



# La mutualisation et l'inclusion à l'épreuve de la segmentation

Arthur Charpentier (Montréal, Canada)

3<sup>e</sup> Colloque International de l'Actuariat Francophone



International Actuarial Association  
Association Actuarielle Internationale



APSA.bj  
L'Asbl pour la promotion des Sciences Actuarielles  
Association pour la Promotion des Sciences Actuarielles

Institut des Actuaires en Belgique  
Instituut van de Actuariassen in België  
Institute of Actuaries in Belgium



Institut canadien  
des actuaires  
Canadian  
Institute  
of Actuaries



Institut des  
**ACTUAIRES**



Schweizerische  
Aktuarvereinigung  
Association Suisse  
des Actuaires  
Associazione Svizzera  
degli Attuari

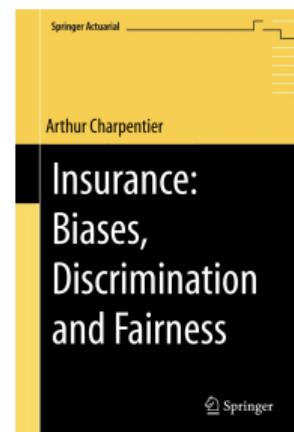
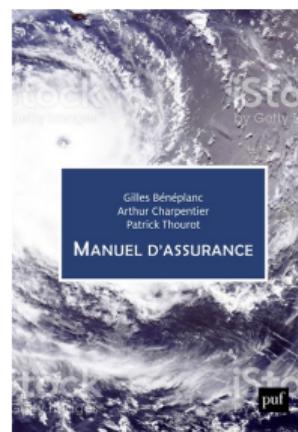
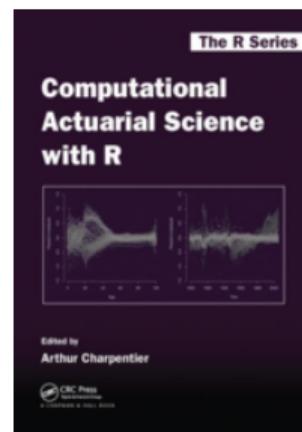
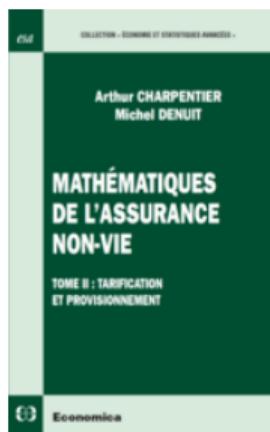
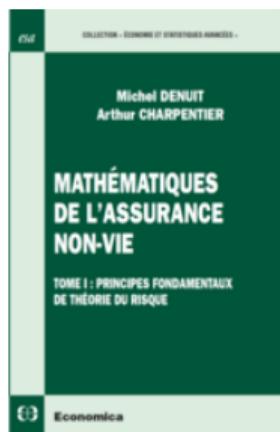


Association Tunisienne des Actuaires

# bio (succinte)

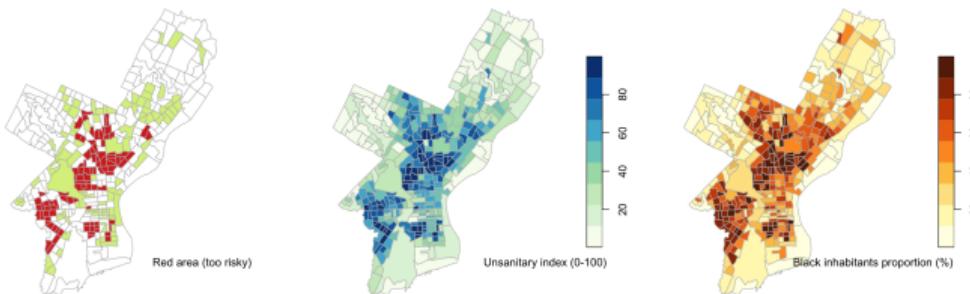
Professeur à l'Université du Québec à Montréal (<https://freakonometrics.github.io/>)

- › Denuit and C. (2004, 2005) Mathématiques de l'Assurance Non-Vie,
- › C. (2014) Computational Actuarial Science with R,
- › Bénéplanc et al. (2022) Manuel d'Assurance,
- › C. (2024) Insurance: Biases, Discrimination and Fairness.

