Modéliser la sinistralité en risques industriels
L'apport de la théorie des valeurs extrêmes

Soutenance de mémoire pour l'obtention du diplôme de Statisticien
Mention Actuariat et l'admission à l'Institut des Actuaires

18 mai 2016

Introduction

Enjeux

- Enjeux majeurs en risques industriels :
  - **Segmenter** pour gagner en compétitivité dans un marché concurrentiel
  - Utiliser des méthodes simples en interprétation et calibrage
- Approche intuitive : utiliser les méthodes inspirées du risque de particuliers
  - Modèle classique de fréquence coût
  - Utilisation de GLM sur les deux modèles séparés
- Adapter cette méthode aux risques industriels ?
  - Appliquer cette méthode en **vérifiant le cadre des hypothèses** dans un contexte de risques industriels
  - Proposer des adaptations **concrètes et industrialisables** s'adaptant à ce risque.

Problématique

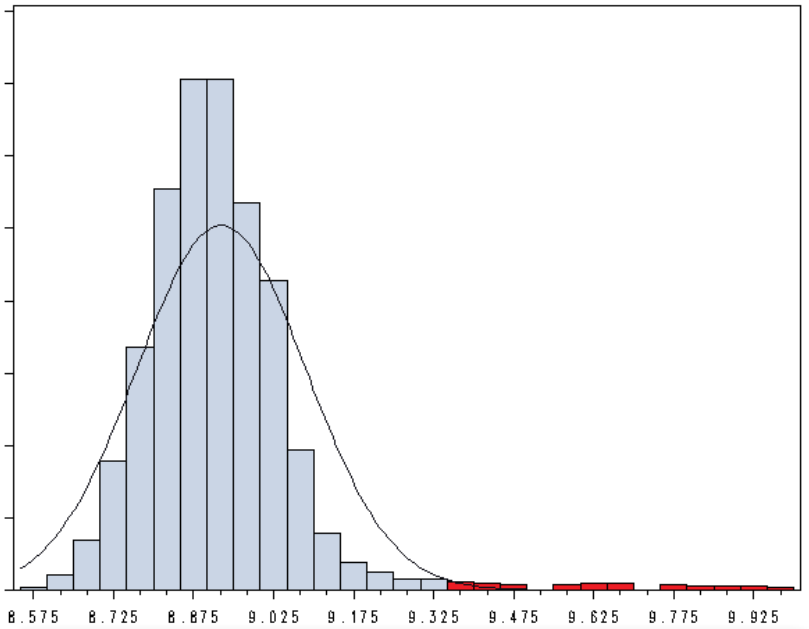
Les limites du GLM

➲ Mais les GLM reposent sur des hypothèses fortes, notamment deux contestables en risques industriels :

- ➲ La variable réponse suit une loi de la **famille exponentielle**
- ➲ Les observations sont **indépendantes**

➲ Un premier constat en utilisant le GLM en présence de queue épaisse

➲Instabilité des coefficients



Distribution intercept coût moyen (échantillonnage)

➲Mauvais pouvoir prédictif

Garantie	Obs/Predit
Incendie	80,86 %
Bris de machine	252,01 %
Vol	96,24 %
Bris de glace	76,24 %
Dégâts des eaux	133,66 %

S/P sur la base de prédiction

➲ La présence d'une sinistralité d'intensité créée de l'instabilité

GENERALI

Problématique

Quel modèle pour les risques industriels ?

➲ Utiliser la théorie des valeurs extrêmes et un modèle de propension :

$$\boldsymbol{E[S] = E[N] * \left(P_g * E[C_g] + (1 - P_g) * E[C_{\bar{g}}]\right)}$$

- ➲ Conserver la **distinction** fréquence et coût dans la modélisation
- ➲ Séparer sinistralité **attritionnelle et grave**

➲ La démarche :

- ➲ Présentation des données
- ➲ Hypothèse de distribution de la famille exponentielle : **théorie des valeurs extrêmes**

$$E[S] = E[N] * (P_g * \boldsymbol{E[C_g]} + (1 - P_g) * E[C_{\bar{g}}])$$

- ➲ Hypothèse d'indépendance : les **équations d'estimation généralisées.**

$$E[S] = \boldsymbol{E[N]} * (\boldsymbol{P}_g * E[C_g] + (1 - \boldsymbol{P}_g) * \boldsymbol{E[C_{\bar{g}}]})$$

- ➲ Quel impact sur la modélisation ?

