



Mémoire présenté le : 08 Mars 2016

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA  
et l'admission à l'Institut des Actuaire

Par : Mustapha AOUICHI

Titre Analyse de l'assurance voyage à travers la tarification du risque

Confidentialité :  NON  OUI (Durée :  1 an  2 ans)

*Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus*

*Membre présents du jury de l'Institut des Actuaire*

signature

*Entreprise*

**F. PLANCHET**

*Nom : Mondial Assistance*

**D. VALEE**

*Signature :*

*Membres présents du jury de l'ISFA*

*Directeur de mémoire en entreprise :*

**J.L RULLIERE**

*Nom : S. BONNEAU*

*Signature :*

*Invité :*

*Nom :*

*Signature :*

**Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité)**

Signature du responsable entreprise

Secrétariat

Signature du candidat

Analyse de l'assurance voyage à travers la  
tarification du risque.

*Résumé*

L'assistance relève de la branche 18 du secteur de l'assurance. Le marché de l'assurance non-vie étant actuellement dans un contexte de plus en plus standardisé et concurrentiel, l'assistance apparaît comme un réel relais de croissance et de démarcation entre les sociétés d'assurance.

C'est dans ce contexte que les sociétés d'assistance sont amenées depuis peu à acquérir une véritable expertise des produits d'assurance vendus sur le marché notamment au travers de compétences actuarielles.

Afin d'accroître ses compétences, Mondial Assistance qui appartient au groupe Allianz Assistance, a fait le choix de créer récemment le service actuariat technique.

Dans cette étude réalisée au sein du service actuariat de la direction technique, une refonte des produits d'assurance voyage de Mondial Assistance France commercialisés via notre site, nous a semblé nécessaire, d'autant plus que notre dernière revue tarifaire datait de 2010.

Avec pour objectif de segmenter davantage la tarification au sein du portefeuille d'assurés, afin de ne pas pénaliser les assurés de la même manière et proposer un tarif plus juste pour un profil de risque donné.

Ce mémoire s'inscrit aussi dans les travaux de construction d'un nouveau modèle de prime pure en assurance voyage, ventilée par type de produits Assistance et Assurance pour nos ventes en BtoC. Compte tenu de la stabilité au niveau de nos sources de données depuis 2010 sur ce secteur, et de la volonté de développer de nouvelles méthodes de tarification, il nous a paru nécessaire de tester l'application des Modèles Linéaires Généralisés (GLM) et des Modèles Additifs Généralisés (GAM) afin d'avoir un axe d'analyse différent.

Ces méthodes se sont imposées depuis plusieurs années en Europe dans la tarification de contrats en assurance non-vie, comme une bonne alternative aux méthodes déterministes classiquement utilisées. Elles nous permettront aussi d'avoir une meilleure connaissance de notre sinistralité afin de mieux orienter notre offre, tant au niveau tarifaire que marketing.

Pour cela, nous effectuerons tout au long de ce mémoire plusieurs analyses statistiques sur l'ensemble de nos variables afin de proposer une grille tarifaire que nous soumettrons aux services souscription et marketing, de façon à vérifier la pertinence des tarifs déterminés.

*Résumé*

The assistance comes under class 18 in the insurance industry. The non-life insurance market is currently in a context of increasingly standardization; assistance appears to be a real growth driver and demarcation between insurance companies.

In this context assistance companies must increase level of expertise for insurance products they sold in particular through actuarial skills.

In view of strengthening its own skills, Mondial Assistance which belongs to Allianz Group, has recently made the choice to create a technical actuarial service.

In this study within the actuarial department of Technical Direction, we thought to redesign travel insurance product Mondial Assistance France provide through our site, especially as our last pricing review was made in 2010.

The aim is to go in a further price segmentation of our insured portfolio, in order to deliver the best estimate price for each given risk profile.

This memoir takes part in the construction of a new model for determining the travel insurance pure premium and allocates it between assistance and insurance products for our BtoC sales.

Given the stability in our data sources in this sector since 2010, and willingness to develop new pricing methods, we felt necessary to put to test the application of Generalized Linear Models (GLM) and Models Generalized Additive (GAM) in order to have an analysis from another point of view.

Since several years in Europe, these methods have imposed themselves in the pricing of non-life insurance contracts as a good alternative to deterministic methods conventionally used. They will allow us to have a better knowledge of our claims and have a better orientation of our offer in both pricing and marketing level.

From this perspective, we will perform throughout this memoir several statistical analyzes on all of our variables to propose a pricing structure we will submit to underwriting and marketing department, in order to verify the relevance of the rates determined.

### **Remerciements**

Je tiens à adresser mes remerciements à Monsieur Jean-Claude Guerin, Stephanie Morlane et Severine Bonneau, membres de la direction technique et Bottais Souad Responsable formation, de m'avoir permis de suivre la formation d'actuariat dispensée par l'ISFA et de m'avoir donné la possibilité de réaliser ce challenge dans les meilleures conditions.

Je tiens également à remercier Monsieur Christophe Neves, actuaire et directeur du groupe actuariel, pour ses nombreuses remarques et propositions pertinentes.

J'adresse également ma reconnaissance à l'ensemble des collaborateurs de la direction technique de Mondial Assistance pour leur lecture et conseils précieux. Mais aussi à Carlos Isturiz, responsable projet marketing et son équipe pour nos nombreuses discussions concernant les tendances de nos données.

Je remercie également Madame Esterina Masiello et Christian Robert, enseignant à l'ISFA, pour leurs conseils et le suivi qu'ils m'ont accordé. Et de façon plus générale l'ensemble du personnel de l'ISFA notamment Clemens Delphine pour leur disponibilité et amabilité constante.

J'adresse enfin ma profonde gratitude à toute ma famille, pour leur soutien, leur bonne humeur et leur confiance totale, malgré les périodes de doute que j'ai pu traverser.

# Contenu

---

Introduction .....	8
Partie 1 : Le marché de l'assurance voyage .....	9
A. Contexte général et juridique .....	9
1. L'assurance voyage.....	9
2. Les acteurs de l'assurance voyage .....	10
3. Les intermédiaires à travers la directive : DIA.....	10
4. Les produits d'assurance voyage.....	11
5. Structure du contrat d'assurance voyage.....	12
B. Les garanties et risques. ....	14
1. Les principaux risques couverts par l'assurance voyage .....	14
2. Les principaux risques couverts par l'assistance voyage .....	16
3. Caractéristiques techniques des produits .....	17
C. Analyse stratégique du secteur.....	19
1. Les menaces .....	19
2. Les opportunités .....	21
3. Les forces et les faiblesses .....	21
D. Conclusion.....	24
Partie 2 : Mécanisme & méthodologie de tarification.....	25
A. Introduction.....	25
B. Collecte et organisation des données .....	26
1. Description de la base.....	27
2. Constitution de la base de travail .....	30
3. Création des variables de sinistralité.....	31
4. Analyse de la base constituée.....	32
C. Statistiques descriptives univariées .....	34
1. Transformation des variables .....	34
2. Etude de la corrélation entre les variables .....	42
D. Théorie des modèles linéaires généralisés (MLG) .....	43
1. Les MLG comme extension des modèles gaussiens.....	43
2. Estimation des paramètres : Méthode du maximum de vraisemblance .....	45
3. Description des variables à expliquer.....	48
E. Sélection des variables de tarification .....	56
1. Etude sur la fréquence des sinistres : Modèle de régression logistique .....	57

2.	Etude sur le coût des sinistres .....	60
3.	Analyse des résidus des régressions logistique et Gamma .....	62
F.	Modèle de tarification .....	63
1.	Principe de base.....	64
2.	Modélisation de la fréquence .....	65
3.	Modélisation du « coût moyen ».....	70
4.	Comparaison entre les coefficients tarifaires observés et ceux prédits .....	73
5.	Modèles tarifaires .....	77
G.	Méthode de tarification alternative : les Modèles Additifs Généralisés (GAM).....	85
1.	Cadre théorique.....	85
2.	Choix des fonctions de lissage de la variable continue durée du voyage.....	88
3.	Résultats de la tarification alternative grâce aux GAM .....	94
4.	Structure tarifaire finale .....	96
Partie 3 :	Conclusion .....	98
Partie 4 :	Bibliographie.....	100
1.	Livres .....	100
2.	Mémoires.....	100
Partie 5 :	Table des illustrations.....	101
Partie 6 :	Table des tableaux .....	103
Partie 7 :	Annexes.....	105
A.	Choix de nombre de classes .....	105
B.	Mesures de corrélation.....	106
C.	Sélection des variables .....	108
1.	Au titre de la fréquence .....	108
2.	Au titre de la sévérité.....	108
D.	Test d'adéquation de loi .....	109
E.	Modèle de tarification actuel.....	110
1.	Présentation des fréquences et des coefficients linéaires de durée (méthode actuel).....	110
2.	Résultats du MLG sur 3 variables sans la durée du voyage.....	111
F.	Synthèse des points d'intention réglementaire pour le secteur d'assurance .....	112
G.	Passage du document de la CAS.....	113

## Introduction

---

Le groupe Allianz Global Assistance International, est le leader dans le domaine de l'assurance et de l'assistance. Ses services rassemblent près de 250 millions de bénéficiaires sur cinq continents différents. Son implantation en France est dynamisée via Mondial Assistance France par divers services dont la vente en ligne via le site mondial-assistance.fr. Mondial Assistance France développe des offres d'assurance voyage visant à couvrir les risques liés aux déplacements personnels (vacances, études...) ou professionnels.

Le groupe Allianz Global Assistance International a réalisé 2,442 milliards d'euros en CA et un revenu net de 67,9 millions d'euros en 2013. Il compte 400 000 prestataires et 180 correspondants qui assurent la gestion des soins dans 1 287 hôpitaux et cliniques, 192 centres de soins dans 122 pays différents. Les retours provenant des assurés témoignent de leur satisfaction quant à la qualité de communication et au professionnalisme des prestataires et des correspondants.

Une enquête, menée par TNS Sofres pour Mondial Assistance, a relevé en 2012, que 83% des personnes de la catégorie de 62 ans et plus se sentent inquiets et deviennent pessimistes lorsqu'il s'agit de voyager. Ce résultat est confirmé par l'Institut Think<sup>1</sup> dans son enquête pour Groupama Assistance et l'Argus de l'assurance. 53% des sujets sondés disent ne pas avoir d'envie pour l'aventure, 57% confient avoir moins de goût pour l'aventure que par le passé et 62% estiment que voyager est devenu très risqué.

La première interprétation penche vers la considération de ce point de vue comme un obstacle à la croissance du marché de l'assurance voyage. La deuxième interprétation renseigne sur l'existence d'un créneau indiquant que plus les personnes se sentent en détresse plus elles ont besoin d'une couverture d'assurance et d'assistance pendant leur déplacement afin d'être rassurées.

Il serait ainsi exigé, de la part des assureurs, la fourniture de services diversifiés et d'une tarification de consensus pour combler les besoins de couverture des clients tout en gardant un seuil de rentabilité acceptable.

Cela nous amène à nous poser la question suivante : dans quelle mesure pourrait-on optimiser la tarification des services d'assurance voyage ?

Pour répondre à cette problématique, nous allons devoir approfondir, voire revoir notre approche d'analyse générale des produits proposés. Notamment, à travers une refonte des modèles économétriques utilisés pour le traitement des données, la tarification et la prévision.

Afin d'étudier le sujet, un plan en deux parties est proposé : une première partie abordant le contexte général de l'assurance voyage à travers les volets réglementaires, techniques et marketing ; et une seconde partie proposant un modèle de tarification basé sur les GLM pour nos produits d'assurance voyage.

---

1 Informations recueillies sur le site <http://www.argusdelassurance.com/mediatheque.pdf>

# Partie 1 : Le marché de l'assurance voyage en France

---

## A. Contexte général et juridique

L'assurance et l'assistance voyage sont des produits standards. Ils reposent dans ce sens sur des bases juridiques et techniques stables. Nous le verrons à travers cette première partie qui aborde successivement les points saillants de ces bases en présentant l'état du marché de l'assurance voyage en France, les caractéristiques techniques de ce type de couverture et enfin une analyse stratégique qui nous permettra d'identifier et de comprendre les éventuels leviers de croissance future pour ce secteur.

### 1. L'assurance voyage.

L'assurance voyage est un ensemble de services d'assurance et d'assistance consistant à couvrir les frais imprévus auxquels chaque voyageur peut avoir à faire face.

Elle désigne l'ensemble des prestations associées aux risques encourus lors d'un déplacement. Elle comporte des garanties d'assurance non vie comme les pertes de bagages (qui relève de la branche 7 « marchandises transportées ») ou l'annulation (qui relève de la branche 16 « pertes financières »). L'analyse du marché de l'assurance voyage est subordonnée à un ensemble de faits et de phénomènes de natures diverses. La croissance démographique mondiale et la libre circulation des personnes multiplient les risques d'accidents et les risques de perte de bagages. Le défi des compagnies d'assurance est de maîtriser au mieux cette situation, en proposant des offres adéquates.

Selon les sources de données du Syndicat National des Sociétés d'Assistance (SNSA), les sociétés d'assurance et d'assistance voyage ont pu réaliser en France, un chiffre d'affaires agrégé de 253 millions d'euros en 2014. Ce chiffre d'affaires est principalement composé des montants de primes d'assurance et d'assistance collectées auprès des assurés tout au long de l'année. Le syndicat constate aussi que ce chiffre d'affaires est en recul de 10.7% par rapport aux primes d'assurance et d'assistance collectées en 2013. Pourtant, deux autres réalités vont à l'encontre de ce fait. Les sociétés françaises d'assistance regroupées au sein de la SNSA ont réalisé un chiffre d'affaires global de 2,57 milliards d'euros, soit une augmentation de 7,2% par rapport à l'année 2012. Cette situation explique le défi auquel les sociétés d'assurance et d'assistance voyage sont confrontées. Les résultats du secteur dans son ensemble affichent une hausse tandis que le sous-secteur du voyage est dans un état critique de rentabilité.<sup>2</sup>

---

2 Informations recueillies sur le site <http://www.lechotouristique.com/article/le-ca-voyage-des-societes-d-assistance-a-recule-de-9-en-2013,63912>

## 2. Les acteurs de l'assurance voyage

Sur le marché de l'assurance, il y a trois grandes catégories d'offreurs à savoir les compagnies d'assurance, les mutuelles et les institutions de prévoyance. Parallèlement, nous rencontrons d'autres acteurs clés, qui travaillent en collaboration avec les compagnies d'assurance dans la distribution des contrats, ce sont les intermédiaires d'assurance : agents généraux d'assurance, courtiers en assurance, mandataires d'assurance et mandataires intermédiaires. Ces derniers exercent leurs activités sans nécessité d'autorisation octroyée auprès des institutions de régulation et de contrôle tels que l'Autorité de Contrôle Prudentiel et de Résolution (ACPR) contrairement aux sociétés d'assurance et d'assistance.

Le tableau suivant illustre la typologie et l'effectif des entreprises et intermédiaires qui travaillent sur le secteur de l'assurance.

**Tableau 1: Effectif des entreprises et intermédiaires d'assurance**

Sociétés d'assurance et d'assistance		Sociétés intermédiaires d'assurance et d'assistance	
Types d'organisme	Nombre recensé le 1er janvier 2013	Types d'organisme	Nombre recensé le 1er janvier 2013
Entreprise de réassurance	16	Courtier en assurance	21550
Entreprise d'assurance vie	64	Agent général en assurance	11844
Entreprise d'assurance non vie	215	Mandataire d'assurance	2731
Mutuelle	625	Mandataire intermédiaire en assurance	15689
Société de groupe d'assurance	19	Courtier en opérations de banque	4439
Institution de prévoyance	49	Mandataire non exclusif en opération de banque	7802
Entreprise d'assurance mixte	39	Mandataire exclusif en opérations de banques	4239
SGAM	8	Mandataire d'IOBSP	6977
Groupement paritaire de prévoyance	6	Conseiller en investissement financier	4866
Union mutualiste de groupe	3	Agent lié de PSI	3345
Succursales d'entreprise des pays tiers	5		
<b>Total</b>	<b>1049</b>	<b>Total</b>	<b>83482</b>

En analysant le tableau provenant du secrétariat général de l'ACPR et ORIAS, les chiffres confirment que le marché de l'assurance et de l'assistance est dominé par les sociétés ou les personnes physiques exerçant des activités d'intermédiation d'assurance. Leur effectif est très majoritairement supérieur aux 1049 entreprises d'assurance, ce qui explique la férocité de la concurrence et certainement l'arrivée à saturation du marché.

## 3. Les intermédiaires à travers la directive : DIA

### Définition du code des assurances – Article L511-1 :

« L'intermédiation en assurance ou en réassurance est l'activité qui consiste à présenter, proposer ou aider à conclure des contrats d'assurance ou de réassurance ou à réaliser d'autres travaux préparatoires à leur conclusion. N'est pas considérée comme de l'intermédiation en assurance ou en réassurance l'activité consistant exclusivement en la gestion, l'estimation et la liquidation des sinistres. Est un intermédiaire d'assurance ou de réassurance toute personne qui, contre rémunération, exerce une activité d'intermédiation en assurance ou en réassurance. »

Cette directive vise à établir un véritable marché unique de l'intermédiation en coordonnant et en harmonisant les réglementations nationales pour exercer l'activité d'intermédiation en assurance.

Elle permet ainsi à des intermédiaires de commercialiser des produits d'assurance voyage. Nous pouvons subdiviser ces intermédiaires en deux catégories :

- Les courtiers et agents généraux considérés comme une profession assez spécialisée dans le secteur de l'assurance.
- Les tours opérateurs et les agences de voyage assimilés à des « non spécialistes » de l'assurance.

Par définition, ces deux catégories d'intermédiaires distribuent les produits d'assurance et d'assistance suivant un schéma en BtoBtoC ou en BtoB qui représentent chez Mondial Assistance plus de 90% du CA des produits vendus. Les proportions de vente pour ces intermédiaires sont très nettement en faveur des tours opérateurs et agences de voyage avec une part de vente atteignant plus de 9 ventes sur 10.

Ces chiffres démontrent l'importance de ces deux intermédiaires dans la distribution de nos produits mais aussi l'appréhension face à ce schéma largement dominé par des acteurs nouveaux et non spécialisés sur le marché de l'assurance.

Notons que cette loi n'est pas la seule pouvant impacter la mise en place de la tarification des produits d'assurance voyage. Ainsi, un descriptif des principaux textes de lois sera disponible dans l'Annexe F.

#### 4. Les produits d'assurance voyage

Allianz Global Assistance, via le site de Mondial Assistance, propose plusieurs produits (ou solutions) temporaires ou annuels à tacite reconduction tels que :

- la solution Voyages Affaires pour les déplacements professionnels ;
- la solution Vacances pour les personnes souhaitant partir en vacances (croisière, séjour, circuit...);
- la solution Voyages Globe-Trotter pour les voyages multi-destination de plus de 3 mois ;
- la solution Voyages groupes de plus de 10 personnes ;
- et la solution Jeune/Études à l'étranger pour les voyages d'études, stages et déplacements au pair à l'étranger ;

Ces produits couvrent en général les frais de rapatriement médical, les soins hospitaliers à l'étranger, la responsabilité civile à l'étranger, le vol de bagages, les retards ou annulations de vol.

Sous la pression de la concurrence et la multiplication des acteurs opérant sur le secteur, les produits d'assurance sont distribués à travers le marché via un certain nombre de canaux :

- les canaux traditionnels de vente animés par les réseaux de commerciaux salariés, les bureaux et succursales d'assurance, les agences générales, les cabinets de courtage et les guichets des établissements bancaires et financiers ;
- la vente par Internet pratiquée pour lancer les produits à l'international qui permet de proposer des offres globalisées par le biais de schémas d'assurance en Libre Prestations de Service (LPS) ;

- les offres en package proposées par les entreprises d'autres secteurs comme les compagnies aériennes, les agences de voyages et touristiques, etc... ;

Quant aux produits offerts par mondial-assistance, les canaux de distribution utilisés sont nombreux et diversifiés. Ils comportent :

- la vente en ligne sur le site internet mondial-assistance.fr et via une plateforme téléphonique dédiée (SSCI) ;
- la vente indirecte en collaboration avec les intermédiaires d'assurance ou assimilés : agence de voyage situé en France, les courtiers, les compagnies aériennes telles qu'Air France, les banques et assureurs ;
- la vente indirecte par les affiliés : entreprises, artisans, groupement d'artisans et associations organisateurs de voyages ;

La pratique des grandes entreprises ou des fédérations d'assurance tend vers l'utilisation d'Internet pour publier les informations détaillant leurs offres, les modalités de souscription et toutes autres informations utiles en alternance avec les canaux de distribution traditionnels. Pour les assurés, ce canal facilite l'accès à l'information et représente un gain en termes de coût de transaction (temps, déplacement, etc.). Pour les entreprises d'assurance, il permet de réduire dans une certaine mesure les coûts internes.

## 5. Structure du contrat d'assurance voyage

Le contrat d'assurance voyage ou police d'assurance est un « *document qui constate les obligations réciproques du groupe Allianz Global Assistance et de l'assuré (ou souscripteur)* » découlant de leurs engagements respectifs de couvrir certains risques pour le premier en contrepartie du paiement d'une prime pour le second. Il formalise au minimum des conditions générales et des conditions particulières concernant le produit et aux plus d'autres informations utiles pour les assurés. Il doit retracer la situation et le choix de l'assuré tel que la nature du risque souscrit, les renseignements sur l'assuré, les garanties choisies, la durée du contrat et les cotisations (primes).

La constitution d'un contrat d'assurance invite l'assuré à effectuer préalablement une évaluation de ses besoins d'assurance et à se documenter sur les services couverts par Mondial Assistance.

### 5.1 La prime pure et la prime nette

L'étude de la tarification des contrats d'assurance voyage renvoie toujours à la définition de la prime pure. Il s'agit de la plus petite partie de la prime d'assurance évaluée comme le coût du risque, c'est à dire le montant moyen nécessaire à l'assureur pour payer ses obligations envers l'assuré sans profit ni perte pour la compagnie. Elle peut être évaluée de façon actuarielle lorsque les données à la disposition de l'analyste sont suffisamment complètes et de bonne qualité.

La prime nette désigne le montant restant à l'assureur, pour rembourser ses obligations envers l'assuré, comportant le coût du risque (prime pure) et l'ensemble des frais annexes (frais de gestion + Marge). Dans ce sens, la détermination de la cotisation ou prime nette prend en compte des éléments de base (charges fixes) pour l'ensemble des produits et des éléments distinctifs de chaque produit (charges variables). Ce sont donc les charges de fonctionnement évaluées par la compagnie qui différencient la prime pure et la prime nette.

$$\text{Prime nette} = \text{Prime Pure} + \text{charges de fonctionnement} + \text{Marge}$$

## 5.2 Les coûts internes

Les charges fixes représentent les coûts internes de la compagnie. Contrairement à une compagnie d'assurance classique, les compagnies proposant de l'assistance doivent faire face à des coûts internes élevés qui peuvent être supérieurs aux coûts externes. Les prestations d'assistance sont gérées en interne par des salariés comme des médecins urgentistes ou spécialistes, des assistantes sociales et des infirmières, ces frais, notamment ceux des médecins, peuvent s'avérer onéreux dans la mesure où les prestations sont délivrées 24H/24, 365 jours par an. Nous pouvons citer, comme exemple, le cas d'une personne ayant eu un sinistre à l'étranger et qui a besoin d'un rapatriement d'urgence. Dans ce cas, un médecin, un infirmier et un logisticien se chargent de faire le déplacement pour permettre une bonne prise en charge de l'assuré. Cela se traduit parfois par de nombreuses heures comptabilisées ainsi que des frais associés (hébergement, transport...). Cette notion de coûts internes sera importante pour notre tarification, notamment dans l'interprétation de certains de nos résultats.

## 5.3 Les charges diverses et indices

### Les charges des éléments accessoires

Ce sont des charges supplémentaires rattachées à de petites dépenses courantes comme les dépenses de courriers, les dépenses liées à l'édition des contrats. Elles font partie des charges incorporables au coût de revient et devront être prises en compte dans la détermination du prix de chaque produit d'assurance.

### Les charges fiscales

Les entreprises d'assurances sont assujetties à un certain nombre de taxes comme les autres entreprises telles que la Taxe sur les conventions d'assurance TCA et la TVA. Les taux, indiqués par les dispositions du Code des assurances et du Code général des impôts, varient en fonction de la nature de l'activité. Les charges fiscales sont calculées en prenant en compte le montant de la prime nette et les éléments accessoires. Selon les dispositions comptables et fiscales, elles sont aussi incorporables au prix de revient des produits d'assurance voyage proposés par les assureurs.

## Les indices

L'indice est un multiplicateur appliqué à certains coûts de prestations d'assurance voyage. Il sert à calculer et à appliquer l'élasticité du montant des cotisations et des primes par rapport à l'évolution du niveau de garanties à la charge de l'assureur. L'indice choisi par l'assureur est inscrit dans une clause nommée clause d'indexation qui doit être portée à la connaissance de l'assuré, validée et acceptée par celui-ci. En général, les indices sont applicables aux contrats à tacite reconduction annuelle comme les packs annuels proposés sur le site mondial-assistance.

## B. Les garanties et risques.

En assurance, le terme risque désigne un « phénomène aléatoire correspondant à une situation où le futur n'est prévisible qu'avec des probabilités ». Il peut affecter une personne, ses biens et sa responsabilité civile.

Le groupe Allianz propose via Mondiale assistance trois types de produit regroupant différentes garanties :

- une formule assistance qui inclut uniquement des garanties d'assistance ;
- une formule assurance qui inclut uniquement des garanties d'assurance ;
- et une formule multirisque, qui comprend des garanties d'assistance et d'assurance.

Afin de mieux cerner la composition de ces formules, une définition des différentes garanties proposées semble nécessaire. Nous verrons également, les différents impacts à prendre en compte lors de l'établissement du tarif .En effet, les périodes de garantie n'étant pas toujours linéaires au sein d'un même produit, il sera nécessaire de préciser les délais d'exposition aux risques pour chaque garantie.

### 1. Les principaux risques couverts par l'assurance voyage

#### 1.1 Définition d'un sinistre de type annulation de voyage et interruption

De nombreuses circonstances qui dépendent de la personne ou de l'emploi de l'assuré peuvent être à l'origine de la réalisation du risque d'annulation de voyage. Ainsi, le sinistre peut être par exemple défini si l'une des circonstances suivantes est observée, vérifiée et acceptée par l'assureur :

- l'assuré, le compagnon de voyage, un membre de la famille de l'assuré, un membre de la famille du compagnon de voyage, la personne sous tutelle légale de l'assuré, la personne s'occupant de la garde des enfants à charge de l'assuré, l'homme clé employé par l'assuré et la personne hôte à destination tombent malades, ou subissent un accident grave ou un décès ;
- l'assuré ou son compagnon de voyage, salarié ou professionnel dans une entreprise, se retrouve sans emploi suite à un licenciement économique soudain justifié ;

- les propriétés de l'assuré ou de son compagnon de voyage subissent un dommage majeur à cause d'un incendie ou d'un problème de canalisation d'origine naturelle ou provoqué par des actes délictueux par une tierce personne comme le vol ;
- l'assuré ou le compagnon de voyage doit effectuer un examen de rattrapage dans une école ou une université pour obtenir un diplôme de fin d'étude ou dans l'objectif d'éviter tout prolongement de la scolarité à condition que le voyage ait été réservé et le contrat conclu avant la date de l'examen raté et que la date de l'examen de rattrapage ait été annoncée de manière soudaine ;
- l'assuré ou le compagnon de voyage a reçu une convocation imprévue à un tribunal après la date de mise en vigueur de la couverture et la date de la réservation et que le tribunal ne reconnaît pas le départ en voyage comme un argument suffisant de reporter la durée de prescription de la convocation ;
- l'assuré ou le compagnon de voyage perd les documents nécessaires à leur voyage ou à leur séjour à l'étranger parce qu'ils ont été volés et ne peuvent pas être remplacés avant la date de départ en voyage.

Le choix et l'intégration de ces arguments pour l'annulation du voyage dépendent de l'étendue de la négociation entre l'assuré et l'assureur soit en termes de personnes soit en termes de circonstances. Ils devront tenir compte d'une analyse détaillée des critères spéciaux pour chaque pays de destination comme la réglementation du transport, le fonctionnement des universités et des grandes écoles et la politique publique en matière de santé.

Pour ces garanties, nous aurons un impact différent dans l'approche tarifaire dans la limite où l'assureur sera exposé :

- aux sinistres annulations pour la période qui court entre la date de souscription et la date de départ,
- aux sinistres interruptions pour la période qui court entre la date de départ et la date de retour.

## 1.2 Définition d'un sinistre sur les bagages

Certains risques liés aux bagages se réalisent et deviennent des sinistres couverts par les compagnies d'assurance dans le cadre de l'assurance voyage. Ces sinistres sont bien définis suite à la constatation objective. Les sinistres aux bagages sont limités à trois types d'incidents que sont la perte ou la disparition, la détérioration et la destruction. Les risques liés aux bagages des professionnels peuvent être couverts jusqu'à 5 000 euros avec Mondial-assistance.

L'exposition à prendre en compte dans ce cas par l'assureur lors de la tarification, correspond à la période qui court entre la date de départ et la date de retour.

## 1.3 Les sinistres engageant la responsabilité civile

Tout acte civil peut engendrer des conséquences négatives sur la personne d'autrui ou sur ses biens. À partir du moment où cela engendre des pertes financières ou des dommages à autrui pendant la durée du séjour, cela engage la responsabilité civile de l'assuré ou de son compagnon de voyage. Les produits vendus sur le site mondial assistance sont annexés d'une garantie contre ce type de risque.

Cette garantie couvre les conséquences des dommages corporels, matériels ou immatériels consécutifs, causés par l'assuré dans le respect de la jurisprudence du pays dans lequel il se trouve. La limite de couverture est généralement fixée à 4 500 000 euros pour un assuré qu'il soit un particulier ou un professionnel.

L'exposition à prendre en compte dans ce cas par l'assureur lors de la tarification, correspond à la période qui court entre la date de départ et la date de retour.

## 2. Les principaux risques couverts par l'assistance voyage

### 2.1 Un sinistre de type individuel accident

Il n'existe pas de définition uniforme de l'accident pour tous les assureurs. Pour exemple certains assimilent à un accident des événements tels que des brûlures, des déchirures ou certaines intoxications. Nous définirons donc de façon globale un accident comme un événement exogène survenu d'une manière soudaine et imprévue lors du déplacement ou du séjour de l'assuré à l'étranger. En cas d'accident, cette garantie prévoit le versement d'un capital en cas de décès ou d'invalidité permanente et définitive.

L'exposition à prendre en compte, dans ce cas par l'assureur lors de la tarification, correspond à la période qui court entre la date de souscription et la date de fin du contrat souscrit.

### 2.2 Un sinistre de type maladie

Le sinistre de type maladie inclut les maladies survenues de façon imprévue et soudaine et les maladies dues à un accident pendant le voyage ou le séjour à l'étranger. Pour être défini comme un sinistre, il faut que la maladie ou la conséquence d'un accident nécessite un traitement thérapeutique ou médical. La qualification du sinistre commence donc avec le début de traitement et finit dès que le médecin responsable du traitement prescrit la fin du celui-ci.

Par ailleurs, la définition du sinistre intègre également les troubles dus au déplacement ou aux conditions de confort proposées pendant le séjour. Néanmoins, de nombreux cas de maladie ou cas d'accident sont écartés de la définition de sinistre maladie. Cela résulte de l'appréciation de leur caractère prévisible comme les allergies.

Ce risque aura un impact différent dans l'approche tarifaire dans la limite où la maladie peut entraîner une annulation du déplacement ou se manifester lors du déplacement, dans ce cas la période d'exposition au risque à prendre en compte court entre la date de souscription et la date de retour.

### 2.3 Les frais médicaux

La prise en charge des frais médicaux est une opération qui consiste à financer les frais relatifs aux soins, à l'hospitalisation ou au rapatriement médical d'un assuré suite à un accident ou à une urgence médicale survenu pendant son déplacement (voyage, vacances, etc.). Plusieurs circonstances peuvent être à l'origine d'une prise en charge comme :

- l'opération consistant à la prise en charge sur place de l'assuré par une personne mandatée par la compagnie d'assurance ;
- l'opération d'hospitalisation d'urgence de l'assuré suite à une maladie ou à un accident ;
- l'opération de rapatriement de l'assuré vers un hôpital régional disposant de la spécialisation nécessaire à l'intervention médicale adéquate ou à un retour au domicile avec accompagnement médical ;

Le montant de frais pris en charge dépend de la nature de chaque situation, des dispositions prévues par le contrat sur la rubrique « frais médicaux, rapatriement, soins d'urgences » et enfin du type de contrat conclu entre les deux parties. Quel que soit le cas, les risques maladies et soins d'urgences peuvent engendrer des frais médicaux importants. A titre d'exemple, Mondial Assistance a pu intervenir dans un dossier concernant un trouble de la digestion due à une consommation de fast-food à New York pour un cout de 8 000 euros ou une hospitalisation d'urgence à Miami pour un cout de 53 000 euros. La survenance de ces évènements peut concerner tout type de personne. L'intérêt d'un contrat assurance voyage pour l'assuré, est que l'assureur supporte la grande partie ou la totalité de ces charges.

On peut se demander toutefois, si la prise en charge des frais médicaux relève de l'assistance ou de l'assurance. Pour la suite de notre étude et suivant la configuration des produits de Mondial Assistance, celle-ci sera considérée comme une prestation d'assistance car associée à une situation d'urgence à l'étranger (au contraire d'une assurance de frais de santé classique qui couvre la vie quotidienne).

Au niveau tarifaire, l'exposition à prendre en compte, dans ce cas par l'assureur, correspond à la période qui court entre la date de départ et la date de retour.

### 3. Caractéristiques techniques des produits

La typologie de risque constitue une variable à la base de la segmentation des clients souscrivant des contrats via Mondial-assistance. Nous distinguons

#### a) Les risques à long terme et les risques à court terme

Les risques à long terme ont une plus forte probabilité de survenance. Leur anticipation est fortement influencée par la considération du facteur temps. Les autres facteurs sont minimisés. Le cas des risques maladies et accidents constitue un cas typique. À court terme, il est moins probable que l'assuré subisse un accident mortel ou une maladie soudaine et imprévue parce qu'avant son départ, sa vigilance minimise les facteurs habituels qui produisent ces risques (par exemple apporter des vêtements chauds, reporter son séjour en cas de doute sur l'état de santé). Par contre, au fur et à mesure que le séjour se prolonge, cette vigilance peut être détournée. En effet, même les risques à faible probabilité comme les maladies et les accidents se réalisent et entraînent des coûts financiers importants. La définition des contrats temporaires proposés par mondial-assistance.fr repose sur cette logique de couvrir l'assuré pour une période précise définie.

## **b) Les risques individuels et les risques de groupe**

Les risques individuels sont les risques qui surviennent à une personne qui voyage seule. Tous les facteurs de risque se concentrent autour d'une seule personne, et pourraient donc augmenter les probabilités d'accidents, de maladie ou d'autres types de sinistres. Cette hypothèse, de concentration du risque pourra être validée ou réfutée plus loin dans ce mémoire, lors de l'analyse de nos produits. Néanmoins, ce principe doit être pris en compte lors de la tarification et dans la stratégie marketing de proposition d'une couverture à titre individuel.

Les risques de groupe reposent sur la théorie de la dépendance en termes de probabilité. Autrement dit, les risques d'accident ou de maladies des personnes qui voyagent ensemble ont une corrélation non négligeable parce que les uns veillent sur les autres. La pertinence des facteurs de risque diminue ou augmente pour chaque individu inclus dans le groupe. On peut imaginer que le facteur de risque diminue dans le cas d'un voyage en famille car généralement les parents se montrent plus vigilants envers leurs enfants pendant les activités de vacances et leurs interdisent par exemple de s'aventurer loin du camping ou dans les endroits difficiles. Il y a alors une plus forte probabilité que les sinistres ne se réalisent pas.

En se basant sur cette approche, Mondial Assistance propose deux types de produits selon le groupement des assurés : l'assurance pour une seule personne et l'assurance pour un groupe de personnes. Les caractéristiques de ces deux offres sont similaires à quelques exceptions près.

## **c) Les risques conditionnels**

L'appréhension des risques conditionnels est une technique à prendre en compte dans la tarification en assurance voyage. On constate, que si le risque avant le départ se réalise (annulation), il n'y a pas de risque pendant le voyage. Dans ce cas, l'assureur s'occupe seulement de rembourser le montant de la couverture liée à ce risque sans influence avec la réalisation des risques pendant le voyage.

Si au contraire, le risque pendant le voyage se réalise, la couverture engagée porte seulement sur ce risque sans aucune corrélation avec la couverture du risque avant le voyage. La dissection de ces deux types de risque repose alors sur leur caractère conditionnel. Ils ne se réalisent pas en même temps et si le premier surgit, le second sera nul.