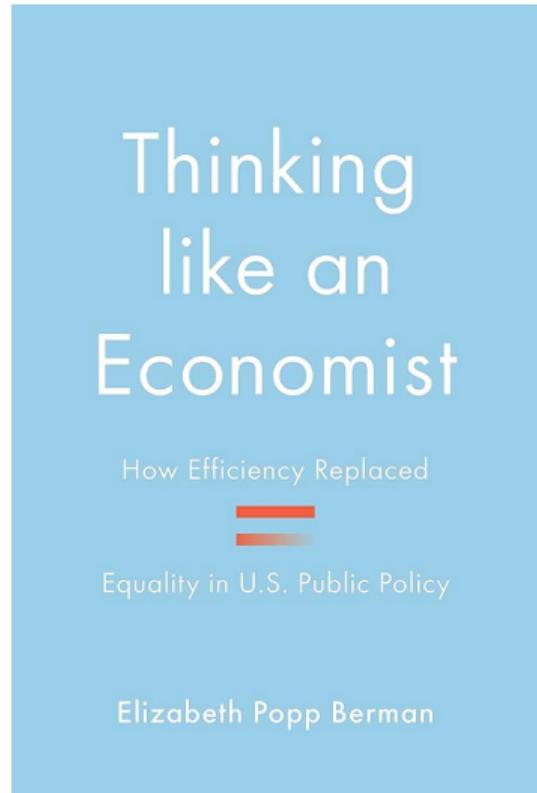


# justification économique de la segmentation

- > notion de rationalité et d'actuarial fairness,
- > nécessaire de tenir compte de l'hétérogénéité
- > cf efficacité en économie, Berman (2022)
- > quelle hétérogénéité ? est-ce juste ?



- > cf "redlining" (assurance ou crédit)  
la prime d'assurance est-elle plus chère
  - parce que les maisons sont vétustes ?
  - parce que la population est noire ?



# de la spirale de la segmentation

- › "Insurance is the contribution of the many to the misfortune of the few"
- › Mathématiques de la segmentation (De Wit and Van Eeghen (1984), C. et al. (2014))

$$\text{Var}[Y] = \underbrace{\text{Var}[\mathbb{E}[Y|\mathbf{X}]]}_{\rightarrow \text{assuré}} + \underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\rightarrow \text{modèle "parfait"}} + \underbrace{\mathbb{E}\{\text{Var}[\mathbb{E}[Y|\Theta]]|\mathbf{X}\}}_{\rightarrow \text{imperfection}}.$$

## SEGMENTATION ET MUTUALISATION LES DEUX FACES D'UNE MÊME PIÈCE ?

**Arthur Charpentier**  
Professeur à l'Université du Québec, Montréal

**Michel Demut**  
Professeur à l'Université catholique de Louvain

**Romuald Elie**  
Professeur à l'Université de Marne-la-Vallée

L'assurance repose fondamentalement sur l'idée que la mutualisation des risques entre des assurés est possible. Cette mutualisation, qui peut être vue comme une réflection actuarielle de la loi des grands nombres, n'a de sens qu'en tant d'une population de risques « homogènes » [Charpentier, 2011]. C'est condition nécessaire pour assurer de segmenter, ce qui correspond plusieurs fois au sens commun. Avec l'explosion du nombre de données, et donc de variables possibles, certaines assureurs devraient l'idée d'un tarif individualisé, souhaitant rester en cause l'idée de mutualisation des risques. Ensuite cette frise qui pose à segmenter et le faire de sorte que tout (pour des raisons sociales mais aussi actuarielles) ou au moins de robustesse statistique (2) à imposer une solidarité renversée entre les assurés, quel qu'il soit à se rendre dans un contexte de forte concurrence avec le risque d'assurance.

### Tarification sans segmentation

Sous segmentation, le « pris juste » d'un risque est l'espérance de la perte qu'il entraîne sur sa durée de vie. C'est l'efficacité d'assurance dividende de la valorisation actuarielle : on moyenne, la somme des prises d'assurance permet d'indemniser l'intégralité des autres personnes dans

l'assurance. Ainsi d'illustre les différents aspects de la construction du tarif et ses implications sur sa solidarité : il faut faire en sorte de faire le meilleur (tarif juste) qui respecte la fréquence d'assurance.

Le taux de risque correspond au taux d'indemnisation et l'espérance de l'assurance devient alors la somme des taux de risque multipliés par leur probabilité.

Le primaire est alors  $E[Y] = 1000 \times 0.001$ . Dans cet exemple, la primaire pure segmentation, soit de 82,00 euros.

	Assuré	Age	Habitation	Agri	Salarié	Médecin
J/V (2000)	82,3	30	99	120	82,3	
J/C (2000)	82,3	30	63	80	63,3	
K/V (2000)	82,3	30	99	90	82,3	
E/C (1000)	82,3	32,5	63,5	66,7	63,5	
V/W (1000)	82,3	32,5	63,5	99	45	
P/W (1000)	82,3	32,5	63,5	66,7	45	
Promo	41,17	196,94	99	20	253,19	
Statut	60	225	106,67	20	411,67	
XP	1437,5%	114,2%	112,3%	100,6%	116,6%	
EC: 90-90	100,0%	100,0%	100,0%	100,0%	100,0%	
EC: 90-91	100,0%	100,0%	100,0%	100,0%	100,0%	
Avr de mortalité	11,6%	55,8%	26,0%	5,7%	10,1%	

Segmentation et mutualisation, les deux faces d'une même pièce ?

