beta}_0 + \sum_{j=1}^p \sum_{l=1}^{L_j} \mathbb{I}_{\{x_{i,j} \in I_{j,l}\}} \hat{\beta}_{j,l} \right) = \hat{\beta}_{j,l}$$

where  $l$  is the rank of the discretized interval of  $X_j$  in which the observation  $x_{i,j}$  falls. For all insureds and all risk factors, a high dimensional cloud of contributions is obtained  $\mathcal{C}_{N^{Ass}} = \{c_i \in \mathbb{R}^p, i = 1, \dots, N^{Ass}\}$ .

The dispersion of this cloud reflects the structural risk transfer decided by the insurer. Similar and slightly disparate contributions reflect little differentiation of risk profiles by the technical premium. Inversely, a very sparse contribution cloud reflects a very granular premium segmentation. Using the **inertia** to evaluate such dispersion leads to the following indicator:

#### DI : Part of structural risk ceded generalized

$$ID_{Int||\partial\mu_i} = \frac{I(\mathcal{C}_N^{Ass})}{I(\mathcal{C}_N^\Omega)} = \frac{\sum_{i=1}^{N^{Ass}} \|c_i^{Ass} - \bar{c}^{Ass}\|_2^2}{\sum_{i=1}^{N^{Ass}} \|c_i^\Omega - \bar{c}^\Omega\|_2^2}$$

$c_i^{(a)}$  : contributions of the insured  $i$  for model (a) and  $\bar{c}^{(a)}$ : average contributions by variable.

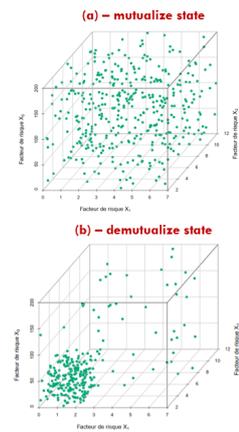
#### Algorithmic entropy of a portfolio description

**Algorithmic entropy** (Zurek, 1989) measures the heterogeneity in a physical system given his partial description (denoted  $x_d$ ). It is decomposed as the sum  $\mathcal{S}_{Gamma_{x_d}} = K(x_d) + H(\Gamma_{x_d})$  of :

**Complexity of Kolmogorov** (Kolmogorov, 1968; Li and Vitányi, 2019) of the system description that measures the heterogeneity between particles described by  $x_d$ ;  $K(x_d) := \min_p \{|p| : U(p) = x\}$ , the length of the smallest program  $p$  that generate the description  $x_d$  when run on a universal Turing machine  $U(\cdot)$  (Turing, 1937). It is not computable but can be approximated by empirical methods;

**Shannon entropy** of the macro-state  $\Gamma_{x_d}$  quantifies the residual heterogeneity i.e. the uncertainty on the microscopic configurations of the system when only the macroscopic description  $x_d$  is available.

To quantify mutualization, the physical system is replaced by the space generated by the risk factors and particles by the insureds. The  $x_d$  description of the physical system is replaced by an iterative synthesis of the dispersion of the insureds according to the risk factors sorted by importance for the risk modeling (denoted  $x_d^{freq}$ ).



**Fig. 4.** Insureds in risk factor space: (a) insureds in the entire space, (b) insureds huddled

### Example of portfolio description

$x_1^{freq} =$  "40 % of the vehicles in the portfolio are TOYOTA or FORD"

$x_2^{freq} = x_1^{freq}$  and "20 % of TOYOTA or FORD vehicles are less than 6 months old at subscription"

$x_3^{freq} = x_2^{freq}$  and ...

A description  $x_d^{freq}$  is obtained using a CART regression tree calibrated by regressing the individualized risk profiles on all of risk factors available with the *market distribution*. It results in a sequence of couples (variable threshold) that iteratively partition the space into nodes or sub-space. This tree, called *splitter*, encrypts the risk structure of the market.

If the OL are observed,  $x_d^{freq}$  is completed with a description of the importance of the solidarity margins raised by the insurer to cross-subsidize insureds in its portfolio (denoted  $x_d^{OL}$ ).  $x_d^{OL} := \alpha_1^{OL} | \alpha_2^{OL} | \dots | \alpha_d^{OL}$  where  $\alpha_d^{OL}$  measures the solidarity created between the policyholders of the node  $\mathcal{N}_d$  by the gain in OL stability resulting from partitioning the node  $\mathcal{N}_d$  into  $\mathcal{N}_{d,<}$  and  $\mathcal{N}_{d,>}$  computed as follows:

$$\alpha_d^{OL} := \frac{\text{Var}_{\mathcal{N}_d}(\mathbf{OL}) - \left( \frac{|\mathcal{N}_{d,<}|}{|\mathcal{N}_d|} \text{Var}_{\mathcal{N}_{d,<}}(\mathbf{OL}) + \frac{|\mathcal{N}_{d,>}|}{|\mathcal{N}_d|} \text{Var}_{\mathcal{N}_{d,>}}(\mathbf{OL}) \right)}{\text{Var}_{\mathcal{N}_0}(\mathbf{OL})}$$

where  $\text{Var}_{\mathcal{N}_d}(\mathbf{OL})$  is the volatility of OL of the policyholders of node  $\mathcal{N}_d$  and  $\text{Var}_{\mathcal{N}_0}(\mathbf{OL})$  is the variability of the OL in the whole portfolio. Normalizing the algorithmic entropy by the maximum heterogeneity of the insureds in the space, we obtain a mutualization indicator :

### MI : Generalized mutualization ratio

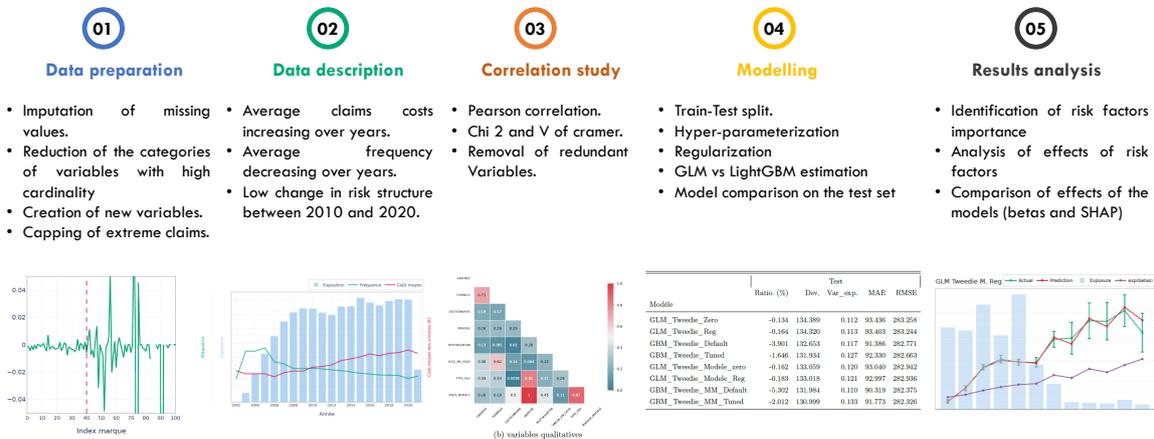
$$IM_{CK||Ent} = \frac{\min_d \{K(x_d) + H(\Gamma_{x_d})\}}{N^{Ass} \log_2(L_{0,0})}$$

où  $x_d$  : description (freq or LO) of the portfolio and  $L_{0,0}$  : number of terminal leaves of the *splitter*.

## 4 Numerical application of mutualization indicators

### 4.1 Modeling of the most individualized risk profiles

On the available data history, the 2010 - 2019 range is used to calibrate the risk profile models (figure 5) and the 2020 is used for mutualization quantification purpose. GLM and LightGBM models are implemented to predict as accurately as possible the expected loss  $BC_i^\Omega$  for each insured  $i$ .



**Fig. 5.** The GLM-based models are better at reconstructing the total loss paid ; in fact, the best GLM model predicts an overall loss with an absolute error of about 0.2 % versus 2.0 % for the best LightGBM. Inversly, the LightGBM model has better scores on the metrics build on individual accuracy such as MAE, RMSE and deviance. The models agree on the most important factors and their effects.

The model for evaluate the technical premiums  $BC^{Ass}$  is also calibrated on variables formula, age of the vehicle, type of vehicle and power, each with 3 modalities.

## 4.2 Level of mutualization of the motor extended warranty portfolio in 2020

This pricing strategy of BNP Paribas Cardif/Icare leads to a significant aggregation of individualisable risk profiles corresponding to a creation of solidarity between the insureds in the portfolio (figure 6). The mutualization indicators (table 2) corroborate this graphic observation by quantifying the mutualization induced by this strategy. The demutualization of the portfolio is quite low, since this strategy uses only 20.5 % of the information necessary to individualize the technical premiums, i.e. 5 different risk profiles on average per technical premium amount. The solidarity premium is also quite high, about

Indicator		Value
Part of structural risk ceded	$DI_{Var  BC}$	0.175
Level of individualization	$DI_{Ent  BC}$	0.205
Solidarity premium	$MI_{L1  BC}$	34.215 €
Margin profile volatility	$MI_{Var  OL}$	0.388
Individualization level of OL	$MI_{Ent  OL}$	0.782
Part of structural risk ceded generalized	$ID_{Int  \partial\mu}$	0.184

34 €.

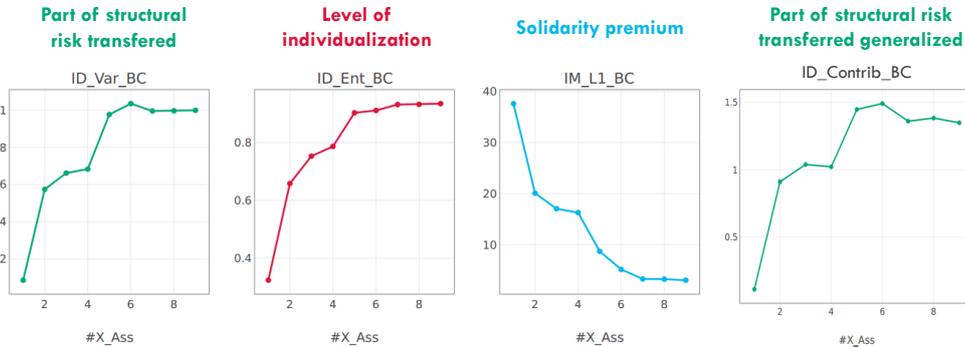


**Tab. 2.** Mutualization measures on vehicles covered in 2020. ID are quite low while MI are quite high.

**Fig. 6.** Distribution of the most individualized burning costs ( $BC^{Omega}$ ) and priced burning costs ( $BC^{Ass}$ ).

## 4.3 Mutualization indicators to support pricing strategy

Mutualization indicators can be used to drive premium segmentation when constructing the technical premium. Indeed, in addition to the usual adequacy metrics, the insurer could add a constraint on the minimum solidarity desired in its portfolio (figure 7).



**Fig. 7.** The demutualization (resp. mutualization) of the portfolio evaluated on the technical premiums with the proposed indicators increases (resp. decreases) when the pricing variables become more and more complete ( $X_{Ass} \rightarrow X_{\Omega}$ ).