L'exemple du pack temporaire de l'offre assurance voyage d'affaires illustre fidèlement l'appréhension des risques conditionnels. Si les facteurs de risque d'annulation apparaissent avant le départ et le sinistre étant défini, mondial-assistance s'acquitte d'une couverture maximale de 5 000 euros maximum suivant les dispositions mentionnées dans le contrat. Le reste des garanties réservées aux dommages bagages (3 000 euros), aux frais médicaux et hospitalisation d'urgence à l'étranger (150 000 euros), à la responsabilité civile à l'étranger (4 500 000 euros) et à l'individuel accident (15 000 euros) ne seront pas engagés.

## C. Analyse stratégique du secteur

Les travaux antérieurs traitant directement le domaine de l'assurance voyage sont peu nombreux dans les bibliothèques universitaires en ligne. Les travaux déjà effectués par les professionnels et les institutions publiques ou privées sont limités à des explications, sur les offres d'assurance voyage d'une compagnie, les rapports annuels d'activités et les publications officielles. Parmi les institutions qui ont déjà mené des études, la SNSA, la FFSA et l'ORIAS sont les plus connues ainsi que l'ARGUS.

Nous avons pu voir que le secteur de l'assurance voyage est animé par un très grand nombre et types d'acteurs. En général, cette situation entraîne un effet de stabilisation du prix des contrats d'assurance voyage autour d'un prix de marché. En revanche, l'interdépendance et la superposition des activités de ces acteurs transforment le mode de distribution des produits d'assurance voyage en circuit indirect. Les marges se superposent en cascade et les conséquences tombent sur les consommateurs et réduisent la marge de manœuvre des compagnies d'assurance. Une analyse stratégique est nécessaire afin de détecter les menaces/opportunités et les forces/faiblesses du secteur de l'assurance voyage.

### 1. Les menaces

Le nombre de concurrents dans le secteur de l'assurance en général est assez élevé. Pourtant, l'inexistence de statistiques « propres » sur les prestataires d'assurance voyage ne permet pas de mesurer l'influence d'une société sur le secteur ni la nature de la concurrence sur les produits d'assurance voyage.

*Les manques à gagner sont accaparés par les intermédiaires d'assurance ou assimilés :* Comme ils entrent en concurrence avec les compagnies d'assurance voyage, la prospérité de leur activité entraîne des impacts négatifs sur celles des compagnies. Ils perçoivent d'importantes commissions allant jusqu'à 80% de la prime finale. Même si le mécanisme d'intermédiaire est bénéfique pour l'assureur (force de vente, animation commerciale, gestion des polices d'assurance ...), ce taux de commission compromet les marges de manœuvre des compagnies dans la procédure de tarification des offres d'assurance et d'assistance. Dans ce sens, l'assureur se met dans une position critique ou la pression de la concurrence incite à la baisse même si les charges liées à chaque prestation contiennent des consommations incompressibles. Par conséquent, l'effervescence et la stratégie des intermédiaires d'assurances constituent une menace.

Les plafonds de prise en charge des différentes sociétés d'assurance voyage restent assez flous en raison du manque de statistiques disponibles et exhaustives du secteur. Néanmoins, Il y a quelques organismes privés qui établissent des comparaisons de prix et de prestations offertes par quelques compagnies d'assurance de taille et de positionnement jugées similaires. Sont représentés ci-dessous des données non exhaustives qui illustrent les plafonds de prises en charge par garantie :

#### a. Frais médicaux : intervention chirurgicale et hospitalisation

- Globe Partner : 100% des frais réels sans franchise ni délai de carence
- April Magellan : 100% des frais réels
- ISIC : plafond de 150 000 euros
- AVI : pas de plafond

- Europ Assistance : plafond de 300 000 euros plus franchise de 30 euros
- Mondial Assistance : plafond de 300 000 euros plus franchise de 30 euros
- Chapka : 100% des frais réels
- Globe Partner : 3 000 euros
- April Magellan : 5 000 euros / personnes et 15 000 euros/événement
- ISIC : plafond de 12 000 euros
- AVI : 2 000 euros
- Europ Assistance : 15 250 euros
- Mondial Assistance : 1500 euros/personne/sinistre, frais de secours : 1 500 euros/personne/sinistre
- Chapka : 5 000 euros/personnes et 15 000euros/événement

*b. Plafond de prestations d'assurance bagages : pendant le trajet*

- Globe Partner : 1 150 euros
- April Magellan : 1 600 euros avec franchise de 30euros/sinistres, 800 euros pour objets précieux
- ISIC : 230 euros si retard de livraison de 24 heures
- AVI : 2 000 euros ; 300 euros pour objet de valeur
- Europ Assistance : 3 000 euros
- Mondial Assistance : 1 000 euro/personne/sinistre et 3 000euros/personne/sinistre pour le pack premium ;
- Chapka : 2 000 euros, objet précieux 1 000 euros

D'après ces items, on observe que de nombreux paramètres et détails sont en jeu (produit, formule etc...) et compromettent ainsi la comparaison des clients avant de contracter. Néanmoins, il y a une possibilité de comparaison avec les chiffres exacts présentés pour certaines garanties. En prenant par exemple les plafonds en assurance bagages, les chiffres sont plus lisibles et leur représentation simplifie la comparaison. La concurrence en termes de prix reste malgré tout difficile à évaluer dans le domaine de l'assurance voyage.

*Avantages comparatifs des concurrents* : les sociétés du marché de l'assurance voyage sont amenées à concevoir un ou plusieurs produits de distinction par rapport aux autres concurrents. Ce ou ces produits sont les fruits d'une étude marketing qui précède la segmentation de leur client et la personnalisation de leur produit. Pour Mondial Assistance, les offres en termes de prise en charge des frais médicaux sont moins attractives. En effet, plusieurs concurrents proposent une prise en charge sans plafond des frais médicaux contrairement à mondial assistance. Dans ce sens, les potentiels clients pourraient préférer souscrire auprès des autres concurrents qui se montrent plus persuasifs pour certaines de leurs offres.

## 2. Les opportunités

*Développement du marché* : le développement du marché actuel d'assurance tend vers le développement de l'utilisation d'Internet. Les compagnies d'assurance se lancent progressivement dans la conception et la diffusion d'un nouveau type de marketing nommé marketing affinitaire. Il s'agit en fait d'une nouvelle stratégie de segmentation des offres d'assurance en combinant l'approche marketing traditionnelle et la technique de vente en ligne. Une étude de Xerfi en 2011 pour l'argus de l'assurance montre que la croissance de l'assurance affinitaire est de l'ordre de 15% par an. Ce propos est justifié par une étude de FEVAD de 2009 sur les indicateurs de vente de police d'assurance, mentionnant une croissance de 25% sur les visites des sites d'assurance. Le marché de l'assurance affinitaire et les ventes en ligne sont des véritables opportunités à saisir par Mondial Assistance.

*Extension de marché* : La France constitue une part de marché importante du groupe Allianz global assistance. Elle apporte une croissance de 9% sur l'ensemble du marché européen de la compagnie en 2009. On note aussi de fort développement sur des marchés émergents à fort potentiel tels que les marchés chinois et indiens dont le taux avoisine 43%. Le Brésil ou les Etats-Unis possèdent des taux de performance des prestations d'assurance voyage respectivement de 19,3% et 24,7%.

Par ailleurs, le groupe a commencé une stratégie d'extension de marché en 2010 et prévoit une hausse des consommateurs de produits d'assurance voyage en ligne et off line sur ces différents marchés. L'ouverture de ces marchés offre de grande opportunité au groupe pour le lancement de nouveaux produits qui permettront de renforcer la maîtrise du métier (Données, segmentation, tarification..).

## 3. Les forces et les faiblesses

*Expertise marketing* : L'analyse de la conception des produits permet de confirmer la maîtrise du concept marketing pour les responsables de Mondial Assistance. Les produits d'assurance voyage proposés sur le site renferment plusieurs types d'assurance en fonction des exigences anticipés des assurés. Ils sont constitués de plusieurs packs en fonction du besoin et du budget des souscripteurs. Cette capacité d'anticiper les besoins des assurés et de répondre convenablement témoigne de la maîtrise marketing et guidera toutes les stratégies et techniques de conception des nouveaux produits. De surcroît, le groupe a choisi de diversifier ses canaux de distribution.

*Produits et service innovants* : Les services de communication avec les clients ont été nettement améliorés depuis quelques temps. L'avènement des applications avec les Smartphones a permis au groupe de moderniser la communication avec ses clients. Il existe désormais des applications mobiles accessibles à tout moment via les Smartphones :

- en Australie, l'application MyNissan sur iPhone permet de localiser la voiture de l'assuré, chercher des parkings, indiquer les stations d'essence les plus proches, etc... ;
- en France, l'application Travel permet de déclarer et de suivre la gestion des sinistres mais aussi de présenter des informations utiles et pratiques aux assurés ;
- en Royaume-Uni, l'application mini assistance peut fournir quelques services à l'assuré comme la localisation d'un restaurant, les fonctions GPS l'aident à indiquer la direction et permettent de le localiser à tout instant.

Avec ces services, la communication devient de plus en plus efficace et en temps réel. Cela facilite l'intervention des spécialistes au secours de l'assuré. C'est dans ce sens que ces services sont classés services novateurs.

*Procédure de qualité* : Mondial Assistance France a une technique de gestion interne de contrat très au point. La base de données comprend un fichier contrat dans lequel sont expressément décrites toutes les caractéristiques techniques et informatiques nécessaires pour identifier et individualiser un contrat. Les tables contrat sont en relation avec les tables sinistres, lesquelles sont éditées à partir d'un fichier de format unique indiquant les caractéristiques techniques et informatiques de tous les sinistres enregistrés. Cette technique de gestion des contrats facilite l'édition, l'insertion et le contrôle des items caractérisant chaque contrat et permet d'avoir une vision plus globale de la liaison entre contrat-sinistre. Des états et des formulaires croisés peuvent être édités afin d'analyser toutes les facettes de relation existant entre ces deux rubriques. Cette gestion dynamique des bases de données contrats, sinistres constitue une force pour Mondial Assistance.

### **Concernant les faiblesses**

*Produits et services*: la plupart des offres proposées par Mondial Assistance sont semblables aux produits ou aux lignes de produits proposés par les concurrents tels que le rapatriement médical, l'annulation de voyage et la responsabilité civile à l'étranger. Dans ce cas, les nouveaux clients potentiels se montrent indifférents face à plusieurs offres portées à leur connaissance. Ils ont tendance à prioriser la variable de base, le prix pour justifier leur choix parce qu'ils ne sont pas en mesure de dénicher les détails minutieux concourant à la différenciation du produit de Mondial Assistance. C'est une autre raison de stagnation, voire de diminution de la rentabilité des entreprises d'assurance voyage comme Mondial Assistance, et par extension du secteur en général.

## En résumé, une analyse SWOT nous donne :

Forces	Faiblesses
<ul style="list-style-type: none"> <li>- Expertise marketing dans la conception des offres en pack</li> <li>- Disposition de service innovant concernant les applications mobiles</li> <li>- Procédure de qualité concernant la gestion et la conception de la base de données contrat-affilié-sinistre et suppression indirecte des activités de sous-traitances liées à la collecte d'information sur les sinistres</li> </ul>	<ul style="list-style-type: none"> <li>- Produits et services sans point de différenciation perceptible aux égards des clients potentiels empêchant l'extension normale de parts de marché.</li> <li>- Produits et services distribués en très large partie par des tours opérateur et agences de voyage sans spécialisation du domaine de l'assurance.</li> </ul>
Opportunités	Menaces
<ul style="list-style-type: none"> <li>- Développement du marché actuel vers la vente directe via internet (BtoC).</li> <li>- Présence du groupe à l'international : le groupe a largement étendu ses marchés à l'international en souscrivant des nouveaux contrats sur les marchés chinois, allemand, australien et américain ;</li> <li>- Développement démographique et donc du voyage dans sa globalité</li> </ul>	<ul style="list-style-type: none"> <li>- Le nombre de concurrent sur le marché de l'assurance est très important mais la concurrence directe ne peut pas être évaluée par l'inexistence de statistiques propres à l'assurance voyage.</li> <li>- Contexte de récession économique et d'évolution des comportements de voyages (voyages plus courts, moins loin qui nécessitent moins d'assurance. Ex : partir en France plutôt qu'à l'étranger)</li> <li>- La stratégie de tarification des commissions des intermédiaires d'assurance compromet la tarification des offres du groupe parce que le taux de leur commission s'élève jusqu'à 60% et plus.</li> <li>- L'analyse concurrentielle des prix reste floue car les informations concernant les prix des produits proposés ne sont pas délivrées de la même façon et sont donc difficile à comparer.</li> </ul>

## D. Conclusion

Tout au long de cette première étape, nous avons tenté de faire un panorama général du secteur de l'assurance voyage et des risques qu'il présente en nous appuyant sur différentes études publiées et en analysant des exemples de risques couverts.

Les résultats de l'analyse stratégique indiquent que la superposition des activités des acteurs en assurance entraîne la perte de rentabilité des compagnies d'assurance voyage comme global Allianz Assistance.

Cette situation figure parmi les menaces qui pèsent sur l'évolution de l'assurance voyage. Néanmoins, des opportunités ressortent aussi de l'analyse stratégique comme le développement de nouvelles solutions entrant dans le cadre de l'assurance voyage pour exploiter les services internet, combinés avec l'approche marketing. Elle prend en compte les variables traditionnelles pour la segmentation des clients.

Les schémas de distribution de l'assurance voyage nous ont amenés à nous intéresser à des acteurs incontournables : les intermédiaires, par qui la majorité de nos ventes passe.

Nous avons vu que ces intermédiaires représentent deux grandes menaces qui sont : la vente par des non spécialistes de l'assurance mais aussi des taux de commission très élevés.

C'est pourquoi pour la suite de ce mémoire, il nous a semblé intéressant d'exploiter un secteur moins important en termes de vente mais qui tend de plus en plus à se développer : la vente directe en BtoC. Cette méthode présente l'avantage d'éviter l'intermédiation et d'avoir une meilleure maîtrise de son risque.

Afin d'y parvenir, nous analyserons l'ensemble des données disponibles via nos ventes directes sur internet (BtoC), afin de proposer un nouveau modèle de tarification basé sur le modèle linéaire généralisé pour nos formules d'assistance et d'assurance. Ce qui nous permettra également d'avoir une meilleure connaissance de notre sinistralité sur ce secteur et ainsi proposer une conception stratégique plus dynamique en terme d'offre marketing et tarifaire.

# Partie 2 : Mécanisme & méthodologie de tarification

---

## A. Introduction

Nous avons pu étudier tout au long de la première partie, la nécessité d'exploiter tous les leviers potentiels de croissance sur le secteur de l'assurance voyage. Nous avons aussi souligné la complexité du marché de l'assurance voyage notamment à travers la multitude des :

- Circuits de ventes : ventes en BtoC, ventes en BtoBtoC via divers intermédiaires
- Produits proposés suivant les différents profils d'assurés : Vacancier, Homme d'affaires, globe-trotter, étudiants, etc...
- Formules proposées : Assurance, Assistance et Multirisque avec des franchises et des plafonds de prise en charge différents.

Afin d'avoir une lecture simplifiée de notre exposition aux risques et une meilleure connaissance de notre sinistralité, nous nous intéresserons dans cette seconde partie à l'application d'une tarification de notre formule assistance commercialisée en vente directe (BtoC) pour nos différents produits.

L'objet de cette seconde partie portera uniquement sur le calcul de la prime pure appliquée à l'assuré. La valorisation des chargements et de la marge de l'assureur ne sera pas traitée. Néanmoins à titre d'exemple, une petite illustration de l'élaboration de la prime nette incluant les frais de chargement sera effectuée dans la dernière partie de ce mémoire (partie GAM).

Plus en détails, nous verrons : une première section consacrée à l'exploration « subjective » des données disponibles et les différents traitements qui y sont liés afin d'assainir la base face aux différentes données erronées ou aberrantes.

Une deuxième section regroupe les diverses analyses « objectives » et statistiques de notre base. Notamment à travers l'analyse « tri à plat » de quelques variables explicatives.

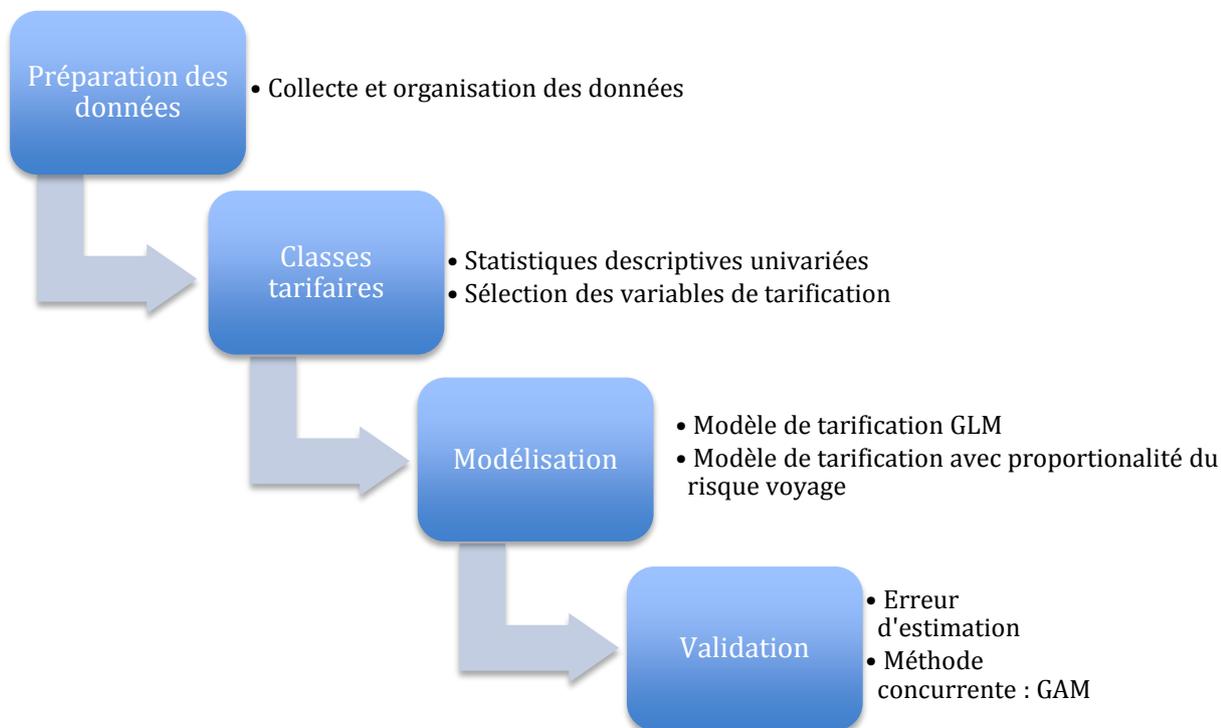
La troisième section expose la modélisation par 2 approches. Une approche en « Fréquence x Coût moyen » classiquement utilisée en tarification et une approche alternative basée sur la prise en compte de la proportionnalité du risque (« Fréquence x Coût moyen x Durée du voyage »). Ces approches, relativement aisées à mettre en œuvre, nous permettront d'avoir une estimation cohérente des risques considérés. Sachant que dans un contexte concurrentiel fort, le tarif proposé se doit d'être le plus adapté possible à la population visée par le contrat. Une segmentation de la population selon plusieurs variables explicatives sera alors réalisée, dans le but de calculer une série de coefficients qui permettront d'adapter la tarification à tout bénéficiaire.

Pour cela, nous utiliserons une méthode de tarification bien connue en assurance non-vie : la théorie des Modèles Linéaires Généralisés.

L'application de cette théorie nous permettra de modéliser la fréquence et le coût moyen des prestations par des lois de probabilité usuelles et ainsi de construire des intervalles de confiance qui fourniront une vision plus précise de la répartition des risques.

Et enfin, dans la dernière étape, nous présenterons différentes validations de notre modèle, notamment à travers une méthode concurrente : les Modèles Additifs Généralisés (GAM). Cette partie nous permettra de proposer une nouvelle tarification avec des résultats que nous analyserons afin de statuer sur la pertinence de son application.

Figure 1: Logigramme de la méthodologie générale



## B. Collecte et organisation des données

La première étape porte sur la collecte des données indispensables à la tarification. L'objectif est de constituer une base fiable sur l'activité constatée via nos ventes en BtoC. Nos fournisseurs de données ont été le Marketing, le Contrôle de gestion et l'Informatique. L'une des premières difficultés de cette étape a été de consolider ces données issues de plusieurs sources. En effet, la configuration informatique actuelle, ne nous permettait pas d'extraire les données sinistres assurance et assistance via une source unique.

Nous verrons lors de cette partie, qu'il a fallu faire des choix de prise en compte des pratiques opérationnelles (en collaboration avec le service souscription), sur le périmètre à étudier, afin que celui-ci soit le plus fiable possible.

Nous nous intéressons donc à déterminer la prime pure sur des contrats d'assurance et d'assistance voyage. De cette façon, nous allons sélectionner les contrats ayant fait l'objet ou non de sinistres.

La fenêtre de notre étude s'ouvre entre le 01 janvier 2010 et le 30 avril 2015. Ce choix de période s'explique par une contrainte opérationnelle. En effet, la base à notre disposition correspond à un flux déjà pré établi entre la Direction marketing et la Direction des systèmes d'information qui n'était pas en capacité d'extraire un nouveau flux avec un historique plus conséquent et fiable pour notre étude.

Pour l'année 2010, il nous faudra tenir compte de la refonte complète de nos produits (suppression de produits, amélioration d'autres) qui nous amène à analyser différemment cette performance

contrairement aux autres années où les données sont à périmètre stable et égal. Pour 2015, seul le premier trimestre est pris en compte.

La tarification impose de bien connaître la population à couvrir. La phase de constitution des données est donc déterminante pour le tarif. Une part très importante du temps consacré à l'étude a été réservée à l'analyse et au nettoyage des bases fournies par les directions métiers.

Notre base se décompose en 2 fichiers :

- Un fichier « contrats », contenant toutes nos expositions : il contient les éléments permettant de comprendre les caractéristiques qui impactent le tarif futur. Ce fichier contient les informations relatives à l'ensemble des assurés.
- Et un fichier « sinistres », où sont répertoriés tous les sinistres : cette base renseigne les informations sur les sinistres survenus de 2006 à 2015.

Le portefeuille de Mondiale Assistance se divise en 3 catégories de formule : la formule « Assurance », la formule « Assistance » et la formule « Multirisques ». Notre étude portera uniquement sur la formule Assistance car le rattachement des sinistres à la base des polices se fait de manière aisée contrairement aux sinistres assurances qui peuvent provenir de différents outils internes en cours d'analyse.

## 1. Description de la base

### 1.1 Fichier « Contrats »

Le fichier contrats reçu initialement par la Direction des systèmes d'information, contenait 140 000 lignes. Etant donné la prise en compte uniquement de notre formule assistance et de nos contrats temporaires pour la suite de ce mémoire, notre base sera de 60 313 lignes. Nous prenons en considération les contrats en cours mais aussi les contrats qui ont été clôturés (expiration de la période du risque). Ceci nous permet de travailler sur une base d'étude suffisamment conséquente.

Au début de l'étude et après un travail d'homogénéisation des différents fichiers de données, 61 variables sont disponibles. 32 d'entre elles ont été retenues, en fonction du taux de remplissage et de leur incidence éventuelle dans le tarif. Le tableau ci-dessous présente l'ensemble des informations disponibles pour réaliser la tarification. La première colonne du tableau renseigne sur le nom de la variable tel qu'il apparaît dans les outils de gestion. La deuxième colonne donne la signification de la variable. Un premier indicateur sur la qualité de la donnée est donné dans la troisième colonne. Il s'agit du taux de remplissage qui évalue le pourcentage de renseignement de la donnée.

Figure 2: Liste des variables en entrée du fichier Contrat

Nom	Signification	Taux de remplissage
code_formule	Code formule	100%
code_pdt	Code produit	100%
Date création	Date création du contrat	100%
Date départ	Date de départ (en voyage)	100%
Date souscription	Date de la souscription	100%
Durée	Durée de l'exposition	100%
etat_contrat	Etat du contrat	100%
Frequence	Type contrat (ann/temp)	100%
lib_formule	Libellé formule	100%
lib_pdt	Libellé produit	100%
Mnt prime	Montant prime (contrat)	100%
Mnt voyage	Montant voyage	100%
Motif_voyage	Motif du voyage	100%
nb_assure	Nombre d'assurés	100%
nb_choses_assure	Nombre de choses assurées	100%
Num contrat	Numéro de contrat	100%
Type mvt	Type mvt (crea, modif,...)	100%
Nom sous	Nom souscripteur	99%
lib_option	Libellé option	94%
moyen_paiement	Moyen de paiement	94%
Cp	Code postal	86%
Prénom sous	Prénom souscripteur	86%
Ville	Ville	86%
Date retour	Date de retour (du voyage)	83%
pays_dest_lib	Lib pays destination	81%
pays_ori_lib	Lib pays origine	67%
Pays lib	Libellé pays	58%
code_sous	Code client souscripteur	24%
Date anniv	Date anniversaire contrat (annuels)	24%
Date effet	Date d'effet du contrat	24%
type_voyage	Type de voyage	20%
Age	Age du souscripteur	16%

#### Problème de qualité des données du fichier « Contrats »

Au-delà de l'aspect taux de remplissage, nous avons été confrontés à des problèmes de qualité des données plus subtils. En effet, nous avons pu constater des problèmes liés :

- à l'affectation du pays de destination : certaines confusions ont pu être observées dans la destination pays, entre les pays du Maghreb et les Emirats Arabes Unis. Dans ce cas, une modalité les regroupant a été créée appelée Maghreb ;
- au remplissage de certains champs par des valeurs nulles, ce qui répertorie ce champ comme renseigné dans nos bases. Nous rencontrerons ce problème notamment sur les champs concernant les montants de voyage. Afin de corriger ce biais pour cette variable, il a été décidé d'affecter un montant forfaitaire en fonction des destinations suivant le Tableau 2 validé par le service souscription ci-dessous.

Tableau 2 : Montants de voyage forfaitaires attribués à des zones prédéfinies

Destination	Montant Forfaitaire (en €)
Amerique du nord	850
Asie	900
Afrique	950
Europe hors France	500

Nous déciderons tout de même de suivre avec beaucoup d'attention cette variable dans toute la suite de notre étude.

## 1.2 Le fichier « Sinistres »

Il répertorie les informations relatives à tous les sinistres liés aux voyages d'un assuré. Chaque sinistre est décomposé en garanties sinistrées, d'où le fait qu'un sinistre présente autant de lignes dans la base sinistre qu'il a de garanties sinistrées.

Figure 3: Liste des variables en entrée du fichier Sinistre

Nom	Signification	Taux de remplissage
Branche code	Code de la branche	100%
Branche libellé	Libellé de la branche	100%
Date ouverture	Date ouverture dossier	100%
Date survenance	Date survenance dossier	100%
Etat Dossier	Etat du dossier	100%
Num contrat	Numéro de contrat	100%
Num dossier	Numéro de dossier	100%
Origine code	Code origine sinistre	100%
Origine Libellé	Libellé origine sinistre	100%
Type dossier	Type de dossier	100%
Date maj	Date de. Mise à jour dossier	61%
Objet code	Code objet sinistre	17%
Objet libellé	Libellé objet sinistre	17%

Ce fichier étant globalement bien renseigné, nous conserverons les 10 premières variables, pour la suite de notre étude.

## 2. Constitution de la base de travail

Elle porte sur le croisement entre les bases de données de sinistres et celle des contrats décrites plus haut. La base résultante nous permet donc de rattacher à chaque ligne de contrat, un ou plusieurs sinistres.

La jointure entre la base sinistre et la base portefeuille se fait grâce au numéro de contrat (*num\_contrat*).

La constitution de la base de travail a demandé une analyse des données aberrantes et des données manquantes. En fonction de la qualité de l'information fournie, le choix de garder la variable ou non sera fait.

### Suppression et correction des données aberrantes

Pour s'assurer d'une plus grande stabilité et fiabilité du modèle, quelques opérations sur cette nouvelle base sont nécessaires :

- Les sinistres conservés sont les sinistres dont le montant est supérieur ou égal à 0 ;
- La survenance des sinistres importants (souvent appelés « sinistres graves ») perturbe beaucoup la construction d'un tarif. En effet, le montant très élevé de tels sinistres augmente considérablement la « charge sinistres » et pénalise donc tous les segments où se trouvent ces polices sinistrées. Des techniques d'écrêtement des sinistres atypiques existent mais ne feront pas l'objet de ce mémoire. Le seuil entre sinistres « graves » et sinistres « attritionnels » est fixé à 60 000 euros, en accord avec la souscription. Statistiquement, cela correspond à un quantile à 99.5% de la distribution du montant des sinistres.
- Les contrats ayant un montant de prime inférieur à 0 ont également été exclus de l'étude.

### Les valeurs manquantes

Nous allons privilégier la suppression des variables pour lesquels il y a plus de 20% des données manquantes. Nous constaterons que les variables ainsi supprimés concernent des variables importantes avec probablement un fort pouvoir discriminant. Nous l'observons notamment pour le pays d'origine et l'âge de l'assuré.

### Revalorisation des sinistres

Pour déterminer la sinistralité future, il nous a fallu nous appuyer sur l'historique de notre portefeuille. Afin d'être juste sur cette revalorisation, nous devons considérer que toutes les conditions d'une année passée sont identiques aux conditions de l'année étudiée.

Etant donné qu'un sinistre entraîne un certain coût, il nous a fallu déterminer le coût de ce même sinistre s'il était survenu l'année étudiée. Nous parlons de ce cas de « mis as if » des sinistres.

Plusieurs méthodes étant envisageables pour revaloriser les sinistres (méthodes prospectives, rétrospectives), nous avons privilégié les estimations d'inflation transmises par le service de gestion des sinistres. Celle-ci s'établit en moyenne à 5%.

En pratique, nous multiplions les sinistres de 2014 par 1,05, ceux de 2013 par 1,05<sup>2</sup> et ainsi de suite.

### 3. Création des variables de sinistralité

Ce sont les variables de sinistralité qui vont en l'occurrence expliquer le risque soit par la probabilité de survenance d'un sinistre, soit par son montant. Il est important de bien distinguer les 2 phénomènes :

- la fréquence de survenance des sinistres ;
- l'intensité du sinistre, une fois qu'il est survenu.

Tout au long de ce mémoire les deux phénomènes seront donc étudiés en parallèle et simultanément. Nous supposons également une indépendance entre les deux phénomènes.

Les 2 phénomènes sont observables au niveau police et estimable après une agrégation des polices par catégorie de risque.

Dans le cadre de l'analyse exploratoire, nous avons utilisé 3 grandeurs clés. Les 3 mesures ci-dessous permettent ainsi de calibrer la fréquence et le coût moyen par catégorie de risque.

#### 3.1 La population soumise au risque (ou population sous risque)

##### Définition :

Nous définissons la population soumise au risque assistance (notée  $n$ ) comme étant la somme des coefficients de durée d'exposition de l'ensemble des polices couvertes par la garantie assistance.

Le coefficient de durée d'exposition d'une police  $i$  est le nombre de jours pendant lesquels elle est couverte (période d'exposition notée  $Duree_i$ ) sur une période considérée (par exemple, une année).

$$n = \sum_i \frac{Duree_i}{365}$$

##### Exemple :

Soit une police couvrant 2 personnes :

La première personne est couverte du 1er janvier 2014 jusqu'au 31 juin 2014, soit un coefficient de durée d'exposition de 1/2.

La seconde personne a une couverture qui court du 1er janvier 2005 jusqu'au 31 décembre 2005, soit un coefficient de durée d'exposition de 1.

La population soumise au risque de ce groupe est donc de 1,5 personne (au lieu de 2).

#### 3.2 La fréquence de sinistre par catégorie de risque

A la différence de produits d'assurance non-vie traditionnelle, une fois qu'une police a connu un sinistre, celle-ci est immédiatement considérée éteinte. Par conséquent, tous les éléments relatifs à la

quantification de la fréquence de sinistralité s'appuieront sur l'observation ou non d'un événement de sinistre (et non pas au nombre de sinistres).

Pour  $k$  une catégorie de risque (pouvant être assimilée à la modalité d'une variable catégorielle) et  $i$  un contrat donné, la fréquence s'exprime par la formule suivante :

$$Frequence_k = \frac{\sum_i top\_sinistre\_assistance_i * \frac{Duree_i}{365} * \mathbb{I}_{\{i \in k\}}}{n_k}$$

Où :

- $top\_sinistre\_assistance_i$  est l'indicatrice de l'événement de sinistre pour le contrat  $i$  ;
- $\mathbb{I}_{\{i \in k\}}$  est égale à 1 lorsque le contrat  $i$  appartient à la catégorie de risque  $k$  et 0 sinon ;
- Et  $n_k = \sum_i \frac{Duree_i * \mathbb{I}_{\{i \in k\}}}{365}$  représente la population sous risque dans la catégorie de risque  $k$ .

#### Exemple :

Considérons deux polices appartenant à la catégorie de risque dont le motif de leur déplacement est « voyage d'affaires » et dont le nombre de bénéficiaires est strictement égal à 1. Supposons que pour la :

- police 1 :  $top\_sinistre\_assistance_1 = 0$ ,  $Duree_1 = 1$  jour ;  $\mathbb{I}_{\{1 \in \text{« voyage d'affaires » \& nb=1\}} = 1$
- et police 2 :  $top\_sinistre\_assistance_2 = 1$ ,  $Duree_2 = 4$  jours et  $\mathbb{I}_{\{2 \in \text{« voyage d'affaires » \& nb=1\}} = 1$ .

La fréquence la catégorie de risque dont le motif de leur déplacement est « voyage d'affaires » est

$$\text{donc : } Frequence_{\text{« voyage d'affaires » \& nb=1}} = \frac{0 * \frac{1}{365} * 1 + 1 * \frac{4}{365} * 1}{\frac{1}{365} + \frac{4}{365}} = 80\%.$$

### 3.3 Le coût moyen par catégorie de risque

Pour  $k$  une catégorie de risque (pouvant être assimilée à la modalité d'une variable catégorielle) et  $i$  un contrat donné, nous définissons le coût moyen comme le coût total dans la catégorie rapporté au nombre de sinistres total dans la catégorie :

$$intensité_k = \frac{\sum_i cout_i * \mathbb{I}_{\{i \in k\}}}{\sum_i \mathbb{I}_{\{i \in k\}}}$$

Où :

- $cout_i$  est l'indemnité reçue par le contrat  $i$  ;
- $\mathbb{I}_{\{i \in k\}}$  est égale à 1 lorsque le contrat  $i$  appartient à la modalité  $k$  et 0 sinon.

## 4. Analyse de la base constituée

Nous présentons ici l'évolution du portefeuille entre 2010 et 2015, après validation suite à un écart inférieur à 5% des montants sinistres et primes agrégées entre notre base constituée et les chiffres

provenant de *reporting* interne. Ces écarts s'expliquant notamment par des différences de date d'extraction et de méthodologie d'identification des polices d'assistance qui sont en cours de discussion en interne.

**Tableau 3 : Evolution de la fréquence et du coût moyen selon les années**

Années	Fréquence	Coût moyen	Coût moyen as-if*	Prime pure	Nombre de contrats	Primes acquises	Exposition (en jours)	Montant sinistres as-if*
2010	2,38%	3 878	5 045	92.3	6 214	608 865	188 302	363 018
2011	2,01%	2 495	3 119	50.1	10 330	1 001 835	303 622	321 584
2012	2,25%	2 774	3 274	62.4	11 863	1 093 019	355 887	365 208
2013	1.78%	2 956	3 311	52.6	9 065	845 087	263 701	223 386
2014	2.02%	3 054	3 234	63.2	12 700	1 231 684	356 080	403 990
2015	0.84%	1 956	1 956	16.4	4 051	396 928	117 843	96 025

\*après correction de l'inflation

L'analyse du tableau ci-dessus ne peut se faire sans prendre en considération, le contexte économique global externe dans lequel les compagnies d'assurance voyage évoluent maintenant depuis quelques années, à savoir pour rappel :

- La crise économique globale qui touche les potentiels assurés ;
- Les conflits internationaux, qui réduisent l'activité touristique et économique des pays concernés. On peut par exemple citer le Printemps arabe (2012/2013) qui entraîna une baisse brutale des réservations de séjour pour des pays comme la Tunisie ou l'Egypte.

Nous constatons qu'hormis 2010 (refonte des produits vendus), nos coûts moyens ne cessent d'augmenter comme cité dans la première partie de ce mémoire à travers l'analyse du marché de l'assurance voyage par la SNSA.

Dans le même temps, on remarque une augmentation de la volatilité de la fréquence entre les années mais qui reste au global stable autour des 2%.

La fréquence et le coût moyen pour 2015 sont difficilement interprétables dans la limite où nous ne prenons pas en compte la période estivale de juin à septembre (prise en compte uniquement des trois premiers mois de 2015).

