

# Comment trouver la segmentation optimale sur un jeu de données de tarification ?

Groupe de Travail IA Impact du *machine learning*

**Ludovic POIRAUD**

Responsable du GT

Consultant Manager



**Alexandre EBY**

Co Responsable du GT

Consultant Manager



**Baptiste DIELTIENS**

Co Responsable du GT

Consultant Manager



**Ievgen SAVIN**

Participant du GT

Consultant Manager



# Sommaire

**Introduction**

**Construction d'une méthodologie de travail**

**Présentation des premiers résultats et études de sensibilités**

**Conclusion & pistes de réflexion**

# Introduction

## Présentation du groupe de travail

### Présentation du contexte

- Fort développement du Machine Learning
- Quel est l'avenir de la mutualisation ?

### Présentation du groupe de travail

- « Comment trouver la segmentation optimale sur un jeu de données de tarification ? »
- Spécificités :
  - Travaux novateurs et techniques
  - Différents domaines variés : Machine Learning, optimisation informatique, dynamique de population ...
  - Travaux très riches

# Introduction

## Objectifs

Les objectifs de ce groupe de travail sont multiples :

- Proposer une **approche novatrice de modélisation d'un marché concurrentiel**
- Proposer des **indicateurs pertinents** de comparaison de tarifs en fonction de la segmentation
- Conclure sur la **segmentation optimale** sur des exemples, et proposer des généralisations

# Construction d'une méthodologie de travail

## Résumé de la démarche

- **Question sous-jacente** : comment démontrer aux directions générales des organismes assureurs que l'hyper segmentation, indépendamment du surapprentissage, n'est pas toujours pertinente, ni bénéfique ?
- **Résumé de la démarche proposée** :
  1. **Choix des données**
  2. **Tarification à la prime pure et sélection du modèle**
  3. **Segmentation par quantile du modèle sélectionné**
  4. **Mise en concurrence « stochastique » des différentes segmentations**
  5. **Sélection des segmentations optimales au regard de différentes visions (rentabilité, part de marché)**

# Construction d'une méthodologie de travail

## Présentation des données utilisées et tarification

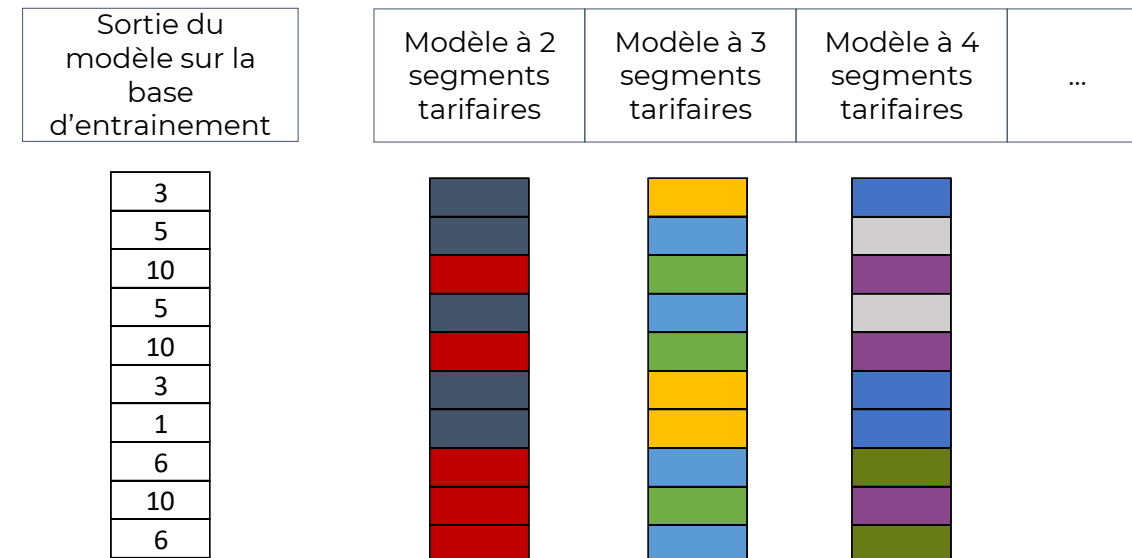
- **Le risque** : responsabilité civile automobile (à l'instar de plusieurs « pricing game »)
- **Les données** : données **MTPL2** contenant les caractéristiques et sinistres d'environ 700 000 polices d'assurance RC auto, observés sur 1 an
- **La tarification** : tarification à la **prime pure, par apprentissage statistique** ; les hyperparamètres sont déterminés par validation croisée pour éviter le surapprentissage
- **Les modèles implémentés** : modèles linéaires (GLM) coût x fréquence et « tweedie », modèles de bagging, modèles de boosting → une dizaine de modèles au total
- **Le modèle retenu** : regroupement de l'ensemble des modèles de tarification sous un **méta modèle par stacking**
- **In fine** : **1 modèle optimisé, hyper segmenté** mais entraîné pour ne pas être en surapprentissage

