Mémoire présenté devant l'ENSAE ParisTech

pour l'obtention du diplôme de la filière Actuariat

et l'admission à l'Institut des Actuaires

le 22/02/2017

Par : **QUATRE Tristan**

Titre: **Développement d'un scoring et de modèles de tarification en assurance-crédit PME.**

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

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

Membres présents du jury de la filière

Entreprise :AXA ASSURCREDIT

Nom : M. Lionel STEMPERT
Signature :

Membres présents du jury de l'Institut des Actuaires

Directeur de mémoire en entreprise :

Nom : M. William BAPTISTE
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

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
3, Avenue Pierre Larousse – 92245, MALAKOFF CEDEX, France

Remerciements

Je souhaite remercier Monsieur Pierre Fournel, Président du Directoire, ainsi que Monsieur Lionel Stempert, son successeur, de m'avoir accueilli au sein d'AXA Assurcrédit et proposé ce sujet.

Je remercie également Monsieur William Baptiste pour le partage de son expertise ainsi que ses précieux conseils au quotidien.

Je tiens aussi à remercier Monsieur Frédéric Eichel, actuaire chez AXA Entreprises, et Monsieur Emmanuel Guiffart, actuaire chez Coface, pour la qualité de leurs suggestions, notamment sur des aspects techniques.

Enfin, je remercie les membres du directoire et l'ensemble des équipes d'AXA Assurcrédit pour le temps qu'ils m'ont consacré ainsi que toutes les explications qu'ils ont pu me fournir sur leur activité et leur métier.

Résumé

L'activité d'assurance-crédit consiste à couvrir les clients contre le risque de défaut de paiement de leurs acheteurs. L'assureur, dans ce cadre, remplit également une fonction de conseil pour l'assuré dans la sélection de ses clients.

Son premier rôle est donc la sélection, dans le portefeuille clients de l'assuré, des risques que l'assureur va vouloir couvrir. Arrivera ensuite la tarification du produit d'assurance contre le risque d'impayés des acheteurs de l'assuré qui auront été agréés.

Nous décrivons dans un premier temps l'élaboration d'un modèle de *scoring* des entreprises françaises spécialisé sur le segment des PME, cœur de métier d'AXA Assurcrédit. A cet effet, après une étape nécessaire de traitement des données, nous mettons en place un premier modèle de *scoring* financier pour ensuite intégrer des variables non financières en vue d'élaborer le modèle de *scoring* final. Le score permet ensuite de sélectionner les acheteurs qui seront couverts par l'assureur.

Nous mettons ensuite en place un premier modèle de tarification dérivé du modèle de *scoring*. Pour ce faire, nous utilisons les probabilités de défaillance précédemment calculées pour en déduire une exposition au défaut en fixant des hypothèses quant à l'utilisation des limites de crédit accordées par l'assureur.

Nous travaillons également un second modèle de tarification, basé sur le « modèle collectif », pour mettre en place une tarification qui pourra être appliquée aux prospects de la compagnie pour lesquels nous ne disposons pas de l'exhaustivité des acheteurs.

Mots clefs : Assurance-crédit, scoring, régression logistique, probabilités de défaut, Use Factor, Exposition au défaut, Taux de perte, Modèle collectif, GLM

Abstract

Credit insurance business covers customers against the payment default of their buyers. The insurer, in this context, also has an advisory function: it advises the insured in the selection of its customers.

Its first task is to select, in the insured's customers portfolio, the risks that the insurer is willing to cover. Then, it seeks to price the insurance product against the risk of default of the approved insured's buyers.

This report first describes the development of a scoring model of French companies, specialised in SMEs, which are at the core of AXA Assurcrédit's business. To achieve this, after first processing data, the report fits a financial scoring model. It integrates non-financial variables to develop the final scoring model. The score is then able to select buyers which will fit the underwriting policy of the insurer.

The report then sets up a first pricing model derived from the scoring model. It uses the previously calculated default probabilities to deduce year exposure to default by setting assumptions about the use of the credit limits granted by the insurer.

The report also works on a second pricing model, based on the collective model, to implement a pricing that can be applied to the prospects of the business for which no complete buyers' portfolio is available.

Keywords: Credit Insurance, scoring, logistic regression, default probabilities, Use Factor, Exposure At Default, Loss Given Default, Collective model, GLM.

Sommaire

Introduction

Nous nous intéressons, dans ce mémoire, à l'élaboration d'outils de *scoring* et tarification en assurance-crédit. En effet, Axa Assurcrédit souhaite refondre sa tarification pour être plus sélectif et plus agressif dans les risques qu'elle couvre, mais également pour pouvoir déléguer une partie de la souscription au réseau. Nous décrirons les étapes de construction des différents modèles.

Après avoir présenté l'assurance-crédit et ses caractéristiques, nous mettrons en place le modèle de *scoring*, premier outil nécessaire à la sélection des risques. Celui-ci va permettre à l'assureur d'arbitrer (c'est-à-dire accepter ou refuser de couvrir) les risques du portefeuille d'acheteurs de l'assuré.

A partir de ce modèle de *scoring* et des probabilités de défaillance qui en découlent, nous mettrons en place un premier modèle de tarification basé sur le calcul de l'exposition au risque de défaut et de la perte attendue.

Ce premier modèle de tarification nécessitant de connaître l'intégralité du portefeuille d'acheteurs de l'assuré, nous utiliserons le modèle collectif pour développer une seconde tarification basée sur l'information disponible dans les questionnaires de souscription.

Pour chacun des deux modèles de tarification, nous définirons le passage de la prime pure à la prime totale afin de mesurer l'impact de ces tarifications par rapport aux primes actuellement perçues.

1 Présentation de l'activité et de la problématique

1.1 L'assurance-crédit

Un assureur crédit propose à ses clients de les couvrir contre le risque d'impayés de leurs acheteurs. Les produits assurent donc le crédit inter-entreprises.

Le crédit inter-entreprises est la somme des crédits clients que les entreprises non financières se consentent entre elles. Le plus souvent, le crédit inter-entreprises nait du décalage de paiement que s'accordent les sociétés entre elles. C'est une source de financement pour l'entreprise cliente, censée donner, à l'origine, le temps au client de vérifier la qualité des biens ou services fournis, mais c'est également un risque pour le fournisseur.

Historiquement, le crédit client est né au Moyen Âge dans les pays d'Europe du Sud. Cette région, très marquée par le catholicisme, était soumise à l'interdiction religieuse du prêt avec intérêt. Le développement du commerce s'est alors accompagné de celui du crédit client permettant de tenir compte des délais d'acheminement et de commercialisation des marchandises.

La loi de modernisation de l'économie de 2008 (LME) définit des délais maximums de paiement et des sanctions en cas de non-respect de ceux-ci, les entreprises portent néanmoins un risque de crédit au bilan lié aux crédits inter-entreprises.