## 4.4 Hyper-segmentation, mutualization and competitive marketplace

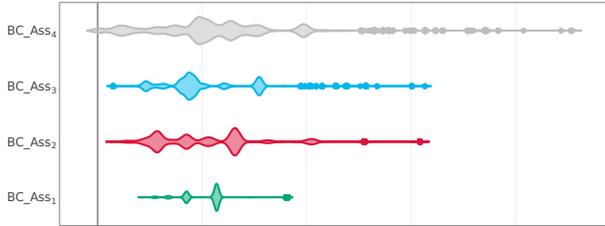
Mutualization indicators can also be used for external communication purposes, in particular for comparison between several insurers. In this way, 4 insurers whose variables used to calibrate their technical premiums models are given by the table 3 are considered.

**Tab. 3.** Pricing strategies used by insurers to construct the technical premium.

Insurer	Factor
$Ass_1$ ■	FORMULE, TYPE_VEH, AGE_VEH_SOUS
$Ass_2$ ■	FORMULE, TYPE_VEH, AGE_VEH_SOUS, PUISSANCE
$Ass_3$ ■	FORMULE, TYPE_VEH, AGE_VEH_SOUS, 'TRANCHE_KM_SOUS'
$Ass_4$ ■	FORMULE, TYPE_VEH, AGE_VEH_SOUS, MARQUE

#### 4.4.1 Comparing the mutualization levels of monopolistic insurers

On a monopolistic market, each of these insurers would offer the technical premiums whose distributions are given in figure 8.  $Ass_4$  by using the MARQUE variable, segments its risk very precisely, resulting in very spread out premiums, unlike  $Ass_1$ , which uses very little information in its technical premium model. The evaluations of the mutualization levels corresponding to these strategies are given in the table 4. The most demutualized insurer is by far,  $Ass_4$  with an individualization level of 36 % and a solidarity premium of 26 €. In contrast, the strategy of insurer 1 is very supportive, with individualization of premiums at only 13 % and a solidarity premium of 35 € per insured.



**Fig. 8.** Technical premiums offered by insurers in monopolistic markets.  $Ass_4$ ’s premiums are highly segmented, while  $Ass_1$ ’s premiums are highly aggregated.

Indicateur	$Ass_1$ ■	$Ass_2$ ■	$Ass_3$ ■	$Ass_4$ ■
DI_Ent_BC	0.131	0.199	0.216	0.360
MI_L1_BC	37.153	33.170	34.421	28.627
DI_Var_BC	0.047	0.175	0.147	0.344
MI_Var_OL	0.943	0.827	0.866	0.634
MI_Ent_OL	0.908	0.909	0.908	0.906
MI_L1_BC_Shift	35.065	29.703	31.473	26.101
Premium	4659776	4529484	4603268	5050225
S/P	0.975	1.003	0.987	0.899

**Tab. 4.** Simple mutualization indices of several insurers in a monopolistic markets. In order of demutualization:  $Ass_4$  ■ >  $Ass_2$  ■  $\succeq$   $Ass_3$  ■ >  $Ass_1$  ■

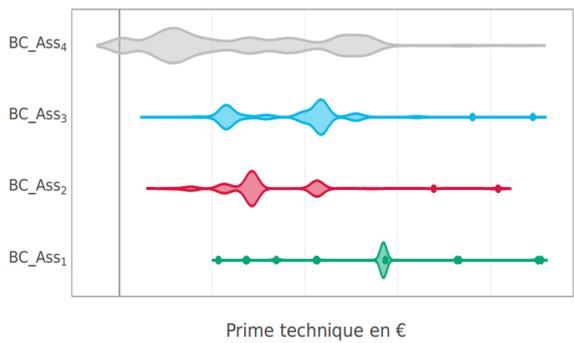
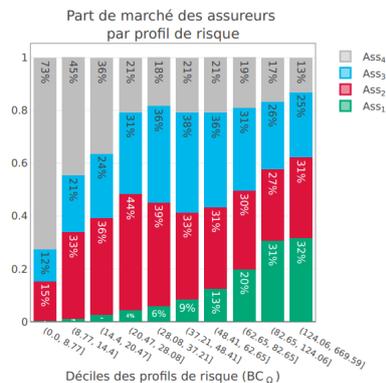
The S/P ratios show that these insurers have positive underwriting results in a monopoly situation with a much more favorable result for the most segmented insurer. This result is due to the improvement in the claims experience (including a Covid-19 effect). The out-performance of  $Ass_4$  is due to the increase in the weight of bad risks in the market which also benefit from a greater improvement in claims. Thus, demutualization allows the insurer to better control its risk and to hedge against a deviation from the market risk structure.

#### 4.4.2 Comparing the mutualization levels of insurers on a competitive market

The previous comparison is carried out in monopolistic markets and does not take into account the influence of the pricing strategy on the profiles subscribed by each insurer. The four previous insurers are placed in a competitive market in which each insured has perfect knowledge of all the contracts offered and subscribes to the cheapest policy. On the one hand, the introduction of competition leads to selection by price; by underwriting only the profiles for which it is competitive, each insurer will recover a more or less heterogeneous fraction of the policyholders. On the other hand, it induces a pooling effort on more restricted risk groups. This leads to a global demutualization of the market (table 5). At the insurer level, mutualization also decreases compared to the monopolistic situation, in particular the solidarity premium (corrected for the absolute value of the expected technical result per policy) decreases by about 22 € per insurer.

**Tab. 5.** Simple mutualization indices of several insurers in a competitive market.

Indicator	$Ass_1$ ■	$Ass_2$ ■	$Ass_3$ ■	$Ass_4$ ■	Marché
DI_Ent_BC	0.086	0.147	0.171	0.384	0.304
MI_L1_BC	52.769	28.484	30.457	16.337	28.469
DI_Var_BC	0.030	0.047	0.043	0.195	0.087
MI_Var_OL	0.922	0.747	0.849	0.594	0.724
MI_L1_BC_Shift	7.926	7.723	10.896	8.148	8.759
$\Delta$ IM_L1_BC_Shift monopole	-27.14	-21.98	-20.58	-17.95	-
Market share (%)	11.780	31.880	28.100	28.241	100.000
Premium	587156	893783	911934	666104	3058979
S/P	1.775	1.345	1.493	1.404	1.485



(a) Insurers' market shares according to individualized risk profiles in a competitive situation.

(b) Premiums collected by insurers in a competitive market.

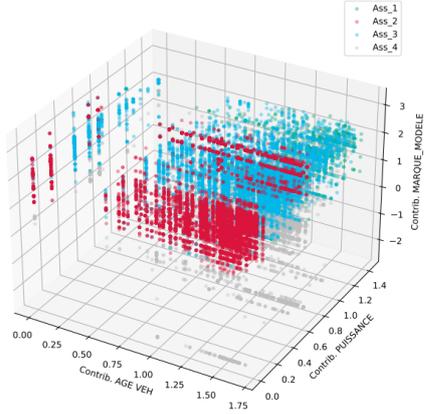
**Fig. 9.** Ass 4 ■ is positioned mostly on very good risks for which it charges very low premiums, Ass 1 ■ is a victim of anti-selection, it underwrites bad risks at under-valued premiums, Ass 2 ■ and Ass 3 ■ are present on all risks.

This competition allows us to empirically highlight the effects of hyper-segmentation on insurers' results and on the market:

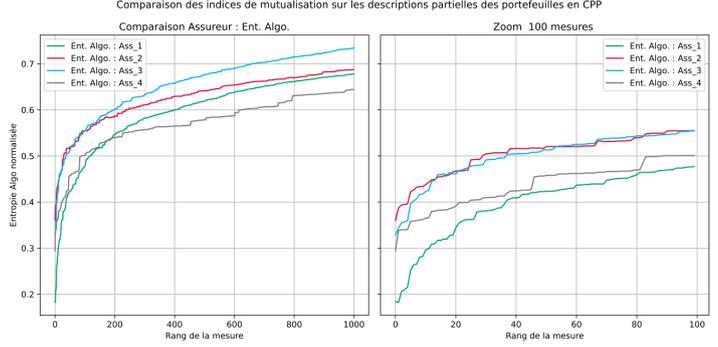
<p>Price segmentation leads to better expected results</p>	<p>Hyper-segmentation is not a guarantee of financial performance</p>	<p>Hyper-segmentation also has other drawbacks (exclusions)</p>	<p>The hyper-segmentation spiral can be detrimental to both the market and the policyholders</p>
<p>Segmentation is a protection against structural risk</p>	<p>Higher segmentation can also lead to worse results in some situations</p>	<p>Higher segmentation can lead to exclusions</p>	<p>Hyper-segmentation drives the market towards demutualization</p>
<ul style="list-style-type: none"> <li>Offering individualized premium allows the insurer to adjust the overall premium level to expected losses while hedging against moral hazard .</li> <li>Higher segmentation allows for a better understanding risks and the bring up ability to capture the best risks in the market.</li> </ul>	<ul style="list-style-type: none"> <li>In particular, high segmentation leads to lower premiums on good risks that make it more difficult to mutualize with other risks, especially when these risks are volatile.</li> </ul>	<ul style="list-style-type: none"> <li>Lowering the price on good risks mechanically increases the price of bad risks, making some categories "uninsurable".</li> <li>Financially, it also means that a hyper-segmented insurer may "miss" opportunities to improve risks on "bad risks".</li> </ul>	<ul style="list-style-type: none"> <li>Higher segmentation pushes the entire market toward more demutualization, making insurance less accessible for certain risks.</li> <li>The entry of a new player with finer segmentation destroys more value for the market than any other competition.</li> </ul>
<p>In a competitive environment, segmentation is a fundamental requirement.</p>	<p>There is no justification to over segment too strongly compared to competitors</p>	<p>Segmenting decisions have to be taken looking at the competition and at potential exclusions.</p>	<p>The value proposal of a new insurer like should not be reduced to more segmented premiums only.</p>

#### 4.5 Exotic indicators reveal the impact of factors on demutualization

Unlike simple indicators, exotic indicators provide representations that allow us to better understand the mutualization and the structure of risks in the portfolio. Figure 10 specifies the contribution of exotic indicators to describe the mutualization of the 4 previous insurers when they are in competition.



(a) Positioning of policyholders in the space of factor contributions to the individualized risk profile w.r.t insurer. The policyholders of Ass 1 ■ have high contributions to the risk according to the power (factor PUISSANCE) and are also risky brands (factor MARQUE). Ass 4 ■ is positioned on brands with lower claims experience. Ass 2 ■ is positioned on all the risks by brand, but makes a selection according to the power. The vehicles of Ass 3 ■ have low or high contributions according to the brand, but have more risky power classes than for Ass 2 ■.



(b) Evolution of GMR - Frequency by insurer by description. The portfolios of Ass 2 ■ and Ass 3 ■ cover the market structure better than those of Ass 1 ■ and Ass 4 ■. Ass 1 ■ is positioned on very specific risks since from the first measurements, the GMR - Freq is very low (and therefore the residual mutualization too); Ass 2 ■ and Ass 3 ■ are positioned on several different profiles and capture a large number of insureds on each of these risk segments ( $K(x_d)$  and  $H(\Gamma_{x_d})$  are both high). Ass 4 ■ also positions itself on several different risk segments ( $H(\Gamma_{x_d})$  high), but only underwrites limited number of policyholders on each of these segments ( $K(x_d)$  low).

**Fig. 10.** Comparison and analysis of mutualization between insurers on a competitive market based on exotic indicators.