# Construction d'une méthodologie de travail

## Segmentation par quantile du modèle retenu

- **Principe** : regrouper les prédictions du modèle de tarification pour obtenir différentes grille de tarification à différents niveaux de segmentation
- **Objectif** : les différentes grilles tarifaires **ne se différencient que par leur niveau de segmentation**, l'effet « modèle » est ainsi gommé
- **Application** : sur la base de d'entrainement, segmentation par quantile en fonction de la valeur prédite pour obtenir un jeu de  **$n$  modèles tarifaires plus ou moins segmentés** sur la base d'un unique modèle.

*Les sorties des modèles segmentés correspondent alors aux prédictions moyennes du modèle initial, par segment.*



# Construction d'une méthodologie de travail

Modélisation d'un marché concurrentiel

**Concept général de la modélisation du marché concurrentiel simplifié :**

*Les acteurs*

- **$n$  assureurs** : avec chacun une des grilles tarifaires construites précédemment
- **$m$  individus à assurer** : les données de la base de test (n'ayant pas servi à la tarification)

*Les modélisations du comportement des acteurs*

- une **fonction de sélection** qui attribue à chaque individu l'assureur chez qui il décide de s'assurer (représente le **niveau de rationalité** de l'assuré)
- des fonctions qui définissent les **mouvements des acteurs dans une dynamique temporelle** (résiliations d'une année sur l'autre, augmentation tarifaires, ...)

*La finalité*

- connaissant la sinistralité de la population et les tarifs assureurs à appliquer : simuler les **résultats, S/P, part de marché (PdM) et autres indicateurs** annuels de chaque assureurs selon leur population sous risque



# Construction d'une méthodologie de travail

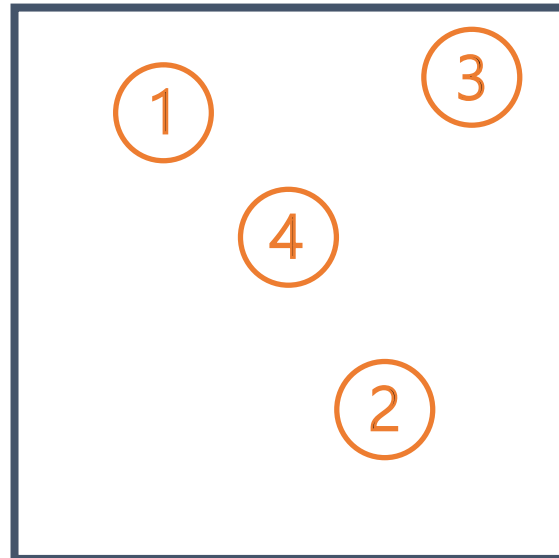
Mise en concurrence « stochastique » - Exemple simple

