Pratiques avancées de tarification et de provisionnement en assurance non vie

Master 2 Actuariat, ISFA, Université Lyon 1

Année universitaire 2020 - 2021

Xavier MILHAUD

(www.xaviermilhaud.fr)

PLAN DU COURS

1. Tarification a priori - concepts avancés
2. Construction d'un zonier
3. Provisionnement

ORGANISATION DU COURS

- 21h de cours magistral : 7 séances de 3h (2 séances en FC) ;
  - cours 1 - introduction : slide 1-25
  - cours 2 - : slide 26-45
  - cours 3 - slide 46-61
  - cours 4 - slide 62-84
  - cours 5 - slide 85-90 + début zonier
  - cours 6 - Fin zonier et début microlevel reserving
  - cours 7 - microlevel reserving
- 16h de travaux dirigés en salle machine (prenez vos ordinateurs pour chaque seance) : 8 séances de 2h.

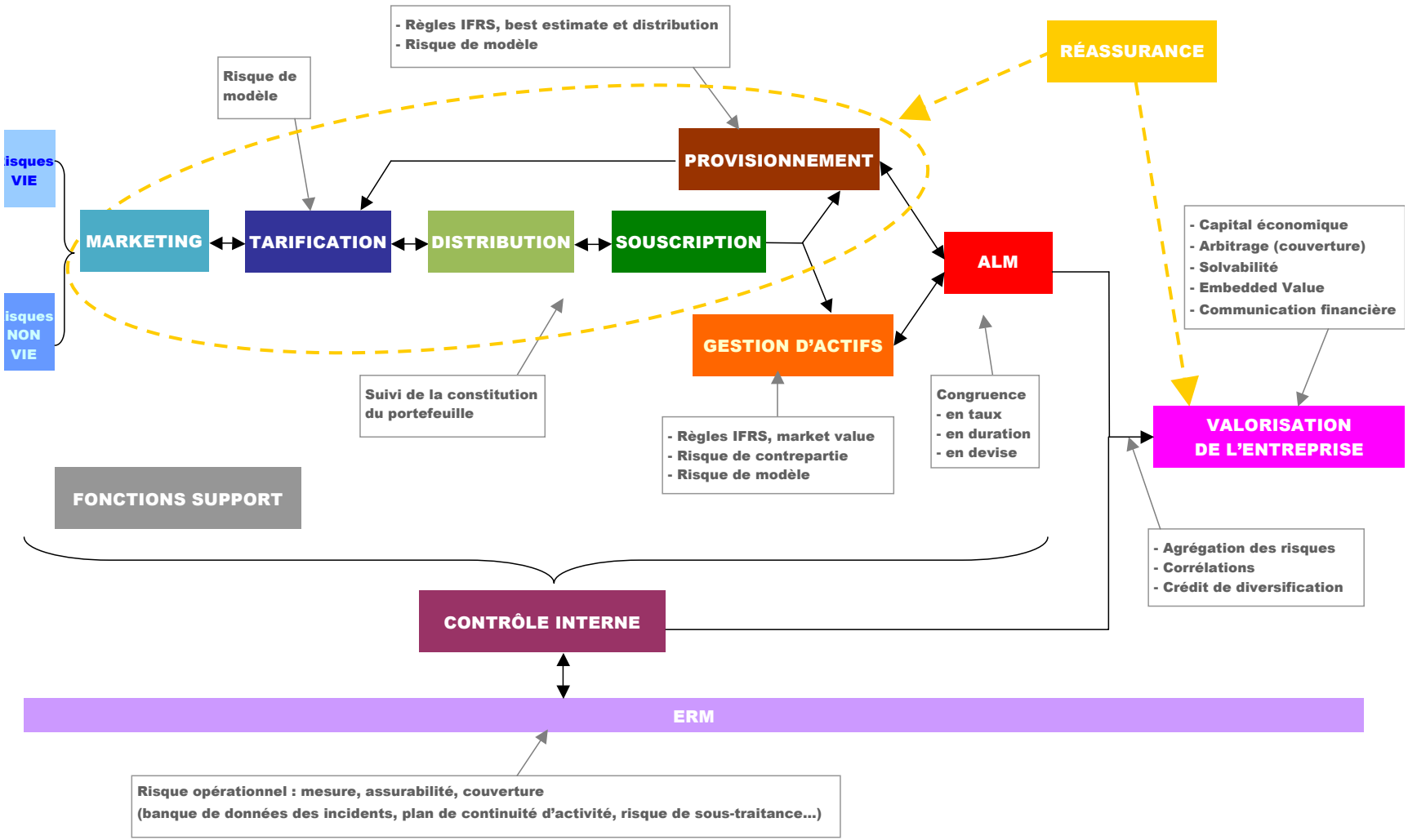
Objectif : confronter la théorie à la pratique !

1 Tarification a priori - concepts avancés

2 Construction d'un zonier

3 Provisionnement

La chaîne de gestion des risques dans l'assurance



ASSURANCE NON-VIE (IARD)

STATS

* **Auto/Moto**
  - Vol
  - Dommages matériel/Accident
  - Bris de glace
* **MRH**
  - Incendie
  - Vol
  - Inondation/dégât des eaux
* **RC**
  - RC corporelle
  - RC matérielle
* **CAT**
  - Cat Nat (sécheresse, tempêtes, ...)
  - Terrorisme ...
* **SANTE**
  - Hospital Costs (hospitalisation)
  - Remboursements (pharmacie, optique)

Rq : avec franchise souvent, annuel

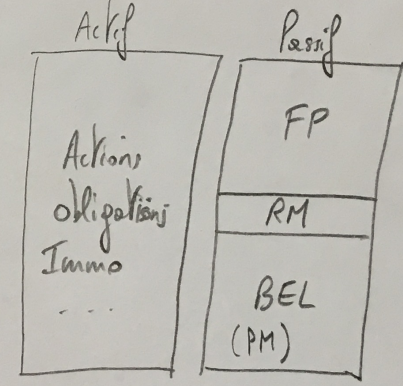
ASSURANCE VIE

PROBAS

* **Prévoyance** : Chaîne Markov
  - Dépendance (LTC) → AVQ/5
  - Maladies graves (Critical Illness) Cancer, Alzheimer, ...
  - Décès
  - Incapacité/Invalidité
    - < 3 ans
    - ↓ ≥ 3 ans
* **Epargne** : maths f
  - Rendement garanti (Euro)
    ⇒ Risque financier porté par l'assureur
  - Rendement non-garanti (Unités de Compte)
    ⇒ Risque financier porté par l'assuré

Rq : problématique d'horizon pluri-annuels, maturité

GESTION ACTIF/PASSIF



Problématique : interactions Actif/Passif à cause principalement :
→ des comportements d'assurés (rachats, ...)
→ des mécanismes de participations aux bénéfices

Finalité : FP > SCR pour exercer

REASSURANCE

POSSIBLE EN VIE-NON VIE
- Proportionnelle
- Non-proportionnelle

1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
- Problématiques opérationnelles pour tarifer
- Résumé

2 Construction d'un zonier

3 Provisionnement

CONTRAT D'ASSURANCE ET TARIFICATION

Une police d'assurance est un contrat entre deux parties :

- → l'assuré, détenteur du contrat ;
- → l'assureur, pourvoyeur du contrat.

En échange de la couverture d'un risque par l'assureur, l'assuré verse une **prime** d'assurance.

En cas de sinistre, le bénéficiaire du contrat reçoit le montant contractuel prévu en cas de survenance du sinistre.

Ainsi le risque économique initialement supporté par l'assuré est transféré vers l'assureur.

La mutualisation induite par la souscription de nombreux contrats au sein d'une compagnie d'assurance permet l'utilisation grossière de la **loi des grands nombres**.

En effet,

- → un portefeuille d'assurance couvre un risque en particulier : les pertes sont considérées être de même loi de probabilité... ⇒ Tarification par garantie !
- → les contrats sont a priori indépendants les uns des autres.

Ces propriétés doivent permettre à l'assureur de **prédire avec une précision relative** les pertes encourues pour une période donnée.

Soit un portefeuille d'assurance contenant I polices. Notons la loi du $i^{\text{ème}}$ contrat S_i (perte), et la loi des pertes agrégées S_I.

La LFGN stipule la CV presque sûre de la moyenne empirique de pertes i.i.d., notée $\bar{S}_I = \frac{1}{I}\sum_{i=1}^{I} S_i$, vers l'espérance de la loi :

$$\boxed{\bar{S}_I \xrightarrow[n\to\infty]{p.s.} \mathbb{E}[S_i] = \mu.}$$

Ou encore : $\mathbb{P}\left(\lim_{I\to\infty} \bar{S}_I = \mu\right) = 1$.

Ce résultat est à l'origine du **principe général de tarification** : la prime vaut au moins μ, aussi appelée prime pure du contrat. C'est cette prime que nous modéliserons.

DANGERS D'UNE MAUVAISE TARIFICATION

Se tromper dans la tarification d'un produit peut avoir plusieurs conséquences dommageables :

- comme cela est souvent lié à la segmentation, il y a un risque de composition du portefeuille (bons et mauvais risques) ;
- investir dans 1 politique de vente (marketing, ...) mal adaptée ;
- impact néfaste sur la concurrence, déficit d'image ;

- mauvaise évaluation de la marge de risque, et donc in fine du provisionnement : (pour rappel, $S_I = \sum_i S_i$)

$$VaR_\alpha(S_I) = \inf\{s \in \mathbb{R}^+ : \ \mathbb{P}(S_I > s) \leq (1 - \alpha)\}$$

RESTRICTIONS DUES A LA REGLEMENTATION

→ La législation a également un impact en termes de segmentation et de tarification.

L'exemple récent le plus célèbre (pas le cas en provisionnement) : *Primes unisexe* : "Les compagnies d'assurances ne pourront plus, à partir du 21 décembre 2012, prendre en considération le critère du sexe pour calculer les primes et prestations d'assurances dans leurs contrats." a jugé la Cour de justice de l'UE.

→ Explication du tarif en assurance : directive DDA (distribution en assurance : éclairage notamment sur marges / commissions).

PRIME COMMERCIALE

En pratique l'assureur applique des chargements à cette prime, car mathématiquement sa ruine est certaine à horizon infini dès lors que la tarification respecte le strict principe d'équivalence.

La prime d'assurance Π_i se décompose donc en $+$sieurs parties :

- $\rightarrow$ la prime technique (provisions techniques dans le bilan économique SII) : comporte la **prime pure** (modèles vus ici) $\mathbb{E}[S_i]+$ chargements techniques ; où les chargements techniques sont issus des principes de prime (cf plus loin).
- $\rightarrow$ la prime d'inventaire composée de la prime technique plus les frais :
  - d'acquisition,
  - d'administration et gestion du contrat,

→ la prime commerciale (prime finale) intègre à la prime d'inventaire la rémunération d'intermédiaires (courtiers, ...).

La stratégie de la compagnie peut également jouer sur la hauteur de ces chargements.

Objectif de l'assureur :

Mettre en place une tarification segmentée tout en conservant le principe de mutualisation.

En effet, nous savons que

→ $E[S] = E[E[S \mid X]]$

→ ce qui se dérive empiriquement $\frac{1}{n}\sum_i S_i \quad \sim \quad \frac{1}{n}\sum_i \pi(X_i)$

MARGE POUR RISQUE (RM)

C'est une notion **différente du chargement technique**.

Elle dépend du risque couvert, et n'entre pas dans le tarif. En revanche, elle fait partie des provisions techniques (BE + RM).

Elle représente le coût du capital appliqué aux flux de SCR futurs actualisé :

$$RM = CoC \times \sum_{t=1}^{T} \frac{SCR_t}{(1+r)^t}.$$

Rq : RM = coût d'immobilisation du capital pr l'activité ($CoC \simeq 6\%$), ou coût de portage du risque (ex : lors d'un rachat du portefeuille).

1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

PRINCIPE DE L'ESPERANCE MATHEMATIQUE

Notons Π la prime, S le montant cumulé des sinistres de la police.

Le principe de la prime pure donne $\boxed{\Pi(S) = \mathbb{E}[S].}$

Le principe de l'espérance mathématique donne

$$\boxed{\Pi(S) = (1+\beta)\,\mathbb{E}[S], \qquad \beta > 0.}$$

$\rightarrow$ Chargement très simple, mais n'apporte aucune information sur les fluctuations de S autour de sa moyenne...

Difficulté de ce principe : choix de β.

Remarque : pour des risques dégénérés ($\mathbb{P}(S = s) = 1$), on devrait avoir $\Pi(S) = s$ ce qui n'est pas vrai ici.

Pour évaluer son risque de perte, l'assureur peut utiliser la théorie des grandes déviations et le **lemme de Chernoff**.

Lemme. (Chernoff). Soient $S_1, S_2, ..., S_n$ des v.a.p. indépendantes et de même loi que S telles que $\mathbb{E}[e^{tS}] < \infty$ pour un $t > 0$. Posons $X_i = S_i - (1+\beta)\mathbb{E}[S_i]$. Alors

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq 0\right) \leq \rho^n \quad \text{et} \quad \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq 0\right) = \log \rho,$$

où $\rho = \inf_t M_X(t) < 1$ et $M_X(t) = \exp(-t(1+\beta)\mathbb{E}[S_i])\, M_S(t)$.

Preuve. En utilisant l'inégalité de Bienaymé-Tchebischev, on déduit

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq 0\right) = \mathbb{P}\left(e^{t\sum_{i=1}^{n} X_i} \geq 1\right) \leq \mathbb{E}[e^{t\sum_{i=1}^{n} X_i}] = (\mathbb{E}[e^{tX_i}])^n = (M_X(t))^n.$$

L'inégalité est vraie $\forall t > 0$, donc en particulier pour celui qui vérifie le minimum de $M_X(t)$.

Remarque :

- La dérivée $M_X'(0)$ est négative car $\mathbb{E}[X_i] = -\beta\,\mathbb{E}[S_i] < 0$, alors même que $M_X(0) = 1$.
- D'autre part, $\mathbb{P}(X_i > 0) > 0$, donc $\lim_{t\to\infty} M_X(t) = +\infty$.

D'où l'existence d'un minimum < 1 (théo. valeurs intermédiaires).

Ainsi, si l'assureur souhaite **majorer par ϵ la probabilité d'un résultat négatif** sur la période, donc

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq 0\right) \leq \epsilon,$$

il choisira β tel que $\boxed{\rho^n(\beta) = \epsilon}$ (ex : $S \sim \mathcal{E}(\lambda) \Rightarrow \rho = e^{-\beta}(1+\beta)$).

PRINCIPE DE LA VARIANCE

Le principe de la variance donne

$$\boxed{\Pi(S) = \mathbb{E}[S] + \beta\, Var(S), \qquad \beta > 0.}$$

Inconvénient : symétrie par rapport à l'espérance.

$\rightarrow$ On comptabilise les valeurs négatives de la v.a. $(S - \mathbb{E}[S])$, pourtant favorables à l'assureur.

Conséquence : on augmente trop les chargements techniques.

i) Application du principe à la réassurance proportionnelle.

Cherche une couverture pour une prop. $\lambda \in [0,1]$ du risque S :

$$\Pi(\lambda S) = \mathbb{E}[\lambda S] + \beta\, Var(\lambda S) = \lambda \mathbb{E}[S] + \lambda^2 \beta\, Var(S) < \lambda \Pi(S).$$

Donc l'assuré aurait intérêt à **diviser son risque initial** en n parties égales car il paierait moins cher : en effet,

$$n\,\Pi\left(\frac{S}{n}\right) < \Pi(S).$$

ii) Principe de la variance et agrégation de risques indépendants.

Si on considère deux risques indépendants S_1 et S_2, on a

$$\boxed{\Pi(S_1 + S_2) = \Pi(S_1) + \Pi(S_2)},$$

ce qui implique que l'accumulation de risques indépendants ne conduit pas au principe de diversification.

PRINCIPE DE L'ECART-TYPE

Le principe de l'écart-type donne

$$\boxed{\Pi(S) = \mathbb{E}[S] + \beta\,\sigma(S), \qquad \beta > 0.}$$

A l'inverse, le découpage du risque ici ne conduit pas à une diminution de la prime :

$$n\,\Pi\left(\frac{S}{n}\right) = \Pi(S).$$

PRINCIPE EXPONENTIEL

Le principe exponentiel donne

$$\boxed{\Pi(S) = \frac{1}{\alpha} \ln(\mathbb{E}[e^{\alpha S}]).}$$

Le paramètre α est appelé **coefficient d'aversion au risque**.

D'après l'inégalité de Jensen, la prime technique est supérieure à la prime pure :

$$\Pi(S) \geq \mathbb{E}[S].$$

En effet, si α est proche de 0, en utilisant les propriétés de la transformée de Laplace :

$$\begin{aligned}
\Pi(S) &= \frac{1}{\alpha}\ln\left(1+\alpha\mathbb{E}[S]+\frac{\alpha^2}{2}\mathbb{E}[S^2]+o(\alpha^2)\right)\\
&= \frac{1}{\alpha}\left(\alpha\mathbb{E}[S]+\frac{\alpha^2}{2}\mathbb{E}[S^2]\right)-\frac{1}{2\alpha}\left(\alpha\mathbb{E}[S]+\frac{\alpha^2}{2}\mathbb{E}[S]\right)^2+o(\alpha)\\
&= \mathbb{E}[S]+\frac{\alpha}{2}\,Var(S)+o(\alpha)
\end{aligned}$$

On retrouve le principe de la variance...

Si

- $\alpha \to 0$: principe de la prime pure ;
- $\alpha \to \infty$: principe de la perte maximale,

$$\Pi(S) \to \sup\{s\ :\ \mathbb{P}(S<s)<1\} = r_s.$$

PRINCIPE D'ESSCHER

Le principe d'Esscher préconise de choisir une prime égale à

$$\boxed{\Pi(S) = \frac{\mathbb{E}[Se^{\alpha S}]}{\mathbb{E}[e^{\alpha S}]}.}$$

On peut montrer que $\Pi(S) \geq \mathbb{E}[S]$ puisque $Cov(S, e^{\alpha S}) \geq 0$.

Cette prime est l'espérance mathématique calculée avec la nouvelle f.d.r. G définie par

$$dG(x) = \frac{e^{\alpha x} dF_S(x)}{\int_0^\infty e^{\alpha x} dF_S(x)},$$

qui est la transformée d'Esscher de F_S.

PRINCIPE DE WANG (Proportional hazard transform)

Le principe de Wang s'appuie sur la définition

$$\boxed{\Pi(S) = \int_0^\infty (\bar{F}_S(x))^r \, dx,}$$

où $\bar{F}_S = 1 - F_S$ (survie), et $r \in [0, 1]$. On a $\Pi(S) \geq \mathbb{E}[S]$.

Ce principe est très utilisé en réassurance.

En effet, la transformée de Wang permet de calculer très simplement les primes des traités en **excédent de sinistre**.

Par exemple, pour un traité (noté dans la pratique : hXSa)

- de priorité a,
- de portée h,

on a :

$$hXSa = \begin{cases} 0 & \text{si} \quad 0 \leq S \leq a \\ S - a & \text{si} \quad a \leq S \leq a + h \\ h & \text{si} \quad a + h \leq S \end{cases}$$

La prime vaut

$$\boxed{\Pi(hXSa) = \int_0^h (\bar{F}_S(x + a))^r \, dx = \int_a^{a+h} (\bar{F}_S(x))^r \, dx.}$$

PRINCIPE DU FRACTILE

Dans le principe du fractile, on adopte la prime Π qui vérifie

$$\boxed{\Pi(S) = \inf\left(p \,|\, F_S(p) \geq 1 - \epsilon\right) = \inf\left(p \,|\, \mathbb{P}(S > p) \leq \epsilon\right).}$$

C'est donc la plus petite prime telle que la probabilité que le sinistre dépasse la prime est au plus de ϵ.

Par exemple,

- $\rightarrow$ si $\epsilon = 1/2$, alors la prime est la médiane de la distribution ;
- $\rightarrow$ si $\epsilon = 0$, alors la prime suit le principe de la perte maximale.

PROPRIETES SOUHAITABLES DES PRINCIPES

Un assureur utilisant une mesure de risque donnée attend d'elle un ensemble de propriétés "naturelles" censées refléter la réalité...

1. La prime vaut au moins la prime pure : $\Pi(S) \geq \mathbb{E}[S]$.

   On peut ajouter que si $\mathbb{P}(S = s) = 1$, alors $\Pi(S) = s$.
   Ceci implique qu'il n'y ait pas de chargement injustifié. *Parfois, le chargement peut même être négatif suivant les conditions de marché (concurrence, ...).*

2. Invariance par translation : $\Pi(S + c) = c + \Pi(S),\ \forall c \geq 0$.

   c est une constante, et en particulier $\Pi(0) = 0$.
   Tout risque déterministe est tarifé à sa propre valeur.

3. Additivité : $\Pi(S_1 + S_2) = \Pi(S_1) + \Pi(S_2)$,

si S_1 et S_2 sont indépendants.

Cependant, cette propriété ne vérifie pas le principe de diversification des risques. On lui préfère la propriété

$$\Pi(S_1 + S_2) \leq \Pi(S_1) + \Pi(S_2).$$

Rappelons au passage que le principe de la variance est additif, alors que celui de l'écart-type est sous-additif.

Cette propriété induit un **gain de diversification**, qui profite

- + soit à l'assuré (prime plus faible),
- + soit à l'assureur (probabilité de ruine moins élevée).

4. Homogénéité : $\Pi(\lambda S) = \lambda\,\Pi(S),\ \forall \lambda \geq 0$.

   $\Rightarrow$ invariance par changement de numéraire, elle est essentielle pour la **réassurance proportionnelle**.

   Propriété remise en cause par quelques auteurs lorsque λ est grand ($\Pi(\lambda S) > \lambda\,\Pi(S)$).

5. Itérativité : $\Pi(S_1) = \Pi(\Pi(S_1 \mid S_2))$.

   On peut calculer la prime du risque S_1 en deux étapes :

   - $\rightarrow$ on applique d'abord la prime Π à la distribution de S_1 conditionnelle à S_2 ;
   - $\rightarrow$ on obtient une v.a.r., fonction de S_2, à laquelle on applique de nouveau le principe de prime.

Exemple.
Le nombre annuel d'accidents d'un chauffeur est modélisé par une loi de Poisson $\mathcal{P}(\lambda)$. Le profil de risque λ est inconnu et différent pour chaque chauffeur, donc la réalisation d'une v.a.r. Λ. La loi du nombre d'accidents conditionnelle à $\Lambda = \lambda$ est de Poisson, et si $\Lambda \sim \mathcal{G}amma$ alors la loi est une binomiale négative.

6. Convexité : $\Pi(\lambda S_1 + (1-\lambda)S_2) \leq \lambda\,\Pi(S_1) + (1-\lambda)\Pi(S_2)$,

   $\forall \lambda \in [0,1]$ et S_1, S_2.

   Cette propriété est utile pour la recherche de décisions optimales dans le choix de contrat d'assurance ou de réassurance.

RESUME DES PROPRIETES DES PRINCIPES

Principes	Propriétés Prime pure	Trans.	Addit.	Itérat.	Homog.
Prime pure	+	+	+	+	+
Espérance	+	–	+	–	+
Variance	+	+	+	–	–
Ecart-type	+	+	–	–	+
Exponentiel	+	+	+	+	–
Utilité	+	+	*e*	*e*	–
Valeur moyenne	+	*e*	*e*	+	–
Esscher	+	+	+	–	–
Fractile	+	+	+	+	–

+ : la propriété est vérifiée ; – : la propriété n'est pas vérifiée ;
e : vérifiée en considérant les fonctions *u* et *f* qui nous permettent de retomber sur les principes exponentiel et prime pure.

1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

PRINCIPE DE PARTAGE DE VARIANCE DU RISQUE

Source : A. Charpentier.

Aucune segmentation, aucun transfert de risque.
→ Tout la partie risquée (contenu dans la variance) est conservée par l'assureur.

No risk classification, identical premium

	Insured	Insurer
Loss	$\mathbb{E}[S]$	$S - \mathbb{E}[S]$
Average Loss	$\mathbb{E}[S]$	0
Variance	0	$\text{Var}[S]$

SEGMENTATION ET INFORMATION COMPLETE

Source : A. Charpentier.

Information complète sur les facteurs de risque.

Perfect classification, (ultra) personalized premium

	Insured	Insurer
Loss	$\mathbb{E}[S\|\mathbf{\Omega}]$	$S-\mathbb{E}[S\|\mathbf{\Omega}]$
Average Loss	$\mathbb{E}[S]$	0
Variance	$\text{Var}\Big[\mathbb{E}[S\|\mathbf{\Omega}]\Big]$	$\text{Var}\Big[S-\mathbb{E}[S\|\mathbf{\Omega}]\Big]$

$$\text{Var}[S]=\underbrace{\mathbb{E}\Big[\text{Var}[S|\mathbf{\Omega}]\Big]}_{\to\text{insurer}}+\underbrace{\text{Var}\Big[\mathbb{E}[S|\mathbf{\Omega}]\Big]}_{\to\text{insured}}.$$

Car $Var(S - E[S\,|\,\Omega]) = E[Var(S\,|\,\Omega)]$

SEGMENTATION ET INFORMATION INCOMPLETE

Source : A. Charpentier.

Imperfect classification, personalized premium

	Insured	Insurer
Loss	$\mathbb{E}[S\|\boldsymbol{X}]$	$S-\mathbb{E}[S\|\boldsymbol{X}]$
Average Loss	$\mathbb{E}[S]$	0
Variance	$\text{Var}\Big[\mathbb{E}[S\|\boldsymbol{X}]\Big]$	$\mathbb{E}\Big[\text{Var}[S\|\boldsymbol{X}]\Big]$

$$\begin{aligned}\text{Var}[S] &= \mathbb{E}\Big[\text{Var}[S|\boldsymbol{X}]\Big]+\text{Var}\Big[\mathbb{E}[S|\boldsymbol{X}]\Big]\\ &= \underbrace{\underbrace{\mathbb{E}\Big[\text{Var}[S|\boldsymbol{\Omega}]\Big]}_{\text{pooling}}+\underbrace{\mathbb{E}\Big[\text{Var}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big]}_{\text{solidarity}}}_{\rightarrow\text{insurer}}+\underbrace{\text{Var}\Big[\mathbb{E}[S|\boldsymbol{X}]\Big]}_{\rightarrow\text{insured}}.\end{aligned}$$

1 Tarification a priori - concepts avancés
- Introduction
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
- Résumé

2 Construction d'un zonier

3 Provisionnement

APPROCHE INDEMNITAIRE

Idée : coûts dépendent de l'occurrence éventuelle d'un sinistre (**au plus un sinistre dans la période**) et du montant qui en résulte.

$$\boxed{S_i = \begin{cases} b & \text{si } I_i = 1 \\ 0 & \text{si } I_i = 0 \end{cases}}$$

où $I_i \sim$ Bernouilli $\mathcal{B}(p_i)$ (occurrence du sinistre), et b déterministe.

$\rightarrow \mathbb{E}[S_i] = \mathbb{E}[I_i] \times b$

$\rightarrow Var(S_i) = Var(I_i) \times b^2$

Exemple : coût en sinistre d'un contrat d'assurance vie sur un an.

APPROCHE FORFAITAIRE

Idée : S_i est définie par 2 composantes. Une **masse en 0**, et une **composante continue** pour le coût si un sinistre survient.

$$\boxed{S_i = \begin{cases} Y & \text{si } I_i = 1, \\ 0 & \text{si } I_i = 0 \end{cases}}$$

où $I_i \sim$ Bernouilli $\mathcal{B}(p_i)$ (occurrence du sinistre), et $Y \perp\!\!\!\perp I_i$.

$\rightarrow \mathbb{E}[S_i] = p_i\, \mathbb{E}[Y], \quad Var(S_i) = \mathbb{E}[I_i] Var(Y) + Var(I_i)\mathbb{E}[Y]^2$

$\rightarrow F_{S_i}(s) = q_i + p_i F_Y(s) \quad (s \geq 0).$

$\rightarrow M_{S_i}(t) = M_{I_i}(\ln(M_Y(t)))$

Exemple : le coût en sinistres pour le contrat santé i sur un an.

APPROCHE FREQUENCE - COUT MOYEN

La $+$ souvent utilisée en IARD.

Idée : S_i est fonction de 2 aléas, N_i et Y_k, respectivement le nombre de sinistres et les montants unitaires associés.

$$\boxed{S_i = \begin{cases} \sum_{k=1}^{N_i} Y_{ik} & \text{si } N_i > 0, \\ 0 & \text{si } N_i = 0 \end{cases}}$$

où N_i est une v.a. discrète, N_i **et** Y_{ik} **sont** $\perp\!\!\!\perp$ **et les** Y_{ik} **sont i.i.d**.

$\rightarrow \mathbb{E}[S_i] = \mathbb{E}[N_i]\,\mathbb{E}[Y_{ik}]$,

$\rightarrow Var(S_i) = \mathbb{E}[N_i]Var(Y_{ik}) + Var(N_i)\mathbb{E}[Y_{ik}]^2$

$\rightarrow M_{S_i}(t) = M_{N_i}(\ln(M_{Y_{ik}}(t)))$

$\rightarrow F_{S_i}(s) = \mathbb{P}(N_i = 0) + \sum_{m=1}^{\infty} \underbrace{F_{Y_{i1}+\ldots+Y_{im}}(s)}_{\text{inconnu}} \times \mathbb{P}(N_i = m)$

MODELE CLASSIQUE DE PRIME PURE

Soit S_i la somme annuelle des sinistres du contrat i.
Le nb N_i de sinistres est une v.a. considérée $\perp\!\!\!\perp$ des coûts Y_{ik}, eux-même i.i.d. :

$$S_i = \begin{cases} 0 & \text{si } N_i = 0 \\ Y_{i1} + \ldots + Y_{in} & \text{si } N_i = n. \end{cases} \quad \Leftrightarrow \quad S_i = \sum_{k=1}^{N_i} Y_{ik}$$

Ainsi, $\mathbb{E}_{\mathbb{P}}[S_i] = \mathbb{E}_{\mathbb{P}}[N_i] \times \mathbb{E}_{\mathbb{P}}[Y_{ik}]$.