## 5 Conclusion, limits and perspectives

Measurement of mutualization raises expectations to which the proposed indicators are quite favorable. The solidarity premium and the level of individualization are our main proposals for evaluating the mutualization of a portfolio: they are complementary and relatively easy to calculate and explain. The GMR is quite relevant when it comes to studying how interactions between insurers modify the structure of the portfolio in a competitive market.

The main limitations of the indices are that they do not take directly into account the actuarial fairness in their formulation, and that the indicators are dependent on how actuaries decide to estimate the individualised risk profile that should be the same to ensure indicator comparison. Indeed, the indicators must be complemented with other measures of value to make sure that the solidarity created is socially acceptable.

This study is conducted in order to encourage insurers to evaluate their level of mutualization on various products such as borrowers' insurance. This will eventually allow to better delimit the fears linked to demutualization. In this context, the Python implementation to calculate indicators is available on <https://github.com/B-Noumedem/Indices-de-mutualisation.git>. Further improvements to this study include : analysis of the influence of individual risk profile modeling on the indicators ; direct integration of actuarial fairness in indicators definition ; deeper study of the links between mutualization and inclusion to insurance.

## BIBLIOGRAPHIE

- George A. AKERLOF : The Market for “Lemons” : Quality Uncertainty and the Market Mechanism\*. *The Quarterly Journal of Economics*, 84(3):488–500, août 1970. ISSN 0033-5533. URL <https://doi.org/10.2307/1879431>.
- Takuya AKIBA, Shotaro SANO, Toshihiko YANASE, Takeru OHTA et Masanori KOYAMA : Optuna : A Next-generation Hyperparameter Optimization Framework. *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, pages 2623–2631, Anchorage, AK, USA, juillet 2019. Association for Computing Machinery. ISBN 9781450362016. URL <https://doi.org/10.1145/3292500.3330701>.
- Faruk ARSLAN : Social Physics : How Good Ideas Spread-The Lessons from a New Science, by Alex Pentland. *Journal of Information Privacy and Security*, 12(4):218–220, octobre 2016. ISSN 1553-6548. URL <https://doi.org/10.1080/15536548.2016.1243849>.
- Tom BAKER et Jonathan SIMON : *Embracing Risk : The Changing Culture of Insurance and Responsibility*. University of Chicago Press, 2001. ISBN 9780226035192 9780226035185 9780226035178. URL <http://www.bibliovault.org/BV.landing.ep1?ISBN=9780226035192>.
- Laurence BARRY et Arthur CHARPENTIER : Personalization as a promise : Can Big Data change the practice of insurance? *Big Data & Society*, 7(1):205395172093514, janvier 2020. ISSN 2053-9517, 2053-9517. URL <http://journals.sagepub.com/doi/10.1177/2053951720935143>.
- Alberto CEVOLINI et Elena ESPOSITO : From pool to profile : Social consequences of algorithmic prediction in insurance. *Big Data & Society*, 7(2):205395172093922, juillet 2020. ISSN 2053-9517, 2053-9517. URL <http://journals.sagepub.com/doi/10.1177/2053951720939228>.
- Gregory J. CHAITIN : On the Length of Programs for Computing Finite Binary Sequences. *Journal of the ACM*, 13(4):547–569, octobre 1966. ISSN 0004-5411. URL <https://doi.org/10.1145/321356.321363>.
- Arthur CHARPENTIER : La loi des grands nombres et le théorème central limite comme base de l’assurabilité? *Risques*, (86):111–116, juin 2011. ISSN 978-2-3588-038-4. URL [https://freakonometrics.hypotheses.org/files/2020/12/Charpentier\\_86.pdf](https://freakonometrics.hypotheses.org/files/2020/12/Charpentier_86.pdf).
- Arthur CHARPENTIER, Michel DENUIT et Romuald ELIE : Segmentation et mutualisation les deux faces d’une même pièce? *Risques*, (103):57–61, septembre 2015. ISSN 978-2-35588-068. URL <https://www.revue-risques.fr/2015/09/risques-n-103/>.

- 
- Tianqi CHEN et Carlos GUESTRIN : XGBoost : A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, août 2016. ACM. ISBN 9781450342322. URL <https://dl.acm.org/doi/10.1145/2939672.2939785>.
- Olivier DARRIGOL : *Lectures on Gas Theory (1896, 1898)*, volume 1. Oxford University Press, mars 2018. URL <https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198816171.001.0001/oso-9780198816171-chapter-9>.
- La Fabrique D’ASSURANCE : Intelligence artificielle et éthique dans le secteur de l’assurance. Rapport technique, Livre blanc 2019, décembre 2019.
- Dimitri DELCAILLAU : Contrôle et Transparence des modèles complexes en actuariat, Mémoire d’actuariat EURIA, 2019. URL <https://www.institutdesactuaires.com/docs/mem/af306e55a66c5ce6bcf6696cd3df9e0a.pdf>.
- Michel DENUIT et Arthur CHARPENTIER : *Mathématiques de l’assurance non-vie. T. 1 : Principes fondamentaux de théorie du risque*. Collection économie et statistiques avancées. Economica, Paris, 2004. ISBN 9782717848540.
- Alain DESROSIERES : *L’argument statistique*. Sciences sociales. Presses de l’école des mines, 2008. ISBN 9782356710024 9782356710055.
- Sean DEVINE : The Insights of Algorithmic Entropy. *Entropy*, 11(1):85–110, mars 2009. URL <https://www.mdpi.com/1099-4300/11/1/85>.
- Sean DEVINE : *Algorithmic information theory for physicists and natural scientists*. IOP Publishing, 2020. ISBN 9780750326407 9780750326384.
- Sander DEVRIENDT, Katrien ANTONIO, Tom REYNKENS et Roel VERBELEN : Sparse regression with Multi-type Regularized Feature modeling. *Insurance : Mathematics and Economics*, 96:248–261, janvier 2021. ISSN 01676687. URL <https://linkinghub.elsevier.com/retrieve/pii/S0167668720301608>.
- Tomasz DOWNAROWICZ : Entropy. *Scholarpedia*, 2(11):3901, 2007. ISSN 1941-6016. URL <http://www.scholarpedia.org/article/Entropy>.
- Christophe DUTANG : Actuariat de l’Assurance Non-Vie, Notes de cours ENSAE IP Paris M2, 2020.
- FCA : General insurance pricing practices. Market Study MS18/1.3, décembre 2020. URL <https://www.fca.org.uk/publication/market-studies/ms18-1-3.pdf>.
- Richard P. FEYNMAN : The Feynman Lectures on Physics, mainly mechanics, radiation, and heat. URL [https://www.feynmanlectures.caltech.edu/I\\_toc.html](https://www.feynmanlectures.caltech.edu/I_toc.html). 1963.
-

- 
- Richard P. FEYNMAN : *Feynman lectures on computation*. Advanced book program. Perseus, paperback print., [repr.] édition, 2001. ISBN 9780738202969.
- Sylvestre FREZAL et Laurence BARRY : Fairness in Uncertainty : Some Limits and Misinterpretations of Actuarial Fairness. *Journal of Business Ethics*, 167(1):127–136, novembre 2020. ISSN 0167-4544, 1573-0697. URL <http://link.springer.com/10.1007/s10551-019-04171-2>.
- P. GACS : The Boltzmann entropy and randomness tests. *In Proceedings Workshop on Physics and Computation. PhysComp '94*, pages 209–216. IEEE Comput. Soc. Press, 1994. ISBN 9780818667152. URL <http://ieeexplore.ieee.org/document/363679/>.
- Ian GOODFELLOW, Yoshua BENGIO et Aaron COURVILLE : *Deep learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, 2016. ISBN 9780262035613.
- François-Xavier HAY : La mutualisation est-elle soluble dans le big data? *Risques*, (103):57–61, septembre 2015. ISSN 978-2-35588-068. URL <https://www.revue-risques.fr/2015/09/risques-n-103/>.
- E. T. JAYNES : Gibbs vs Boltzmann Entropies. *American Journal of Physics*, 33(5):391–398, mai 1965. ISSN 0002-9505. URL <https://aapt.scitation.org/doi/10.1119/1.1971557>.
- Guolin KE, Qi MENG, Thomas FINLEY, Taifeng WANG, Wei CHEN, Weidong MA, Qiwei YE et Tie-Yan LIU : LightGBM : A Highly Efficient Gradient Boosting Decision Tree. *In Advances in Neural Information Processing Systems*, pages 3146–3154. Curran Associates, Inc., 2017.
- Benno KELLER, Eling MARTIN, Schmeiser HATO, Christen MARKUS et Loi MICHELE : Big Data and Insurance : Implications for Innovation, Competition and Privacy, mars 2018.
- A. N. KOLMOGOROV : Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1-4):157–168, janvier 1968. ISSN 0020-7160. URL <https://doi.org/10.1080/00207166808803030>.
- Edouard LEDERER : Les nouvelles technologies mettent sous pression le modèle économique des assureurs. *Les Echos*, juillet 2019. URL [https://www.lesechos.fr/finance-marches/banque-assurances/les-nouvelles-\\*-1037468](https://www.lesechos.fr/finance-marches/banque-assurances/les-nouvelles-*-1037468).
- Turo-Kimmo LEHTONEN et Jyri LIUKKO : The Forms and Limits of Insurance Solidarity. *Journal of Business Ethics*, 103(S1):33–44, avril 2011. ISSN 0167-4544, 1573-0697. URL <http://link.springer.com/10.1007/s10551-012-1221-x>.
- Turo-Kimmo LEHTONEN et Jyri LIUKKO : Producing Solidarity, Inequality and Exclusion Through Insurance. *Res Publica*, 21(2):155–169, mai 2015. ISSN 1572-8692. URL <https://doi.org/10.1007/s11158-015-9270-5>.
-

- 
- Leonid Anatolevich LEVIN : Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problemy Peredachi Informat-sii*, 10(3):30–35, 1974. URL [http://alexander.shen.free.fr/library/Levin74\\_LawsOfInformationConservation.pdf](http://alexander.shen.free.fr/library/Levin74_LawsOfInformationConservation.pdf).
- Ming LI et Paul VITÁNYI : *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer International Publishing, Cham, 2019. ISBN 9783030112974 9783030112981. URL <http://link.springer.com/10.1007/978-3-030-11298-1>.
- Scott LUNDBERG et Su-In LEE : A Unified Approach to Interpreting Model Predictions. *arXiv :1705.07874 [cs, stat]*, novembre 2017. URL <http://arxiv.org/abs/1705.07874>. arXiv : 1705.07874.
- Scott M. LUNDBERG, Gabriel G. ERION et Su-In LEE : Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv :1802.03888 [cs, stat]*, mars 2019. URL <http://arxiv.org/abs/1802.03888>. arXiv : 1802.03888.
- Liz MCFALL : Personalizing solidarity? The role of self-tracking in health insurance pricing. *Economy and Society*, 48(1):52–76, janvier 2019. ISSN 0308-5147. URL <https://doi.org/10.1080/03085147.2019.1570707>.
- Liz MCFALL et Liz MOOR : Who, or what, is insurtech personalizing? : persons, prices and the historical classifications of risk. *Distinktion : Journal of Social Theory*, 19(2):193–213, mai 2018. ISSN 1600-910X, 2159-9149. URL <https://www.tandfonline.com/doi/full/10.1080/1600910X.2018.1503609>.
- Johannes PAEFGEN, Thorsten STAAKE et Frédéric THIESSE : Evaluation and aggregation of pay-as-you-drive insurance rate factors : A classification analysis approach. *Decision Support Systems*, 56:192–201, décembre 2013. ISSN 01679236. URL <https://linkinghub.elsevier.com/retrieve/pii/S0167923613001723>.
- Roger PENROSE : *À la découverte des lois de l'univers : la prodigieuse histoire des mathématiques et de la physique*. O. Jacob, 2007. ISBN 9782738118400.
- Foster PROVOST et Tom FAWCETT : *Data science for business : what you need to know about data mining and data-analytic thinking*. O'Reilly, 1. ed., 2. release édition, 2013. ISBN 9781449361327.
- Joel RATSABY : An Algorithmic Complexity Interpretation of Lin's Third Law of Information Theory. *Entropy*, 10(1):6–14, mars 2008. URL <https://www.mdpi.com/1099-4300/10/1/6>.
- Christian-Yann ROBERT : Théorie des valeurs extrêmes, Notes de cours ENSAE IP Paris M2, 2020.

