# La tarification des produit d’assurance IARD à l’ère du Big Data

Sommaire

[La tarification des produit d’assurance IARD à l’ère du Big Data 1](#_Toc519094412)

[Introduction 2](#_Toc519094413)

[Première Partie : Principes de la tarification des contrats IARD 4](#_Toc519094414)

[La struture du Tarif 4](#_Toc519094415)

[L’approche actuarielle classique 6](#_Toc519094416)

[Classification des risques en catégories homogènes 6](#_Toc519094417)

[La dimension géographique 7](#_Toc519094418)

[Deuxième Partie : Modèlisation avec les outils big-data et la data-science 8](#_Toc519094419)

[Préparation des Données 8](#_Toc519094420)

[Construction des modèles 8](#_Toc519094421)

[Réconciliations comptable 9](#_Toc519094422)

[Documentation 9](#_Toc519094423)

[Validation 10](#_Toc519094424)

[Troisième Partie : Usages 11](#_Toc519094425)

[Pilotage d’activité 11](#_Toc519094426)

[Valeur client 13](#_Toc519094427)

[Politique Commerciale 13](#_Toc519094428)

## Introduction

Les assureurs détiennent des millions de contrats et souhaitent modéliser l’impact de dizaines, voire centaines de variables pour tarifer au juste prix chaque contrat. Pour avoir un ordre de grandeur, nous pouvons estimer qu’un assureur qui détient un portefeuille de quatre millions de contrats et qui souhaite mesurer l’impact de cent variables sur trois ans d’activité doit traiter 1,2 milliard de point de données. Cela représente 9,6 Go de donnée brute (en comptant une moyenne de 8 octets pour le stockage d’une donnée atomique). Ce volume croit avec la taille du portefeuille, la profondeur de l’historique et le nombre de variables étudiées. Les outils classiques de l’actuaire et du statisticien ne permettant pas de traiter des volumes aussi importants. La maitrise des technologies big data est donc un facteur clef de succès pour la tarification des contrats IARD.

Face à ce large volume de données, les principaux acteurs du secteur ont historiquement fait le choix de recourir à des solutions externes développées par des vendeurs spécialisés. Le bénéfice principal de cette approche est de fournir des résultats fiables en mutualisant les couts de recherche et développement. En revanche cette approche conduit à une logique de “boite noire” et limite les capacités d’innovations pour exploiter de nouveaux axes de différenciations tarifaires. Tous les assureurs utilisent le même modèle et la compétition se fait principalement sur la qualité et la quantité des données disponibles. Cela avantage les acteurs qui disposent des plus gros portefeuilles.

Le big data et la data-science changent la donne en mettant à la disposition des actuaires des solutions en kit pour créer des modèles adaptés au business modèle et à la stratégie de leur compagnie.

L’objectif de ce livre est de fournir les bases méthodologiques et pratiques aux actuaires et data-scientistes qui souhaitent développer des outils de tarifications innovants tout en adhérants aux meilleurs pratiques développées par la communauté actuarielle.

Les méthodes proposées sont mis en œuvre dans des exemples de code disponibles sur la platforme de partage collaborative github.

Ce livre contient intentionnellement peu de formules mathématiques car l’objectif est plutôt de proposer un cadre méthodologique permettant aux praticiens de développer leur propres algorithmes adaptés au contexte commercial et stratégique de leur entreprise.

## Première Partie : Principes de la tarification des contrats IARD

### La struture du Tarif

On distingue traditionnellement trois éléments dans la structure tarifaire :

* le **tarif technique** : cout des sinistres ;
* le **tarif analytique** : cout de production incluants les sinistres et les autres charges allouées à la branche par la comptabilité analytique ;
* le **tarif commercial** : cout de production + marge.