En réalité, N_i est souvent **conditionnellement** $\perp\!\!\!\perp$ à Y_i, donc

$$\mathbb{E}_{\mathbb{P}}[S_i \,|\, \mathcal{X}_i] = \mathbb{E}_{\mathbb{P}}[N_i \,|\, \mathcal{X}_i] \,.\, \mathbb{E}_{\mathbb{P}}[Y_{ik} \,|\, \mathcal{X}_i],$$

où $\mathcal{X}_i$ est un ensemble d'informations.

APERCU D'UNE BASE DE DONNEES

```
> head(myData, n=16)
```

	PERMIS	ACV	SEX	STATUT	CSP	USAGE	AGECOND	...	GARAGE	CHARGE
1	245	10	F	C	50	2	40	...	3	0
2	348	10	F	A	50	1	63	...	3	0
3	16	10	F	C	26	2	20	...	3	0
4	291	10	F	A	50	1	56	...	3	0
5	123	10	F	A	50	1	29	...	3	0
6	295	10	F	A	37	1	43	...	3	0
7	24	10	F	A	50	2	21	...	3	0
8	181	9	F	A	50	3	35	...	3	0
9	157	10	M	C	55	1	31	...	3	0
10	338	10	M	C	1	2	48	...	2	179
11	20	10	M	C	26	2	19	...	3	0
12	208	10	F	A	50	2	39	...	3	0
13	127	10	F	A	37	1	29	...	1	0
14	93	7	F	C	50	2	39	...	3	0
15	134	10	F	A	50	1	36	...	3	0
16	416	10	F	C	50	1	60	...	3	0

Le principe de la tarification est d'approcher $\mathcal{X}$ par un **proxy** (variables tarifaires).
Ce proxy correspond aux info. indiv. $\rightarrow$ variables explicatives :

$$\Rightarrow \text{c'est le contexte des modèles de régression.}$$

Supposons que l'assureur dispose de J facteurs explicatifs du risque, notés $\{X_1, \ldots, X_J\}$, on obtient alors la formule

$$\mathbb{E}_{\mathbb{P}}[S \,|\, X_1, \ldots, X_J] = \mathbb{E}_{\mathbb{P}}[N \,|\, X_1, \ldots, X_J] \,.\, \mathbb{E}_{\mathbb{P}}[Y \,|\, X_1, \ldots, X_J].$$

Le problème est donc d'obtenir (tarification a priori, VS a posteriori en crédibilité)

- $\mathbb{E}_{\mathbb{P}}[N \,|\, X_1, \ldots, X_J]$: estimation de la loi de N.
- $\mathbb{E}_{\mathbb{P}}[Y \,|\, X_1, \ldots, X_J]$: idem.

En économétrie, on cherche à estimer $\mathbb{E}_{\mathbb{P}}[Z \,|\, X_1, \ldots, X_J]$ par une fonction des facteurs explicatifs notée $\Phi(X_1, \ldots, X_J)$.

En économétrie **linéaire**, on a coutûme de supposer que

$$Z \,|\, X_1, \ldots, X_J \quad \sim \quad \mathcal{N}(\beta_0 + \beta_1 X_1 + \ldots + \beta_J X_J,\ \sigma^2).$$

En notant $\mathbf{X} = (1, X_1, \ldots, X_J)^T$ le **vecteur des facteurs de risque** et $\beta = (\beta_0, \beta_1, \ldots, \beta_J)^T$ les **coefficients** de régression, on peut simplifier cette écriture sous forme matricielle :

$$Z \,|\, \mathbf{X} \quad \sim \quad \mathcal{N}(\mathbf{X}^T \beta,\ \sigma^2).$$

Problème : le modèle linéaire est rarement adapté en assurance...
Alternative : besoin de supposer relations non-linéaires $\Rightarrow$ GLM.

ETAPES STATISTIQUES DE TARIFICATION

1. Statistiques descriptives univariées et bivariées ;
2. Modélisation de la fréquence par un GLM adapté (choix d'une loi pour la réponse, intégration des covariables), cela donne

$$\mathbb{E}[N \mid \mathbf{X}] = f_1(\mathbf{X}\beta)$$

3. Modélisation du coût par un autre GLM adapté, on obtient

$$\mathbb{E}[Y \mid \mathbf{X}^{'}] = f_2(\mathbf{X}^{'}\beta)$$

4. Synthèse pour en déduire la prime (pure) :

$$\mathbb{E}[S_i \mid \mathbf{X}, \mathbf{X}^{'}] = E[N \mid \mathbf{X}] \times E[Y \mid \mathbf{X}^{'}]$$

PROPAGATION D'ERREUR ?

En construisant deux modèles (1 pour la fréquence et 1 pour la sévérité), on prend le risque de propager des erreurs...

Parfois il vaut mieux essayer de construire un unique modèle qui rende compte à la fois de la fréquence et de la sévérité : cela **dépend de la qualité d'adéquation de la loi de fréquence notamment**.

En réalité dans cette ultime approche, on perd l'info sur le nb de sinistres et on s'intéresse à la charge totale par contrat. La masse en 0 (contrats non-sinistrés) induit des difficultés de calibration, ce qui explique la décomposition fréquence - coût moyen en pratique.

1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

EXEMPLES CLASSIQUES D'APPLICATION

L'usage des GLM est ancré depuis longtemps dans les moeurs.
On peut citer parmi les domaines concernés :

- **assurance santé** : remboursements soins, frais d'hospitalisation ;
- **assurance auto** / **moto** : dommages matériels, vol, ... ;
- **assurance Multi-Risques Habitation** (MRH) : incendie, vol, dégâts des eaux, ...
- **assurance Responsabilité Civile** (RC) : dommages à autrui.

Les cas de la RC, de l'assurance CATNAT et de la réass. IARD sont un peu $\neq$ car font intervenir des montants CAT en général.

APPLICATIONS EN VIE

On se sert aussi des GLM en Vie, notamment en

- **épargne** : essentiellement du risque comportemental sur les produits en taux garantis (euro) ou non (UC) ;
- **prévoyance** : DC, LTC (Long-Term Care : dépendance), CI (Critical Illness : maladies redoutées), incap/inval. ;
- **réassurance vie** : même remarque qu'en non vie.

Remarque : de par la nature des contrats, il y a souvent une dimension temporelle dans la modélisation en Vie qui $\nexists$ en non-vie $\rightarrow$ modèles de durée.

RISQUE DE LONGEVITE [Lee and Carter, 1992]

C'est le **modèle le plus utilisé en mortalité** (longévité) :

$$log(\mu_x(t)) = \alpha_x + \beta_x \kappa(t) + \epsilon_x(t)$$

- x est l'âge, t l'année ;
- $\mu_x(t)$ est le taux de mortalité instantané l'année t à l'âge x ;
- α_x : structure de la mortalité en fonction de l'âge ;
- $\kappa(t)$: vitesse d'amélioration de la mortalité (série temp.) ;
- β_x : la vitesse d'amélioration a des impacts $\neq$ selon l'âge ;
- les résidus $\epsilon_x(t) \sim \mathcal{N}(0, \sigma^2)$.

RISQUE DE MORTALITE : MODELE DE BRASS [Brass, 1964], [Brass and Macrae, 1984]

C'est un **modèle relationnel** basé sur la régression logistique :

$$ln\left(\frac{q^{exp}(x,t)}{1-q^{exp}(x,t)}\right) = a + b \times ln\left(\frac{q^{ref}(x,t)}{1-q^{ref}(x,t)}\right)$$

où

- x est l'âge de la personne, t est le facteur temporel,
- q^{ref} est une table de mortalité de référence,
- q^{exp} est la table de mortalité d'expérience.

Calibre les coef. (a, b) pour établir le passage d'1 table à l'autre, par ex. d'une population nationale à une population d'assurés.

INTERET DES GLM

Les GLM permettent de

- modéliser des réponses diverses $\in \mathbb{R}, \mathbb{R}^+, \mathbb{N}, [0,1], ...$;
- intégrer toute type d'information exogène susceptible d'influer sur la variable dépendante (réponse Y),
- quantifier l'impact des facteurs de risque $\mathbf{X}$ (sens/intensité),
- résidus hétéroscédastiques (la loi varie par profil).

Ils nécessitent d'introduire deux hypothèses fondamentales :

- les individus Y_i sont $\perp\!\!\!\perp$ entre eux (rq : si les indiv. étaient corrélés, cela résulterait aussi à avoir – d'indiv., donc $n \searrow$) ;
- les variables explicatives $\mathbf{X}$ sont $\perp\!\!\!\perp$ deux à deux.

POURQUOI CES HYPOTHESES ?

Vision géométrique du modèle linéaire, voir l'article Econometrie et Machine Learning d'Antoine Ly et Arthur Charpentier pour expliquer la nécessité d'indépendance entre les X_j...

ATTENTION A LA NOTION DE CORRELATION

$\exists$ plusieurs mesures de dépendance, e.g. corrélation de rang (Kendall, Spearman). La $+$ répandu est Pearson,

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

où $\mu_X = E[X]$ et σ_X est l'écart-type de X.

Mesure la corrél. **linéaire**. En effet, considérons la v.a. X telle que $X \sim \mathcal{N}(0,1)$. Ainsi $\mu_X = 0$, et $\mu_{X^3} = 0$. Notons $Y = X^2$, on a

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(X^2-\mu_{X^2})]}{\sigma_X \sigma_{X^2}} = \frac{\mu_{X^3} - \mu_X \mu_{X^2}}{\sigma_X \sigma_{X^2}} = 0.$$

Corrélation nulle alors que X et X^2 parfaitement corrélées !

COMPOSANTS D'UN GLM [McCullagh and Nelder, 1989]

Pour l'individu i...

1. La loi de la réponse aléatoire Y_i : par hyp. elle $\in$ à une *distribution de la famille exponentielle.*

2. Le prédicteur $\eta_i = \sum_{j=1}^{J} \beta_j X_{ij}$, linéaire et déterministe : *les facteurs de risque explicatifs le constituent.*

3. La fonction de lien g : monotone, dérivable, inversible. En pratique, n'importe quelle FdR, t.q.

$$g(\mathbb{E}[Y_i \,|\, \mathbf{X}_i]) = \eta_i.$$

Ex. du modèle linéaire : $g = Id \quad \eta_i = \sum_{j=1}^{J} \beta_j X_{ij} \quad Y_i \sim \mathcal{N}(\eta_i, \sigma^2)$.

LOI DE L'ERREUR / FONCTION DE LIEN

Adapter le lien en fonction du domaine de définition de Y.

Loi	Lien naturel	Moyenne	Utilisation
$\mathcal{N}(\mu, \sigma^2)$	Id : $\eta = \mu$	$\mu = X\beta$	Rég. lin.
$\mathcal{B}(\mu)$	logit : $\eta = ln(\frac{\mu}{1-\mu})$	$\mu = \frac{exp(X\beta)}{1+exp(X\beta)}$	Taux
$\mathcal{P}(\mu)$	log : $\eta = ln(\mu)$	$\mu = exp(X\beta)$	Fréquence
$\mathcal{G}(\alpha, \beta)$	inverse : $\eta = \frac{1}{\mu}$	$\mu = (X\beta)^{-1}$	Sévérité
$\mathcal{IN}(\mu, \lambda)$	inverse^2 : $\eta = -\frac{1}{\mu^2}$	$\mu = (X\beta)^{-2}$	Sévérité

COEFFICIENTS ESTIMES ET IMPACTS

En général, on interprète les résultats de la manière suivante :

- $\hat{\beta}_j > 0$: $\nearrow$ du facteur de risque X_j provoque $\nearrow$ de $g(\mathbb{E}[Y \mid \mathbf{X}_i])$;
- $\hat{\beta}_j < 0$: $\nearrow$ du facteur de risque X_j provoque $\searrow$ de $g(\mathbb{E}[Y \mid \mathbf{X}_i])$;
- $\hat{\beta}_j = 0$: effet nul de la variation dudit X_j.

Evidemment, cela **dépend aussi du type de modélisation** !

- Pour des modèles à effets additifs, la valeur de réf. sera 0 ;
- Pour des modèles multiplicatifs, la valeur de référence sera 1 (à une transformation près parfois, cf modèle log-Poisson).

Pour connaitre le type d'effet, on réécrit le modèle sous la forme

$$\mathbb{E}[Y \mid \mathbf{X}] = g^{-1}(\mathbf{X}^T \beta).$$

RAPPORT DE COTE (ODD-RATIO ou OR)

En souscrivant en ligne, vous pouvez par ex. avoir une idée de la calibration de certains assureurs pour certains facteurs de risque : comparer le tarif en faisant évoluer **1 seule** caractéristique (ex : âge, ancienneté du permis, couleur de la voiture, ...)

Cela correspond à l'**odd-ratio**, un rapport sur la quantité d'intérêt :

$$\frac{\mathbb{E}[Y \,|\, X_j = x_j + 1]}{\mathbb{E}[Y \,|\, X_j = x_j]} = h(\beta_j),$$

avec h une fonction à déterminer.

Exemple log-poisson : $Y \sim \mathcal{P}(\lambda)$, donc $\lambda = e^{\mathbf{X}^T\beta} \;\Rightarrow\; h(\beta_j) = e^{\beta_j}$.

VALIDATION D'UN GLM - ETAPES

1. Construction de 2 échantillons $\perp\!\!\!\perp$ par tirage aléatoire : un d'apprentissage (construction) et un de validation ;
2. Validation de la significat. globale du modèle (déviance, LRT) : déviance $2(\ln L(Y \mid Y) - \ln L(\hat{\mu} \mid Y)) \sim \chi^2(n - p - 1)$
3. Validation de la significativité des coef. de régression un à un ;
4. Résidus : homoscédasticité (pour un segment donné), doit être aléatoire (test des signes ? on ne connait pas la loi des résidus dans un cas général à cause du lien...) ;
5. Confrontation "modélisé / empirique" sur l'éch. de validation.

COMPARAISON MODELE - EXPERIENCE

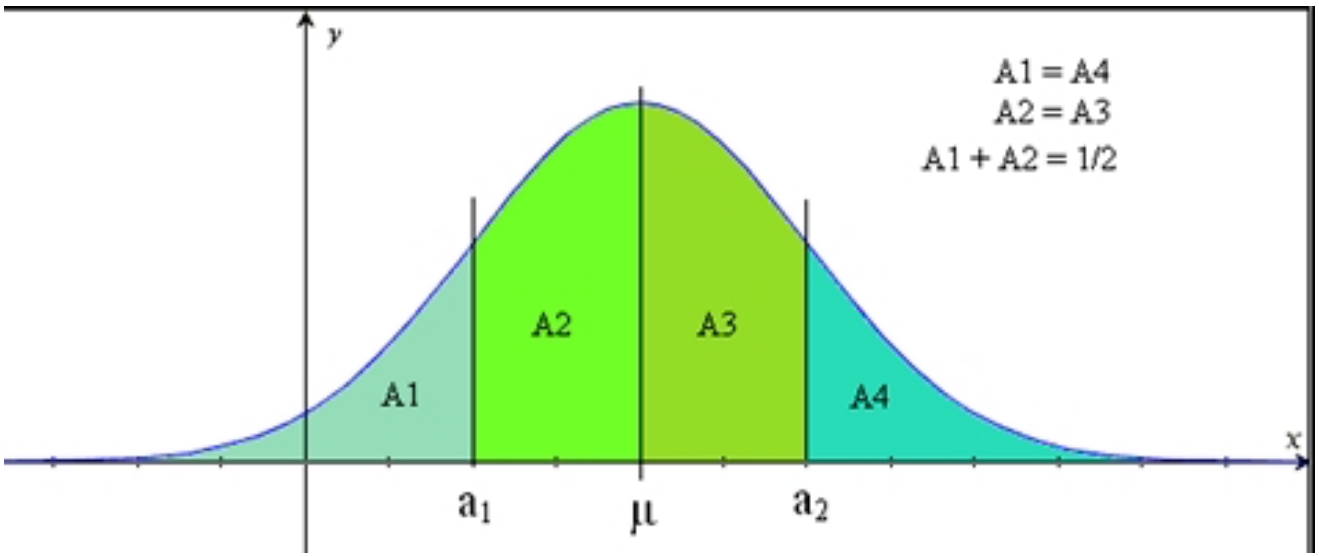
Pour le dernier point mentionné précédemment, on peut recourir par exemple à :

- Ex : indice de Gini à minimiser (aire), montants sinistres en fonction des primes, tout normalisé entre 0 et 1 par la transformation $\frac{x-min}{max-min}$).
- Q-Q plot (par ex. sévérité Gamma) doit se faire par segment car les lois sont conditionnelles...(si $Y|X \sim \mathcal{G}(\alpha, \lambda)$, alors Y n'est pas Gamma).

LIMITES DE LA GAUSSIENNE

L'utilisation d'une loi Normale est encore très répandue... Mais cela implique des erreurs fondamentales de raisonnement, notamment

- la densité de la loi est symétrique,
- sa queue de distribution est **fine**,
- support non adapté à des charges sinistres $\Rightarrow \mathbb{P}(Y < 0)$.



EFFETS DES FACTEURS DE RISQUE

Inutile de modéliser sans réflexion préalable sur les données...

En ce sens, il est essentiel de faire des **statistiques descriptives** afin de déterminer l'intérêt éventuel de

- discrétiser une variable continue : par des stats descriptives bivariées, par des arbres CART, par des modèles GAM (optimisation faite par méthode semi-paramétriques de lissage, par ex. les splines, cf [Pouna Siewe, 2010]), ...

- rendre continue une variable catégorielle (ordonnée) : si l'effet est monotone en fonction des modalités.

C'est la vision "ingénieur" couplée à la vision statisticien !

TRANSFORMATION DU PREDICTEUR ?

Il peut être utile d'**introduire une transfo. dans le prédicteur** sur certaines covariables en fonction du type d'impact sur Y.

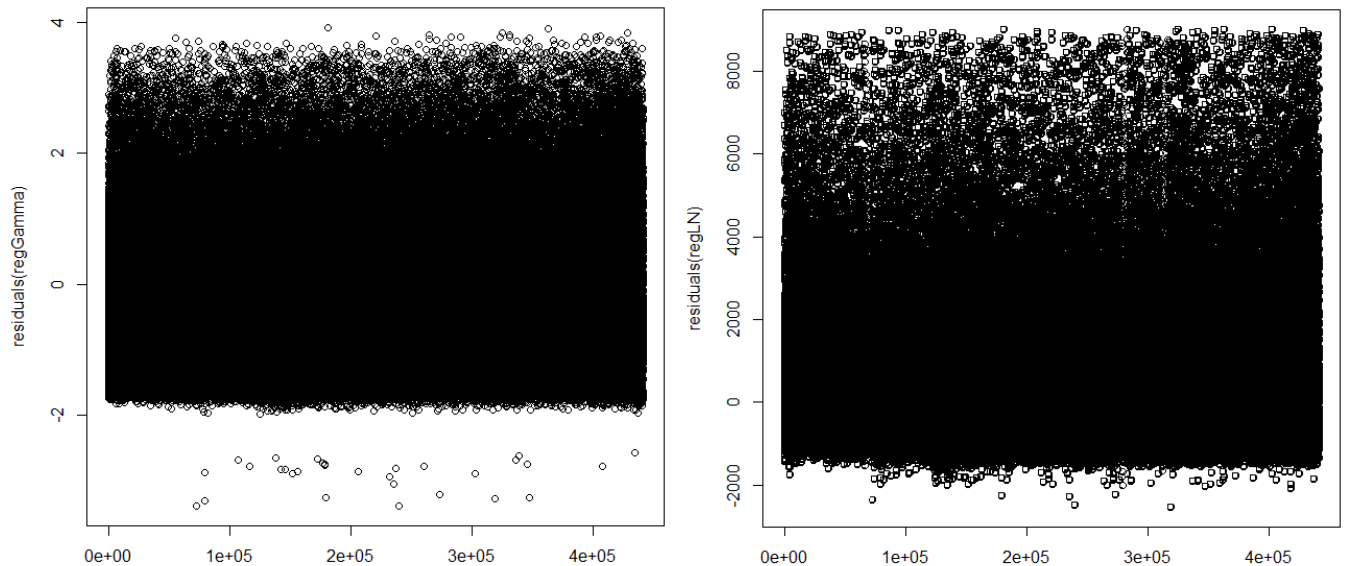
Cette transformation sera choisie en fonction de l'effet du facteur de risque sur Y lors de la visualisation des statistiques desc.

Prenons un ex. concret : supposons que l'âge x a un impact exponentiel sur le taux de mortalité q_x, mais que la CSP joue de manière linéaire. Ainsi on posera un modèle de la forme

$$ln(q_x) = a + b\,x + ln(c\,CSP) \;\Leftrightarrow\; q_x = A \times \exp(bx) \times c\,CSP$$

LES RESIDUS

L'exemple ci-dessous montre que le modèle Gamma est bien mieux adapté que le modèle lognormal dans cet exemple...



Dans le cas d'une loi continue (coût moyen), on peut tester ces résidus grâce au test des signes.

TWEEDIE or not TWEEDIE ? [Boucher and Danail, 2011]

La densité est donnée par

$$f(y;\mu,\phi) = a(y,\phi)\,\exp\left(\frac{1}{\phi}\Big[y\theta(\mu) - \kappa(\theta(\mu))\Big]\right),$$

$$\theta(\mu) = \begin{cases} \frac{\mu^{1-p}}{1-p} & \text{si } p \neq 1 \\ \log\mu & \text{si } p = 1 \end{cases} \qquad \kappa(\theta(\mu)) = \begin{cases} \frac{\mu^{2-p}}{2-p} & \text{si } p \neq 2 \\ \log\mu & \text{si } p = 2 \end{cases}$$

Dans cette formalisation, $\mathbb{E}[Y] = \mu$ et $Var(Y) = \psi\mu^p = \psi\mathbb{E}[Y]^p$, avec ψ un parametre de dispersion > 0.

L'ordre $p \in \mathbb{R}^+$ (*paramètre d'indice*), choisi (en fonction de l'application) avant d'estimer μ et ϕ, définit le **type de distribution** :

- $\rightarrow$ $p < 0$: réalisations dans $\mathbb{R}$; $p = 0$: loi gaussienne,
- $\rightarrow$ $0 < p < 1$: pas de distribution (pas de modèle Tweedie),
- $\rightarrow$ $p = 1$ avec $\phi = 1$: loi de Poisson,
- $\rightarrow$ $1 < p < 2$: loi composée Poisson-Gamma (réalisations ≥ 0),
- $\rightarrow$ $2 < p < 3$ ou $p > 3$: positive stable distributions ($x > 0$),
- $\rightarrow$ $p = 2$: loi Gamma, $p = 3$: loi inverse gaussienne.

En pratique, $1 < p < 2$ pour modéliser fréq. et coût en mm tps !
Inconvénient : mêmes var. explicatives prises en compte dans les lois de fréq. et de coût, or les praticiens savent qu'elles sont $\neq$.

1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

AVANTAGES

L'avantage essentiel des modèles non-paramétriques réside dans la flexibilité de la forme de dépendance entre la réponse Y et les facteurs de risque $\mathbf{X}$.

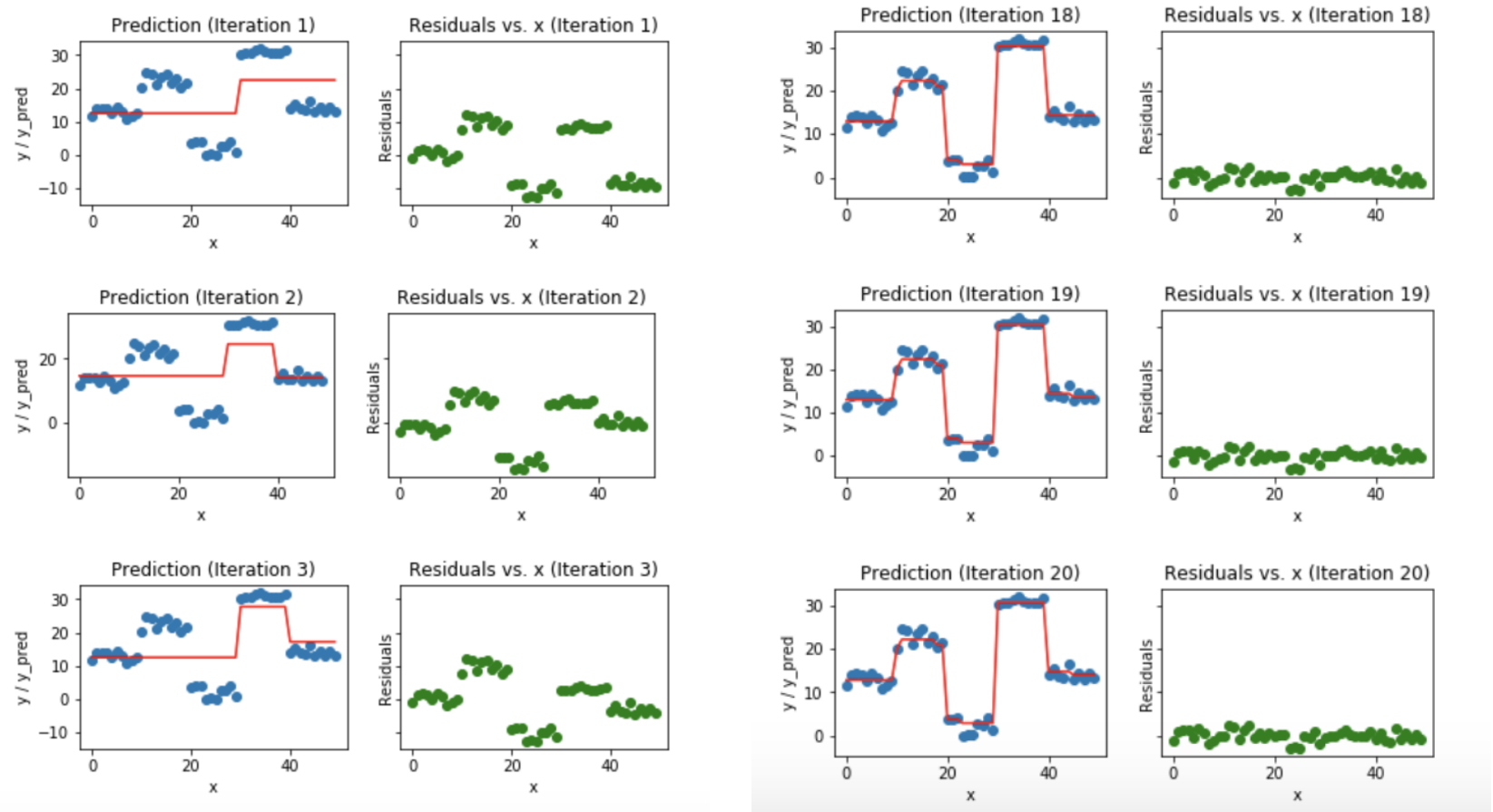
Ils permettent naturellement de traiter :

- les effets de seuil,
- les effets non-monotones,
- la dépendance entre les variables explicatives.

Il sont donc une excellente alternative aux GLM.

ILLUSTRATION AVEC UN ALGORITHME GBM

Gradient Boosted Trees (GBM) : effets seuil, non-monotones !



INCONVENIENTS

La difficulté de la manipulation de ces modèles réside dans :

- le manque d'interprétabilité,
- la gestion du surappentissage qui parfois est complexe.

En effet, certaines modélisations nécessitent de bien maitriser le choix des paramètres de tuning, qui peuvent en nombre assez grand (GBM par exemple).

TYPES DE MODELE

Parmi les approches non-paramétriques, on peut notamment utiliser :

- les arbres de décision CART,
- les modèles ensemblistes de type bagging ou boosting,
- les modèles à effet additif GAM.

Quelques références intéressantes :
- mémoire IA de C. Dutang sur les GAM,
- de nombreux mémoires IA sur le bagging et le boosting.

1 Tarification a priori - concepts avancés
- Introduction
- Modèles de tarification
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

2 Construction d'un zonier

3 Provisionnement

CREATION DE POCHES D'ASSURES

La segmentation amène à créer des poches d'assurés ayant les mêmes caractéristiques. Il y a un arbitrage naturel entre

- une segmentation **"grossière"** : peu de tarifs $\neq$;
- une segmentation **précise** : beaucoup de profils de risque considérés $\neq$, des tarifs très personnalisés.

La question essentielle liée à la segmentation est l'exposition :
$\rightarrow$ Remise en cause du principe de mutualisation (LFGN)...
$\rightarrow$ Attention pour les GLM (MLE asymptotique), voire même pour le calcul de la sinistralité globale en espérance par agrégation...
$\rightarrow$ Bc segmenter ne fait pas forcément $\searrow$ tarif car prime de risque (composant la prime technique) $\nearrow$ (incertitude des estimateurs).

MODELE PARCIMONIEUX

On a tjs 2 effets inverses en modélisation (cf théorie de Vapnik) :

- adéquation du modèle : + la dimension du modèle est grande, + l'adéquation aux données est bonne ;
- qualité prédictive : + la dimension du modèle est grande, + sa capacité prédictive est mauvaise (bruit au lieu du signal).

L'idée est donc de rechercher un arbitrage dans la dimension qui permette d'obtenir un bon compromis dans ces 2 objectifs.

C'est ce qu'on appelle un modèle parcimonieux.

Critères de sélection de modèles emboîtés : AIC, BIC, ...
Econométrie : pénalité ex-post / Machine-learning : pénalité dans l'optimisation (LASSO, ...).

PENALITES EX-POST ET EX-ANTE

Usuellement, on utilise des pénalisations a posteriori...

Bien que conduisant potentiellement à des estimateurs biaisés, on peut préférer au regard d'un critère d'erreur quadratique moyenne des estimateurs pénalisés ex-ante : cf article Econométrie et Machine Learning p.15 !