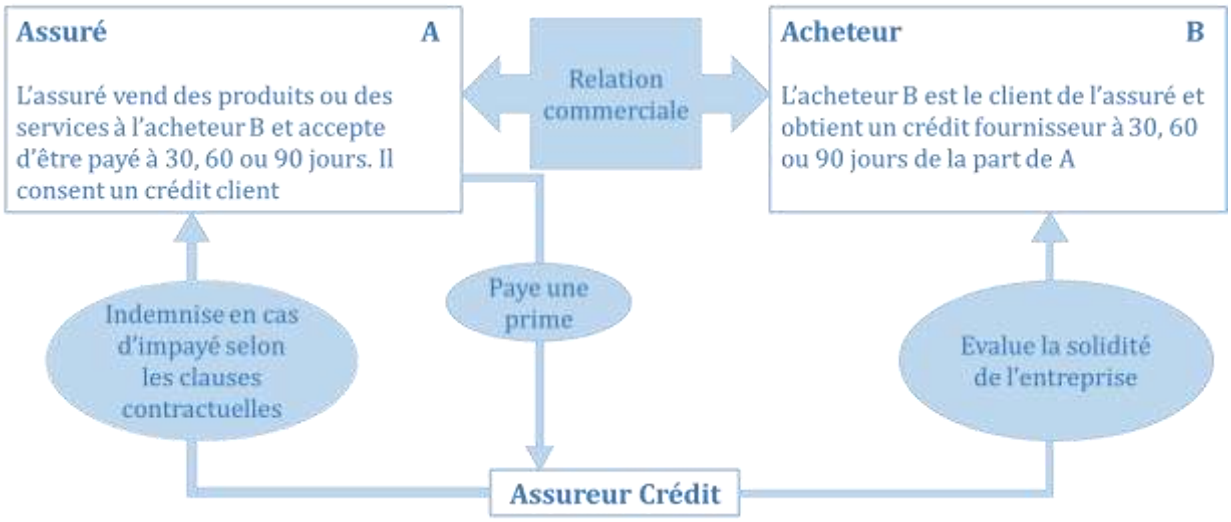
L'une des possibilités dont dispose une entreprise pour réduire ce risque est de souscrire une assurance-crédit qui aura deux objectifs principaux :

- L'assurer sur tout ou partie de son risque lié aux crédits consentis à ses acheteurs ;
- Lui faire bénéficier d'outils de gestion du risque de crédit de ses acheteurs.

Les autres possibilités de gérer son risque clients sont soit l'affacturage (financement par cession des factures à un factor), soit choisir de ne pas s'assurer (ou « s'auto-assurer ») et alors gérer en interne le risque de crédit lié aux acheteurs, éventuellement en s'abonnant à une source d'information censée alerter en cas de dégradation de la qualité du risque crédit. Cette dernière option nécessite qu'une entreprise se dote d'une équipe de crédit management ainsi que d'outils permettant de suivre ses risques.

Par ailleurs, les assureurs crédit suivent et traitent l'information financière. Certaines grandes entreprises se tournent vers les assureurs crédits, non pas pour bénéficier des produits d'assurance, mais de leur analyse crédit et du traitement qu'ils font d'informations confidentielles grâce à leur propre base de clientèle.

Fonctionnement de l'assurance crédit



L'assureur crédit propose de couvrir le risque acheteur de la société assurée, selon une limite fixée par l'assureur.

Les contrats d'assurance-crédit présentent différentes options. Le plus souvent, l'assuré déclare auprès de son assureur ses nouveaux clients afin d'obtenir un agrément (c'est-à-dire une limite de crédit pour laquelle il aura l'assurance que cet acheteur sera couvert en cas de défaut de paiement sur des factures émises, tant que cet agrément est maintenu). Les contrats font par ailleurs mention d'une quotité garantie qui fixe le pourcentage de l'agrément qui sera effectivement indemnisé, également une limite de décaissement sur l'ensemble de la police (multiple de la prime). Si un acheteur n'est pas agréé, l'assuré n'est alors pas couvert contre un défaut de paiement de ce dernier, sauf s'il dispose d'une clause d'auto-arbitrage (ou clause de « non dénommés » sur laquelle nous reviendrons).

Ce système d'agréments monitorés en permanence permet, d'une part à l'assuré d'avoir une vision sur la santé financière de ses potentiels acheteurs ; d'autre part à l'assureur de gérer de façon dynamique, pendant la durée de vie du contrat, sa sinistralité potentielle.

Un contrat d'assurance-crédit fait ainsi l'objet d'une gestion active après sa souscription. En effet, l'assureur va devoir étudier chacun des nouveaux acheteurs de son client pour accorder ou non la couverture (délivrer l'agrément). Les acheteurs déjà couverts sont monitorés et leurs agréments peuvent faire l'objet de modification en cours de contrat (réduction ou résiliation).

L'assureur crédit peut ainsi gérer sa sinistralité en revoyant sa politique d'agrément pendant la durée du contrat. Il peut alors avoir une politique de sélection plus restrictive en annulant ou limitant les agréments déjà octroyés. Le rôle de l'assureur crédit est donc également le monitoring des acheteurs déjà garantis. L'arbitre (analyste crédit) va avoir un rôle actif tout au long de la durée de la police en permettant de réduire l'intensité du risque (fréquence et coût).

Le contrat va ainsi indiquer la nature des opérations assurées, les montants garantis et les services associés (recouvrement par exemple). La prime tient compte de ces éléments et sera souvent forfaitaire pour les petites entreprises et fixée en fonction d'un pourcentage du chiffre d'affaires ou de la masse assurable (part du chiffre d'affaires faisant l'objet du contrat d'assurance). Ce pourcentage varie généralement entre 0,1% et 1% (Banque de France 2013).

Le marché est relativement oligopolistique avec trois acteurs principaux contrôlant plus de 80% du marché mondial hors Chine (Banque de France 2012) :

- Euler Hermès, filiale du groupe allemand Allianz, numéro un mondial avec 34% de part de marché,
- Atradius, filiale du groupe espagnol Catalana Occidente, numéro 2 mondial avec 28% de part de marché,
- Coface, filiale du groupe français Natixis, numéro 3 mondial avec 20% de part de marché.

Les trois premiers acteurs mondiaux totalisent ainsi 82% de part de marché. En France, en plus de ces trois acteurs sont présents Groupama sur le créneau de l'agro-alimentaire et AXA Assurcrédit, spécialisé sur les PME/TPE.

Le marché français de l'assurance-crédit est mature, extrêmement concurrentiel et concentré. La concurrence est donc essentiellement liée aux prix, d'autant plus que la souscription par les entreprises de ce produit, contrairement à beaucoup de produits d'assurance dommage ou de responsabilité civile, n'est pas obligatoire. Pour autant, sur le segment PME/TPE, une grande majorité des entreprises, 90% environ, entrant dans le champ des risques de l'assurance-crédit (commerce BtoB hors groupe et hors entités publiques) restent encore non détentrices d'une police d'assurance-crédit.