Le Tableau 3 résume donc parfaitement la difficulté à laquelle sont confrontées les compagnies d'assurance à savoir une hausse des coûts moyens combinée à une volatilité de la fréquence sinistre. D'où la difficulté pour les compagnies d'assistance de préserver une certaine rentabilité sans revoir leur tarification.

L'analyse de la prime pure doit être faite en tenant compte comme vu auparavant dans la première partie (section 1), que les compagnies d'assistance délivrent des prestations d'assistance qui peuvent représenter dans certains cas uniquement des coûts internes.

Cela pose le problème de l'interprétation de la définition classique du S/P en assurance, qui lui, analyse les coûts externes par rapport à la prime. L'intégration des coûts internes sur une « adaptation du S/P pour l'assistance », aurait été plus pertinente mais celle-ci reste sensible, étant donné les discussions engagées actuellement sur les coûts internes au sein de notre compagnie.

De façon plus générale, nous effectuerons la suite des analyses de ce mémoire uniquement sur la prime pure dans un premier temps. Et dans un deuxième temps nous intégrerons les frais divers internes afin de pouvoir proposer une prime commerciale qui nous permettra de comparer de manière aisée les tarifs proposés.

## C. Statistiques descriptives univariées

Les variables porteuses d'une information sont appelées variables explicatives. De cette manière, une fois les données clairement déterminées, il convient d'effectuer des travaux d'analyses préliminaires des dites variables avant même de commencer la modélisation. Malgré un nettoyage préalable des données, il est possible que certaines variables ne soient pas pertinentes à la segmentation et à la tarification. Une première analyse descriptive permet de les repérer. On s'intéressera plus particulièrement aux variables prédéfinies par les directions métier, soit afin de confirmer une intuition, soit dans le but d'en obtenir une.

En ce sens, les enjeux de l'analyse descriptive sont alors multiples. Elle permet tout d'abord de déterminer les caractéristiques d'un individu moyen et ensuite de vérifier la pertinence des variables utilisées opérationnellement.

En pratique, nous allons commencer par travailler sur les variables présélectionnées et couramment utilisées dans l'assurance voyage.

### 1. Transformation des variables

L'objectif de cette partie est de décrire la méthodologie employée pour transformer les variables potentiellement intégrables dans la tarification.

Les variables explicatives à notre disposition sont de nature continue, qualitatives ordinales ou qualitatives nominales.

Or, nous souhaitons que le tarif soit appliqué pour toute une catégorie de risques, c'est-à-dire que nous voudrions établir des cases tarifaires dans lesquelles chaque nouvelle souscription (ou nouveau remplacement) pourra être inscrite.

Nous avons donc transformé les variables potentiellement explicatives en des tranches de valeurs de telle sorte qu'elles soient à la fois suffisamment peuplées et suffisamment discriminantes. Ainsi, ces tranches contiennent chacune au minimum 0.1% d'exposition. Un seuil au-dessus engendrerait une inhomogénéité du segment, d'où une pertinence limitée dans l'analyse technique des résultats. A l'inverse, un seuil inférieur entraînerait une instabilité (donc grande incertitude) des résultats observés.

De plus, cette étape est primordiale pour la suite de l'étude. D'un point de vue statistique, on se concentrera sur deux éléments : le choix de nombres de classes et la méthodologie de discrétisation des variables continues.

Il est important de noter que la discrétisation ne s'appuie pas uniquement sur une vision statistique. Elle prend également en considération les avis des experts dans chaque domaine d'activité.

#### 1.1 Etude de la fréquence

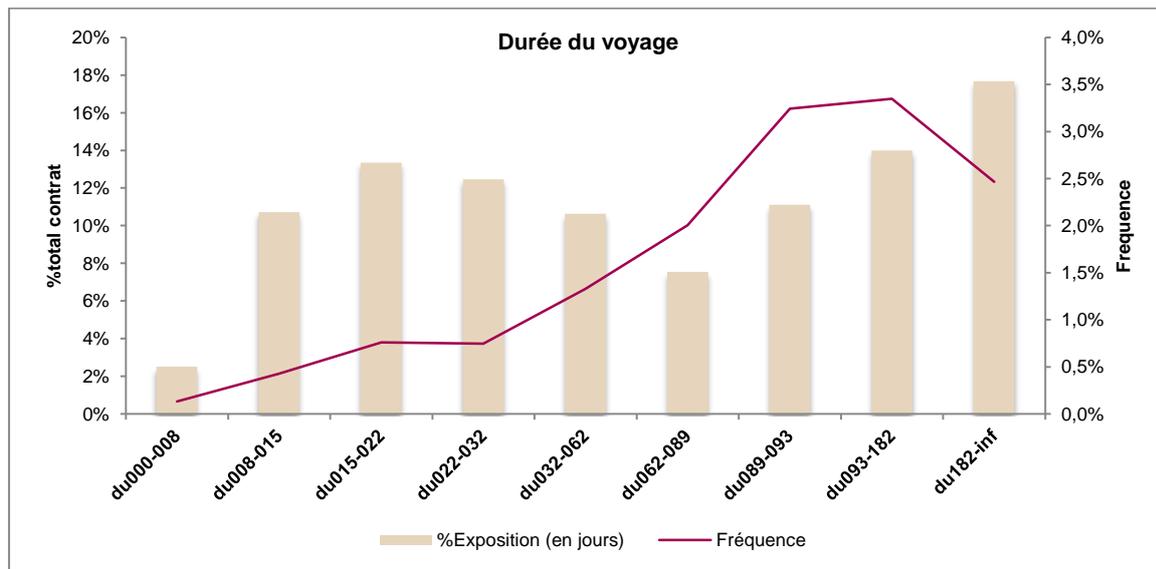
Nous traiterons dans cette sous-partie les 5 variables les plus caractéristiques du risque d'après le service souscription définies par la variable durée du voyage, destination du voyage, nombre d'assurés, montant du voyage et le motif du voyage.

### La variable « durée du voyage »

Pour l'ensemble de nos contrats temporaires, l'assuré doit renseigner ces dates de départ et de retour.

D'où le découpage ci-dessous de la durée de voyage des souscripteurs, effectué à l'aide de la méthode des Quantiles. Celle-ci a été transformée entre tranches de 9 modalités.

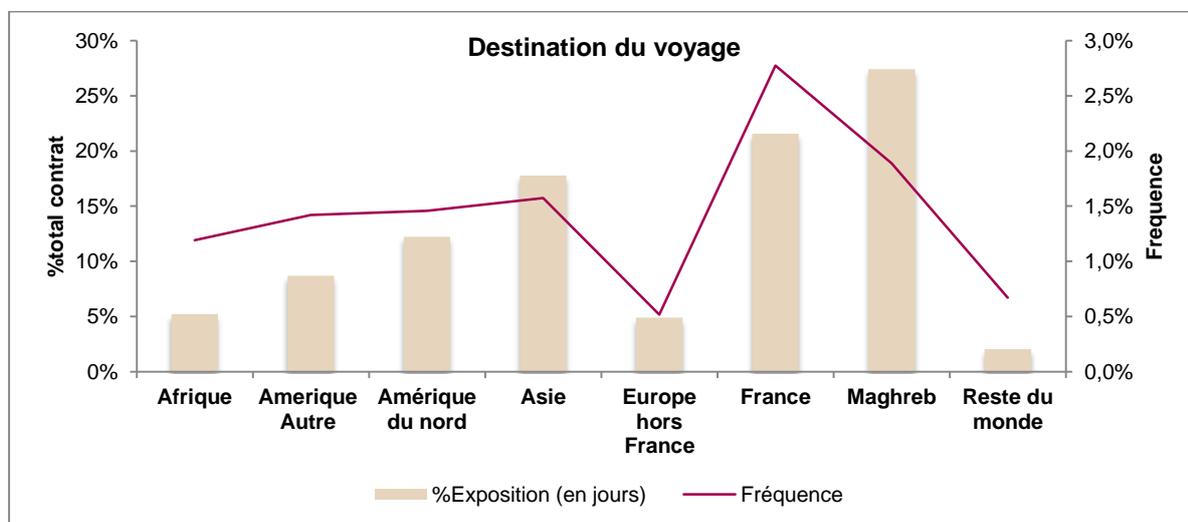
**Figure 4: Fréquence de sinistres en fonction de la durée de voyage**



L'analyse de ce graphique nous renseigne sur l'exposition au risque de sinistre suivant la durée du voyage. Nous constatons, comme nous aurions pu le prévoir, que le risque d'avoir un sinistre assistance croît avec la durée de l'exposition. En effet, nous constatons qu'au fur et à mesure du prolongement du séjour, la vigilance diminue et accroît donc le risque d'incident.

### Variable « Destination géographique »

**Figure 5: Fréquence de sinistres en fonction de la destination du voyage**



Le graphique ci-dessus montre que la majorité de nos contrats sont répartis sur quelques zones géographiques précises. En effet, nous constatons que le Maghreb, la France, l'Amérique du nord et l'Asie représentent une part très significative des contrats souscrits. Les 2 points suivants permettent en partie de justifier cette répartition :

- Certains pays comme l'Algérie (Maghreb) impose d'avoir une couverture d'assurance voyage pour fournir un visa aux personnes pénétrant sur le territoire. Ceci rend obligatoire pour les assurés désirant se rendre dans ces pays de souscrire un produit voyage.
- Le coût important qu'un accident peut engendrer dans certains pays ne présentant pas une couverture santé optimale comme nous pouvons avoir en France. Le prix de l'accès aux soins est également un levier de souscription, c'est le cas des Etats-Unis ou un accident/une maladie, peut très rapidement entrainer des coûts élevés. C'est pour cela que les personnes souhaitant s'y rendre préféreront souscrire une couverture d'assurance voyage optimale.

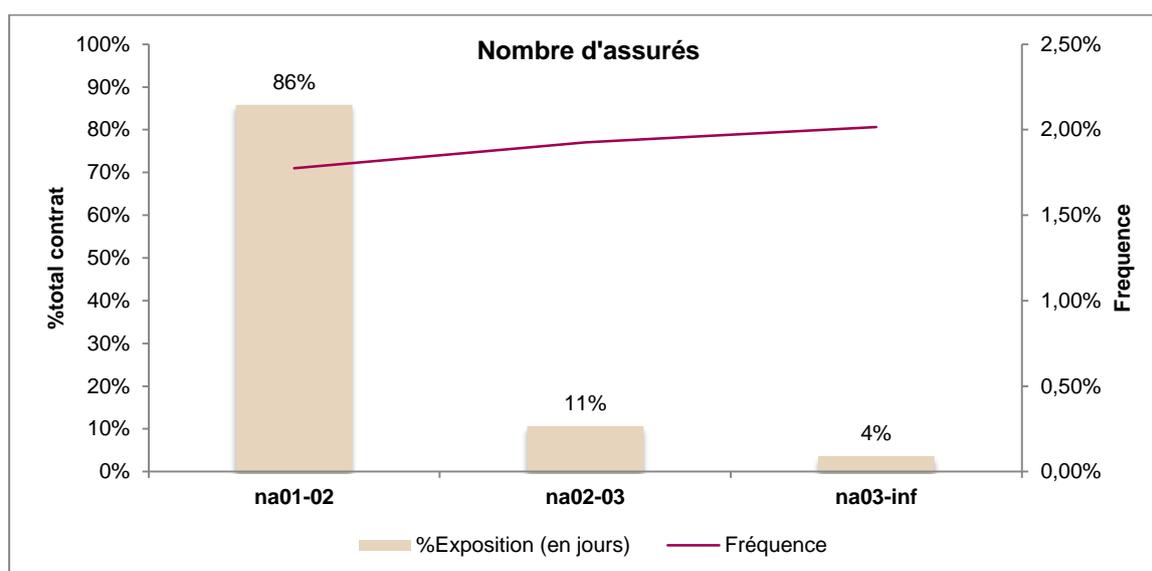
### Variable « nombre d'assurés »

Ce découpage correspond au nombre d'assurés par contrat souscrit. En effet, nous avons vu dans la première partie que certains produits pouvaient couvrir plusieurs bénéficiaires, notamment sur nos produits Assistance.

Nous avons fait le choix de découper cette variable en 3 modalités :

- **na01-02** : cette modalité désigne les polices dont le nombre de bénéficiaires est inclus dans l'intervalle [1 ; 2[ ;
- **na02-03** : cette modalité désigne les polices dont le nombre de bénéficiaires est inclus dans l'intervalle [2 ; 3[ ;
- **na03-inf** : cette modalité désigne les polices dont le nombre de bénéficiaires est inclus dans l'intervalle [3 ; infini [.

**Figure 6: Fréquence de sinistres en fonction du nombre d'assurés**



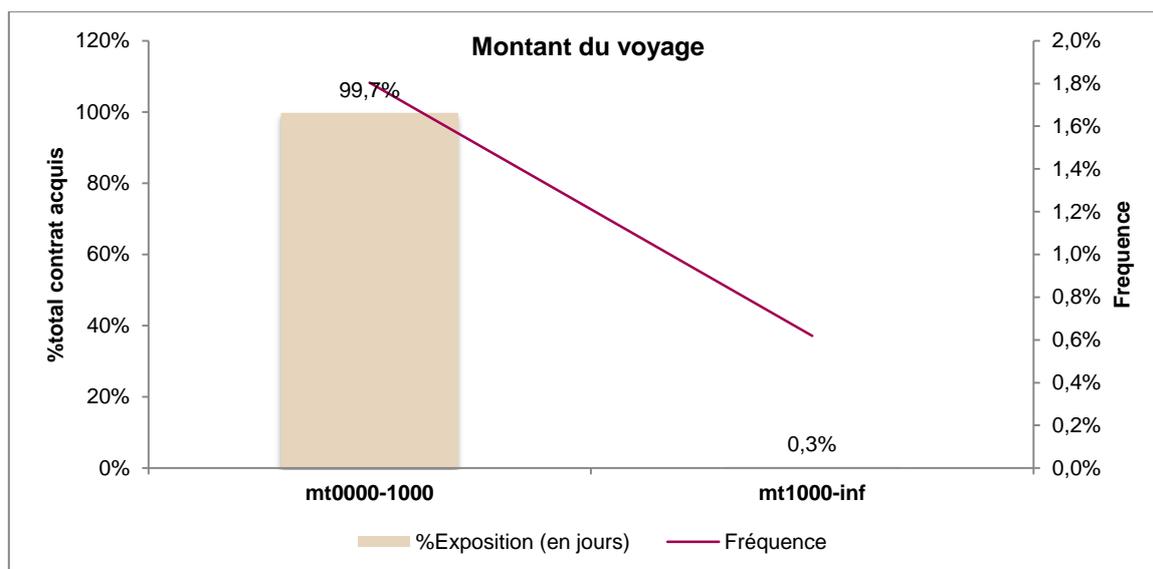
Dans ce graphique, nous pouvons remarquer que plus de 80 % de nos contrats, concerne qu'un ou deux bénéficiaires. Nous observons également que la fréquence augmente sensiblement suivant le nombre d'assurés par contrat. Ce graphe nous permet aussi de réfuter dans une moindre mesure, l'hypothèse de concentration du risque individuel (cf. : première partie).

### Variable « montant du voyage »

Celui-ci correspond au montant du voyage souscrit. En effet, cette variable étant renseignée à 100% (modulo les retraitements présentés dans le Tableau 2), elle est surtout nécessaire pour les garanties d'assurance en cas d'annulation ou d'interruption de séjours. Elle se décompose en 2 modalités comme on l'observe ci-dessous :

- **mt0000-1000** : cette modalité désigne les polices dont le montant du voyage est inclus dans l'intervalle [0 ; 1000[. Les montants en dessous d'un certain seuil ne font pas l'objet d'un suivi systématique. Ces montants de voyage non renseignés sont également intégrés dans cette modalité ;
- **mt1000-inf** : cette modalité désigne les polices dont le montant du voyage est inclus dans l'intervalle [1000 ; infini [.

**Figure 7: Fréquence de sinistres en fonction du montant du voyage**

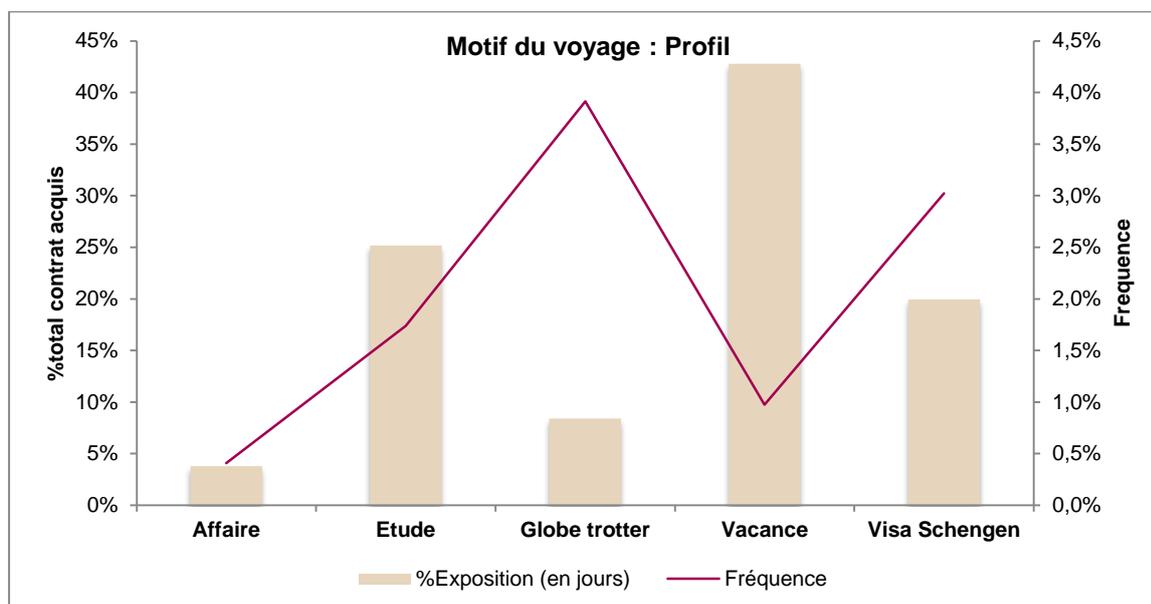


Ce graphique nous indique que le montant par contrat de l'ensemble de notre portefeuille est compris entre 0 et 1000€. Avec une exposition très faible sur notre portefeuille de l'ordre de 0.3%, pour les autres tranches. L'incertitude (cf. Tableau 2), liée à l'affectation des montants, peut fausser la répartition moyenne des contrats souscrits par montant. Nous prendrons malgré cela la décision de tenir compte de cette variable pour la suite de notre étude. En contrepartie, nous y accorderons une attention particulière au moment du choix des variables tarifaires.

### Variable « motif du voyage »

Cette variable nous renseigne sur le type de voyage. En effet, pour la suite de cette étude nous déduirons à travers le type de voyage, un profil de voyageur. Par exemple un assuré qui souscrit un contrat solution vacance sera assimilé à une personne qui voyage pour des vacances. Nous aurons suivant ce principe des « vacanciers », « étudiants », « globe-trotter », « hommes d'affaires » et « voyageurs dans l'espace Schengen ». Cette variable se décompose en 5 modalités comme on l'observe ci-dessous.

**Figure 8: Fréquence des sinistres en fonction du motif du voyage**



Les déplacements à titre professionnel « Affaire » représentent 4% de notre exposition (en nombre de jours) de risque assistance à l'étranger et la fréquence de sinistres quant à elle est de l'ordre de 2%. Les déplacements à titre privé avec pour objectif des vacances ou des études, représentent la majorité de notre portefeuille. Concernant la fréquence, nous observons clairement aussi que nous avons une sinistralité très élevée pour les « Globe trotter » et les voyageurs de l'espace Schengen en comparaison avec les autres profils.

L'analyse de ce graphe nous laisse penser que cette variable est discriminante pour la fréquence. Nous confirmerons par la suite notre intuition.

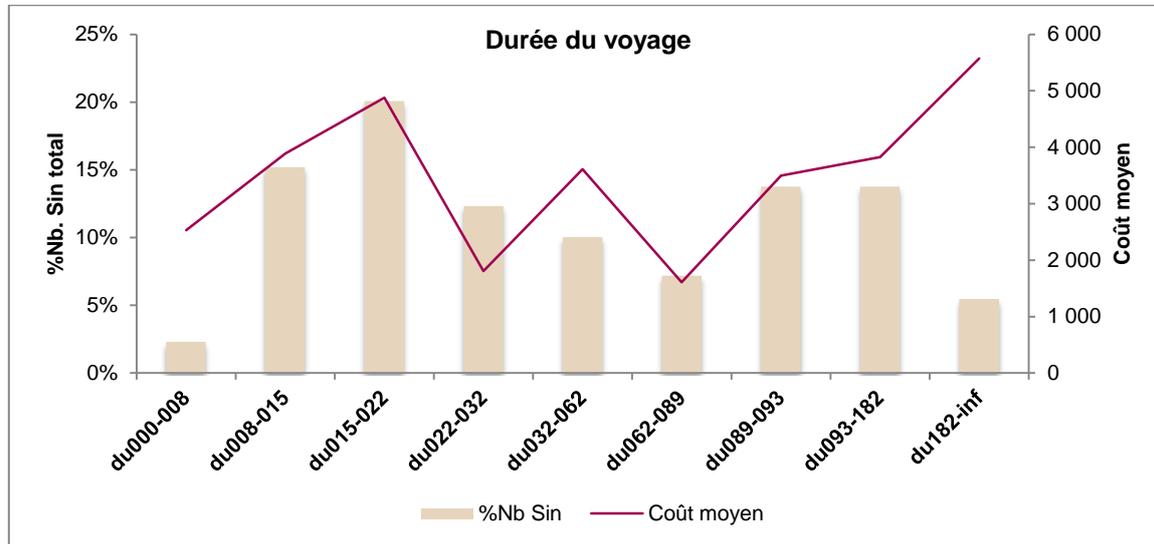
## 1.2 Etude sur le coût moyen

Dans ce cas, nous répèterons l'analyse effectuée pour la fréquence, pour le coût moyen en conservant les mêmes découpages afin d'éviter des complications de méthodes lors de l'élaboration de la tarification finale.

### Variable « durée du voyage »

Nous pouvons grâce au schéma ci-dessus regrouper cette variable en 9 modalités,

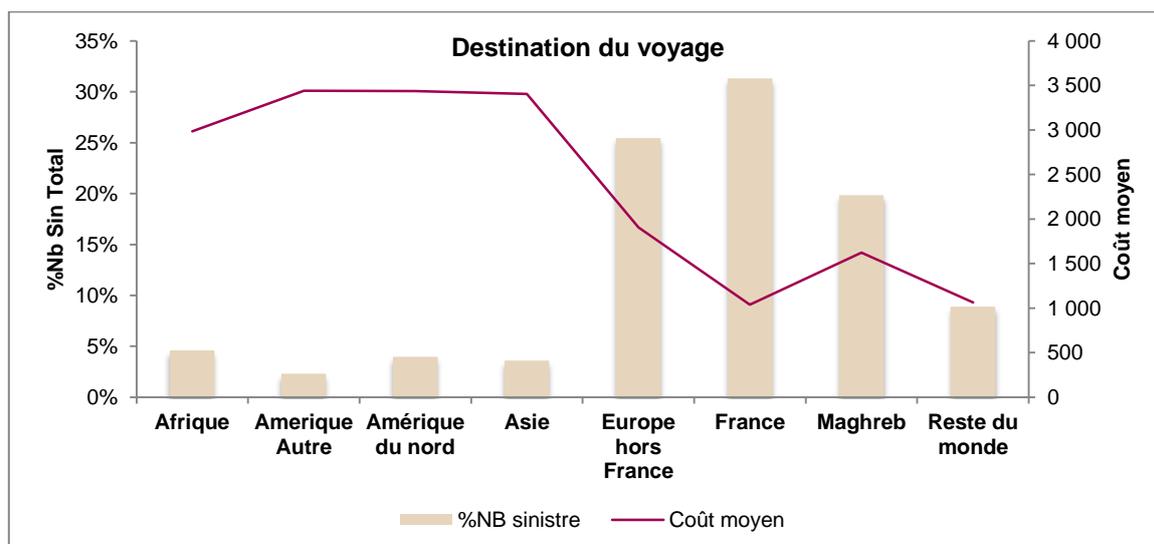
**Figure 9: Coût moyen en fonction de la durée du voyage**



Nous pouvons dire qu'il y a une dépendance faible entre le coût moyen et la durée du voyage. En effet, nous ne constatons pas de véritable tendance entre le coût du voyage et la durée du voyage dans la limite où nous observons des coûts moyens très volatiles suivant la durée du voyage. La variable « durée du voyage », semble ne pas apporter d'informations sur nos coûts moyens.

### Variable « destination géographique »

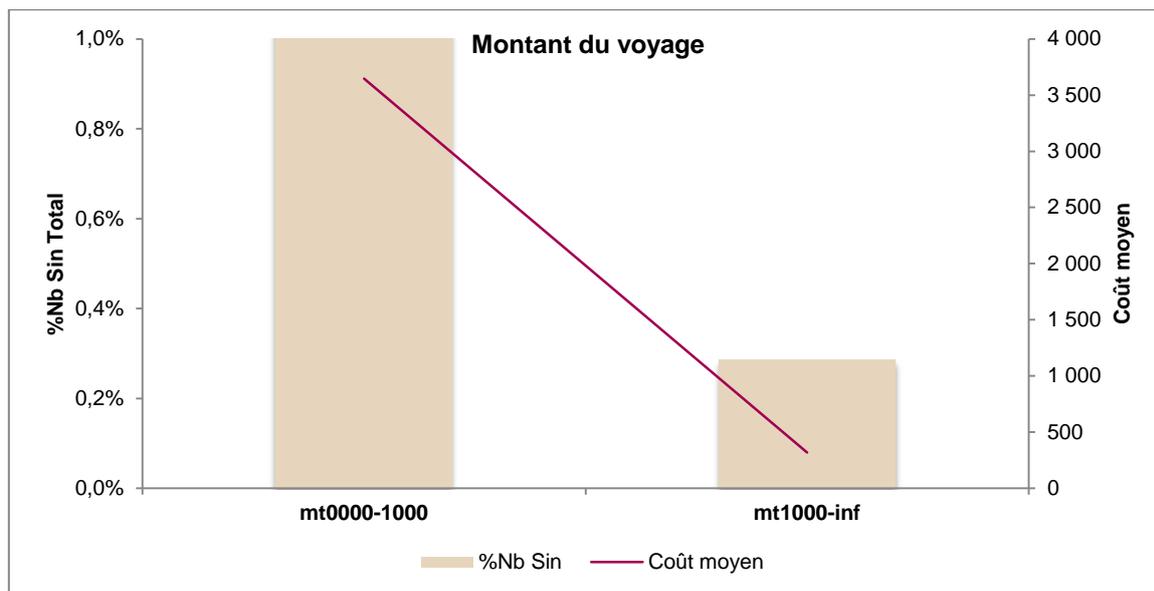
**Figure 10: Coût moyen en fonction de la destination du voyage**



A l'observation du graphique ci-dessus, nous constatons que l'Afrique, l'Amérique et l'Asie sont les zones géographiques ayant l'intensité de sinistre la plus élevée. Nous pouvons expliquer ces coûts élevés par l'absence de moyens techniques (transport, logistique, frais de santé à l'intérieur du pays) permettant une prise en charge moins onéreuse et optimale des assurés. Nous constatons également que le Maghreb présente un coût moyen assez élevé qui est en partie justifié par notre décision de regrouper le Maghreb avec les pays du Moyen-Orient.

### La variable « montant du voyage »

Figure 11: Coût moyen en fonction du montant du voyage

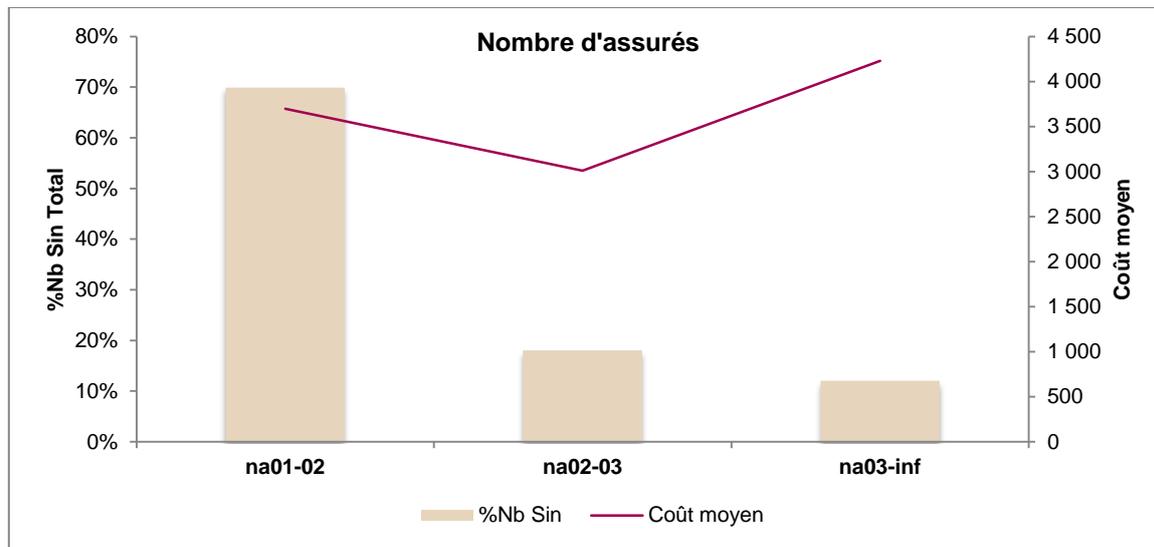


Ce graphe nous indique une intensité très élevée pour nos montants de voyage inférieur à 1000 euros. L'intensité des montants de voyage supérieur à 1000 euros est moins élevée mais sur une exposition marginale de notre portefeuille.

### La variable « nombre d'assurés »

Le nombre d'assurés par contrat souscrit comprend 3 modalités :

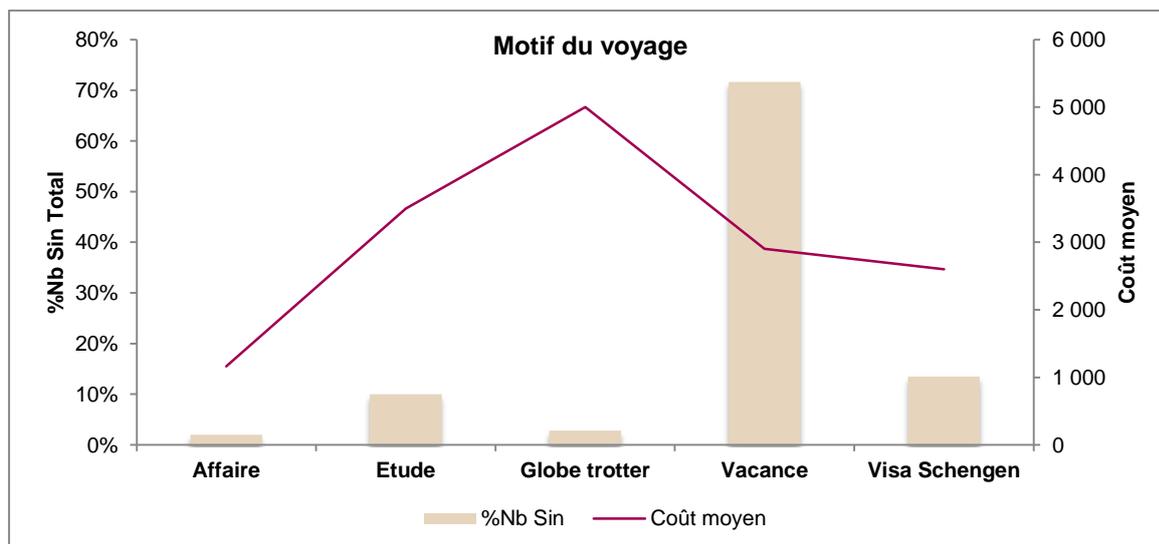
Figure 12: Coût moyen en fonction du nombre d'assurés



Nous observons que la grande majorité de nos sinistres sont portés par des polices avec un unique bénéficiaire pour un coût moyen de 3 700 euros. En revanche, environ 35% de nos contrats couvrent plusieurs bénéficiaires avec des coûts moyens très élevés pour les contrats contenant plusieurs assurés. La variable « nombre d'assurés », semble ne pas apporter d'information sur nos coûts moyens.

#### Variable « motif du voyage »

Figure 13: Coût moyen en fonction du motif du voyage



Au vu de graphique, le motif du voyage apparaît clairement discriminant pour la sévérité des sinistres. Les voyages de type « Globe-trotter » représentent la plus forte sinistralité en fréquence et en coût moyen derrière les déplacements de type « Vacances ». Cela peut se justifier en partie par le profil « aventurier » des assurés souscrivant le produit « Globe trotter ».

## 2. Etude de la corrélation entre les variables

Une fois le tri à plat effectué (analyse subjective), nous pouvons dès à présent passer à une analyse plus objective des variables retenues en observant les corrélations qu'elles ont entre elles. Cette deuxième approche, plus « statistique », nous permettra de confirmer ou d'infirmer les premières remarques effectuées précédemment.

Pour cela, le V de Cramer et le Gamma de Goodman – Kruskal décrits en annexe B vont être utilisés pour évaluer la corrélation entre les variables. Les variables les plus corrélées ne seront pas à insérer ensemble dans le modèle puisque cela biaiserait les résultats. Pour cela, nous définissons 50% (en valeur absolue) comme seuil critique de corrélation.

A partir de la procédure « proc freq » avec l'option « chisq » de SAS, les corrélations suivantes sont obtenues :

**Tableau 4: Tableau des corrélations entre variables explicatives**

	Destination géographique	Durée de l'exposition	Montant du voyage	Nombre d'assurés	Profil
Destination géographique	100%	26%	2%	11%	50%
Durée de l'exposition		100%	-70%	-44%	58%
Montant du voyage			100%	80%	10%
Nombre d'assurés				100%	25%
Profil					100%

La corrélation entre la variable profil de l'assuré et la destination géographique a été calculée grâce au V de Cramer, car les 2 variables sont des variables nominales. En revanche, la corrélation entre la variable durée de l'exposition et la variable nombre d'assurés a été calculée grâce au Gamma de Goodman, car il s'agit en effet de 2 variables ordinales. Dans le cas de la corrélation entre une variable ordinale et une variable nominale, nous avons considéré que le V de Cramer était plus adapté.

Du tableau de corrélations ainsi obtenu, nous pouvons en tirer les conclusions suivantes :

- Le montant du voyage est principalement corrélé au nombre d'assurés.
- la variable « Profil » est fortement corrélée à la durée de l'exposition. Ceci fait effectivement sens commercialement dans la mesure où les déplacements assimilés « Globe-trotter » se déroulent systématiquement sur de longues périodes à contrario des déplacements « Affaire » qui ont lieu sur de courtes périodes.
- Nous constatons également une corrélation entre le montant du voyage et la durée de l'exposition.

Une fois la corrélation entre les variables explicatives étudiées, il faut observer la corrélation entre les variables explicatives et la variable cible. Les résultats obtenus en fonction de la base utilisée sont présentés ci-dessous :

Tableau 5: Tableau des corrélations sur la fréquence et l'intensité des sinistres

	Fréquence des sinistres	Intensité des sinistres
Destination géographique	14%	29%
Durée du voyage	61%	15%
Montant du voyage	-33%	-19%
Nombre d'assurés	-7%	-11%
Profil	10%	22%

Concernant la fréquence des sinistres :

Suite à l'analyse de ce tableau, nos intuitions confirment un pouvoir discriminant pour les variables : « Durée du voyage » et « Montant du voyage ». La variable « Profil / Motif du voyage » ne semble pas très corrélée à la fréquence de sinistralité. Mais en revanche, il serait intéressant d'évaluer la corrélation Durée x Profil qui elle serait beaucoup plus importante.

Concernant l'intensité des sinistres :

Suite à l'analyse de ce tableau, nos intuitions confirment un pouvoir discriminant pour la variable : Destination géographique.

Toutefois, cela ne nous permet pas de valider définitivement cette variable pour notre modèle.

## D. Théorie des modèles linéaires généralisés (MLG)

L'objectif de cette partie est de rappeler rapidement les concepts mathématiques liés à la théorie des modèles linéaires généralisés.

### 1. Les MLG comme extension des modèles gaussiens

Pour mieux comprendre la généralisation des modèles linéaires, rappelons le cadre du modèle linéaire gaussien.

Nous cherchons à modéliser une variable  $Y$  (appelée variable à expliquer ou réponse) à l'aide d'un certain nombre de variables  $X$  (appelées variables explicatives).