- 
- Claude SHANNON : A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, juillet 1948. ISSN 0005-8580.
- Lloyd SHAPLEY : *A Value for n-Person Games*. Princeton University Press, mars 1953. ISBN 9781400881970. URL <https://www.degruyter.com/document/doi/10.1515/9781400881970-018/html>.
- Fernando SOLER-TOSCANO, Hector ZENIL, Jean-Paul DELAHAYE et Nicolas GAUVRIT : Calculating Kolmogorov Complexity from the Output Frequency Distributions of Small Turing Machines. *arXiv :1211.1302 [nlin]*, mai 2014. URL <http://arxiv.org/abs/1211.1302>. arXiv : 1211.1302.
- R. J. SOLOMONOFF : A formal theory of inductive inference. Part I. *Information and Control*, 7(1):1–22, mars 1964. ISSN 0019-9958. URL <https://www.sciencedirect.com/science/article/pii/S0019995864902232>.
- Jérôme SPERONI : Emprunteur : la commission de suivi Aeras pointe le risque de démutualisation. *L'Argus de l'assurance*, septembre 2016.
- Andreia TEIXEIRA, Armando MATOS, André SOUTO et Luís ANTUNES : Entropy Measures vs. Kolmogorov Complexity. *Entropy*, 13(3):595–611, mars 2011. URL <https://www.mdpi.com/1099-4300/13/3/595>.
- Alan TURING : On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265, 1937. ISSN 1460-244X. URL <https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-42.1.230>.
- Sybille VIÉ : Assurance emprunteur : la résiliation à tout moment, simple évolution ou vraie révolution ? *L'Argus de l'assurance*, décembre 2021.
- Hector ZENIL, Santiago HERNÁNDEZ-OROZCO, Narsis A. KIANI, Fernando SOLER-TOSCANO, Antonio RUEDA-TOICEN et Jesper TEGNÉR : A Decomposition Method for Global Evaluation of Shannon Entropy and Local Estimations of Algorithmic Complexity. *Entropy*, 20(8):605, août 2018. URL <https://www.mdpi.com/1099-4300/20/8/605>.
- W. H. ZUREK : Algorithmic randomness and physical entropy. *Physical Review A*, 40(8):4731–4751, octobre 1989. ISSN 0556-2791. URL <https://link.aps.org/doi/10.1103/PhysRevA.40.4731>.

# Annexes

---

# A Maximisation de la variance et de l'entropie

## A.1 Maximisation de la variance

**Proposition 1 :** De toutes les variables aléatoires discrètes ayant pour support réel  $\{x_1, \dots, x_n\}$ , celle qui a la plus grande variance est l'unique variable aléatoire vérifiant :  $\mathbb{P}(X = x_{(1)}) = \mathbb{P}(X = x_{(n)}) = \frac{1}{2}$ .

*Démonstration.* Soit  $n \in \mathbb{N}^*$ . Toute variable aléatoire discrète à valeur dans  $\{x_1, \dots, x_n\}$  est définie par la donnée de  $p = (p_1, \dots, p_n)$  vérifiant  $(c_1) : \sum_{k=1}^n p_k = 1$  et  $(c_2) : p_k \geq 0 \forall k \in \{1, \dots, n\}$ .

On cherche à maximiser :

$$\mathbb{V}(X) = \sum_{k=1}^n p_k x_k^2 - \left( \sum_{k=1}^n p_k x_k \right)^2 = V(p)$$

sous les contraintes  $(c_1)$  et  $(c_2)$ .

Le lagrangien de ce problème s'écrit :

$$L(p, \lambda, \mu) = V(p) + \lambda \left( \sum_{k=1}^n p_k - 1 \right) - \sum_{k=1}^n p_k \mu_k$$

Ses dérivées partielles par rapport aux  $p_k$  s'écrivent :

$$\frac{\partial L}{\partial p_k}(p, \lambda, \mu) = \left( x_k^2 - 2x_k \sum_{k=1}^n p_k x_k \right) + \lambda - \mu_k$$

Pour satisfaire les conditions de premier ordre de Kuhn-Karush-Tucker, ces dérivées doivent être nulles à l'optimum ; plus généralement ces conditions imposent que :

$$\begin{aligned} \forall k \in \{1, \dots, n\}, 2x_k \sum_{k=1}^n p_k x_k &= x_k^2 + \lambda - \mu_k \\ \forall k \in \{1, \dots, n\}, \mu_k p_k &= 0 \\ \sum_{k=1}^n p_k - 1 &= 0 \end{aligned}$$

Posons alors  $m = \sum_{k=1}^n p_k x_k$ . On a pour tout  $k$  tel que  $p_k > 0$ ,

$$x_k^2 - 2x_k m + \lambda = 0 \text{ soit encore } m^2 - (x_k - m)^2 = \lambda \quad (1)$$

Cette équation ne peut être vérifiée que pour une valeur unique de  $k$  ou exactement deux valeurs distinctes symétriques par rapport à  $m$ . Le premier cas n'est pas intéressant puisque la

variance est nulle. On se ramène alors au problème de maximisation plus simple suivant :

$$x_k^2 p_k + x_l^2 p_l - (x_k p_k + x_l p_l)^2 = x_k^2 (p_k - p_k^2) + x_l^2 (p_l - p_l^2) - 2x_k x_l p_k p_l$$

sous les contraintes  $p_k + p_l = 1$  et  $p_k, p_l \geq 0$ . En égalisant  $\lambda$  dans les 2 équations définies par 1, on peut poser sans perte de généralité  $x_l = 2m - x_k$ . On se ramène après simplification à un problème de maximisation à une seule variable :

$$\min_{k \in [0,1]} 4(x_k - m)^2 (p_k - p_k^2)$$

Lorsque  $k$  est fixé, cette fonction est maximale pour  $p_k = \frac{1}{2}$  ; à  $p_k$  fixé, il faudrait que  $|x_k - m|$  soit le plus grand possible, c'est-à-dire que  $x_k$  soit le plus éloigné de son symétrique  $x_l$  ; il vient alors que  $x_k, x_l \in \{x_{(1)}, x_{(n)}\}$ .  $\square$

## A.2 Maximisation d'entropie

**Proposition 2 :** De toutes les variables aléatoires discrètes ayant pour support  $\{x_1, \dots, x_n\}$ , celle qui a la plus grande entropie est l'unique variable aléatoire vérifiant :  $\mathbb{P}(X = x_i) = \frac{1}{n}$  pour tout  $i$ , c'est-à-dire la loi uniforme.

*Démonstration.* On cherche à maximiser

$$H(p) = - \sum_{k=1}^n p_k \log_2(p_k)$$

sous les contraintes  $(c_1)$  et  $(c_2)$ . Le lagrangien de ce problème s'écrit :

$$L(p, \lambda, \mu) = H(p) + \lambda \left( \sum_{k=1}^n p_k - 1 \right) - \sum_{k=1}^n p_k \mu_k$$

Ses dérivées partielles par rapport aux  $p_k$  s'écrivent :

$$\frac{\partial L}{\partial p_k}(p, \lambda, \mu) = -(\log_2(p_k) + 1) + \lambda - \mu_k$$

Pour tout  $k$ , on a  $p_k > 0$  sinon le logarithme serait mal défini. Les conditions de premier ordre de Kuhn-Karush-Tucker imposent alors que :  $\lambda = \log_2(p_k) + 1$ , donc les  $p_k$  sont tous égaux à  $\frac{1}{n}$  puisque les  $n$  probabilités  $p_k$  somment à 1.

De plus, la dérivée seconde du lagrangien par rapport aux  $p_k$  est strictement négative donc l'entropie est bien maximale lorsque  $p_k = \frac{1}{n}$  pour tout  $k$ .  $\square$

## B Analyses descriptives complémentaires du portefeuille d'extension de garantie automobile

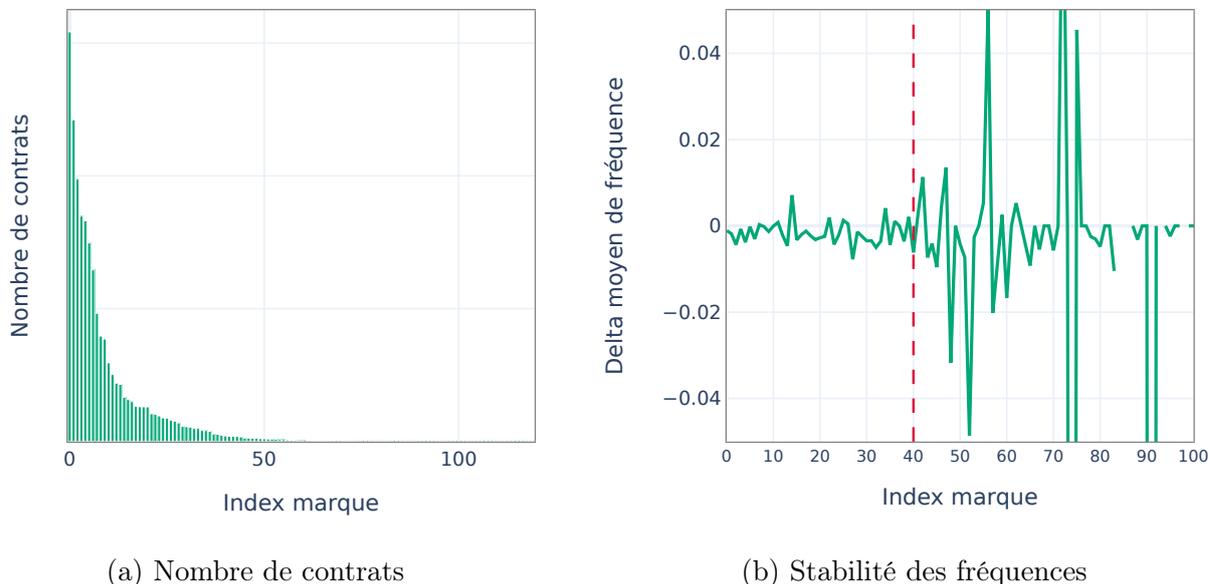


FIGURE 5 – Nombre de contrats par marque de véhicule et stabilité des fréquences de sinistres par marque



FIGURE 6 – Évolution de la distribution du portefeuille suivant la motorisation entre 2002 et 2021