- monde paramétrique : régressions pénalisées (pénalisations ex-ante)
- monde Machine Learning : gestion des paramètres de tuning (pénalisations ex-ante)

GLM : DIFFICULTES D'ESTIMATION

Il arrive souvent en pratique que des coefficients de régression calibrés ne soient **pas significatifs**. Cela correspond au test :

$$H_0 : \hat{\beta}_j = 0 \qquad VS \qquad H_1 : \hat{\beta}_j \neq 0.$$

But : rejeter H_0 à un certain niveau de confiance α, en se basant sur le test de Fisher (ou Wald) $(\hat{\beta}_j/\sigma(\hat{\beta}_j))^2 \quad (\sim \chi^2(1))$.

Lorsque l'exposition est faible dans une poche, la calibration des coefficients de régression affectés à cette poche devient ardue...

Cela est dû au fait que le MLE est **asymptotiquement gaussien** :

$$\hat{\beta}_j^{MLE} \quad \sim \quad \mathcal{N}(\beta_j,\ 1/\mathcal{I}(\beta_j)).$$

$\Rightarrow$ La variance de l'estimateur peut devenir grande si l'information de Fisher est faible (quantité d'info contenue dans les données, petite dans le cas de trop peu d'individus).

La technique consiste alors à **regrouper certaines modalités de covariables** qualitatives (ou catégorielles). La démarche statistique "propre" s'y rapportant :

1. calibration du modèle complet,
2. pour le test de chaque coef. associé aux covariables, repérer la pire "p-valeur" au-dessus du seuil α,
3. agréger la modalité correspondante avec une autre "intelligemment" ;
4. recalibrer le modèle, et revenir à l'étape 2 tant que le modèle n'est pas satisfaisant.

POURQUOI PARTIR DU MODELE COMPLET ?

Lors de l'étape de sélection de modèle, on conseille généralement de partir du modèle complet, puis d'en chercher un sous-modèle optimal. Cela est dû au théorème de Frish-Waugh (voir aussi article Econométrie et Machine Learning d'Antoine Ly et Arthur Charpentier, section 2.9).

En effet, imaginons les 2 cas suivants :

- underfit, i.e. le vrai modèle (inconnu en pratique) s'écrit

$$y_i = \beta_0 + x_1^T \beta_1 + x_2^T \beta_2 + \epsilon_i$$

et que l'on estime

$$y_i = \beta_0 + x_1^T \beta_1 + \eta_i.$$

Alors

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T \epsilon$$

Et donc $E[\hat{\beta}_1] = \beta_1 + E[(X_1^T X_1)^{-1} X_1^T X_2 \beta_2] \quad \neq \quad \beta_1$ (biais !).

- overfit, i.e. le vrai modèle (inconnu) s'écrit

$$y_i = \beta_0 + x_1^T \beta_1 + \epsilon_i$$

et que l'on estime

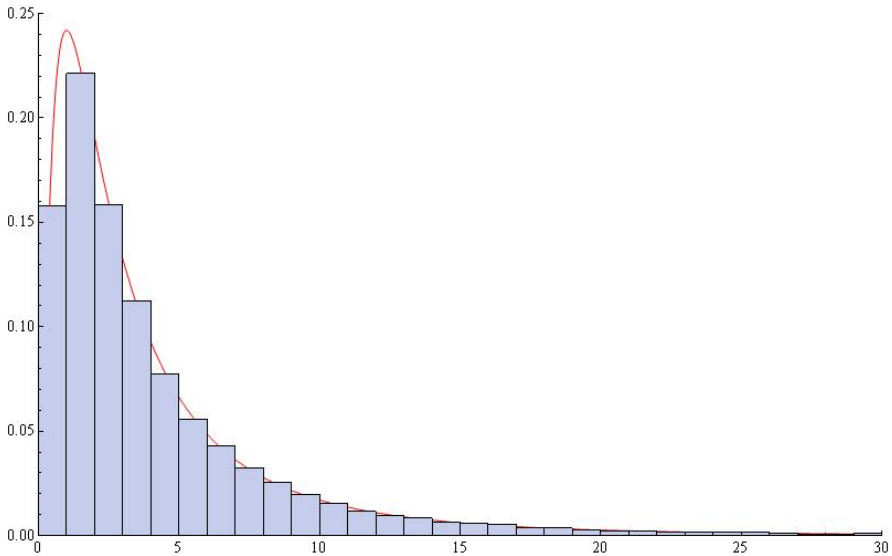
$$y_i = \beta_0 + x_1^T \beta_1 + x_2^T \beta_2 + \eta_i.$$

Alors $E[\hat{\beta}_1] = \beta_1$,
mais perte d'efficacité car overfitting ! D'où pénalisation par complexité du modèle.

DISTRIBUTION DE SINISTRALITE PAR POCHE

Au final, une question importante est d'**identifier les poches pour lesquelles la modélisation marche bien ou non** : il vaut mieux se tromper sur certains profils que sur d'autres...

Pour cela, on confronte la densité théo. construite par GLM à la densité empirique du profil et on espère une bonne adéquation (ex : rootogram) !



1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

PRATIQUE COURANTE

Dans les compagnies d'assurance, on penche souvent pour la loi de Poisson dans la modélisation de la fréquence des sinistres lorsqu'on adopte une modélisation de type fréquence-coût.

En effet,

- la survenance des sinistres est considérée sans mémoire„
- la Poisson ne dépend que d'un paramètre donc est simple
- cela simplifie le calcul global de sinistralité à l'échelle du portefeuille : loi Poisson composée stable par addition.

Souvent la variance empirique du nombre de sinistres est bien supérieure à sa moyenne empirique : cela va à l'encontre de la propriété fondamentale de cette loi $\Rightarrow$ pas adapté !

SURDISPERSION : BINOMIALE-NEGATIVE

Elle peut être construite comme un mélange de lois de Poisson :

$$(N \,|\, \Lambda = \lambda) \sim \mathcal{P}(\lambda) \qquad \text{et} \qquad \Lambda \sim \mathcal{G}a(\alpha, \delta).$$

La densité jointe de N et Λ vaut

$$f_{N,\Lambda}(n,\lambda) = f_{N|\Lambda=\lambda}(n)\, f_\Lambda(\lambda) = e^{-\lambda}\frac{\lambda^n}{n!}\,\frac{\delta^\alpha \lambda^{\alpha-1} e^{-\delta\lambda}}{\Gamma(\alpha)} \quad (\lambda, \alpha, \delta > 0, n \in \mathbb{N}).$$

Λ est continue et N discrète : la distribution marginale de N est

$$\begin{aligned}
\mathbb{P}(N=n) &= \int_0^\infty f_{N,\Lambda}(n,\lambda)\, d\lambda = \int_0^\infty e^{-\lambda}\frac{\lambda^n}{n!}\,\frac{\delta^\alpha \lambda^{\alpha-1} e^{-\delta\lambda}}{\Gamma(\alpha)}\, d\lambda \\
&= \frac{\delta^\alpha}{n!\,\Gamma(\alpha)} \int_0^\infty \lambda^{n+\alpha-1} e^{-(\delta+1)\lambda}\, d\lambda = \frac{\delta^\alpha \Gamma(\alpha+n)}{n!\,\Gamma(\alpha)\,(\delta+1)^{\alpha+n}}
\end{aligned}$$

Posons ensuite $p = \frac{\delta}{\delta+1}$, et $q = 1 - p = \frac{1}{\delta+1}$. Alors

$$\boxed{\mathbb{P}(N = n) = \frac{\Gamma(\alpha + n)}{n!\,\Gamma(\alpha)} p^{\alpha} q^{n}.}$$

La v.a. $N \sim \mathcal{NB}(\alpha; p)$ prend ses valeurs dans $\{0, 1, 2, ...\}$.

Remarques :

- La queue de distribution est plus épaisse que celle d'une loi de Poisson.
- Sa variance est plus grande qu'une loi de Poisson : loi utilisée en cas de surdispersion des observations.

AUTRES LOIS SURDISPERSEES

Une autre loi potentiellement utille pour traiter le phénomène de surdispersion est la loi de Borel-Tanner.

Elle fait partie des EDF (Exponential Distribution Functions)...donc appartient à la famille des GLM !

MODELES INFLATES
[Frees, 2009], [Vasechko et al., 2009]

Mélange discret à 2 composantes (**grande masse en 1 point**)...

Les “0” observés viennent de loi de comptage + masse en 0 (ex : “vrais” 0 pr pas de sinistre, et “faux” 0 provenant de recours...) :

- deux “sources” de 0, proportion du Dirac égal à $f_{zero}(0)$
- l’autre regroupe les obs. $\neq 0$ provenant de la loi de comptage.

$$\boxed{\mathbb{P}(N=k) \quad = \quad f_{zero}(0)\, dirac_{(0)} + (1 - f_{zero}(0))\, f_{\text{count}}(k).}$$

Ex : $N \sim ZI\mathcal{P}(\lambda)$: $\mathbb{P}(N=k) = \begin{cases} \pi_0 + (1-\pi_0)\, e^{-\lambda} & \text{si } k = 0, \\ (1-\pi_0)\, e^{-\lambda}\, \dfrac{\lambda^k}{k!} & \text{si } k > 0. \end{cases}$

Régression (N continue) (cf formation comportements chris à la fin). π_0 peut resulter d’une binomiale par ex. **Offset ?**

MODELES TRONQUES
[Frees, 2009], [Vasechko et al., 2009]

Mélange à 2 composantes (“hurdle-at-zero”), 1 seule source de 0 :

- loi de type binomiale par exemple qui génère les 0 (ne proviennent plus du tout de la loi comptage),
- à laquelle on ajoute une loi de comptage tronquée.

$$\mathbb{P}(N=k) = \begin{cases} f_{zero}(0) & \text{si } k=0, \\ (1-f_{zero}(0))\,\dfrac{f_{\text{count}}(k)}{1-f_{\text{count}}(0)} & \text{si } k>0. \end{cases}$$

Zero-trunc. $\mathcal{P}$: $\mathbb{P}(N=k) = \begin{cases} \pi_0 & \text{si } k=0, \\ (1-\pi_0)\,\dfrac{e^{-\lambda}\lambda^k}{(1-e^{-\lambda})k!} & \text{si } k>0. \end{cases}$

REMARQUES

- On pourrait voir le modele zero-inflate comme un modele dans lequel les couts sont parfois egaux a zero a cause de recours par exemple...alors qu'initialement ils n'étaient pas nuls !
- Si les zéros n'ont qu'une provenance, $+$ robuste d'utiliser un modèle hurdle concernant l'estimation statistique des param. (car estimation isolée pour chacune des 2 parties du modèle : logit et modèle de comptage).

Ex. : data "AutoClaim" dans la librairie R *cplm* (Yip and Yau, 2005).
Computational tools for such models (zero-altered models, ...) :
librairie *mboost* et *countreg*.
$\rightarrow$ Voir l'article Boosting actuarial regression models (IME 2019).

1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

TYPOLOGIES DE SINISTRE

La sinistralité se décompose généralement en trois typologies de sinistre :

- attritionnels : haute fréquence, petite sévérité ;
- graves : basse fréquence, grande sévérité ;
- CAT : très basse fréquence, sévérité extrême.

Nécessité de séparer ces données car les **modèles classiques ne fonctionnent que sur les sinistres attritionnels** (à cause des queues des distributions des lois utilisées) $\Rightarrow$ écrêtement.

Rq : utiliser techniques de Théorie des Valeurs Extrêmes.

DONNEES ATYPIQUES

Malgré l'écrêtement des sinistres, on observe parfois de la sinistralité un peu atypique au sein de l'échantillon...

On peut traiter ce problème avec des approches un peu plus sophistiquées comme celle présentée dans l'article

Computational Bayesian Credibility for GLMs, de José Garrido (Concordia University, Montreal)

Il s'agit d'estimer la prime en 2 étapes...En moyennant 2 fois.

DISCONTINUITE - DISTRIBUTION DES MONTANTS

On observe parfois (surtout pour les branches à développement long) des pics de densité pour certaines valeurs de coût de sinistre unitaire.

Cela est dû par exemple à des forfaits à l'ouverture (de sinistre), type convention IRSA ou forfait IDA en assurance automobile.

Ces montants forfaitaires doivent être exclus de l'étude !

Rq : cette suppression fait souvent baisser le coût moyen, suggérant que les forfaits d'ouverture sont prudents.

GESTION D'UNE HETEROGENEITE INOBSERVABLE

Une approche potentielle pour gérer l'hétérogénéité **inobservable** des données consiste à considérer des modèles mélanges finis. Ils peuvent être discrets ou continus (ex : mélange Poisson-Gamma).

Admettons que l'on observe l'échantillon $x = (x_1, ..., x_d)^T$, réalisations iid de $X = (X_1, ..., X_d)^T$.

La densité mélange de X s'écrit comme suit dans le cas discret :

$$p(x; \Theta) = \sum_{j=1}^{M} \pi_j f_j(x; \theta_j), \quad \text{avec} \sum_{j=1}^{M} \pi_j = 1, \quad \pi_j > 0.$$

En termes d'estimation des paramètres, on se base généralement sur l'algorithme Espérance-Maximisation (EM).

ILLUSTRATION

On se propose ici d'afficher l'aspect caractéristique d'une densité de probabilité d'une loi mélange discret.

PRINCIPE DE L'ALGORITHME EM

Complétion artificielle des données pas à pas (on n'observe pas le label Y d'appartenance des indiv. aux composantes).

Soit $Z = (X, Y)$ les données (X est observé, au contraire du label Y). L'algorithme se décompose en 2 étapes à chaque itération k :

- E-step : calcule log-vraisemblance espérée des données fictives :
$$Q(\Theta; \Theta^{(k)}) = \mathbb{E}_{\Theta^{(k)}}[\ln L_c(\Theta) \mid X]$$
- M-step : met à jour les paramètres en maximisant Q, donc
$$\Theta^{(k+1)} = \arg\max_{\Theta} Q(\Theta; \Theta^{(k)}).$$

Au final : attribution de l'obs. à l'une des composantes (Bayes).

1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

FRANCHISE ET EFFET DE SEUILLAGE

Franchise : impacte la loi de fréquence et de coût.

Historiquement, la franchise a été instaurée afin de

- **diminuer** l'aléa moral (comportement moins prudent car assuré) ;
- **l'antisélection** (délai de carence par exemple en Prévoyance).

D'un point de vue statistique, cette approche doit être adaptée pour tenir compte des contraintes liées au dispositif de collecte des données, à savoir qu'il existe un **seuil de collecte** des pertes.

Seules les pertes $> H$ (où H est la franchise) sont collectées.

IMPACT SUR LA SEVERITE

On observe un échantillon $(X_1, ..., X_n)$ de pertes i.i.d. au delà du seuil de collecte H.

On obtient donc une distribution modifiée par rapport à la distribution théorique sans seuillage, donnée par

$$\tilde{f}_{\theta|H}(x) = \frac{f_\theta(x)}{\mathbb{P}(X > H)} \mathbb{1}_{x>H} = \frac{f_\theta(x)}{\int_H^\infty f_\theta(u)\, du} \mathbb{1}_{x>H}.$$

Estim. des paramètres θ : méthode des moments généralisée (minimise l'écart entre moments théo / moments empiriques), ...

EXEMPLE : $X_{i,j}^k \sim X \sim \mathcal{LN}(\mu, \sigma)$

Besoin : au – autant de moments théo. que de param. à estimer...

En notant les moments $m_p(\theta) = \mathbb{E}[X^p \mid X > H] = \int_{-\infty}^{\infty} x^p \, \tilde{f}_{\theta|H}(x) \, dx$,

$$
\begin{aligned}
m_1(\mu, \sigma) &= \frac{1 - \Phi\left(\frac{\ln H - (\mu + \sigma^2)}{\sigma}\right)}{1 - \Phi\left(\frac{\ln H - \mu}{\sigma}\right)} e^{\mu + \sigma^2/2} \\
m_2(\mu, \sigma) &= \frac{1 - \Phi\left(\frac{\ln H - (\mu + 2\sigma^2)}{\sigma}\right)}{1 - \Phi\left(\frac{\ln H - \mu}{\sigma}\right)} e^{2(\mu + \sigma^2)}
\end{aligned}
$$

où Φ désigne la fonction de répartition d'une loi $\mathcal{N}(0, 1)$.

Puis on inverse le système en remplaçant m_1 et m_2 par $\tilde{\mu}_n$ et $\tilde{\sigma}_n$ (EMM), et on trouve $\hat{\mu}$ et $\hat{\sigma}$!

IMPACT SUR LA FREQUENCE - EXEMPLE POISSON

Souvent modélisée par la loi de Poisson ($N \sim \mathcal{P}(\lambda)$) :

$$P(N = n) = e^{-\lambda} \frac{\lambda^n}{n!}$$

- Simple (EMV = moy. empirique).
- Calibration de la fréquence après celle de la sévérité pour prendre en compte la présence du seuil de collecte :

$$\hat{\lambda} = \frac{\hat{\lambda}_H}{\mathbb{P}(X > H)} = \frac{\hat{\lambda}_H}{1 - F_{\hat{\theta}}(H)}$$

En pratique donc : calculer la moyenne empirique du nb de pertes annuel (λ_H) et utiliser l'estimateur de θ pour obtenir le vrai λ.

RECOURS ET REASSURANCE

Concernant les recours, il y a 2 solutions :

- soit les recours se traitent en amont de la modélisation,
- soit on modélise la probabilité de recours, puis combien cela rembourse (approche PD-LGD en crédit)

La réassurance peut également intervenir dans le tarif : elle s'intégre après estimation des modèles et déduction de la prime pure.

PROVISION AJOUTEE AU TARIF

Idée : il manque de l'information dans la sinistralité observée dans la base, car certains sinistres ne sont pas déclarés/clos...

Le provisionnement peut donc jouer dans la valeur de la prime, en l'occurence la baisser si l'activité fait des bénéfices ou la monter pour des branches à développement long.

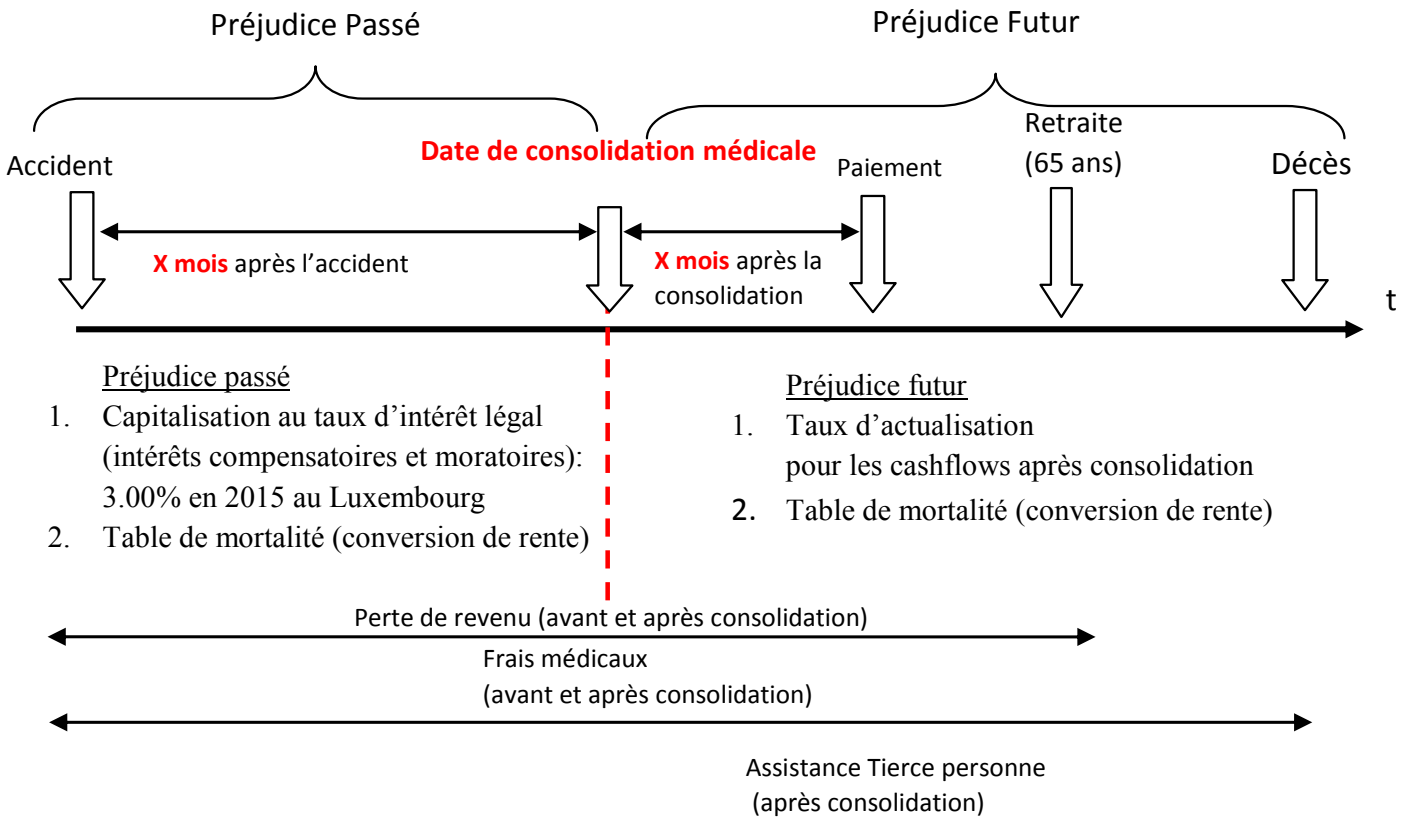
Une manière d'intégrer le provisionnement serait de faire d'abord un Chain Ladder pr évaluer la charge ultime puis utiliser le volume de prime pour en déduire un Loss-Ratio.

Ce Loss-Ratio est ensuite appliqué à la prime déterminée par GLM.

EXEMPLE : DOMMAGES EN RC CORPORELLE

Difficultés : consolidation médicale ⇒ rapport AGIRA par ex.

Rq : attention donc à l'inflation, notamment médicale.



Source : mémoire actuariat de Nicolas Faugère.

VISION “AS-IF” DES MONTANTS

En principe, les données répertorie les montants de sinistre relativement à une certaine date... qui peut être ancienne ! Attention donc à l’inflation.

Afin de tarifer pour les années à venir, il est **important de ramener ces montants au moment de la tarification** (en ramenant ces montants à des coûts “actuels”)

C’est ce qu’on appelle la mise en “as-if” : cela revient en général à capitaliser les montants sur une ou plusieurs périodes.

LISSAGE DU TARIF

En réalité, une refonte tarifaire amène quasi-systématiquement à un **écart de tarif significatif** entre l'existant et le nouveau.

Une manière de combler cet écart en pratique est d'estimer le modèle GLM sans en tenir compte, puis on compare la nouvelle et l'ancienne prime. Cela nous permet de déterminer une constante permettant de passer d'une prime à l'autre.

Cette constante est ensuite réintégrée dans la modélisation via un nouvel offset ; puis on re-estime le modèle avec cet offset.

N.B. : les méthodes différent suivant les compagnies...

1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

DANS LES GLM : QU'EST CE QU'UN OFFSET ?

L'offset représente une sorte d'**exposition**.

C'est une constante qui va venir modifier le risque de base, donc le risque qui n'est pas lié au profil de l'assuré en particulier.

Exemples d'offset :

- assurance auto indiv. : nb d'années d'assurance du véhicule ;
- assurance collective auto : taille de la flotte assurée ;
- incapacité-invalidité : effectif salariés, masse salariale ;
- réassurance : taille du portefeuille, ...

Calcul du tarif : bien fixer l'offset à 1 (si unité de mesure en année, car durée d'assurance de 1 an par défaut).

INTEGRATION D'UN OFFSET DANS UN GLM

Tout simplement ! C'est un terme commun à tous les individus, mais dont la valeur va changer en fonction des individus.

En terme explicite, l'équation devient

$$g(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]) = \mathit{offset} + \mathbf{x}^T\beta.$$

- on **contraint le coefficient de l'offset à valoir 1** (c'est pourquoi il n'apparait pas dans l'équation !) ;
- pour la calibration, on régresse $g(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]) - \mathit{offset} = \mathbf{x}^T\beta$.

EXEMPLE AVEC LE MODELE LOG-POISSON

L'idée globale de l'offset est que la réponse **y est proportionnelle**.

Donc l'offset s'exprime sur la même échelle que la réponse. Dans le cas du modèle log-Poisson de paramètre λ, on aurait donc

$$ln(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]) = ln(exposition) + \mathbf{x}^T\beta.$$

Soit le modèle suivant à calibrer : $ln\left(\frac{\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]}{exposition}\right) = \mathbf{x}^T\beta.$

On remplace donc la fréquence (au sens nb de sinistres) par une fréquence standardisée !

ET DANS LES AUTRES MODELES ?

Cf TP sur le modele binomial.

Modèle CART ?

CONTRAINDRE DES COEFFICIENTS

Si l'on veut intégrer dans le modèle des facteurs de risque dont les coefficients ont déjà une valeur (estimée par ailleurs), on peut donc utiliser la même idée que l'offset...

Ainsi, si l'on souhaite intégrer un zonier dans le modèle tarifaire, on introduira Z comme un offset. Ex : si 3 zones de risque :

- zone 1 : $z = -5\%$
- zone 2 : rien.
- zone 3 : $z = +5\%$

Ex. GLM log-Poisson : on introduit l'offset $\log(z)$, donc $\log(1.05)$ pour la zone 3...

1 Tarification a priori - concepts avancés
- Introduction
  - Notions de base en tarification
  - Chargements techniques et principes de prime
  - Segmentation et partage du risque
- Modèles de tarification
  - Approches de tarification
  - Modèles paramétriques : les GLM
  - Modèles non-paramétriques
- Problématiques opérationnelles pour tarifer
  - Segmentation et modélisation : limites à garder en tête
  - Surdispersion pour la loi de fréquence
  - Difficultés liées aux données d'assurance
  - Autres problématiques opérationnelles
  - Tenir compte de l'exposition au risque
  - Réponse catégorielle : sur-représentation d'une modalité
- Résumé

TAUX DE REPONSE FAIBLE

On cherche parfois à modéliser un **événement binaire "rare"** en utilisant des modèles GLM.

Quel(s) problème(s) cela pose ?

Difficultés énoncées précédemment sur la calibration notamment
$\rightarrow$ $+$sieurs poches où on observe (très) peu ou pas l'événement...

Exemples concrets (souvent en risque comportemental) :

- taux de résiliation en assurance vie et non-vie (surtout en vie où les taux de résiliation annuels sont $+$ faibles) ;
- taux de conversion en assurance directe par exemple.

SEUIL D'AFFECTATION ET COURBE ROC

Dans ce type de problématique, on a coutume d'évaluer la performance d'un modèle grâce à la courbe ROC.

Celle-ci permet également de voir que dans un tel cas, le meilleur seuil d'affectation de la réponse à l'une ou l'autre des modalités possibles pour la réponse ne se situe pas à une probabilité égale à 0,5...

FORMALISATION DU CONTEXTE

Plaçons nous dans le cadre de risque comportemental pour présenter le concept (ex : taux de conversion). Cela nous amène à considérer un modèle GLM de type logistique, à savoir

$$ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T \beta.$$

Rappelons que

- $\mathbf{X}_i^T = (1, X_{i1}, ..., X_{iJ})$ et $\beta^T = (\beta_0, \beta_1, ..., \beta_J)$;
- $i \in 1, ..., I$: $Y_i \in \{0, 1\}$ $\Rightarrow$ $Y_i \sim \mathcal{B}(p_i)$;
- $p_i = P(Y_i = 1)$.

En pratique, $\bar{p} = \frac{1}{I} \sum_i \mathbb{1}_{y_i=1}$ est **de l'ordre de quelques % au** $+$.

UN APARTE SUR LA FONCTION DE LIEN

Dans le cadre du modèle logistique, 3 fonctions de lien possibles. Liées aux 3 fonctions de répartition possibles pour Y^* (continue) non observable (cf TP) :

- FdR loi logistique (modèle logit) :

$$F(x) = \frac{1}{1+e^x}, \qquad g(p) = ln\left(\frac{p}{1-p}\right)$$

- FdR loi normale centrée réduite (modèle probit) :

$$F(x) = \Phi(x) = \int_{-\infty}^{x} e^{-t^2/2} dt, \qquad g(p) = \Phi^{-1}(p)$$

- FdR loi Gumbel II (modèle complementary log-log) :

$$F(x) = 1 - \exp(-\exp(x)), \qquad g(p) = log(-log(1-p)).$$

PROBLEMES THEORIQUES ASSOCIES
[Albert and Anderson, 1984]

1. La séparabilité : en fait, l'existence d'un estimateur du maximum de vraisemblance est conditionné par le problème de séparation. Il n'∄ de MLE en cas de séparation complète.

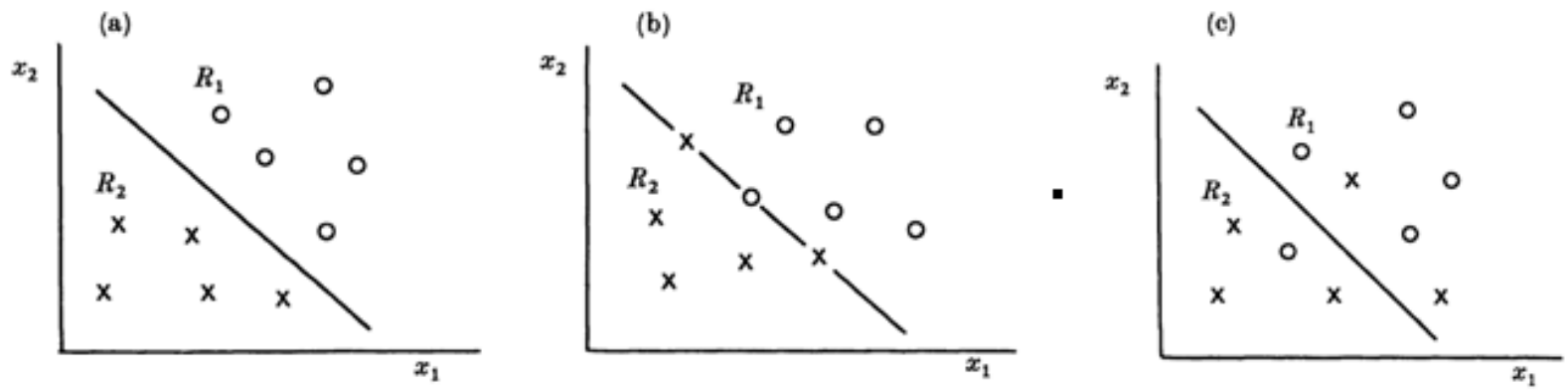


Figure 2 Possible configuration of sample points in the case of two variables, x_1 and x_2, and two groups, E_1, shown by circles, and E_2, shown by crosses. Regions R_1 and R_2 define corresponding allocation rule. (a) Complete separation. (b) Quasi-complete separation. (c) Overlap.