Considérons 4 assureurs avec 4 segmentations tarifaires différentes : ① ② ③ ④

## Modélisation d'un marché concurrentiel unique

- Calcul d'indicateurs (résultat, PdM, S/P) associés à chaque assureur

- **Inconvénient : aucune information sur la distribution de ces indicateurs (outre la distribution temporelle, limitée)**

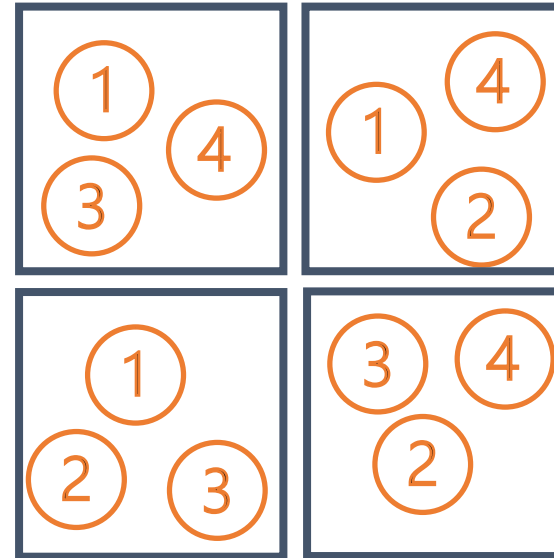


## Simulation de multiples « sous marchés » concurrentiels

- Par dénombrement, il existe 4 scénarios mettant en concurrence 3 assureur ; chaque assureur intervient dans 3 scénarios

- Calcul des indicateurs associés à chaque assureur par scénario

- **Avantage : information sur la distribution des indicateurs par assureur (moyenne et volatilité notamment)**



# Construction d'une méthodologie de travail

Mise en concurrence « stochastique »

- A plus grande échelle, en appliquant cette méthodologie et en considérant  **$n$  organismes assureurs à comparer** et des **sous marchés de  $p$  assureurs en concurrence** :
  - il existe  $\binom{n}{p}$  scénarios de sous marchés (soit 16M avec  $n = 50$  et  $p = 6$ , par exemple)
  - et chaque assureur apparait dans  $\binom{n-1}{p-1}$  scénarios (soit 2M avec  $n = 50$  et  $p = 6$ )
- **Choix de la segmentation optimale** : peut être fait selon différentes visions/métriques (S/P, part de marché) en sélectionnant la segmentation qui permet :
  - d'**optimiser la moyenne  $\mu$**  : minimiser le S/P moyen ou maximiser la PdM moyenne
  - de **minimiser la volatilité  $\sigma$** , qui représente la composante « risque »
  - on peut s'intéresser aux ratios suivants, à maximiser :