Le modèle présenté par Legendre et Gauss à la fin du XVII siècle, se présente de la façon suivante :

$$Y \sim \mathcal{N}(\mu, \sigma^2) \text{ où } \mu = {}^t X \beta$$

Avec  $\beta$  les paramètres du modèle

La base des modèles linéaires généralisés (MLG) étend le modèle gaussien à une famille de lois particulières, appelée famille exponentielle, en définissant le modèle suivant :

$$Y \sim \text{Loi}(\mu) \text{ où } \mu = \mathbb{E}[Y] = g^{-1}(X^t \beta)$$

Le MLG est donc composé 3 éléments :

- la composante aléatoire  $Y$  ;
- la composante déterministe  $X$  ;
- la fonction de lien  $g$  .

### **Notations du modèle :**

Dans cette partie, nous adopterons les notations suivantes :

- $Y$  la variable à expliquer qui est un vecteur de dimension  $n$   $Y = {}^t(Y_1, \dots, Y_n)$  ;
- $(X_i)_{1 \leq i \leq p}$  les variables explicatives qui sont représentées par une matrice de dimension  $(n, p+1)$  :  $X = (1, X_1, \dots, X_p)$  ;
- $\beta$  les coefficients tarifaires qui sont un vecteur de dimension  $p+1$   $\beta = (\beta_0, \dots, \beta_p)$  ;
- $n$  le nombre de polices d'assurance ;
- $p$  le nombre de variables explicatives.

Et en admettant pour notre étude, les hypothèses « qualitatives » inhérentes aux GLM suivantes :

- Hypothèse 1 (Indépendance entre les observations) : Les observations  $(Y_i)_{1 \leq i \leq n}$  permettant l'estimation<sup>3</sup> du modèle linéaire doivent être indépendantes. Dans le cadre de ce mémoire, les observations sont données pour chaque police d'assurance (dans certains cas elles peuvent être de niveau risque – assuré) ;
- Hypothèse 2 (Indépendance et constance dans le temps des variables explicatives) : le MLG suppose également que les  $(X_i)_{1 \leq i \leq p}$  les variables explicatives de calibrage sont constantes dans le temps et indépendantes entre elles ;
- Hypothèse 3 (caractère iid. des résidus<sup>4</sup>): les résidus sont indépendants et identiquement distribués.

Nous considérerons dans la toute suite de ce mémoire que ces trois hypothèses sont respectées.

<sup>3</sup> Par estimation, on désigne le calcul des coefficients tarifaires

<sup>4</sup> La notion de résidu sera définie par la suite

**Rappel sur la famille exponentielle :**

Nous allons définir brièvement la famille de lois des modèles linéaires généralisés.

Le fait de dire que la densité doit appartenir à la famille exponentielle, signifie que celle-ci doit pouvoir s'écrire sous la forme suivante :

$$f(y|\theta, \varphi) = \exp\left(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right), y \in \mathbb{S}$$

Où  $\mathbb{S}$  est un sous-ensemble de  $\mathbb{N}$  ou de  $\mathbb{R}$ .

Il faut noter que :

- les fonctions  $a$ ,  $b$  et  $c$  et  $c$  sont spécifiques à chaque distribution de probabilité ;
- les fonctions  $a$  et  $c$  sont telles que  $a(\varphi) = \varphi/\omega$  et  $c = c(y, \varphi/\omega)$  où  $\omega$  est vecteur le poids connu à priori et affecté au vecteur  $Y$ ;
- $\theta$  est appelé paramètre canonique ;
- $\varphi$  est dit paramètre de dispersion ;

La moyenne et la variance d'une variable  $Y$  dont la densité peut se mettre sous cette forme sont alors :

$$\begin{aligned}\mathbb{E}[Y] &= b'(\theta) \\ \text{Var}(Y) &= \frac{b''(\theta)\varphi}{\omega}\end{aligned}$$

Où les signes « prime » ici désignent les dérivées première et seconde par rapport à  $\theta$ . Si l'on note  $\mu$  la moyenne de  $Y$ , alors la variance peut s'écrire comme une fonction de la moyenne  $\text{Var}(Y) = \frac{V(\mu)\varphi}{\omega}$  où  $V$  est la fonction variance.

## 2. Estimation des paramètres : Méthode du maximum de vraisemblance

Toute l'étude relative à la sélection des variables sera réalisée sur SAS. Nous présentons ici les techniques utilisées par l'outil pour calculer les paramètres du modèle.

### Ajustement du modèle

Avant d'arriver à la description de la métrique utilisée pour la sélection des variables, il est important de présenter la manière dont sont estimés les coefficients tarifaires (ou de régression)  $(\beta_i)_{0 \leq i \leq p}$ .

Les coefficients de régression sont estimés dans SAS en maximisant la log-vraisemblance du MLG. Une fois le modèle mis sous forme exponentielle, la log-vraisemblance s'écrit alors :

$$L(\beta) = \sum_{i=1}^n \log(f(y_i, \mu_i, \varphi))$$

La log-vraisemblance est paramétrée dans SAS non pas avec son paramètre canonique  $\theta_i$  mais plutôt la moyenne  $\mu_i$ . SAS utilise ensuite une procédure itérative appliquée à l'algorithme de Newton-Raphson pour maximiser la log-vraisemblance.

Par exemple, à la  $r$ -ième itération, l'algorithme remplace le vecteur  $(\beta)_r$  par :

$$(\beta)_{r+1} = (\beta)_r - H^{-1}s$$

avec  $H$  la matrice Hessienne (dérivée seconde) et  $s$  le gradient (dérivée première) de la log-vraisemblance par rapport aux coefficients de régression. Ils peuvent se mettre sous la forme

$$s = (s_j)_{0 \leq j \leq p} = \left( \frac{\partial L}{\partial \beta_j} \right)_{0 \leq j \leq p} \quad \text{et} \quad H = (h_{i,j})_{0 \leq i,j \leq p} = \left( \frac{\partial^2 L}{\partial \beta_j \partial \beta_i} \right)_{0 \leq i,j \leq p}.$$

Dans toute la suite, nous noterons  $\hat{\beta}$  le vecteur des estimateurs des coefficients de régression et  $L = L(\hat{\beta})$  la log-vraisemblance maximisée du MLG.

### Qualité de l'ajustement du modèle : la déviance

La qualité d'ajustement du modèle ainsi estimée est mesurée grâce à la déviance. La déviance est une mesure statistique qui se définit comme  $D = 2 \log \frac{L_{SAT}}{L}$  (déviance réduite) avec  $L_{SAT} (= \sum_{i=1}^n \log(f(y_i, y_i, \varphi)))$  la vraisemblance du modèle saturé (c'est-à-dire le modèle qui possède autant de paramètres que d'observations).

Plus la déviance sera proche de 0, meilleure sera la qualité de l'ajustement.

Par ailleurs, la déviance réduite  $D$ , sous hypothèse de significativité du modèle, suit une loi du  $\chi_2$  à  $n-p-1$  degrés de liberté où  $n$  est le nombre d'observations et  $p$  le nombre de paramètres à estimer. De plus sa valeur espérée est égale au nombre de degrés de liberté résiduel. Ainsi, il faut, pour que le modèle soit acceptable, que le rapport  $D/(n-p-1)$  ne soit pas grand devant 1.

### Test de significativité des variables tarifaires : le test du rapport de vraisemblance

Le test du rapport de vraisemblance permet de tester la significativité des variables tarifaires prises en compte dans le modèle linéaire. Cela revient à comparer la déviance de 2 modèles l'un avec contraintes et l'autre sans contraintes. La contrainte consiste à retirer la variable dont on veut tester la significativité. L'hypothèse s'écrit alors :

$H_0$ : variable  $X$  non significative

$H_1$ : variable  $X$  significative

La statistique  $\Delta = 2[L - \tilde{L}]$  suit une loi du Chi-deux à  $k$  degrés de liberté. Avec  $k$  le nombre de coefficients de régression liés à la variable  $X$ ,  $L$  le log-vraisemblance du modèle sans contrainte et  $\tilde{L}$  la log-vraisemblance du modèle avec contrainte.

Ainsi, l'hypothèse  $H_0$  sera rejetée au niveau  $\alpha$  si  $\Delta > \chi_{k,1-\alpha}^2$ .

### Significativité des coefficients estimés : test de Wald

Pour tester la significativité des coefficients estimés, nous aurons recours au test de Wald. La significativité d'un seul paramètre est analysée. En conséquence, les coefficients sont considérés

séparément contrairement au test du rapport de vraisemblance où nous passons à 0 l'ensemble des coefficients liés à une variable. Pour une variable  $X_j$  l'hypothèse s'écrit alors :

$$\begin{aligned} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0 \end{aligned}$$

La statistique  $T = \frac{\widehat{\beta}_j}{\widehat{\sigma}_j}$  suit une loi du Chi-deux. Avec  $\widehat{\beta}_j$  l'estimateur de  $\beta_j$  et  $\widehat{\sigma}_j$  l'écart-type d'estimation de  $\beta_j$ . Ainsi, l'hypothèse  $H_0$  sera rejetée à  $1 - \alpha\%$  si la valeur  $t_0$  de T est que  $\mathbb{P}[T > t_0] < \alpha$  ce qui signifie que le coefficient est significativement non nul.

## 2.1 Critères de sélection des variables du modèle

C'est dans cette étape que SAS détermine grâce à des indicateurs synthétiques quel modèle est meilleur qu'un autre. Pour cela, il utilise plusieurs types d'indicateurs :

### Le critère AIC

Par définition l'AIC (Akaike Information Criterion) pour un modèle à p paramètres est défini par :

$$AIC = -2L + 2p$$

L est la vraisemblance maximisée et p le nombre de paramètres dans le modèle. Avec ce critère, la déviance du modèle est pénalisée par deux fois le nombre de paramètres. L'AIC représente donc un compromis entre le biais et la parcimonie. En effet, le biais va diminuer avec le nombre de paramètres tandis que la parcimonie nécessite de décrire les données avec le plus petit nombre de paramètres possible. Ainsi, plus ce critère sera petit, meilleur sera le modèle.

### Le critère SC / BIC

Un autre critère de choix de modèle est le SC / BIC (Schwarz / Bayesian Informative Criterion). Pour un modèle à p paramètres estimés sur n observations, il est défini par :

$$BIC = -2L + p \log n$$

Tout comme pour l'AIC, plus le critère de choix BIC est faible et meilleur sera le modèle.

### Courbe ROC :

Un indicateur de la capacité du modèle à discriminer est la courbe ROC (*Receiver-Operating Characteristic*) Cette technique récente est empruntée à la bio-statistique. Lorsque l'on fait varier le seuil pour la règle de décision, la sensibilité et la spécificité changent et évoluent en sens inverse. Il s'agit donc de tracer la sensibilité en fonction de (1-spécificité) puis de calculer l'aire située en dessous de cette courbe. Ceci permet d'évaluer la précision du modèle pour discriminer les résiliations des non résiliations. Si la courbe se situe sur la bissectrice du repère du graphique, alors le modèle n'a aucun pouvoir discriminant. Sa capacité à discerner les contrats n'ayant pas été sinistrés des sinistrés est la même que si l'on faisait un tirage au sort. En revanche, plus la courbe s'écarte de

cette bissectrice, et meilleur est le pouvoir discriminant du modèle. On retiendra la règle suivante :

**Tableau 6: Interprétation de la valeur de l'aire sous la courbe ROC**

Aire ROC	Pouvoir discriminant
Egal à 50%	Pas de discrimination
Entre 50% et 60%	La discrimination est mauvaise
Entre 60% et 70%	La discrimination est acceptable
Entre 70% et 80%	La discrimination est bonne
Entre 80% et 90%	La discrimination est excellente
Supérieur à 90%	La discrimination est exceptionnelle

SAS fournit en sortie la statistique C, qui est une estimation de l'aire ROC.

### 3. Description des variables à expliquer

#### 3.1 Modélisation de la fréquence des sinistres

Afin de suivre et piloter le suivi du risque assistance sur les contrats « temporaires », il est nécessaire d'analyser la première composante explicative de la charge Sinistres du portefeuille : la fréquence des sinistres. En effet, la fréquence des sinistres de type assistance sur les contrats « temporaires » étant très faible, il nous est d'autant plus important d'ajuster le plus finement possible cette fréquence empirique.

A ce titre, la modélisation de la fréquence des sinistres s'effectue en trois étapes :

- Choix du modèle ;
- Estimation des paramètres de ce dernier ;
- Validation.

Le modèle sera choisi à travers différentes lois : loi binomiale et binomiale négative, loi de Poisson. Le test de Kolmogorov-Smirnov sera effectué pour valider ou non le modèle étudié.

#### Choix du modèle

Afin de déterminer la loi qui décrit le mieux la fréquence de sinistralité, commençons par déterminer le tableau des fréquences empiriques.

**Tableau 7: Tableau des fréquences empiriques**

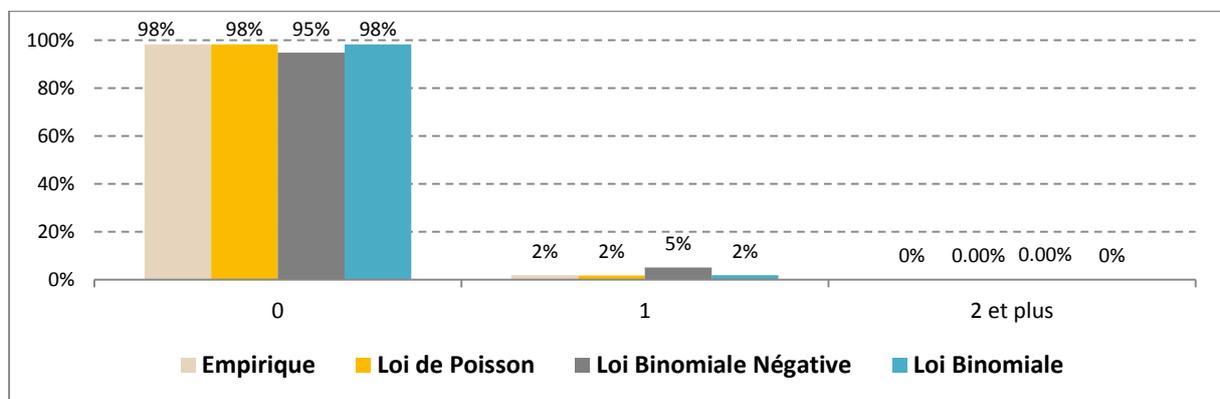
Nombre de sinistres Assistance	Nombre de contrats	Fréquence empirique en jours	Fréquence empirique relative
0	50232	1 427574	98,1%
1	416	26390	1,9%

**Note :** La fréquence ici est calculée en reprenant les formules explicitées en préambule. A savoir, pour connaître par exemple la fréquence des contrats ayant connu 1 sinistre, nous faisons le rapport le nombre total de jours des contrats ayant connu 1 sinistre par le nombre total de jours du portefeuille.

Pour les contrats temporaires sur le périmètre assistance, nous ne pouvons observer au maximum qu'un seul sinistre par an et par contrat. Cette information nous suggère *a priori* de modéliser la fréquence à partir d'une loi binomiale. Malgré tout, nous tentons de modéliser la variable aléatoire nombre de sinistres par une loi de Poisson ou une loi binomiale négative. L'objectif ici est de s'assurer que ces lois ont néanmoins des probabilités d'avoir plus de 2 sinistres par an et par contrat relativement faibles.

Pour cela, nous comparons graphiquement les estimations de lois probabilistes à la réalité historique. Ce qui revient à comparer la fréquence d'apparition d'un nombre de sinistres empiriques à celui modélisé.

Figure 14 : Distribution du nombre de sinistres par contrat estimé avec un modèle probabiliste



Grâce au test d'adéquation de Kolmogorov – Smirnov (en annexe), nous avons calculé les *D – Statistics* pour la loi binomiale, la loi de Poisson et la loi Binomiale Négative. Les résultats sont les suivants :

Tableau 8: Statistiques de Kolmogorov - Smirnov pour la loi de Poisson et la loi Binomiale Négative

Loi	D	p value
Binomiale	0,90	0.3567
Poisson	0,98	<0.001
Binomiale négative	0,99	<0.001

Bien que les *D statistics* apparaissent proches, les *p values* des lois de Poisson et Binomiale Négative sont inférieures au seuil de 5%. Les tests d'ajustement présenté dans ce tableau nous permettent de valider notre choix concernant la distribution à retenir.

Par conséquent, nous faisons le choix de modéliser la fréquence de sinistres comme l'espérance de l'indicatrice de survenance d'un sinistre  $I$ , prenant la valeur 1 ou 0 suivant que l'assuré a connu un sinistre ou non. Autrement dit, cette variable suit donc une loi de Bernoulli de paramètre (et d'espérance)  $p = \mathbb{P}[I = 1]$ .

Par conséquent, la loi de Bernoulli (cas particulier de la loi binomiale) sera retenue pour la suite du mémoire. Cette loi de probabilité permet de modéliser une variable binaire. La variable à modéliser sera re-codifiée sous la forme binaire "aucun sinistre" ou "au moins un sinistre", et une régression logistique sera faite.

### Estimations des paramètres de la loi retenue

La loi de Bernoulli peut se mettre sous la forme exponentielle<sup>5</sup>. Considérons une loi de Bernoulli  $Y$  de paramètre  $p = \mathbb{P}[Y = 1] = \mathbb{E}[Y]$ . La densité s'écrit alors

$$f(y) = \exp \left[ y \ln \left( \frac{p}{1-p} \right) + \ln(1-p) \right].$$

On reconnaît là une densité de la forme exponentielle avec :

- Le paramètre  $\theta = \ln \left( \frac{p}{1-p} \right)$  ;
- $b(\theta) = -\ln \left( \frac{1}{e^\theta + 1} \right)$  ;
- $\phi = 1$  et  $\omega = 1$  pour toutes les observations.

Elle découle directement de l'expression de la log-vraisemblance :

$$L = \sum_{i=1}^n y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

En se basant sur la méthode des moments (qui coïncide avec la méthode du maximum de vraisemblance dans le cas de la loi Binomiale), le paramètre  $p$  est déterminé de la façon suivante :

$$\hat{p} = \frac{\sum_i top\_sinistre\_assistance_i * Duree_i}{\sum_i Duree_i}$$

Où  $Duree_i$  est la durée maximale de couverture de la garantie pour le contrat  $i$  ;  $top\_sinistre\_assistance_i$  est l'indicatrice de l'événement de sinistre pour le contrat  $i$ .

<sup>5</sup> Familles exponentielles au sens des modèles linéaires généralisés

### 3.2 Modélisation du coût moyen

La modélisation du coût moyen se détermine de manière similaire à la fréquence des sinistres en respectant les trois étapes explicitées dans la partie précédente. Après indexation des sinistres, les distributions que l'on envisage pour le montant des sinistres sont les suivantes :

- Log-normale ;
- Gamma ;
- Gaussienne inverse.

#### Recherche de lois sous-jacentes

Pour tester l'adéquation entre nos observations et les lois suggérées, nous commençons par observer les QQ-plot. Le QQ-plot est un graphique représentant les quantiles théoriques en fonction des quantiles empiriques. On conclut à un ajustement correct lorsque les points sont alignés sur la première bissectrice. La production des QQ-Plots est faite automatiquement grâce à la procédure « univariate » de SAS. Elle détermine, par estimation du maximum de vraisemblance, les paramètres des lois Gamma et Log-normale. La méthode du maximum de vraisemblance sera détaillée dans la suite de cette partie.

- Loi log-normale :

La loi Lognormale a pour densité :

$$f(x) = -\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2}$$

Avec  $\mu \in \mathbb{R}$  et  $\sigma$  un réel positif.

Commençons par observer la répartition des sinistres selon leur coût :

Figure 15: Histogramme du montant des sinistres ajusté par une loi log-normale

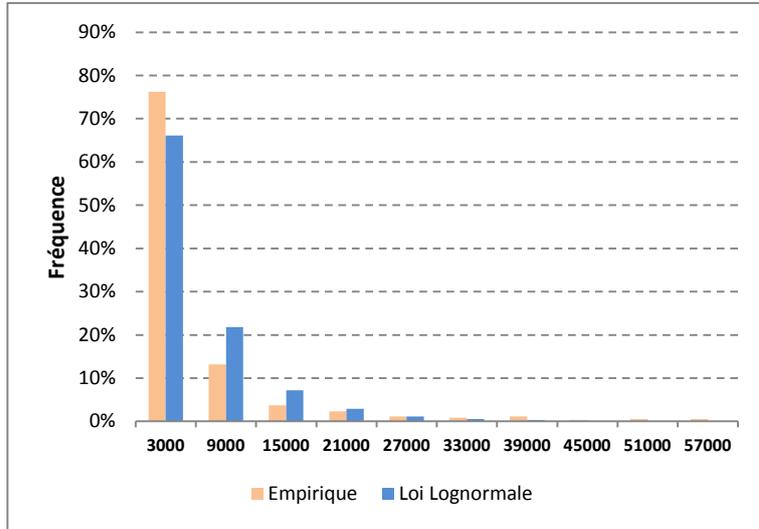
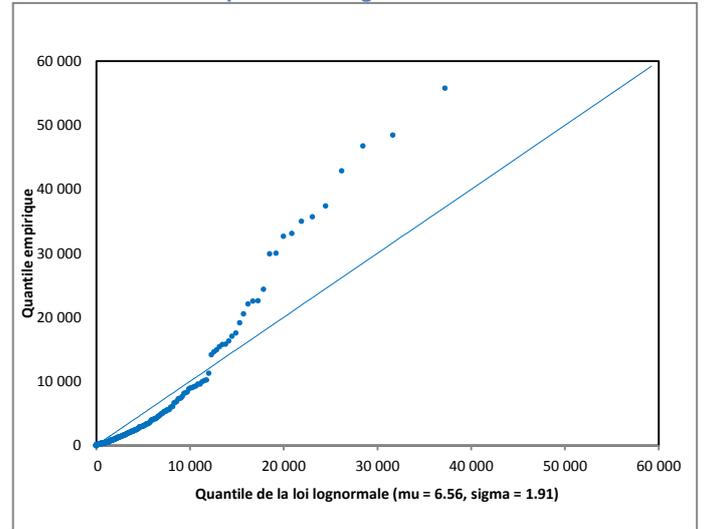


Figure 16: QQ plot Montant des sinistres empirique / Montant modélisé par une loi log-normale



Grâce à l'histogramme, nous remarquons globalement que la loi log-normale n'arrive pas à capter les variations des montants des sinistres. La Figure 15 permet de faire un zoom sur la manière de modéliser les faibles valeurs de montants des sinistres (< 3 000€). C'est effectivement cette partie qui nous intéresse en raison du niveau moyen des montants de sinistres. Nous nous rendons compte que la loi log-normale (histogramme bleue) est en-dessous de l'histogramme orange.

A présent, nous allons faire un test d'ajustement (par la méthode des distances de Kolmogorov-Smirnov).

Tableau 9 : Résultat de l'ajustement du montant des sinistres par une loi log-normale

Goodness-of-Fit Tests for Lognormal Distribution		
Test	Statistic	
Kolmogorov-Smirnov	D	0.34985522
	p value	0.0205

La lecture sur une table de valeurs nous donne, pour un seuil fixé à 5% d'erreur la p-valeur. Elle apparaît sur ce tableau étant inférieure à ce seuil, l'ajustement de la loi n'est pas valable, au moins du point de vue théorique.

- Loi Gamma :

La loi Gamma a pour densité :

$$f(x) = \frac{z^\alpha e^{-z}}{x\Gamma(\alpha)}$$

Avec  $\theta > 0$ ,  $\alpha > 0$  et  $z = x/\theta$ .

De même pour la loi log-normale, nous représentons l'histogramme et le qq-plot pour la loi Gamma :

Figure 17 Histogramme du montant des sinistres ajusté par une loi Gamma

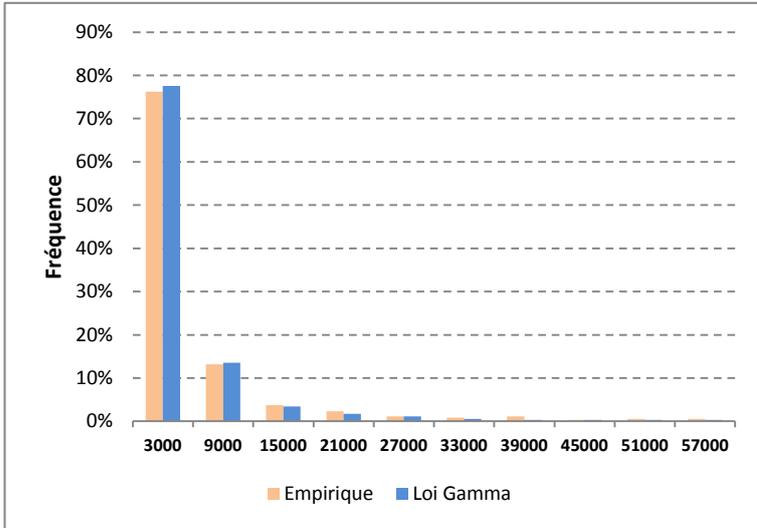
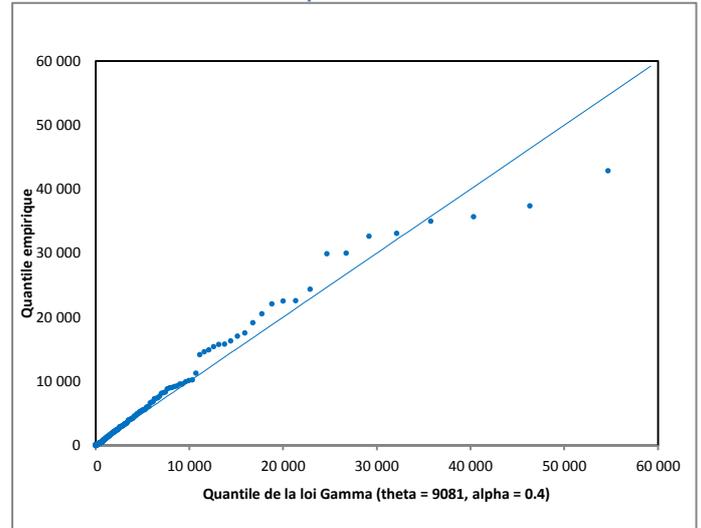


Figure 18: QQ plot Montant des sinistres empirique / Montant modélisé par une loi Gamma



Dans le cas de l'utilisation de la loi Gamma, nous avons un effet inverse à celui obtenu pour la loi log-normale. En ce sens que, selon le QQ-Plot, entre 0 et 10 000€, la loi gamma estime parfaitement la distribution empirique du montant des sinistres. Et au-delà de 15 000€ nous obtenons une sous-estimation des montants d'indemnité.

Nous calculons la statistique de KS :

Tableau 10: Résultat du test d'ajustement du montant des sinistres par une loi Gamma

Goodness-of-Fit Tests for Gamma Distribution		
Test	Statistic	
Kolmogorov-Smirnov	D	0.1145635
	p value	0.376

La lecture sur une table de valeurs nous donne, pour un seuil fixé à 5% d'erreur la p-valeur. Elle apparaît sur ce tableau comme étant supérieure à ce seuil, l'ajustement de la loi est donc valable, au moins du point de vue théorique.

- Loi gaussienne inverse

La loi Gaussienne Inverse a pour densité :

$$f(x) = \frac{1}{\theta} \sqrt{\frac{\alpha}{2\pi z^3}} e^{-\frac{\alpha(z-1)^2}{2z}}$$

Avec  $\theta > 0$ ,  $\alpha > 0$  et  $z = x/\theta$ .

Figure 19 Histogramme du montant des sinistres ajusté par une loi Gaussienne Inverse

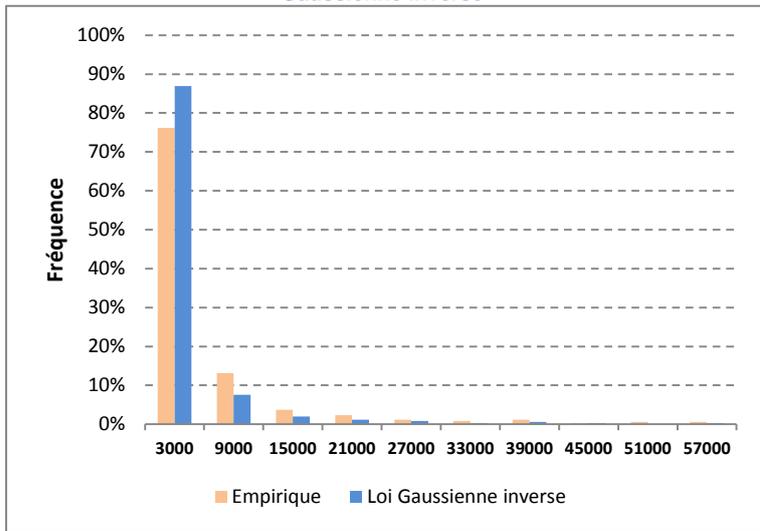
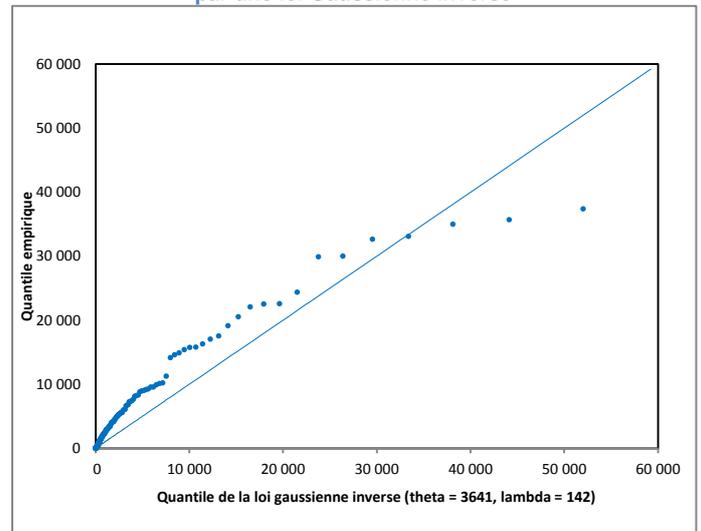


Figure 20: QQ plot Montant des sinistres empirique / Montant modélisé par une loi Gaussienne Inverse



Dans le cas de l'utilisation de la loi Gaussienne Inverse, les montants modélisés sont globalement sous-estimés.

Tableau 11: Résultat du test d'ajustement du montant des sinistres par une loi Gaussienne Inverse

Goodness-of-Fit Tests for Inverse Gaussian		
Test	Statistic	
Kolmogorov-Smirnov	D	0.740550721
	p value	<0.001

La lecture sur une table de valeurs nous donne, pour un seuil fixé à 5% d'erreur la p-valeur. Elle apparaît sur ce tableau comme étant très inférieure à ce seuil, l'ajustement de la loi n'est donc pas valable, au moins du point de vue théorique.

La statistique de Kolmogorov pour loi Gamma est supérieure à celles de la log-normale et de la Gaussienne Inverse.

Toutefois, l'ajustement à la loi Gamma semble beaucoup mieux adapté à celui de la loi log-normale et de la Gaussienne Inverse. En effet, la loi Gamma capte mieux les effets sur les petits montants que sur les gros montants.

En conclusion, lors du paramétrage de notre modèle, nous prendrons le soin de préciser que nous souhaitons une modélisation des valeurs du coût moyen selon une loi Gamma.

### Estimation des paramètres de la loi retenue

La fonction de densité de la loi Gamma, tels que les paramètres  $\alpha$  et  $\theta$  soient strictement positifs, est définie par l'expression suivante :

$$f(x) = \frac{z^\alpha}{x \Gamma(\alpha)} e^{-z}$$

Avec  $z = x/\theta$ ,  $\alpha > 0$ ,  $\theta > 0$  et  $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ .

La procédure « Univariée » de SAS fait appel à la méthode d'estimation du maximum de vraisemblance pour estimer les paramètres de la loi Gamma retenue pour la modélisation du montant des sinistres Assistance. Nous allons décrire les grandes lignes des principes de cette méthode.

Pour estimer les paramètres de la loi Gamma, SAS utilise un maximum de vraisemblance approché. Le log du maximum de vraisemblance pour une distribution Gamma s'écrit pour un échantillon de  $n$  variables aléatoires iid  $y_i$  ( $1 \leq i \leq n$ ) de la façon suivante :

$$\begin{aligned} L &= \sum_{i=1}^n \log \left( y_i^{\alpha-1} \frac{e^{-\frac{y_i}{\theta}}}{\theta^\alpha \Gamma(\alpha)} \right) \\ &= (\alpha - 1) \sum_{i=1}^n \log(y_i) - \frac{1}{\theta} \sum_{i=1}^n y_i - n\alpha \log(\theta) - n \log(\Gamma(\alpha)) \end{aligned}$$

Pour des raisons d'allègement des écritures, nous remplacerons la notation  $\sum_{i=1}^n$  par  $\Sigma$ .

- Approximation de la fonction Digamma

En dérivant l'équation précédente (le log de la vraisemblance) par rapport à  $\theta$  et en résolvant l'équation  $\frac{\partial l}{\partial \theta} = 0$ , on arrive à l'estimateur de  $\theta$  par maximum de vraisemblance :

$$\hat{\theta} = \frac{\Sigma y_i}{n\alpha} = \frac{m_1}{\alpha}$$

En remplaçant  $\theta$  par son estimateur  $\hat{\theta}$  dans l'expression de  $l$  et en simplifiant, on obtient :

$$L = (\alpha - 1) \Sigma \log(y_i) - n\alpha - n\alpha \log(m_1) + n\alpha \log(\alpha) - n \log(\Gamma(\alpha))$$

Soit  $d$  un réel défini comme suit :  $d = \log(m_1) - \frac{1}{n} \Sigma \log(y_i)$ .

En résolvant l'équation  $\frac{\partial l}{\partial \alpha} = 0$ , on arrive à la fonction digamma  $\psi(\alpha)$  :

$$\log(\alpha) - \psi(\alpha) = d$$

En première approximation, la fonction digamma peut-être approché de la manière suivante :

$$\hat{\psi}(\alpha) \approx \log(\alpha) - \frac{1}{\alpha} \left( 0.5 + \frac{1}{12\alpha + 2} \right)$$

Cette approximation est 1.4% proche de la valeur exacte pour des valeurs de  $\alpha > 0$ . Lorsque  $\alpha$  est proche d'une racine positive de la fonction digamma (qui est approximativement de 1.46) cette valeur de 1.4% devient plus importante tout en restant à de seuil de différence significativement bas.

- Initialisation des valeurs par la méthode des moments

Si la méthode d'approximation du maximum de vraisemblance est impossible, la procédure « univariate » de SAS utilise alors la méthode des moments.

Pour rappel, le moment d'ordre  $k$  de la distribution Gamma :

$$\mathbb{E}[X^k] = \theta^k \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)}, k > -\alpha$$

Considérons par exemple les moments d'ordre 1 et 2 que l'on note  $m_1$  et  $m_2$ , une des solutions à ces 2 équations est la valeur initiale de  $\alpha$  utilisée par SAS :

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}$$

Si la différence  $m_2 - m_1^2$  est proche de 0, alors SAS initialise  $\hat{\alpha}$  à 1. Après avoir exhibé un estimateur de  $\alpha$ , l'estimateur de  $\theta$  s'écrit :

$$\hat{\theta} = \frac{m_1}{\hat{\alpha}}$$

### **Validation du choix de la distribution retenue**

Après une première analyse graphique (QQ-Plot), l'analyse de la p-valeur afférente au test de Kolmogorov-Smirnov permet de valider le choix de la loi Gamma.

## **E. Sélection des variables de tarification**

*Note :* Cette partie présentera, dans un premier temps, la théorie sous-jacente à la méthodologie de sélection des variables tarifaires. Les résultats seront présentés au fur et à mesure.

Les parties précédentes ont permis l'analyse uni-variée de la sinistralité du risque assistance. Avant l'arrivée des outils informatiques puissants, les actuaires avaient une approche uni-variée de la tarification. L'objectif de cette partie est de mettre en avant la tarification dans son approche multi-variée. En effet, l'examen en multi-variée de la tarification a pour principal atout la prise en compte des interactions entre les différentes variables tarifaires. Une interaction peut se produire lorsque le coefficient tarifaire d'une variable dépend de l'effet combiné de plusieurs autres. C'est dans ce même état d'esprit que nous allons dans la suite sélectionner les variables tarifaires.

La méthodologie adoptée pour la sélection des variables consiste à tester la combinaison de variables qui a le plus grand pouvoir explicatif de la prime pure tout en garantissant une estimation des coefficients tarifaires robuste et cohérente. En pratique, à partir de l'ensemble des variables préalablement sélectionnées, nous aurons recours à des procédures de sélection automatique pour la construction du modèle tarifaire. Ces procédures visent à identifier les variables qui, en tenant compte des autres, permettent d'expliquer de façon significative la déviance totale.

Toutefois, étant donné le nombre limité des variables explicatives de notre étude, nous aurions pu nous passer de l'étape de sélection automatique en intégrant directement l'ensemble de nos variables dans notre modèle. Nous ferons toutefois le choix d'utilisation de cette procédure automatique afin de tester toutes les significativités entre nos variables (significativité des variables restantes lors de la non prise en compte d'une variable non significative) et par souci de présentation de cette méthodologie très largement utilisée lors des études comportant un nombre important de variables explicatives.