PRÉSENTATION DE LA BASE DE DONNÉES

Présentation de la base de données

Définition du périmètre

- Les risques industriels dans l'univers entreprise dommage :
  - Un élément clé de la relation assureur-assuré
  - Une branche particulièrement touché par la sinistralité grave

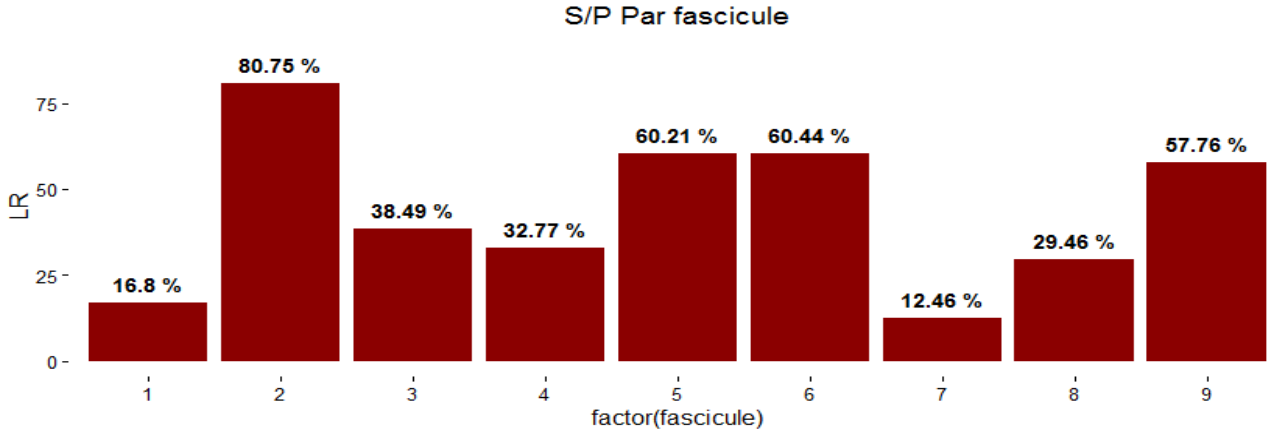
- Périmètre de notre étude :
  - Périmètre contrat :
    - Contrats en cours de validité entre 2007 et 2014
    - Relevant du périmètre de souscription 2014
    - Suffisamment renseignés
  - Périmètre sinistre :
    - Tous les sinistres associés à ces contrats
    - Exclusions des sinistres RC & Climatiques

GENERALI

Présentation de la base de données

Redressements sur les variables

- L'activité de l'entreprise :
  - L'activité de l'entreprise représente une **information capitale**
    - Importante pour la **segmentation** du risque



    - Permet de **redresser** et **lier** d'autres variables
  - En présence de multi-activités nous retenons la principale
  - Si l'information d'une année est absente, nous utilisons celle d'une autre
- Redressement des clauses associées au contrat remontées en texte libre :
  - Elles représentent l'engagement des entreprises dans la prévention des risques
  - Elles ont une influence sur le **coût attritionnel**

GENERALI

Présentation de la base de données

Redressements sur les sinistres

➲ Evaluer le coût réel des sinistres :

- ➲ La franchise **sous-évalue les coûts** et doit être recalculée
  - ➲ Estimée en pourcentage du sinistre, elle vaut alors :

$$\tilde{F} = \frac{\alpha}{1-\alpha} * (\text{Dépenses} - \text{Recettes} + \text{Provisions})$$

- ➲ Mise à la même base monétaire grâce à **l'inflation** et la **conversion** en euro
- ➲ Prise en compte de la **coassurance fréquente** en risques industriels

$$C = \text{inflation} * \text{taux de change} * \frac{\text{Franchise} + \text{Dépenses} - \text{Recettes} + \text{Provisions}}{\text{Taux de coassurance}}$$