$$ratio_i^{SP} = \frac{\mu_{k_{SP}} - \mu_i}{\sigma_i}$$

$$ratio_i^{PdM} = \frac{\mu_i - \mu_{k_{PdM}}}{\sigma_i}$$

avec  $k_{SP}$  tel que  $\mu_{k_{SP}} = \max_j \mu_j$

Avec  $k_{PdM}$  tel que  $\mu_{k_{PdM}} = \min_j \mu_j$

# Construction d'une méthodologie de travail

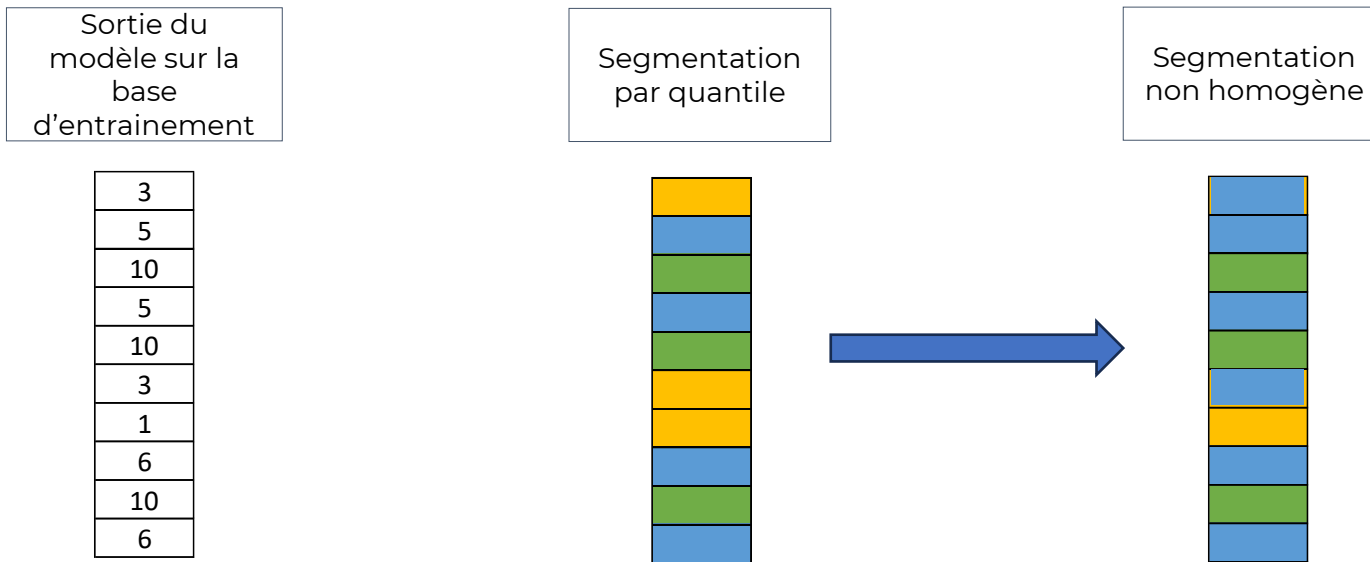
Modélisation du niveau de rationalité des assurés

- **Question sous-jacente** : comment l'assuré choisit-il l'assureur **étant donné l'information dont il dispose** ?
- **Différentes hypothèses** :
  - **« Rationalité parfaite »** : sélection du tarif minimum parmi tous les choix proposés
  - **« Rationalité imparfaite »** : sélection aléatoire parmi les  $n$  tarifs les moins chers
  - **« Information incomplète »** : sélection du tarif minimum parmi  $n$  tarifs tirés aléatoirement
  - **« Effet moutonnier »** : sélection du tarif en fonction de la part de marché de l'assureur
- Possibilité de **combiner** plusieurs de ces hypothèses :
  - Pour créer de nouvelles fonctions de sélection
  - Au sein de la population : chaque assuré a une fonction de sélection différente

# Construction d'une méthodologie de travail

Alternatives à la segmentation par quantile

- **Regroupements tarifaires non homogènes :**



- **Création de modèles différents pour chaque assureur :**
  - Se rapproche d'une vision réelle (chaque assureur à son propre modèle)
  - Mais les résultats ne portent plus uniquement sur la segmentation