Les assureurs connaissent une baisse de leur ratio sinistres / primes sur ces dernières années alors que les défaillances d'entreprises restent à des niveaux très élevés même si la tendance récente est orientée à la baisse. La conséquence est un sentiment, de la part des assurés, de porter seuls le risque alors qu'ils détiennent une police d'assurance. Ceci résulte du fait de nombreux refus ou restrictions de garanties. Ceci entraine un comportement d'auto-assurance des entreprises qui choisissent alors de ne plus souscrire de contrat. Les chiffres d'affaires des assureurs crédit sont donc aussi en décroissance.

Axa Assurcrédit est filiale du groupe AXA à hauteur de 60% et du groupe Coface pour les 40% restant. L'entreprise utilise des moyens informatiques de Coface et notamment son *scoring* (sous forme de notes de 1 à 10). Une partie de l'arbitrage est également sous traitée à Coface.

Axa Assurcrédit, dans ce contexte, souhaite refondre sa tarification pour être plus sélective et plus agressive dans l'approche tarifaire des risques qu'elle accepte en fonction des secteurs d'activité ou des qualités de risques clients. Elle souhaite également pouvoir

déléguer une partie de la souscription auprès de ses commerciaux et intermédiaires en leur fournissant un outil de cotation automatique.

1.2 L'offre d'AXA Assurcrédit

1.2.1 Les types de garanties accordées

Trois types de garanties sont accordées via les différents contrats d'assurance crédit :

- Les crédits « Non Dénommés » (ND) : ce sont des crédits qui ne dépassent pas, par acheteur, une limite définie à la signature du contrat. Ces crédits sont assurés à hauteur de cette limite sans que l'assuré ait besoin d'interroger l'assureur pour avoir un agrément. La quotité garantie est faible sur le ND (le plus souvent 60%).
  Généralement, le plafond des crédits ND est de 5 000€, c'est-à-dire que l'assuré peut traiter avec autant de clients qu'il le souhaite pour un montant inférieur ou égal à cette limite, il bénéficiera alors d'une couverture pour 60% des montants traités en cas de défaut de l'acheteur.
  Une limite maximum de décaissement au titre des crédits « non dénommés » est en général intégrée au contrat.

- Les crédits « Non Dénommés Surveillés » (NDS) : ce sont des crédits qui ne dépassent pas, par acheteur, un seuil défini à la signature du contrat. Mais contrairement aux ND, ils sont soumis à accord de l'assureur. L'assuré va donc demander, pour chacun de ses acheteurs dont l'encours est inférieur au seuil déterminé, l'accord pour être couvert. Le NDS est assimilable à un « agrément express » car c'est un automate qui acceptera ou non la limite de crédit NDS sur la base du score de l'acheteur (score fourni par le modèle de scoring).
  Dans le cadre du contrat Global Kup, le seuil du NDS est de 5 000€ et la quotité garantie associée de 80%. Les assurés n'ont donc qu'à déclarer le nom du client à leur assureur ; celui-ci donnera alors ou non son accord pour couvrir cet acheteur. En cas de défaut, l'assuré est couvert à hauteur de 80% de ses factures dans la limite de 4 000€ (80% de quotité garantie avec 5 000€ de limite).

- Les crédits « dénommés » (Agréments) : cette catégorie regroupe le reste des crédits de l'assuré. Ils font l'objet d'une demande d'agrément de l'assuré auprès de l'assureur pour le montant souhaité par l'assuré. L'assureur va alors y répondre favorablement, réduire le montant ou refuser de couvrir cet acheteur. C'est un arbitre (analyste crédit) qui traite la demande sur la base de son analyse et du traitement qu'il va faire de l'information disponible.
  A titre d'exemple, un assuré demandera un agrément pour un acheteur X et pour un montant de 10 000€. L'assureur peut accepter les 10 000€, réduire ou refuser ce montant. En cas d'accord, l'assuré sera couvert contre le défaut de l'acheteur X à

hauteur de 9 000€ (90% de quotité garantie x 10 000€ d'agrément), la quotité garantie étant souvent fixée à 90% contractuellement pour les agréments.

La quotité garantie permet de limiter le risque à un pourcentage de l'agrément. Maintenir cette quotité garantie inférieure à 100% permet à l'assureur d'éviter de subir un aléa moral avec ses assurés, en faisant porter une part du risque par l'assuré lui-même.

La limite de décaissement (l'indemnisation est limitée à 20 ou 30 fois la prime selon les contrats) permet de plafonner les sinistres graves ou l'impact d'une fréquence de sinistres élevée.

1.2.2 Les principaux produits

1.2.2.1 Global Kup

Ce produit fait l'objet d'une grille tarifaire simple, fonction de la tranche du chiffre d'affaires de l'assuré. Le produit propose essentiellement une couverture via des « agréments express » dans la limite de 5 000 € (ces agréments sont alors accordés ou refusés par l'automate en fonction du score de la société concernée). L'assuré peut opter pour un nombre défini d'agréments (accord sur un acheteur avec une limite) de 25 à 500. Le tarif va alors en tenir compte pour refacturer à l'assuré les frais de suivi des acheteurs soumis à agrément. La prime sera composée alors d'une prime d'assurance de base, fonction du chiffre d'affaires, à laquelle sont ajoutés les coûts d'enquête et surveillance des acheteurs faisant l'objet d'une demande de limite de crédit.

A la souscription, un assuré va indiquer son chiffre d'affaires assurable (celui issu de l'activité avec d'autres entreprises hors activité B2C, hors intra-groupe et hors paiement cash), il sera alors positionné dans une classe tarifaire (il existe onze classes actuellement de 0 à 7,5M€). L'assuré a ensuite la possibilité de sélectionner le nombre d'agréments qu'il envisage de demander dans l'année. En dessous de 5 000 € d'encours, la demande n'est pas soumise à agrément mais à une autorisation en « non dénommés surveillés » (agrément express). Au-delà de 5 000 €, l'assuré doit demander un agrément. Il y a donc une notation, issue du score, utilisée par un automate qui accepte ou refuse les agréments express. Il y a également une notation, liée au score, mais aussi issue d'une analyse de l'acheteur effectuée par l'arbitre qui doit se prononcer sur l'agrément demandé.

L'agrément automatique des demandes en NDS est basé sur un score qui n'est mis à jour que lors de la publication d'informations de la part de l'acheteur. Il s'agit le plus souvent de la publication des comptes annuels, le cas échéant d'informations sur des procédures légales ou des impayés constatés par la Banque de France à travers son system FIBEN, auquel les assureurs crédit et les banques ont accès, ou des défauts constatés sur d'autres fournisseurs par des agences de recouvrement.