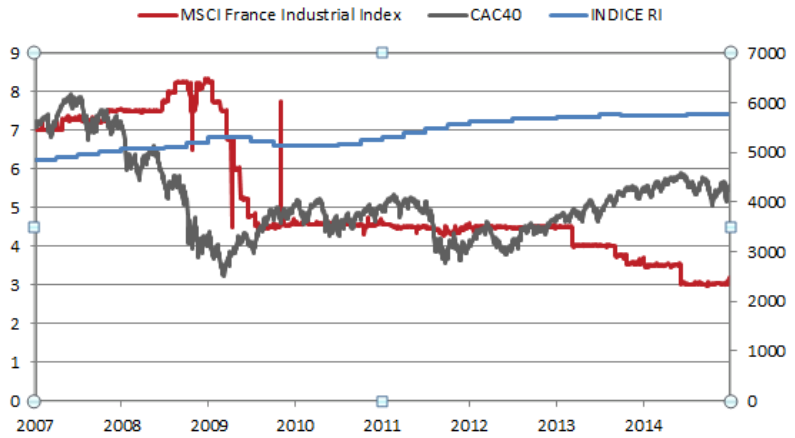
- ➲ Nous obtenons donc des sinistres **comparables** à leur **juste valeur**

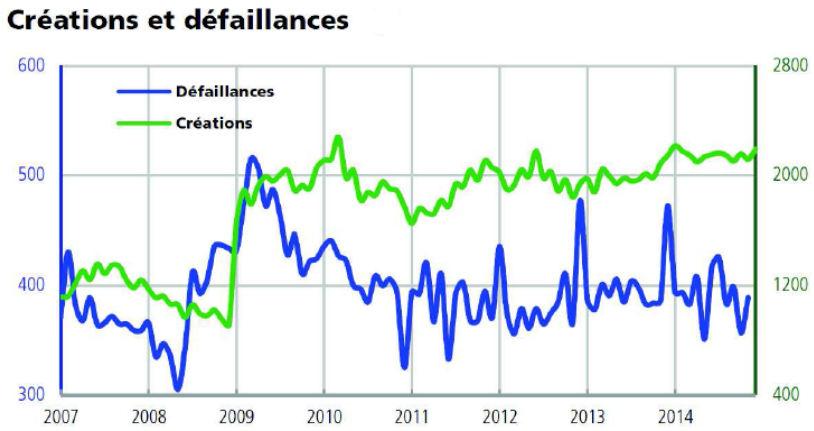
Présentation de la base de données

Les données externes

➲ Utilisation d'indices

- ➲ CAC 40, MSCIFII et indice RI
- ➲ Pas de distinction pour l'ensemble des contrats d'une même période





➲ Données INSEE

- ➲ Publiées annuellement par secteur
- ➲ Notamment défaillance et création d'entreprises

➲ Dans les modèles le MSCI FII est une variable explicative pour la propension de grave

GENERALI

ESTIMATION DE LA SINISTRALITÉ GRAVE

Estimation de la sinistralité grave

Objectif

➲ Dans cette partie nous verrons comment modéliser à l'aide de la théorie des valeurs extrêmes :

$$P_p = E[N] * (P_{\boldsymbol{g}} * \boldsymbol{E[C_g]} + (1 - P_{\boldsymbol{g}}) * E[C_{\overline{\boldsymbol{g}}}])$$

➲ Il faudra déterminer un seuil pour la sinistralité grave et automatiser le processus

GENERALI

Théorie des valeurs extrêmes

Principes fondamentaux

➲ Méthode par dépassement de seuil :

➲ Les lois des Pareto Généralisées ont pour fonction de répartition :

$$F_{\mu,\sigma,\xi}(x) = \begin{cases} (1-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\chi_{(x\in]\mu;\frac{\mu-\sigma}{\xi}[)} & si\ \xi<0 \\ (1-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\chi_{(x>\mu)} & si\ \xi>0 \\ (1-\exp(\frac{x-\mu}{\sigma}))\chi_{(x>\mu)} & si\ \xi=0 \end{cases}$$