# Construction d'une méthodologie de travail

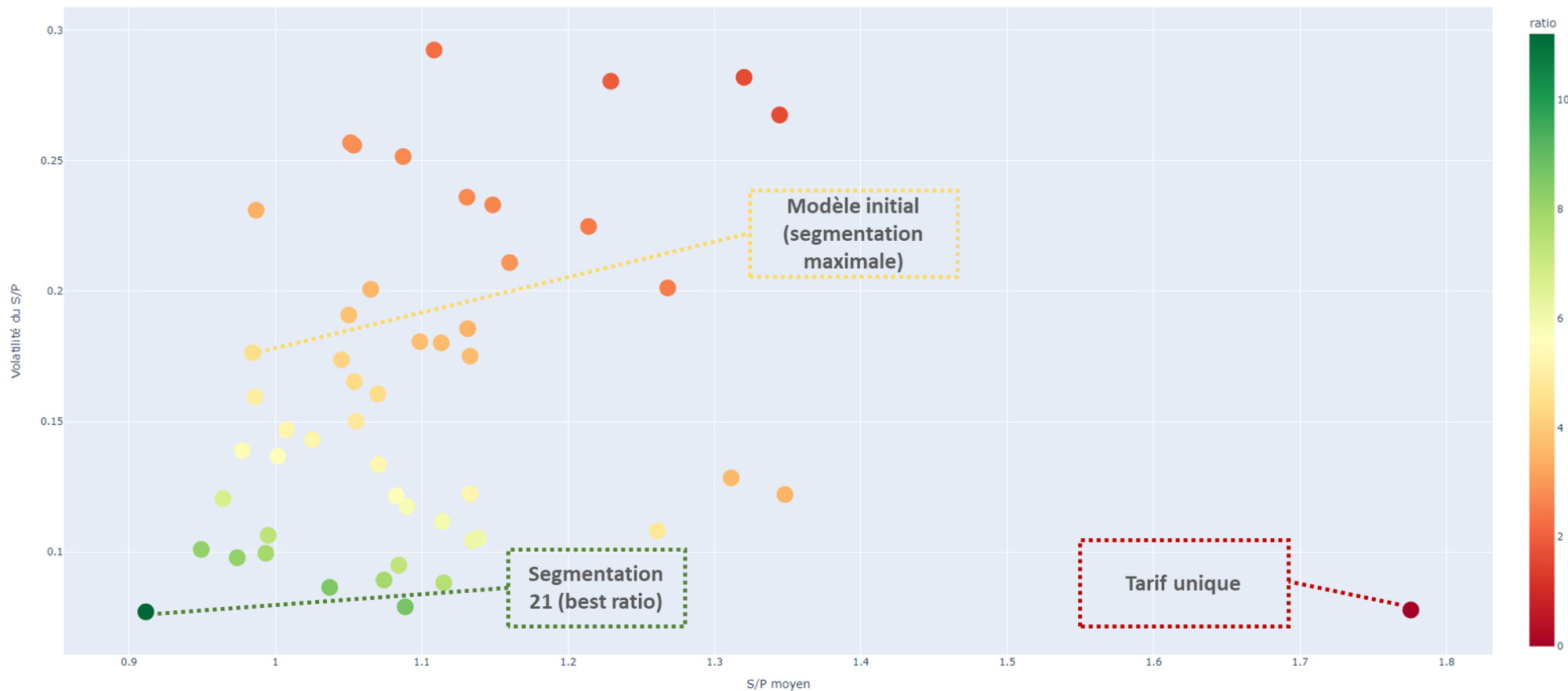
## Intégration d'une dynamique temporelle

- **Objectif** : intégrer une **dynamique temporelle** au marché concurrentiel pour évaluer la **stabilité** des indicateurs
- **Principe** : l'assureur revoit son tarif et l'assuré reste chez le même assureur avec un certain **taux de maintien**
- Ce taux de maintien dépend de :
  - L'évolution des **tarifs** de l'**assureur**
  - L'évolution des **tarifs** de la **concurrence**
  - L'évolution des **parts de marché**
  - Du nombre de d'années d'**ancienneté** (fidélité)
  - De l'intégration de **remises commerciales**
- Ici aussi, possibilité de combiner différentes composantes