Le **tarif technique**, également appelé **prime pure**, reflète le coût du risque, c’est à dire le coût des sinistres attendus sur la période de couverture. Le tarif technique dépend à la fois du profil du client et du produit vendu. Par exemple le niveau des franchises affecte de manière directe le montant des sinistre. En conséquence le tarif technique dépend du niveau de franchise choisi. Le tarif technique est obtenu à partir d’un modèle d’apprentissage sur les sinistres passés et d’une série d’ajustement pour tenir compte des hypothèses d’évolutions des principaux paramètres.

Le **tarif opérationnel** reflète le coût de production d’un contrat, c’est à dire la somme du coût des sinistres modélisé par le tarif technique et des autres charges imputées en comptabilité analytique. Ces charges comprennent notamment les frais généraux, les charges financières, les couts d’acquisitions, … Pour obtenir le tarif opérationnel il est nécéssaire de modéliser les charges par contrat et d’ajouter le montant obtenu au tarif technique. Par construction le tarif opérationnel est toujours supérieur au tarif technique.

Le **tarif commercial** est le tarif réellement proposé au client. Il est généralement supérieur au tarif opérationnel (et donc au tarif technique) afin de dégager une marge pour l’assureur. Toutefois il peut-être utile d’avoir sur certains segments un tarif commercial inférieur au tarif opérationnel. Par exemple un assureur pourrait envisager une stratégie consistant à proposer des tarifs d’assurance automobile très avantageux aux jeunes conducteurs qui ont une sinistralité élevée en pariant sur le fait que leur cout du risque va diminuer au fil des ans et que le contrat sera profitable à terme. Cela permet de renouveller le portefeuille en faisant payer la sinistralité des jeunes conducteurs par les clients plus expérimentés.

La **marge technique** d’un contrat est égale à la différence entre le tarif commercial et le tarif technique.

La **marge opérationnelle** est égale à la différence entre le tarif commercial et le tarif opérationnel. La marge opérationnelle peut-être négative pour un contrat.

@TODO : graphique sur les différentes primes

#### Modèle Technique

Le modèle technique a pour objectif de prédire pour chaque contrat le cout des sinistres sur la période de couverture de référence, en général un an. Pour cela il faut réaliser un apprentissage statistique sur les contrats des années antérieures.

Il existe souvent des contraintes légales ou opérationnelles qui obligent à exclure certaines variables dans les processus de tarification opérationnels, à titre d’illustration voici deux exemples :

* contrainte légale : dans le cadre de l’égalité des genre le législateur français interdit de prendre en compte la variable “sexe” dans la tarification des contrats d’assurance ;
* contrainte opérationnelle : certaines variables peuvent être inconnues ou invérifiables à la souscription. Il est préférable de ne pas les utiliser dans le tarif opérationnel. Par exemple le nombre de kilomètres parcourus avec le véhicule est une information peu vérifiable et il convient de l’utiliser avec prudence.

Dans la pratique est il donc possible de construire plusieurs modèles techniques pour différents usages :

* modèle technique non contraint : utilisé pour le suivi de l’activité et de la rentabilité réelle du portefeuille ;
* modèle technique contraint : utilisé pour les processus opérationnels de gestion des contrats : tarification, souscription et éventuellement résiliation.

Dans la suite nous nous intéresserons particulièrement aux méthodes de calcul du tarif technique.

#### Modèle de Cout

Le modèle de cout @TODO

#### Modèle Commercial

Le modèle de tarification commercial mets en oeuvre la stratégie commerciale de l’assureur. Il n’est pas obligatoirement le reflet des modèles techniques, toutefois il est très important de disposer d’outils permettant de mettre en regard le tarif commercial avec le tarif technique afin de garantir la maitrise des risques.

### L’approche actuarielle classique

#### Modélisation des fréquence par une loi de Poisson

#### Modélisation des couts par une loi Gamma

### Classification des risques en catégories homogènes

Afin de réaliser une bonne modélisation il est important de catégoriser les risques dans des catégories homogènes afin de pouvoir isoler les facteurs de risques spécifique à chaque type de risque et identifier la loi de probabilité réelle de chaque catégorie.