L'agrément manuel fait l'objet d'un suivi plus approfondi et plus dynamique par les arbitres qui vont exploiter toute information disponible sur les acheteurs.

Tout est compris dans le tarif, il n'y a donc pas pour l'assuré de frais de suivi supplémentaires liés aux agréments ou aux NDS. L'assuré peut demander une limite de 5 000 € sur autant de NDS qu'il le souhaite et sans frais supplémentaires. Ce produit est donc également utilisé par les assurés pour interroger le modèle de *scoring* de l'assureur sur la qualité du crédit d'un acheteur potentiel. Ce dernier point est important car il devra être pris en compte dans le calcul de la véritable exposition au défaut à partir du calcul du cumul des encours garantis dans la mesure où certaines limites de crédit accordées ne correspondent pas à de réels encours de risque encourus par l'assureur.

La limite (5 000 € pour le NDS) est soumise à une quotité garantie. Celle-ci est de 80% pour le NDS et 90% pour les agréments.

L'assureur fixe également une limite maximum d'indemnisation de vingt fois la prime.

1.2.2.2 Globale

C'est le principal produit vendu par Axa Assurcrédit. Les paramètres et options sont beaucoup plus nombreux que sur la police Global Kup :

- Certains contrats (20%) disposent d'une option « non dénommés » (ND) pour des encours généralement inférieurs à 5 000 € (dans le cas standard). De sorte que l'assuré n'a pas besoin de déclarer ses factures ou clients pour cette partie.
- Certains contrats disposent de NDS ou Agréments Express.
- L'option principale est l'agrément. Chaque agrément fait l'objet de frais facturés auprès de l'assuré.
- La quotité garantie est fixée par les clauses particulières et varie de 60% à 100% (généralement 60% pour le ND, 80% pour le NDS et 90% pour les agréments).
- Le maximum d'indemnité annuelle représente vingt ou trente fois la prime.

Le tarif fait l'objet d'une négociation commerciale entre l'assuré et l'assureur. La qualité de crédit de l'assuré est alors prise en compte, le plus souvent via son secteur, et celui de ses acheteurs via une étude d'une sélection de sa clientèle.

Le produit « Globale » peut être tarifé de deux manières par les souscripteurs, soit via une prime forfaitaire déterminée en début d'année, soit via un taux de prime et un montant minimum à payer en forme d'acompte, qui sera ensuite ajusté en fonction du chiffre d'affaires réellement assuré multiplié par le taux de prime.

Le tarif sera exprimé en taux appliqué au chiffre d'affaires assuré pour la prime classique et en montant fixe pour la prime forfaitaire. Les déterminants de la prime sont les mêmes. Dans le cas d'une prime non forfaitaire, un minimum de prime à payer est défini avec un

ajustement en fin de période en fonction du chiffre d'affaires déclaré par l'assuré. Ce minimum à payer peut-être déconnecté du chiffre d'affaires dans le cadre d'une prime non forfaitaire.

1.2.2.3 Global Kup Excess

Ce contrat ne fera pas l'objet d'étude dans ce rapport. Il s'agit d'une couverture d'un excédent de perte annuelle au-delà d'une franchise élevée (environ 4 à 5 fois le montant de la prime). Il a été relativement peu vendu pour l'instant.

2 Le modèle de *scoring*

La défaillance légale regroupe les procédures de liquidation et de redressement judiciaires. En tout état de cause, c'est une cessation de paiement qui va provoquer le déclenchement de l'une ces procédures et ainsi activer les clauses du contrat de l'assureur crédit sur cette contrepartie.

Pour un assureur crédit, le développement d'un score va avoir une double utilité :

- Pouvoir classer les risques sur lesquels les assurés souhaitent être couverts et accepter ou non de les assurer. C'est sur la base du score, puis d'une analyse crédit si besoin est, que l'assureur va accorder, réduire ou refuser l'agrément demandé par l'assuré. C'est également l'évolution de ce score qui va permettre à l'assureur de gérer de façon dynamique ses risques couverts et éventuellement réduire ou retirer des agréments préalablement accordés s'il s'est produit un changement chez les acheteurs concernés. Chez AXA Assurcrédit, plus de 90% des limites sont accordées par un « automate » sur la base du score de l'acheteur.
- Le modèle de *scoring* va également permettre de calculer une probabilité de défaillance pour les entreprises couvertes ; elle sera ensuite utilisée dans le modèle de tarification.

Le modèle de *scoring* permet d'affecter un score à une entreprise. Le score est un indicateur synthétique auquel peut être associé un degré de défaillance. Le modèle de *scoring* va donc, par des méthodes quantitatives, permettre de quantifier le risque de défaillance d'une entreprise donnée.

L'objectif est d'obtenir un modèle capable d'évaluer précisément le risque associé à un acheteur (Boisselier et Dufour 2003). Il est donc nécessaire que cette estimation soit robuste temporellement, c'est-à-dire qu'elle dépende peu de l'échantillon utilisé. D'autre part, le modèle doit être accessible pour que ses résultats soient interprétables et utilisables facilement par les arbitres de l'assureur.

La régression logistique, modèle visant à expliquer une variable dichotomique, est toute indiquée puisqu'elle répond à ces exigences.

2.1 Méthodologie

2.1.1 La régression logistique

La variable à expliquer, Y, est dichotomique (soit il y a défaillance et elle prend la valeur de un, soit il n'y a pas défaillance et elle est égale à zéro).

Nous allons donc utiliser la régression logistique pour estimer l'ensemble des modélisations des différentes étapes de construction du *scoring*.

Nous cherchons à estimer la probabilité conditionnelle d'une variable dichotomique (Rakotomalala 2015) :

$$P(Y = y_k/X)$$

D'après le théorème de Bayes, nous pouvons l'exprimer de la façon suivante :

$$P(Y = y_k/X) = \frac{P(Y = y_k) \times P(X/Y = y_k)}{P(X)}$$

Soit

$$P(Y = y_k/X) = \frac{P(Y = y_k) \times P(X/Y = y_k)}{\sum_{k=1}^{k=K} P(Y = y_k) \times P(X/Y = y_k)}$$

Dans le cas où Y est une variable dichotomique, cela revient à la relation suivante :

$$\frac{P(Y = 1/X)}{P(Y = 0/X)} = \frac{P(Y = 1)}{P(Y = 0)} \times \frac{P(X/Y = 1)}{P(X/Y = 0)}$$

La règle de décision sera alors :

$$Si\ \frac{P(Y = 1/X)}{P(Y = 0/X)} > 1, alors\ Y = 1$$

Le rapport $\frac{P(Y=1/X)}{P(Y=0/X)}$ est simple à estimer car c'est le rapport entre le nombre d'observations égales à 1 (entreprises en défaillance dans notre cas) et le nombre d'observations égales à 0 (entreprises ne faisant pas défaut).