Segmentation et mutualisation, les deux faces d'une même pièce ?

Segmentation et mutualisation, les deux faces d'une même pièce ?

Tableau 1 - Décomposition de la variance de la dépense entre l'assurance et l'assuré en présence d'une segmentation parfaite

	Assuré	Assuré
Dépense	82,3	82,3
Dépense moyenne	82,3	0
Covar (E[X])	0	0
Total	100%	82,3%
	(1000)	(1000)

Sources : annexe, exemple final.

Tableau 2 - Décomposition de la variance de la dépense entre l'assurance et l'assuré sans segmentation

Sources : annexe, exemple final.

Tableau 3 - Décomposition de la variance de la dépense entre l'assurance et l'assuré en présence d'une segmentation parfaite

Sources : annexe, exemple final.

On voit alors, dans un exemple, que dire que la segmentation est une chose et la mutualisation une autre chose est une chose de bien peu de sens. En effet, lorsque la société d'assurance utilise le tiers d'indemnisation et par l'agé, et se retrouve en concurrence face à une société qui ne segmente pas (telle le cas du tableau 6). Dans le second cas, la société d'assurance utilise l'âge comme variable tarifaire (cas 2 du tableau 6).

Ce graphique de variance pour l'assurance traduit le fait que la segmentation utilisée en cas de des assurés différenciés homogène.

Dans un autre exemple illustratif, on peut regarder le tableau 6. On voit que la première société d'assurance utilise le tiers d'indemnisation et par l'âge, et se retrouve en concurrence face à une société qui ne segmente pas (telle le cas du tableau 6). Dans le second cas, la société d'assurance utilise l'âge comme variable tarifaire (cas 2 du tableau 6).

### Tarification avec segmentation (imparfaite)

La méthode n'est pas tout simple. En particulier, la connaissance des classes de risque est souvent imprécise. L'assurance n'a pas accès à des données parfaitement détaillées sur le portefeuille des risques, mais alors que l'assurance utilise ces données pour élaborer des modèles de prévision, mais pas pour élaborer des modèles de segmentation. Ces deux parties sont donc éloignées l'une de l'autre.

Les résultats de segmentation sont alors classés dans la mesure où l'assurance connaît ces dernières, et classe également les risques sous forme d'intervalles de risque qui dépendent pas, mais à des intervalles plus larges que ceux utilisés par l'assurance. Cela va de pair avec l'assurance qui connaît les intervalles de risque mais qui ne connaît pas les intervalles de segmentation.

On voit alors que l'assurance utilise une « segmentation parfaite », mais que les intervalles de risque sont assez larges, et que l'assurance qui ne connaît pas les intervalles de segmentation, mais qui connaît les intervalles de risque, utilise une « segmentation parfaite ».

Tableau 4 - Répartition des classes d'âges lorsqu'un assureur propose de segmenter ses clients en fonction de leur risque

Sources : annexe, exemple final.

On voit alors que l'assurance utilise l'âge comme variable tarifaire (cas 1 du tableau 6).

Tableau 5 - Répartition des classes d'âges lorsqu'un assureur utilise une information imprécise pour tarifier

Sources : annexe, exemple final.

Tableau 6 - Répartition des classes d'âges lorsqu'un assureur utilise une information complète pour tarifier

Sources : annexe, exemple final.

### Tarification avec segmentation (parfaite)

Supposons maintenant qu'il existe une assurance sociale qui n'a pas encore qu'il dispose d'une connaissance parfaite des classes de risque lorsque l'assurance sera mise en œuvre. Cet assureur utilise alors l'âge comme variable tarifaire (cas 1 du tableau 6).

En effet, on retrouve ici la relation classique de décomposition de la variance.

$\text{Var}(E[X|Z]) + \text{Var}(E[Y|Z|X]) = \text{Var}(Y)$

Sources : annexe, exemple final.

On retrouve alors que la variance de l'assurance est ici  $\text{Var}(V(S|Z)) + \text{Var}(V(Y|S|Z))$ .

avec un terme qui correspond à la variance que nous avons en situation d'information parfaite, mais aussi un terme additionnel que l'assureur peut interpréter comme étant la variance due à l'incertitude de l'assurance sur la segmentation. Cela permet alors de décomposer la variance totale des dépenses sous la forme

$\text{Var}(E[X|Z]) + \text{Var}(E[Y|Z|X]) = \text{Var}(V(S|Z)) + \text{Var}(V(Y|S|Z))$

Sources : annexe, exemple final.