Un contrat d’assurance IARD couvre généralement plusieurs garanties. Par exemple un contrat Multi-Risque Habitation couvre l’assuré contre les dégâts des eaux, vol, incendie. Chaque garantie a des facteurs explicatifs différents et il est donc recommandé de modéliser chaque type de risque séparément.

Au sein d’une même garantie il est parfaois possible d’isoler plusieur risques de nature différente. Par exemple la garantie de responsabilité civile en assurance automobile couvre à la fois les dommages corporels et matériels des tiers, or le développement de ces deux types de sinistres sont totalements distincts. Il convient donc, si possible de les modéliser séparément. En assurance habitation il peut-être utile de modéliser de manière séparée les appartement et les maisons. De la même manière sur les garanties des professionnels il peut être utile de modéliser séparément certains secteurs d’activités tels que l’agriculture, l’industrie et le secteur tertiaire (ou tout autre disctinction permettant de classer les sinistres dans des catégories homogènes).

Enfin, certaines garanties sont susceptibles de subir des typologies de sinistres distinctes. Par exemple pour les garanties “incendie”, on peut constater généralement deux types de sinistres :

* des sinistres attritionnels relativement fréquents et peu graves : dommages limités à une pièce et quelques meubles ;
* des sinistres rares mais d’une gravité exceptionnelle : perte totale du logement et de son contenu.

Dans cette situation il est également recommandé de modéliser de manière distincte les différents types de sinistres. Dans l’exemple précédent il serait judicieux de modéliser séparément les sinistres attritionnels et les graves.

Le jugement du praticien est bien sûr nécéssaire pour distinguer quand il est utile de distinguer les risques et de réaliser des modèles séparé et lorsque cela aboutirait à une complexité inutile au vu de la matérialité des risques.

Pour résumer il est recommandé de modéliser séparément :

* chaque type de garanties ;
* au sein de chaque garantie les risques différents ;
* lorsque nécéssaire : les sinistres attritionnels et les sinistres graves ;

#### Les modèles linéaires généralisés (GLM)

### La dimension géographique

##### Zonier

Lissage spatial

Données externes

## Deuxième Partie : Modèlisation avec les outils big-data et la data-science

### Préparation des Données

Bases Contrats

Bases Sinistres

Base d’événements

Base Image

Selection des données

### Construction des modèles

Méthodologie

Metrique

Optimisation convexe

Méthode d'ordre 1  
  
 Descente du Gradient  
  
 Descente par Coordonées  
  
 Descente du Gradient Stochastique  
  
Méthode d'ordre 2  
  
 Newton-Raphson  
  
 ...

Selection des variables

Greedy  
  
Pénalisation

Traitement des modalités à faible exposition

Lissage  
  
Smoothing

Interactions entre variables

Lissage spatial

### Réconciliations comptable

Réintégration des sinistres exclus

Projections à l’ultime

Tendances et Inflation

Ajustement des niveaux de base

### Documentation

Dictionnaire des données

Distribution des données modélisées

Modèles

Relativities  
  
Base levels

Courbe de gain (lift curve)

Résidus

Courbe des relativités par variable

### Validation

Seconde Opinion

## Troisième Partie : Usages

### Pilotage d’activité

#### Calcul de la Prime Pure

#### Estimated Loss Ratio

##### Principes

L’ELR signifie « Estimated Loss Ratio ». Il est calculé comme le ratio de la prime pure modélisée divisée par la prime commerciale observée.

ELR = Prime Pure / Prime Commerciale

L’ELR est une modélisation du ratio S/C. Le S/C est défini comme le ratio du coût des sinistres observés sur la cotisation acquise sur la période d’observation. L’ELR est donc un indicateur avancé de la rentabilité d’un contrat, d’un segment de client ou du portefeuille.