Pour estimer le rapport de probabilité, la régression logistique passe par un lien logarithmique :

$$ln\left[\frac{P(X/Y = 1)}{P(X/Y = 0)}\right] = b_0 + b_1X_1 + \cdots + b_JX_J \qquad (1)$$

Cette hypothèse fondamentale permet de couvrir un grand nombre de lois de distribution de données (loi normale, les lois exponentielles, les lois discrètes, les lois Béta, les lois Gamma, les lois de Poisson ainsi qu'un mélange de variables explicatives binaires et continues).

La régression logistique est une méthode semi-paramétrique car l'hypothèse porte uniquement sur le rapport des probabilités et non la distribution de chacune d'elle.

Nous allons passer par un modèle de type LOGIT. Nous posons pour cela la probabilité π :

$$\pi(X) = P(Y = 1/X)$$

Nous avons donc :

$$ln\left[\frac{\pi(X)}{1-\pi(X)}\right] = a_0 + a_1X_1 + \cdots + a_JX_J \qquad (2)$$

La quantité $\frac{\pi(X)}{1-\pi(X)}$ est un rapport de chance (appelé odds). S'il est égal à 0,5 ; cela indique qu'un individu a deux fois plus de chances d'être égal à 0 qu'à 1.

Cette relation aboutit à la fonction de répartition de la loi logistique :

$$\pi(X) = \frac{e^{a_0+a_1X_1+\ldots+a_JX_J}}{1+e^{a_0+a_1X_1+\ldots+a_JX_J}} = \frac{1}{1+e^{-(a_0+a_1X_1+\ldots+a_JX_J)}}$$

Il est à noter que nous avons les relations suivantes :

$$\begin{aligned} ln\left[\frac{\pi(X)}{1-\pi(X)}\right] &= a_0 + a_1X_1 + \cdots + a_JX_J \\ &= ln\left[\frac{P(Y=1)}{P(Y=0)} \times \frac{P(X/Y=1)}{P(X/Y=0)}\right] \\ &= ln\left[\frac{P(Y=1)}{P(Y=0)}\right] + ln\left[\frac{P(X/Y=1)}{P(X/Y=0)}\right] \\ &= ln\left[\frac{p}{1-p}\right] + \left(b_0 + b_1X_1 + \cdots + b_JX_J\right) \end{aligned}$$

Les relations (1) et (2) sont donc identiques à une constante près :

$$a_0 = ln\left[\frac{p}{1-p}\right] + b_0$$

L'estimation des paramètres sera faite par le maximum de vraisemblance.

2.1.2 Les données disponibles

Nous disposons de trois jeux de données :

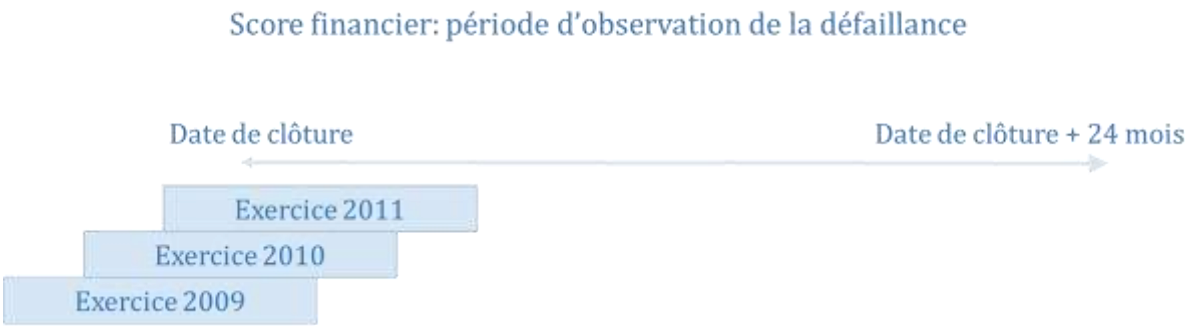
- Les bilans et comptes de résultats d'entreprises françaises sur 10 ans (2005 à 2015). Ce sont les champs des 4 premières pages du formulaire CERFA de déclaration des comptes annuels auprès du greffe (192 champs).
- Les défaillances d'entreprises sur la même période avec leur cause (redressement judiciaire ou liquidation judiciaire)
- Des variables de profil des entreprises françaises (Siren, Date de naissance du dirigeant, département de l'entreprise, date d'immatriculation de l'entreprise, date de création de l'entreprise, effectif de l'entreprise, le code APE du secteur et la situation juridique indiquant si la société est en redressement ou non à la date des données) à deux dates données qui nous serviront pour calibrer le score final.

En amont du calibrage du modèle, nous avons une étape importante de « nettoyage » et fiabilisation des données.

2.2 Le score financier

Le score financier va être calibré sur les comptes de 2009, 2010 et 2011 et les défaillances qui les suivent. L'objectif de cette première étape est d'obtenir un modèle permettant de déterminer une probabilité de défaillance uniquement à partir des données financières d'une entreprise (bilan et compte de résultat).

Nous choisissons d'observer les défaillances intervenant 24 mois suivant la date de clôture d'un exercice. Le point de départ est donc la date de clôture de chacune des sociétés dans une année donnée et dont les comptes sont disponibles. Cette observation peut se faire sur 12, 24, 36 mois. Il s'avère, en pratique, que le résultat est plus discriminant à 24 mois dans le choix des variables explicatives. La probabilité de défaut pourra ensuite être calculée sur l'horizon souhaité.



2.2.1 Traitement des données et sélection des variables explicatives

La première étape est celle du traitement des données disponibles ainsi que la sélection des données qui seront utilisées. Nous « nettoyons » les bases disponibles en excluant les sociétés dont les bilans et comptes de résultats sont insuffisants (avec trop de données manquantes) pour l'analyse ou celles présentant trop de valeurs aberrantes (total bilan ou chiffre d'affaires négatifs par exemple). Cette première étape nous conduit à retenir dans notre étude environ 750 000 comptes de sociétés par année sur les trois années servant à calibrer le score.

L'étude des comptes des entreprises est basée sur des ratios financiers. Une centaine de ratios peut être calculée à partir des données disponibles. Nous examinons pour chacun de ces ratios leur disponibilité, leur cohérence et leur comportement par rapport à la variable que nous souhaitons modéliser (la fréquence des défaillances).

Il est à noter que chaque ratio financier fait l'objet de tests sur les agrégats qui le composent. Ceci est nécessaire afin de ne pas prendre en compte des éléments qui pourraient être l'inverse des phénomènes que nous souhaitons mettre en exergue. À titre d'exemple, lorsque nous calculons le ratio *Résultat net / Fonds propres*, nous analysons le signe de l'agrégat *Résultat net* et celui de *Fonds Propres*. En effet, une entreprise en pertes et qui a des fonds propres négatifs est dans une situation financière très dégradée contrairement à une entreprise qui dégage un résultat avec des fonds propres positifs. Ces situations sont testées pour pouvoir les distinguer et ne pas prendre en compte un ratio qui mélangerait des réalités différentes.