On retrouve alors que la variance de l'assurance est ici  $\text{Var}(V(S|Z)) + \text{Var}(V(Y|S|Z))$ .

avec un terme qui correspond à la variance que nous avons en situation d'information parfaite, mais aussi un terme additionnel que l'assureur peut interpréter comme étant la variance due à l'incertitude de l'assurance sur la segmentation. Cela permet alors de décomposer la variance totale des dépenses sous la forme

$\text{Var}(E[X|Z]) + \text{Var}(E[Y|Z|X]) = \text{Var}(V(S|Z)) + \text{Var}(V(Y|S|Z))$

Sources : annexe, exemple final.

On retrouve alors que la variance de l'assurance est ici  $\text{Var}(V(S|Z)) + \text{Var}(V(Y|S|Z))$ .

avec un terme qui correspond à la variance que nous avons en situation d'information parfaite, mais aussi un terme additionnel que l'assureur peut interpréter comme étant la variance due à l'incertitude de l'assurance sur la segmentation. Cela permet alors de décomposer la variance totale des dépenses sous la forme

$\text{Var}(E[X|Z]) + \text{Var}(E[Y|Z|X]) = \text{Var}(V(S|Z)) + \text{Var}(V(Y|S|Z))$

Sources : annexe, exemple final.

On retrouve alors que la variance de l'assurance est ici  $\text{Var}(V(S|Z)) + \text{Var}(V(Y|S|Z))$ .

avec un terme qui correspond à la variance que nous avons en situation d'information parfaite, mais aussi un terme additionnel que l'assureur peut interpréter comme étant la variance due à l'incertitude de l'assurance sur la segmentation. Cela permet alors de décomposer la variance totale des dépenses sous la forme

$\text{Var}(E[X|Z]) + \text{Var}(E[Y|Z|X]) = \text{Var}(V(S|Z)) + \text{Var}(V(Y|S|Z))$

Sources : annexe, exemple final.

# de la théorie aux expériences, les premiers "pricing games"

## ► 2015-2017, "pricing games" (en lien avec l'institut des actuaires)

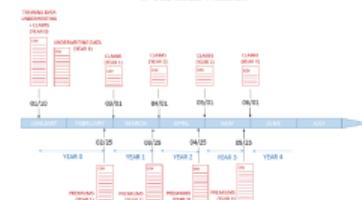
- Trois compétitions (2015, 2016 et 2017)
- De 10 à 25 participants (qui donnaient des prix, pas des modèles)

### THIRD ACTUARIAL PRICING GAME

From January 2017 till June 2017, we organize the Third Actuarial Pricing Game, as part of a research project conducted by Arthur Charpentier, Université de Bourgogne 1 (France) & Quantact (Montreal, Canada), with the support of the ACTINFO chair of the Institut Louis Bachelier, and the Institut des Actuaires, the French Institute of Actuaries.



#### 1. THE RULES : AGENDA



On January 10th, datasets with 100,000 insured is provided to all (potential) players. There are two datasets for each player:  
— an underwriting dataset ( $U_i$ ), with information about insurance policies, insured drivers and their cars  
— see page 5 for description of the variables)  
— a claims dataset ( $C_i$ ), with all claims occurred during the year 2016 for all policyholders  
An underwriting dataset ( $U_i$ ) with the same 100,000 policies/drivers willing to purchase insurance for Year 1 is provided. Players must provide prices for those 100,000 insured.  
A player can be a dataset or not. It can be sent by individuals, or groups, practitioners, students or academics.

For February 25th, players must send a dataset with 100,000 prices to pricing game@freakonometrics.fr in a zip file with two columns: the Id\_policy, and the premium. Let  $\hat{p}_{i,j}$  denote the initial premium for policy  $i$ , offered by player  $j$  (for year 1). Let  $t_{i,j}$  denote the losses for policy  $i$  (unknown by players when submitting their premiums).

1

lightgray Num	Family Type	Name	Label	Format
1	ID	id_client	ID: Client ID	string
2	ID	id_vehicle	ID: Vehicle ID	string
3	ID	id_year	ID: Year	string
4	Claims	clm.claim	Claims: Claim ID	string
5	Claims	clm.clm_nb	Claims: Number of Claims	int
6	Claims	clm.amount	Claims: Total Claims Amount	int