2. La dimensionnalité (“curse of dimensionality”).
   On dispose souvent de bc de covariables : la dim. de l’espace ↗ vite et les données peuvent rapidement devenir “sparse”.

Pour toute procédure statistique, la sparsité est un problème important. On entend parfois parler de

“Small N large P”

Pour avoir un résultat fiable dans la plupart des modèles statistiques, la taille des données dont nous avons besoin croit souvent exponentiellement en fonction de la dimension du modèle.

Remarque : dans le cadre de données “sparse”, on utilise plutôt la régression ridge, lasso, elastic net...

SOLUTIONS THEORIQUES POSSIBLES

Pour éviter le problème de sparsité ou de non-existence du MLE pour des données qui seraient séparées (ou quasi-séparées), il existe deux principales méthodes :

- la vraisemblance pénalisée (penalized likelihood method) ;
- la régression logistique conditionnelle exacte (exact conditional logistic regression).

Rq : la 3^e alternative est le response-based sampling, artifice pour retomber sur un problème plus facile à traiter mais qui n'est pas applicable directement sur le problème d'origine (cf $+$ loin).

UN MOT SUR LA VRAISEMBLANCE PENALISEE [Firth, 1993]

C'est une technique adaptée au problème de petit échantillon (peu de réponses observées égales à 1 entre dans ce cadre).

L'idée est de **corriger le biais des estimations MLE** (biais en $o(n^{-1})$) dû au manque de données. Pour corriger ce biais, on optimise la vraisemblance pénalisée de l'information de Fisher :

$$L^*(\beta) \quad = \quad L(\beta)\sqrt{\mathcal{I}(\beta)}.$$

Cette fonction de pénalité est appelée l'a priori de Jeffrey. Asymptotiquement, son influence est négligeable.

LA REGRESSION LOGISTIQUE CONDITIONNELLE EXACTE [Mehta and Patel, 1995]

Considérons un coefficient de régression β_j ($j = 1, ..., J$).
Introduisons la statistique exhaustive (ou suffisante) de β_j

$$T_j = \sum_{i=1}^{I} y_i \, x_{ij},$$

L'inférence est basée sur la distribution exacte sous hypothèse nulle de T_j, conditionnellement au vecteur de statistiques exhaustives des autres coefficients :

$$T_{j^-} = (T_k)_{k \in [1,J],\, k \neq j}$$

On maximise ensuite la vraisemblance conditionnelle

$$\mathbb{P}(T_j = t_j \,|\, \beta_j, T_{j^-} = t_{j^-}) = \frac{\exp(\beta_j\, t_j)}{\sum_{\Omega_j} \exp(\beta_j \sum_i y_i^* x_{ij})}$$

où Ω_j est l'ensemble des permutations y^* de y **telles que** pour chaque $y^* \in \Omega_j$

$$\sum_i y_i^* x_{ij'} = T_{j'} \qquad \forall j' \in j^-.$$

- Fonctionne bien pour des données mal séparées ;
- Consommateur de ressources calcul (mal adapté si big BdD).

VARIANCE DE L'ESTIMATEUR MLE

Rappel : l'erreur d'estimation de β est composée de 2 termes : le biais au carré, plus la variance de l'estimateur.

Estimation classique : on estime le vecteur β de paramètres par maximum de vraisemblance, où la vraisemblance vaut

$$L(\beta; y = (y_1, ..., y_I)) = f_{(Y_1,...,Y_I)}(y_1, ..., y_I; \beta).$$

Grâce à l'indépendance, $L(\beta; y = (y_1, ..., y_I)) = \prod_i f_{Y_i}(y_i; \beta)$,
et donc

$$L(\beta; y) = \prod_i p_i^{y_i} (1 - p_i)^{1-y_i}$$

où β est caché dans p_i.
$(p_i = \exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{iJ})/(1 + \exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{iJ})))$

Ainsi, on cherche à **résoudre le problème de minimisation**

$$(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_J) = \underset{\beta=(\beta_0,...,\beta_J)}{\arg\min} \; (-\log \, L(\beta; y))$$

avec $\log \, L(\beta; y) = -\sum_i \, ln\Big(1 + \exp((1 - 2y_i)\mathbf{x}_i^T \beta)\Big)$.

[Greene, 2008] montre que la variance de l'estimateur est donnée par

$$Var(\hat{\beta}) = \left(\sum_i \, p_i(1 - p_i)\mathbf{X}_i^T \mathbf{X}_i \right)^{-1} .$$

- La proportion de 1 intervient dans le terme $f(p_i) = p_i(1 - p_i)$;
- $p_i = \mathbb{P}(Y_i = 1 \,|\, \mathbf{X})$ est petit pour la plupart des individus ;

On peut faire quelques remarques :

- $f(p_i)$ est maximale pour $p_i = 0.5$;
- si le modèle a un pouvoir explicatif convenable, p_i sera plus grand pour les individus dont la réponse observée y_i vaut 1 que pour les autres ;
- donc $p_i(1 - p_i)$ sera plus grand pour ces individus ($y_i = 1$) $\Rightarrow$ leur variance sera $+$ faible.

Ce raisonnement explique pourquoi augmenter la proportion de réponses égales à 1 améliore l'estimation des coefficients de régression.

BIAIS DU MLE SUR DONNEES DESEQUILIBREES [McCullagh and Nelder, 1989]

Après avoir vu l'expression de la variance de l'estimateur, on peut en estimer le biais (évidemment ces 2 quantités sont à minimiser).

Rappel : pour un estimateur $\hat{\beta}$ de β, on définit le biais comme

$$\text{biais}(\hat{\beta}) \quad = \quad \mathbb{E}[\hat{\beta} - \beta] \quad = \quad \mathbb{E}[\hat{\beta}] - \beta.$$

Dans le cadre du MLE dans le modèle logistique, il est **estimé par la quantité**

$$\text{biais}(\hat{\beta}^{MLE}) = \frac{\mathbf{X}^T W \xi}{\mathbf{X}^T W \mathbf{X}}.$$

où W et ξ sont liés aux poids des observations et aux $\hat{p}_i$.

De manière plus précise, on a

- w_i est le poids accordé à l'observation i ;
- $\hat{p}_i$ est l'estimation fournie par la modélisation ;
- $\xi_i = 0.5 \times Q_{ii} \times [(1 + w_i)\,\hat{p}_i - w_i]$;
- W est la matrice telle que $W = diag(\hat{p}_i\,(1 - \hat{p}_i)\,w_i)$;
- Q est la matrice donnée par

$$\frac{\mathbf{X}\,\mathbf{X}^T}{\mathbf{X}^T W \mathbf{X}};$$

- Q_{ii} sont les éléments diagonaux de la matrice Q ;

Rq : dans le cadre de petits échantillons avec peu de "succès" ($y_i = 1$), c'est $\hat{\beta}_0$ qui est affecté en premier. Par propagation, tous les $\hat{\beta}_j$ sont ensuite affectés.

EXEMPLE DE BIAIS [King and Zeng, 2001]

Considérons la modélisation suivante : $p_i = \dfrac{\exp(\beta_0 + \beta_1 X_1)}{1 + \exp(\beta_0 + \beta_1 X_1)}$.

Dans ce cas, on peut approximer le biais de $\hat{\beta}_0$ par

$$\mathbb{E}[\hat{\beta}_0 - \beta_0] \quad = \quad \frac{\bar{p} - 0.5}{n\,\bar{p}\,(1 - \bar{p})}.$$

Clairement, le biais sera donc négatif car $\bar{p}$ est petit dans notre cas
$\Rightarrow$ on aura tendance à systématiquement sous-estimer β_0 !

En revanche, ce biais diminue à la vitesse n^{-1}...

PROPAGATION DU BIAIS

Le biais dans l'estimation des paramètres induit automatiquement un biais dans l'estimation des probabilités p_i. On montre que la proba. p_i est sous-estimée dans le contexte du modèle logistique (avec peu de succès observés), et que le biais peut être estimé par

$$p_i = \mathbb{P}(Y_i = 1 \,|\, \mathbf{X}) \quad = \quad \hat{p}_i + C_i$$

où le facteur de correction C_i vaut

$$C_i = (0.5 - \hat{p}_i)\, \hat{p}_i\, (1 - \hat{p}_i)\, \mathbf{X}\, Var(\hat{\beta}^{MLE})\, \mathbf{X}^T.$$

$\rightarrow$ $C_i > 0$ car $\hat{p}_i$ petit : on **sous-estime systématiquement** p_i ;
$\rightarrow$ biais $\searrow$ si la variance de l'estimateur diminue, ou si $\hat{p}_i \nearrow$...
$\Rightarrow$ Lien entre biais de la proba estimée et variance de l'estim. $\hat{\beta}$.

REECHANTILLONNAGE

Les 2 approches théo. de correction du biais (vraisemb. pénal. / reg. log. cond. exacte) étant difficiles à mettre en oeuvre, on opte en pratique pour la méthode de type “importance sampling”.

Nous avons au départ un jeu de données dont le taux de conversion vaut τ (ex : $\tau = 2\%$).

Pour éviter les pb de calibration avec ces données, on rééquilibre l'échantillon en termes de nb d'événements d'intérêt observés.

C'est la **response-based sampling method** (ou **choice-based sampling method**). Notons τ^c le nv taux de conversion ($\tau^c > \tau$).

TYPE RESPONSE-BASED SAMPLING

Nous devons donc construire un **response-based dataset**.
Cette méthode soulève 2 questions sans réponse évidente :

- si nous changeons la proportion des modalités observés dans l'échantillon d'apprentissage, le modèle construit sera différent. Comment ensuite retrouver des résultats cohérents pour la population d'origine ?

- lors du rééchantillonnage, il faut choisir un taux arbitraire de représentation des modalités de la réponse. Par ex., on choisit 30% (τ^c) de contrats souscrits. Comment fixer ce taux ?

FORMALISATION DU CONTEXTE

On dispose des données et du problème suivant :

- I est la taille de l'échantillon initial ;
- $Y_i \sim \mathcal{B}(p_i) \quad \Rightarrow \quad y_i \in \{0, 1\}$;
- $\mathbf{X}_i = (X_{i1}, ..., X_{iJ}) \in \mathbb{R}^J$;
- on note $f_{\mathbf{X}}$ la densité de $\mathbf{X}$, et f_Y celle de Y.

On cherche à estimer le paramètre p_0 de la loi de Y, après avoir supposé un modèle paramétrique (logistique) :

$$f_Y(y \,|\, \mathbf{x}) = f_Y(y \,|\, \mathbf{x}, p_0).$$

Notons que $f(y, \mathbf{x}) = f_Y(y \,|\, \mathbf{x}, p_0)\, f_{\mathbf{X}}(\mathbf{x})$ (Bayes).

AUTRES TECHNIQUES DE REECHANTILLONNAGE

- **Simple random sampling** : vraisemblance d'1 seule obs. :

$$L^{sr}(p; (y, \mathbf{x})) = f(y, \mathbf{x}) = f_Y(y \,|\, \mathbf{x}, p)\, f_{\mathbf{X}}(\mathbf{x}).$$

  $\rightarrow$ Propriétés estimateur identiques que sur la population globale (maximise la même forme de vraisemblance).

- **Exogenous stratified sampling** : stratifie l'échantillon sur **x**. On a dc une nvelle densité $g(\mathbf{x})$, et la vraisemblance s'écrit

$$L^{es}(p; (y, \mathbf{x})) = f(y, \mathbf{x}) = f_Y(y \,|\, \mathbf{x}, p)\, g(\mathbf{x}).$$

  $\rightarrow$ Adapté pour sur-représenter des catégories de personnes.
  $\rightarrow$ Ne modifie pas le maxim. de la vraisembl. (se fait sur p).

RESPONSE-BASED SAMPLING ET MODELE LOGIT [Xie and Manski, 1989]

Stratification sur la réponse Y : on modifie le taux d'occurence de l'événement de la population d'origine. I désigne la taille de l'échantillon, f_Y la densité de Y dans la population d'origine.

$Y \in \{0, 1\}$: notons $1 - \tau^c$ le taux moyen (dans le nouvel échantillon) de non occurence de l'événement, et τ^c son complémentaire.
On y associe le nb d'événements (ou pas événement) I_0 et I_1 t.q.

$$1 - \tau^c = (I_0/I) \quad \text{et} \quad \tau^c = (I_1/I).$$

Rappelons que f_Y désigne la densité de la réponse Y. On a

$$L^{rb}(p; (y, \mathbf{x})) = f(\mathbf{x} \,|\, y) \frac{I_y}{I} = \frac{f_Y(y \,|\, \mathbf{x}, p) f_{\mathbf{X}}(\mathbf{x})}{\int_{\mathcal{X}} f_Y(y \,|\, \mathbf{x}, p) f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}} \frac{I_y}{I}$$

Ici l'optimisation est modifiée, contrairement à précédemment où la vraisemblance à optimiser était directement une fonction de p à travers le noyau $f_Y(y \mid \mathbf{x}, p)$.

Ainsi le paramètre d'intérêt p sur lequel optimiser **intervient différemment dans le noyau** qui devient

$$\frac{f(x,y)}{f(y)} = \frac{f_Y(y \mid \mathbf{x}, p) f_{\mathbf{X}}(\mathbf{x})}{\int_{\mathcal{X}} f_Y(y \mid \mathbf{x}, p) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}}.$$

Ici, la densité marginale de Y (au dénominateur) dépend de p : on va donc modifier $\hat{p}$ en maximisant la vraisemblance (estimation de p) et obtenir un estimateur qui n'est pas robuste pour la population globale.
Comment le rendre donc robuste ?

METHODE 1 : WEIGHTING METHOD [Manski and Lerman, 1977]

Il suffit de pondérer la vraisemblance avant de l'optimiser sur l'échantillon response-based.

Ils définissent ainsi la **weighted maximum likelihood estimation**, basée sur la log-vraisemblance

$$\log L_w(p;(y,\mathbf{x})) = \sum_{i=1}^{I} w(y_i)\, \ln(f(y_i \mid \mathbf{x}_i, p))$$

avec

$$w(y) = \frac{f_Y(y)}{(I_y/I)}.$$

On conserve au final l'estimateur

$$\hat{p}^{MLE} = \arg\max_{p} \log L_w(p; (y, \mathbf{x})).$$

Remarque : les poids font intervenir

- la quantité (I_y/I) : proportion de 0-1 dans la popu. créée.
  $\rightarrow$ Directement observable à partir des données.

- $f_Y(y) = \int_{\mathcal{X}} f_Y(y \,|\, \mathbf{x}, p) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$: choix crucial si proportion du phénomène non-observée en pratique.
  $\rightarrow$ issu en général d'une connaissance / information extérieure (survey, ...) si non-observée (mais prop. observée pr nous ds la pop. globale car classif. supervisée).

Rq : méthode qui fonctionne qlq soit le lien (logit, probit, ...).

EN PRATIQUE

On a

$$w(y) = \frac{f_Y(y)}{(I_y/I)}.$$

Au numérateur, il s'agit de la proportion de 1 (respectivement 0) dans la population d'origine. Au dénominateur, il s'agit de la proportion de 1 (respectivement 0) dans la pop. response-based. Donc

- pour $y_i = 1$: les poids sont $w(1) = \frac{\tau}{\tau^c} < 1$
- pour $y_i = 0$: les poids sont $w(0) = \frac{1-\tau}{1-\tau^c} > 1$

On surpondère les observ. égales à 0 : logique puisque l'échant. response-based contient bien moins de 0 que celui d'origine...

ECART DE TARIF

Ce biais peut rapidement mener à une sous-estimation importante de la sinistralité dans le cas de gros portefeuille...

Prenons par ex. le portef. avec caractéristiques suivantes :

- 1 000 000 d'assurés,
- une fréquence moyenne de survenance des sinistres de 10%,
- un coût moyen du sinistre de 2000 euros.

Admettons que le biais de la probabilité de survenance soit de 1%, donc sous-évaluée à 10% plutôt que 11%.
Grossièrement, il faudrait donc ajouter 10 000 sinistres dans l'année, soit une charge totale de

$$10000 \times 2000$$

soit 20 000 000 d'euros à payer en plus !

METHODE 2 : PRIOR CORRECTION [Xie and Manski, 1989]

Uniquement dans le cas du modèle logit (lien logit et bonne spécification du modèle). Estimer par MLE sur l'éch. **response based** conduit à bien estimer ts les coef. de régr., excepté β_0.

On "corrige" donc l'estimation $\hat{\beta}_0$ de β_0 comme suit :

$$\tilde{\beta}_0 = \hat{\beta}_0 - \ln\left(\frac{1-\tau}{\tau}\,\frac{\tau^c}{1-\tau^c}\right),$$

avec τ la prop. de 1 ds la pop., et τ^c celle ds l'éch. response-based.

Ainsi, on estime coef. par max de vrais. sur l'échantillon response based en introduisant avant un offset valant $\ln\left(\frac{1-\tau}{\tau}\,\frac{\tau^c}{1-\tau^c}\right)$.

JUSTIFICATION DE LA CORRECTION

Soit C_1 l'événement $Y = 1$ et C_0 l'événement $Y = 0$.

$$
\begin{aligned}
P(C_1|x) &= \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_0)P(C_0)} \\
&= \frac{1}{1 + \frac{P(x|C_0)P(C_0)}{P(x|C_1)P(C_1)}} \\
&= \frac{1}{1 + \exp\left(\ln\left(\frac{P(x|C_0)P(C_0)}{P(x|C_1)P(C_1)}\right)\right)} \\
&= \frac{1}{1 + \exp\left(-\ln\left(\frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)}\right)\right)} \\
&= \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 \times \mathbf{x})\right)}
\end{aligned}
$$

Or,

$$\beta_0 + \beta_1 \times \mathbf{x} = \ln\left(\frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)}\right),$$

donc les paramètres de régression sont estimés sous l'hyp. que les probabilités a priori de chaque classe sont équilibrées voire égales... On peut ainsi re-introduire le odd-ratio a priori dans l'intercept comme ceci :

$$\begin{aligned}
\beta_0 + \beta_1 \times \mathbf{x} + \ln\left(\frac{P(C_1)}{P(C_0)}\right) &= \ln\left(\frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)}\right) + \ln\left(\frac{P(C_1)}{P(C_0)}\right) \\
\beta_0 + \ln\left(\frac{P(C_1)}{P(C_0)}\right) + \beta_1 \times \mathbf{x} &= \ln\left(\frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)}\right) + \ln\left(\frac{P(C_1)}{P(C_0)}\right)
\end{aligned}$$

D'où $\tilde{\beta}_0 = \beta_0 + \ln\left(\frac{P(C_1)}{P(C_0)}\right)$.

1. Tarification a priori - concepts avancés
   - Introduction
   - Modèles de tarification
   - Problématiques opérationnelles pour tarifer
   - Résumé
2. Construction d'un zonier
3. Provisionnement

RESUME DES ETAPES DE CREATION D'UN TARIF

On résume ici les principales étapes à exécuter dans une optique de tarification.

Dans l'ordre :

1. Importation des données et premiers traitements (données aberrantes, valeurs manquantes, transformation de types, . . .)
2. Extraction des bases par garantie assurée
3. Traitement des données (nettes de franchises, recours, forfait type IDA, mise en as-if pour l'inflation, dvp des sinistres pour prise en compte de provision ds tarif, réass. à répercuter ?)

4. Statistiques descriptives (exposition, fréquence et cout moyen par variable explicative, tests de corrélation, . . .) et premiers choix de travail sur les modalités
5. Extraction des seuils et écrêtement : isolement des extrêmes
6. Détermination de l'individu de référence (si GLM) ;

7. Création d'échantillons d'apprentissage et de validation ;
8. Modélisation (hypothèse, adéquation aux lois choisies, . . .) ;
9. Optimisation du modèle et travail manuel sur les variables et les modalités ;
10. Validation du modèle (résidus, comparaison à l'empirique sur l'échantillon de validation) ;
11. Détermination des primes ;
12. Viabilité des primes segmentées définies.

CONCLUSION

De nombreux écueils à la mise en place opérationnelle d'une tarification en assurance...

Principalement :

- travail sur les covariables (regroupement de modalités, catégorisation) en amont de la modélisation / optimisation ;
- la segmentation et ce qu'elle induit (attention à ne pas trop segmenter !) ;
- le choix paramétriques éventuels (lois, liens, ...) ;
- la calibration des modèles (convergence MLE, bornitude vraisemblance, initialisation de l'algo. Newton-Raphson, ...) ;

- la validation d'un modèle ;
- la gestion de la surdispersion des données ;
- la potentielle (très) faible sinistralité...

Il est primordial de bien être conscient de ces limites.

La qualité du tarif peut être apprécié par une courbe de Lorenz (en abscisses : % population triée par primes estimées classées par ordre décroissant, en ordonnées les pertes cumulées empiriques correspondantes...⇒ loi du 20-80 : 20% des contrats engendrent 80% de la perte globale)

Une alternative serait d'adopter une approche non-paramétrique ⇒ Machine Learning ([Paglia and Phelippe-Guinvarc'h, 2011], [Aouizerate, 2012], [Leroy and Planchet, 2016]...)

1 Tarification a priori - concepts avancés

2 Construction d'un zonier

3 Provisionnement

ARTICLES DE REFERENCE

[Boskov and Verrall, 1994]

[Taylor, 2001]

[Taylor, 1999]

[MATHIS, 2009]

[Brouhns et al., 2002]

[Mahy and Denuit, 2002]

1 Tarification a priori - concepts avancés

2 Construction d'un zonier
- Introduction
  - Généralités
  - Données à disposition
  - Pré-modélisation : une étape commune
- Zonier administratif
- Zonier par lissage spatial
- Zonier prédictif
- Conclusion

3 Provisionnement

DEFINITION D'UN ZONIER

Définition (Larousse) du mot "Zonier, zonière". *Adj, nom*.
Relatif à la zone autour de Paris ; habitant de cette zone.

Il semble que ce ne soit pas très adapté... En revanche, *zonaire* (adj.) est affecté à un nom et désigne un ensemble qui présente des zones.

Remarque importante : un véhiculier peut être rapproché d'un zonier : on explique la sinistralité spécifique par le type de véhicule. Par exemple, un certain modèle de moto est très présent dans les motos école. La fréquence de sinistre observée sera alors plus grande, cela est dû à l'utilisation du véhicule.

OBJECTIF PRINCIPAL

A la base, le zonier en assurance a été introduit essentiellement pour des raisons commerciales.

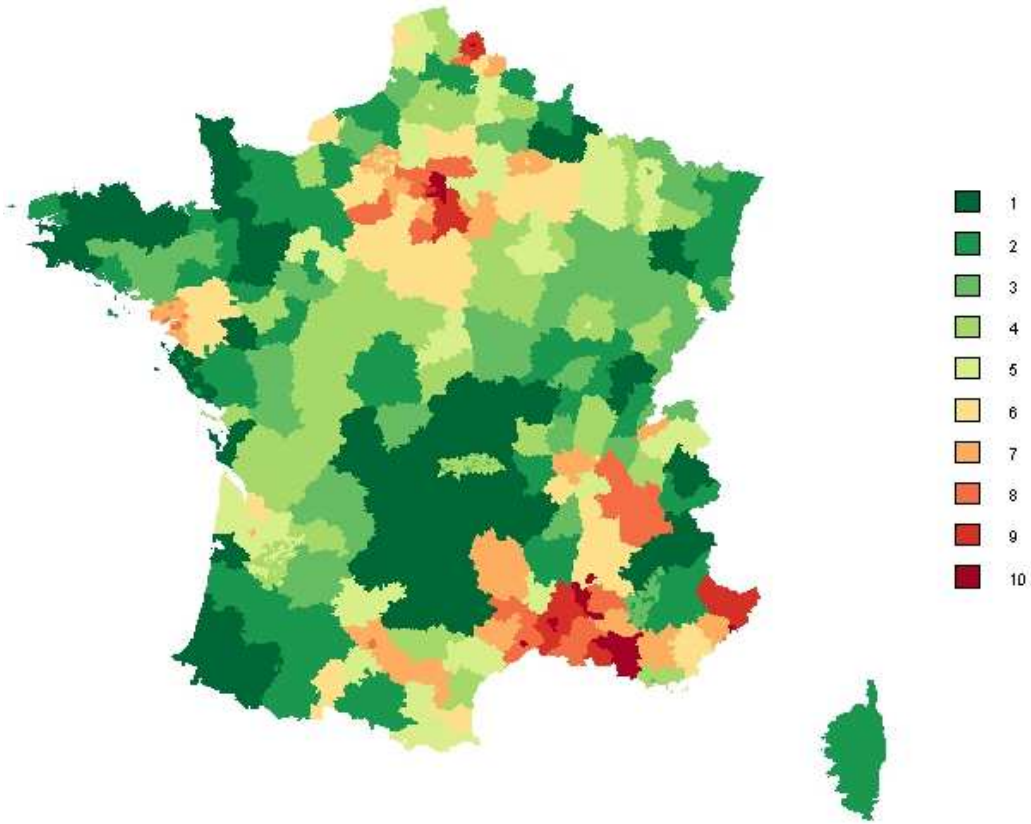
Objectif : éviter des “sauts” de tarif sur deux zones géographiques voisines, tous critères égaux par ailleurs.

$\Rightarrow$ Vente par les agents rendue plus facile... Moins de plaintes des assurés.

Autre avantage : création de classes de risque géographiques. On diminue le nombre de modalités par rapport à si l’on avait introduit la variable comme facteur de risque dans un modèle.

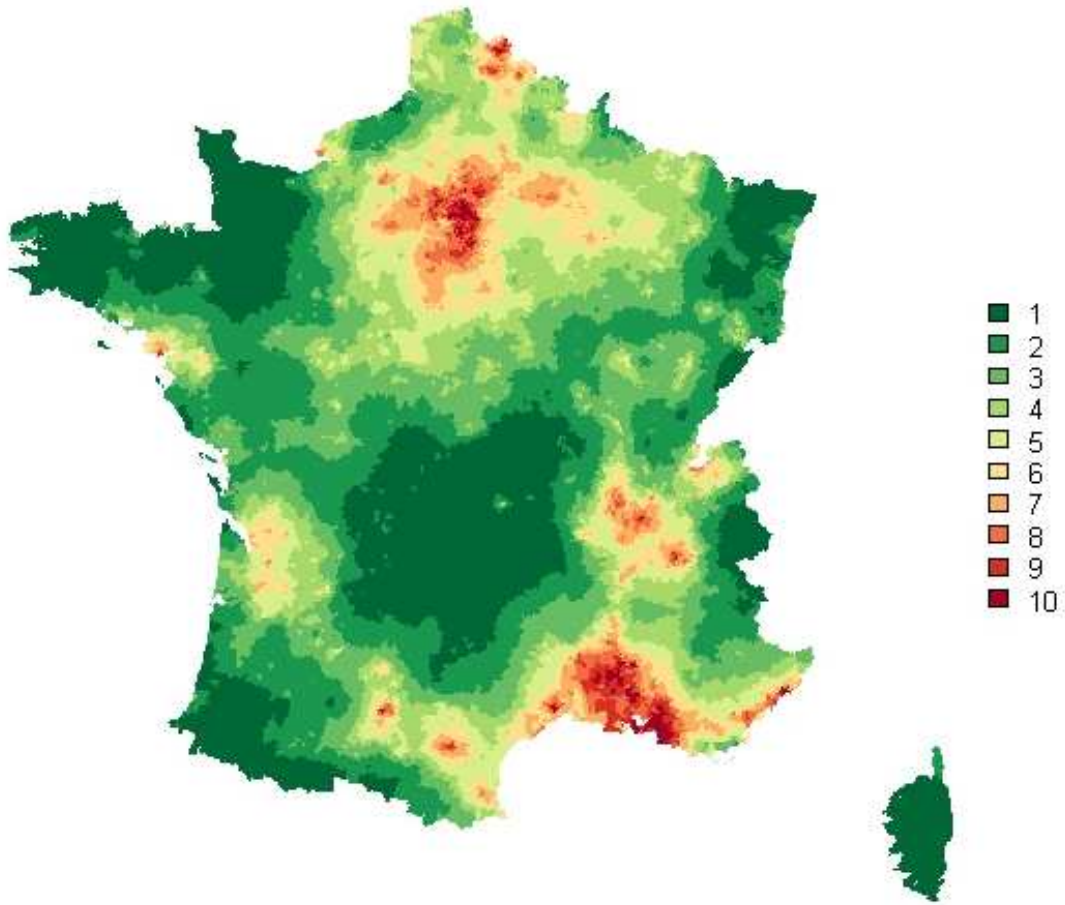
TYPES DE ZONIER

Le zonier administratif :



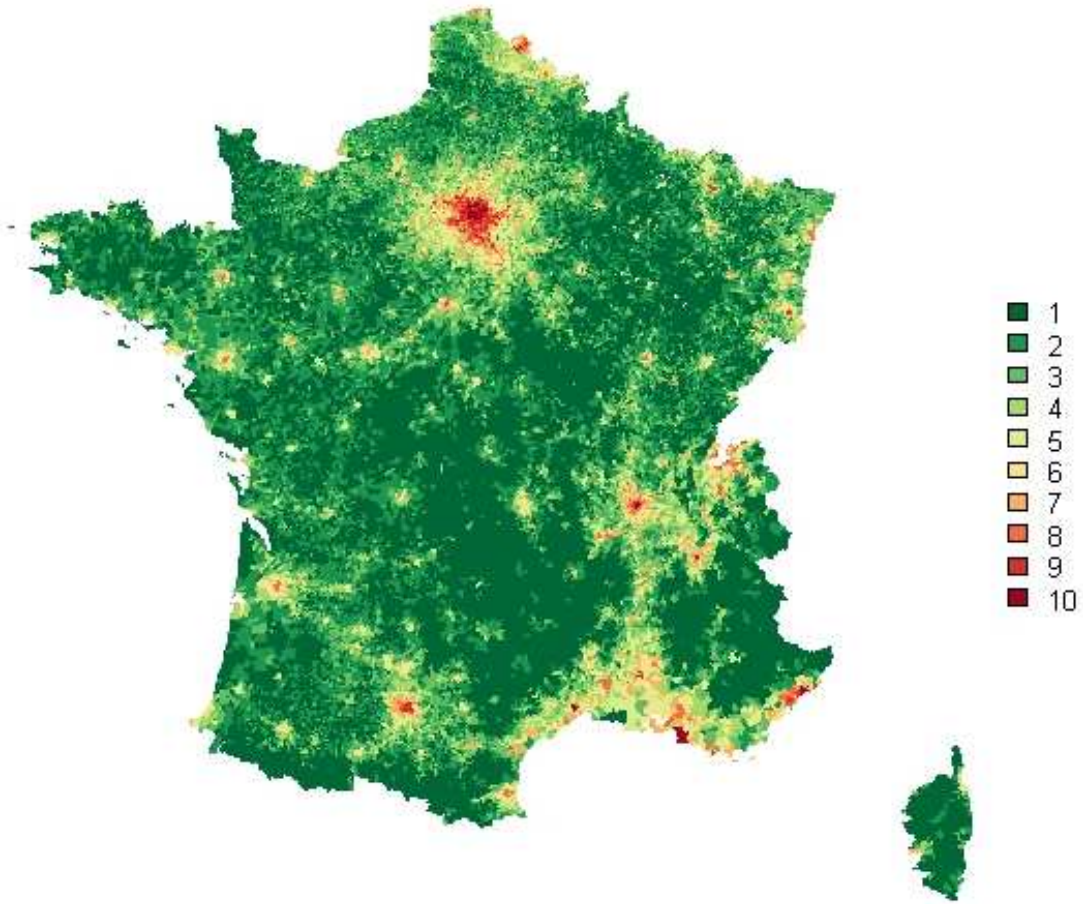
LISSAGE SPATIAL

Le zonier par lissage spatial :



ZONIER PREDICTIF

Le zonier prédictif :



CARACTERISTIQUES D'UN ZONIER

Les zoniers se construisent en général par garantie !