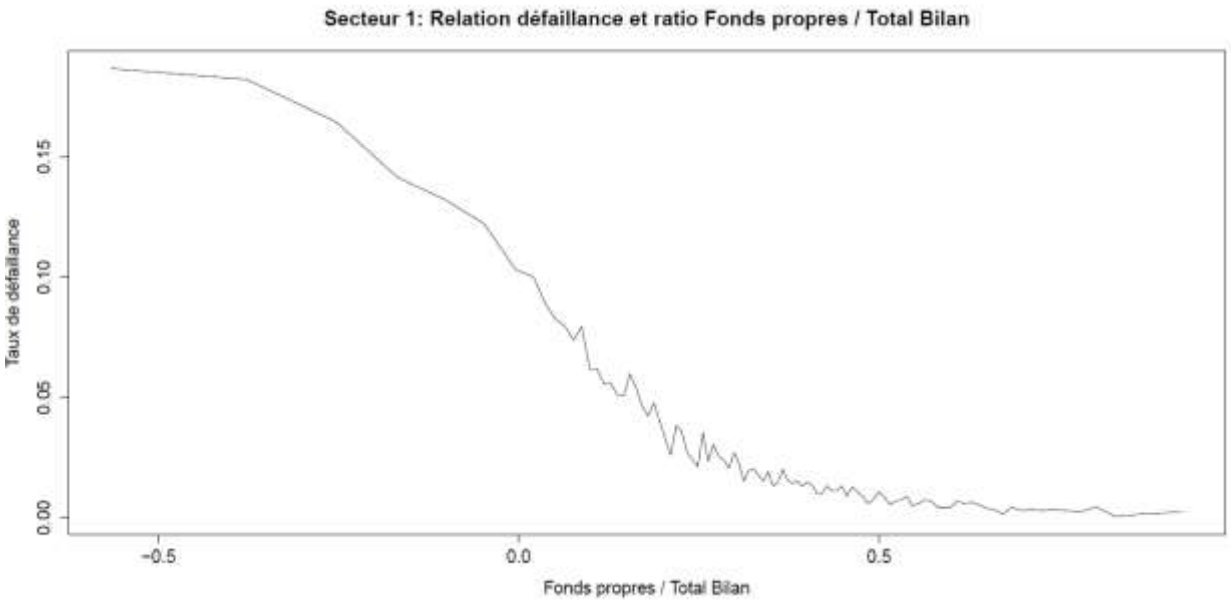
La première sélection de ratios est faite sur des critères financiers et une observation graphique pour identifier ceux qui influent sur la fréquence de sinistralité des entreprises.

Il est réalisé également dans cette phase une première segmentation sectorielle. Les ratios financiers n'ayant pas le même comportement selon le secteur d'appartenance peuvent avoir des conséquences sur les modélisations par la suite. Nous regroupons donc les secteurs dont les comportements sont similaires. Les critères de regroupement sont l'analyse graphique et la proximité des taux de défaillance.

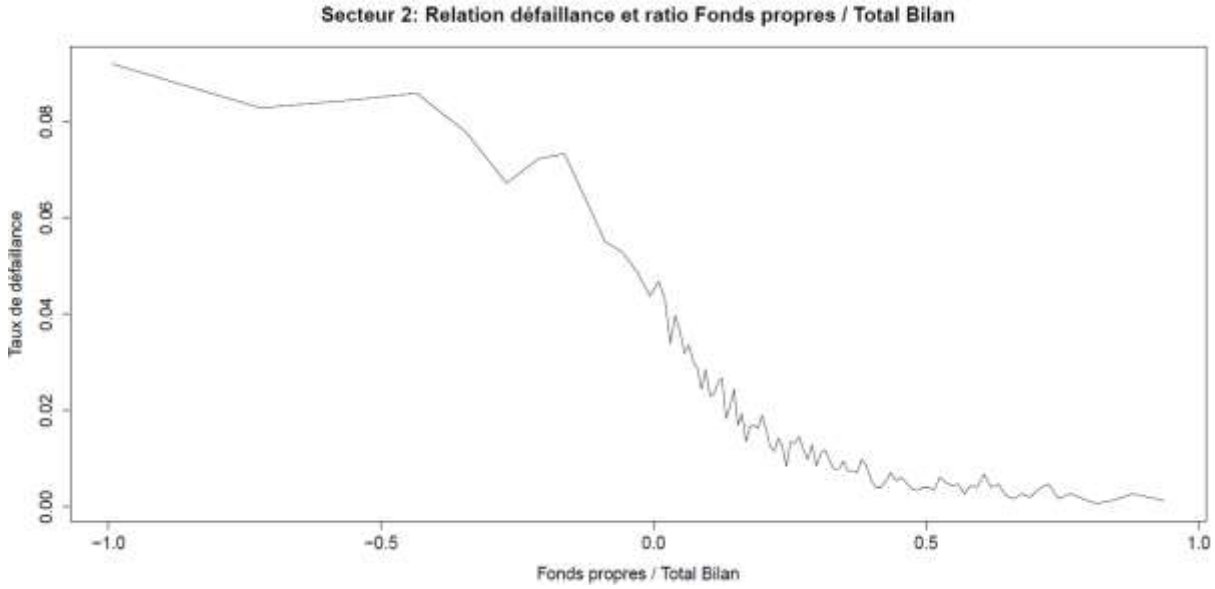
Ces observations nous conduisent à agréger les secteurs d'activités en quatre « macro-secteurs » :