➲ La distribution conditionnelle des excès est définie par :

$$F_u(x) = P(X \leq x | X > u)$$

➲ Théorème principal

Théorème de Pickands-Balkema-De Haan

Soient $X_1, ..., X_n$ des variables aléatoires indépendantes et identiquement distribuées et soit F_u la fonction de distribution conditionnelle des excès de X au-delà du seuil u. Alors F appartient au domaine d'attraction du maximum de G_ξ si et seulement s'il existe une fonction positive σ telle que :

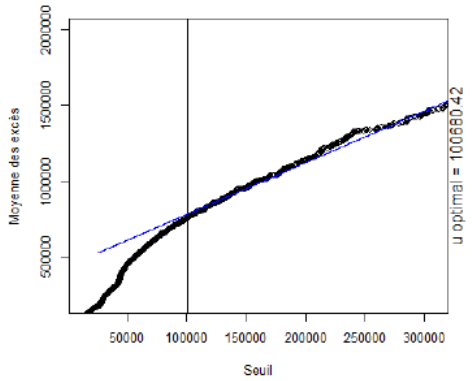
$$\lim_{u\to+\infty} F_u(x) = F^{GPD}_{\mu,\sigma,\xi}(x)$$

Etude du seuil de sinistre grave

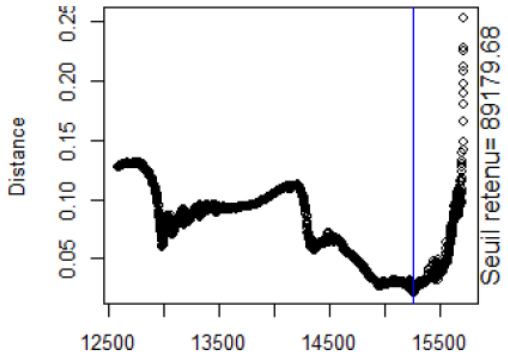
Méthodes de déterminations de seuil (1/2)

➲ Méthodes les plus utilisées et applicables en assurance :

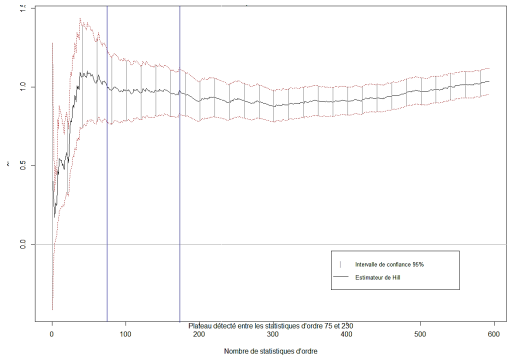
➲ Moyenne des excès



➲Distance (KS)



➲ Estimateur de Hill



➲Avantages

- ➲ Utilisent la distribution des données
- ➲ Bases théoriques
- ➲ Fournissent des informations

➲Inconvénients

- ➲ Méthodes asymptotiques
- ➲ Méthodes **graphiques**

GENERALI

Etude du seuil de sinistre grave

Déterminer un seuil (2/2)

➲Comment choisir le seuil de manière automatique ?

- ➲ Les méthodes graphiques ne sont pas **automatisables** sans prise de risques
- ➲ Chaque méthode peut apporter un seuil **différent**
- ➲ On souhaite trouver un seuil qui soit robuste

➲ Benlagha et al proposent d'utiliser une combinaison convexe de ces seuils qui minimise la variance :

- ➲ Bootstrap pour estimer la matrice de covariance
- ➲ Sélection de la combinaison linéaire minimisant la variance

➲ Avantages de cette méthode :

- ➲ Evaluer la précision des méthodes de détection automatiques
- ➲ Minimiser l'importance des méthodes inadaptées