Commençons par décrire le principe général de sélection des variables tarifaires.

### **Principe de la méthode de sélection automatique (Stepwise) :**

Le premier modèle testé est celui ne comportant que le terme constant, puis on intègre pas à pas les variables exogènes les plus influentes, c'est-à-dire celles dont l'introduction apporte un meilleur pouvoir discriminant. A chaque entrée de variables, le caractère significatif de toutes les variables déjà présentes dans le modèle est testé et comparé à un seuil de confiance (5% dans ce mémoire compte tenu du nombre d'observations). A chaque pas, cette méthode permet un réexamen des variables introduites dans le modèle aux étapes antérieures. Autrement dit, elle consiste à une sélection ascendante avec une élimination descendante. Le processus est itératif et continue jusqu'à ce qu'aucune variable additionnelle ne soit acceptée et qu'aucune variable sélectionnée précédemment ne soit éliminée.

## 1. Etude sur la fréquence des sinistres : Modèle de régression logistique

*Note : Il est important de noter que cette partie ne traite pas de l'estimation du modèle tarifaire à proprement parler. Nous n'y ferons donc pas apparaître les coefficients de régression mais exclusivement les tests de significativité des variables candidates.*

La démarche adoptée pour sélectionner les variables tarifaires au titre de la fréquence est assez structurante. Nous commençons par estimer le modèle avec toutes les variables que nous avons à notre disposition. En cas de non significativité de l'une des variables, nous appliquons la procédure de sélection automatique *stepwise*.

L'objectif de cette partie est alors double :

- de déterminer le paramétrage choix de la fonction de lien ;
- et d'analyser la significativité de chaque variable afin d'avoir une première idée des résultats de la sélection automatique.

### **1.1 Paramétrage du modèle de régression logistique sous SAS**

L'une des premières difficultés d'ordre pratique dans l'exécution d'un GLM est souvent liée au choix de la fonction de lien. Comme nous l'avons déjà souligné plus haut, modéliser un phénomène contenu dans un certain intervalle de valeurs, par un modèle qui, lui, ne le serait pas, n'est pas acceptable. L'utilisation d'une fonction de lien bien définie est une réponse à ce problème. Pour le nombre de sinistres, nous avons fait le choix de le modéliser par une loi de Bernoulli de paramètre  $p \in [0; 1]$ . La fonction de lien doit donc avoir la propriété de conserver l'intervalle de valeurs  $[0; 1]$ . La fonction logit a cette propriété :  $\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta X \Leftrightarrow p = \frac{\exp(\beta X)}{1 + \exp(\beta X)} \in [0; 1]$  Avec  $p = \mathbb{P}(Y = 1|X)$  représentant la probabilité d'avoir un sinistre connaissant la réalisation des variables explicatives.

Le choix du poids dans notre paramétrage est également primordial. Le poids ici est la durée du séjour.

La log-vraisemblance pondérée équivaut alors à :

$$L(\beta) = \sum_{i=1}^n \omega_i [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

$$= \sum_{i=1}^n \omega_i \left[ y_i \left( \beta_0 + \sum_{j=1}^n \beta_j x_{ij} \right) - \log \left( \beta_0 + \sum_{j=1}^n \beta_j x_{ij} \right) \right]$$

Où  $\omega_i$  représente le poids de chaque observation,  $y_i$  représente les observations et  $x_{ij}$  sont les réalisations des variables explicatives. Rappelons que les paramètres  $\beta_j$  sont estimés par maximisation de la fonction L.

## 1.2 Modèle avec toutes les variables disponibles : le « modèle initial »

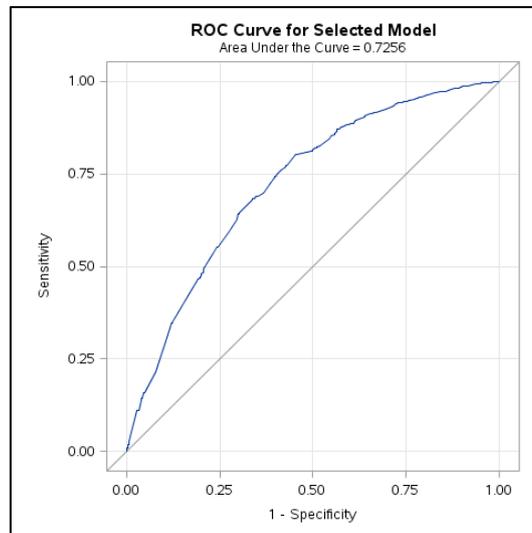
La table ci-dessous restitue les résultats des tests de significativité pour chacune des variables grâce au test du rapport de vraisemblance. Selon les tests, la variable montant du voyage n'est pas significative pour expliquer le nombre de sinistres. Cette observation sera validée par la suite (partie suivante) grâce à la procédure de sélection automatique des variables. La statistique de Wald permet d'avoir une idée sur la contribution de chaque variable à la capacité prédictive du modèle. Nous observons ainsi que la variable durée du séjour est la plus contributrice à l'explication du phénomène de la fréquence des sinistres. Elle est suivie par le profil du voyageur. La variable la moins contributrice est le montant du voyage.

Tableau 12: Tests de significativité (type 3) des variables du MLG "initial" du nombre de sinistres

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
DU	8	5879.14	<.0001
MT	1	6.09	0.0735
NA	2	789.36	<.0001
Zone	7	11354.52	<.0001
Profil	4	26745.65	<.0001

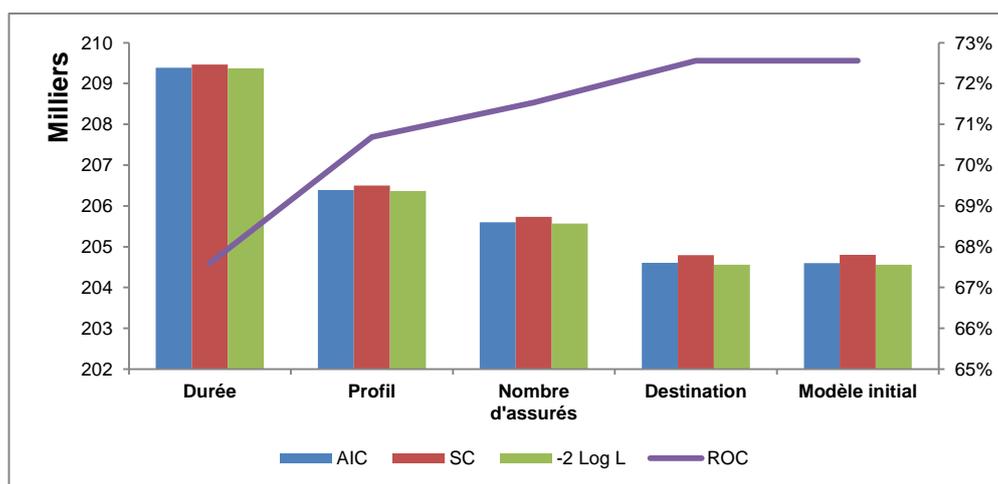
SAS nous permet également d'obtenir la courbe ROC et calcule automatiquement l'aire sous la courbe qui est une excellente mesure de la capacité prédictive du modèle. Ici, l'aire sous la courbe est de 73%. La discrimination est plutôt bonne selon la grille de décision établie par ailleurs (Tableau 6).

Figure 21: Courbe ROC du modèle initial de fréquence



### 1.3 Sélection automatique des variables explicatives

Nous testons les variables sélectionnées avec la méthode *stepwise* afin de vérifier et confirmer le pouvoir discriminant de nos variables sélectionnées sans l'intégration de la variable « Montant du voyage ». Pour ceci, nous utilisons la procédure PROC LOGISTIC sous SAS paramétrée avec l'option *stepwise*. Ci-dessous l'évolution des critères AIC, BIC et  $-2\log L$  est représentée en fonction des variables dans leur ordre de sélection :

Figure 22: Evolution des métriques de sélection des variables avec la méthode *stepwise*

Dès la première étape de la *stepwise* (introduction de la variable « Durée du voyage »), l'aire sous la courbe ROC mesurée par la statistique C dans SAS est 68%. L'aire sous ROC nous permettant d'évaluer la performance de règle de classement. Le pouvoir discriminant du modèle est alors considéré comme « acceptable ». Cela montre que le phénomène de fréquence de la sinistralité

s'explique essentiellement par cette variable. L'ajout de la variable « profil » apporte un gain en aire sous la courbe ROC de 3%

Le modèle initial, c'est-à-dire celui avec les 5 variables possède un pouvoir discriminant égal à un modèle avec les 4 variables Durée, Profil, Nombre d'assurés et Destination. Autrement dit, la variable Montant du voyage n'est pas discriminante pour le nombre de sinistres. C'est d'ailleurs ce qui ressort de la sélection stepwise qui n'a retenu que les 4 variables citées précédemment.

**Tableau 13: Tests de significativité des variables du MLG sans le montant du voyage**

Type 3 Analysis of Effects			
Effect	DF	Wald	Pr > ChiSq
		Chi-Square	
Profil	4	19987.3408	<.0001
DU	8	57357.4389	<.0001
NA	2	903.3289	<.0001
Zone	7	780.7463	<.0001

En résumé, les variables sélectionnées par la procédure stepwise sont donc :

- la durée du voyage ;
- le profil des voyageurs ;
- la destination géographique ;
- et le nombre d'assurés.

## 2. Etude sur le coût des sinistres

De même que pour le nombre de sinistres, nous commençons par justifier le paramétrage de la procédure « GENMOD » de SAS et ensuite nous présentons les résultats de la sélection des variables pour le coût moyen.

### 2.1 Paramétrage du modèle de régression gamma sous SAS

La fonction de lien canonique de la loi gamma est la fonction inverse. L'utilisation de la fonction de lien canonique facilite les calculs théoriques mais en pratique, elle présente quelques inconvénients :

- le premier est qu'elle ne garantit pas la prédiction de coûts de sinistres positifs *a contrario* de la loi Gamma qui elle ne prend en compte que des valeurs positives ;
- le deuxième inconvénient est que les coefficients des co-variables sont à interpréter à l'inverse de l'ordre de la logique habituelle. En ce sens que plus le coefficient est élevé en valeur absolue, moins son influence est grande sur la valeur prédite.

Pour toutes ces raisons, nous avons fait le choix d'utiliser la fonction logarithme népérien comme fonction de lien.

## 2.2 Modèle avec toutes les variables disponibles : le « modèle initial »

Nous estimons notre régression Gamma et la procédure ainsi paramétrée, nous obtenons les résultats suivants :

Tableau 14: Statistiques d'ajustement du MLG "initial" du coût moyen

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	416	1054.7608	3.4357
Scaled Deviance	416	319.0000	1.0391
Log Likelihood		-643.3840	
AIC (smaller is better)		1312.7679	
BIC (smaller is better)		1361.7154	

Notre attention se portera uniquement sur les statistiques de déviance, de déviance standardisée (*scaled deviance*, en anglais) et de log-vraisemblance (*Log Likelihood*). Leurs valeurs se lisent dans la colonne Value. La colonne DF (*Degrees of Freedom*) contient le nombre de degrés de liberté résiduels. La déviance standardisée est égale à 319. Cette valeur, rapportée au nombre de degrés de liberté, fait 1.0391, ce qui est pas trop éloigné de 1 pour en effet considérer que le modèle est recevable.

Pour être plus précis, nous pouvons calculer la p-value du test du chi 2. Celle-ci peut s'exprimer par  $\mathbb{P}[\chi_{n-p-1}^2 > Value]$ , avec Value égale à la valeur de la déviance standardisée. En pratique, nous l'obtenons à partir de la fonction « CDF » de SAS, qui renvoie la valeur de la fonction de répartition d'une distribution donnée. La p-value obtenue est de l'ordre de 6%.

Tableau 15: Tests de significativité des variables du MLG "initial" du coût moyen

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Zone	7	33.22	<.0001
DU	8	15.16	0.1058
MT	1	0.35	0.4961
NA	2	2.89	0.2463
Profil	4	12.99	0.0086

Au seuil de confiance de 5%, les variables montant du voyage, nombre d'assurés ne sont pas discriminantes. La variable Durée pourrait être retenue au même titre que la variable Destination et la variable Profil de l'assuré. Nous effectuons alors la sélection automatique.

## 2.3 Sélection automatique des variables explicatives

Au titre de la sévérité des sinistres, la sélection automatique des facteurs explicatifs a été effectuée avec la procédure GLMSELECT du logiciel SAS au moyen d'une procédure itérative de type «

stepwise », En réalité, la procédure GLMSELECT effectue une régression linéaire entre les facteurs explicatifs et le montant d'indemnité. Or précédemment nous avons vu qu'il n'existait pas de lien linéaire entre la variable à expliquer et les variables explicatives (la validation de cette hypothèse sera faite grâce aux modèles additifs généralisés). Une astuce permettant d'utiliser proprement la procédure GLMSELECT serait de prendre le logarithme népérien du montant des sinistres Assistances. La loi du coût pouvant être, dans une moindre mesure, considérée comme log-normal, cela aura pour effet de recentrer les résidus. Et dans le même sens de créer un lien linéaire entre les facteurs explicatifs et le coût.

Au terme de la procédure automatique, les variables significativement représentatives du coût moyen sont alors :

- la destination géographique ;
- et le profil de l'assuré.

### 3. Analyse des résidus des régressions logistique et Gamma

Pour valider la pertinence du modèle, nous analysons les résidus.

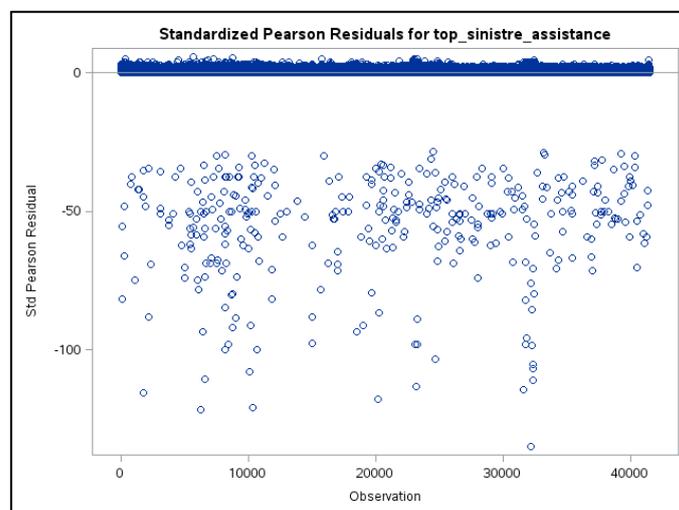
La procédure « GENMOD » de SAS introduit trois types de résidus construits sur l'écart entre l'observé et l'estimé :

- les résidus bruts : ils se définissent comme la différence entre l'observé et l'estimé  $r_i = y_i - \mu_i$  ;
- les résidus de Pearson : eux représentent la racine carrée de la ième contribution du Chi 2 de Pearson  $r_{Pi} = (y_i - \mu_i) \sqrt{\frac{\omega_i}{V(\mu_i)}}$  ;
- et les résidus de la Déviance quant à eux ils sont définis ici comme la racine carrée de la ième contribution de la déviance prenant le signe des résidus bruts  $r_{Di} = \sqrt{d_i} * \text{sign}(y_i - \mu_i)$ .

Avec  $y_i$  la ième réalisation de la variable à expliquer Y,  $\mu_i$  la ième estimation de Y,  $\omega_i$  le poids de chaque observation,  $d_i$  la déviance et  $V(\mu_i)$  la fonction variance introduite dans la théorie des MLG.

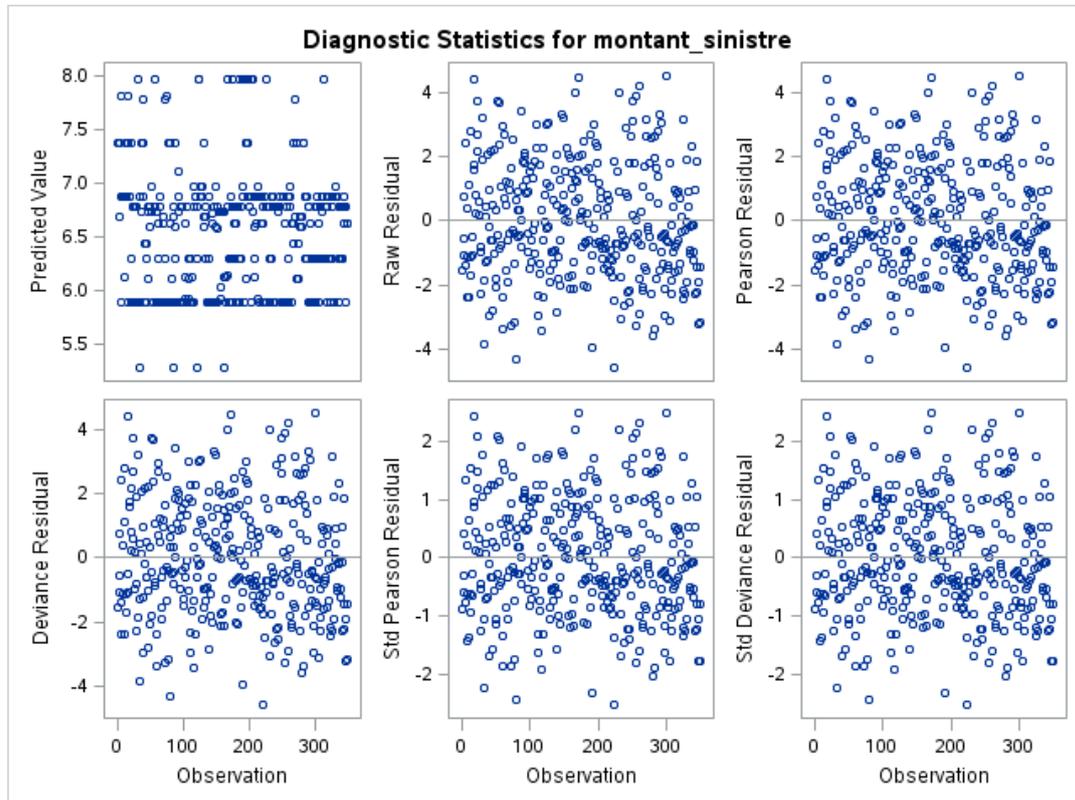
Nous aborderons ce point de manière graphique et nous nous concentrerons sur les résidus de Pearson normalisés.

Figure 23: Résidus sortis du modèle de régression du nombre de sinistres



La validité du modèle pour la fréquence des sinistres est jugée bonne ici car les résidus observés se situent autour de l'axe des abscisses et le nuage de points est de forme cylindrique autour de l'axe des abscisses.

Figure 24 : Résidus sortis du modèle de régression du coût moyen



Pour le coût moyen, le graphique ci-dessus ne met pas en évidence une structure particulière. L'asymétrie apparente dans les résidus de Pearson n'invalide pas le modèle Gamma tel qu'il a été construit. Les résidus de Pearson sont centrés avec une moyenne proche de 0.

## F. Modèle de tarification

L'objectif de cette partie est de rappeler les fondements théoriques de l'estimation de la prime pure. Le principe de base y est énoncé, ainsi que les hypothèses sous-jacentes.

## 1. Principe de base

Considérons une entreprise d'assurance Non-Vie, dont on supposera pour simplifier :

- Qu'elle se limite à la branche assistance ;
- Qu'elle commercialise un contrat unique mono-garantie (exemple : Assurance annulation, assistance ou assurance bagages)
- Qu'elle possède un portefeuille constitué de  $n$  risques

Notations :

- $N_i$  : variable aléatoire (v.a.) du nombre de sinistres survenant pendant une année sur le contrat n° $i$ , avec  $i \in \llbracket 1 \dots n \rrbracket$  ;
- $Y_{i,j}$  : v.a. du coût unitaire du  $j$ -ième sinistres sur le contrat n°  $i$ , avec  $i \in \llbracket 1 \dots n \rrbracket$  et  $j \in \llbracket 1 \dots N_i \rrbracket$  ;
- $X_i$  : v.a. du coût total des sinistres survenant pendant une année sur le contrat n° $i$ , avec  $i \in \llbracket 1 \dots n \rrbracket$ .

On peut alors écrire :

$$X_i = \sum_{j=1}^{N_i} Y_{i,j}.$$

Les deux hypothèses classiques sur les deux premières variables aléatoires sont :

- Indépendance et distribution identique pour les  $Y_{i,j}$  :  $\mathbb{E}[Y_{i,j}] = \mathbb{E}[Y]$  et  $\text{Var}[Y_{i,j}] = \text{Var}[Y]$  en notant la loi commune aux  $Y_{i,j}$  (v.a. du coût unitaire d'un sinistre) ;
- Indépendance des  $N_i$  et des  $Y_{i,j}$ .

Nous pouvons alors démontrer les relations suivantes :

- $\mathbb{E}[X_i] = \mathbb{E}[N_i]\mathbb{E}[Y_{i,j}]$  (1)
- $\text{Var}[X_i] = \text{Var}[N_i](\mathbb{E}[Y_{i,j}])^2 + \mathbb{E}[N_i]\text{Var}[Y_{i,j}]$  (2)

Dont nous déduisons, en posant :  $X = \sum_{i=1}^n X_i$  et  $N = \sum_{i=1}^n N_i$  :

- $\mathbb{E}[X] = \mathbb{E}[N]\mathbb{E}[Y]$  (1')
- $\text{Var}[X] = \text{Var}[N](\mathbb{E}[Y])^2 + \mathbb{E}[N]\text{Var}[Y]$  (2')

Les variables aléatoires  $N$  et  $Y$  sont généralement estimées par des estimateurs sans biais suivants :

- Fréquence :  $f^k = \frac{\text{nombre de sinistres}}{\text{exposition}}$  (sur une période donnée)
- Coût moyen :  $C^k = \frac{\text{coût total des sinistres}}{\text{nombre de sinistres}}$  (sur une période donnée).

Nous pouvons alors déterminer la prime pure comme la valeur probable du coût du sinistre, soit :

$$\pi = \mathbb{E}[X] = f * C$$

La modélisation ci-dessus est évidemment très simplifiée. En réalité, le contrat considéré comporte un certain nombre de garanties.

En reprenant les notations précédentes et en intégrant en indice  $k$  la garantie ( $k$  variant de 1 à  $G$  nombre total de garanties du contrat, en supposant que chaque contrat inclut l'ensemble des garanties, ce qui revient à supposer qu'elles sont toutes obligatoires – hypothèse pas nécessairement vraie dans la pratique), nous pouvons alors écrire :

$$\mathbb{E}[X_i^k] = \mathbb{E}[N_i^k] \mathbb{E}[Y_{i,j}^k]$$

Les variables aléatoires  $N^k$  et  $Y^k$  sont généralement estimées par les estimateurs sans biais suivants :

- Fréquence :  $f^k = \frac{\text{nombre de sinistres}^k}{\text{exposition}^k}$  (moyenne sur l'ensemble des années)
- Coût moyen :  $C^k = \frac{\text{coût total des sinistres}^k}{\text{nombre de sinistres}^k}$  (moyenne sur l'ensemble des années).

On peut alors déterminer la prime pure relative à la  $k$ -ième garantie comme la valeur probable du coût du sinistre sur cette garantie, soit :

$$\pi^k = \mathbb{E}[X^k] = f^k * C^k$$

Et l'on obtient par sommation la prime pure totale du contrat, soit :

$$\pi = \sum_{k=1}^G \pi^k = \sum_{k=1}^G f^k * C^k$$

Il apparaît clairement sur cette expression de la prime qu'en matière de tarification, il sera indispensable de disposer de statistiques fiables pour estimer les fréquences par garantie ainsi que les coûts moyens par garantie.

## 2. Modélisation de la fréquence

### 2.1 Modèle suite à la sélection des variables tarifaires

Cette partie recense l'ensemble des résultats d'estimation du modèle tarifaire au titre de la fréquence de sinistralité. C'est dans cette partie que nous donnerons les coefficients tarifaires estimés grâce à la régression logistique et les incertitudes d'estimation associées

A ce stade, nous rappelons que les variables tarifaires retenues sont les suivantes :

- Durée ;
- Profil ;
- Nombre d'assurés ;
- Destination.

Les trois tests rejettent l'hypothèse de la nullité de tous les coefficients et montrent que l'apport des variables explicatives permet de gagner en qualité par rapport au modèle minimal de manière significative.

Tableau 16 : Test de la significativité globale du modèle sans regroupement des modalités

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14325.9760	21	<.0001
Score	12785.9361	21	<.0001
Wald	11652.0836	21	<.0001

### Significativité des variables

Tableau 17: Significativité (type 3) des variables retenues pour la modélisation la fréquence

Type 3 Analysis of Effects			
Effect	DF	Wald	Pr > ChiSq
		Chi-Square	
Profil	4	2068.6692	<.0001
DU	8	5152.4191	<.0001
NA	2	908.3357	<.0001
Zone	7	842.3360	<.0001

Les variables sont toutes significatives avec un risque de première espèce quasi nul. D'autre part, nous pouvons constater que la variable « durée du voyage » est la plus discriminante, en effet elle possède la valeur de test la plus élevée (Khi 2 de Wald).

### Estimation du modèle de fréquence

En rassemblant l'ensemble des éléments décrits jusqu'à la partie 3, le modèle de tarification prend alors la forme suivante :

Pour un segment de risque ou case tarifaire  $i$ , la fréquence est donnée par la formule

$$\log \left[ \frac{\hat{\mathbb{P}}[\text{verser une indemnité}_i]}{1 - \hat{\mathbb{P}}[\text{verser une indemnité}_i]} \right] = \hat{\beta}_0 + \hat{\beta}_{\text{durée},i} * \text{durée}_i + \hat{\beta}_{\text{Profil},i} * \text{Profil}_i + \hat{\beta}_{\text{nb\_assuré},i} * \text{nb\_assuré}_i + \hat{\beta}_{\text{Destination},i} * \text{Destination}_i$$

Où les coefficients  $\hat{\beta}_0$ ,  $\hat{\beta}_{\text{durée}}$ ,  $\hat{\beta}_{\text{Profil}}$ ,  $\hat{\beta}_{\text{nb\_assuré}}$  &  $\hat{\beta}_{\text{Destination}}$  sont ceux estimés par la régression logistique.

### Estimation des coefficients avant regroupement des modalités.

- Choix des modalités de référence : le choix de la modalité de référence de chaque variable s'est porté sur la modalité ayant connu historiquement la plus grande fréquence de sinistres. Toutefois, ce choix n'influe en rien sur les résultats obtenus, il ne modifie que le nombre de coefficients significativement non nuls.

Tableau 18: Estimation du modèle de fréquence suite à la sélection automatique des variables

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald	Pr > ChiSq
					Chi-Square	
Intercept		1	-3.2743	0.0441	5414.53	<.0001
Profil	Affaire	1	-1.0269	0.0796	163.51	<.0001
	Etude	1	-1.3334	0.0376	1234.08	<.0001
	Globe trotter	1	-0.3159	0.0438	50.97	<.0001
	Visa Schengen	1	-0.0015	0.0434	0.0012	0.7836
Durée	du000-008	1	-4.0977	0.1708	566.22	<.0001
	du008-015	1	-3.1649	0.0597	2762.01	<.0001
	du015-022	1	-2.6463	0.0496	2796.36	<.0001
	du022-032	1	-2.5573	0.0487	2707.35	<.0001
	du032-062	1	-1.7362	0.0423	1655.72	<.0001
	du062-089	1	-1.1265	0.0385	841.44	<.0001
	du089-093	1	-0.6149	0.0332	335.19	<.0001
	du093-182	1	0.0591	0.0225	6.79	0.0121
Nombre d'assurés	na02-03	1	0.5418	0.0233	511.90	<.0001
	na03-inf	1	0.9500	0.0384	601.01	<.0001
Destination	Afrique	1	0.40513	0.0445	81.56	<.0001
	Amerique Autre	1	0.21351	0.0352	36.05	<.0001
	Amérique du nord	1	0.55506	0.0318	299.11	<.0001
	Asie	1	0.56749	0.0286	385.84	<.0001
	Europe hors France	1	-0.94853	0.0628	224.16	<.0001
	France	1	0.22264	0.0405	29.09	<.0001
	Reste du monde	1	-0.13959	0.0844	2.66	0.1058

Nous n'interpréterons les résultats qu'en terme de qualité relative d'explication de la variable cible (sinistre / pas de sinistre) par les variables explicatives.

Nous définissons le pouvoir discriminant de chaque variable explicative à partir des coefficients associés à ses modalités : on le considère égal à la différence entre le coefficient estimé le plus élevé et le coefficient estimé le plus faible. Plus les coefficients estimés sont dispersés, plus les probabilités estimées le sont, et plus les modalités de la variable considérée se distinguent les unes des autres et permettent d'expliquer les variations de la variable réponse étudiée.

Le pouvoir discriminant ainsi défini n'a qu'une valeur comparative entre les variables.

#### Les odds ratios

L'*odds ratio* également appelé rapport des chances ou risque relatif rapproché, est une mesure statistique exprimant le degré de dépendance entre nos variables. Il est souvent utilisé en inférence bayésienne et en régression logistique, et permet de mesurer l'effet d'un facteur.

Tableau 19: Odds ratio modèle de fréquence juste après la sélection automatique des variables

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald	
		Confidence Limits	
Profil Affaire vs Vacance	0.393	0.338	0.456
Profil Etude vs Vacance	0.299	0.279	0.321
Profil Globe trotter vs Vacance	0.752	0.693	0.817
Profil Visa Schengen vs Vacance	1.001	0.923	1.086
DU du000-008 vs du182-inf	0.024	0.018	0.033
DU du008-015 vs du182-inf	0.056	0.050	0.063
DU du015-022 vs du182-inf	0.090	0.082	0.099
DU du022-032 vs du182-inf	0.098	0.089	0.107
DU du032-062 vs du182-inf	0.207	0.191	0.224
DU du062-089 vs du182-inf	0.360	0.335	0.386
DU du089-093 vs du182-inf	0.573	0.538	0.610
DU du093-182 vs du182-inf	1.056	1.013	1.102
NA na02-03 vs na01-02	1.634	1.563	1.708
NA na03-inf vs na01-02	2.370	2.206	2.547
Zone Afrique vs Maghreb	1.446	1.330	1.571
Zone Amerique Autre vs Maghreb	1.214	1.137	1.297
Zone Amérique du nord vs Maghreb	1.655	1.560	1.756
Zone Asie vs Maghreb	1.683	1.595	1.775
Zone Europe hors France vs Maghreb	0.422	0.375	0.474
Zone France vs Maghreb	1.223	1.133	1.321
Zone Reste du monde vs Maghreb	0.933	0.797	1.093

Odds ratio =  $\{P(Y=1/X=1) \div P(Y=0 / X=1)\} / \{P(Y=1 / X=0) \div P(Y=0 / X=0)\}$

RR (risque relatif) =  $P(Y=1 / X=1) \div P(Y=1 / X=0)$

Si  $P(Y=1/X)$  est petite, le risque relatif est sensiblement égal au odds ratio. Dans ce cas  $P(Y=1/X1)$  et  $P(Y=1/X0)$  sont pratiquement égales et petites. Le produit  $P(Y=1/X1) \times P(Y=1/X0)$  est négligeable.

Odds ratio  $\approx$  RR =  $P(Y=1/X1) \div P(Y=1/X0)$

- Interprétation des *odds ratios* : Sur la variable « durée du voyage » : Un assuré qui souhaite voyager pour 3 mois (90 jours) à 2 fois moins de chance de se faire payer une indemnité qu'un voyageur souhaitant s'exiler pour une durée plus courte de 6 mois, toutes choses égales par ailleurs. Pour la variable « nombre d'assurés » ; un groupe de plus de 3 personnes a sa probabilité de survenance d'un sinistre 2.3 fois plus grande qu'un individu parti tout seul.

## 2.2 Modèle final

Le modèle final qui sera retenu pour prédire la fréquence de sinistre par classes de risque est le modèle pour lequel nous aurons regroupé toutes les modalités non-significatives.

Regroupements des modalités sous représentées coefficients

De nombreuses sous classes apparaissent non significatives (surlignées ci-dessus), des regroupements sont alors nécessaires. Pour cela, nous utiliserons à nouveau les résultats préliminaires d'analyse uni-variée (tri à plat). Nous décidons à ce titre de regrouper les classes qui ont des fréquences à peu près proches.

La démarche adoptée pour le regroupement des modalités est une démarche pas à pas de type *backward*. Elle consiste à éliminer une à une les variables les moins significatives du modèle jusqu'à obtenir un modèle où toutes les variables sont significatives. Dans notre cas, il ne s'agit pas exactement de variables mais de modalités.

Notre première tentative est celle qui nous a paru la plus vraisemblable. En ce sens, nous décidons donc de regrouper Maghreb avec le Reste du monde, Vacances avec Visa Schengen. L'agrégation de ces modalités nous donne les résultats suivants.

**Tableau 20: Estimation du modèle de fréquence final (après regroupement des modalités non significatives)**

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
<b>Intercept</b>		1	-2.9826	0.0384	6046.6661	<.0001
<b>Durée</b>	<b>du000-008</b>	1	-3.7226	0.1620	528.2299	<.0001
	<b>du008-015</b>	1	-2.8796	0.0549	2747.8596	<.0001
	<b>du015-022</b>	1	-2.4079	0.0450	2862.6576	<.0001
	<b>du022-032</b>	1	-2.3253	0.0445	2732.3887	<.0001
	<b>du032-062</b>	1	-1.5766	0.0388	1653.6086	<.0001
	<b>du062-089</b>	1	-1.0235	0.0364	792.7337	<.0001
	<b>du089-093</b>	1	-0.5577	0.0317	310.0171	<.0001
	<b>du093-182</b>	1	0.0544	0.0214	6.4487	0.0111
<b>Nombre d'assurés</b>	<b>na02-03</b>	1	0.4909	0.0226	470.9201	<.0001
	<b>na03-inf</b>	1	0.8628	0.0366	556.1602	<.0001
<b>Zone</b>	<b>Afrique</b>	1	0.3733	0.0420	79.1442	<.0001
	<b>Amerique Autre</b>	1	0.1986	0.0330	36.2093	<.0001
	<b>Amérique du nord</b>	1	0.5085	0.0296	295.1286	<.0001
	<b>Asie</b>	1	0.5251	0.0265	391.4982	<.0001
	<b>Europe hors France</b>	1	-0.8601	0.0596	208.5122	<.0001
	<b>France</b>	1	0.2075	0.0286	52.7619	<.0001
<b>Profil</b>	<b>Affaire</b>	1	-0.9341	0.0754	153.4647	<.0001
	<b>Etude</b>	1	-1.2070	0.0331	1327.3291	<.0001
	<b>Globe trotter</b>	1	-0.2805	0.0380	54.4866	<.0001

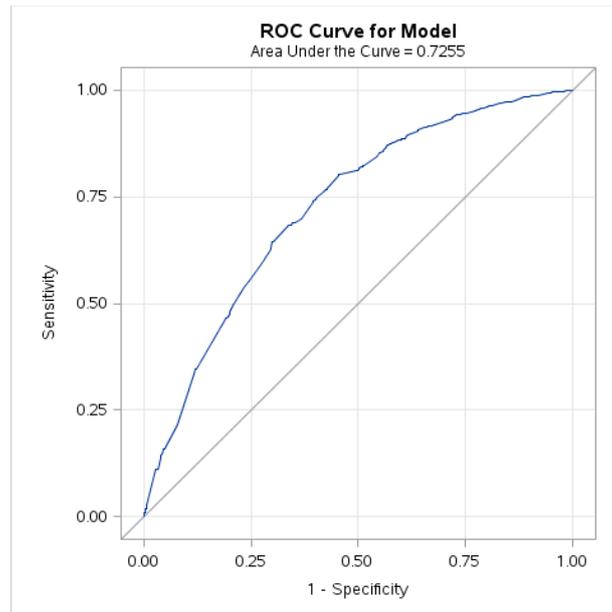
Les coefficients de régression montrent que toutes les durées de voyage inférieures à 3 mois sont moins risquées que la durée de voyage de référence (plus de 6 mois). Il apparaît également que les assurés ayant une durée de voyage comprise entre 3 et 6 mois sont plus susceptibles de générer un sinistre que les assurés de référence.

#### Performance et capacité prédictive du modèle

La valeur de l'aire sous la courbe ROC (73%) traduit une bonne association entre la probabilité prédite et la valeur de la variable explicative.

L'allure de la courbe ROC (figure 25), qui est largement au-dessus de la diagonale, confirme la performance du modèle. L'axe des abscisses correspond au pourcentage cumulé du nombre acquis classé par ordre décroissant de la fréquence de sinistralité observée empiriquement. L'axe des ordonnées correspond au pourcentage cumulé des fréquences de sinistres prédits par le modèle.