Variables List : Claims database

lightgray Num	Family Type	Name	Label	Format
1	ID	id_client	ID: Client ID	string
2	ID	id_vehicle	ID: Vehicle ID	string
3	ID	id_policy	ID: Policy ID	string
4	ID	id_year	ID: Year	string
5	Policy	pol.bonus	Policy: Bonus Coefficient	float
6	Policy	pol.coverage	Policy: Coverage	string
7	Policy	pol.duration	Policy: Duration	int
8	Policy	pol.st_duration	Policy: Current Endorsement Duration	int
9	Policy	pol.pay.freq	Policy: Payment Frequency	string
10	Policy	pol.premium	Policy: Premium	float
11	Policy	pol.usage	Policy: Usage	string
12	Policy	pol.issue_code	Policy: Issue Town Code	string
13	Drivers	drvdrv2	Drivers: Secondary Driver Presence Indicator	string
14	Drivers	drv.age1	Drivers: First Driver Age	int
15	Drivers	drv.age2	Drivers: Secondary Driver Age	int
16	Drivers	drv.sex1	Drivers: First Driver Gender	string
17	Drivers	drv.sex2	Drivers: Secondary Driver Gender	string
18	Drivers	drv.agelcl1	Drivers: First Driver License Age	int
19	Drivers	drv.agelcl2	Drivers: Secondary Driver License Age	int
20	Vehicle	vltage	Vehicle: Vehicle Age	int
21	Vehicle	vlctyl	Vehicle: Engine Capacity	int
22	Vehicle	vltdin	Vehicle: DIN Power	int
23	Vehicle	vltfuel	Vehicle: Fuel Type	string
24	Vehicle	vlvtype	Vehicle: Type	string
25	Vehicle	vlmodel	Vehicle: Model	string
26	Vehicle	vl.sale.begin	Vehicle: Sales Date Beginning	int
27	Vehicle	vl.sale.end	Vehicle: Sales Date End	int
28	Vehicle	vl.speed	Vehicle: Max Speed	int
29	Vehicle	vl.type	Vehicle: Type	string
30	Vehicle	vl.value	Vehicle: Value	int
31	Vehicle	vl.weight	Vehicle: Weight	int

Variables List : Underwriting database

## 2 Instructions, partie 1 (15 Juillet - 15 Septembre)

Le but est de proposer une prime pure pour les 37 772 contrats de la base **pricing**.  
Sont attendus

- une base (un fichier csv) constituée de deux colonnes (seulement): la première n'a affa sera le numéro de police et la seconde Premium contiendra la prime pure proposée. La base sera constituée de 37 772 lignes (il est impératif de proposer une prime pour *tous* les contrats).
- un rapide descriptif de la méthodologie utilisée, décrivant les variables retenues.

Merci d'indiquer un nom déscriptif pour chaque modèle proposé, et d'envoyer le fichier csv avec un rapide descriptif à [PricingGame@institutdesactuaires.com](mailto:PricingGame@institutdesactuaires.com), avant le **15 Septembre**.

## 3 Instructions, partie 2 (20 Septembre - 20 Octobre)

Le 20 Septembre, tous les participants recevront les primes de 2 de leurs compétiteur au moins.

Les participants auront alors la possibilité de réviser (ou pas) la prime qu'ils proposent. Un fichier csv révisé - et un rapide descriptif de la méthodologie adoptée - devront être envoyé à [PricingGame@institutdesactuaires.com](mailto:PricingGame@institutdesactuaires.com) pour le **20 Octobre**, au plus tard.

## 4 Règles du Jeu

Une fois les primes collectées, nous fonctionnerons comme un agrégateur de prix: chacun des xxx assurés se verra affecter une compagnie selon une m'thode qui sera basée sur une sélection aléatoire parmi les primes les moins chères. Les organisateurs normalisent les prix pour éviter les stratégies de dumping. En l'occurrence, la somme des primes proposées par un assureur sera égale à 2,5 millions d'euros (par un facteur d'inflation appliqué uniformément).

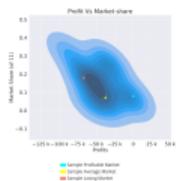
Les résultats seront présentés lors de la journée 100% Data Science organisée à Paris (voir également sur <http://freakonometrics.hypotheses.org>).

Pour plus d'information, [PricingGame@institutdesactuaires.com](mailto:PricingGame@institutdesactuaires.com) ou @freakonometrics sur Twitter.

# retour sur la compétition de 2021

## ► 2021, "IA Crowd pricing game"

- 11 semaines ≈ 2,000 participants, 10,000 modèles
- retours sur les performances en compétition



Driver Summaries

Policies won	drv_sex1	drv_age1	drv_age_ic1	drv_drv2	drv_sex2	drv_age2	drv_age_ic2
others 54.5 - 100.0%	M (53%)	65.57	38.70	No (73%)	F (66%)	58.00	27.74
sometimes: 18.2 - 54.5%	M (55%)	58.19	36.87	No (72%)	F (62%)	51.05	29.14
rarely: 9.1 - 18.2%	M (53%)	68.92	38.35	No (71%)	F (62%)	51.09	29.12
never:	M (58%)	58.17	36.08	No (64%)	F (60%)	51.16	29.31
In sample profitable market	M (54%)	61.05	38.33	No (74%)	F (64%)	49.29	25.25
In sample average market	M (59%)	61.73	38.95	No (75%)	F (65%)	50.43	28.29
In sample losing market	M (53%)	59.89	37.37	No (73%)	F (64%)	50.70	26.65

- compétition avec un prix à gagner  
(donc une métrique cible de succès)

Insurance Pricing Game      Active tracks      Timelines      Organizers

a global  
data science competition  
with real motor insurance data



Real motor  
insurance data



Build a prediction  
model for claims



Play in a simulated  
marketplace

#track A

### Motor insurance market simulation

Play as an insurance company, using historical data in a competitive market with other players. See if you can make a profit with realistic market conditions.