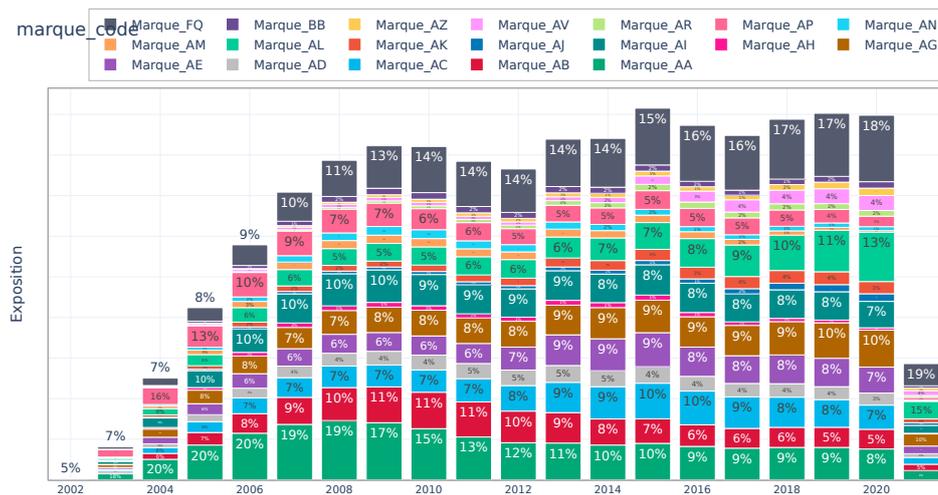


FIGURE 7 – Évolution de la distribution du portefeuille suivant la marque entre 2002 et 2021

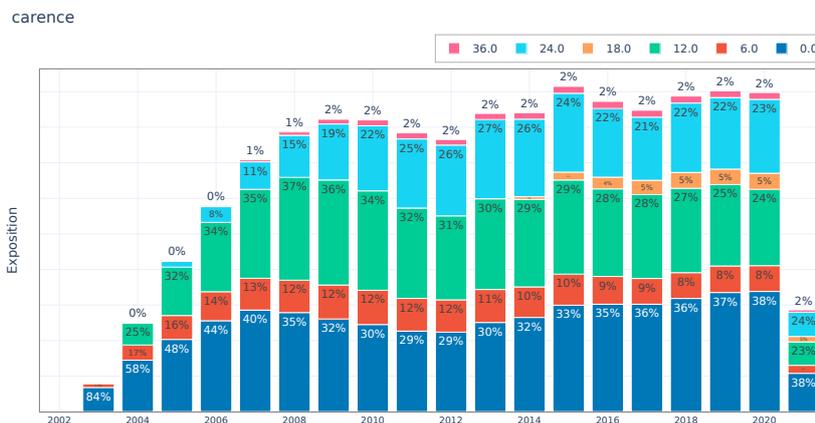


FIGURE 8 – Évolution de la distribution du portefeuille suivant la carence entre 2002 et 2021



FIGURE 9 – Évolution de la distribution du portefeuille suivant le type de véhicule entre 2002 et 2021

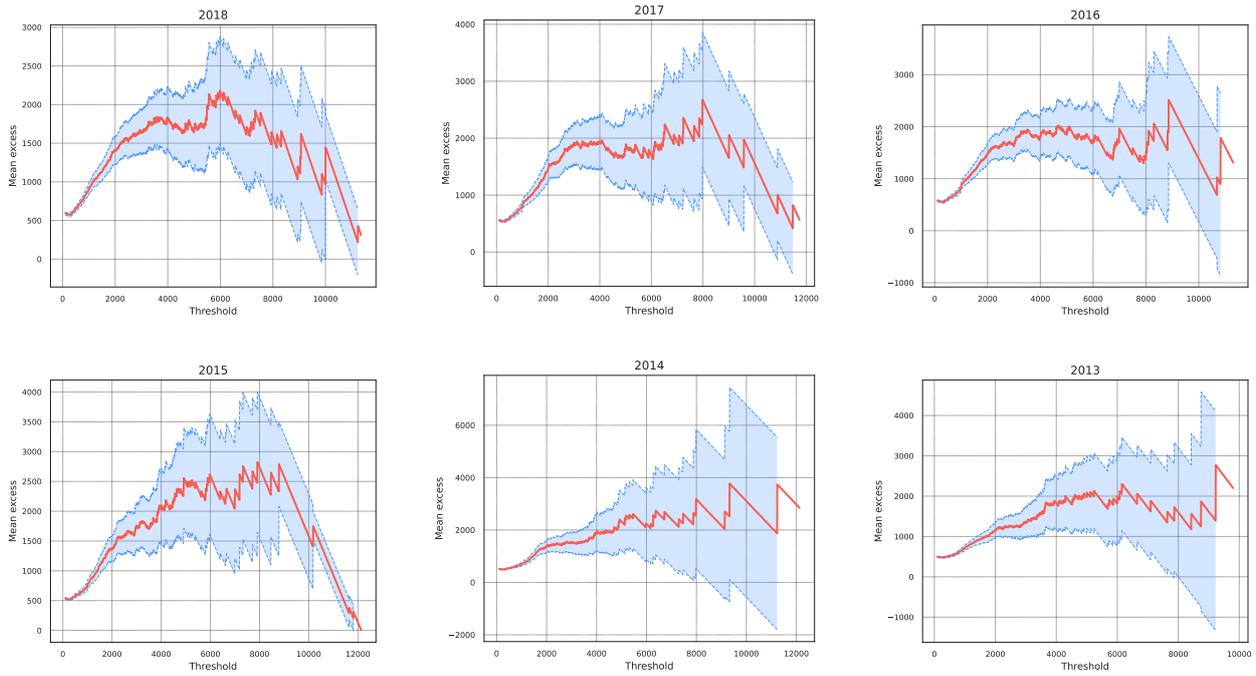
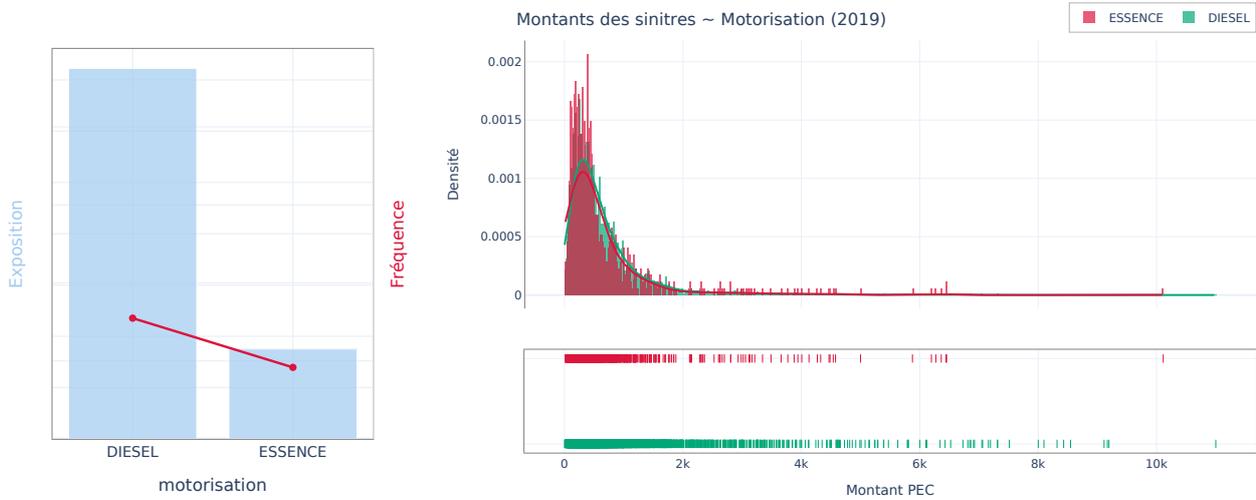


FIGURE 10 – Identification des seuils d’écèlement des sinistres extrêmes sur quelques années de l’historique

TABLEAU 4 – Seuils d’écèlement des sinistres extrêmes et ordres du quantile associés entre 2010 et 2019

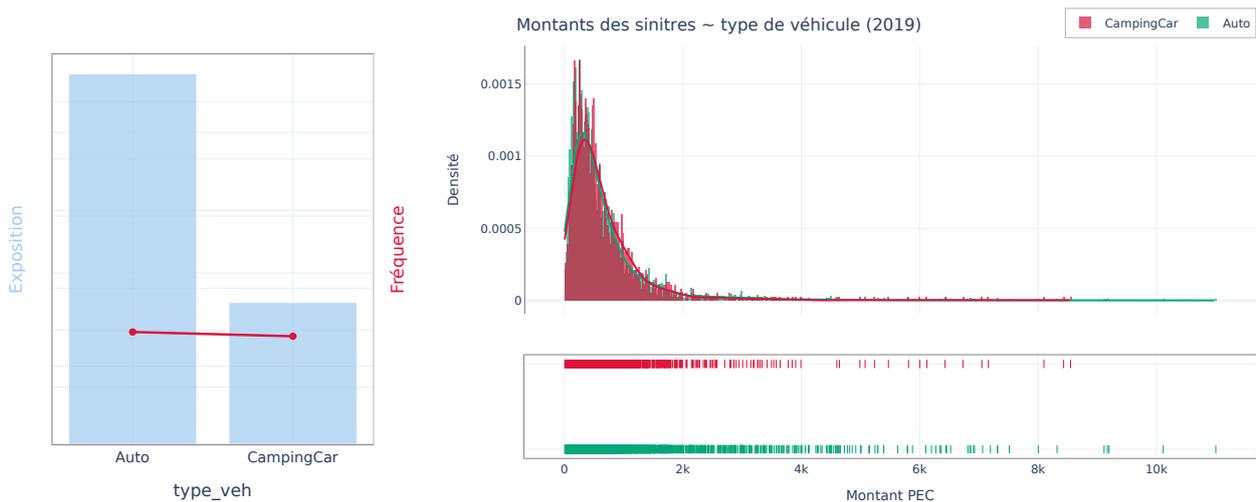
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Seuil	2500	3200	2500	2700	2600	3200	2900	3100	3400	3800
Ordre Quantile	0.986	0.995	0.989	0.987	0.984	0.989	0.981	0.986	0.985	0.987



(a) Fréquence ~ motorisation

(b) Sévérité ~ motorisation

FIGURE 11 – Fréquence et sévérité des sinistres sur le portefeuille en 2019 suivant la motorisation



(a) Fréquence ~ *type véhicule*

(b) Sévérité ~ *type véhicule*

FIGURE 12 – Fréquence et sévérité des sinistres sur le portefeuille en 2019 suivant le type de véhicule