Estimation de la sinistralité grave

Autres paramètres et coût moyen

- Estimation classique des autres paramètres :
  - Par maximum de vraisemblance, moments pondérés et estimateurs dédiés.
- Mais un problème de convergence de l'intégrale pour le calcul de $E[C_g]$:

$$\mathbb{E}[X] = \int_0^A (\sigma z + \mu)(1+\xi z)^{-\frac{1}{\xi}-1} dz = \mu - (\sigma A + \mu)(1+\xi A)^{-1/\xi} + \frac{\sigma}{1-\xi}\left(1 - \frac{1}{(1+\xi A)^{\frac{1-\xi}{\xi}}}\right)$$

- L'intégrale **diverge** la plupart du temps et ne permet pas de calculer de coût moyen
- En pratique la limite **infinie** de l'intégrale n'est **pas pertinente**
- En utilisant des limites d'engagement, on obtient les estimations suivantes

Garantie	Empirique/Estimé
Incendie	95%
Bris de machine	90%
Vol	90%
Bris de glace	102%
Dégâts des eaux	84%

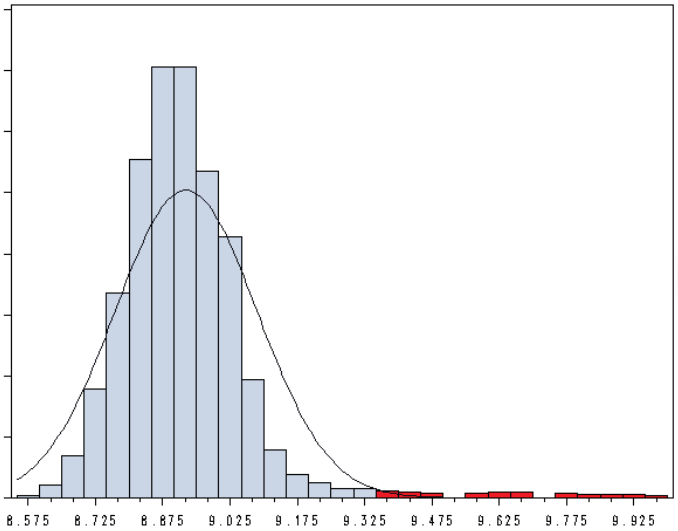
Estimation de la sinistralité grave

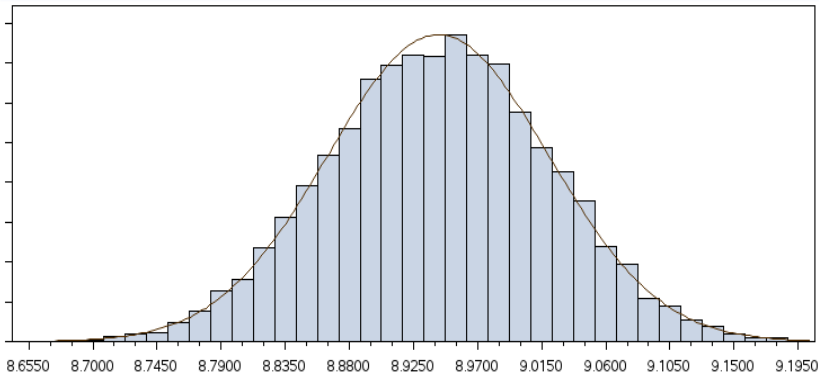
Conclusion partielle

➲ La théorie des valeurs extrêmes pour distinguer la sinistralité grave :

- ➲ La méthode par **combinaison convexe** propose une méthode industrialisable
- ➲ L'estimation des coûts moyens graves est correcte avec des bornes bien choisies

➲ Un effet sur l'estimation de la sinistralité attritionnelle :





Goodness-of-Fit Tests for Normal Distribution				
Test	Statistique		P-value	
Kolmogorov-Smirnov	D	0.00716668	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.11057513	Pr > W-Sq	0.085
Anderson-Darling	A-Sq	0.66752000	Pr > A-Sq	0.085

Distribution intercept coût moyen (échantillonnage)

- ➲ La distribution des coûts attritionnels est à **queue fine**
- ➲ La distribution des coefficients sur les échantillons est bien **normale**

GENERALI

INDÉPENDANCE ET MODÈLES LINÉAIRES GÉNÉRALISÉS

Indépendance et modèles linéaires généralisés

Objectif

➲ Dans cette partie, on s'intéresse à la modélisation de :

$$P_p = \boldsymbol{E[N]} * (\boldsymbol{P}_g * E[C_g] + (1 - \boldsymbol{P}_g) * \boldsymbol{E}[\boldsymbol{C}_{\bar{g}}])$$

➲ Les GLM permettent de modéliser ces éléments

➲ Mais quid de l'indépendance lorsque les mêmes individus se retrouvent plusieurs fois dans la même base de données avec des volumes restreints ?