Exemples :

- garantie vol,
- garantie CAT NAT (zonier inondation, sécheresse),
- zonier santé (prix de la santé assez différent en fonction des régions),
- ...

Idée sous-jacente : le risque de vol est fortement lié au lieu d'habitation.
Agrégation de zoniers : question délicate !

2 Construction d'un zonier
- Introduction
  - Généralités
  - Données à disposition
  - Pré-modélisation : une étape commune
- Zonier administratif
  - Etapes de l'agrégation territoriale
  - Agrégation territoriale : choix du seuil d'exposition minimale
  - Zones de risque spatial : classification
- Zonier par lissage spatial
  - Le modèle de Boskov et Verrall
  - Algorithme de Gibbs
- Zonier prédictif
- Conclusion

SYSTEME D'INFORMATION GEOGRAPHIQUE (SIG)

Ensemble de données repérées ds l'espace (référence) : ex,

- données géographiques : un code postal, ... ;
- données localisées : nb de sinistres dans ce code postal.

On a des référentiels de données géographiques :

- Code Officiel Géographique (COG) : codification communes, cantons, arrondissements, départements, ..., DOM-TOM ;
- Référentiel GEOFLA : géré par Institut Géog. Natio. (IGN) ;
- Référentiels postaux : Hexaposte, Hexavia, Hexaclé, Hexaligne3, Cedexa ;
- Norme AFNOR : pour normaliser les adresses pour l'Europe.

2 Construction d'un zonier
- Introduction
  - Généralités
  - Données à disposition
  - Pré-modélisation : une étape commune
- Zonier administratif
  - Etapes de l'agrégation territoriale
  - Agrégation territoriale : choix du seuil d'exposition minimale
  - Zones de risque spatial : classification
- Zonier par lissage spatial
  - Le modèle de Boskov et Verrall
  - Algorithme de Gibbs
- Zonier prédictif
- Conclusion

STRATEGIE GENERALE

En commun de toutes les techniques de zonage, il existe une étape préliminaire permettant **d'"isoler"** l'effet du risque géographique.

Considérons par exemple un modèle de fréquence. On note N_i le nombre de sinistres de l'individu i, et on connait son exposition notée e_i.

Supposons que $N_i \sim \mathcal{P}(\lambda_i)$. En spécifiant un GLM log-Poisson :

$$\ln(E[N_i \mid X_i]) = \log(e_i) + \beta_0 + X_i^T \beta,$$

avec $\beta^T = (\beta_1, ..., \beta_p)$, et $X_i^T = (X_1^i, ..., X_p^i)$.

Dans un cas classique, X contient une information sur le lieu où se trouve le risque (ex : lieu d'habitation).

On constituera comme dans le cadre général

- un échantillon d'apprentissage pour construire du modèle,
- un échantillon de validation pour valider le modèle.

On peut procéder par échantillonnage stratifié (sur l'expo. par ex. : 2/3 de l'expo dans l'éch. d'apprentissage et 1/3 ds validation) : l'idéal est d'avoir une exposition uniformément répartie sur le territoire (parfois utopique !).

Stratégie pour construire un zonier : ne pas intégrer le facteur de risque géographique dans la calibration du modèle, puis travailler sur les résidus pour faire ressortir cet effet.

Les méthodes de zonage consistent à mesurer le niveau de risque par “région” $\Rightarrow$ on obtient une partition en zones de risque homogène.

Point de vocabulaire : on distingue dans les méthodes de zonage deux types de données :

- les données laticielles : données observées sur une partition du territoire (ex : exposition par commune) ;
- les données ponctuelles : données géocodées (ex : ensemble de sinistres à des lieux précis).

MANQUE D'INFORMATION

Une des principales problématiques concerne le manque d'information.
Exemple : si la "région" considérée est une commune, on peut ne pas disposer d'information à ce niveau.

Comment mesurer alors le risque relatif à cette commune ?

Cela dépend du type de zonier que nous construisons :

- avec un zonier administratif, il faudrait considérer une "région" plus grande, et accentuer ainsi la mutualisation. Cela induit :
  - \- une perte de précision dans le zonier,
  - \+ un gain dans la robustesse de la mesure du risque car on a plus de données ;

- procéder par lissage spatial (on mutualise les risques proches, ex : Boskov-Verrall (1994), Taylor (2001)). Cela induit notamment :
  - \+ une extraction des petites fluctuations aléatoires du risque pour en révéler la structure spatiale sous-jacente.
  - \- une difficulté de calibration pour les paramètres de lissage, difficulté d'arbitrer dans le niveau de précision du zonage.

- procéder par introduction de variables externes prédictives du risque géographique (sociodémographiques, topographiques, de population, ...). Cela induit :
  - \+ on peut extrapoler le niveau de risque d'une région non exposée à partir de ses caractéristiques,
  - \- choix complexe dans la multitude des indicateurs potentiels pour la construction du modèle.

1 Tarification a priori - concepts avancés

2 Construction d'un zonier
- Introduction
- Zonier administratif
  - Etapes de l'agrégation territoriale
  - Agrégation territoriale : choix du seuil d'exposition minimale
  - Zones de risque spatial : classification
- Zonier par lissage spatial
- Zonier prédictif
- Conclusion

3 Provisionnement

CONTEXTE

Principe du zonier administratif : le zonier administratif correspond à un zonage par agrégation territoriale.

On prendra ici l'exemple d'un zonier fréquence (mais il existe des zoniers de coût aussi !).

Evidemment, il existe d'autres facteurs de risque que la région expliquant la fréquence $\Rightarrow$ trouver une mesure du niveau de risque d'une région qui ne dépende **que** du facteur spatial (isoler l'influence du facteur géo. toutes choses égales par ailleurs).

Rappelons que

$$N_i = e_i \times e^{\beta_0} \times e^{\beta_1 X_1^i} \times ... \times e^{\beta_p X_p^i} + \epsilon_i$$

2 Construction d'un zonier
- Introduction
  - Généralités
  - Données à disposition
  - Pré-modélisation : une étape commune
- Zonier administratif
  - Etapes de l'agrégation territoriale
  - Agrégation territoriale : choix du seuil d'exposition minimale
  - Zones de risque spatial : classification
- Zonier par lissage spatial
  - Le modèle de Boskov et Verrall
  - Algorithme de Gibbs
- Zonier prédictif
- Conclusion

FACTEUR DE RISQUE SPATIAL

Soit X_1 le critère géographique, alors β_1 est le facteur spatial.

1. On modélise N sans $X_1 \Rightarrow$ on obtient $\hat{\beta}_2, ..., \hat{\beta}_p$ (si GLM) ;
2. $\rightarrow$ Lorsque l'exposition est différente de 0, on pose :

$$R_i = \frac{N_i}{e_i \times e^{\hat{\beta}_2 X_2^i} \times ... \times e^{\hat{\beta}_p X_p^i}} = e^{\beta_0} \times e^{\beta_1 X_1^i} \times \frac{e^{\beta_2 X_2^i}}{e^{\hat{\beta}_2 X_2^i}} \times ... \times \frac{e^{\beta_p X_p^i}}{e^{\hat{\beta}_p X_p^i}}.$$

Sous l'hyp. $\hat{\beta}_2 = \beta_2, ..., \hat{\beta}_p = \beta_p$, on définit le risque spatial par

$$R_i = e^{\beta_0} e^{\beta_1 X_1^i} + \epsilon' ;$$

$\rightarrow$ Lorsque l'exposition est nulle, on prend $R_i = 0$.
On appelle R_i le **risque spatial résiduel**.

Rq : dans la suite, k est une “région”. En pratique, on aura donc déjà agrégé les observations des assurés par “région”.
Rq 2 : on aurait aussi pu considérer $N_i - \hat{N}_i$ plutôt que $N_i \,/\, \hat{N}_i$.

3. Admettons que nous travaillons au niveau commune ici. On peut déduire l'estimateur $\hat{r}_i$ de R_i pour chaque assuré :

$$\hat{r}_i = \frac{n_i}{offset_i \; e^{\hat{\beta}_2 x_2^i} ... e^{\hat{\beta}_p x_p^i}}$$

où n_i est le nombre de sinistres observés pour l'assuré i.
On peut maintenant définir l'estimateur du risque spatial résiduel au niveau de la commune k par

$$\hat{r}_k^c = \frac{\sum_{i=1}^{I^k} e_i \, \hat{r}_i}{\sum_{i=1}^{I^k} e_i},$$

avec e_i l'exposition, I^k nb assurés ds la commune k.

4. On créé une nouvelle base de données où chaque ligne est une commune, avec :
   - un code commune fourni par l'INSEE par exemple,
   - l'exposition e_k de cette commune k,
   - le risque résiduel spatial $\hat{r}_k^c$,
   - le nombre de sinistres prédits, $\hat{n}_k^c$.

5. Enfin, on procède à l'agrégation territoriale au besoin.
   $\rightarrow$ Si le niveau choisi est trop fin (pas d'exposition), on agrège alors au niveau d'au-dessus (ici le département par exemple).

Idée générale : la statistique de risque spatial résiduel doit pouvoir être considérée robuste. Elle doit donc excéder un certain seuil d'exposition minimal, noté e dans la suite.

2 Construction d'un zonier
- Introduction
  - Généralités
  - Données à disposition
  - Pré-modélisation : une étape commune
- Zonier administratif
  - Etapes de l'agrégation territoriale
  - Agrégation territoriale : choix du seuil d'exposition minimale
  - Zones de risque spatial : classification
- Zonier par lissage spatial
  - Le modèle de Boskov et Verrall
  - Algorithme de Gibbs
- Zonier prédictif
- Conclusion

CHOIX DU SEUIL D'EXPOSITION MINIMALE

On se rend compte que le risque spatial résiduel de chaque commune k peut correspondre :

- soit à son propre risque spatial résiduel évalué comme décrit précédemment,
- soit au risque spatial résiduel du niveau d'agrégation au-dessus (si l'exposition était trop faible),
- soit au risque spatial résiduel du niveau d'agrégation encore au-dessus si cette dernière exposition n'était pas suffisante, et ainsi de suite...

On stocke pr chq commune les niveaux de risque spatiaux possibles suivant niveau d'agrégation (commune, canton, ...).

$\Rightarrow$ On se sert de ce tableau pour définir le seuil d'exposition minimale, noté e dans la suite.

En résumé donc, on procède comme suit :

1. sur la base d'apprentissage A, construire GLM puis calculer le niveau spatial résiduel par commune $\hat{r}_k^A$ (r_k^c précédemment) ;
2. sur la base de validation, on estime également le niveau de risque spatial résiduel $\hat{r}_k^T$ par commune ;
3. pour trouver le seuil d'expo. minimale e, on optimise

$$\min_e \left(\sum_{k=1}^{n} e \, (\hat{r}_k^T(e) - \hat{r}_k^A(e))^2 \right)$$

avec n le nombre de communes.

En pratique, suivant la valeur de e, $\hat{r}_k^A$ et $\hat{r}_k^T$ diffèrent.

Pour tester $\neq$ valeurs de e, on se définit une grille de valeurs possibles (par ex. de l'exposition minimale à l'exposition maximale avec un certain pas).

Si $e_k > e$, on conserve les risques spatiaux résiduels courants $\hat{r}_k^A$ et $\hat{r}_k^T$. Sinon, on prend les valeurs pour l'agrégation d'au-dessus.

2 Construction d'un zonier
- Introduction
  - Généralités
  - Données à disposition
  - Pré-modélisation : une étape commune
- Zonier administratif
  - Etapes de l'agrégation territoriale
  - Agrégation territoriale : choix du seuil d'exposition minimale
  - Zones de risque spatial : classification
- Zonier par lissage spatial
  - Le modèle de Boskov et Verrall
  - Algorithme de Gibbs
- Zonier prédictif
- Conclusion

CLASSIFICATION PAR ZONE

Une fois e déterminé, on ré-affecte le bon niveau de risque spatial résiduel pour chq commune (celui de la commune si $e_k > e$, sinon au niveau d'agrégation supérieur tel que expo $> e$).

Cette affectation est réalisé pour l'ensemble des données (apprentissage et validation).

Les niveaux de risque par commune ont maintenant été calculés : il faut regrouper les communes avec niveau de risque similaire afin d'avoir un zonier.

En fonction du nombre de zones voulu (disons Z zones), on peut faire une classification en Z classes.

En général, cette classification se fait par quantile d'exposition : on veut créer Z classes avec même niveau d'exposition.

Notons a l'exposition de chacune des classes créées, ainsi

$$a = \frac{\textit{expo totale}}{Z}.$$

En pratique, on veut satisfaire le critère "avoir au moins a en termes d'exposition".

Concrètement, la 1ère classe contient l'ensemble des communes avec plus faible niveau de risque spatial dont la somme des expositions soit au moins égale à a, et ainsi de suite.

On obtient ainsi la carte du zonier avec Z couleurs...

1 Tarification a priori - concepts avancés

2 Construction d'un zonier
- Introduction
- Zonier administratif
- Zonier par lissage spatial
  - Le modèle de Boskov et Verrall
  - Algorithme de Gibbs
- Zonier prédictif
- Conclusion

3 Provisionnement

DIFFERENCE PRINCIPALE DE LA METHODE

L'approche par lissage spatial ([J. Besag and Mollie, 1991], [Boskov and Verrall, 1994], [Taylor, 2001], ...) a toujours pour but d'estimer le facteur de risque spatial d'une région.

Ici le zonier **ne correspond pas** à une découpe administrative.

Le principe de base est de considérer la sinistralité liée à un lieu ainsi que celle des "régions" alentours.

Hypothèse implicite de cette approche : 2 régions proches ont des facteurs de risque spatiaux similaires.

2 Construction d'un zonier
- Introduction
  - Généralités
  - Données à disposition
  - Pré-modélisation : une étape commune
- Zonier administratif
  - Etapes de l'agrégation territoriale
  - Agrégation territoriale : choix du seuil d'exposition minimale
  - Zones de risque spatial : classification
- Zonier par lissage spatial
  - Le modèle de Boskov et Verrall
  - Algorithme de Gibbs
- Zonier prédictif
- Conclusion

MODELE BAYESIEN DE BOSKOV-VERRALL [Boskov and Verrall, 1994]

Modèle de référence pour ceux voulant mettre en oeuvre une approche basée sur l'expérience (mise à jour paramètres).

On reprend les notations et le modèle précédent :

- $N = (N_1, N_2, ..., N_r) = (N_i)_{1 \leq i \leq r}$;
- N_i : nb sinistres ds "région" i, n_i est la version observée ;
- $i \in \{1, 2, ..., r\}$: il y a r "régions" ;
- e_i est l'exposition de la région i,
- un modèle GLM log-Poisson pr le nb de sinistres :

$$N_i = e^{\ln(e_i)} \times e^{\eta_i + \mu_i + \nu_i}.$$

LES TERMES DU MODELE

On peut donner une interprétation aux différents termes du modèle de Boskov et Verrall.

Signification de chacun des termes de la modélisation :

- η_i représente les facteurs de risque non spatiaux (âge, ...) ;
- μ_i est l'effet du risque spatial ;
- ν_i sont les résidus du modèle.

Ainsi, on décompose les différents effets en fonction de leur aspect spatial ou non.

ETAPES DE MODELISATION

On liste ici les étapes nécessaires à la mise en place du modèle de Boskov-Verrall, dont voici un résumé.

1. On estime un GLM log-Poisson sans μ_i, le facteur de risque spatial. On obtient ainsi $\hat{\eta}_i$.

2. Il reste deux quantités aléatoires dans le modèle d'origine :
   - μ_i pour l'effet du risque spatial,
   - ν_i pour les résidus du modèle.

   On doit maintenant spécifier des distributions de probabilités **a priori** pour ces 2 quantités (pour utiliser la théorie bayésienne).

3. Supposons que :

- **l'effet spatial** μ_i est lissé, en introduisant une dépendance spatiale entre les régions voisines.

  Notons δ_i l'ensemble des régions dans le voisinage de la région i.

  Une loi possible peut être

  $$\mu_i \sim U_i \sim \mathcal{L}(\tau), \quad \text{avec } f(\mu_i;\tau) \sim \tau^{-1/2}\, e^{-\frac{1}{2\tau}\sum_{j\in\delta_i}(\mu_i-\mu_j)^2},$$

  (Ressemble à un noyau gaussien centré sur la région i)

  $\Rightarrow$ Seuls les voisins ont donc une influence sur la densité (on pourrait même introduire une dépendance en fonction de la distance entre "région").

Pour trouver la loi du vecteur, on tient compte de cette dépendance.

Donc

$$\begin{aligned} f(\mu;\tau) &= f((\mu_1,\mu_2,...,\mu_r);\tau) \\ &= f(\mu_r\,|\,\mu_1,...,\mu_{r-1};\tau) f((\mu_1,...,\mu_{r-1});\tau) \\ &= f(\mu_r\,|\,\mu_1,...,\mu_{r-1};\tau) f(\mu_{r-1}\,|\,\mu_1,...,\mu_{r-2};\tau) f((\mu_1,...,\mu_{r-2});\tau) \\ &= ... \\ &= f(\mu_r\,|\,\mu_1,...,\mu_{r-1};\tau)\,...\,f(\mu_2|\mu_1;\tau) f(\mu_1;\tau) \\ &= \tau^{-r/2}\, e^{-\frac{1}{2\tau}\sum_{i\sim j}(\mu_i-\mu_j)^2} \end{aligned}$$

où $i \sim j$ désigne l'ensemble des couples (i,j) voisins.

- les résidus ν_i sont **indépendants**, centrés, et de type gaussien, i.e.

$$\nu_i \sim V_i \sim \mathcal{L}(\lambda), \quad \text{avec } f(\nu_i;\lambda) \sim \lambda^{-1/2}\, e^{-\frac{1}{2\lambda}\nu_i^2}.$$

On obtient donc

$$f(\nu;\lambda) = \prod_i f(\nu_i;\lambda) \sim \lambda^{-r/2}\, e^{-\frac{1}{2\lambda}\sum_{i=1}^r \nu_i^2}.$$

- la loi a priori des paramètres est donnée par

$$(\tau,\Lambda) \sim \mathcal{L}(\xi), \quad \text{avec } f(\tau,\lambda;\xi) = e^{-\frac{\xi}{2\tau}-\frac{\xi}{2\lambda}},$$

avec $\xi > 0$ et petit.
C'est une distribution dite "peu informative" (donne peu d'information sur la distribution du paramètre a priori).

4. On sait que $N_i \,|\, \eta_i, \mu_i \sim \mathcal{P}(e_i \times e^{\eta_i + \mu_i})$.

   Donc $N \sim \mathcal{L}(\Theta)$ avec $\Theta = (U, V, \tau, \Lambda)$.

5. Détermination de la loi a posteriori des paramètres.
   Pour prédire le nombre de sinistres, on cherche la loi de

$$N \,|\, (U, V, \tau, \Lambda).$$

   Notons $(U, V) = ((\mu_1, \nu_1), (\mu_2, \nu_2), ..., (\mu_r, \nu_r))$

$$P(N|(U,V)) = \frac{P(N,(U,V))}{P(U,V)} = \frac{P((U,V)|N)P(N)}{P(U,V)} = \frac{P((U,V)|N)P(N)}{\sum_n P((U,V)|N)P(N)}$$

   Ce qui nous amène à devoir connaitre la loi de $(U, V) \,|\, N$ (ou $(U, V, \tau, \Lambda) \,|\, N$ puisque (U, V) dépend de (τ, Λ)...)

Problème : $(U, V, \tau, \Lambda) \,|\, N$ n'a pas de forme connue.

En effet,

$$
\begin{aligned}
f(\mu, \nu, \tau, \lambda \,|\, n) &\sim P(N_1 = n_1, ..., N_r = n_r \,|\, U = \mu, V = \nu, \tau = \tau, \Lambda = \lambda)\, f(\mu, \nu, \tau, \lambda) \\
&= P(N_1 = n_1, ..., N_r = n_r \,|\, \mu, \nu, \tau, \lambda)\, f(\mu, \nu \,|\, \tau, \lambda)\, f(\tau, \lambda) \\
&= P(N_1 = n_1, ..., N_r = n_r \,|\, \mu, \nu, \tau, \lambda)\, f(\mu \,|\, \tau = \tau)\, f(\nu \,|\, \Lambda = \lambda)\, f(\tau, \lambda) \\
&= \prod_{i=1}^{r} P(N_i = n_i \,|\, \mu_i, \nu_i, \tau, \lambda)\, f(\mu \,|\, \tau = \tau)\, f(\nu \,|\, \Lambda = \lambda)\, f(\tau, \lambda) \\
&= \prod_{i=1}^{r} e^{-\theta_i} \frac{\theta_i^{n_i}}{n_i!}\, f(\mu \,|\, \tau = \tau)\, f(\nu \,|\, \Lambda = \lambda)\, f(\tau, \lambda) \quad \text{(forme inconnue !)}
\end{aligned}
$$

$\Rightarrow$ On a besoin d'une méthode type Monte Carlo Markov Chain (MCMC).

RESUME DU RAISONNEMENT

Voici donc les étapes qui conduisent au résultat :

1. On spécifie les lois a priori :
   - couple de paramètres (τ, Λ) (loi peu informative) ;
   - la dépendance spatiale via la loi de $\mu_i \sim U_i \sim \mathcal{L}(\tau)$;
   - le bruit (résidus) via la loi de $\nu_i \sim V_i \sim \mathcal{L}(\lambda)$;
   - le nombre de sinistres via la loi de $N \sim \mathcal{L}(\Theta)$ avec $\Theta = (\tau, \Lambda, U, V)$, plus précisément

$$N \sim \mathcal{P}(E[N]) \sim \mathcal{P}(exposition \times e^{\hat{\eta}+U+V})$$

2. On cherche la loi a posteriori des paramètres via l'échantillonneur de Gibbs (méthode MCMC, algo. de metropolis-Hastings).

2 Construction d'un zonier
- Introduction
  - Généralités
  - Données à disposition
  - Pré-modélisation : une étape commune
- Zonier administratif
  - Etapes de l'agrégation territoriale
  - Agrégation territoriale : choix du seuil d'exposition minimale
  - Zones de risque spatial : classification
- Zonier par lissage spatial
  - Le modèle de Boskov et Verrall
  - Algorithme de Gibbs
- Zonier prédictif
- Conclusion

ECHANTILLONNEUR DE GIBBS

On utilise l'échantillonneur de Gibbs car :

- les lois ne sont pas conjuguées : loi a posteriori $\neq$ loi a priori (pas seulement mise à jour des param.) ;
- on connait les densités univariées conditionnelles ;
- il n'est pas possible de trouver explicitement la loi a posteriori.

L'échantillonneur de Gibbs va permettre de déterminer un échantillon de la densité a posteriori.

Principe : exploiter les densités conditionnelles (simu d'1 fonction multivariée décomposable en $+$sieurs simus fonctions univariées) : `https://www.youtube.com/watch?v=ER3DDBFzH2g`

Donnons nous un vecteur pour nos variables aléatoires :

$$X = (\tau, \Lambda, \mu_1, \mu_2, ..., \mu_r, \nu_1, \nu_2, ..., \nu_r, N)$$

de densité conditionnelle $f(\tau, \lambda, \mu, \nu \,|\, n)$ (densité a posteriori).

On a donc observé un nombre de sinistres n, et on cherche l'information que cela peut nous amener sur les autres paramètres. **L'échantillonneur de Gibbs permet d'obtenir des réalisations de X.**

C'est une procédure itérative où l'on va fixer tous les paramètres sauf un : celui-ci est tiré au sort avec la distribution associée, puis on actualise !

ALGORITHME : MISE EN PRATIQUE

A partir de l'étape k, on tire pour l'étape $(k+1)$:

1. $\tau^{(k+1)} \sim f(\tau \mid \lambda^{(k)}, \mu^{(k)}, \nu^{(k)}, n)$
2. $\lambda^{(k+1)} \sim f(\lambda \mid \tau^{(k+1)}, \mu^{(k)}, \nu^{(k)}, n)$
3. $\mu_1^{(k+1)} \sim f(\mu_1 \mid \lambda^{(k+1)}, \tau^{(k+1)}, \mu_{-1}^{(k)}, \nu^{(k)}, n)$, où $\mu_{-1}^{(k)} = (\mu_2^{(k)}, ..., \mu_r^{(k)})$

   ...

   $\mu_r^{(k+1)} \sim f(\mu_r \mid \lambda^{(k+1)}, \tau^{(k+1)}, \mu_{-r}^{(k+1)}, \nu^{(k)}, n)$

4. $\nu_1^{(k+1)} \sim f(\nu_1 \mid \lambda^{(k+1)}, \tau^{(k+1)}, \mu^{(k+1)}, \nu_{-1}^{(k)}, n)$

   ...

   $\nu_r^{(k+1)} \sim f(\nu_r \mid \lambda^{(k+1)}, \tau^{(k+1)}, \mu^{(k+1)}, \nu_{-r}^{(k+1)}, n)$

Il faut donc fixer des valeurs initiales en définissant un vecteur $X^{(0)}$.

Après l'étape k, l'étape $(k+1)$ se finit quand les $(2r+2)$ valeurs ont été simulées, donnant

$$X^{(k+1)} = (\tau^{(k+1)}, \lambda^{(k+1)}, \mu_1^{(k+1)}, ..., \mu_r^{(k+1)}, \nu_1^{(k+1)}, ..., \nu_r^{(k+1)}, n).$$

On vient donc d'obtenir un nouvel état de la chaine de Markov. Ce nouvel état est un nouveau jeu de paramètres, donc une nouvelle observation de la densité a posteriori.

Cette chaine de Markov converge vers une distribution stationnaire $f(\tau, \lambda, \mu, \nu \,|\, n)$ après quelques centaines / milliers d'itérations.

La valeur des états qui suivent cette convergence permettent de construire la densité empirique conditionnelle recherchée.
La chaine simulée à K états s'écrit alors

$$\left\{(\tau^{(0)}, \lambda^{(0)}, \mu^{(0)}, \nu^{(0)}, n), ..., (\tau^{(K)}, \lambda^{(K)}, \mu^{(K)}, \nu^{(K)}, n)\right\}$$

En supposant que cette chaine de Markov atteint son état stationnaire après T itérations, l'estimateur de la densité sera donné par une approximation Monte Carlo, i.e.

$$\hat{X} = \frac{1}{K - T} \sum_{T+1}^{K} X^{(k)}.$$

On remarque cependant que la simulation des densités conditionnelles univariées nécessite de les connaitre !

Par exemple,

$$\mu_i^{(k+1)} \sim f(\mu_i \,|\, \tau^{(k+1)}, \lambda^{(k+1)}, \mu_{-i}^{(k+1)}, \nu^{(k)}, n)$$

Avec les choix faits ici pour les lois a priori, les densités conditionnelles univariées sont fournies dans [J. Besag and Mollie, 1991].

Par exemple, la loi qui permet de mettre à jour le paramètre τ est explicite (loi du χ^2).

Dans les cas où on ne peut les déterminer, elles sont alors évaluées via l'algorithme *Adaptive Rejection Sampling* (ARS).

1 Tarification a priori - concepts avancés

2 Construction d'un zonier

- Introduction
- Zonier administratif
- Zonier par lissage spatial
- Zonier prédictif
- Conclusion

3 Provisionnement

CONSTRUCTION D'UN ZONIER PREDICTIF

Comme dans le cadre du zonier administratif, on isole l'effet du risque géographique via la construction en amont d'un GLM ne contenant pas de facteur de risque géographique...