Not yet launched! →

**AIcrowd**

- \$6000 Leaderboard
- \$3500 PartnerRe
- École Polytechnique Montréal
- actuaris tech
- \$1500 Actuaries Institute Australia
- \$1000 ACTUARIES

Learn more about AIcrowd Partners

#track B

### Worker Compensation Claim Prediction

sponsored by

- Actuaries Institute Australia
- Canadian Institute of Actuaries
- argenesis

Not yet launched! →

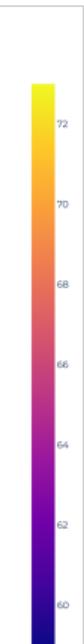
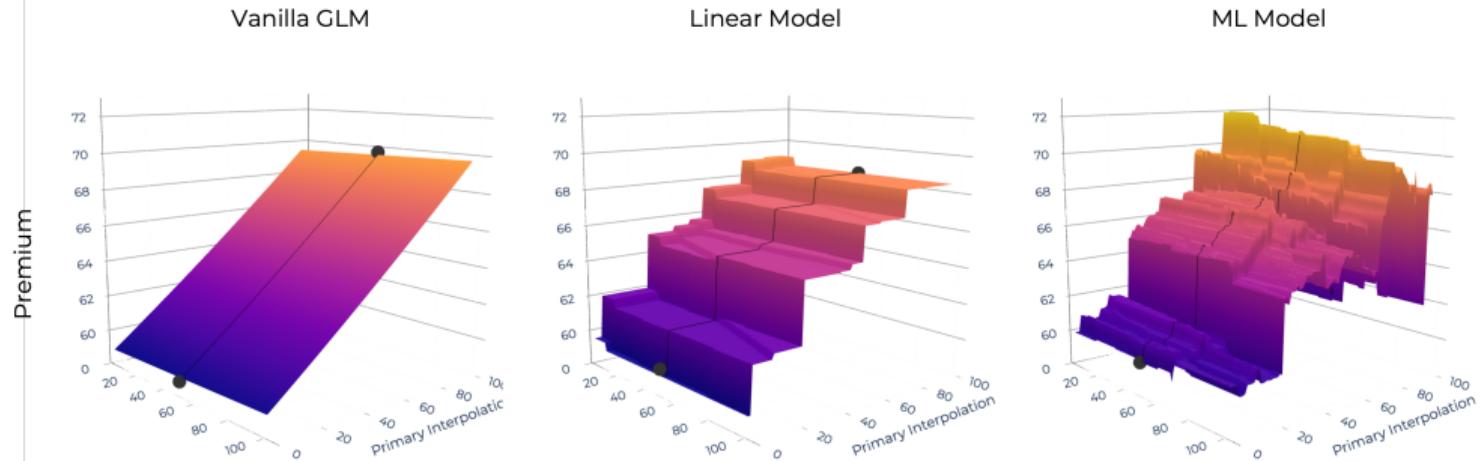
Learn more about Actuaries Institute Australia

# compétition entre modèles, quel équilibre ?

- modèles d'apprentissage machine souvent **mal calibrés**, cf **Denuit et al. (2021)**