GENERALI

Indépendance et modèles linéaires généralisés

Principe des équations d'estimation généralisées

➲ Equations d'Estimation Généralisées :

- ➲ Les principes et avantages des GLM sont conservés
- ➲ La présence d'une matrice de **covariance** dans la vraisemblance tient compte d'un lien entre les observations d'un même individu
- ➲ Introduction d'une métrique pour **pénaliser** la matrice de covariance :

$$QIC(R) = -2Q(\widehat{\beta}(R), \phi) + 2\mathrm{Tr}(\widehat{\Omega}_I \widehat{V}_R)$$

- ➲ Plus la matrice est **contrainte**, plus la pénalité est **faible**
- ➲ Permet de donner une chance à l'indépendance

Indépendance et modèles linéaires généralisés

Impact de la structure de dépendance sur les modèles

➲ Comparaison des structures de dépendance :

Covariance	Signification sur le modèle	QIC Fréquence	QIC Couts attritionnels	QIC Propension
Indépendante	Pas de corrélation apparente dans le temps	51 301	70 356	6 453
Autorégressive	Les résultats d'une année sont influencés par ceux des autres avec décroissance dans le temps	51 256	71 027	6 477
Echangeable	Les résultats d'un individu sont corrélés sans perte d'information dans le temps	50 799	72 521	6 500
Non contrainte	Corrélation sans structure	50 825	71 930	6 +512

➲ Pour la fréquence on préférera une structure de covariance échangeable : le nombre de sinistres d'un client apporte une information forte

➲Les modèles de coûts attritionnels et propension s'adaptent bien à une structure classique

GENERALI

ANALYSE DES RÉSULTATS ET CONCLUSION

Analyse des résultats

Un modèle plus stable

➲ Des coefficients plus stables mais des erreurs équivalentes :

- ➲ Les coefficients sont **stables** sur la base d'apprentissage
- ➲ Comparaison des erreurs et prédictions sur la base de validation :

Garantie	RMSE		Obs/Predit	
	Modèle proposé	GLM	Modèle proposé	GLM
Incendie	**11,23**	11,30	**93 %**	81 %
Bris de machines	5,84	**5,83**	**94 %**	252 %
Vol	**5,64**	5,65	**100 %**	96 %
Bris de glaces	3,86	**3,83**	70 %	**76 %**
Dégâts des eaux	**9,50**	9,51	**95%**	134 %

- ➲ Le modèle proposé n'améliore pas grandement le RMSE sauf en incendie
- ➲ **Meilleure stabilité** des modèles sur les prédictions

Analyse des résultats

Conclusion

- Une méthode de tarification innovante
  - Utilisation concrète de la **théorie des valeurs extrême**
  - Prise en compte des structures de dépendance entre les observations grâce aux **équations d'estimation généralisées**.
- Une méthode pleinement opérationnelle
  - Méthodes d'analyses similaires (AIC/QIC, résidus, modélisation stepwise)
  - Facilement interprétable et calibrable
  - Simple à reproduire grâce aux travaux d'automatisation
- Un modèle plus robuste
  - **L'indépendance** entre les observations n'est plus nécessaire
  - **Plus robuste** face à la sinistralité d'intensité

Analyse des résultats

Pour continuer

- Un modèle applicable à d'autres typologies d'assurance :
  - Gain de RMSE (2%) et Gini (4 points) lors d'un pricing game en assurance auto

- Améliorer la modélisation du coût grave ?
  - Calculer des estimations sur des ensembles homogènes ?
  - Modèles VGAM (Vecteur additifs Généralisés) ?

Merci
de votre attention