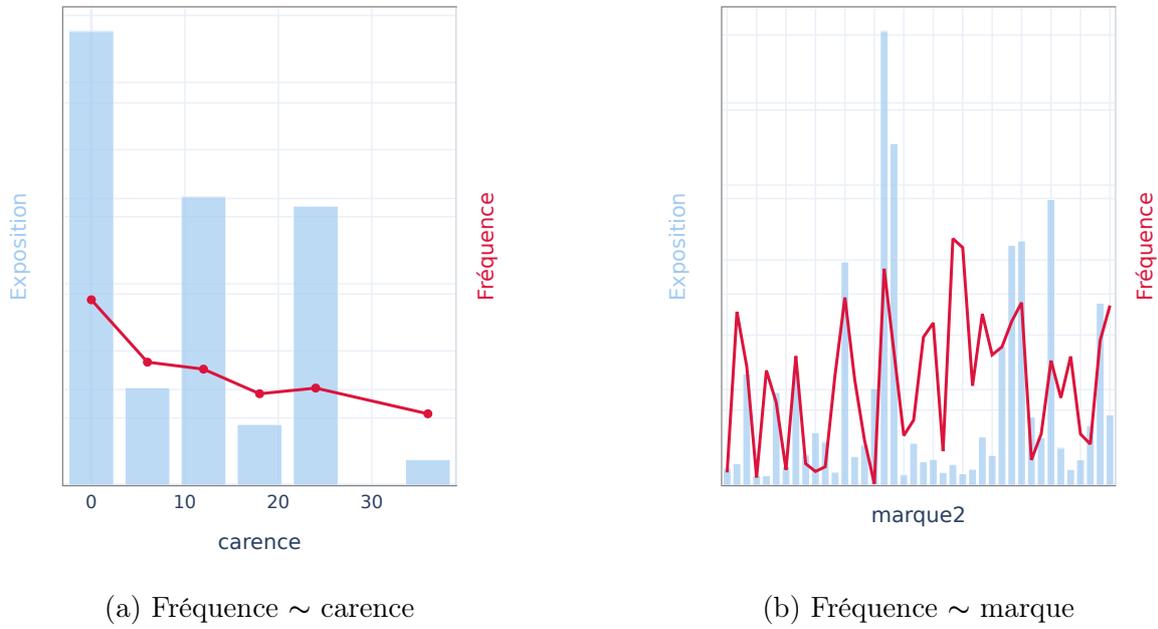


FIGURE 13 – Fréquence des sinistres sur le portefeuille en 2019 suivant la durée de carence et la marque

## C Analyses complémentaires pour la modélisation du Burning Cost à partir d'un modèle de régression

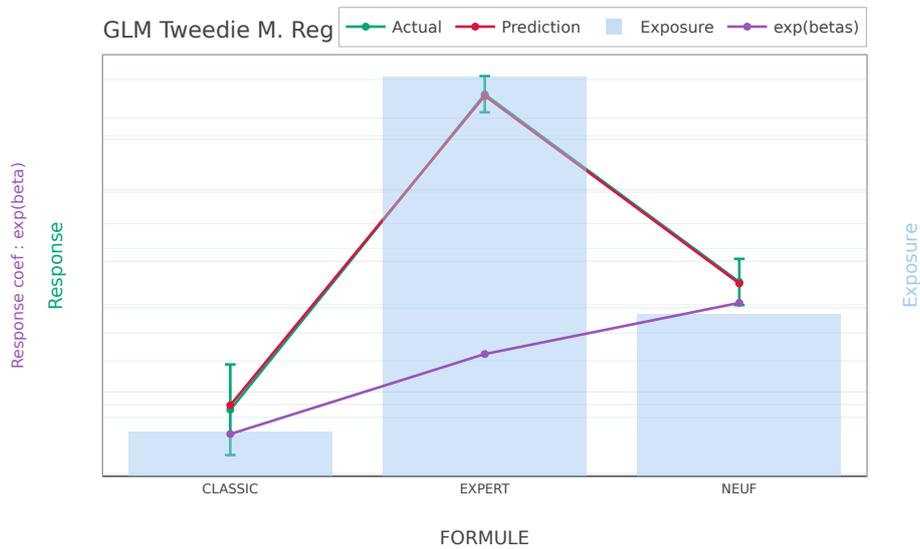


FIGURE 14 – Effet estimé par le modèle GLM de la formule souscrite sur la sinistralité

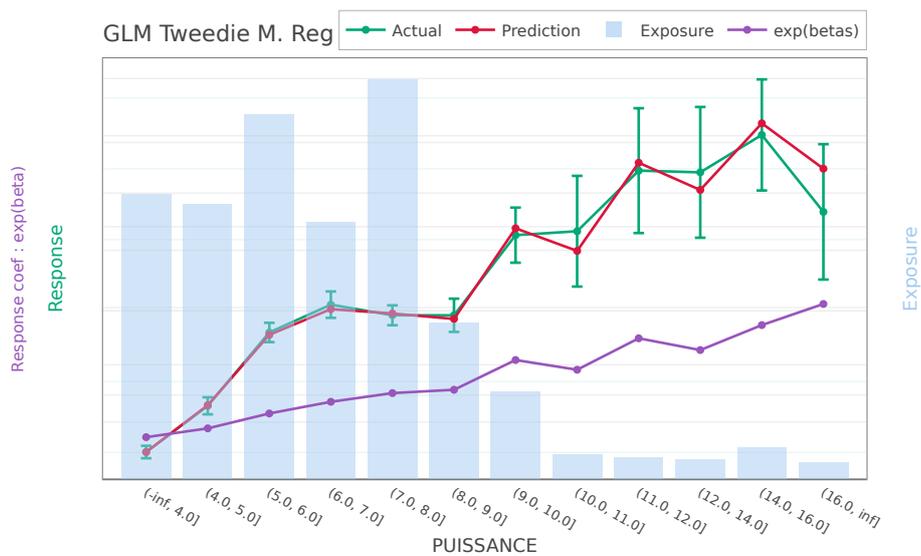


FIGURE 15 – Effet estimé par le modèle GLM de la puissance souscrite sur la sinistralité

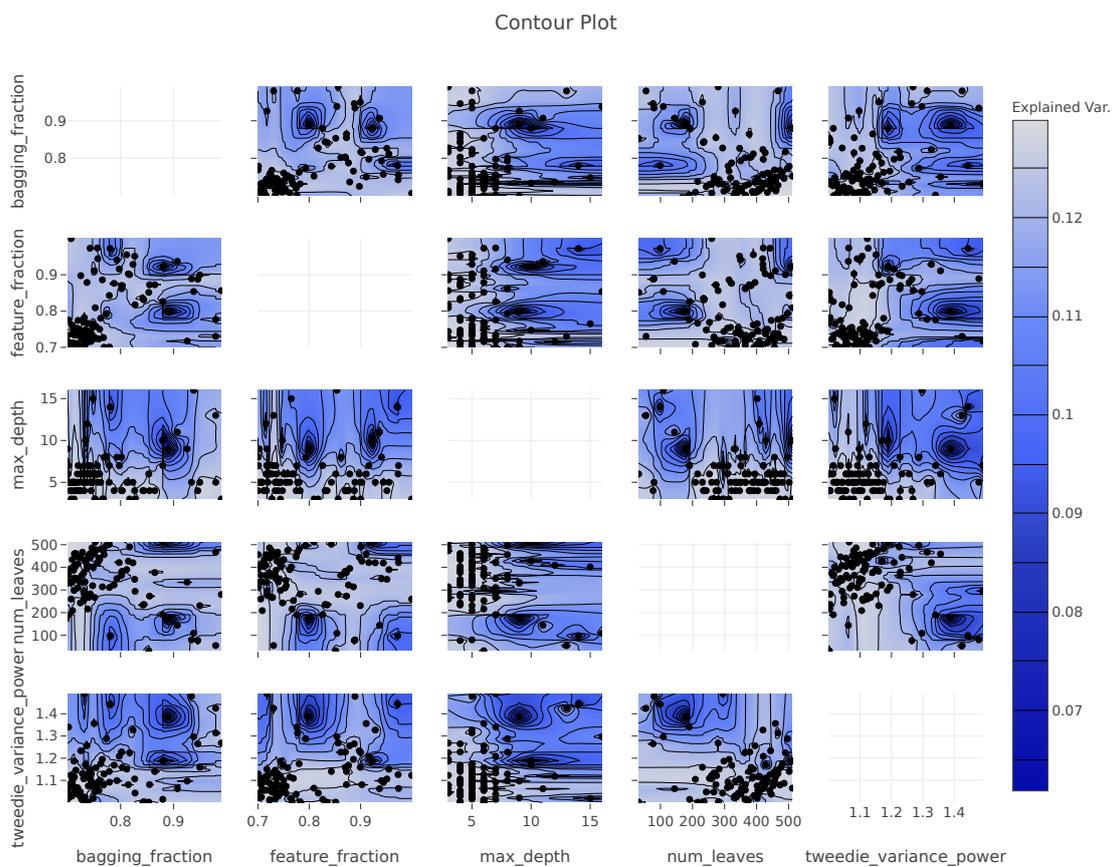


FIGURE 16 – Exploration de l'espace des hyper-paramètres par l'algorithme optuna pour la modélisation du burning cost

TABLEAU 5 – Résultats de régression du coût moyen des sinistres annuel sur l'année d'exercice

Dep. Variable :	y_cms	R-squared :	0.935			
Model :	OLS	Adj. R-squared :	0.927			
Method :	Least Squares	F-statistic :	115.7			
Date :	Mon, 22 Nov 2021	Prob (F-statistic) :	4.91e-06			
Time :	08 :09 :55	Log-Likelihood :	-43.224			
No. Observations :	10	AIC :	90.45			
Df Residuals :	8	BIC :	91.05			
Df Model :	1					
	coef	std err	t	P>  t	[0.025	0.975]
Intercept	████████	4522.359	████████	0.000	████████	████████
ANNEE	24.1502	2.245	10.758	0.000	18.973	29.327
Omnibus :	1.549	Durbin-Watson :	1.346			
Prob(Omnibus) :	0.461	Jarque-Bera (JB) :	0.923			
Skew :	0.421	Prob(JB) :	0.630			
Kurtosis :	1.774	Cond. No.	1.41e+06			

TABLEAU 6 – Résultats de régression de la fréquence annuelle sur l'année d'exercice

Dep. Variable :	y_freq	R-squared :	0.975			
Model :	OLS	Adj. R-squared :	0.972			
Method :	Least Squares	F-statistic :	308.2			
Date :	Mon, 22 Nov 2021	Prob (F-statistic) :	1.13e-07			
Time :	08 :19 :21	Log-Likelihood :	52.718			
No. Observations :	10	AIC :	-101.4			
Df Residuals :	8	BIC :	-100.8			
Df Model :	1					
	coef	std err	t	P>  t	[0.025	0.975]
Intercept	████████	0.308	████████	0.000	████████	████████
ANNEE	-0.0027	0.000	-17.555	0.000	-0.003	-0.002
Omnibus :	0.107	Durbin-Watson :	0.829			
Prob(Omnibus) :	0.948	Jarque-Bera (JB) :	0.324			
Skew :	0.056	Prob(JB) :	0.850			
Kurtosis :	2.125	Cond. No.	1.41e+06			

TABLEAU 7 – Évaluation du modèle de pricing sur une restriction des variables connues par l'assureur

-	sample	Ratio_Est_actual (%)	Deviance	Var_explained	MAE	RMSE
pred_train_BC_Ass	Train	11.102975	129.257361	-0.037979	85.466272	225.368749
pred_train_BC_Ass	Test	10.455032	129.459488	-0.039989	85.478293	226.431894

TABLEAU 8 – Évaluation des modèles de tarification des quatre assureurs étudiés, en train, test et sur le portefeuille en 2020

Assureur	sample	Ratio_Est_actual (%)	Deviance	Var_explained	MAE	RMSE
Ass_1	Train	-0.001	149.836	0.007	96.870	290.483
Ass_1	Test	-0.218	149.875	0.007	96.962	286.191
Ass_1	2020	2.604	154.255	0.015	96.460	325.005
Ass_2	Train	-0.004	145.851	0.032	96.180	289.769
Ass_2	Test	-0.184	145.900	0.032	96.283	285.489
Ass_2	2020	-0.265	151.061	0.034	94.722	324.488
Ass_3	Train	0.000	146.816	0.025	96.201	289.919
Ass_3	Test	-0.064	146.802	0.025	96.363	285.618
Ass_3	2020	1.359	152.646	0.024	95.492	324.670
Ass_4	Train	-0.005	142.106	0.064	95.375	289.339
Ass_4	Test	-0.174	142.409	0.063	95.537	285.092
Ass_4	2020	11.201	148.588	0.057	98.892	324.435