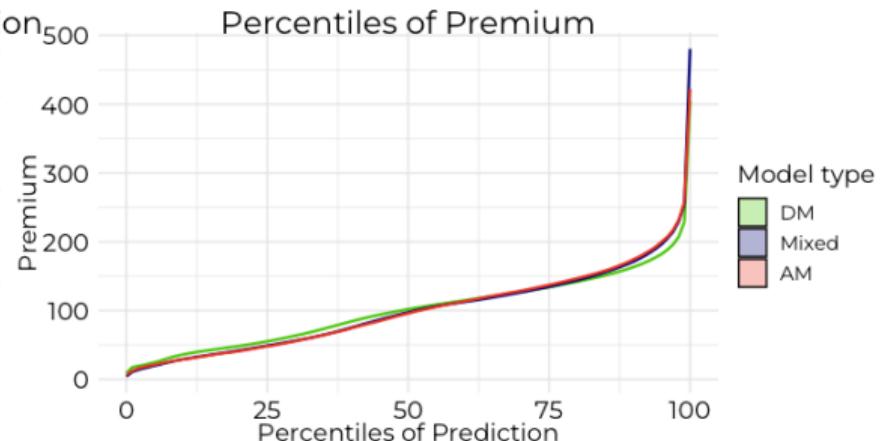
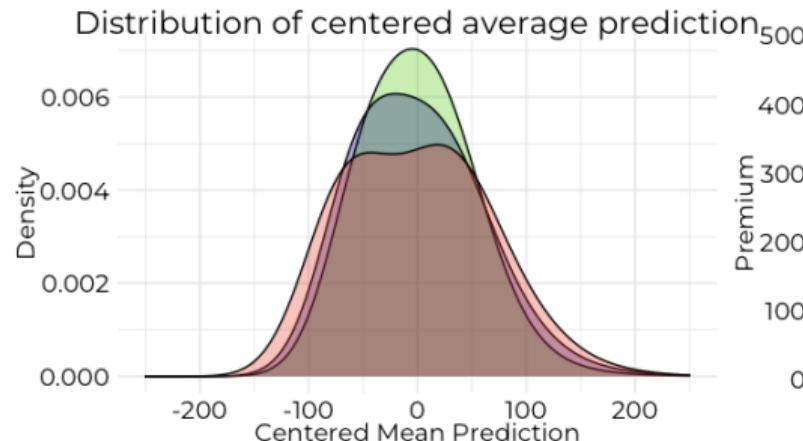
$$\mathbb{E}[Y|\hat{Y} = y] = y, \forall y$$

- apprentissage machine = plus de **variabilité** (avant la compétition)



# compétition entre modèles, quel équilibre ?

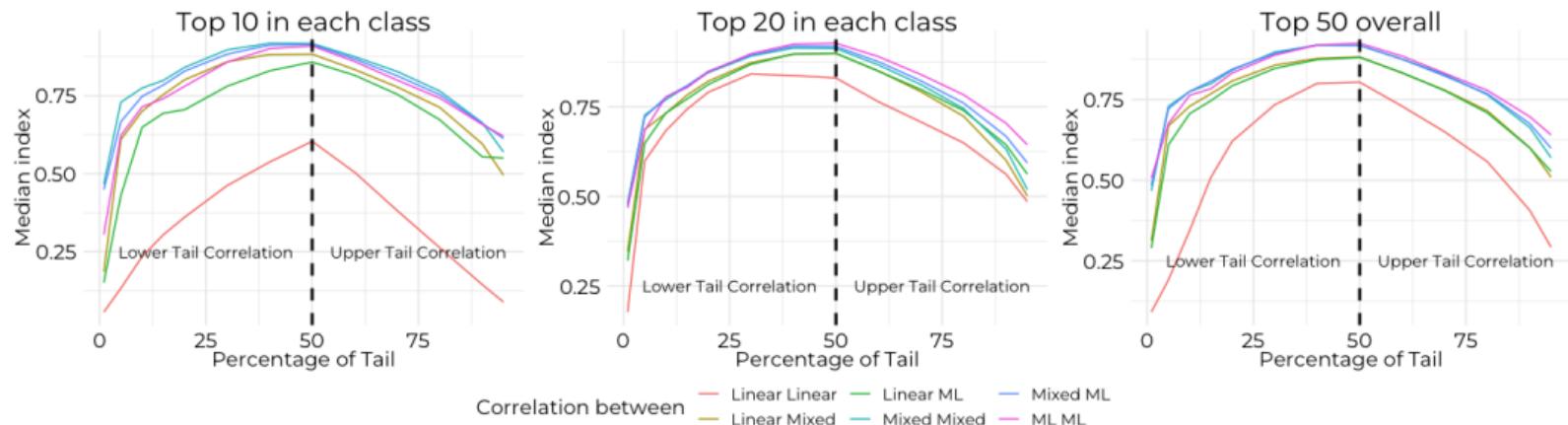
- apprentissage machine = plus de **variabilité** (avant la compétition)



(via Ratz et al. (2023))