Logiquement, on procède ensuite de la manière suivante :

1. on récupère les résidus du modèle,
2. on essaie de construire un modèle prédictif de ces résidus (par ex. un autre GLM) avec des variables explicatives pertinentes.

Question clef : choix des variables explicatives du risque géographique.

Inconvénient : si le choix n'est pas judicieux...

Avantage : on peut effectuer des prévisions pour de nouvelles zones non exposées et sur lesquelles on ne détient pas d'historique de sinistralité...

1 Tarification a priori - concepts avancés

2 Construction d'un zonier

- Introduction
- Zonier administratif
- Zonier par lissage spatial
- Zonier prédictif
- Conclusion

3 Provisionnement

CONCLUSION

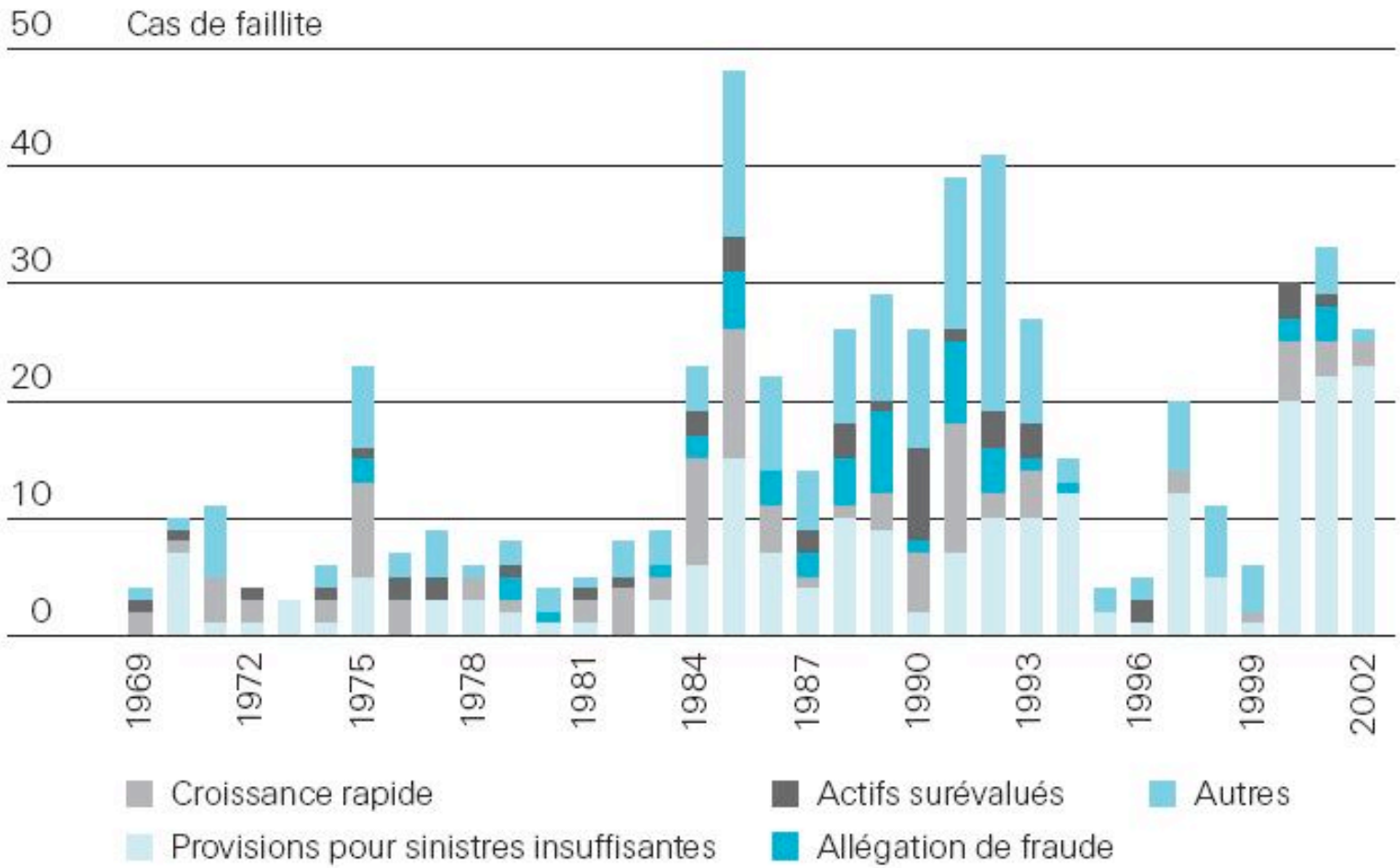
- Il existe 3 grandes manières de construire un zonier.
- Les modèles décrits ici ne sont **pas exhaustifs** et certains acteurs en utilisent des variantes (par ex. classifier directement suivant la taille des résidus après la première modélisation).
- Certains modèles nécessitent une maitrise technique importante (lissage spatial de Boskov et Verrall), ou une connaissance du risque affinée pour le choix des paramètres (lissage spatial de Wittaker).

1 Tarification a priori - concepts avancés

2 Construction d'un zonier

3 Provisionnement

IMPORTANCE DU CALCUL DES PSAP



Source : A.M. Best : Best's Insolvency Study, Property/Casualty U.S. Insurers 1969–2002, mai 2004, p. 34

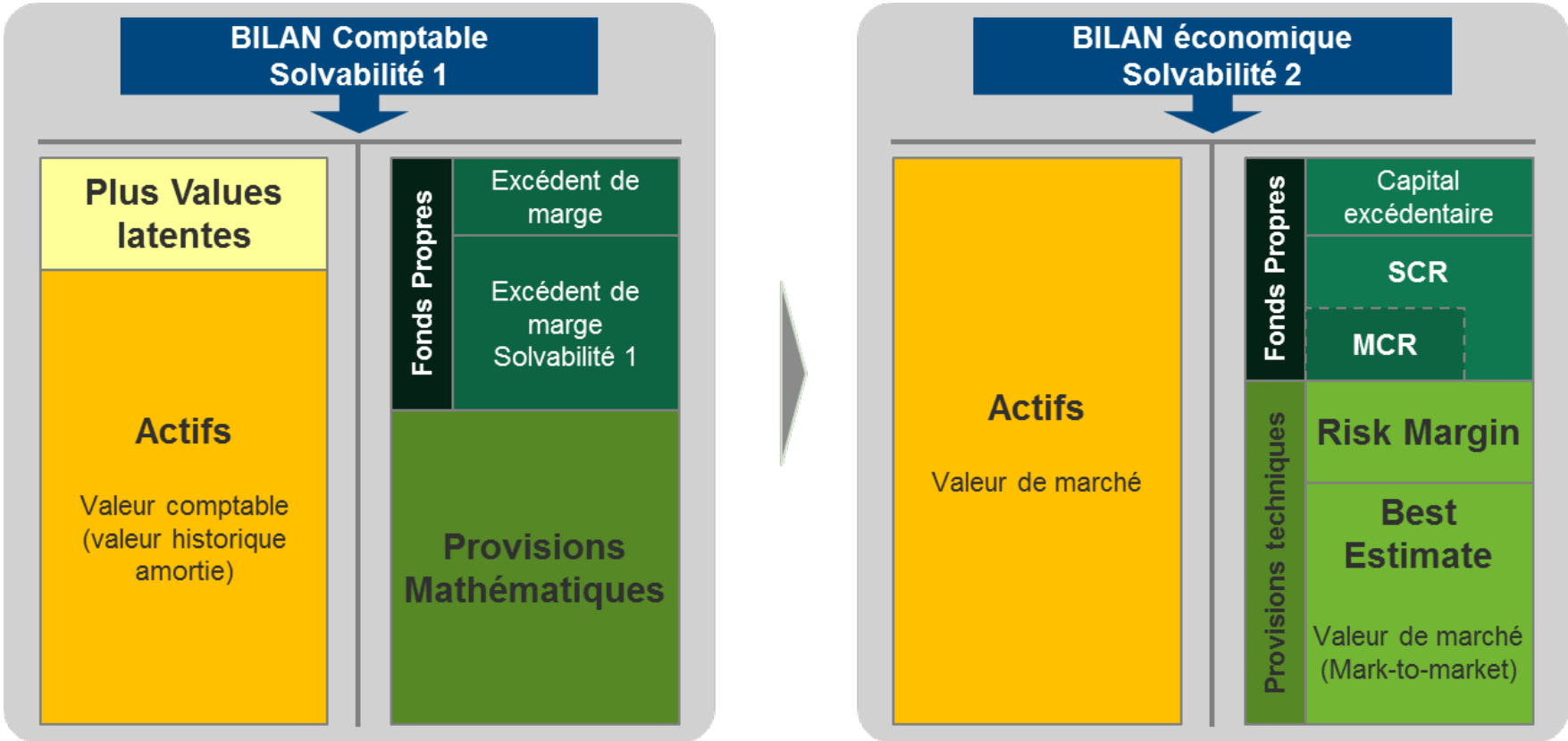
1 Tarification a priori - concepts avancés

2 Construction d'un zonier

3 Provisionnement

- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
- Illustrations de l'intérêt de la méthode sur des cas pratiques

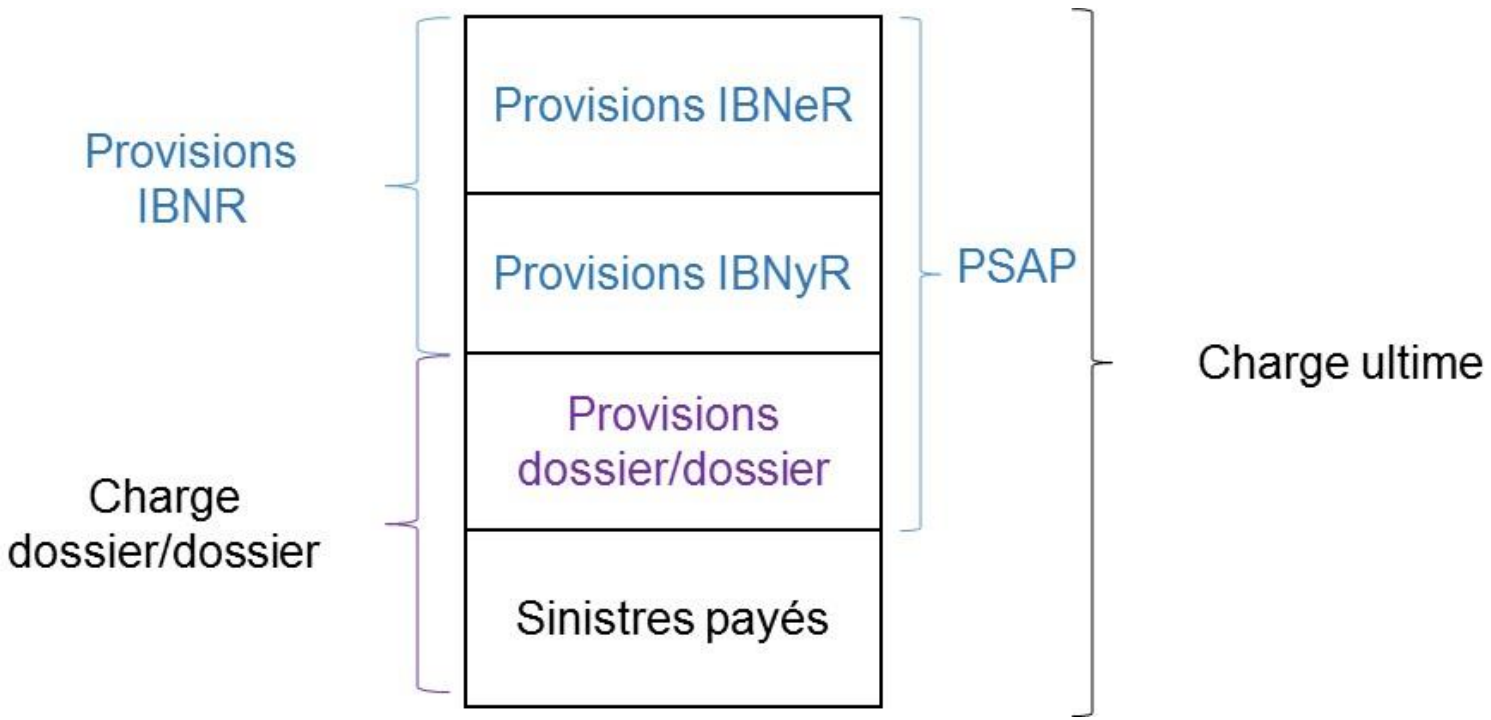
BILAN SOLVABILITE 2 ET PROVISION



Interactions actif-passif ⇒ BEL/PM varie !

SCR provision lié à la variation du BEL d'un exercice sur l'autre.

DECOMPOSITION DE LA CHARGE D'UN SINISTRE



TECHNIQUES DE PROVISIONNEMENT

Il y a 2 grandes approches pour calculer les provisions.

1. Modèles sur données agrégées (ex : **Chain Ladder**) : macrolevel reserving.
   - Travail sur paiements stockés par période de survenance *i* et délai de réglement *j*.
   - Hypothèse sous-jacente : stationarité.
2. Estimation par sinistre : microlevel reserving !
   - Utilisation des caractéristiques des sinistres pour les sinistres en cours de paiement.
   - Anticipation des tardifs (non encore déclarés).

1. Tarification a priori - concepts avancés

2. Construction d'un zonier

3. Provisionnement
   - Problématique du provisionnement
   - Données agrégées et triangle de liquidation
   - Provisionnement stochastique MCMC
   - Mise en lumière des limites de ces modèles
   - Provisionnement ligne-à-ligne (microlevel reserving)
   - Illustrations de l'intérêt de la méthode sur des cas pratiques

CAS DE MODELES SUR DONNEES AGREGEES

Au 31/12/I, les données sont stockées dans un triangle de liquidation :

Année de survenance	0	1	...	Années j	de ...	développement $J-i$	...	$J-1$	J
0	$x_{0,0}$	$x_{0,1}$	...	$x_{0,j}$	...	...	...	$x_{0,J-1}$	$x_{0,J}$
1	$x_{1,0}$	$x_{1,1}$	...	$x_{1,j}$				$x_{1,J-1}$	
⋮	⋮	⋮	...	⋮	...		⋱		
i	...	...	...	$x_{i,j}$	...	$x_{i,J-i}$			
⋮	⋮	⋮		⋮					
$I-j$	...	...	...	$x_{I-j,j}$					
⋮									
$I-1$	$x_{I-1,0}$	$x_{I-1,1}$							
I	$x_{I,0}$								

EXEMPLE : CHAIN LADDER SUR DONNEES CUMULEES

Année de survenance	Années de développement					
	0	1	2	3	4	5
1988	3209	4372	4411	4428	4435	4456
1989	3367	4659	4696	4720	4730	
1990	3871	5345	5398	5420		
1991	4239	5917	6020			
1992	4929	6794				
1993	5217					

Ce qui donne les facteurs communs de développement

j	0	1	2	3	4
	(0-1)	(1-2)	(2-3)	(3-4)	(4-5)
f_j	1.38	1.01	1.0043	1.0018	1.0047

Et les cadences cumulées de règlement :

j	0	1	2	3	4	5
pc_j	70.8	97.8	98.9	99.3	99.5	100

On en déduit le triangle inférieur de liquidation et les provisions

Exercice	i	0	1	2	3	4	5	Provisions
1988	0						4456	0
1989	1					4730	4752	22
1990	2				5420	5430	5456	36
1991	3			6020	6046	6057	6086	66
1992	4		6794	6872	6902	6914	6947	153
1993	5	5217	7204	7287	7318	7332	7367	2150
							Total	2427

Rq : dernière prov. représente 89% de la prov. globale (short-tail).

EXTENSIONS

On dénombre bc de méthodes dérivant du modèle déterministe de Chain Ladder, afin d'intégrer une dimension stochastique :

- le modèle de Mack, avec hypothèse sur les 2 premiers moments,
- le modèle de Merz-Wüthrich, pour une vision à un an plutôt qu'à l'ultime,
- les approches GLM-bootstrap, pour obtenir une distribution complète de la provision.

Tous ces modèles ont déjà été vus en cours... On aborde dans la suite une vision bayésienne du provisionnement.

1. Tarification a priori - concepts avancés

2. Construction d'un zonier

3. Provisionnement
   - Problématique du provisionnement
   - Données agrégées et triangle de liquidation
   - Provisionnement stochastique MCMC
     - Algorithme de Metropolis-Hastings
     - Modèle CRC
   - Mise en lumière des limites de ces modèles
   - Provisionnement ligne-à-ligne (microlevel reserving)
   - Illustrations de l'intérêt de la méthode sur des cas pratiques

RETOUR SUR LE CONTEXTE BAYESIEN

On stipule un modèle qui régit les observations, $X \sim f(\theta)$.

Connaissant un échantillon observé x,

- les statisticiens fréquentistes testent $\theta = \theta_0$;
- alors que les bayésiens calculent la distribution a posteriori $(\Theta \,|\, X)$ du paramètre, notée $f(\theta \,|\, x)$, étant donné une distribution a priori $\Theta \sim \pi$. Ainsi, ils cherchent

$$f(\theta \,|\, x) = \frac{f(x \,|\, \theta)\, \pi(\theta)}{\int_\nu f(x \,|\, \nu)\, \pi(\nu)\, d\nu}.$$

Question : comment choisir l'a priori Θ ? $\Rightarrow$ non-informative prior...

3 Provisionnement
- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
  - Algorithme de Metropolis-Hastings
  - Modèle CRC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
  - Idée du provisionnement par arbre de décision
  - Formalisation : construction de l'arbre
  - Extension de CART aux données censurées
- Illustrations de l'intérêt de la méthode sur des cas pratiques
  - Application 1 : comparaison aux prévisions d'experts
  - Application 2 : assurance de revenus

CALCUL NUMERIQUE : LE PB DE LA DIMENSION

Imaginons que le paramètre θ est multidimensionnel, de dimension n avec n grand (ex : n est le nb de facteurs de développement).

Soit X les observations, par ex. les données du triangle. Ainsi,

$$f(\theta \mid x) = \frac{f(x \mid \theta)\, \pi(\theta)}{\int_{\nu_1} \ldots \int_{\nu_n} f(x \mid \nu)\, \pi(\nu)\, d\nu}$$

avec

- $f(x \mid \theta)$ est la vraisemblance de X sachant $\Theta = \theta$,
- $\pi(\theta)$ est l'a priori sur θ,
- $f(\theta \mid x)$ est l'a posteriori sur θ.

Le calcul de l'intégrale multidimensionnelle est très complexe...

SIMULATION ET CHAINE DE MARKOV

Issue de "Sampling Based Approach to Calculating Marginal Densities", Gelfand and Smith, JASA (1990).

Une chaine de Markov satisfait

$$\mathbb{P}(X_t = y \,|\, X_{t-1} = x_{t-1}, ..., X_1 = x_1) = \mathbb{P}(X_t = y \,|\, X_{t-1} = x_{t-1}).$$

L'état courant ne dépend que de l'état précédent !

La théorie ergodique stipule, sous certaines conditions, l'existence d'une mesure stationnaire, g, telle que

$$\mathbb{P}(X_t = y \,|\, X_{t-1}) \underset{t \to +\infty}{\longrightarrow} g(y)$$

Interp. : pour T grand, $\{X_{T+n}\}_{n=1}^{N}$ est un N-échantillon de loi $g(X)$.

ALGORITHME DE METROPOLIS-HASTINGS

L'algorithme de Metropolis-Hastings est une chaine de Markov fondamentale.

Il se décompose en les étapes suivantes :

1. soit $f(\theta^\star \mid x)$ la densité de $\Theta \mid X$;
2. au temps $t = 1$: fixer une position initiale θ_1 dans l'espace des paramètres ;
3. fixer une distribution $p(\theta \mid \theta_{t-1})$ permettant de proposer une nouvelle valeur du paramère connaissant la précédente valeur ;
4. à partir de $t = 2$, répéter jusqu'à convergence de la chaine :
   - à l'étape t, simuler une proposition $\theta^\star \sim p(\theta \mid \theta_{t-1})$;
   - simuler $U \sim \mathcal{U}(0, 1)$;

- calculer
$$R = \frac{f(\theta^\star \,|\, x)}{f(\theta_{t-1} \,|\, x)} \, \frac{p(\theta_{t-1} \,|\, \theta^\star)}{p(\theta^\star \,|\, \theta_{t-1})}$$
- Si $U < R$, alors $\theta_t = \theta^\star$. Sinon $\theta_t = \theta_{t-1}$.

Astuce : en considérant ce ratio, l'intégrale multidimensionnelle disparait ! En effet,

$$R = \frac{\dfrac{f(x \,|\, \theta^\star) \, \pi(\theta^\star)}{\int_{\nu_1} \dots \int_{\nu_n} f(x \,|\, \nu) \, \pi(\nu) \, d\nu}}{\dfrac{f(x \,|\, \theta_{t-1}) \, \pi(\theta_{t-1})}{\int_{\nu_1} \dots \int_{\nu_n} f(x \,|\, \nu) \, \pi(\nu) \, d\nu}} \, \frac{p(\theta_{t-1} \,|\, \theta^\star)}{p(\theta^\star \,|\, \theta_{t-1})}$$

et les termes se simplifient...

Rq : le Gibbs sampler est un cas particulier en dim. 1 de l'algo.

UTILITE ET NAISSANCE DE LA METHODE MONTE CARLO MARKOV CHAIN (MCMC)

Il suffit donc pour implémenter cet algorithme de disposer de

- la distribution conditionnelle $f(x \mid \theta)$,
- la distribution a priori $\pi(\theta)$.

On obtient une distribution limite (après CV de la chaine de Markov) qui est la distribution a posteriori.
En pratique, elle est donnée par un échantillon de réalisations !

Théoriquement,

- la distribution limite est la même, $\forall p(\theta \mid \theta_{t-1})$!
- pas de limite sur le nb de paramètres,
- les conditions d'application de l'algorithme sont satisfaites dans le cadre du provisionnement.

3 Provisionnement
- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
  - Algorithme de Metropolis-Hastings
  - Modèle CRC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
  - Idée du provisionnement par arbre de décision
  - Formalisation : construction de l'arbre
  - Extension de CART aux données censurées
- Illustrations de l'intérêt de la méthode sur des cas pratiques
  - Application 1 : comparaison aux prévisions d'experts
  - Application 2 : assurance de revenus

PROVISIONNEMENT : EXEMPLE DU MODELE CRC

Idée : on se base sur la réalité des données dont on dispose.
En l'occurence,

- volume de primes récolté par survenance,
- on a une idée du Loss-Ratio attendu,
- on dispose du triangle de liquidation historique.

On propose d'ajuster les sinistres par un modèle à facteurs :

$$C_{wd} \sim \textit{lognormal}(\mu_{wd}, \sigma_d),$$

avec un facteur dépendant de l'année de survenance, et un facteur dépendant du délai de réglement.

QUANTITES DU MODELE CRC

Dans le modèle CRC (CRoss-Classified), on spécifie les paramètres ainsi :

- C_{wd} : montant cumulé de sinistres pr l'année "w" et délai "d",
- μ_{wd} : moyenne de la distribution lognormale de l'année "w" avec délai "d",
- σ_d : l'écart-type de la distribution lognormale pour le délai "d", avec la contrainte :

$$\sigma_1^2 > \sigma_2^2 > ... > \sigma_d^2$$

Rq : contrainte logique car $+$ délai $\nearrow$, $+$ proportion connue de la sinistralité $\nearrow$, et donc plus la variance $\searrow$.

PARAMETRAGE POUR FAIRE DU BAYESIEN

En réalité, on spécifie la moyenne de la loi lognormale comme

$$\mu_{wd} = \log(\mathit{Premium}_w) + \log(\mathit{ExpectedLossRatio}) + \alpha_w + \beta_d.$$

On a besoin de spécifier une loi a priori sur (triangle taille 10×10)

- $\log(\mathit{ExpectedLossRatio}) \sim \mathcal{N}(-0.4, \sqrt{10})$;
- $\alpha_w \sim \mathcal{N}(0, \sqrt{10})$ pour $w = 2, ..., 10$, avec $\alpha_1 = 0$;
- $\beta_d \sim \mathcal{N}(0, \sqrt{10})$ pour $d = 1, ..., 9$, avec $\beta_{10} = 0$;
- la variance de la lognormale (en satisfaisant la contrainte) :

$$\sigma_d^2 = \sum_{i=d}^{10} a_i, \quad \text{avec } a_i \sim \mathcal{U}(0, 1).$$

RESULTATS PAR MCMC

En sortie du calibrage, on obtient N (ex : $N = 10000$) réalisations de la loi a posteriori des paramètres.

Puis on calcule pour chaque état après convergence de la chaine de Markov (= chaque simulation) :

1. $\{\mu_{w,10}\}_{w=1}^{10} = \{\log(\mathit{ExpectedLossRatio})\} + \{\alpha_w\} + \{\beta_{10}\}$
2. ce qui permet de resimuler les ultimes :

$$\{C_{w,10}\}_{w=2}^{10} \sim \{\mathit{lognormal}(\mu_{w,10}, \sigma_{10})\}_{w=2}^{10},$$

3. reconstruire le total des charges ultimes $\{\sum_{w=1}^{10} C_{w,10}\}$;

On peut ensuite calculer les statistiques d'intérêt sur la charge ultime globale, par exemple : moyenne($\{\sum_{w=1}^{10} C_{w,10}\}$), ...

COMMENTAIRES

L'avantage est de pouvoir enrichir le paramétrage des modèles à partir de données historiques, avec mise à jour...

De nombreuses extensions du modèle CRC ont été proposées, en particulier

1. pour gérer des corrélations entre survenance et développement,
2. pour gérer des corrélations entre plusieurs triangles,
3. ...

Cf TP !

1 Tarification a priori - concepts avancés

2 Construction d'un zonier

3 Provisionnement

- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
- Illustrations de l'intérêt de la méthode sur des cas pratiques

AVANTAGES ET INCONVENIENTS

- Chain Ladder :
  - \+ données simples/compactes, facile à implémenter,
  - – ne se sert pas des informations précises sur les sinistres,
  - – nécessite des hypothèses (très) fortes...

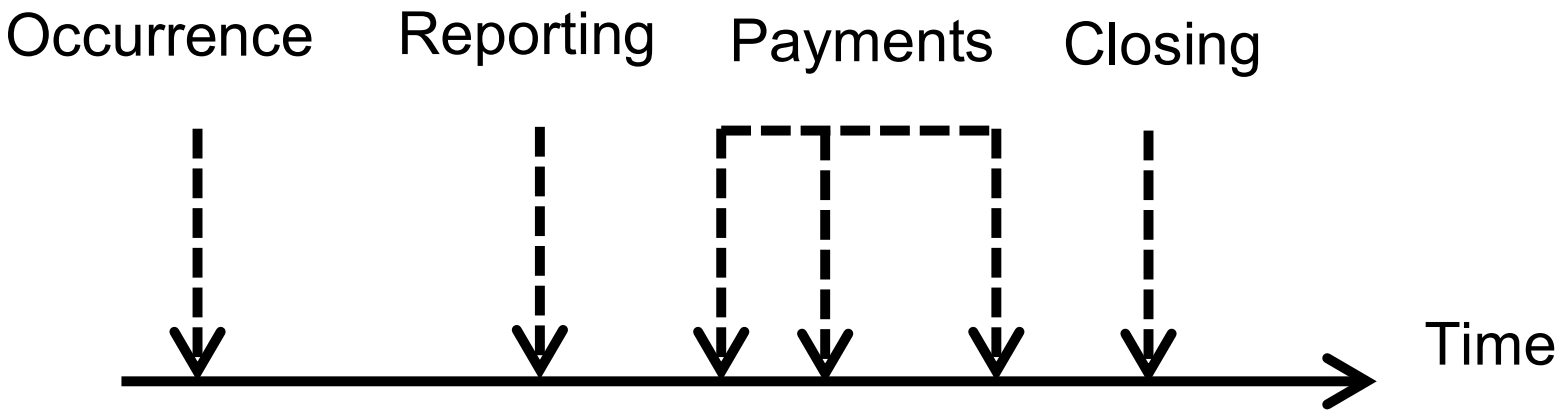
- Micro-level reserving :
  - \+ utilise les données individuelles sur les sinistres : prise en compte de l'hétérogénéité, de la vie du sinistre ;
  - \+ adapté potentiellement à des branches longues,
  - – plus difficile à implémenter,
  - – nécessite de gérer à part les IBNyR.

GESTION DE L'HÉTÉROGÉNÉITÉ

Etant donné que l'on "mélange" toutes les données en vision agrégée, la qualité de l'estimation de la provision repose sur la qualité et la stabilité des données... Il faut identifier :

1. Les facteurs internes qui pourraient impacter la provision :
   - évolution du portefeuille,
   - politique de souscription, tarification et réassurance,
   - politique de gestion des sinistres (cadence de réglement).
2. Et les facteurs externes :
   - pratiques de marché, cycles économiques, inflation,
   - évolution de la sinistralité (fréquence, sévérité),
   - modifications réglementaires et comptables.

NON PRISE EN COMPTE DE LA VIE DU SINISTRE



Elle a un impact majeur sur la provision à constituer... Notamment,

- la durée de vie du sinistre, s'il a été ré-ouvert ou non,
- la typologie du risque sous-jacent (ex : assurance construction décennale),
- le nombre de paiements...

1 Tarification a priori - concepts avancés

2 Construction d'un zonier

3 Provisionnement

- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
  - Idée du provisionnement par arbre de décision
  - Formalisation : construction de l'arbre
  - Extension de CART aux données censurées
- Illustrations de l'intérêt de la méthode sur des cas pratiques

ATOUTS DES MODELES DE REGRESSION

- Etre capable de gérer l'hétérogénéité des données :
  - du fait du temps de développement du sinistre,
  - de ses caractéristiques (type de risque, ex : construction), ...
- Utiliser des techniques d'apprentissage statistique pour privilégier un estimateur non-paramétrique :
  - flexibilité de la forme de dépendance entre T et $\mathbf{X}$;
  - ici on prend les arbres CART sur lesquels on retravaillera.
- Disposer de résultats de convergence des estimateurs :
  - [Lopez et al., 2016] : *Tree-based censored regression with applications in insurance*, `EJS`.