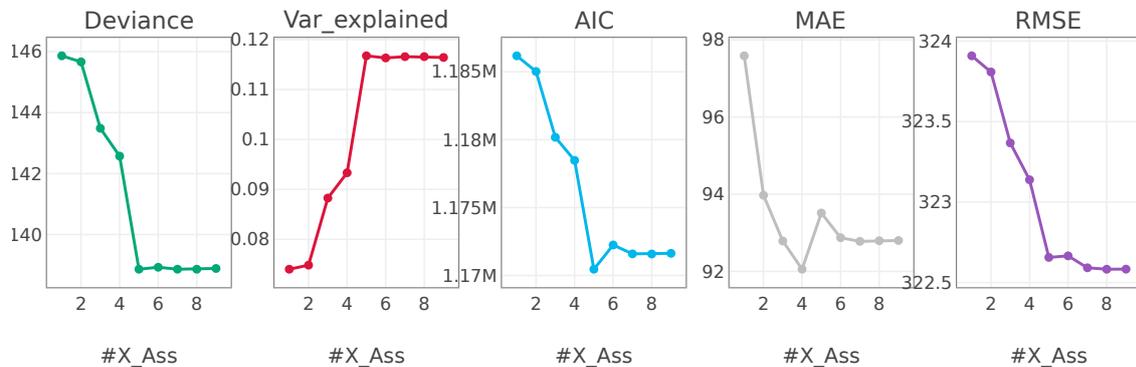


FIGURE 17 – Évolution des métriques d'adéquation lorsque :  $X_{Ass} \rightarrow X_{\Omega}$

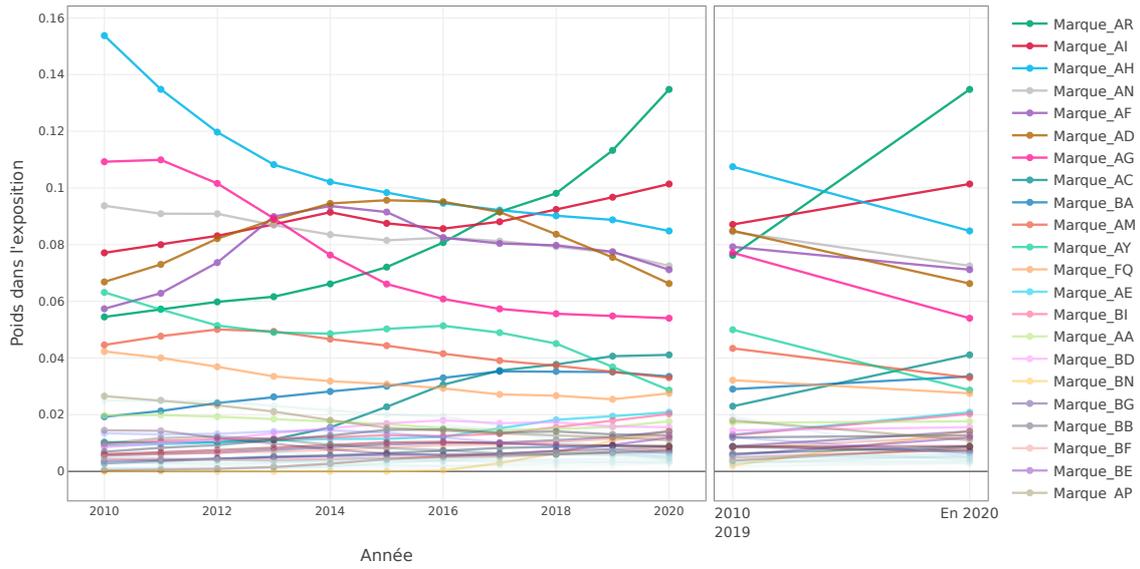


FIGURE 18 – Évolution du poids des marques dans le portefeuille par année

Montant de sinistres PEC (y1) et exposition(y2) en 2020 suivant le profil de risque

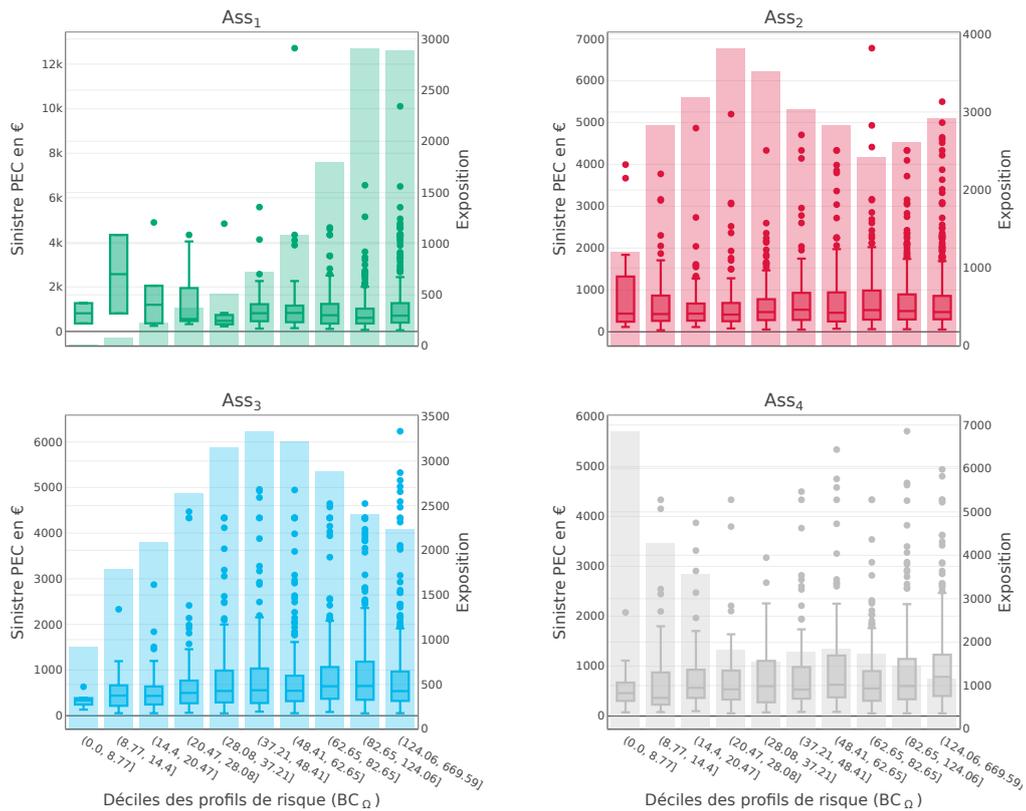


FIGURE 19 – Distributions des montants de sinistres PEC (boxplot) et exposition (barplot) suivant les profils de risques souscrits par assureur en concurrence

# D Approximation de la complexité de Kolmogorov de la description du portefeuille par la BDM et la Lossless compression

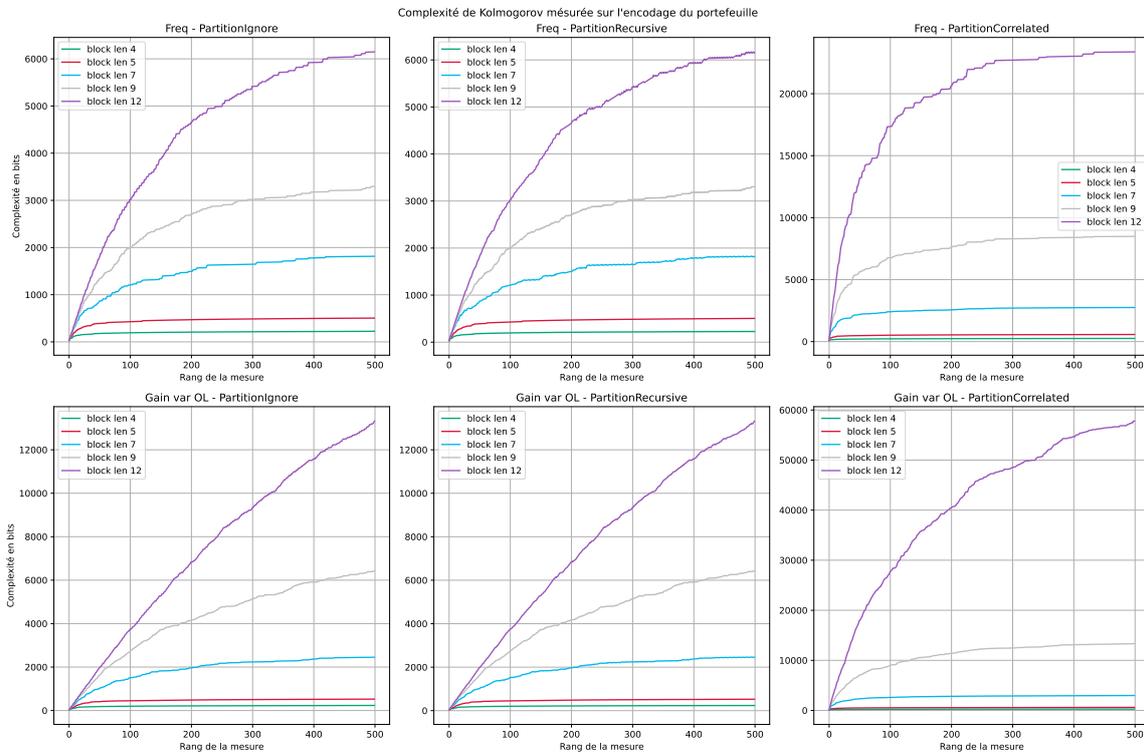


FIGURE 20 – Les valeurs estimées de la CK par le BDM en fonction des conditions de bords et de la longueur du bloc