Figure 25 : Courbe ROC du modèle GLM sur la fréquence



### 3. Modélisation du « coût moyen »

#### 3.1 Construction du modèle

La procédure GENMOD de SAS va nous permettre de construire un modèle de prédiction du coût moyen à partir de la variable choisie précédemment.

Rappelons à ce stade que le modèle Gamma avec une fonction de lien logarithme népérien a été préféré au modèle normal avec une fonction de lien logarithme népérien du fait des problèmes de convergence que ce dernier présentait.

#### Modèle Gamma avec un lien logarithme népérien

En premier lieu, SAS nous donne quelques informations sur la procédure utilisée ainsi que quelques informations relatives à l'adéquation du modèle aux données. Le nombre de degrés de liberté est égal au nombre d'observations ( $n$ ) auquel nous soustrayons le nombre de paramètres ( $p$ ).

Nous nous intéressons ici à la déviance normalisée (*Scaled Deviance*) ainsi qu'au khi-deux de Pearson normalisé (*Scaled Pearson X2*). Le rapport de la déviance normalisée par le nombre de

degré de liberté n'est pas très supérieur à 1, ce qui se traduit par une qualité de la régression acceptable (Tableau 21)

**Tableau 21 : Critère permettant d'évaluer la qualité d'ajustement du modèle de régression du coût moyen**

Criterion	DF	Value	Value/DF
Deviance	416	1006.7547	2.3546
Scaled Deviance	416	220.7757	1.2232
Log Likelihood		-489.2229	
AIC (smaller is better)		982.4457	
BIC (smaller is better)		989.9761	

Pour être plus précis, nous pouvons néanmoins calculer la p-value du test du chi 2. Après calcul sous SAS, nous obtenons 0,24%. Compte tenu de la profondeur de notre échantillon, nous considérons cette valeur suffisamment grande pour admettre la validité du modèle.

L'estimation des paramètres du modèle

$$\ln(\mathbb{E}[\text{coût\_moyen}]) = \beta_0 + \beta_{\text{profil}} * \text{Profil} + \beta_{\text{zone\_geo}} * \text{zone\_geographique}$$

Après les critères d'adéquation du modèle aux données, on trouve l'estimation des paramètres du modèle avec leur erreur standard, les intervalles de confiance et les tests de Wald associés.

Enfin, on trouve les tests de Wald pour chaque modalité. On note que plus le paramètre estimé est faible et son erreur standard élevée, moins il est significatif.

**Tableau 22: Coefficients estimés par le modèle de régression**

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	6.6915	0.3804	5.9459	7.4371	309.41	<.0001
Zone	Afrique	1	1.4146	0.5532	0.3304	2.4988	6.54	0.0106
	Amérique Autre	1	-0.0797	0.5524	-1.1624	1.0029	0.02	0.0852
	Amérique du nord	1	-0.4310	0.4671	-1.3464	0.4845	0.85	0.0351
	Asie	1	0.0939	0.4539	-0.7956	0.9835	0.04	0.0360
	France	1	-0.2473	0.6533	-1.5276	1.0331	0.14	<.0001
	Maghreb	1	0.0349	0.4286	-0.8051	0.8749	0.01	<.0001
	Reste du monde	1	0.2805	0.7156	-1.1220	1.6831	0.15	<.0001
Profil	Vacance	1	0.5563	0.8284	-1.0674	2.1800	0.45	0.5019
	Etude	1	-0.1344	0.3679	-0.8554	0.5866	0.13	0.7148
	Globe trotter	1	0.9157	0.5434	-0.1493	1.9808	2.84	0.0920
	Visa Schengen	1	-0.9594	0.3980	-1.7395	-0.1792	5.81	0.0159
Scale		1	1.8184	0.0720	1.6826	1.9651		

Pour finaliser la construction du modèle et en assurer sa robustesse, nous devons regrouper certaines modalités superflues. En effet, les modalités sont qualifiées de superflues lorsque leurs p-value sont supérieures à 5%. L'objectif à première vue de ce tableau est de regrouper :

- Pour la variable Destination géographique :
  - Amérique Autre ;
- Pour la variable Profil des voyageurs :
  - Etude ;
  - Globe-trotter ;
  - Vacance ;

Les modalités de référence sont pour la variable Destination géographique la modalité Europe Hors France et pour la variable Profil des voyageurs, la modalité Affaire (Pour rappel, les modalités de référence jouent un rôle primordial dans l'évaluation de la significativité car la significativité est en effet évaluée par rapport à la modalité de référence). Autrement dit, si nous changions la modalité de référence certains coefficients qui apparaissaient significatifs dans le modèle actuel pourraient ne plus l'être.

Malgré la non-significativité de certaines modalités, les contraintes opérationnelles nous imposent un niveau de granularité particulier que nous tiendrons en compte. Les regroupements retenus sont alors : Le profil Vacance à regrouper avec le profil Globe-Trotter. Le profil Visa Schengen a quant à lui été rapproché du profil Etude. Pour ce qui est de la variable Destination, aucun regroupement n'a été réalisé mais justement une modification de la modalité de référence.

### 3.2 Modèle final pour le coût moyen

Le modèle retenu pour la suite est le suivant :

$$\ln(\mathbb{E}[\text{coût\_moyen}]) = \beta_0 + \beta_{\text{profil}} * \text{Profil} + \beta_{\text{zone\_geo}} * \text{zone\_geographique}$$

Tableau 23: Estimation du modèle final pour le coût moyen

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	8.0814	0.3233	7.9165	8.2462	766.96	<.0001
Zone	Amerique Autre	1	0.1631	0.0326	0.1465	0.1797	8.3	0.004
	Amérique du nord	1	0.1414	0.0283	0.1270	0.1558	12.44	0.0004
	Asie	1	0.1422	0.0284	0.1277	0.1567	20.05	<.0001
	Europe hors France	1	-0.4572	0.0914	-0.5038	-0.4106	19.67	<.0001
	France	1	-1.0659	0.2132	-1.1746	-0.9572	7.85	0.0051
	Maghreb	1	-0.6107	0.1221	-0.6730	-0.5484	9.71	0.0018
	Reste du monde	1	-1.1335	0.2267	-1.2491	-1.0178	4.59	0.0321
Profil	Affaire	1	-0.9513	0.5919	-1.2532	-0.6494	2.58	0.0108
	Visa Schengen & Etude	1	-0.5767	0.2300	-0.6940	-0.4594	6.29	0.0122
Scale		1	0.4299	0.0265	0.3810	0.4851		

En comparaison au Tableau 22, le signe des coefficients évoluent en sens contraire du fait du changement de la modalité de référence.

**En résumé suite à nos regroupements nous choisirons comme modèle :**

- La variable « Destination géographique » avec les modalités :
  - o Afrique (référence) ;
  - o Amérique Autre ;
  - o Amérique du Nord ;
  - o Asie ;
  - o Europe hors France ;
  - o France ;
  - o Maghreb ;
  - o Reste du monde.
- La variable « motif du voyage » avec les modalités :
  - o Affaire ;
  - o Vacances & Globe-Trotter (référence) ;
  - o Visa Schengen & Etude.

#### 4. Comparaison entre les coefficients tarifaires observés et ceux prédits

Afin de valider définitivement la pertinence des coefficients obtenus, dans cette partie, nous confrontons les résultats des estimations des GLM grâce aux observations empiriques. Nous étudions également graphiquement les intervalles de confiance de chaque paramètre estimé. En effet, en pratique, il est souvent préférable de considérer la représentation graphique de l'intervalle de confiance plutôt que son interprétation numérique.

##### Lecture des graphiques

- Base 100 : Les graphiques de comparaison des coefficients estimés par rapport aux observés sont représentés en base 100. C'est-à-dire que la fréquence de sinistre ou le coût moyen de chaque modalité est rapporté à la fréquence / coût moyen de la modalité de référence
 
$$coeff\_base\_100 (modalité\ k) = \frac{coeff (modalité\ k)}{coeff (modalité\ reference)}$$
- Les modèles de fréquence et de coût moyen sont multiplicatifs donc les coefficients s'obtiennent en faisant :
  - o Pour le coût moyen  $\exp(\beta_0 + \beta_{modalite\_k})$  ;
  - o Pour la fréquence  $1 - \exp(-\exp(\beta_0 + \beta_{modalite\_k}))$  ;
- Code couleur : la ligne rouge correspond à l'analyse uni-variée empirique, la ligne bleue foncé correspond aux prédictions et la bande bleue clair indique l'intervalle de confiance de l'estimation du paramètre. En faisant l'hypothèse que le modèle est approprié et correct, cela signifie que les données suggèrent que la vraie relativité pour chaque niveau du facteur se situe entre les deux lignes bleues avec 95% de certitude.

#### 4.1 Etude sur la fréquence

Pour la fréquence de sinistres, les deux bandes sont donc très proches, indiquant une bonne qualité de l'estimation des paramètres.

Cependant la ligne rouge correspondant à l'analyse uni-variée diffère des estimations du MLG. Cela signifie qu'il existe des corrélations significatives entre les facteurs de tarification. C'est plutôt bon signe car cela valide la pertinence de l'utilisation des modèles linéaires généralisés.

Figure 26: Comparaison entre fréquences observée et prédite pour la variable "durée du voyage"

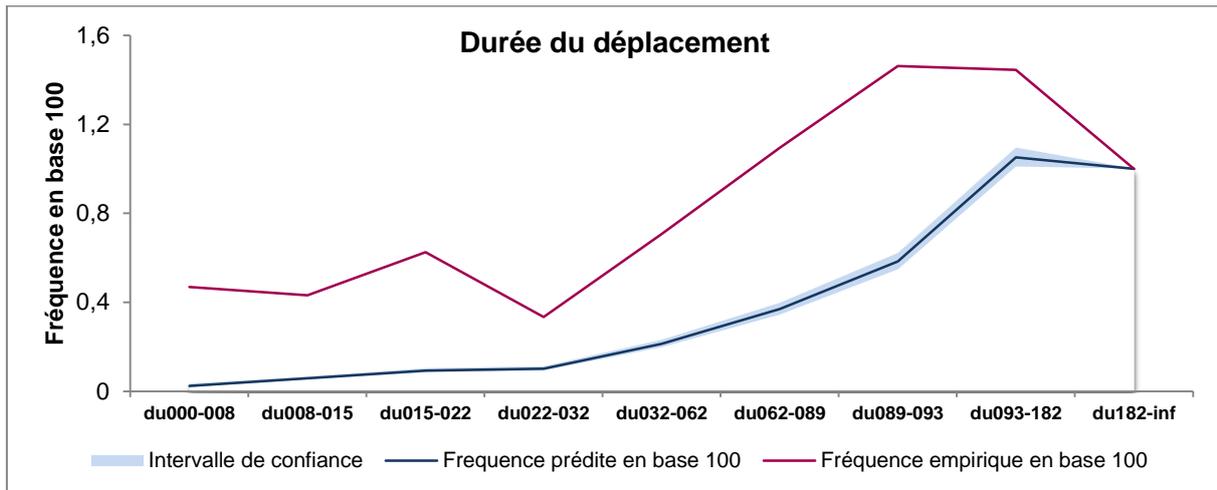


Figure 27: Comparaison entre fréquences réelle et prédites pour la variable "nombre d'assurés "

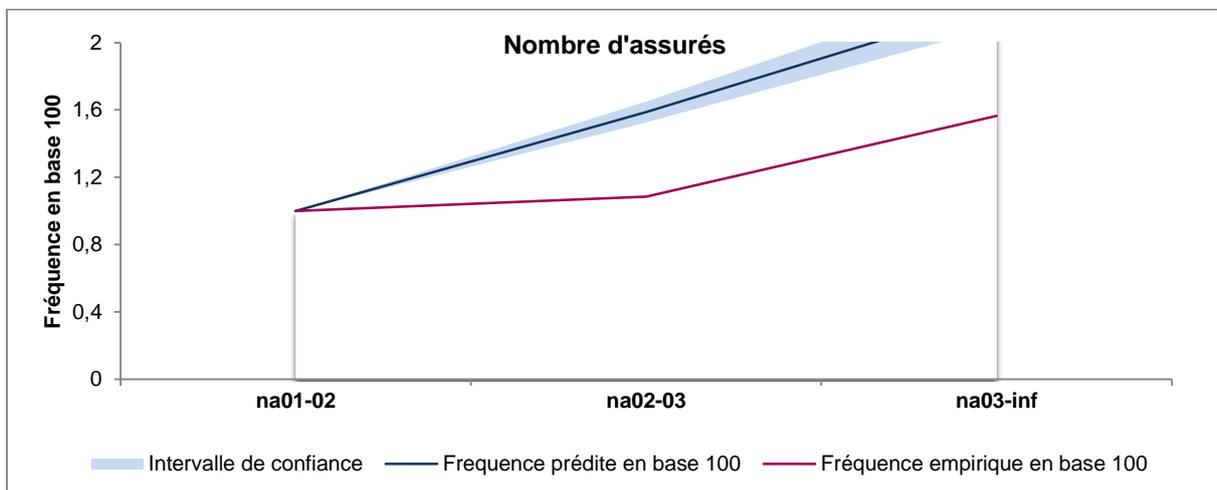


Figure 28: Comparaison entre fréquences réelle et prédites pour la variable "Profil"

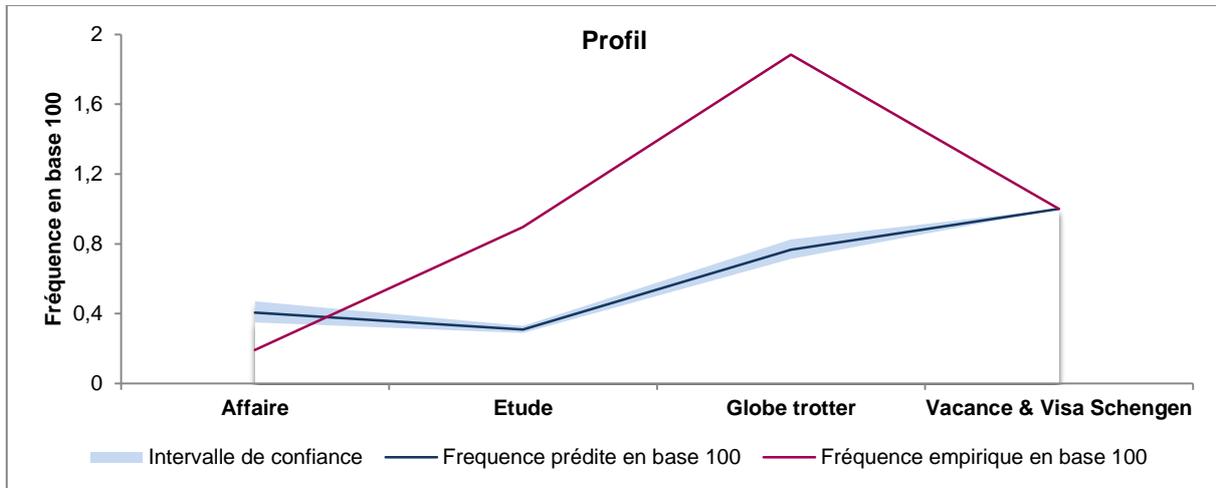
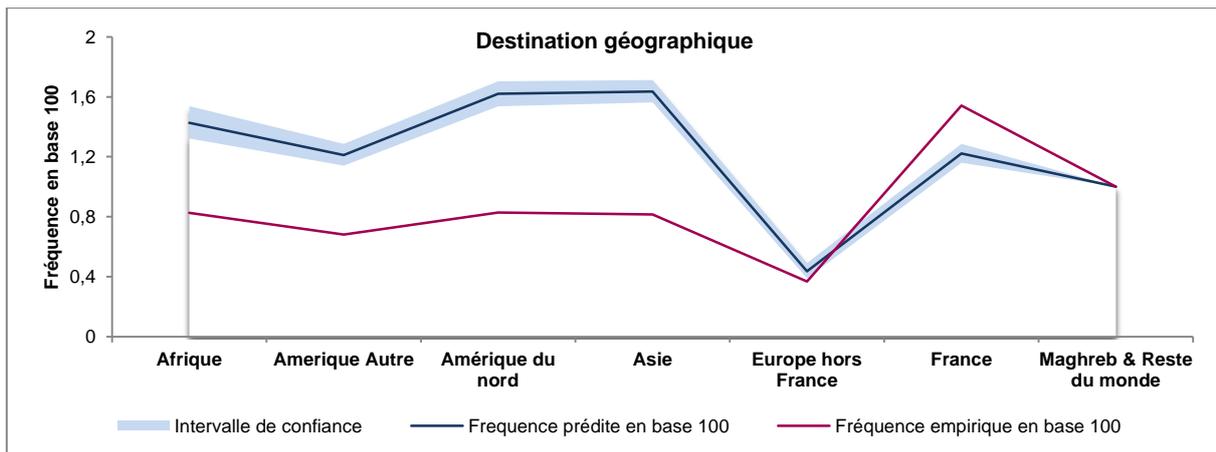


Figure 29: Comparaison entre fréquences réelle et prédites pour la variable "Destination géographique"



## 4.2 Etude sur le coût moyen

Pour le coût moyen, les deux bandes seront très éloignées, indiquant une grande incertitude dans l'estimation des paramètres, lorsque le niveau d'exposition est trop faible.

La courbe des coûts moyens estimés et celle des observés sont quasiment confondues. Cette faible différence entre les paramètres issus de l'analyse uni-variée (en rouge sur les graphiques) et ceux provenant de la résolution du MLG permet de mesurer la faible importance des interactions entre les facteurs. Autrement dit, nous aurions pu estimer ces coefficients sans l'aide d'un modèle de régression.

L'incertitude des estimations est principalement due au faible nombre de points utilisés pour calibrer le modèle de tarification.

Figure 30: représentation des coefficients MLG pour la variable "Destination géographique"

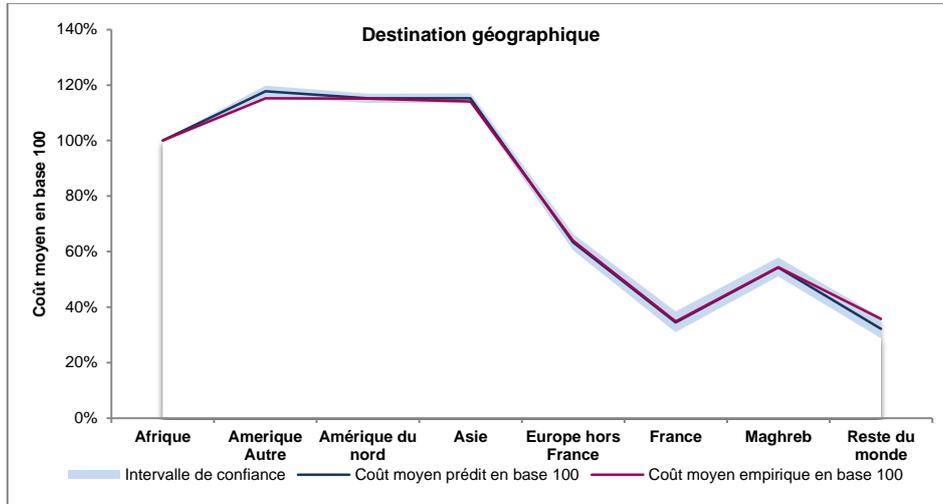
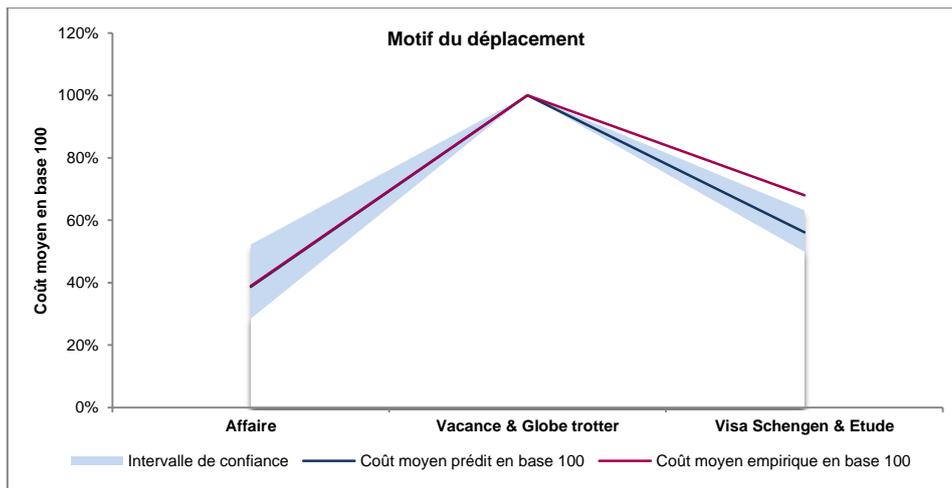


Figure 31: représentation des coefficients MLG pour la variable "Motif du voyage"



Nous concluons donc à la pertinence globale de l'utilisation des GLM avec des niveaux d'intervalle de confiance acceptable pour l'établissement de la tarification via les GLM.

## 5. Modèles tarifaires

Avant la présentation de notre grille tarifaire, il est nécessaire de présenter brièvement la méthode de tarification actuelle qui a servi à faire la tarification lors de la refonte de 2010. Et ensuite de résumer les avantages et les inconvénients de la méthode GLM lors de la détermination de la prime pure en assurance non-vie.

### 5.1 Modèles tarifaire actuel

La tarification actuelle, porte sur une méthode déterministe qui consiste à observer les sinistres et les polices de l'année n-1 pour chacune des branches de sinistre (bagage, annulation, frais médicaux etc...) et pour chaque produit vendu sur le site Mondial-assistance.fr. Une fois cette observation faite, un calcul de coût moyen et de fréquence était effectué pour chaque segment (produit et garantie) en prenant en compte certaines variables clés lorsqu'elles étaient disponibles (âge, destination, etc...).

L'approche utilisée par les services de distribution des produits consiste donc à déterminer des coefficients tarifaires à partir des analyses uni-variées menées par ailleurs. L'avantage de cette approche est qu'elle prend en compte toutes les variables à notre disposition. Il suffit juste de choisir pour chaque variable une modalité de référence qui servira à établir le coefficient tarifaire relatif et la base tarifaire.

Afin de pouvoir comparer les différents modèles tarifaires énoncés, nous rappelons que nous nous sommes restreints aux variables résultantes de la régression stepwise avec les regroupements effectués dans la partie précédente.

Les coefficients tarifaires obtenus par approche directe sont présentés ci-dessous pour le coût moyen et en annexe le tableau des fréquences et coefficients suivant la durée défini et validé avec le service souscription suivant nos regroupements effectués lors de la partie précédente (tableau 35) :

**Tableau 24 : Coefficients tarifaires pour le coût moyen par approche directe sur un modèle à risque proportionnel**

	Coût moyen empirique en base 100		Coût moyen empirique
<b>Afrique</b>	100%	<i>Affaire</i>	39%
<i>Amérique Autre</i>	115%	<b>Vacance &amp; Globe-trotter</b>	100%
<i>Amérique du nord</i>	115%	<i>Visa Schengen &amp; Etude</i>	68%
<i>Asie</i>	114%		
<i>Europe hors France</i>	64%		
<i>France</i>	35%		
<i>Maghreb</i>	54%		
<i>Reste du monde</i>	36%		

La base est égale à 2 985€

Pour déterminer la base, nous prenons le coût moyen du croisement de toutes les modalités prises comme référence. Les modalités de référence sont les mêmes que celles retenues pour les MLG de la partie précédente, elles sont matérialisées en gras.

### Calcul de la prime pure: Cas pratique

La prime pure annuelle est obtenue en faisant le produit des paramètres de la fréquence par le coût moyen. La prime pure s'obtient alors par application de la formule suivante :

$$Prime\_Pure_i = (Base^{freq.} * \hat{\beta}_1^{freq.}) * (Base^{cout} * \hat{\beta}_1^{cout} * Z_{i,1} * \dots * \hat{\beta}_q^{cout} * Z_{i,q})$$

Dans cette formule, le p désigne le nombre de variables explicatives sélectionnées pour la fréquence q pour le coût moyen,  $\hat{\beta}_p^{freq.}$  représente le vecteur des estimations uni-variées des paramètres pour chacune des modalités de la variable  $X_p$  et  $X_{i,p}$  désigne le vecteur où la modalité de la variable p correspondant à l'individu i vaut 1 et les autres coordonnées valent 0.  $Base^{freq.}$  est la fréquence de sinistre du croisement des modalités de référence de toutes les variables retenues pour la fréquence et t la durée maximale du voyage.

Pour donner une indication sur l'utilisation des grilles tarifaires, nous allons prendre un exemple concret.

Un assuré célibataire qui souhaiterait se rendre en vacances en Afrique pour une durée maximale d'une semaine paierait une prime pure de :

$$\left\{ \begin{array}{l} Prime\_Pure_{Afri/du089-093/na01-02/vacs} = (0.48\% * 0.6) * (2\ 985 * 100\% * 100\%) \\ = 0.288\% * 2\ 985\text{€} \\ = 8.64\ \text{€} \end{array} \right.$$

Afin d'apprécier l'ancienne méthode de tarification, un exemple comparatif sera illustré plus loin lors de l'établissement des différents modèles.

Cette méthode diffère de celle que nous proposons dans ce mémoire. En effet, nous tenterons de proposer une tarification regroupant l'ensemble des produits suivant la formule : Assistance.

## 5.2 Modèle tarifaire basé sur les GLM

Le GLM est un modèle paramétrique et bien connu en sciences actuarielles. Il permet de réduire le temps de calcul des algorithmes de résolution utilisés pour l'estimation des paramètres du modèle qui motive donc sa large utilisation en tarification. Ce modèle pourrait donc être utilisé au quotidien par les services de la direction technique de MAF afin de travailler sur des estimations jugées robustes.

Toutefois, nous devons prendre en compte que les prévisions de prime pure découlant du GLM dépendent fortement des données. Les sorties du modèle peuvent être biaisées par des données aberrantes. En effet, un sinistre d'un montant significatif pourrait entraîner par exemple la surtarification d'une classe tarifaire dans le cas d'un mauvais écrêtement des sinistres.

De plus, nous avons vu aussi que les GLM nous impose de faire des hypothèses sur la forme de la loi conditionnelle de la variable expliquée Y en fonction des explicatives. Ces hypothèses peuvent s'avérer fausses et nous exposer à un risque de modèle.

En outre, le GLM ne suit pas le cycle économique, donc les ajustements tarifaires devront être étudiés en amont.

Afin de proposer une tarification, via les GLM, nous débuterons par une description de notre modèle de calcul de la fréquence et du coût moyen.

### Grille de fréquence

Les coefficients ci-dessus issus des prédictions de la régression logistique nous permettent de calculer la probabilité d'avoir un sinistre dans l'année. Nous calculons cette probabilité grâce aux formules ci-dessous :

$$\left\{ \begin{array}{l} \text{Probabilité } (i) = \frac{e^{\hat{\beta}_0} e^{\hat{\beta}_{Duree} X_{Duree}^i} e^{\hat{\beta}_{Nb\ Ass} X_{Nb\ Ass}^i} e^{\hat{\beta}_{Dest} X_{Dest}^i} e^{\hat{\beta}_{Profil} X_{Profil}^i}}{1 + e^{\hat{\beta}_0} e^{\hat{\beta}_{Duree} X_{Duree}^i} e^{\hat{\beta}_{Nb\ Ass} X_{Nb\ Ass}^i} e^{\hat{\beta}_{Dest} X_{Dest}^i} e^{\hat{\beta}_{Profil} X_{Profil}^i}} \\ \text{IC} + 95\% (i) = \frac{e^{\hat{\beta}_0^+} e^{\hat{\beta}_0^+ \hat{\beta}_{Duree} X_{Duree}^i} e^{\hat{\beta}_0^+ \hat{\beta}_{Nb\ Ass} X_{Nb\ Ass}^i} e^{\hat{\beta}_0^+ \hat{\beta}_{Dest} X_{Dest}^i} e^{\hat{\beta}_0^+ \hat{\beta}_{Profil} X_{Profil}^i}}{1 + e^{\hat{\beta}_0^+} e^{\hat{\beta}_0^+ \hat{\beta}_{Duree} X_{Duree}^i} e^{\hat{\beta}_0^+ \hat{\beta}_{Nb\ Ass} X_{Nb\ Ass}^i} e^{\hat{\beta}_0^+ \hat{\beta}_{Dest} X_{Dest}^i} e^{\hat{\beta}_0^+ \hat{\beta}_{Profil} X_{Profil}^i}} \\ \text{IC} - 95\% (i) = \frac{e^{\hat{\beta}_0^-} e^{\hat{\beta}_0^- \hat{\beta}_{Duree} X_{Duree}^i} e^{\hat{\beta}_0^- \hat{\beta}_{Nb\ Ass} X_{Nb\ Ass}^i} e^{\hat{\beta}_0^- \hat{\beta}_{Dest} X_{Dest}^i} e^{\hat{\beta}_0^- \hat{\beta}_{Profil} X_{Profil}^i}}{1 + e^{\hat{\beta}_0^-} e^{\hat{\beta}_0^- \hat{\beta}_{Duree} X_{Duree}^i} e^{\hat{\beta}_0^- \hat{\beta}_{Nb\ Ass} X_{Nb\ Ass}^i} e^{\hat{\beta}_0^- \hat{\beta}_{Dest} X_{Dest}^i} e^{\hat{\beta}_0^- \hat{\beta}_{Profil} X_{Profil}^i}} \end{array} \right.$$

Avec  $i$  la case tarifaire (exemple durée inférieure à 8j x 1 assure x Afrique x Affaire)  $\hat{\beta}$  les estimateurs issus de la régression logistique,  $\hat{\beta}^-$  la borne basse de l'intervalle de confiance à 95% associé à l'estimateur et  $\hat{\beta}^+$  la borne haute de l'intervalle de confiance à 95% associé à l'estimateur.

Le tableau (grille de tarification de la fréquence) recensant toutes les probabilités d'avoir un sinistre dans l'année et par case tarifaire est donné en annexe (tableau 37).

### Grille des coûts moyens

Les coûts présentés dans le tableau 25 sont issus des prédictions des GLM.

Pour obtenir ces chiffres, nous effectuons les calculs suivants :

$$\left\{ \begin{array}{l} \text{Coût moyen Prédit}(i) = e^{\hat{\beta}_0} e^{\hat{\beta}_{Destination} X_{Destination}^i} e^{\hat{\beta}_{Profil} X_{Profil}^i} \\ \text{IC} - 95\% = e^{\hat{\beta}_0^-} e^{\hat{\beta}_0^- \hat{\beta}_{Destination} X_{Destination}^i} e^{\hat{\beta}_0^- \hat{\beta}_{Profil} X_{Profil}^i} \\ \text{IC} + 95\% = e^{\hat{\beta}_0^+} e^{\hat{\beta}_0^+ \hat{\beta}_{Destination} X_{Destination}^i} e^{\hat{\beta}_0^+ \hat{\beta}_{Profil} X_{Profil}^i} \end{array} \right.$$

Avec  $i$  la case tarifaire (exemple Afrique x Affaire),  $\hat{\beta}$  les estimateurs issus de la régression Gamma,  $\hat{\beta}^-$  la borne basse de l'intervalle de confiance à 95% associé à l'estimateur et  $\hat{\beta}^+$  la borne haute de l'intervalle de confiance à 95% associé à l'estimateur.

Tableau 25 : Grille de tarification du coût moyen

Destination	Profil	Coût moyen prédict	IC-95%	IC+95%
Afrique	Affaire	1 249 €	783 €	1 992 €
	Visa Schengen & Etude	1 817 €	1 370 €	2 409 €
	Vacance & Globe-trotter	3 234 €	2 742 €	3 813 €
Amerique Autre	Affaire	1 470 €	907 €	2 384 €
	Visa Schengen & Etude	2 138 €	1 586 €	2 883 €
	Vacance & Globe-trotter	3 807 €	3 175 €	4 564 €
Amérique du nord	Affaire	1 439 €	889 €	2 328 €
	Visa Schengen & Etude	2 092 €	2 092 €	2 815 €
	Vacance & Globe-trotter	3 725 €	3 114 €	4 456 €
Asie	Affaire	1 440 €	890 €	2 330 €
	Visa Schengen & Etude	2 094 €	1 557 €	2 817 €
	Vacance & Globe-trotter	3 728 €	3 116 €	4 460 €
Europe hors France	Affaire	791 €	473 €	1 321 €
	Visa Schengen & Etude	1 150 €	828 €	1 598 €
	Vacance & Globe-trotter	2 047 €	1 657 €	2 529 €
France	Affaire	430 €	242 €	765 €
	Vacance & Globe-trotter	626 €	423 €	925 €
	Visa Schengen & Etude	1 114 €	847 €	1 464 €
Maghreb	Affaire	678 €	400 €	1 151 €
	Visa Schengen & Etude	986 €	699 €	1 392 €
	Vacance & Globe-trotter	1 756 €	1 399 €	2 203 €
Reste du monde	Affaire	402 €	225 €	720 €
	Visa Schengen & Etude	585 €	393 €	870 €
	Vacance & Globe-trotter	1 041 €	786 €	1 378 €

### Calcul de la prime pure : Cas pratique

La prime pure annuelle est obtenue en faisant le produit des paramètres de la fréquence par le coût moyen. Elle fait ainsi appel à tous les paramètres  $\beta$  (coefficients tarifaires) estimés grâce aux MLG. La prime pure s'obtient alors par application de la formule suivante :

$$Prime\_Pure_i = \left[ \frac{e^{(\hat{\beta}_0^{freq.} + \hat{\beta}_1^{freq.} * X_{i,1} + \dots + \hat{\beta}_p^{freq.} * X_{i,p})}}{1 + e^{(\hat{\beta}_0^{freq.} + \hat{\beta}_1^{freq.} * X_{i,1} + \dots + \hat{\beta}_p^{freq.} * X_{i,p})}} \right] e^{(\hat{\beta}_0^{cout} + \hat{\beta}_1^{cout} * Z_{i,1} + \dots + \hat{\beta}_q^{cout} * Z_{i,q})}$$

Dans cette formule, le p désigne le nombre de variables explicatives sélectionnées pour la fréquence et q pour le coût moyen,  $\hat{\beta}_p^{freq.}$  représente le vecteur des estimations des paramètres pour chacune des modalités de la variable  $X_p$  et  $X_{i,p}$  désigne le vecteur où la modalité de la variable p correspondant à l'individu i vaut 1 et les autres coordonnées valent 0.

Pour donner une indication sur l'utilisation des grilles tarifaires, nous allons prendre un exemple concret.

Un assuré célibataire qui souhaiterait se rendre en vacances en Afrique pour une durée maximale d'une semaine paierait une prime pure de :

$$\left\{ \begin{array}{l} Prime\_Pure_{Afri/du000-008/na01-02/vacs} \\ \\ \end{array} \right. = \left[ \frac{e^{(-2,98-3,72+0,37)}}{1 + e^{(-2,98-3,72+0,37)}} \right] e^{(8,08)}$$

$$= 0,18\% * 3\,234\text{€}$$

$$= 5,82 \text{ €}$$

Nous compléterons par la suite cette grille, afin de discuter avec le service souscription et marketing de la pertinence des primes pures calculées.

### 5.3 Modèle tarifaire GLM avec utilisation d'une variable Offset

Jusqu'ici, nous avons analysé la fréquence de sinistralité comme étant une fonction monotone de la durée du voyage. Par ailleurs, nous avons en effet remarqué que plus la durée du voyage est grande plus la probabilité d'avoir un sinistre devenait grande.

L'objectif de cette partie est donc d'arriver à expliciter un modèle tarifaire qui prenne en entrée la durée du voyage avec les autres variables tarifaires obtenues précédemment.

#### Avec l'utilisation des MLG

La prise en compte de la linéarité du risque de fréquence en fonction de la durée du voyage peut se faire via les GLM. Pour cela, il suffit de paramétrer le programme d'exécution en spécifiant la variable *offset* comme étant la durée du voyage rapportée à 360 jours.

Le modèle linéaire généralisé prend la forme suivante :

$$\log \left[ \frac{\mathbb{P}[\text{verser une indemnité}_i]}{1 - \mathbb{P}[\text{verser une indemnité}_i]} \right] = \log \left( \frac{\text{Durée}}{360} \right)$$

$$+ \hat{\beta}_0$$

$$+ \hat{\beta}_{\text{profil},i} * \text{profil}_i$$

$$+ \hat{\beta}_{\text{nb\_assuré},i} * \text{nb\_assuré}_i$$

$$+ \hat{\beta}_{\text{Destination},i} * \text{Destination}_i$$

Tableau 26 : Coefficients estimés par le modèle GLM intégrant le logarithme de la durée en variable offset

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		-2.1900	0.1315	277.2536	<.0001	<.0001
Nombre d'assurés	na02-03	0.4981	0.1460	11.6298	0.0006	0.0283
	na03-inf	1.1858	0.1739	46.5175	<.0001	<.0001
Zone	Afrique	0.1163	0.2496	0.2170	0.6413	0.0036
	Amerique Autre	-0.1421	0.2503	0.3224	0.5702	0.0004
	Amérique du nord	0.0256	0.1896	0.0183	0.8925	<.0001
	Asie	0.0405	0.1757	0.0532	0.8176	<.0001
	Europe hors France	-0.3900	0.3376	1.3346	0.2480	<.0001
	France	0.1124	0.1602	0.4925	0.4828	<.0001
Profil	Affaire	-0.8103	0.3874	4.3759	0.0365	0.0017
	Etude	-1.1127	0.2045	29.6023	<.0001	0.0014
	Globe trotter	-0.4907	0.2597	3.5706	0.0588	<.0001
Ln (Durée / 360)		1.0000	0	.	.	.