# compétition entre modèles, quel équilibre ?

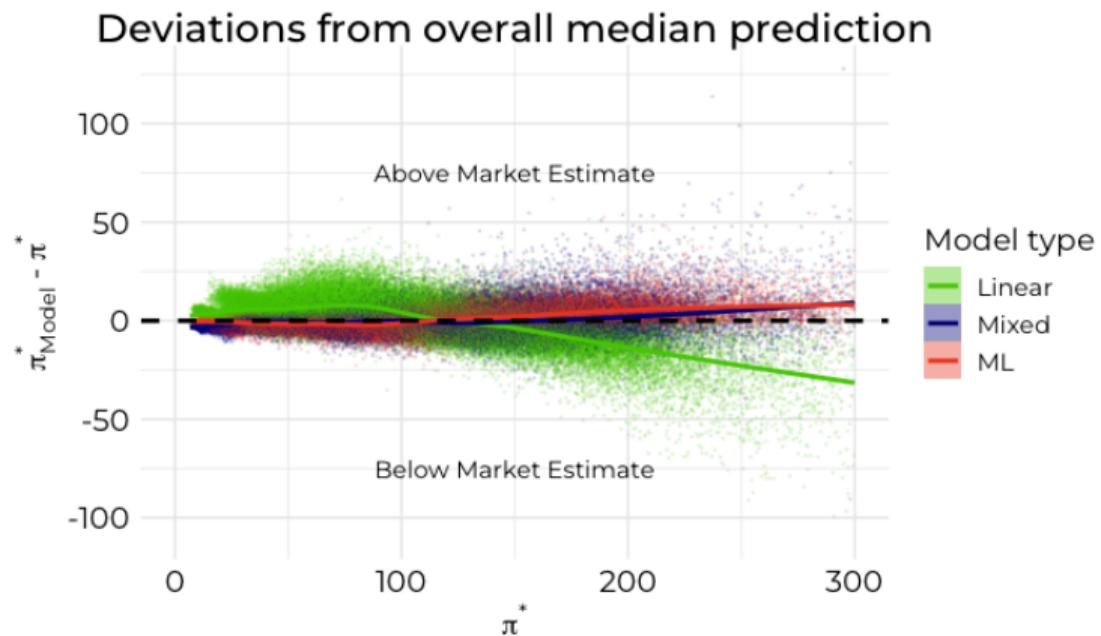
- peu de consensus entre modèles d'apprentissage machine
- ici corrélation (moyenne) de queues entre les modèles
  - à gauche,  $u \rightarrow \mathbb{P}[X_i \leq F_i^{-1}(u)|X_j \leq F_j^{-1}(u)]$
  - à droite,  $u \rightarrow \mathbb{P}[X_i > F_i^{-1}(u)|X_j > F_j^{-1}(u)]$



(via Ratz et al. (2023))

## compétition entre modèles, quel équilibre ?

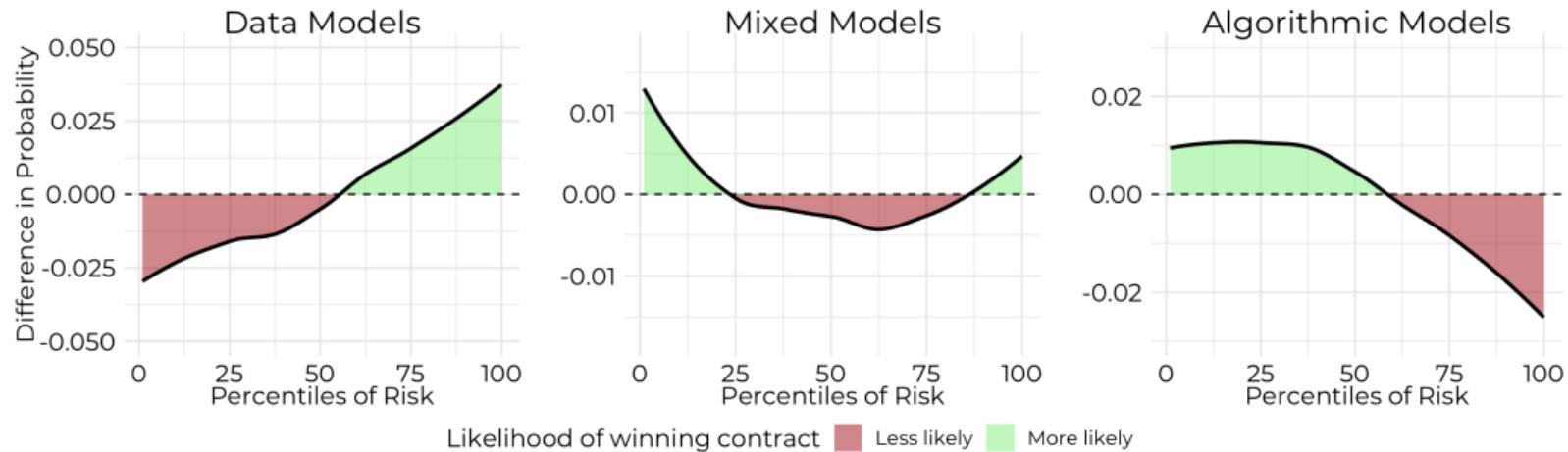
- les modèles linéaires sont chers sur les bas risques, pas chers sur les hauts risques



(via Ratz et al. (2023))

# compétition entre modèles, quel équilibre ?

- les modèles linéaires sont chers sur les bas risques, pas chers sur les hauts risques



(via Ratz et al. (2023))