# Présentation des premiers résultats

## Etude du critère S/P – Rationalité maximale

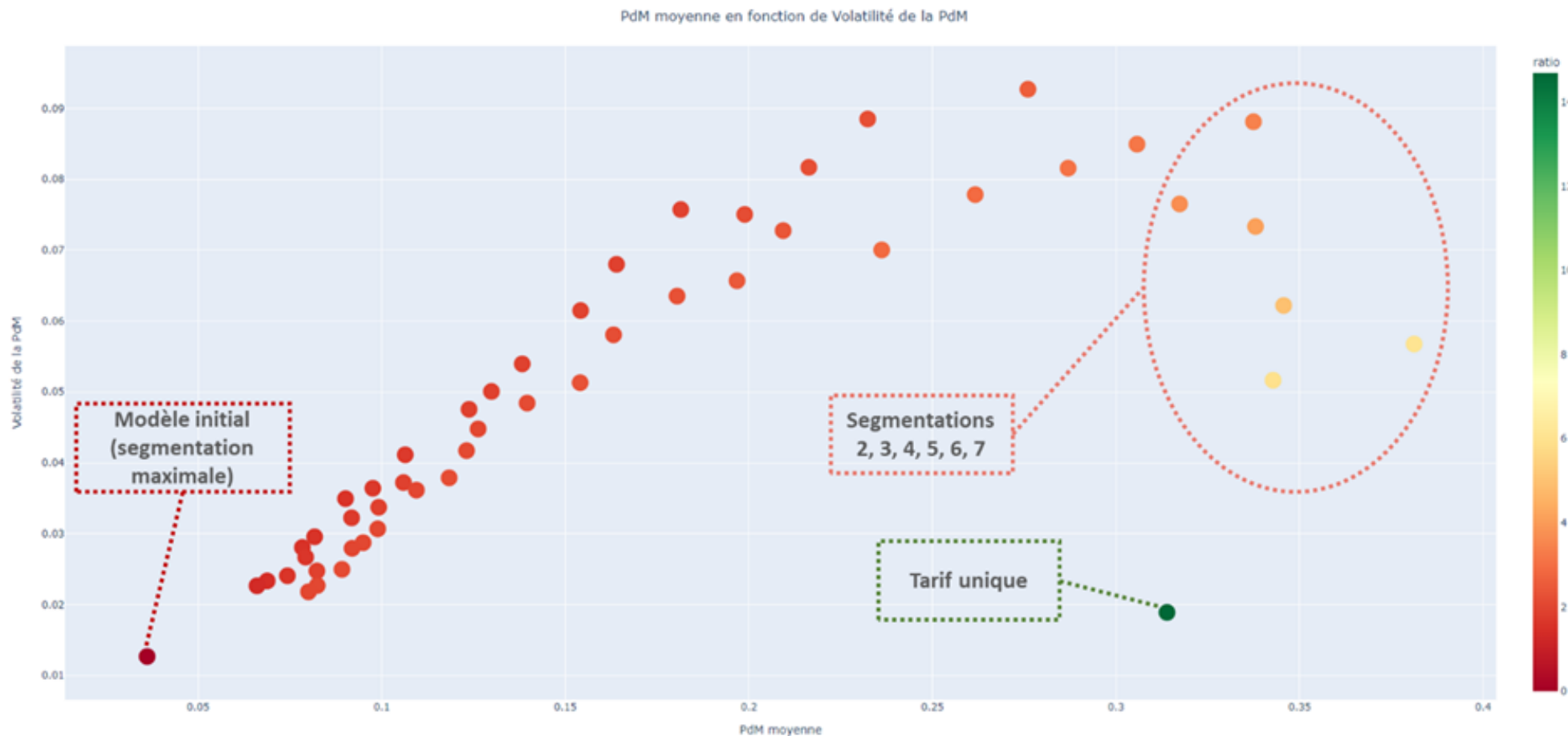
S/P moyen en fonction de Volatilité du S/P



- **50 niveaux de segmentation**
- **6 assureurs par sous marché**
- **1 500 scénarios par assureurs**
- **Modélisation d'une année de concurrence uniquement**
- **Niveau de rationalité maximal**

# Présentation des premiers résultats

Etude du critère PdM – Rationalité maximale



- 50 niveaux de segmentation
- 6 assureurs par sous marché
- 1 500 scénarios par assureurs
- Modélisation d'une année de concurrence uniquement
- Niveau de rationalité maximal

# Présentation des premiers résultats

## Sensibilité à taille des sous marchés

### Objectif :

- Etudier la sensibilité des résultats par rapport aux nombres de concurrents dans les sous-marchés aléatoires générés