### Calcul de la prime pure avec une approche MLG en présence d'une variable offset : Cas pratique

La prime pure annuelle est obtenue en faisant le produit des paramètres de la fréquence par le coût moyen. Elle fait ainsi appel à tous les paramètres  $\beta$  (coefficients tarifaires) estimés grâce aux MLG. La prime pure s'obtient alors par application de la formule suivante :

$$Prime\_Pure_i = \left[ \frac{\frac{Durée}{360} \times e^{(\hat{\beta}_0^{freq.} + \hat{\beta}_1^{freq.} * X_{i,1} + \dots + \hat{\beta}_p^{freq.} * X_{i,p})}}{1 + \frac{Durée}{360} \times e^{(\frac{Durée}{360} + \hat{\beta}_0^{freq.} + \hat{\beta}_1^{freq.} * X_{i,1} + \dots + \hat{\beta}_p^{freq.} * X_{i,p})}} \right] e^{(\hat{\beta}_0^{cout} + \hat{\beta}_1^{cout} * Z_{i,1} + \dots + \hat{\beta}_q^{cout} * Z_{i,q})}$$

Dans cette formule, le p désigne le nombre de variables explicatives sélectionnées pour la fréquence et q pour le coût moyen,  $\hat{\beta}_p^{freq.}$  représente le vecteur des estimations des paramètres pour chacune des modalités de la variable  $X_p$  et  $X_{i,p}$  désigne le vecteur où la modalité de la variable p correspondant à l'individu i vaut 1 et les autres coordonnées valent 0.

Pour donner une indication sur l'utilisation des grilles tarifaires, nous allons prendre un exemple concret.

Un assuré célibataire qui souhaiterait se rendre en vacances en Afrique pour une durée maximale d'une semaine paierait une prime pure de :

$$\left\{ \begin{aligned} Prime\_Pure_{Afri/du000-008/na01-02/vacs} &= \left[ \frac{\frac{7}{360} \times e^{(-2,19+0,12)}}{1 + \frac{7}{360} \times e^{(-2,19+0,12)}} \right] e^{(8,08)} \\ &= 0,24\% * 3\,234\text{€} \\ &= 7,89 \text{€} \end{aligned} \right.$$

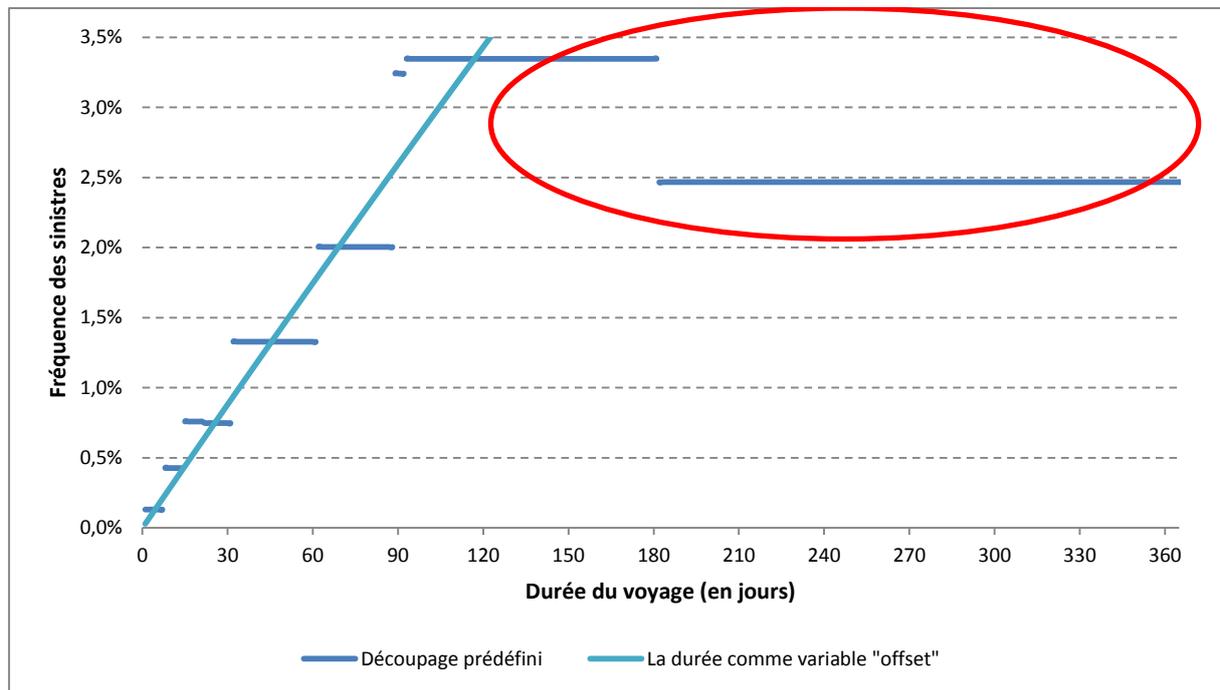
## 5.4 Comparaison entre les 3 modèles de tarification évoqués jusqu'ici

Jusqu'ici nous avons implémenté 3 modèles tarifaires. Le premier fait appel aux résultats des analyses uni-variées et s'apparente davantage au modèle actuellement en place. La deuxième repose sur une utilisation exclusive des MLG. Et enfin, le dernier modèle est une combinaison des deux premiers modèles avec notamment un paramétrage spécifique de la variable durée du voyage.

Dans la Figure 32, nous illustrons l'influence de la durée du voyage sur la fréquence suivant les alternatives de modélisation décrites ci-dessus. Ce graphique permet également de mieux appréhender les écarts de primes pures observées lors de la mise en œuvre des modèles. Nous mettons en avant également les avantages et inconvénients d'implémentation de chaque alternative.

- **Alternative 1 (Modèle actuel) / 8,64 €** : bien qu'il soit facile à implémenter, il n'est validé par aucun standard statistique. Il présente l'avantage de prendre en compte l'évolution de la sinistralité et de la structure du portefeuille années après années. En effet, la mise à jour (au sens réajustement) d'un modèle tarifaire GLM requiert quant à elle la prise en compte des nouvelles années présentes dans l'historique.  
De plus, au travers de cette approche, nous adoptons une démarche similaire à celle actuellement en vigueur au sein de la direction technique. C'est-à-dire qu'à partir des statistiques directement extraites des bases de données, nous élaborons la tarification. Néanmoins, une particularité demeure. En effet, nous introduisons une dépendance linéaire entre la fréquence de sinistres et la durée du voyage très souvent regroupées de façon intuitive ou commerciale, parfois indépendamment d'une étude fine de notre sinistralité. Les alternatives suivantes nous permettent une segmentation avec une meilleure justification technique
- **Alternative 2 (GLM pur / 5,82 €)** : ce modèle s'appuie fondamentalement sur un GLM multivarié. Nous avons pu voir qu'il demande un travail précis de sélection des variables, des tests d'hypothèses afin de trouver un modèle qui s'ajustera au mieux aux données. Il présente le principal inconvénient de mêler des durées de voyage très disparates. En particulier, nous attribuons la même fréquence de sinistres à un assuré en voyage pour 6 mois qu'à un assuré en voyage pour 2 ans. La courbe en escalier de la Figure 32 illustre bien le phénomène.  
En définitive, nous observons une croissance linéaire de la fréquence des sinistres en fonction de la durée du voyage sauf pour les durées comprises entre 3 mois et 3 ans. Ce constat a donc motivé la troisième alternative.
- **Alternative 3 (risque proportionnel à la durée avec MLG) / 7,89 €** : cette alternative a pour principal objectif de palier aux lacunes de l'alternative précédente. Elle consiste à exécuter un modèle linéaire généralisé avec la durée du voyage prise comme variable *offset*, c'est-à-dire comme une variable explicative pour laquelle aucun coefficient de régression n'est estimé. L'introduction de cette variable *offset* nous permettra d'estimer les coefficients des autres variables en tenant compte des contraintes tarifaires liées à la durée du voyage (prix croissant suivant la durée du séjour).

Figure 32 : Différentes modélisations de la fréquence de sinistres en fonction de la durée du voyage



L'hypothèse de linéarité de la fréquence des sinistres peut-être remise en question à la vue de la courbe en escalier de la Figure 32. C'est pour cela qu'une quatrième approche alternative sera explorée afin de valider ou de faire diminuer le risque de notre modèle. En effet, nous analyserons dans la dernière partie via les GAM la forme réelle de la relation entre chaque variable continue et la réponse. Les hypothèses reposant sur l'évaluation de la prime pure étant moins restrictives que pour le GLM. Différentes courbes du modèle nous permettrons de proposer un meilleur découpage de la variable continue par le lissage de la relation entre cette variable et la réponse.

## G. Méthode de tarification alternative : les Modèles Additifs Généralisés (GAM)

Les modèles additifs généralisés (GAM) ont été introduits dans les années 90 par Hastie et Tibshirani (1986) grâce à la combinaison des modèles linéaires et des modèles additifs. Ainsi, les modèles additifs généralisés résultent de l'association de deux méthodes efficaces et puissantes que sont : (i) la famille exponentielle qui permet de travailler avec plusieurs distributions de variable réponse et (ii) les modèles additifs qui eux offrent une souplesse dans l'hypothèse de linéarité des prédicteurs.

Dans le but de proposer une meilleure segmentation de notre risque et de regrouper les modalités des variables continues de manière optimale, nous nous servirons des résultats des GAM pour valider ou à minima critiquer les choix effectués pendant la procédure de tarification. Il s'agira en particulier d'ajuster les découpages intuitifs par les directions métier en analysant la variable « durée du voyage ». En ce sens, la méthodologie des GAM suppose une indépendance de la variable « durée du voyage » avec une quelconque forme rigide paramétrique. Au travers de sa fonction de lissage non paramétrique, les GAM appréhendent mieux la dépendance du risque assistance à la durée du voyage.

Dans cette partie, nous présenterons les modèles additifs généralisés en 3 étapes. Nous commencerons par présenter le cadre théorique des GAM en faisant le lien entre les modèles additifs et les GLM. Nous présenterons ensuite les quelques lisseurs testés sur la variable continue « durée du voyage ». Nous terminerons par la proposition d'un tarif via les GAM.

### 1. Cadre théorique

#### 1.1 Modèles additifs généralisés comme modèles semi-paramétriques

Les modèles linéaires généralisés permettent d'associer aux variables à expliquer (nombre de sinistres et montant des sinistres) une loi et ensuite d'y établir un lien avec les variables explicatives (lien logit pour le nombre de sinistre et log pour le montant des sinistres). Les modèles additifs généralisés ne remettent pas en cause tous ces choix. Ils offrent une marge de manœuvre concernant l'hypothèse forte des GLM selon laquelle il existe une relation linéaire entre chaque variable explicative et les variables à expliquer. En ce sens, le GLM est qualifié de modèle linéaire et paramétrique. Les GAM sont quant à eux des modèles non-linéaires et non-paramétriques.

Le modèle additif généralisé peut-être définie de la façon suivante :

$$g(E[Y]) = \eta = \alpha + \sum_{j=1}^q \beta_j X_j^{QUAL} + \sum_{j=1}^p f_j(X_j^{CONT})$$

Cette définition d'un modèle GAM porte souvent le nom de semi-paramétrisation du modèle. Les spécifications possibles d'un modèle GAM sont les suivantes :

Tableau 27 : Exemple de spécification d'un modèle GAM

Type de modèle	Formule mathématique
Paramétrique	$g(\mathbb{E}[Y]) = \alpha + \beta_1 X_1^{QUAL} + \beta_2 X_2^{QUAL}$
Non paramétrique	$g(\mathbb{E}[Y]) = \alpha + f_1(X_1^{CONT}) + f_2(X_2^{CONT})$
Semi-paramétrique	$g(\mathbb{E}[Y]) = \alpha + \beta_1 X_1^{QUAL} + f_2(X_2^{CONT})$

Avec :

- Une composante aléatoire  $Y$  qui suit une distribution appartenant à la famille exponentielle  $\mathcal{F}_{exp}(\theta, \phi, a, b, c)$  ;
- Une composante systématique composée de :
  - des variables explicatives qualitatives  $X_j^{QUAL}$  pour lesquels il faut estimer les coefficients  $\beta_j$  ;
  - des variables explicatives continues  $X_j^{CONT}$  pour lesquelles il faut estimer les lisseurs  $f_j$  ;
- Une fonction de lien  $g: \mathbb{R} \rightarrow \mathcal{S}$  qui est monotone, dérivable telle que  $\mathbb{E}(Y) = g^{-1}(\eta)$ .

Il faut noter que les modèles linéaires (et les GLM) sont des cas particuliers des modèles additifs (et GAM) avec  $f_j(x) = \beta_j x$ . Mais des fonctions bien plus sophistiquées peuvent-être utilisées.

La partie suivante présente les choix possibles des fonctions de lissage  $f_j$ . Seules les variables continues seront concernées à savoir la durée du voyage. Par conséquent, les variables qualitatives ne feront pas l'objet d'une étude par les GAM.

*Note :* Dans le cadre de l'estimation du modèle GAM, nous nous limiterons au lissage de la durée du voyage – variable supposée continue. Pour rappel, cette variable n'intervient que dans l'estimation de la fréquence des sinistres. Les autres variables seront paramétrées dans leur version « linéaire ». En effet, les autres sont qualitatives et garderont en ce sens leur découpage prédéfini (nombre d'assurés, profil du voyageur et destination géographique).

## 1.2 Fonctions de lissage dans le cas uni-varié

Cette sous-partie décrit brièvement les procédures classiques de choix des fonctions de lissage.

L'une des spécificités des Modèles Additifs Généralisés concerne les fonctions non-paramétriques  $f_j$  des variables prédictives  $X_{ij}$ . Plus précisément, au lieu d'une forme plus ou moins complexe de fonctions paramétriques, Hastie et Tibshirani (1986) évoquent divers lissages des nuages de points que nous pouvons appliquer aux valeurs de la variable  $X_{ij}$ , avec l'objectif de maximiser la qualité de la prévision des valeurs des variables  $Y_i$  (transformées). L'un de ces lissages de nuages de points est le lissage par les *splines* cubiques de lissage, qui permettent de produire une généralisation lissée des relations entre les deux variables du nuage de points. Toutefois, nous pourrions trouver plus d'informations sur ce lissage dans les ouvrages de Hastie et Tibshirani (1986) et de Schimek (2012).

### Les splines de lissage

Les splines de lissage sont des fonctions qui minimisent la somme des carrés des résidus, mais cette somme est « pénalisée » par un terme représentant la courbure de la fonction. Il s'agit ainsi de minimiser l'expression suivante :

$$\sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 dt \quad (1)$$

Où  $\lambda$  est un paramètre réel qui joue le rôle d'un facteur de lissage. Plus sa valeur est élevée, plus le lissage est important.  $f$  est la fonction spline recherchée et  $f''$  est la dérivée seconde de  $f$ .

Il a été démontré (Hastie et Tibshirani, 1999, p.27) que le problème général d'optimisation soulevé précédemment admet, pour un  $\lambda$  donné, une spline naturelle comme solution. En revanche, il n'est pas évident d'expliciter de manière analytique la fonction  $f$ .

### Les fonctions de régression locale pondérée

La méthode la plus naturelle pour lisser une variable est de la rendre polynomiale. Considérons des observations  $x_1, \dots, x_n$  et  $y_1, \dots, y_n$ , un modèle de régression multiple peut donc s'écrire de la façon suivante :

$$Y = \alpha_0 + \alpha_1 X + \dots + \alpha_p X^p$$

L'utilisation de fonction polynomiale reste cependant très forte en termes d'hypothèse. Un moyen d'être plus flexible est de découper la variable  $X$  en plusieurs intervalles représentant ainsi des segments de valeurs.

Ces fonctions sont, avec les fonctions *splines*, les outils les plus utilisés concernant la modélisation de séries temporelles. Car elles se fondent sur un procédé similaire à celui énoncé précédemment.

Les fonctions de régression locale pondérée, aussi appelées fonctions *loess*, remplacent un point  $(x_0, y_0)$  par une régression linéaire sur les points  $(x_i, y_i)$  appartenant au voisinage de  $(x_0, y_0)$ , affectés d'une pondération dépendant de l'éloignement  $|x_i, y_0|$  par rapport au point  $(x_0, y_0)$ .

Les fonctions *loess* réalisent ainsi un lissage non-paramétrique déterminé par l'étendue du voisinage de points participant aux régressions locales. Cette étendue est appelée fenêtre de lissage. Le lissage est d'autant plus important que la fenêtre est large.

## 1.3 Ajustement des modèles GAM

### Maximum de vraisemblance et algorithme numérique

Nous utilisons généralement deux méthodes pour estimer les paramètres d'un modèle linéaire généralisé : la méthode du maximum de vraisemblance et la méthode des moindres carrés. Nous ne nous intéresserons ici qu'à la méthode du maximum de vraisemblance dont le principe a déjà été présenté dans le chapitre précédent. Les algorithmes de résolution s'écartent des standards du domaine des GLM. Nous distinguons deux grandes méthodes de résolution d'un modèle additif généralisé :

- l'algorithme d'ajustement arrière ou *backfitting algorithm* : l'algorithme converge la plupart du temps dans le cas des *splines* de lissage mais dans le cas des fonctions *loess*, ceci est moins sûr ; et ;
- l'algorithme *local scoring algorithm* : cet algorithme reprend, en le généralisant, celui des scores de Fischer.

### Nombre de degrés de libertés et degré de lissage

Pour un GLM, le nombre de degré de liberté représente la différence entre le nombre d'observations et le nombre de paramètres à estimer. Dans un modèle additif généralisé, il sert à comparer le niveau de complexité de deux modèles d'estimation.

Le degré de pénalité ou degré de lissage dans un modèle additif généralisé est mesuré par  $\lambda$  (cf Formule (1)). Quand  $\lambda$  tend vers 0, il n'y a pas de pénalité de lissage et le modèle fournit un ajustement parfait : les valeurs ajustées correspondent alors aux données. En revanche, quand  $\lambda$  tend vers l'infini, l'ajustement est qualifié de « lissage parfait ».

## 2. Choix des fonctions de lissage de la variable continue durée du voyage

Pour le phénomène de fréquence, nous construisons le modèle GAM avec une distribution de type Bernoulli et un lien *logit*. Le modèle avec la variable « durée » prend alors la forme suivante :

$$\boxed{\text{logit}(E[\text{top\_sinistre\_assistance}]) = \alpha_0 + f_{\text{duree}}(\text{duree}_i)}$$

### 2.1 Lecture des résultats

L'estimation GAM du modèle ci-dessus est :

$$\log\left(\frac{\widehat{\mathbb{P}}[\text{verser une indemnité}_i]}{1 - \widehat{\mathbb{P}}[\text{verser une indemnité}_i]}\right) = \widehat{\alpha}_0 + \widehat{\beta} \times \text{duree}_i + \widehat{f_{\text{duree}}}(\text{duree}_i)$$

où  $\widehat{f}_{duree}$  est une fonction non-paramétrique estimateur de  $f$  (composante non-linéaire) et  $\hat{\beta}$  est l'estimateur de la composante linéaire du modèle GAM.

Le modèle converge, le tableau ci-dessous donne les détails des résultats de la résolution numérique du modèle. Comme précisé dans la partie méthodologie, l'algorithme de calibration utilisé ici est le « local fitting ».

Tableau 28: Synthèse des itérations et statistiques d'ajustement

Iteration Summary and Fit Statistics	
Number of local scoring iterations	9
Local scoring convergence criterion	2.0028628E-9
Final Number of Backfitting Iterations	1
Final Backfitting Criterion	3.5083721E-9
The Deviance of the Final Estimate	3785.6754157

Les GAM font une distinction entre l'influence linéaire  $\hat{\beta}$  notée « Linear (duree) » de la variable « durée du voyage » et son influence non-linéaire  $\hat{f}$  notée « Spline (duree) ». A ce titre, un ajustement spécifique y est effectué. Le tableau ci-dessous donne alors l'ajustement de la partie linéaire. La p-value est inférieure à 0.0001.

Tableau 29: Estimation des paramètres intervenant dans la partie linéaire du modèle Binomial GAM

Regression Model Analysis				
Parameter Estimates				
Parameter	Parameter	Standard	t Value	Pr >  t
	Estimate	Error		
Intercept	-4.99328	0.07333	-68.09	<.0001
Linear (duree)	0.00900	0.00080167	11.23	<.0001

La partie non-linéaire est également significativement différente de 0.

Tableau 30 : Estimation des paramètres intervenant dans la partie non-linéaire du modèle Binomial GAM

Smoothing Model Analysis				
Fit Summary for Smoothing Components				
Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline (duree)	6.00000	141.817162	141.8172	<.0001

Le modèle résultant se présente de la façon suivante :

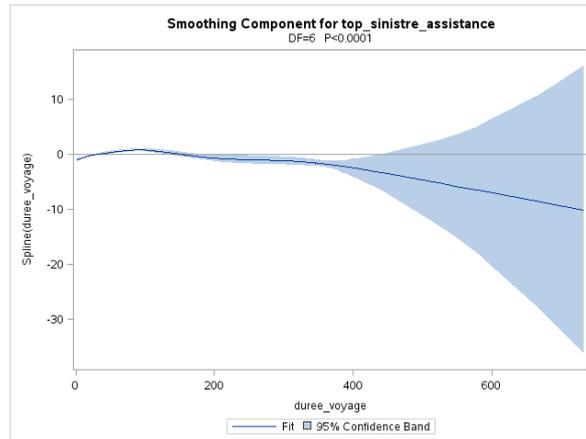
$$\log\left(\frac{\widehat{\mathbb{P}}[\text{verser une indemnité}_i]}{1 - \widehat{\mathbb{P}}[\text{verser une indemnité}_i]}\right) = -4.99 + 0.009 * \widehat{duree}_i + \widehat{f}_{duree}(duree_i)$$

Avec  $\widehat{f}_{duree}$  un spline de degré de lissage 6.

SAS permet également obtenir les valeurs de  $f_{duree}(duree_i)$ . La prédiction de la fréquence de sinistres peut aisément se calculer sous Excel.

La courbe de la Figure 33 laisse bien apparaître une non-linéarité de la fréquence de sinistralité en fonction de la durée du voyage. Ce qui valide ainsi l'hypothèse de non-linéarité de l'influence de la variable « durée du voyage ».

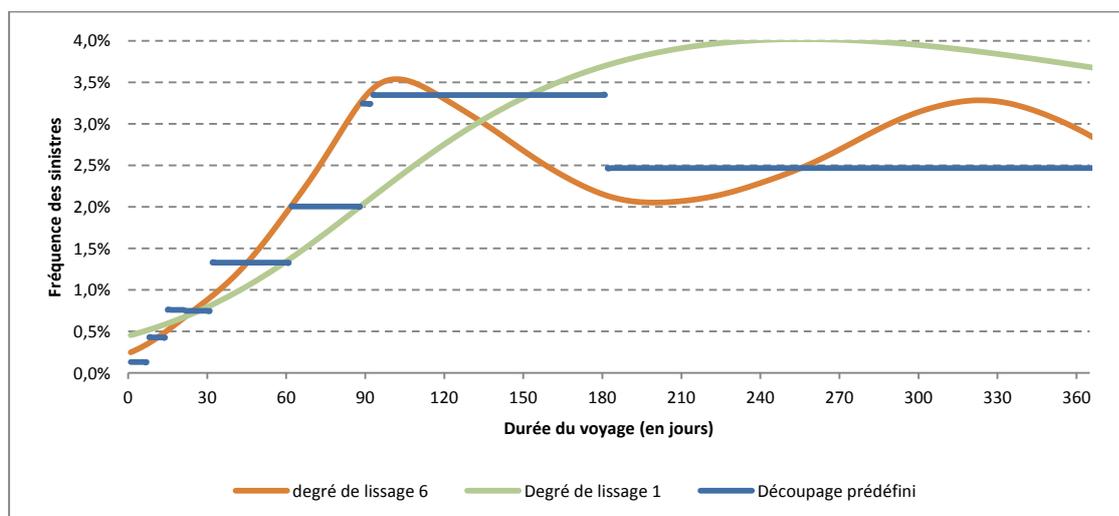
Figure 33: Lisseur de la durée du voyage estimée par la procédure GAM



## 2.2 Sensibilité au degré de lissage

Les lisseurs ont également l'avantage d'éviter des sauts de tarifs dans les cas où un assuré veuille changer de case tarifaire. Par exemple, l'utilisation d'un découpage prédéfini conduit à doubler la fréquence de sinistres suivant que l'assuré passerait d'une formule à 80 jours à une formule à 86 jours sans pour autant qu'il n'y ait une modification significative du risque. Nous analysons alors l'influence du degré de lissage sur le lisseur de la durée de voyage.

Figure 34 : Influence du degré de lissage sur l'estimation du lisseur de la durée du voyage



Pour un degré de lissage égal à 1, nous ne captions pas tous les effets de montée et de redescende de la fréquence de sinistralité en fonction de la durée du voyage. Un degré de lissage égal à 6 nous a paru mieux correspondre aux effets observés empiriquement. C'est donc celui-là que nous utiliserons pour tarifier le risque assistance grâce au GAM. En effet entre 0 et 100 jours, nous considérons que la durée du voyage à une tendance quasi-linéaire sur la fréquence des sinistres. Le comportement des voyageurs entre 3 mois et 6 mois est assez particulier. Nous considérons que les voyageurs commencent à s'acclimater sont donc moins sujet à des sinistres pendant cette période car font de plus en plus attention pendant leur séjour. Au-delà de un an, nous pouvons supposer que les voyageurs reprennent des habitudes normales d'où la (re)croissance de la fréquence des sinistres (Cette supposition devra être revue ou validée avec une exposition plus conséquente que celle de notre étude).

### 2.3 Validation du choix de la fonction de lissage de la variable continue durée du voyage

*Note : Dans l'idéal, nous ne souhaitons pas nous éloigner des modèles linéaires généralisés car ils sont simples et facilement interprétables et ont un temps de calcul plus rapide que les GAM. Dans cette partie, nous nous attacherons à revoir le découpage de la variable durée du voyage en nous appuyant notamment sur les GAM.*

Les variables retenues dans la tarification ont fait l'objet de découpages ou discrétisation selon la nature de la variable. Cette partie permet de critiquer les découpages effectués ci-dessous. En effet, la plupart des transformations faites sur les variables découlent majoritairement de prises en compte de pratiques opérationnelles. Le cas échéant, certaines variables ont été découpées grâce à la méthode des quantiles.

La variable « durée du voyage » est celle qui a été retenue dans la tarification du phénomène de fréquence. Le modèle GAM nous permettra ici de valider le choix des découpages effectués.

Précisons comment s'effectuera le choix des classes :

- à partir des intervalles de confiance issus des modèles GAM, nous allons définir des niveaux moyens grâce à des traits rouges ;
- les classes seront disjointes ;et ;
- la mesure de l'intervalle de confiance nous donne une indication sur la volumétrie de la classe. Afin d'éviter des estimations trop volatiles, nous regrouperons les classes à faibles effectifs.

Le modèle que nous avons présenté plus haut estime le nombre de sinistres en fonction de la durée d'exposition au risque. A présent, nous étudierons directement l'influence de la durée sur la fréquence de sinistralité annuelle.

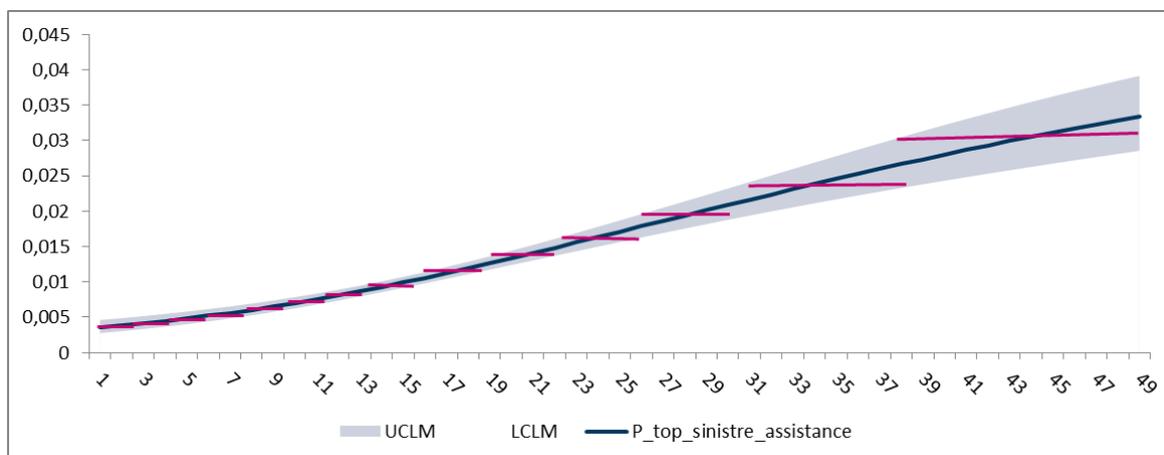
Les intervalles de confiance s'élargissent pour les valeurs supérieures à 104 jours. Cela est dû à la faible distribution sur ce segment. Pour bien travailler le choix des classes, nous allons zoomer sur des tranches de durée d'exposition.

Entre 1 jour et 50 jours, nous observons une forte croissance de la fréquence de la sinistralité. Nous faisons donc le choix de la discrétisation suivante :

Table 1 : Découpage de la variable durée du voyage comprise entre 1 jour et 50 jours

minimum	maximum	Modalités
0	2	du000-002
2	3	du002-003
3	4	du003-004
4	5	du004-005
5	6	du005-006
6	7	du006-007
7	8	du007-008
8	10	du008-010
10	12	du010-012
12	14	du012-014
14	17	du014-017
17	20	du017-020
20	23	du020-023
23	26	du023-026
26	29	du026-029
29	40	du029-040
40	50	du040-050

Figure 35: Durée comprise entre 1 jour et 50 jours

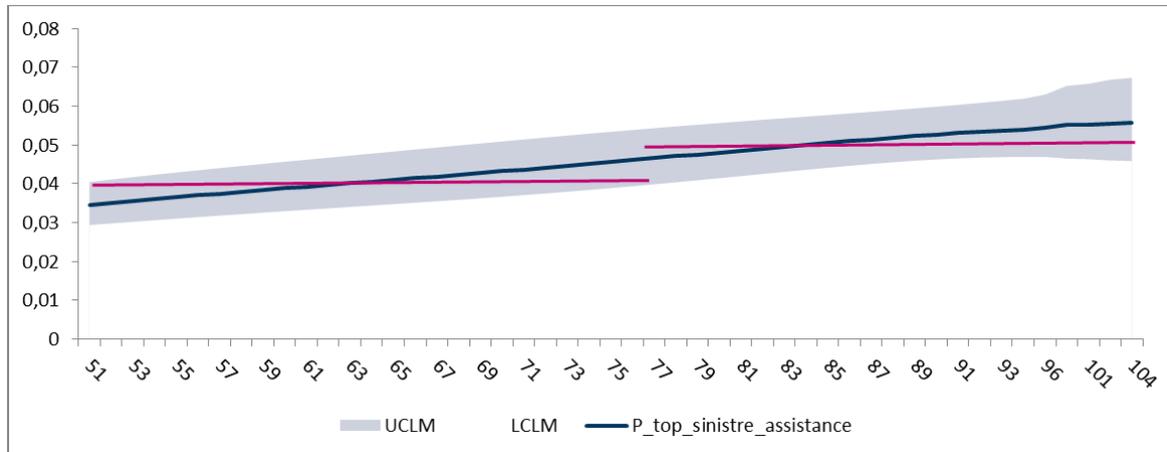


Pour les durées comprises entre 51 jours et 105 jours, les intervalles de confiance sont relativement plus larges. Nous faisons alors le choix de découper la variable en 2.

Table 2: Découpage de la variable durée du voyage comprise entre 51 jours et 105 jours

minimum	maximum	Modalités
50	75	du050-075
75	100	du075-100

Figure 36 : durées comprises entre 51 jours et 105 jours



Et enfin, pour les durées d'exposition supérieures à 100 jours, nous observons une nette dégradation de la précision d'estimation des modèles GAM. Un saut dans la fréquence des sinistres est également à noter pour les durées supérieures à 200 jours. Autrement, la fréquence de sinistres reste stable.

minimum	maximum	Modalités
100	150	du100-150
150	10000	du150-inf

### Performance des GLM sur le nouveau découpage

Nous allons nous intéresser dans cette partie à déterminer la performance en termes de pouvoir discriminant du nouveau découpage en comparaison à l'ancien découpage.

Pour cela, nous lançons une régression logistique en uni-varié sur le nouveau découpage et l'ancien découpage. Comme précisé dans la partie « comparaison des modèles », le modèle le plus performant est celui qui a le critère AIC le plus bas.

Tableau 31 : Performance du nouveau découpage de la variable durée du voyage

	AIC	Aire sous la courbe ROC	Nombre de modalités
<b>Ancien découpage</b>	6028	68%	9
<b>Nouveau découpage</b>	5908	72%	21

Le nouveau découpage semble plus discriminant comparé au découpage basé sur les pratiques opérationnel au sens du critère de l'aire sous la courbe ROC et de l'AIC. Mais il est également difficile de critiquer l'ancien découpage qui n'a que 9 modalités et atteint tout de même un pouvoir discriminant acceptable.

De façon plus générale, la procédure GAM n'est pas adaptée aux études comportant plusieurs variables explicatives car les temps de calculs et d'estimation seraient très longs étant donné la fonction non-paramétrique des GAM. Nous préférons donc écarter définitivement l'hypothèse de remplacer les modèles GLM par celui des GAM.

### 3. Résultats de la tarification alternative grâce aux GAM

La procédure GAM telle que nous l'utilisons c'est-à-dire avec un nombre limité de variables explicatives (4 variables pour la fréquence des sinistres) est relativement longue en temps de calcul d'estimation de la fonction non-paramétrique. Nous avons estimé le modèle tarifaire en utilisant les découpages des variables Profil, Nombre d'assurés et Destination tels qu'ils apparaissent dans le Tableau 20.

Le modèle estimé pour la fréquence par la procédure GAM de SAS est :

Pour un segment de risque ou case tarifaire  $i$ , la fréquence est donnée par la formule

$$\log \left[ \frac{\mathbb{P}[\text{verser une indemnité}_i]}{1 - \mathbb{P}[\text{verser une indemnité}_i]} \right] = \hat{\beta}_0 + \hat{f}_{\text{durée},i}(\text{durée}_i) + \hat{\beta} \times \text{duree}_i + \hat{\beta}_{\text{Profil},i} * \text{Profil}_i + \hat{\beta}_{\text{nb\_assuré},i} * \text{nb\_assuré}_i + \hat{\beta}_{\text{Destination},i} * \text{Destination}_i$$

Nous utilisons à nouveau les résultats du Tableau 23 pour le coût moyen. Nous rappelons le modèle :

$$\ln(\mathbb{E}[\text{coût\_moyen}]) = \beta_0 + \beta_{\text{Profil}} * \text{Profil} + \beta_{\text{zone\_geo}} * \text{zone\_géographique}$$

Les coefficients tarifaires pour la fréquence sont restitués dans le tableau ci-dessous :

**Tableau 32 : Estimation des composantes paramétriques (linéaires) du modèle tarifaire par les GAM**

Regression Model Analysis					
Parameter Estimates					
Parameter		Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept		-5.40606	0.21472	-25.18	<.0001
Nombre d'assurés	na01-02	0.47386	0.19717	2.40	0.0163
	na02-03	0.30623	0.21260	1.44	<.0001
	na03-inf	0	.	.	.
Zone	Afrique	0.21864	0.25003	0.87	0.0038
	Amerique Autre	-0.03985	0.25119	-0.16	0.0087
	Amérique du nord	0.14023	0.19008	0.74	0.0047
	Asie	0.13534	0.17647	0.77	0.0443
	Europe hors France	-0.47670	0.33806	-1.41	<.0001
	France	-0.01787	0.16593	-0.11	<.0001
	Maghreb & Reste du monde	0	.	.	.
Profil	Affaire	-0.95026	0.38761	-2.45	0.0142
	Etude	-1.34610	0.24811	-5.43	<.0001
	Globe trotter	-0.60008	0.30405	-1.97	<.0001
	Vacance & Visa Schengen	0	.	.	.
	Linear (Duree)	0.01006	0.00084364	11.93	<.0001

**Tableau 33 : Estimation de la composante non-paramétrique (non-linéaire) du modèle tarifaire par les GAM**

Smoothing Model Analysis				
Analysis of Deviance				
Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline (Duree)	6.00000	141.711333	141.7113	<.0001

### Calcul de la prime pure : Cas pratique

Pour donner une indication sur l'utilisation des grilles tarifaires, nous allons prendre un exemple concret.