La prime pure modélisée est le cout statistique des sinistres pour une période donnée, avec les informations disponibles à un instant déterminé. Les informations utilisées pour le calcul de la prime pure modélisée doivent être identiques à celles utilisées lors de la phase de modélisation.

La prime commerciale observée est le prix hors taxe qui est facturé au client. La prime commerciale doit être observée au même moment que le risque utilisé pour le calcul de la prime pure modélisée.

L’ELR s’exprime par un nombre : 0.7 ou un pourcentage : 70%. Il exprime la part de la prime commerciale nécessaire pour couvrir les sinistres. Par exemple un ELR de 70% signifie que 70% de la prime perçue sera nécessaire pour couvrir les sinistres. Un ELR de 1,3 ou 130% signifie qu’il faudra mobiliser 130% de la prime pour couvrir les sinistres attendus.

L’ELR est un indicateur instantané sans dimension. Ainsi l’ELR est indépendant de la devise utilisée et de la période de temps choisie. Cela découle du fait que l’ELR est un ratio de deux données qui sont homogènes à un prix sur une durée temporelle. En revanche il faut veiller à ce que la prime pure et la prime commerciale soient exprimées dans les mêmes unités de devise et de durée. Ainsi il est possible d’utiliser la prime pure mensuelle en centimes divisée par la prime commerciale mensuelle en centime. La recommandation du Blueprint AXA est d’utiliser une prime annuelle en euro.

L’ELR peut être calculé à différents niveaux de granularité : contrat, segment de client, portefeuille. Si l’ELR est calculé sur plusieurs contrats alors il faut diviser la somme des primes pures par la somme des primes commerciales (et non pas faire la moyenne des ELR des contrats individuels).

L’ELR peut également être calculé à tout instant de la vie du contrat, et en particulier, mais pas uniquement en date de souscription, résiliation, renouvellement à l’échéance, avenant.

ELR par garantie : il est possible de calculer un ELR par garantie ou un ELR total.

L’ELR a vocation à être suivi dans le temps pour :

* Evaluer la politique tarifaire de souscription ;
* Evaluer la politique tarifaire de renouvellement au terme ;
* Evaluer la politique tarifaire des avenants ;
* Suivre la rentabilité par segment ;
* Suivre la rentabilité par génération ;

##### Suivi par Segment

Le suivi de l’ELR par segment permet de suivre dans le temps la rentabilité attendue sur un segment de client. Pour cela il faut calculer l’ELR en vision « photo » à chaque fin de période par groupe de clients regroupés en segment. L’ELR du segment est égale à la somme des primes pures modélisées sur tous les contrats du segment divisée par la somme des primes commerciales observées sur ce même périmètre de contrat.

##### Suivi par fait de production

Le suivi de l’ELR par fait de production permet de suivre dans le temps la rentabilité attendue sur les contrats ayant réalisé un mouvement spécifique : Affaire Nouvelle, Résiliation, Changement de véhicule, … Pour cela il faut calculer l’ELR en vision « photo » à chaque fin de période par groupe de clients ayant réalisés l’un des mouvements suivis. L’ELR du fait de production sur la période est égale à la somme des primes pures modélisées sur tous les contrats ayant eu le fait de production visé divisée par la somme des primes commerciales observées sur ce même périmètre de contrat.

##### Suivi par cohorte

Le suivi de l’ELR par cohorte permet de suivre l’évolution dans le temps d’une population donnée. Par exemple il est utile de suivre les contrats par génération. Pour cela il faut calculer l’ELR en vision « photo » à chaque fin de période par groupe de clients appartenant à chaque cohorte. L’ELR de la cohorte sur la période est égale à la somme des primes pures modélisées sur tous les contrats de la cohorte divisée par la somme des primes commerciales observées sur cette cohorte.

### Valeur client

Modèle de Marge commerciale

Modèle de résiliation

### Politique Commerciale

#### Affaire Nouvelle

#### Renouvellement

####Avenant