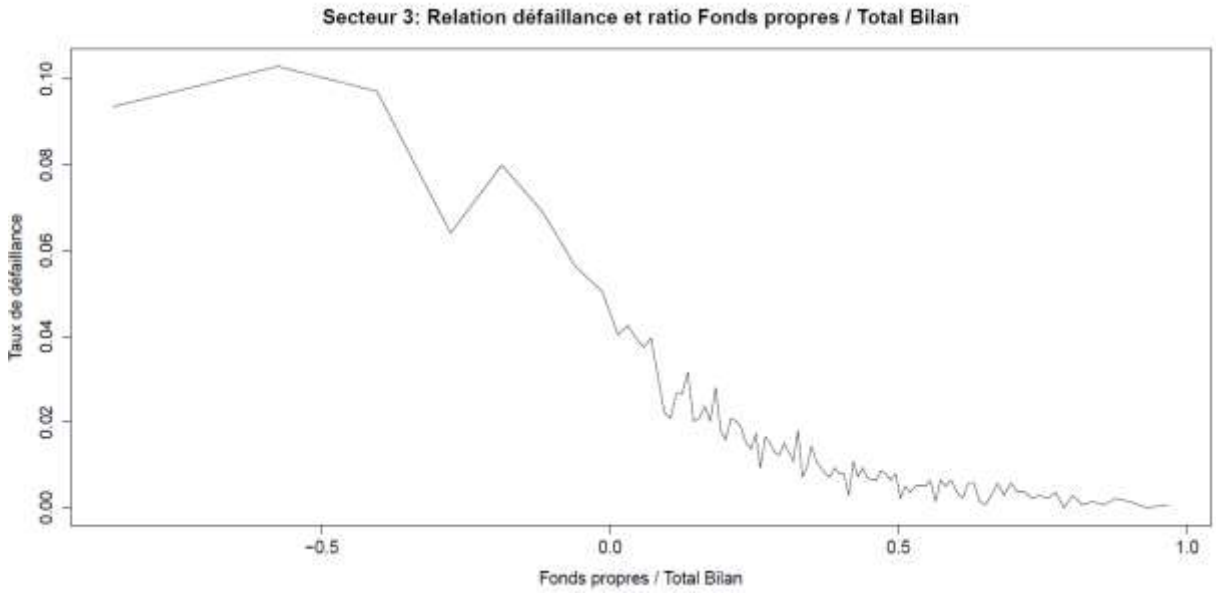
- Secteur 1 : Industrie, transport, construction
- Secteur 2 : Commerce, hébergement, restauration
- Secteur 3 : Services aux entreprises
- Secteur 4 : Autres secteurs

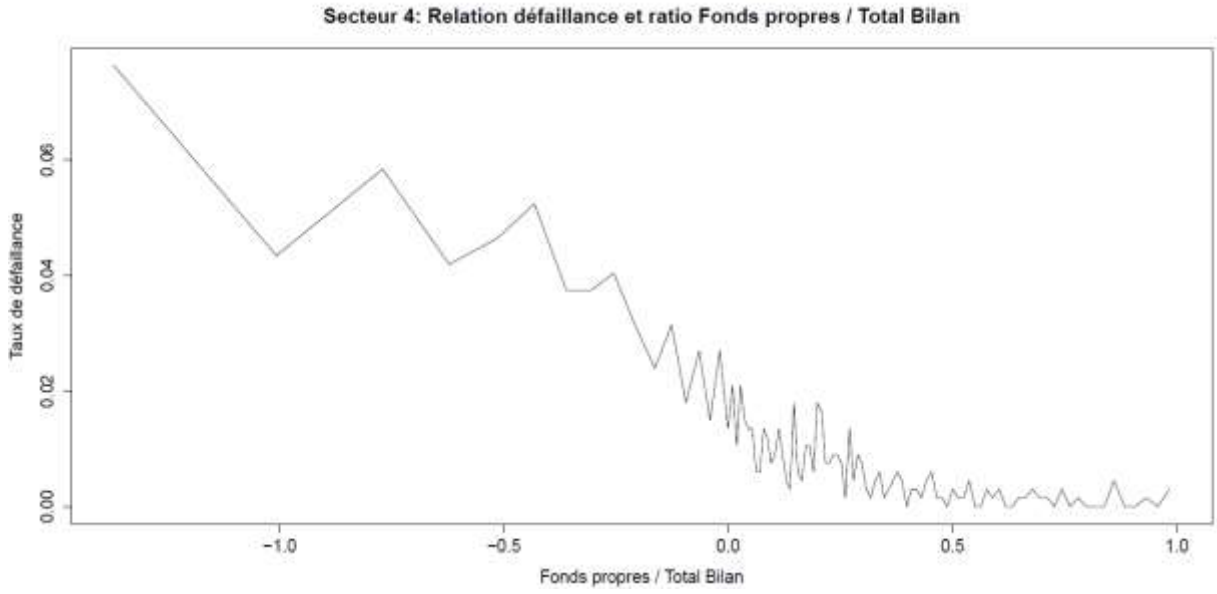
A titre d'exemple, nous observons le ratio *Fonds propres / Total Bilan*. Ce ratio fait clairement apparaitre une relation décroissante entre son niveau et le taux de défaillance constaté dans ce secteur. Il sera retenu dans notre étude sur chacun des secteurs.



Logiquement, ce ratio confirme que moins une entreprise est dotée en fonds propres, plus sa probabilité de défaillance va être élevée. Cela se vérifie quel que soit le secteur.

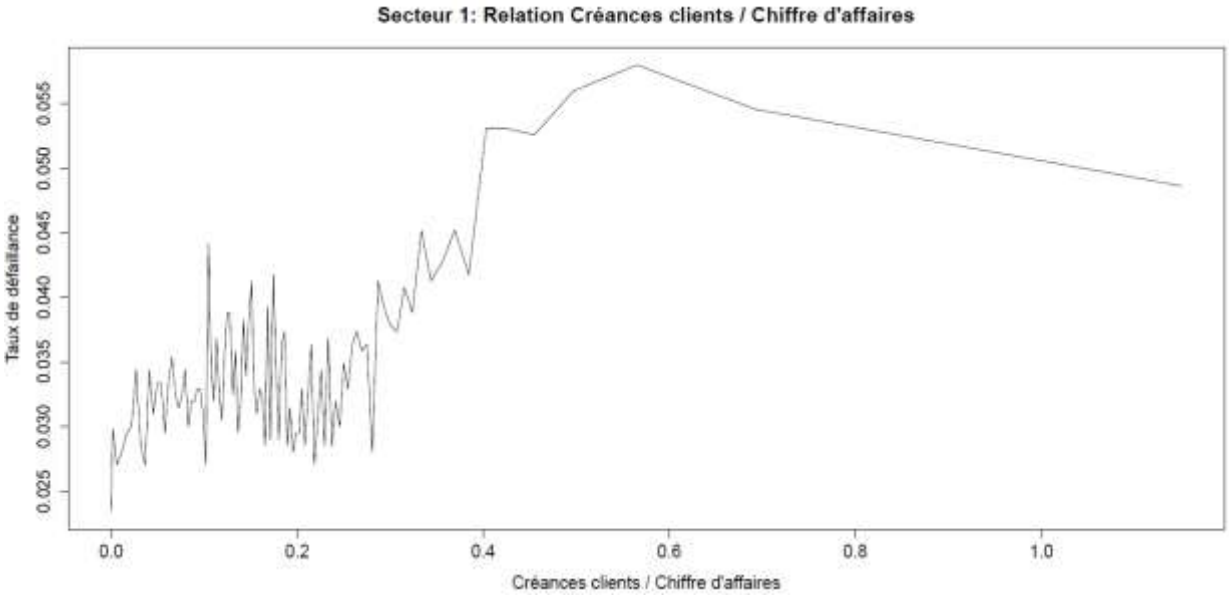






Il est à noter que le quatrième secteur est un regroupement par défaut intégrant les catégories non retenues dans les trois premiers. Le comportement de cet échantillon de secteurs mixtes est en conséquence plus erratique.