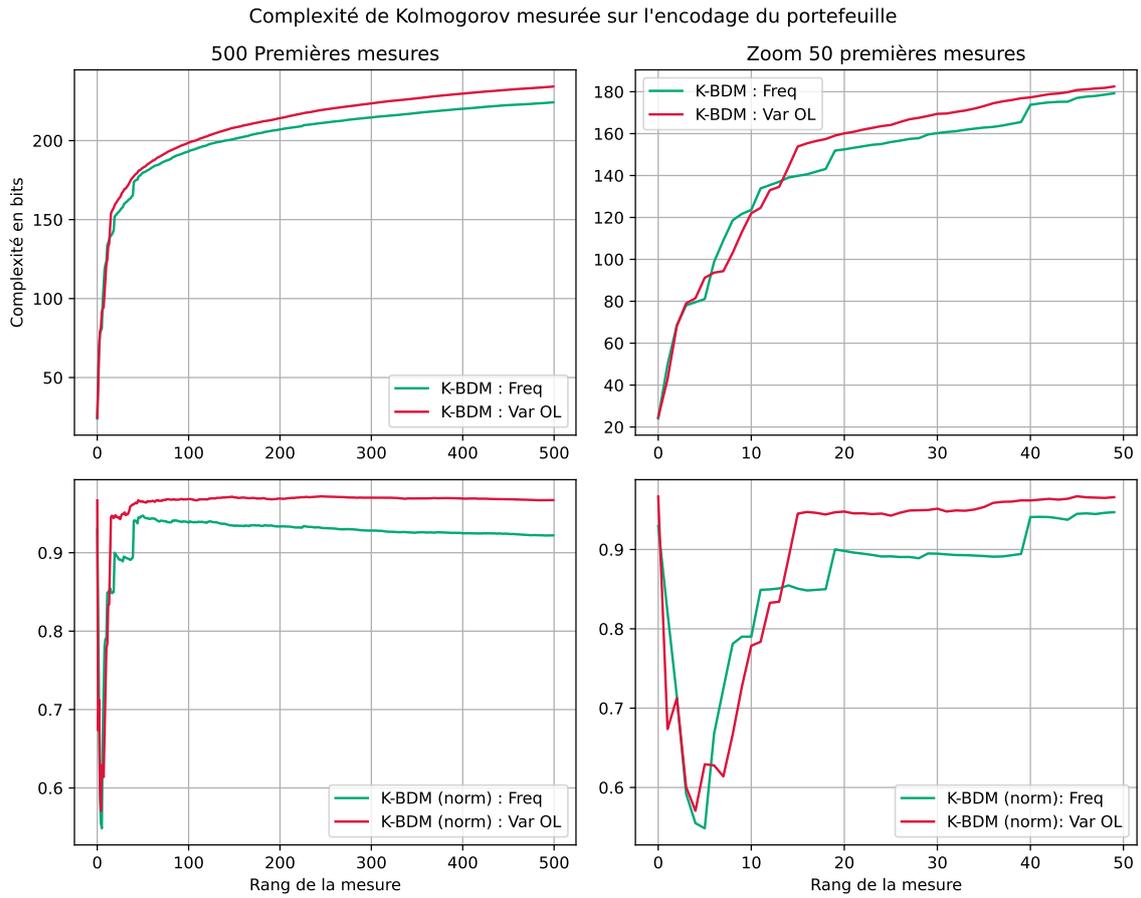
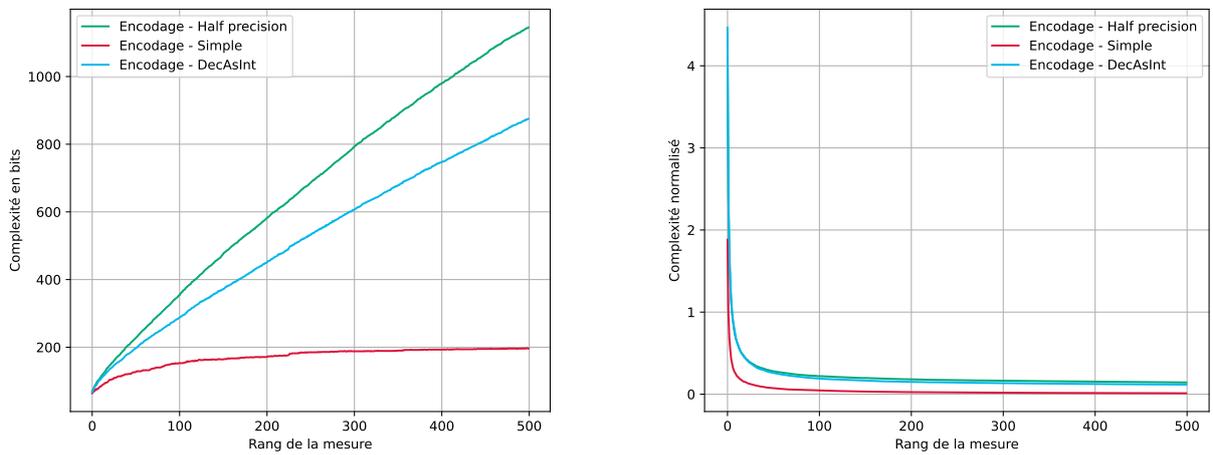


FIGURE 21 – Échec de la normalisation de la BDM et similarité entre OL et fréquence



(a) Codage binaire

(b) Échec de la normalisation

FIGURE 22 – Échec des méthodes de compression sans perte

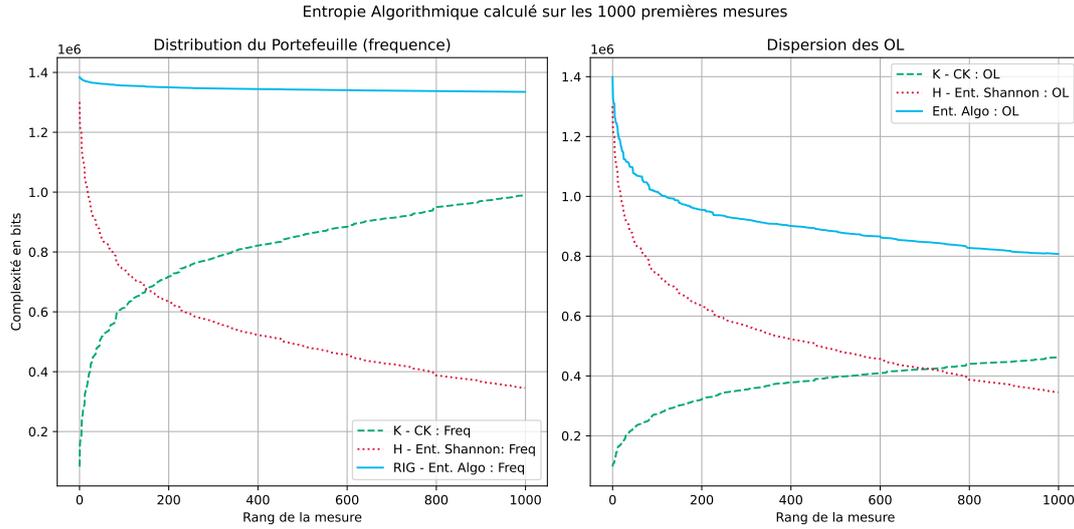


FIGURE 23 – Mesure de la démutualisation à travers le RMG sur le portefeuille de BNP Paribas Cardif (indicateur non normalisé)

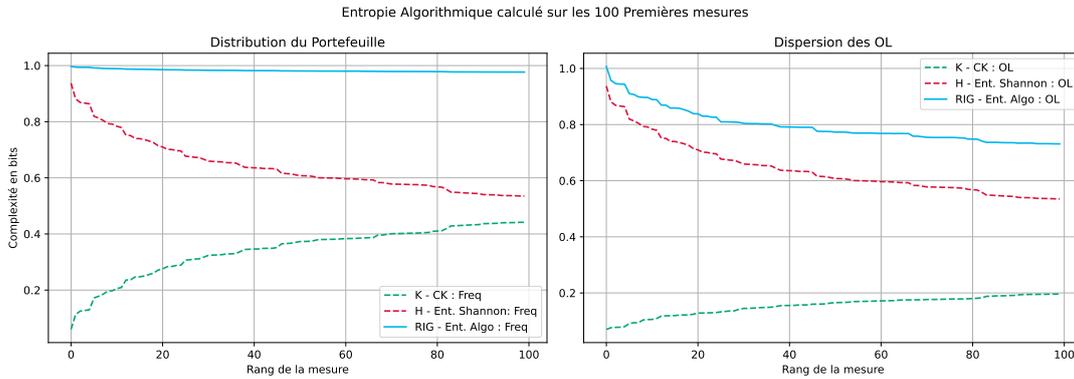


FIGURE 24 – Mesure de la mutualisation à travers le RMG sur le portefeuille de BNP Paribas Cardif - 100 premières mesures

## E Approximation de la complexité de Kolmogorov par l'entropie : robustesse à la taille de la décomposition

Soit  ${}^b\alpha_d = y_1 y_2 \dots y_n$  où chaque  $y_i$  est un bit égal à 1 avec probabilité  $\alpha_d$  et 0 avec probabilité  $\bar{\alpha}_d = 1 - \alpha_d$ . Soit  $r$  un entier plus petit que  $n$ . L'entropie  $H_r$  de la décomposition de  ${}^b\alpha_d$  en sous-chaînes de longueur  $r$  est liée à l'entropie de sa décomposition en sous chaînes de longueur 1 par :  $H_r = rH_1$

*Démonstration.* La décomposition de  ${}^b\alpha_d$  en sous chaînes de longueur  $r$  s'écrit :

$${}^b\alpha_d = \widetilde{y}_1 \widetilde{y}_2 \widetilde{y}_3 \dots \widetilde{y}_{n/r} \text{ avec } \widetilde{y}_j \in \{0, 1\}^r$$

On a  $2^r$  chaînes uniques possibles qui surviennent avec des probabilités différentes. Si  $r_1$  désigne le nombre de bits de 1 dans la sous chaîne  $\widetilde{y}_j$  alors, les occurrences des sous-chaînes suivant le

$r_1$  sont :

$$\begin{array}{l}
 r_1 = 0 : 1 \quad \text{sous-chaîne obtenue avec probabilité } (1 - \alpha_d)^r \\
 r_1 = 1 : r \quad \text{sous-chaînes, chacune obtenue avec probabilité } (1 - \alpha_d)^{r-1} \alpha_d^1 \\
 r_1 = 2 : \binom{r}{2} \quad \text{sous-chaînes, chacune obtenue avec probabilité } (1 - \alpha_d)^{r-2} \alpha_d^2 \\
 \dots \\
 r_1 = j : \binom{r}{j} \quad \text{sous-chaînes, chacune obtenue avec probabilité } (1 - \alpha_d)^{r-j} \alpha_d^j \\
 \dots \\
 r_1 = r : 1 \quad \text{sous-chaîne obtenue avec probabilité } \alpha_d^r
 \end{array}$$

Par définition,

$$\begin{aligned}
 H_r &= - \sum_{j=0}^r \binom{r}{j} (1 - \alpha_d)^{r-j} \alpha_d^j \log_2((1 - \alpha_d)^{r-j} \alpha_d^j) \\
 &= -\log_2(1 - \alpha_d) \sum_{j=0}^r (r - j) \binom{r}{j} (1 - \alpha_d)^{r-j} \alpha_d^j - \log_2(\alpha_d) \sum_{j=0}^r j \binom{r}{j} (1 - \alpha_d)^{r-j} \alpha_d^j \\
 &= -r \log_2(1 - \alpha_d) + [\log_2(1 - \alpha_d) - \log_2(\alpha_d)] \sum_{j=0}^r j \binom{r}{j} (1 - \alpha_d)^{r-j} \alpha_d^j
 \end{aligned}$$

Posons,  $f(a) = \sum_{j=0}^r \binom{r}{j} (1 - \alpha_d)^{r-j} \alpha_d^j a^j = (1 - \alpha_d + \alpha_d a)^r$ . La dérivée de cette fonction se calcule de deux façons différentes :

$$\begin{aligned}
 f'(a) &= \sum_{j=0}^r j \binom{r}{j} (1 - \alpha_d)^{r-j} \alpha_d^j a^{j-1} = r \alpha_d (1 - \alpha_d + \alpha_d a)^{r-1} \\
 f'(1) &= \sum_{j=0}^r j \binom{r}{j} (1 - \alpha_d)^{r-j} \alpha_d^j = r \alpha_d
 \end{aligned}$$

Il vient alors que :

$$H_r = -r(1 - \alpha_d) \log_2(1 - \alpha_d) - r \alpha_d \log_2(\alpha_d) = r H_1$$

□