### Paramètres :

- Assureurs par sous marché : de 4 à 7;
- 50 niveaux de segmentation ;
- 1 500 scénarios par assureurs ;
- Modélisation d'une année de concurrence uniquement ;
- Niveau de rationalité maximal ;

Nombre de concurrents par sous-marché	Top 1			Top 2			Top 3		
	Modèle	S/P	Part de marché	Modèle	S/P	Part de marché	Modèle	S/P	Part de marché
4	21	0,923	25%	25	0,925	23,80%	22	0,958	25%
5	25	0,902	18,10%	21	0,904	19,60%	22	0,944	19,30%
6	25	0,888	14,10%	21	0,897	16,10%	22	0,933	15,60%
7	25	0,876	11,80%	21	0,891	13,40%	22	0,93	12,80%





# Présentation des premiers résultats

## Sensibilité à la rationalité – plusieurs approches

### Objectif :

- Etudier la sensibilité des résultats par rapport aux différents comportements des assurés.

### Les fonctions de rationalité :











- « **Information incomplète** » : Choix du tarif minimal parmi les  $n$  tarifs tirés aléatoirement.
- « **Rationalité imparfaite** » : Choix aléatoire du tarif parmi les  $n$  tarifs minimaux.
- Choix du tarif minimal avec une probabilité  $p$ .
- Passage au tarif minimal si l'écart par rapport au tarif actuel est supérieur à un seuil de tolérance  $t$ .

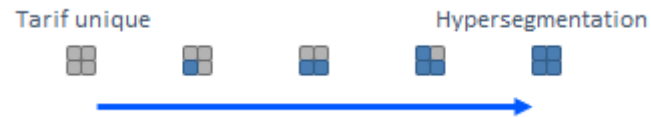
### Paramètres :

- 6 assureurs par sous marché ;
- 50 niveaux de segmentation ;
- 1 500 scénarios par assureurs ;
- Modélisation d'une année de concurrence uniquement ;
- Les diverses **fonctions** de **rationalité**.

# Présentation des premiers résultats

## Sensibilité à la segmentation

Fonction de rationalité	Paramètres	Top 1 Modèle			Moyenne 5 top modèles		
		Modèle	S/P	Part de marché	Modèle	S/P	Part de marché
Rationalité maximale	-	 21	0,896	16,3%	 19,2	0,958	18,3%
Choix du tarif minimal parmi les $n$ tarifs tirés aléatoirement	3	 25	0,945	32,8%	 31,6	0,974	31,6%
	4	 21	0,924	25,3%	 27,4	0,956	23,4%
Choix aléatoire du tarif parmi les $n$ tarifs minimaux	3	 49	0,969	17,5%	 36,0	0,972	17,5%
	4	 49	0,969	18,6%	 45,2	0,977	19,0%



# Conclusion

## Éléments de conclusion

- Ces travaux ont permis de montrer que l'hypersegmentation ne semble pas optimale pour les deux indicateurs étudiés (S/P et Part de marché).
- Selon le critère d'intérêt, les premiers résultats indiquent une "zone d'optimalité".
- Toutefois, caractériser le nombre optimal de segments n'est pas évident et dépendra du critère à optimiser ainsi que de la fonction de rationalité ainsi que de la performance du modèle initial.

# Conclusion

## Piste de réflexion

### Prochaines étapes :

- Renforcer l'indépendance des modèles
  - Limiter la concurrence de modèles proches
  - Prise en compte de quantiles pour la distinction des segments
- Améliorer la dynamique de population
  - Prise en compte de comportement plus complexe mais plus naturelle
  - Prise en compte de différents groupes de population avec des comportements différenciés
  - Réalisation d'une analyse sur plusieurs années (comportement temporel)
- Réflexion autour de la généralisation de notre approche
- Réflexion autour de l'aléa moral à la sur-segmentation