3 Provisionnement
- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
  - Algorithme de Metropolis-Hastings
  - Modèle CRC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
  - Idée du provisionnement par arbre de décision
  - Formalisation : construction de l'arbre
  - Extension de CART aux données censurées
- Illustrations de l'intérêt de la méthode sur des cas pratiques
  - Application 1 : comparaison aux prévisions d'experts
  - Application 2 : assurance de revenus

CENSURE ET PROVISIONNEMENT

Objectif : estimer les montants (ou durées) de sinistres individuels T sachant les caractéristiques $\mathbf{X}$, en utilisant un arbre CART.

On observe parfois seulement le montant payé jusqu'à aujourd'hui, Y : censure droite !

Si le sinistre est censuré :

- le sinistre est encore ouvert et a commencé à être payé (il n'est pas clos $\Rightarrow$ IBNeR).
- le montant total final T reste inconnu : on a payé $Y \leq T$.

Rq : le sinistre est aussi parfois tronqué à gauche.

CLUSTERING PAR ARBRE SUR DONNEES COMPLETES

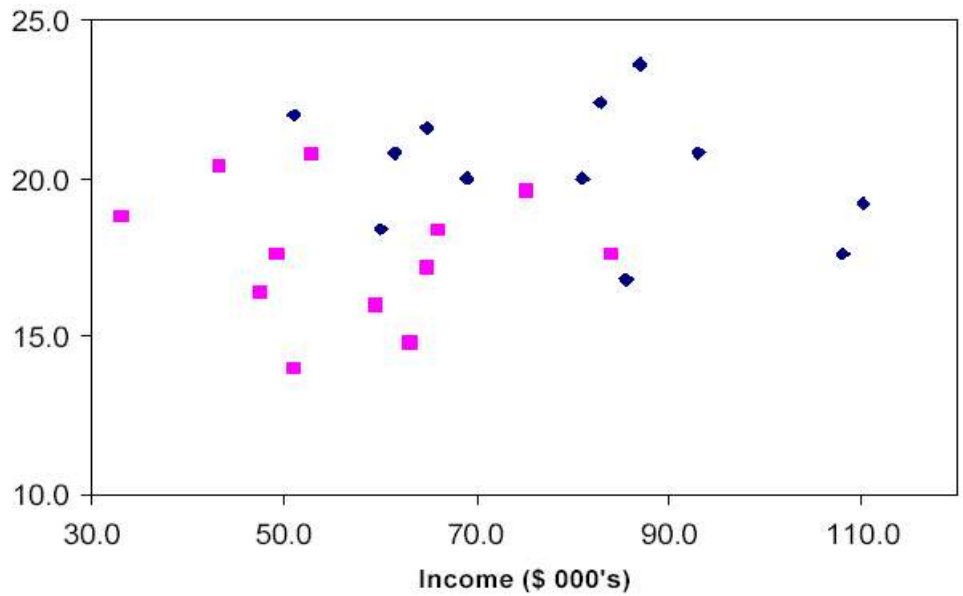
Pour estimer notre quantité d'intérêt, on considère un modèle de segmentation fourni par un arbre de décision où :

1. la racine : population entière (montants) à segmenter $\Rightarrow$ point initial ;
2. les branches : règles de segmentation ;
3. les feuilles : sous-populations homogènes $\Rightarrow$ donne l'estimation de la réponse.

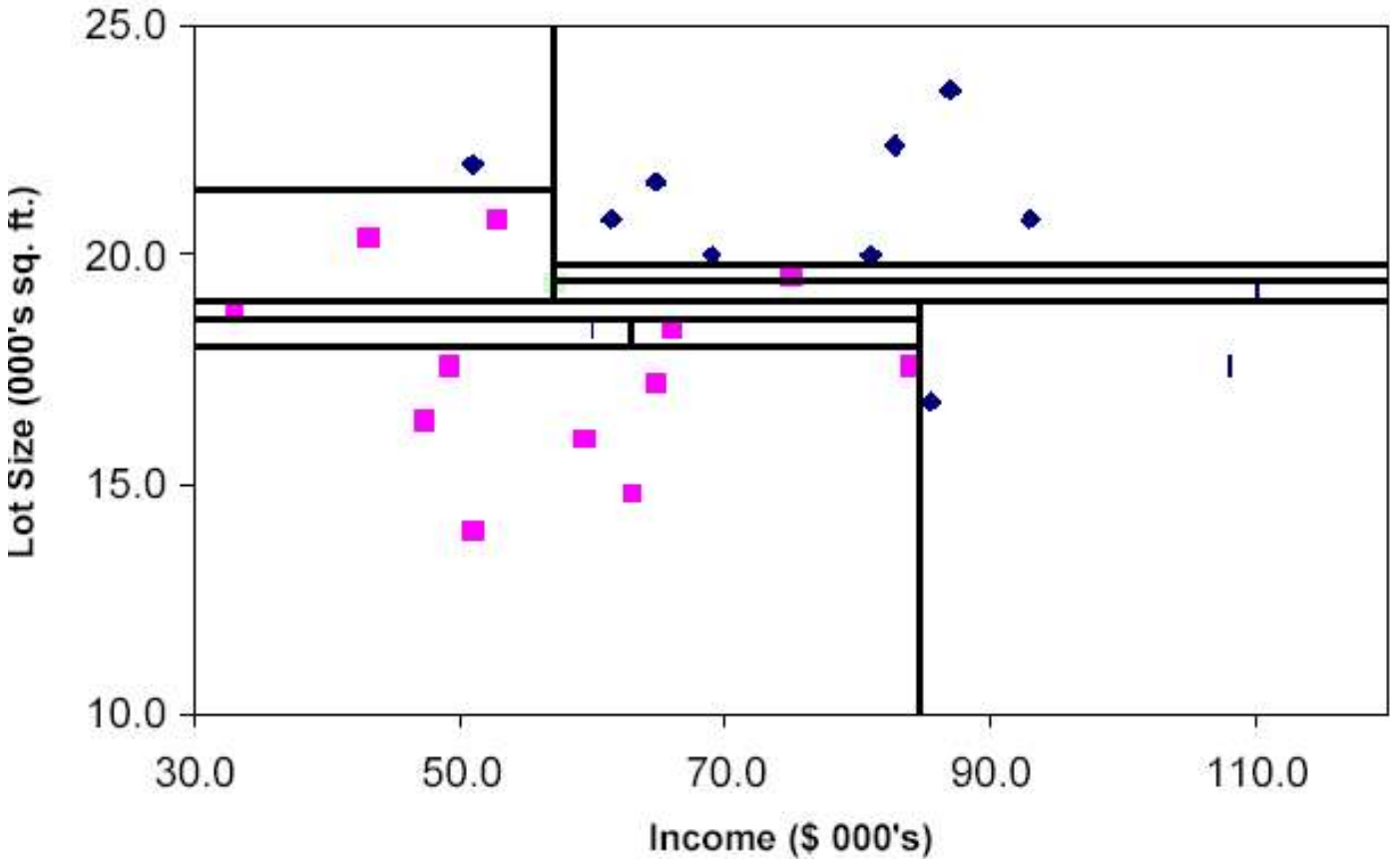
Une référence en actuariat $\rightarrow$ [Olbricht, 2012] (tables de mortalité).

EXEMPLE CART : prévoir propriétaire | revenu et taille

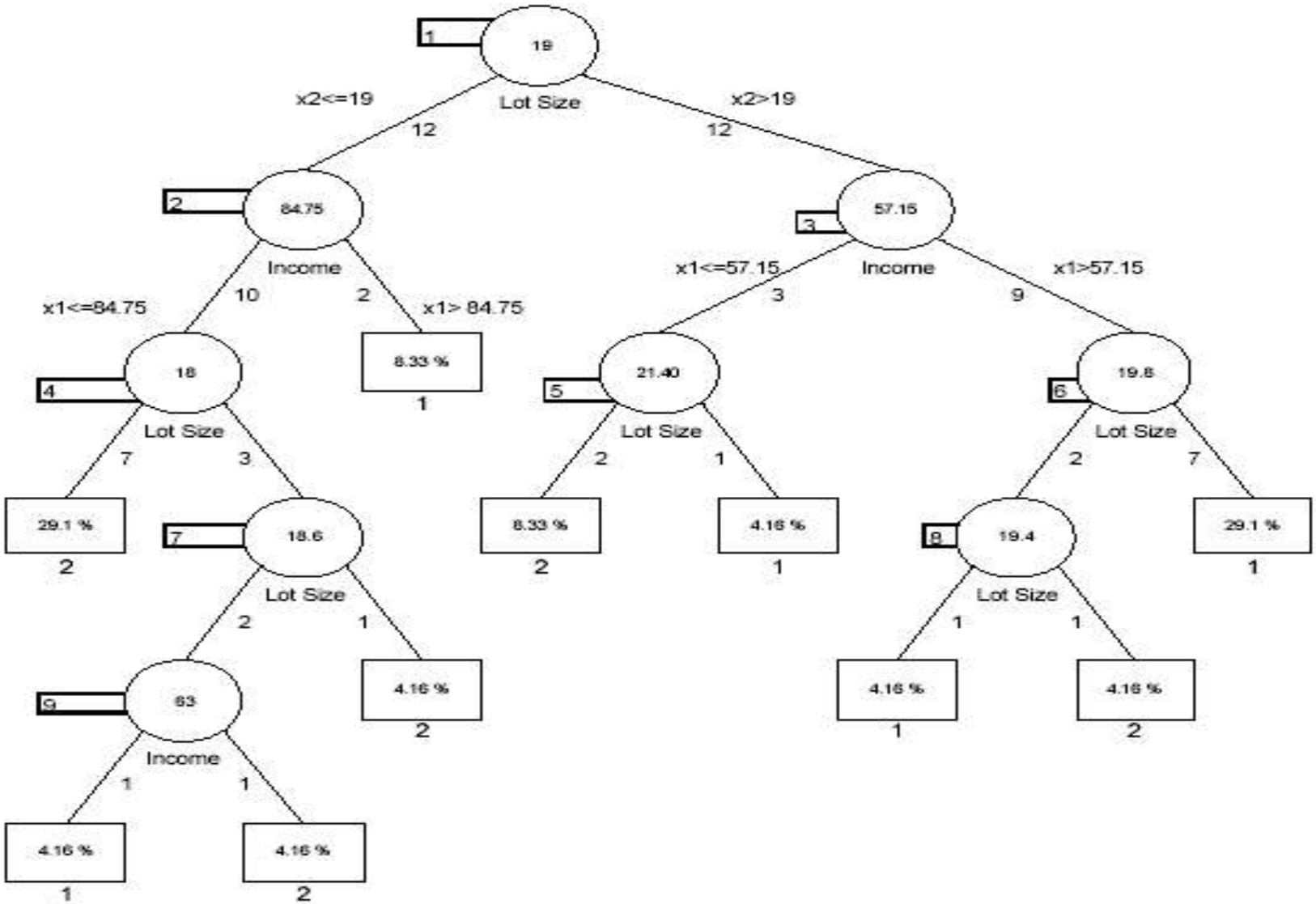
Income ($ 000's)	Lot Size (000's sq. ft.)	Owners=1, Non-owners=2
60	18.4	1
85.5	16.8	1
64.8	21.6	1
61.5	20.8	1
87	23.6	1
110.1	19.2	1
108	17.6	1
82.8	22.4	1
69	20	1
93	20.8	1
51	22	1
81	20	1
75	19.6	2
52.8	20.8	2
64.8	17.2	2
43.2	20.4	2
84	17.6	2
49.2	17.6	2
59.4	16	2
66	18.4	2
47.4	16.4	2
33	18.8	2
51	14	2
63	14.8	2



PARTITION ET ARBRE



But : créer des partitions d'homogénéité maximale.


1
19
Lot Size
x2<=19
12
x2>19
12
2
84.75
Income
10
2
x1<=84.75
x1> 84.75
3
57.15
Income
x1<=57.15
3
x1>57.15
9
4
18
Lot Size
7
3
8.33 %
1
5
21.40
Lot Size
2
1
6
19.8
Lot Size
2
7
29.1 %
2
7
18.6
Lot Size
2
1
8.33 %
2
4.16 %
1
8
19.4
Lot Size
1
1
29.1 %
1
9
63
Income
1
1
4.16 %
2
4.16 %
1
4.16 %
1
4.16 %
2
4.16 %
2

3 Provisionnement
- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
  - Algorithme de Metropolis-Hastings
  - Modèle CRC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
  - Idée du provisionnement par arbre de décision
  - Formalisation : construction de l'arbre
  - Extension de CART aux données censurées
- Illustrations de l'intérêt de la méthode sur des cas pratiques
  - Application 1 : comparaison aux prévisions d'experts
  - Application 2 : assurance de revenus

ARBRE DE RÉGRESSION : Y TOTALEMENT OBSERVEE

$$\pi_0(\mathbf{x}) = E_0[T \,|\, \mathbf{X} = \mathbf{x}] \tag{1}$$

$\rightarrow$ Lien le plus utilisée : relation linéaire entre T et $X \Rightarrow$ EQM.
$\rightarrow$ En pratique, on ne peut pas considérer ts les estimateurs possibles de $\pi_0(\mathbf{x}) \Rightarrow$ CART est une autre classe d'estimateurs :

$$\hat{\pi}(\mathbf{x}) := \hat{\pi}^L(\mathbf{x}) = \sum_{l=1}^{L} \hat{\gamma}_l \, R_l(\mathbf{x}) \tag{2}$$

- L : nombre de feuilles de l'arbre, l leur indice,
- $R_l(\mathbf{x}) = \mathbb{1}(\mathbf{x} \in \mathcal{X}_l)$: appartenance à partition $\mathcal{X}_l$,
- $\hat{\gamma}_l = E_n[Y \,|\, \mathbf{x} \in \mathcal{X}_l]$: moy. empirique de T dans la feuille l.

CONSTRUCTION DE L'ARBRE : CRITERE DE DIVISION

→ Doit être adapté à notre objectif.
→ Pour résoudre (1), *MCO* utilisés car solution donnée par

$$\pi_0(\mathbf{x}) = \arg\min_{\pi(\mathbf{x})} E_0\Big[\phi(T, \pi(\mathbf{x})) \,|\, \mathbf{X} = \mathbf{x}\Big] \tag{3}$$

où $\phi(T, \pi(\mathbf{x})) = (T - \pi(\mathbf{x}))^2$ (ϕ fonction de perte)

→ Conduit à minimiser la variance intra-noeud à chaque étape / maximiser la variance inter.

→ Si T est totalement observé, construire l'arbre avec ce critère donne un estimateur convergent ([Breiman et al., 1984]).

ELAGUER : PENALISER PAR LA COMPLEXITE

Principe de l'algorithme CART : ne pas arrêter la segmentation, construire l'arbre *"maximal"* (taille $K(n)$), puis l'élaguer.

$\rightarrow$ On obtient une suite d'estimateurs $(\hat{\pi}^K(\mathbf{x}))_{K=1,\ldots,K(n)}$.

Eviter surapprentissage $\Rightarrow$ sélectionner le meilleur sous-arbre de l'arbre max., arbitrage entre adéquation et capacité prédictive :

$$R_\alpha(\hat{\pi}^K(\mathbf{x})) = E_n[\,\Phi(Y, \hat{\pi}^K(\mathbf{x}))\,] + \alpha\,(K/n).$$

α coût de complex., K nb de feuilles ([Gey and Nedelec, 2005]).

3 Provisionnement
- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
  - Algorithme de Metropolis-Hastings
  - Modèle CRC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
  - Idée du provisionnement par arbre de décision
  - Formalisation : construction de l'arbre
  - Extension de CART aux données censurées
- Illustrations de l'intérêt de la méthode sur des cas pratiques
  - Application 1 : comparaison aux prévisions d'experts
  - Application 2 : assurance de revenus

RAPPEL : DONNEES ET OBJECTIF

On observe un échantillon iid de v.a. $(Y_i, \delta_i, X_i)_{1 \leq i \leq n}$ de distribution (Y, δ, X), où

$$\begin{cases} Y &= \inf(T, C) \\ \delta &= \mathbf{1}_{T \leq C} \end{cases}$$

Montant courant Y, sinistre ouvert : $\delta = 0$.
C : **variable de censure**.

- On cherche $T^* = E[T \mid \delta = 0, Y, \mathbf{X}]$.
- But : trouver un estimateur de T^*, sachant que l'on n'a pas d'observations iid de $T \Rightarrow$ pas de LGN, ...

COMMENT GÉRER LES SINISTRES OUVERTS ?

- *Mauvaise solution* : ne considérer que les sinistres clos pour construire l'arbre de décision afin d'estimer la réponse.
  $\rightarrow$ On sous-estimera montants finaux, dc la provision.

  Cependant, les sinistres ouverts donnent également une information biaisée $\Rightarrow$ à corriger !

- Une solution possible : surpondérer les sinistres clos avec dével. long pour compenser leur sous-représentation...

  $\Rightarrow$ **Question : quels poids ?**

INGREDIENTS : ESTIMATEUR KAPLAN-MEIER ET IPCW

L'algorithme CART peut être adapté ([Lopez et al., 2016]) avec les outils suivants. Hypothèse : T est indépendant de C.

- Soit $\hat{F}(t) = 1 - \prod_{Y_i \leq t} \left(1 - \frac{\delta_i}{\sum_{j=1}^n \mathbf{1}_{Y_j \geq Y_i}}\right)$.
  $\rightarrow$ Cet estimateur tend vers $F(t) = \mathbb{P}(T \leq t)$.

- Version additive : $\hat{F}(t) = \sum_{i=1}^n W_{i,n} \mathbf{1}_{Y_i \leq t}$, avec les poids Kaplan-Meier

$$W_{i,n} = \frac{\delta_i}{n[1 - \hat{G}(Y_i-)]},$$

  où $\hat{G}(t)$ est l'estimateur Kaplan-Meier de $G(t) = \mathbb{P}(C \leq t)$.

Voir aussi cours du premier semestre de Data Sciences.

1 Tarification a priori - concepts avancés

2 Construction d'un zonier

3 Provisionnement

- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
- Illustrations de l'intérêt de la méthode sur des cas pratiques
  - Application 1 : comparaison aux prévisions d'experts
  - Application 2 : assurance de revenus

3 Provisionnement
- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
  - Algorithme de Metropolis-Hastings
  - Modèle CRC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
  - Idée du provisionnement par arbre de décision
  - Formalisation : construction de l'arbre
  - Extension de CART aux données censurées
- Illustrations de l'intérêt de la méthode sur des cas pratiques
  - Application 1 : comparaison aux prévisions d'experts
  - Application 2 : assurance de revenus

MONTANT ULTIME SUR SINISTRES CENSURÉS

On cherche $E[M \,|\, \delta = 0, X, Y, N]$, avec M le montant du sinistre.

But : revenir à des quantités conditionnées uniquement par **X** !

$$\begin{aligned} E[M \,|\, \delta = 0, X = x, Y = y, N = n] &= E[M \,|\, M \geq n, T \geq y, X = x] \\ &= \frac{E\left[M \mathbf{1}_{M \geq n, T \geq y} \,|\, X = x\right]}{\mathbb{P}(T \geq y, M \geq n \,|\, X = x)}. \end{aligned}$$

Soient $\Phi_1(t, m) = m\mathbf{1}_{m \geq n, t \geq y}$ et $\Phi_2(t, m) = \mathbf{1}_{t \geq y, m \geq n}$.

On veut dc estimer le ratio des 2 quantités suivantes $\Rightarrow$ 2 arbres !

(1) $E[\Phi_1(T, M) \,|\, X = x]$ sur (2) $E[\Phi_2(T, M) \,|\, X = x]$.

EXTRAIT DES DONNEES

Assurance RC médicale aux US : 648 sinistres et leurs caractéristiques (specialité, lieu, statut de réouverture, ...).

	Claim.entry	Indemn.res	ALAE.res	(..)	Cens.	Already.paid	Reserved
47	2000-07-14	0	0.00		1	3456	0
48	2000-07-24	5000	13880.25		0	138435	18880
49	2000-07-31	5000	11304.60		0	7300	16305
50	2000-07-31	5000	103471.31		0	118136	108471
51	2000-08-04	0	0.00		1	46587	0
52	2000-08-14	0	0.00		1	3083	0
53	2000-08-15	0	0.00		1	0	0
54	2000-08-28	0	0.00		1	980	0

```
> summary(myData$Observed.total)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       0    2644   41760   18500 1557000
```

STATISTIQUES DESCRIPTIVES BASIQUES

Statistics on the information selected for our application.

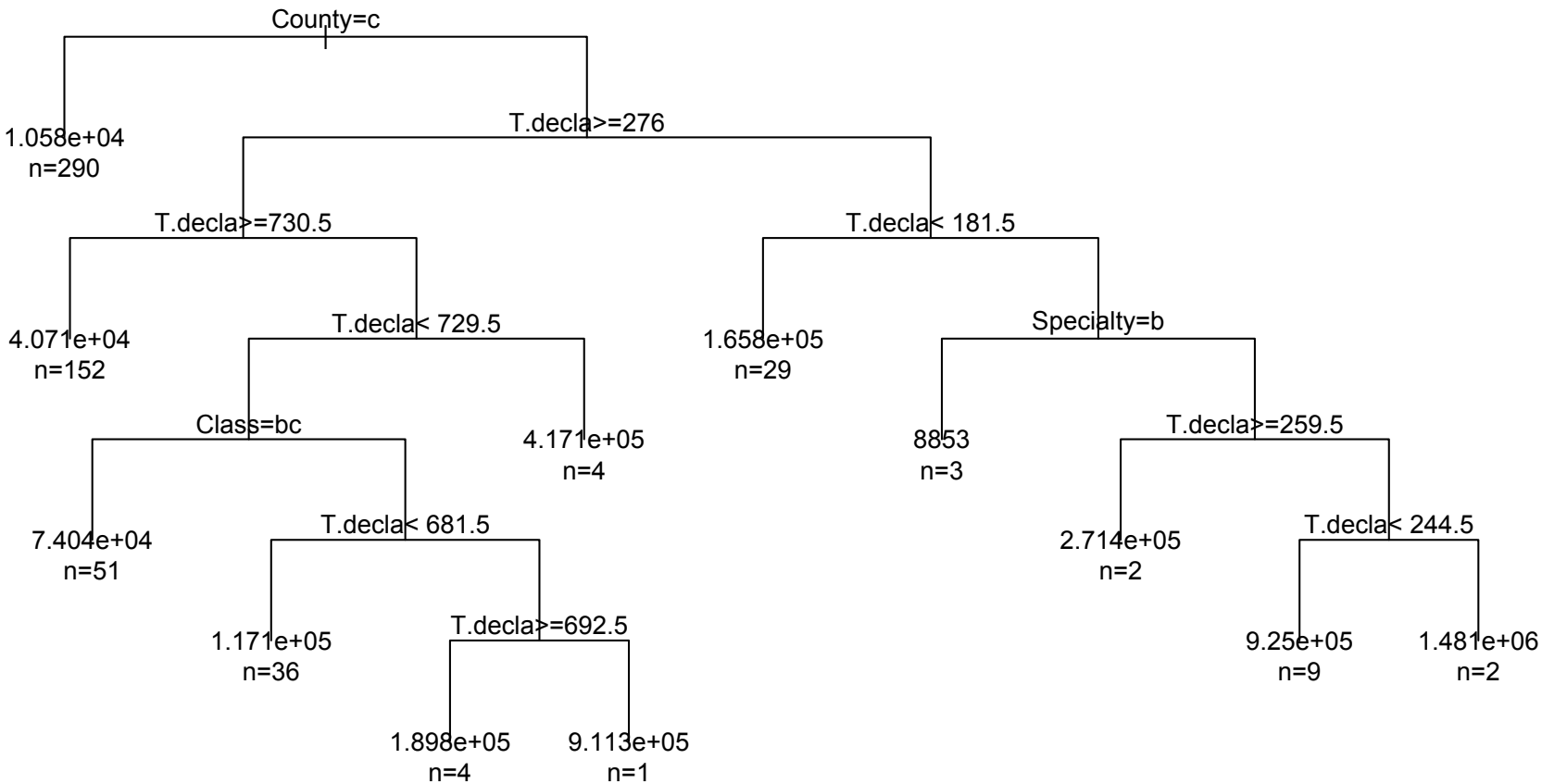
		Statistical indicators					
	Type	Median	Mean	Std.	Min.	Max.	# categories
Insurance type	categorical						2
Specialty	categorical						41
Class	categorical						19
Report date	date				N	N+7	
Area	categorical						30
Closed without payments	boolean						2
Closed without indemnity	boolean						2
Time before opening (days)	continuous	1164	1223	614	2	4728	
Time before declaration	continuous	734	724	560	0	4657	
Reopen status	boolean						2
Cancel status	boolean						2
Reserves	continuous	0	44170	138867	0	1062000	
Development time	continuous	419	606	506	0	2249	
Observed payments	continuous	2617	41810	152319	0	1557000	

→ Données très hétérogènes : beaucoup de montants provisionnés à 0 à cause d'attente de décision judiciaire...

→ Taux de censure important : environ 33% ;

⇒ Un modèle paramétrique serait difficile à estimer !

PRÉVISIONS DE LA QUANTITÉ $E[M1_{(M>n,T>y)} \mid X = x]$
ARBRE ELAGUÉ



PRÉVISIONS DE LA QUANTITÉ $P(M > n, T > y|X = x)$ RESULTATS NUMERIQUES

Error of the tree:

```
> (1.0 - (confusion.matrix[1,1]+confusion.matrix[2,2]) / sum(confusion.matrix))*
> cat("The test sample estimate of the prediction error in the pruned tree is",
The test sample estimate of the prediction error in the pruned tree is 18.6%
```

Predicted probabilities for the denominator:

(..)	Censure	Already.paid	Reserved	Observed.total	KM.weight	Proba.censorship
	1	24	0	24	0.0017	0.1496063
	1	1844	0	1844	0.0017	0.1496063
	1	444	0	444	0.0017	0.1935484
	1	0	0	0	0.0017	0.1496063
	1	3907	0	3907	0.00176	0.2307692
	0	0	81000	0	0	0.7500000
	0	1061	42139	1061	0	0.7400000
	0	1061	79939	1061	0	0.2307692
	0	1061	12439	1061	0	0.7400000

RATIO (1)/(2) ET COHÉRENCE AVEC OPINIONS D'EXPERT

```
>  ##################################################################
>  ## Final prediction of total claim amount for censored claims.
>  ##################################################################
>  ## Comparison b/w predictions from the tree and the one from the expert.

 Censure Already.paid Reserved Adj.predicted.claims Expert.prediction
       0            0    81000             70752.37             81000
       0            0    71600             10585.00             71600
       0            0        0             10585.00                 0
       0            0    13500             10585.00             13500
       0            0    52700             55008.11             52700
       0            0     2500             10585.00              2500
       0            0    55500             70752.37             55500
       0            0    62100             55008.11             62100
       0            0    81000             54274.67             81000
       0         1061    42139             55008.11             43200
       0         4266    57834             70752.37             62100
```

Intérêt : faire des économies en évitant de consulter les experts...

3 Provisionnement
- Problématique du provisionnement
- Données agrégées et triangle de liquidation
- Provisionnement stochastique MCMC
  - Algorithme de Metropolis-Hastings
  - Modèle CRC
- Mise en lumière des limites de ces modèles
- Provisionnement ligne-à-ligne (microlevel reserving)
  - Idée du provisionnement par arbre de décision
  - Formalisation : construction de l'arbre
  - Extension de CART aux données censurées
- Illustrations de l'intérêt de la méthode sur des cas pratiques
  - Application 1 : comparaison aux prévisions d'experts
  - Application 2 : assurance de revenus

ASSURANCE DU RISQUE INCAPACITÉ - INVALIDITÉ

Nous disposons d'un historique de 6 ans d'un portefeuille couvrant le risque incapacité avec les informations suivantes :

- 83 547 sinistres ;
- cause de l'arrêt (maladie ou accident), sexe, CSP, age, durée d'incapacité (censurée ou non), réseau de distribution ;
- le taux de censure vaut 7.2% ;
- durée moyenne en incapacité : 100 jours.

But : trouver une segmentation pour prédire la durée en incapacité, le remboursement étant forfaitaire.