Un assuré célibataire qui souhaiterait se rendre en vacances en Afrique pour une durée maximale d'une semaine paierait une prime pure de :

$$\left\{ \begin{array}{l} Prime\_Pure_{Afri/du000-008/na01-02/vacs} \\ \end{array} \right. = \left[ \frac{e^{(-5.41+0.01 \times \frac{7}{360} - 0.97 + 0.47 + 0.22)}}{1 + e^{(-5.41+0.01 \times \frac{7}{360} - 0.97 + 0.47 + 0.22)}} \right] e^{(8,08)}$$

$$= 0,34\% * 3\ 234\text{€}$$

$$= 10,96\ \text{€}^6$$

Ce tarif basé sur une méthode de segmentation plus « technique » de notre variable durée du voyage sera (au mettre titre que les tarifs définis précédemment) exposé au service souscription afin de juger de la pertinence des sorties.

Les tarifs proposés jusqu'ici sont basés sur la prime pure, qui est essentielle pour la construction d'un tarif actuariel lorsque les données sont en qualité et quantité suffisante. En effet, la prime pure est essentielle dans la limite où elle permet à l'assureur de couvrir le coût du risque. Cependant, elle ne peut être la prime demandée par l'assureur à l'assuré sans intégration des taxes et frais de gestion. Nous proposons donc de déterminer une prime commerciale et une méthode d'ajustement tarifaire.

## 4. Structure tarifaire finale

### 4.1 Intégration des frais internes

Par soucis de confidentialité, nous intégrerons un taux de chargement global (Frais internes + taxes) de 50% sur notre prime pure afin de la convertir en prime commerciale. Nous pouvons donc déterminer la prime proposée aux assurés comme :

$$Prime\_Commerciale_i = Prime\_Pure_i * (1 + Taux\ de\ chargement)$$

Pour exemple, un assuré qui souhaiterait se rendre en vacances en Afrique pour une durée maximale d'une semaine paierait suivant la méthode tarification actuelle une prime commerciale de :

$$\left\{ \begin{array}{l} Prime\_Commerciale_{Afri/du089-093/na01-02/vacs} \\ \end{array} \right. = 0.288\% * 2\ 985\text{€} * (1 + 50\%)$$

$$= 0.288\% * 2\ 985\text{€} * 1.5$$

$$= 12.9\ \text{€}$$

<sup>6</sup> Analyse détaillée : La valeur 8.08 représente le coefficient « intercept »  $\hat{\beta}_0$  de la régression. -5.41 représente le paramètre de lissage  $\hat{f}_{durée, du000-008}$  issu des GAM. 7/360 représente le nombre de jours de voyage que souhaite souscrire l'assuré rapporté à 360 jours. 0.01, -0.97, 0.47 et 0.22 sont les coefficients associés respectivement aux variables Profil, nombre d'assurés et destination du voyage.

## 4.2 Ajustement tarifaire grâce à une projection de la sinistralité future

L'un des objectifs de la tarification est de fixer des taux afin que les primes acquises permettent de couvrir les pertes et les frais futurs tout en réalisant les profits ciblés par les comités de tarification. La détermination de la prime pure a fait l'objet de toutes les parties précédentes. A présent, nous allons nous intéresser à l'ajustement tarifaire à effectuer lorsque les comités tarification ont statué sur une éventuelle anticipation de la sinistralité future. En effet, les conditions économiques futures, le business plan ou d'autres facteurs exogènes peuvent conduire à une sinistralité plus optimiste que celle observée par le passé. Autrement dit, les tarifs en vigueur pourraient s'avérer au-dessus des tarifs marchés.

Dans cette partie, nous allons exhiber des formules fermées permettant d'ajuster les primes pures calculées par segment en fonction d'une anticipation de la sinistralité future.

Il existe plusieurs indicateurs représentatifs de la sinistralité future :

- le ratio sinistre sur prime ;
- ou la prime pure.

En fonction, de l'indicateur utilisé l'ajustement tarifaire a effectué pour chaque segment tarifaire serait alors :

$$Taux\_d\_ajustement = \frac{Sinistralite\_Historique}{Sinistralite\_Future} - 1$$

Avec *Sinistralite\_Future* : le loss-ratio (prime pure) prédit pour l'année future et *Sinistralite\_Historique* le loss-ratio (prime pure) calculé grâce aux données historiques (en général l'exercice comptable qui vient de s'écouler).

En résumé, les grilles de tarification transmises aux services souscription devons prendre en compte les taux de chargements internes ainsi que le taux d'ajustement afin d'être en phase avec une politique de souscription combinée à une tendance réelle du marché à un instant t.

Les grilles pourront aussi être confrontées aux prix du marché en prenant en compte les difficultés que nous énoncions dans notre première partie concernant la mise en place d'études précises et comparatives du marché entre les différents assureurs concurrents (plafonds différents, prix sur devis via mail etc...). Cette étude sera faite par le service Marketing.

## Partie 3 : Conclusion

---

A travers cette étude, nous avons pu compiler les premiers résultats de tarification via l'application de nouvelles méthodes de tarification basée sur les modèles linéaires. Cela nous a permis de mettre en évidence l'importance d'une meilleure segmentation technique de notre activité afin d'éviter la sur ou sous tarification de certaines classes tarifaires.

Mais aussi de souligner l'intérêt d'affiner les résultats de cette étude en remontant certaines informations indispensables, afin que nous puissions les exploiter pour apporter plus de précision à notre tarif. Nous pourrions prendre par exemple, la nécessité de demander à tous nos souscripteurs l'âge des assurés qui nous paraît un critère clé à l'évaluation de nos risques. De même, sensibiliser le souscripteur sur l'importance de l'estimation du montant du voyage lorsqu'il n'est pas connu qui, comme nous l'avons vu, est très rarement correctement renseigné.

Cette étude nous a permis de proposer au service souscription la base d'une tarification plus équilibrée avec une meilleure justification technique. Tous en restant conscient qu'un tarif équilibré ne suffit pas à son propre succès, il doit être également compétitif.

Par ailleurs, cette réflexion nous a permis de mettre en lumière l'évidence que la vente directe était bien un levier de croissance qui nous permet de faire abstraction des intermédiaires et donc des commissions élevées à partir du moment où nous maîtrisons notre risque par une meilleure connaissance de notre sinistralité.

Cependant, nous nous retrouvons d'un autre côté plus exposés au phénomène d'anti sélection via la vente direct sur notre site. Cela se traduit par des fréquences de sinistre plus élevées que si nous passions par nos canaux de distribution classique BtoBtoC, qui eux limitent le phénomène d'anti sélection. A cela, il faut ajouter les couts importants en interne nécessaire au développement de ce canal de distribution (webmaster, étude marketing...).

De plus, une des idées générales de cette étude était de faire un mémoire compréhensible pour tous les profils de métier présent au sein du groupe mondial assistance. C'est dans ce sens que nous avons essayé d'apporter dans ce mémoire une vision d'analyse globale de façon méthodique (contexte + Théorie et pratique), afin de mettre en évidence l'apport fondamental de l'actuariat pour ce secteur de l'assurance.

Il est important aussi de préciser en conclusion que tout au long de cette étude nous avons dû faire différents choix et hypothèses concernant notre modélisation avec de potentielles sources d'erreur. A savoir par exemple en premier lieu la pertinence même de l'application des GLM sur une base représentant une exposition du portefeuille inférieure à 100 000 lignes de données comme le recommande un document de la CAS [5,p40] que l'on trouvera en annexe. Cela résume parfaitement la difficulté de l'actuaire produit ou souscripteur qui lors de la mise en place de sa tarification devra rester conscient des limites que peut représenter le modèle.

Concernant l'extension : Plusieurs pistes intéressantes seront étudiées via le service actuariat. Nous pourrions proposer :

- de poursuivre l'étude en intégrant de nouvelles variables et redéfinir un cout moyen étant donné la configuration des compagnies d'assistance (vu dans notre première partie) ;
- de déployer l'application des GLM aux autres activités du groupe mondial assistance par exemple sur l'auto (en cours), la sante et la MRH ; et ;
- d'évaluer avec précision, le canal de distribution le plus rentable entre le BtoC et le BtoBtoC en incluant l'ensemble des coûts (interne + externe) croissants que nécessitent le développement nouveau de cette branche.

Ces pistes seront centrales, dans le développement futur de ce service.

## Partie 4 : Bibliographie

---

### 1. Livres

DENUIT M., CHARPENTIER A. [2005] Mathématiques de l'assurance non-vie, Tome II : Tarification et Provisionnement, collection « Economie et statistiques avancées », édition Economica, 596 p.

DESCURE C. [2004] Management par la valeur, collection Ecole Nationale d'Assurance, édition Vuibert, 151 p.

EUROSTAT [2002] Les statistiques tome 2: L'analyse de la variance, modélisation, collection Les guides SAS, éditions Eurostat, 519 p.

Hastie, T., & Robert, T. (1986). Generalized Additive Models. *Statistical Science*, 297-318.

Schimek, M., & Turlach, B. (2012). 10. Additive and Generalized Additive Models. Dans *Smoothing and Regression : Approaches, Computation and Application*.

SAPORTA G. [1990] Probabilités, analyse des données et statistique, éditions Technip, 493 p.

Schimek, M., & Turlach, B. (2012). 10. Additive and Generalized Additive Models. Dans *Smoothing and Regression : Approaches, Computation and Application*.

### 2. Mémoires

MILOE Y. [2011] Tarification d'un produit MRH à l'aide de la méthode des marges, mémoire d'actuariat, ISFA

DUPLOUY S. [2008] Etude et modélisation de la sinistralité d'un portefeuille de prévoyance individuelle, mémoire d'actuariat, EURIA

HAMEL S. [2007] Prédiction de l'acte de résiliation de l'assuré et optimisation de la performance en assurance automobile particulier, mémoire d'actuariat, ENSAE

SOULEAU A. [2010] Tarification de la branche d'assurance des accidents du travail, Bonus-malus et Crédibilité, mémoire d'actuariat, EURIA

VANNEAUX J-C. [2010] Modélisation de la valeur contrat (PNPV) pour le pilotage du portefeuille automobile particulier, mémoire d'actuariat, ISFA

## Partie 5 : Table des illustrations

---

Figure 1: Logigramme de la méthodologie générale.....	26
Figure 2: Liste des variables en entrée du fichier Contrat.....	28
Figure 3: Liste des variables en entrée du fichier Sinistre .....	29
Figure 4: Fréquence de sinistres en fonction de la durée de voyage .....	35
Figure 5: Fréquence de sinistres en fonction de la destination du voyage .....	35
Figure 6: Fréquence de sinistres en fonction du nombre d'assurés .....	36
Figure 7: Fréquence de sinistres en fonction du montant du voyage .....	37
Figure 8: Fréquence des sinistres en fonction du motif du voyage.....	38
Figure 9: Coût moyen en fonction de la durée du voyage .....	39
Figure 10: Coût moyen en fonction de la destination du voyage .....	39
Figure 11: Coût moyen en fonction du montant du voyage .....	40
Figure 12: Coût moyen en fonction du nombre d'assurés .....	41
Figure 13: Coût moyen en fonction du motif du voyage.....	41
Figure 14 : Distribution du nombre de sinistres par contrat estimé avec un modèle probabiliste.....	49
Figure 15: Histogramme du montant des sinistres ajusté par une loi log-normale .....	52
Figure 16: QQ plot Montant des sinistres empirique / Montant modélisé par une loi log-normale .....	52
Figure 17 Histogramme du montant des sinistres ajusté par une loi Gamma .....	53
Figure 18: QQ plot Montant des sinistres empirique / Montant modélisé par une loi Gamma .....	53
Figure 19 Histogramme du montant des sinistres ajusté par une loi Gaussienne Inverse.....	54
Figure 20: QQ plot Montant des sinistres empirique / Montant modélisé par une loi Gaussienne Inverse.....	54
Figure 21: Courbe ROC du modèle initial de fréquence .....	59
Figure 22: Evolution des métriques de sélection des variables avec la méthode <i>stepwise</i> .....	59
Figure 23: Résidus sortis du modèle de régression du nombre de sinistres .....	62
Figure 24 : Résidus sortis du modèle de régression du coût moyen .....	63
Figure 25 : Courbe ROC du modèle GLM sur la fréquence.....	70
Figure 26: Comparaison entre fréquences observée et prédite pour la variable "durée du voyage" ...	74
Figure 27: Comparaison entre fréquences réelle et prédites pour la variable "nombre d'assurés" .....	74
Figure 28: Comparaison entre fréquences réelle et prédites pour la variable "Profil" .....	75
Figure 29: Comparaison entre fréquences réelle et prédites pour la variable "Destination géographique" .....	75
Figure 30: représentation des coefficients MLG pour la variable "Destination géographique".....	76
Figure 31: représentation des coefficients MLG pour la variable "Motif du voyage" .....	76

Figure 32 : Différentes modélisations de la fréquence de sinistres en fonction de la durée du voyage	84
Figure 33: Lisseur de la durée du voyage estimée par la procédure GAM.....	90
Figure 34 : Influence du degré de lissage sur l'estimation du lisseur de la durée du voyage.....	90
Figure 35: Durée comprise entre 1 jour et 50 jours.....	92
Figure 36 : durées comprises entre 51 jours et 105 jours.....	93
Figure 38 : Détails des coefficients estimés dans le cadre du modèle « initial » utilisé pour détecter les variables tarifaires .....	108

## Partie 6 : Table des tableaux

---

Tableau 1: Effectif des entreprises et intermédiaires d'assurance .....	10
Tableau 2 : Montants de voyage forfaitaires attribués à des zones prédéfinies .....	29
Tableau 3 : Evolution de la fréquence et du coût moyen selon les années .....	33
Tableau 4: Tableau des corrélations entre variables explicatives .....	42
Tableau 5: Tableau des corrélations sur la fréquence et l'intensité des sinistres .....	43
Tableau 6: Interprétation de la valeur de l'aire sous la courbe ROC .....	48
Tableau 7: Tableau des fréquences empiriques .....	48
Tableau 8: Statistiques de Kolmogorov - Smirnov pour la loi de Poisson et la loi Binomiale Négative	49
Tableau 9 : Résultat de l'ajustement du montant des sinistres par une loi log-normale.....	52
Tableau 10: Résultat du test d'ajustement du montant des sinistres par une loi Gamma .....	53
Tableau 11: Résultat du test d'ajustement du montant des sinistres par une loi Gaussienne Inverse.	54
Tableau 12: Tests de significativité (type 3) des variables du MLG "initial" du nombre de sinistres ....	58
Tableau 13: Tests de significativité des variables du MLG sans le montant du voyage.....	60
Tableau 14: Statistiques d'ajustement du MLG "initial" du coût moyen .....	61
Tableau 15: Tests de significativité des variables du MLG "initial" du coût moyen .....	61
Tableau 16 : Test de la significativité globale du modèle sans regroupement des modalités .....	66
Tableau 17: Significativité (type 3) des variables retenues pour la modélisation la fréquence .....	66
Tableau 18: Estimation du modèle de fréquence suite à la sélection automatique des variables.....	67
Tableau 19: <i>Odds ratio</i> modèle de fréquence juste après la sélection automatique des variables.....	68
Tableau 20: Estimation du modèle de fréquence final (après regroupement des modalités non significatives).....	69
Tableau 21 : Critère permettant d'évaluer la qualité d'ajustement du modèle de régression du coût moyen.....	71
Tableau 22: Coefficients estimés par le modèle de régression .....	71
Tableau 23: Estimation du modèle final pour le coût moyen .....	72
Tableau 24 : Coefficients tarifaires pour le coût moyen par approche directe sur un modèle à risque proportionnel .....	77
Tableau 25:Grille de tarification du coût moyen .....	80
Tableau 26 : Coefficients estimés par le modèle GLM intégrant le logarithme de la durée en variable offset.....	82
Tableau 27 : Exemple de spécification d'un modèle GAM.....	86
Tableau 28: Synthèse des itérations et statistiques d'ajustement .....	89
Tableau 29: Estimation des paramètres intervenant dans la partie linéaire du modèle Binomial GAM	89
Tableau 30 : Estimation des paramètres intervenant dans la partie non-linéaire du modèle Binomial GAM .....	89

Tableau 31 : Performance du nouveau découpage de la variable durée du voyage .....	93
Tableau 32 : Estimation des composantes paramétriques (linéaires) du modèle tarifaire par les GAM .....	95
Tableau 33 : Estimation de la composante non-paramétrique (non-linéaire) du modèle tarifaire par les GAM .....	95
Tableau 34: Statistiques d'ajustement du MLG Log-normal « initial » du coût moyen .....	108
Tableau 35 : Test de significativité des variables retenues pour le modèle à risque proportionnel....	111
Tableau 36: Coefficients estimés dans le cadre du modèle à risque proportionnel.....	111

## Partie 7 : Annexes

---

### A. Choix de nombre de classes

En théorie, le nombre de classes croît avec la population. Les principales méthodes de détermination du nombre de classes sont présentées ci-dessous, où  $k$  est le nombre de classes à déterminer et  $n$  le nombre d'individus :

- Brooks-Carruthers :  $k = 5 \times \log_{10} n$
- Huntsberger :  $k = 1 + 3,3 \times \log_{10} n$
- Sturges :  $k = \log_2(n + 1)$
- Scott :  $k = \frac{(\max(x) - \min(x))}{3,5 \times \sigma(x) \times n^{(-1/3)}}$
- Freedman-Diaconis :  $k = \frac{(\max(x) - \min(x))}{2 \times (Q_3 - Q_1) \times n^{(-1/3)}}$ 
  - Avec :  $Q_i$  : Quartile au  $i^{\text{ème}}$  rang de  $X$

Remarque : Les 3 premières méthodes ne dépendent que du nombre d'individus tandis que les 2 dernières sont fonction du nombre des individus et de leur distribution.

Grâce à la méthode des quantiles, nous avons créé des tranches de valeurs de telle sorte qu'elles soient à la fois suffisamment peuplées et suffisamment discriminantes.

Une fois les tranches de valeurs fixées, un codage spécifique des modalités sera utilisé tout au long de ce mémoire.

## B. Mesures de corrélation

Considérons 2 variables aléatoires X et Y. Le tableau croisé issu des 2 variables se présente sous la forme suivante :

	$Y = y_1$	...	$Y = y_s$	Marg X
$X = x_1$	$n_{11}$	...	$n_{1s}$	$n_{1.}$
...	...	$n_{ij}$	...	...
$X = x_r$	$n_{r1}$	...	$n_{rs}$	$n_{r.}$
Marg. Y	$n_{.1}$	...	$n_{.s}$	n

- Données quantitatives : Coefficient de corrélation linéaire de Pearson

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \text{ où } \rho(X, Y) \in [-1, 1]$$

- Données quantitatives : Tau de Kendall et test de rang se Spearman
  - o Tau de Kendall : Soit deux couples (X, Y) et (X', Y') indépendants et identiquement distribués, le tau de Kendall est défini comme la probabilité de « concordance » moins la probabilité de « discordance » ie :

$$\tau(X, Y) = \mathbb{P}[(X - X')(Y - Y') > 0] - \mathbb{P}[(X - X')(Y - Y') < 0].$$

Si X et Y sont indépendants alors  $\tau(X, Y) = 0$ .

- o Test de rang de Spearman : Soit le couple (X, Y) de fonction de répartition marginale  $F_1$  et  $F_2$  continue et  $(X^\perp, Y^\perp)$  une version indépendante du couple (X, Y) (ie admet  $F_1 F_2$  comme fonction de répartition conjointe). Le rho de Spearman est défini comme suit :

$$\rho_s(X, Y) = 3\{\mathbb{P}[(X - X^\perp)(Y - Y^\perp) > 0] - \mathbb{P}[(X - X^\perp)(Y - Y^\perp) < 0]\}$$

Si X et Y sont indépendants alors  $\rho_s$  est nulle.

- Test d'indépendance du  $\chi^2$  et V de Cramer (entre deux variables nominales)

$$k^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N}$$

où :

- o  $n_{ij}$  est l'effectif dans la cellule (i, j) du tableau de croisement entre les 2 variables X (en ligne  $(X_i)_{1 \leq i \leq r}$ ) et Y en colonne (en ligne  $(Y_j)_{1 \leq j \leq s}$ )
- o  $n_{i.}n_{.j}/N$  est l'effectif théorique sous l'hypothèse d'indépendance ;
- o  $n_{i.} = \sum_{j=1}^s n_{ij}$  somme des lignes ;

- $n_{.j} = \sum_{i=1}^r n_{ij}$  somme des colonnes

La statistique de test  $k^2$  suit une loi  $\chi^2$  sous l'hypothèse d'indépendance entre X et Y, avec  $(r-1)(s-1)$  degrés de liberté où r et s représentent le nombre de modalités respectivement de X et Y. Il faut noter que ce test nécessite un nombre d'observations suffisant dans chaque case (i, j). Classiquement, il doit être supérieur ou égal à 5.

Le V de Cramer est basé sur la statistique  $k^2$ . Il est défini comme suit :

$$\sqrt{\frac{k^2}{N * \inf(s-1, r-1)}}$$

Il est compris entre 0 et 1, et est une forme centrée réduite du  $\chi^2$ .

- Coefficient Gamma de Goodman et Kruskal (entre deux variables ordinales) : c'est une statistique basée uniquement sur le nombre de paires d'observations concordantes et discordantes. Elle ne prend pas en compte les observations égales en X ou en Y. C'est la raison pour laquelle le coefficient Gamma est particulièrement adapté pour les variables de type ordinal. Il est compris entre -1 et 1. Si X et Y sont indépendantes alors le coefficient Gamma est proche de 0. Le coefficient Gamma peut-être calculé de la façon suivante :

$$G = \frac{P - Q}{P + Q}$$

Où P est le nombre de paires concordantes et Q le nombre de paires discordantes.

$$P = \sum_i \sum_j n_{ij} A_{ij}$$

$$Q = \sum_i \sum_j n_{ij} D_{ij}$$

Et

$$A_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} + \sum_{k<i} \sum_{l<j} n_{kl}$$

$$D_{ij} = \sum_{k>i} \sum_{l<j} n_{kl} + \sum_{k<i} \sum_{l>j} n_{kl}$$

## C. Sélection des variables

### 1. Au titre de la fréquence

Figure 37 : Détails des coefficients estimés dans le cadre du modèle « initial » utilisé pour détecter les variables tarifaires

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
<b>Intercept</b>		1	2.9767	0.0420	2.8943	3.0591	5013.44	<.0001
<b>DU</b>	du000-008	1	3.7266	0.1628	3.4076	4.0457	523.98	<.0001
<b>DU</b>	du008-015	1	2.8772	0.0569	2.7656	2.9887	2557.42	<.0001
<b>DU</b>	du015-022	1	2.4058	0.0473	2.3131	2.4985	2589.22	<.0001
<b>DU</b>	du022-032	1	2.3249	0.0464	2.2339	2.4159	2506.80	<.0001
<b>DU</b>	du032-062	1	1.5784	0.0403	1.4994	1.6574	1533.07	<.0001
<b>DU</b>	du062-089	1	1.0241	0.0367	0.9522	1.0960	779.12	<.0001
<b>DU</b>	du089-093	1	0.5590	0.0317	0.4968	0.6212	310.37	<.0001
<b>DU</b>	du093-182	1	-0.0538	0.0215	-0.0958	-0.0118	6.29	0.0121
<b>NA</b>	na02-03	1	-0.4926	0.0226	-0.5369	-0.4482	473.98	<.0001
<b>NA</b>	na03-inf	1	-0.8637	0.0366	-0.9355	-0.7919	556.49	<.0001
<b>Zone</b>	Afrique	1	-0.3683	0.0424	-0.4513	-0.2852	75.52	<.0001
<b>Zone</b>	Amerique Autre	1	-0.1941	0.0336	-0.2599	-0.1282	33.38	<.0001
<b>Zone</b>	Amérique du nord	1	-0.5046	0.0303	-0.5640	-0.4451	276.95	<.0001
<b>Zone</b>	Asie	1	-0.5159	0.0273	-0.5694	-0.4624	357.26	<.0001
<b>Zone</b>	Europe hors France	1	0.8623	0.0599	0.7450	0.9796	207.56	<.0001
<b>Zone</b>	France	1	-0.2024	0.0390	-0.2789	-0.1260	26.94	<.0001
<b>Zone</b>	Reste du monde	1	0.1269	0.0808	-0.0315	0.2853	2.46	0.1165
<b>Profil</b>	Affaire	1	0.9336	0.0759	0.7849	1.0823	151.40	<.0001
<b>Profil</b>	Etude	1	1.2122	0.0359	1.1419	1.2825	1142.67	<.0001
<b>Profil</b>	Globe trotter	1	0.2872	0.0418	0.2053	0.3691	47.20	<.0001
<b>Profil</b>	Visa Schengen	1	0.0014	0.0414	-0.0797	0.0826	0.00	0.9723
<b>Scale</b>		0	1.0000	0.0000	1.0000	1.0000		

### 2. Au titre de la sévérité

Nous avons fait une tentative d'estimation du coût moyen par un modèle de régression log-normal. L'algorithme ne converge pas.

Tableau 34: Statistiques d'ajustement du MLG Log-normal « initial » du coût moyen

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
<b>Deviance</b>	337	21 754 202 313	64 552 529
<b>Scaled Deviance</b>	337	40	0
<b>Log Likelihood</b>		-3 852	
<b>AIC (smaller is better)</b>		7 730	
<b>AICC (smaller is better)</b>		7 731	
<b>BIC (smaller is better)</b>		7 780	
WARNING: Negative of Hessian not positive definite.			

## D. Test d'adéquation de loi

Pour tester l'ajustement de ces lois, la statistique utilisée sera celle de Kolmogorov-Smirnov.

Si  $F$  est la fonction de répartition des coûts de sinistres  $(X_1, \dots, X_n)$  et  $F_0$  la fonction de répartition de la loi supposée, on veut tester :

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$$

$F$  et  $F_0$  sont continues.

Après avoir classé les  $X_i$  par ordre croissant, on mesure grâce à la statistique de Kolmogorov Smirnov la différence maximale entre la fonction de répartition empirique des  $X_i$  et la fonction de répartition de la loi testée :

$$D = \max_{1 \leq k \leq n} \left[ \max \left\{ \frac{k}{n} - F_0(X_k); F_0(X_k) - \frac{k-1}{n} \right\} \right]$$

Nous allons utiliser le théorème suivant :

Si  $F_0$  est continue, alors pour  $x > 0$  :

$$\lim_{n \rightarrow \infty} P_0(\sqrt{n} * D \geq x) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2 x^2)$$

Nous rejetterons  $H_0$  au niveau  $\alpha$ , si  $D$  est supérieure au quantile d'ordre  $1-\alpha$  de la loi de  $D$

## E. Modèle de tarification actuel

### 1. Présentation des fréquences et des coefficients linéaire de durée (méthode actuel)

Tableau 35 : Coefficients et fréquences tarifaires

Motif du voyage	Nombre d'assurés	Destination géographique	Fréquence
Affaire	na01-02	Afrique	1,71%
Affaire	na01-02	Amerique Autre	0,43%
Affaire	na01-02	Amérique du nord	0,43%
Affaire	na01-02	Asie	0,11%
Affaire	na01-02	Europe hors France	0,43%
Affaire	na01-02	France	2,55%
Affaire	na01-02	Maghreb & Reste du monde	0,03%
Affaire	na02-03	Afrique	0,37%
Affaire	na02-03	Amerique Autre	0,37%
Affaire	na02-03	Amérique du nord	5,82%
Affaire	na02-03	Asie	0,37%
Affaire	na02-03	Europe hors France	0,37%
Affaire	na02-03	France	0,37%
Affaire	na02-03	Maghreb & Reste du monde	0,37%
Affaire	na03-inf	Afrique	0,41%
Affaire	na03-inf	Amerique Autre	0,41%
Affaire	na03-inf	Amérique du nord	0,41%
Affaire	na03-inf	Asie	0,41%
Affaire	na03-inf	Europe hors France	0,41%
Affaire	na03-inf	France	0,41%
Affaire	na03-inf	Maghreb & Reste du monde	0,41%
Etude	na01-02	Afrique	2,44%
Etude	na01-02	Amerique Autre	1,83%
Etude	na01-02	Amérique du nord	2,47%
Etude	na01-02	Asie	2,44%
Etude	na01-02	Europe hors France	1,75%
Etude	na01-02	France	1,75%
Etude	na01-02	Maghreb & Reste du monde	1,01%
Etude	na02-03	Afrique	1,74%
Etude	na02-03	Amerique Autre	1,74%
Etude	na02-03	Amérique du nord	1,74%
Etude	na02-03	Asie	1,74%
Etude	na02-03	Maghreb & Reste du monde	1,74%
Etude	na03-inf	Asie	1,74%
Etude	na03-inf	Europe hors France	1,74%
Etude	na03-inf	Maghreb & Reste du monde	1,74%
Globe trotter	na01-02	Europe hors France	4,17%
Globe trotter	na01-02	Maghreb & Reste du monde	4,33%
Globe trotter	na02-03	Europe hors France	4,17%
Globe trotter	na02-03	Maghreb & Reste du monde	2,14%
Globe trotter	na03-inf	Europe hors France	4,17%
Globe trotter	na03-inf	Maghreb & Reste du monde	4,17%
Vacance & Visa Schengen	na01-02	Afrique	0,48%
Vacance & Visa Schengen	na01-02	Amerique Autre	1,06%
Vacance & Visa Schengen	na01-02	Amérique du nord	0,84%
Vacance & Visa Schengen	na01-02	Asie	0,95%
Vacance & Visa Schengen	na01-02	Europe hors France	0,76%
Vacance & Visa Schengen	na01-02	France	2,61%
Vacance & Visa Schengen	na01-02	Maghreb & Reste du monde	0,85%
Vacance & Visa Schengen	na02-03	Afrique	0,39%
Vacance & Visa Schengen	na02-03	Amerique Autre	0,82%
Vacance & Visa Schengen	na02-03	Amérique du nord	1,36%

Vacance & Visa Schengen	na02-03	Asie	2,00%
Vacance & Visa Schengen	na02-03	Europe hors France	1,34%
Vacance & Visa Schengen	na02-03	France	4,20%
Vacance & Visa Schengen	na02-03	Maghreb & Reste du monde	1,21%
Vacance & Visa Schengen	na03-inf	Afrique	2,47%
Vacance & Visa Schengen	na03-inf	Amerique Autre	1,01%
Vacance & Visa Schengen	na03-inf	Amérique du nord	1,78%
Vacance & Visa Schengen	na03-inf	Asie	2,35%
Vacance & Visa Schengen	na03-inf	Europe hors France	5,22%
Vacance & Visa Schengen	na03-inf	France	2,94%
Vacance & Visa Schengen	na03-inf	Maghreb & Reste du monde	1,98%

Tableau 35 : Coefficients du lissage tarifaire pour la durée

Affaire		Vacance&Schengen		Etude		Globe trotter	
Duree	coeff	Duree	coeff	Duree	coeff	Duree	coeff
0_8	0,8	0_8	0,6	0_15	0,5	89_93	0,5
8_15	1	8_15	1	15_89	0,7	93_182	1
15_22	1,1	15_22	1,2	89_93	1	182_sup	2
22_32	1,2	22_32	1,4	93_182	1,2		
32_62	1,3	32_62	1,6	182_sup	1,5		
62_89	1,4	62_89	1,8				
89_93	1,5	89_93	1,9				

## 2. Résultats du MLG sur 3 variables sans la durée du voyage

Tableau 35 : Test de significativité des variables retenues pour le modèle à risque proportionnel

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
NA	2	250.63	<.0001
Zone_Regroupe_F	6	4124.56	<.0001
Profil_Regroupe_F	3	4557.47	<.0001

Tableau 36: Coefficients estimés dans le cadre du modèle à risque proportionnel

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	4.9117	0.0231	4.8664	4.9569	45327.5	<.0001

NA	na02-03	1	-0.2659	0.0223	-0.3097	-0.2221	141.69	<.0001
NA	na03-inf	1	-0.4489	0.0357	-0.5189	-0.3788	157.85	<.0001
Zone_Regroupe_F	Afrique	1	-0.2950	0.0417	-0.3768	-0.2133	49.98	<.0001
Zone_Regroupe_F	Amerique Autre	1	-0.2735	0.0329	-0.3379	-0.2090	69.23	<.0001
Zone_Regroupe_F	Amerique du nord	1	-0.4115	0.0294	-0.4691	-0.3540	196.13	<.0001
Zone_Regroupe_F	Asie	1	-0.4941	0.0263	-0.5457	-0.4425	352.40	<.0001
Zone_Regroupe_F	Europe hors France	1	0.7739	0.0593	0.6576	0.8902	170.20	<.0001
Zone_Regroupe_F	France	1	-1.3147	0.0254	-1.3644	-1.2650	2687.26	<.0001
Profil_Regroupe_F	Affaire	1	0.8778	0.0746	0.7316	1.0239	138.56	<.0001
Profil_Regroupe_F	Etude	1	-0.6292	0.0207	-0.6698	-0.5885	919.82	<.0001
Profil_Regroupe_F	Globe trotter	1	-1.6956	0.0276	-1.7496	-1.6416	3787.37	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

## F. Synthèse des points d'intention réglementaire pour le secteur d'assurance

La réglementation du secteur d'assurance français est très stricte. Elle comporte des textes de bases qui codifient les activités standards admis sur le marché national et les acteurs officiels reconnus par les autorités de contrôle tel que l'ACP (Autorité de contrôle prudentiel). Au fur et à mesure de l'évolution de la conjoncture économique nationale et internationale, des nouveaux textes ont été récemment ajoutés dans l'objectif d'assurer l'actualisation de cette réglementation. La stratification peut être abordée de manière suivante.

### **Les textes assortis de la réglementation communautaire**

- Directive 2012/23/UE du Parlement européen et du Conseil du 12 septembre 2012 sur le report de la date de transposition et d'entrée en application de la directive sur l'accès aux activités d'assurance et de la réassurance et leur exercice
- Règlement n°1215/2012 du Parlement européen et du Conseil du 12 décembre 2012 sur la compétence judiciaire, la reconnaissance et l'exécution des décisions en matière civile et commerciale
- Règlement 524/2013 du Parlement européen et du Conseil du 21 mai 2013 relatif au règlement en ligne des litiges de consommation via la création d'une plateforme européenne en ligne et modifiant le règlement CE 2006/2004 et la directive 2009/22/CE

### **Les textes nouvellement adoptés assortis de la réglementation nationale (2012 et 2013)**

- Décret 2013-100 du 26 janvier 2012 et arrêté du 1<sup>er</sup> mars 2012 relatifs aux modalités de l'inscription sur le registre de l'ORIAS
- Lignes directrices de l'ACP d'avril 2012 relatives à la relation d'affaires et au client occasionnel
- Instruction ACP 2012-I-07 du 13 décembre 2012 relative au questionnaire sur l'application des

règles de protection de la clientèle

- Loi n° 2012-300 du 5 mars 2012 relative aux recherches impliquant la personne humaine
- Décret n° 2012-548 du 23 avril 2012 relatif au fonds de garantie des dommages consécutifs à des actes de prévention, de diagnostic ou de soins dispensés par des professionnels de santé
- Décret 2013-278 du 2 avril 2013 relatif aux modalités d'évaluation des besoins d'assistance par une tierce personne
- Délibération CNIL 2013-212 du 11 juillet 2013 sur les traitements automatisés de données à caractère personnel relatifs à la passation, la gestion et l'exécution des contrats mis en œuvre par les organismes d'assurance, de capitalisation, de réassurance, d'assistance et par leurs intermédiaires

Comme les produits d'assurance concernent les différents secteurs, activités professionnelles et activités personnelles, les textes nouvellement adoptés touchent les nombreux volets de l'activité économique tel les finances, l'environnement, l'emploi, etc. Cela implique aux entreprises d'assurance la recherche constante de la maîtrise de l'aspect juridique de leur environnement et l'actualisation de leurs produits conformément à l'évolution technique du métier

## G. Passage du document de la CAS

Les analyses de sinistres à l'aide de GLM requièrent un certain volume d'historique. Variables selon la fréquence de sinistralité et le nombre de critères retenus, des résultats pertinents sur des portefeuilles de particuliers peuvent généralement être obtenus avec une exposition de l'ordre de 100 000 (par exemple 50 000 par année sur deux ans, etc.). On peut parfois obtenir des résultats significatifs sur des volumes de données plus faibles (en particulier sur des typologies de sinistres avec une sinistralité suffisante), mais il est préférable d'avoir une exposition de l'ordre de quelques centaines de milliers. Comme un modèle ajusté sur une seule année pourrait présenter un biais dû à des événements spécifiques à cette année, il est préférable d'avoir une base de données comprenant deux ou trois ans d'historique.