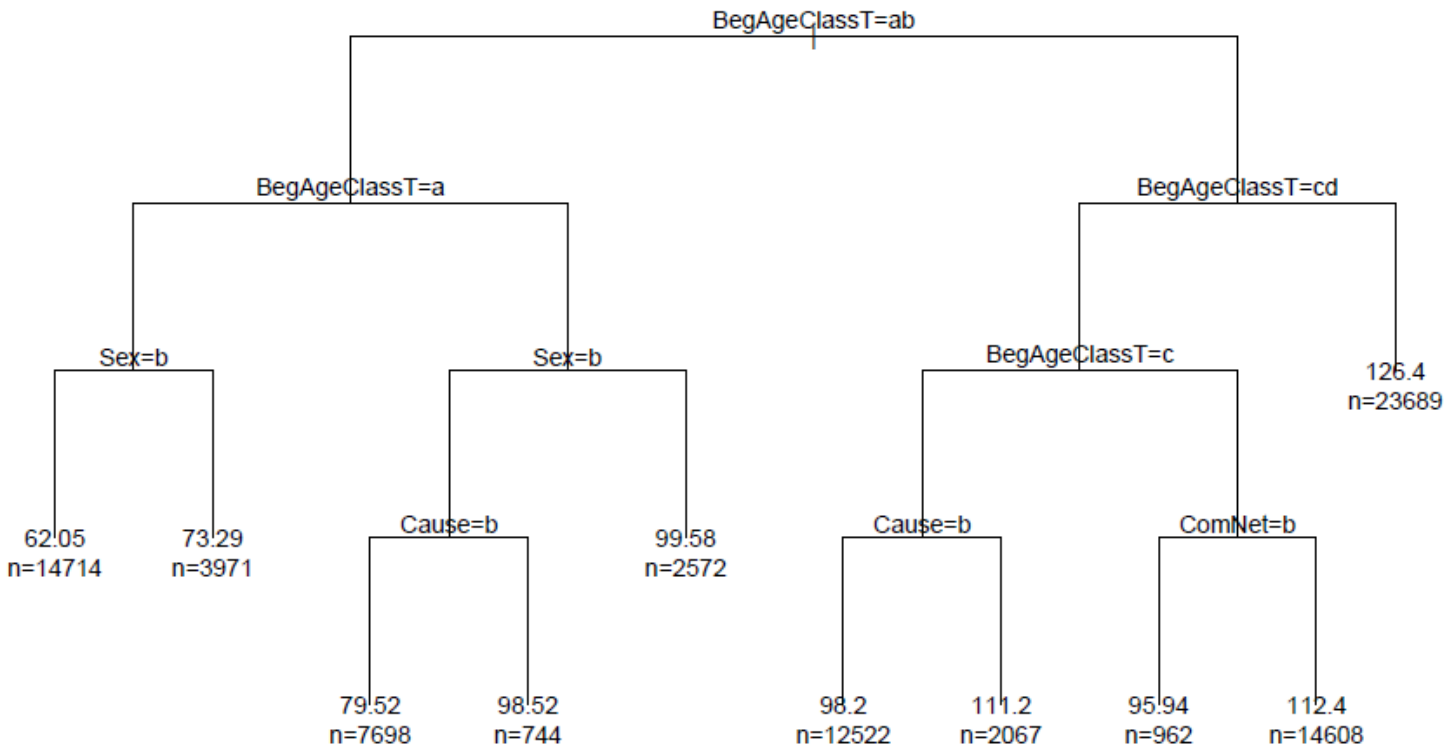
VISUALISATION DES DONNEES

```
> dim(myData)
[1] 83547    20

> summary(myData)
 Sex        TypeEmployee ContractNumber            TYPE_ARRET        SurvDate
 F:65557    CAD: 3074    0725235: 1524   Maladie         :71563   Min.   :2006-01-0
 M:17990    ENP: 5879    0J98706:  879   Acc. Travail    :10644   1st Qu.:2007-05-1
            ETA:  713    0232097:  684   Maladie Hospi.: 1035   Median :2008-08-2
            NCA:73290    0237127:  591   Maternite       :  179   Mean   :2008-07-2
            TNS:  591    0184638:  553   Longue Maladie:    54   3rd Qu.:2009-10-2
                         0448817:  530   Maladie Serv. :    23   Max.   :2010-11-3
                         (Other):78786   (Other)         :    49

    EndObsW          NonCensure              SPC         BegAgeClass BegAgeClassT
 Min.   :   1.00   Mode :logical   Employee:79882   1:21563      1:18685
 1st Qu.:  15.00   FALSE:5991      Manager : 3074   2:19039      2:11014
 Median :  42.00   TRUE :77556     Misc    :  591   3:20496      3:14589
 Mean   :  99.98   NA's :0                          4:22449      4:15570
 3rd Qu.: 106.00                                               5:23689
 Max.   :1578.00
```

	Sex	TypeEmployee	ContractNumber	TYPE_ARRET	SurvDate	BegIndDate	EndIndDat
23	F	NCA	0154496	Acc. Travail	2010-10-12	2010-11-11	2011-01-3
24	F	NCA	0154509	Maladie	2009-09-14	2009-10-14	2011-02-2
33	F	NCA	0154670	Maladie	2010-02-11	2010-03-13	2011-09-3
44	F	NCA	0156555	Maladie	2010-08-24	2010-09-23	2011-04-1
62	F	NCA	0161383	Maladie	2010-03-19	2010-04-18	2012-02-2
68	F	NCA	0161581	Maladie	2010-11-09	2010-12-09	2012-06-2
88	F	NCA	0331202	Maladie	2010-02-12	2010-03-14	2011-04-3
103	F	NCA	0385996	Maladie	2010-11-10	2010-12-10	2012-06-2
136	F	ENP	0725234	Maladie	2010-01-11	2010-02-10	2012-07-1
140	F	ENP	0725235	Maladie	2010-08-23	2010-09-22	2011-01-0

Cause	ComNet	BegAnc	EndAncInd	BegAge	EndObsW	NonCensure	SPC
Accident	Net_C	0	80	47.29363	50	FALSE	Employee
Sickness	Net_C	3	470	41.81246	443	FALSE	Employee
Sickness	Net_A	3	320	39.40041	293	FALSE	Employee
Sickness	Net_A	3	126	50.62286	99	FALSE	Employee
Sickness	Net_C	3	284	46.41752	257	FALSE	Employee
Sickness	Net_A	3	49	51.05544	22	FALSE	Employee
Sickness	Net_C	24	298	52.73374	292	FALSE	Employee
Sickness	Net_A	26	25	45.89733	21	FALSE	Employee
Sickness	Net_C	3	351	51.79466	324	FALSE	Employee
Sickness	Net_A	3	127	54.63107	100	FALSE	Employee

ARBRE ÉLAGUÉ : L'AGE EST CLEF !



La réglementation préconise de calculer les provisions techniques liées à cette durée en fonction de l'âge...Good news !

QUALITÉ DU MODÈLE : COURBE ROC DYNAMIQUE POUR LA CLASSIFICATION À UNE DATE FUTURE

Idée : les courbes ROC donnent une idée du pouvoir prédictif du classificateur. Elles comparent les faux et les vrais positifs de différents modèles, étant donné un seuil de proba. pour l'affectation.

Adaptation : ici le but est de comparer la prévision du modèle à la réalité (sur un échantillon test) à une certaine durée. On veut notamment voir si le modèle détecte les événements déjà survenus à cette date.

POUVOIR PRÉDICTIF : DÉTECTION DES ÉVÉNEMENTS

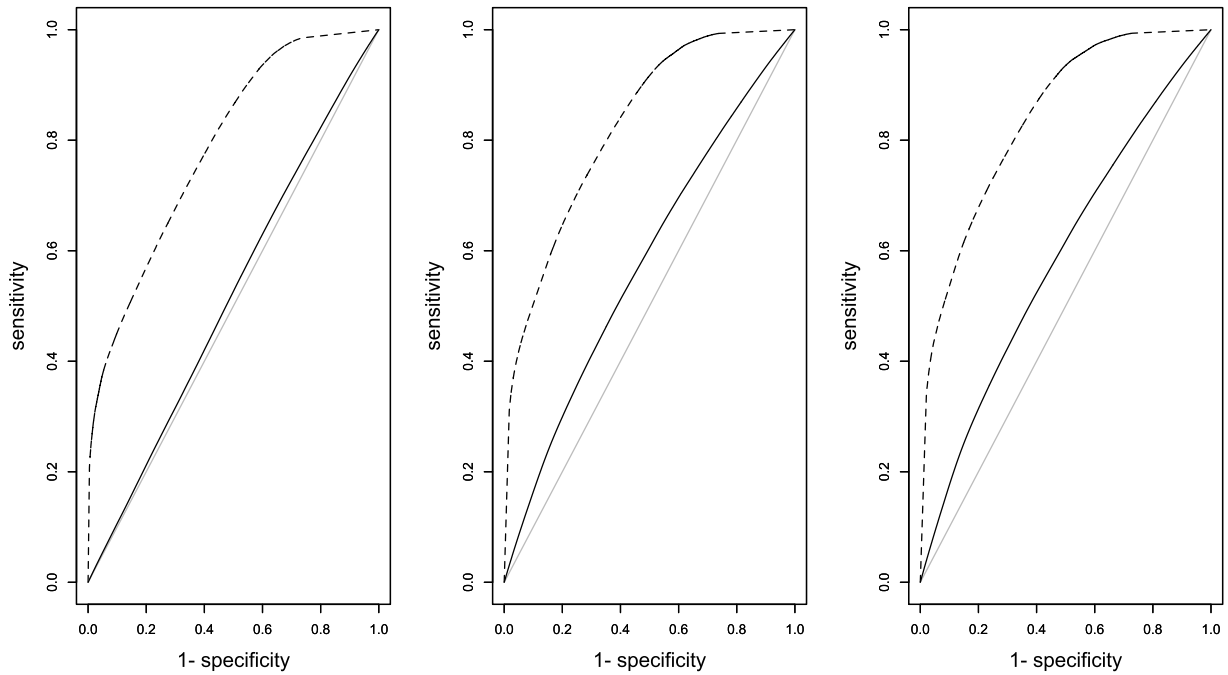


FIG 3. *Dynamic ROC curves at* $t = 15, 100, 110$ *(from left to right). The dotted line corresponds to the CART model and the black line to the Cox model.*

TABLE 6
Dynamic Area Under Curve $AUC(t)$.

	t	15	40	100	110
$AUC(t)$	CART	0.787	0.802	0.824	0.839
	Cox	0.518	0.531	0.576	0.585

GESTION DES DONNÉES
PROVISION À DATE D'ARRÊTÉ

```
> head(myData, n = 6)
  Sex TypeEmployee ContractNumber TYPE_ARRET   SurvDate BegIndDate EndIndDate
1   F          NCA        0001591    Maladie 2007-11-03 2007-12-03 2007-12-21 Si
2   F          NCA        0001591    Maladie 2008-02-04 2008-03-05 2008-08-31 Si
3   M          NCA        0006192    Maladie 2006-12-24 2007-01-23 2007-04-30 Si
4   M          NCA        0006192    Maladie 2009-11-18 2009-12-18 2010-10-01 Si
5   F          NCA        0024191    Maladie 2006-03-20 2006-04-19 2006-09-03 Si
6   F          NCA        0024251    Maladie 2008-06-21 2008-07-21 2010-07-31 Si

  X2006.10.01 X2007.01.01 X2007.04.01 X2007.07.01 X2007.10.01 X2008.01.01 X2008.
1          NA          NA          NA          NA    17.99769          NA
2          NA          NA          NA          NA          NA     91.3125    87.
3          NA     91.3125    5.688769          NA          NA          NA
4          NA          NA          NA          NA          NA          NA
5          NA          NA          NA          NA          NA          NA
6          NA          NA          NA          NA          NA          NA

> dim(learning.sample)
[1] 42523    37


> head(learning.sample)
```

```
   Sex TypeEmployee ContractNumber     TYPE_ARRET   SurvDate BegIndDate EndIndDa
1    F          NCA        0001591        Maladie 2007-11-03 2007-12-03 2007-12-
2    F          NCA        0001591        Maladie 2008-02-04 2008-03-05 2008-08-
3    M          NCA        0006192        Maladie 2006-12-24 2007-01-23 2007-04-
5    F          NCA        0024191        Maladie 2006-03-20 2006-04-19 2006-09-
9    F          NCA        0038268        Maladie 2006-05-02 2006-06-01 2006-07-
10   M          NCA        0064365 Maladie Hospi. 2006-10-30 2006-11-29 2007-02-

Cause ComNet BegAnc EndAncInd   BegAge EndObsW NonCensure      SPC
Sickness  Net_C      3        45 47.69884  0.1971       TRUE Employee
Sickness  Net_C      3       206 47.43053  1.9603       TRUE Employee
Sickness  Net_C      3       124 46.06982  1.0623       TRUE Employee
Sickness  Net_A     30       137 43.63313  1.5003       TRUE Employee
Sickness  Net_A     30        32 35.49897  0.3504       TRUE Employee
Sickness  Net_A      3       107 37.32786  0.8761       TRUE Employee
```

```
> KM.weights <- unlist(aft.kmweight(Y = matrix(data=learning.sample$EndObsW, nro
> sum(KM.weights)
[1] 1
```

```
> head(learning.sample)
```

```
    Sex TypeEmployee ContractNumber TYPE_ARRET   SurvDate BegIndDate EndIndDate
35    F          NCA        0154699    Maladie 2008-04-10 2008-05-10 2008-05-11
173   F          ENP        0729486    Maladie 2006-02-27 2006-03-29 2006-03-30
240   F          NCA        0149036    Maladie 2006-05-18 2006-06-17 2006-06-18
295   F          NCA        0637995    Maladie 2006-06-12 2006-07-12 2006-07-13
299   F          NCA        0637995    Maladie 2007-12-12 2008-01-11 2008-01-12
468   F          NCA        0179261    Maladie 2007-02-01 2007-03-03 2007-03-04

Cause ComNet BegAnc EndAncInd   BegAge EndObsW NonCensure      SPC    KM.weight
Sickness  Net_C      3        28 50.54346   0.011       TRUE Employee 2.351669e-
Sickness  Net_A      3        28 39.60849   0.011       TRUE Employee 2.351669e-
Sickness  Net_A      3        28 54.24778   0.011       TRUE Employee 2.351669e-
Sickness  Net_B      1        30 52.67077   0.011       TRUE Employee 2.351669e-
Sickness  Net_B      1        30 51.94524   0.011       TRUE Employee 2.351669e-
Sickness  Net_A     30         1 44.00000   0.011       TRUE Employee 2.351669e-


> library(rpart)
> formula <- as.formula("EndObsW ~ Sex + TypeEmployee + TYPE_ARRET + Cause + Com
> maximal.tree <- rpart(formula, data = learning.sample, weights = KM.weight, me
```

COÛTS ULTIMES (CENSURÉS OU NON)

```
> dim(validation.sample)
[1] 21261    17

> head(validation.sample)

   Sex TypeEmployee ContractNumber TYPE_ARRET   SurvDate BegIndDate EndIndDate
4    M          NCA        0006192    Maladie 2009-11-18 2009-12-18 2010-10-01 S
6    F          NCA        0024251    Maladie 2008-06-21 2008-07-21 2010-07-31 S
7    F          NCA        0037157    Maladie 2009-09-17 2009-10-17 2009-10-30 S
14   F          NCA        0099654    Maladie 2006-08-17 2006-09-16 2006-09-20 S
16   F          ENP        0119466    Maladie 2007-05-23 2007-06-22 2007-06-24 S
19   F          NCA        0154321    Maladie 2006-09-01 2006-10-01 2006-10-08 S


> predictions.validationSample <- predict(final.tree, newdata = validation.sampl

> proba.nonCensure <- length(which(validation.sample$NonCensure == TRUE)) / nrow
> proba.nonCensure
[1] 0.9244626
```

```
> ## E[T|X] = E[T|delta = 1,X] P(delta=1) + E[T|delta = 0,X] P(delta=0)
> predictionsMoy.sinistresOuverts <- (mean(predictions.validationSample) - mean(
> provisionMoyenne <- predictionsMoy.sinistresOuverts * prestation.timeStep * nr
> provisionMoyenne
[1] 181022.9

> ## To be compared with:
> backtest.provisions.validationSample
[1] 179236.8


> ## Erreur de provision moyenne en pourcentage, backtesting:
> (abs(backtest.provisions.validationSample - provisionMoyenne) / max(c(provisio
[1] 0.9866959
```

ET DANS LE CAS DE CHAIN LADDER ?

```
> triangle.cumule
```

	dev1	dev2	dev3	dev4	dev5	dev6	dev7	dev8	dev9	dev10
2006-01-01	44860	62511	72745	80289	85893	90337	93632	96355	98507	100076
2006-04-01	55982	76905	90518	101090	108863	115069	120081	123873	126825	129345
2006-07-01	49982	71709	84793	93874	100775	106524	110839	114411	117507	119784
2006-10-01	71692	101671	120151	133815	143029	149423	154704	158843	161888	164097
2007-01-01	63976	89524	104879	116125	123886	130064	135364	139655	143139	145725
2007-04-01	62908	87738	102469	113509	121848	128148	132965	136765	140138	143179
2007-07-01	57010	81126	96728	109027	118942	126480	132670	137549	141393	142766
2007-10-01	73432	102235	119236	131857	141478	149142	155474	160262	162374	NA
2008-01-01	69086	95648	111961	123578	131871	138414	143565	145966	NA	NA
2008-04-01	67486	93500	109196	120165	127846	133534	135821	NA	NA	NA
2008-07-01	62748	88588	102430	112289	119677	122728	NA	NA	NA	NA
2008-10-01	77569	107101	124141	136047	140492	NA	NA	NA	NA	NA
2009-01-01	66986	92879	107428	112450	NA	NA	NA	NA	NA	NA
2009-04-01	69909	96281	104723	NA	NA	NA	NA	NA	NA	NA
2009-07-01	58504	70612	NA	NA	NA	NA	NA	NA	NA	NA
2009-10-01	45583	NA	NA	NA	NA	NA	NA	NA	NA	NA

```
> CL.model <- chainladder(triangle.cumule)
```

```
> fact.dev <- sapply(CL.model$Models, coef) # a comparer avec 'fact.dev.CL' calc
       x        x        x        x        x        x        x        x        x
1.384294 1.163524 1.102059 1.067753 1.049660 1.037865 1.029157 1.022532 1.016758

> rectangle.cumule
            dev
origin        dev1      dev2      dev3      dev4      dev5      dev6      dev7
  2006-01-01 44860  62511.00  72745.00  80289.00  85893.00  90337.00  93632.00
  2006-04-01 55982  76905.00  90518.00 101090.00 108863.00 115069.00 120081.00 1
  2006-07-01 49982  71709.00  84793.00  93874.00 100775.00 106524.00 110839.00 1
  2006-10-01 71692 101671.00 120151.00 133815.00 143029.00 149423.00 154704.00 1
  2007-01-01 63976  89524.00 104879.00 116125.00 123886.00 130064.00 135364.00 1
  2007-04-01 62908  87738.00 102469.00 113509.00 121848.00 128148.00 132965.00 1
  2007-07-01 57010  81126.00  96728.00 109027.00 118942.00 126480.00 132670.00 1
  2007-10-01 73432 102235.00 119236.00 131857.00 141478.00 149142.00 155474.00 1
  2008-01-01 69086  95648.00 111961.00 123578.00 131871.00 138414.00 143565.00 1
  2008-04-01 67486  93500.00 109196.00 120165.00 127846.00 133534.00 135821.00 1
  2008-07-01 62748  88588.00 102430.00 112289.00 119677.00 122728.00 127375.09 1
  2008-10-01 77569 107101.00 124141.00 136047.00 140492.00 147468.81 153052.71 1
  2009-01-01 66986  92879.00 107428.00 112450.00 120068.87 126031.47 130803.65 1
  2009-04-01 69909  96281.00 104723.00 115410.90 123230.38 129349.99 134247.82 1
  2009-07-01 58504  70612.00  82158.73  90543.75  96678.40 101479.43 105321.95 1
  2009-10-01 45583  63100.28  73418.67  80911.69  86393.73  90684.03  94117.78
```

```
  > cbind(Provision.parExercice)
 [1,]          0.000000e+00
 [2,]          0.000000e+00
 [3,]         -7.275958e-11
 [4,]         -8.731149e-11
 [5,]          2.209210e+01
 [6,]          7.691398e+02
 [7,]          2.403482e+03
 [8,]          5.500488e+03
 [9,]          8.345020e+03
[10,]          1.195160e+04
[11,]          1.585549e+04
[12,]          2.602862e+04
[13,]          2.986374e+04
[14,]          4.133798e+04
[15,]          4.397777e+04
[16,]          5.681669e+04
> (Provision.globale <- sum(Provision.parExercice))
[1] 242872.1

> ## Erreur de calcul de provision moyenne par Chain Ladder, backtesting:
> (abs(backtest.provisions.validationSample - Provision.globale) / max(c(Provisi
[1] 26.20117
```

COMPARER L'EFFICACITE DES METHODES DE PROVISIONNEMENT ?

Utiliser le backtesting ! Préparer les données comme ceci :

1. ne considérer que des sinistres clos : montant final connu ;
2. introduire une censure (administrative par ex.) pour faire apparaitre artificiellement des sinistres ouverts ;
3. définir un éch. d'apprentissage et un éch. de validation :
   - apprentissage : construire notre arbre par CART pondéré ;
   - validation : pour comparer les prévisions de provision données par l'arbre avec la vraie observation.
4. évaluer la provision relative aux sinistres encore ouvert uniquement : $E[T \mid T > y, \mathbf{X}]$;
5. faire le différentiel.

PROVISIONS À DIFFERENTES DATES D'ARRETE

On a pris ici une date d'arrêté (01/10/2009) qui excède la durée max. du risque (3 ans) $\Rightarrow$ impact de la censure limité...

Plaçons nous maintenant à des dates d'arrêté intermédiaires successives, plus proches du début de la période d'observation...
Voici l'algorithme à implémenter : pour chaque durée atteinte k,

1. sélectionner sinistres (censurés ou non) avec $Y \geq \mathrm{k}$;
2. estimer les poids KM depuis les données ;
3. construire CART pondéré pour estimer $E[T - k \mid T > k, X]$;
4. élaguer l'arbre ;
5. prévoir la durée de vie résiduelle
6. accroitre k et revenir à l'étape 1.

CONSTRUCTION DE LA BASE

On découpe par période les paiements...

Sex	SurvDate	Cause	ComNet	BegIndDateW	EndIndDateW	BegAgeW	EndAncIndW	NonCensure	SPC
F	2008-01-18	Sickness	Net_A	2008-02-17	2008-04-14	51.96441	57	TRUE	Employee
F	2009-05-06	Sickness	Net_C	2009-06-05	2010-07-29	42.68583	419	TRUE	Employee
F	2009-03-16	Sickness	Net_C	2009-04-15	2011-12-31	50.09993	990	FALSE	Employee

2009-12-31.NonCensure	2009-12-31.EndObsW	2009-12-31.PredictCART	2010-03-31.NonCensure	2010-03-31.EndObsW
TRUE	57	NA	TRUE	57
FALSE	209	239.7403	FALSE	299
FALSE	260	234.5195	FALSE	350

2010-03-31.PredictCART	2010-06-30.NonCensure	2010-06-30.EndObsW	2010-06-30.PredictCART	2010-09-30.NonCensure
NA	TRUE	57	NA	TRUE
226.4615	FALSE	390	234.7112	TRUE
232.1991	FALSE	441	225.1316	FALSE

2010-09-30.EndObsW	2010-09-30.PredictCART	2010-12-31.NonCensure	2010-12-31.EndObsW	2010-12-31.PredictCART
57	NA	TRUE	57	NA
419	NA	TRUE	419	NA
533	215.8923	FALSE	625	200.5426

RESULTATS

	Observation dates:							
	01/01/08	04/01/08	07/01/08	10/01/08	01/01/09	04/01/09	07/01/09	10/
Quantities of interest:								
(1) Size of the learning set (backtest data)	20 542	23 370	26 214	28 740	31 962	34 796	37 700	40
(2) Size of the validation set (backtest data)	10 271	11 686	13 107	14 371	15 982	17 399	18 850	20
(3) Corresponding censoring rate in learning set	16.11%	13.94%	12.9%	11.37%	11.97%	10.36%	9.55%	8.
(4) Corresponding censoring rate in validation set	16.24%	13.66%	12.8%	11.4%	11.89%	10.32%	9.27%	8.
(4bis) Number of backtested claims : (4) × (2)	1688							
Application of Section 4.3.1: 1$ a day								
(5) Total paid amount at observation date	818 079	955 809	1 115 449	1 259 591	1 448 942	1 608 799	1 771 356	1 95
(6) Paid amount (censored claims) at observ. date	278 230	286 354	323 982	336 883	378 083	388 346	387 616	39
(7) Final backtested paid amount (censored claims)	657 047	650 253	708 685	719 172	778 448	780 152	768 116	74
(8) Exact global reserve (backtested) : (7) − (6)	378 817	363 899	384 703	382 289	400 365	391 806	380 500	34
(9) Global reserve by Chain Ladder (CL)	151 017	166 614	193 593	207 677	243 701	242 688	254 947	25
(10) Error of CL : ((9) − (8))/(8)	-60.1%	-54.2%	-50%	-45%	-39%	-38%	-33%	-2
(11) Global reserve by weighted CART (wCART)	211 357	227 088	263 030	312 400	402 398	384 361	387 525	37
(12) Error of wCART : ((11) − (8))/(8)	-44,2%	-42%	-31.6%	-18.3%	**0.5%**	-1.9%	1.8%	9.

REMARQUES FINALES

+ Technique particulièrement intéressante pour les secteurs à développement long.

+ Résultats théoriques de convergence.

+ Pouvoir discriminant des facteurs de risque.

+ Extensions possibles en travaillant sur la fonction de perte de l'algorithme.

* Possibilité de remplacer cette technique de provisionnement par tt modèle sur risques individualisés (modèle de Cox, ...)

- Instabilité : typique des CART (random forests, ...).

CONCLUSION GENERALE DU COURS

En décomposant pour chaque grand thème du cours :

1. tarification :
   - modèles de tarification a priori (type GLM) permettent de tenir compte des caractéristiques individuelles des assurés, à l'inverse des modèles de crédibilité ;
   - modèles de crédibilité permettent d'intégrer dans le tarif un historique de sinistres, au contraire des modèles a priori ;
   - questionnement sur les données récoltées est primordial pour une bonne mise en place des modèles (hétérogénéité, surdispersion, exposition au risque, franchise, recours, réassurance, inflation, forfait, ...) ;
   - il est essentiel d'être rigoureux lors de l'étape de statistiques descriptives et d'optimisation des modèles pour trouver le bon niveau de segmentation ;

2. zonier :
   - ils représentent la vision géographique du risque ;
   - ils n'incluent pas la quantification du risque lié aux autres facteurs de risque (âge, ...) ;
   - peuvent être bayésien ou fréquentiste ;
   - s'ajustent au niveau de découpage géographique voulu par son utilisateur ;

3. provisionnement :
   - d'autres méthodes que les méthodes classiques (Mack, ...) permettent d'étendre la gestion de problématiques complexes (corrélation, ...) dans les triangles de liquidation ;
   - les techniques bayésiennes offrent de la flexibilité en termes de modélisation et d'hypothèses, au prix d'une complexité accrue en termes d'implémentation ;
   - le provisionnement individuel, en plein essor, repose sur une vision individuelle de chaque risque ;
   - ce dernier type de provisionnement nécessite de travailler à part sur les IByR...

BIBLIOGRAPHIE


Albert, A. and Anderson, J. A. (1984).
On the Existence of Maximum Likelihood Estimates in Logistic Regression Models.
Biometrika, 71(1) :1–10.


Aouizerate, J. (2012).
Alternative neuronale en tarification sante.
Bull. français d'Actuariat, 12(23).


Boskov, M. and Verrall, R. J. (1994).
Premium rating by geographical area using spatial models.
ASTIN Bull., 24(1) :131–143.


Boucher, J. P. and Danail, D. (2011).
On the Importance of Dispersion Modeling for Claims Reserving : An Application with the Tweedie Distribution.
Variance, 5(2) :158–172.


Brass, W. (1964).
Uses of census and survey data for the estimation of vital rates.
In *African Semin. Vital Stat.*, United Nations document E/ CN .14/CAS .4IVS/7.


Brass, W. and Macrae, S. (1984).
Childhood mortality estimated from reports on previous births given by mothers at the time of a maternity : I. Preceding-births technique.
In *Asian and Pacific Census Forum*, volume 11.


Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984).
Classification and Regression Trees.
Chapman and Hall.


Brouhns, N., Denuit, M., Masuy, B., and Verrall, R. (2002).
Ratemaking by geographical area in the Boskov and Verrall model : a case study using Belgian car insurance data.


Firth, D. (1993).
Bias reduction of maximum likelihood estimates.
Biometrika, 80(1) :27–38.


Frees, E. W. (2009).
Regression Modeling with Actuarial and Financial Applications.
International Series on Actuarial Science. Cambridge University Press, New York.


Gey, S. and Nedelec, E. (2005).
Model selection for CART regression trees.
IEEE Trans. Inf. Theory, 51(2) :658–670.


Greene, W. H. (2008).
Econometric Analysis (6th Edition).
Prentice Hall, New Jersey.

J. Besag, J. Y. and Mollie, A. (1991).
Bayesian image restoration, with two applications in spatial statistics.
Ann. Inst. Stat. Math., 43(1) :1–59.

King, G. and Zeng, L. (2001).
Logistic Regression in Rare Events Data.
Polit. Anal., 9(2) :137–163.

Lee, R. D. and Carter, L. R. (1992).
Modeling and forecasting U.S. mortality.
J. Am. Stat. Assoc., 87(419) :659–671.

Leroy, G. and Planchet, F. (2016).
Un regard actuariel sur les evolutions de l'assurance automobile.
Risques, 105.

Lopez, O., Milhaud, X., and Therond, P.-E. (2016).
Tree-based censored regression with applications in insurance.
Electron. J. Stat., 10 :2685–2716.

Mahy, S. and Denuit, M. (2002).
Decoupage geographique par zones de Voronoi en assurance automobile.

Manski, C. F. and Lerman, S. R. (1977).
Estimation of Choice Probabilities from Choice-based Samples.
Econometrica, 45(8) :1977–1988.

MATHIS, J. (2009).
Elaboration d'un zonier en assurance de vehicules par des methodes de lissage spatial basees sur des simulations MCMC.
PhD thesis.

McCullagh, P. and Nelder, J. A. (1989).
Generalized linear models, 2nd ed.
Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Mehta, C. R. and Patel, N. R. (1995).
Exact logistic regression : Theory and examples.
Stat. Med., 14(19) :2143–2160.

Olbricht, W. (2012).
Tree-based methods : a useful tool for life insurance.
Eur. Actuar. J., 2(1) :129–147.

Paglia, A. and Phelippe-Guinvarc'h, M. V. (2011).
Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique.
Bull. français d'Actuariat, 11(22) :49–81.

Pouna Siewe, V. (2010).
Modeles additifs generalises : Interets de ces modeles en assurance automobile.
PhD thesis, ISFA.

Taylor, G. (2001).
Geographical premium rating by Whittaker spatial smoothing.
ASTIN Bull., 31(1) :147–160.


Taylor, G. C. (1999).
Use of spline functions for premium rating by geographical area.
ASTIN Bull., 19(1) :91–122.


Vasechko, O., Grun-Rehomme, M., and Benlagha, B. (2009).
Moelisation de la frequence des sinistres en assurance automobile.
Bull. français d'Actuariat, 9(18).


Xie, Y. and Manski, F. (1989).
The logit model and response-based samples.
Sociol. Methods Res., 17(3) :283–302.