A contrario, nous constatons que la relation Créances clients / Chiffre d'affaires paraît plus difficilement interprétable.



Ce ratio ne sera donc pas retenu dans le calibrage du score financier.

À l'issue de cette première phase, nous retenons une quarantaine de ratios financiers dont le lien avec le taux de défaillance peut être différent selon les secteurs d'activités des sociétés (liste en annexe 6.2.1.1).

Le *scoring* financier sera donc issu de quatre modèles différents selon le secteur d'activité.

L'objectif du score est d'être utilisé à la fois par les analystes crédit (arbitres) en tant qu'outil d'aide à la décision et comme input de la tarification. Ce modèle se doit donc d'être simple d'interprétation et d'utilisation. Cette contrainte justifie plusieurs choix qui ont été faits :

- Découper, lorsque c'est nécessaire, les variables explicatives par classe. Ce découpage se base sur des études, variable par variable, en fonction de la fréquence de défaillance.
- S'assurer de la cohérence des coefficients issus des estimations.

Le découpage par classe pour les ratios ayant une relation non linéaire avec la probabilité de défaillance est effectué de façon supervisée. Nous souhaitons en effet pouvoir justifier ce découpage sur des critères d'analyses financières utilisables par les arbitres.

2.2.2 Calibrage du modèle et estimations des coefficients

Nous calibrons quatre modèles sur les quatre secteurs d'activité (les ratios financiers retenus ayant des comportements différents, les coefficients estimés seront différents).

Nous disposons de trois années de publications avec environ 200 000 entreprises pour le premier secteur, soit 600 000 observations. Nous séparons l'échantillon en une base de calibrage avec 400 000 observations puis une base de test de 200 000 observations. Il sera fait de même pour chacun des quatre secteurs retenus.

Une première sélection de ratios est réalisée via la procédure *stepwise*[1], l'objectif étant de minimiser *l'Akaïke*[2].

Les ratios retenus sont :

- Fonds propres / passif (R2)
- Dettes fournisseurs / passif (R3)
- Dettes fiscales et sociales / passif (R4)
- Capitaux permanents / passif (R5)
- Fonds de roulement / CA (R8)
- (Fonds de roulement – BFR) /CA (R9)
- Résultat courant avant impôt / Chiffre d'affaires (R12)
- Marge brute d'autofinancement / Chiffre d'affaires (R13)
- Actif circulant / dettes court terme (R18)
- Valeurs réalisables court terme / dettes court terme (R19)
- (Valeurs mobilières de placement + disponibilités) / dettes court terme (R20)
- Excédent Brut d'Exploitation / dettes (R21)
- Ressources stables / emplois stables (R23)
- Charges de personnel / Valeur Ajoutée (R27)
- (Charges fixes + résultat d'exploitation) / (charges fixes + intérêts financiers) (R40)

Une fois cette sélection effectuée, nous utilisons le critère du V de Cramer pour apprécier les colinéarités entre les variables. Nous éliminons les variables trop corrélées entre elles pour réduire le nombre de ratios utilisés et gagner en simplicité d'utilisation. Nous utilisons également ces critères pour retenir les ratios les plus corrélés avec la variable explicative (première colonne du tableau suivant).

[1] La procédure *Stepwise* ou régression pas à pas permet de choisir, parmi les variables explicatives le plus petit nombre d'entre elles qui explique au mieux la variable explicative. Alternativement toutes les variables sont entrées dans le modèle et elles sont progressivement exclues, en fonction du critère de minimisation de l'AIC.

[2] L'AIC (*Akaïke Information Criterion*) est une mesure de la qualité d'un modèle statistique. Le critère est AIC=2k-2ln(L) où k est le nombre de paramètres à estimer du modèle et L le maximum de la fonction de vraisemblance du modèle.

Secteur 1: V de cramer sur les 15 ratios retenus / base de calibrage

	def24	R2_t	R3_t	R4_t	R5_t	R8_t	R9_t	R12_t	R13_t	R18_t	R19_t	R20_t	R21_t	R23_t	R27_t	R40_t
def24	100%	24%	13%	15%	23%	18%	18%	17%	18%	19%	17%	18%	16%	16%	17%	16%
R2_t	24%	100%	20%	22%	46%	36%	22%	20%	22%	38%	32%	33%	22%	60%	17%	22%
R3_t	13%	20%	100%	8%	29%	19%	18%	11%	15%	23%	22%	25%	13%	25%	10%	11%
R4_t	15%	22%	8%	100%	35%	25%	12%	8%	11%	29%	24%	20%	9%	41%	13%	9%
R5_t	23%	46%	29%	35%	100%	39%	20%	15%	21%	48%	37%	34%	18%	77%	18%	17%
R8_t	18%	36%	19%	25%	39%	100%	26%	18%	18%	60%	44%	32%	15%	66%	16%	18%
R9_t	18%	22%	18%	12%	20%	26%	100%	18%	17%	25%	29%	59%	16%	16%	14%	18%
R12_t	17%	20%	11%	8%	15%	18%	18%	100%	56%	15%	16%	21%	44%	23%	51%	79%
R13_t	18%	22%	15%	11%	21%	18%	17%	56%	100%	17%	17%	19%	52%	26%	52%	50%
R18_t	19%	38%	23%	29%	48%	60%	25%	15%	17%	100%	55%	41%	17%	60%	15%	17%
R19_t	17%	32%	22%	24%	37%	44%	29%	16%	17%	55%	100%	45%	19%	43%	15%	18%
R20_t	18%	33%	25%	20%	34%	32%	59%	21%	19%	41%	45%	100%	24%	21%	16%	20%
R21_t	16%	22%	13%	9%	18%	15%	16%	44%	52%	17%	19%	24%	100%	21%	54%	51%
R23_t	16%	60%	25%	41%	77%	66%	16%	23%	26%	60%	43%	21%	21%	100%	22%	22%
R27_t	17%	17%	10%	13%	18%	16%	14%	51%	52%	15%	15%	16%	54%	22%	100%	52%
R40_t	16%	22%	11%	9%	17%	18%	18%	79%	50%	17%	18%	20%	51%	22%	52%	100%

Enfin, nous calibrons le score financier sur les ratios retenus. Nous comparons à chaque étape nos modélisations via une courbe ROC en s'attachant à maximiser l'aire sous la courbe ROC (AUC, voir annexe 6.1).

Le modèle final retient 7 ratios financiers. Le découpage en classes aboutit à 40 variables explicatives.

Les variables sont les suivantes :

- Fonds propres / passif (R2) découpé en 7 classes
- Dettes fournisseurs / passif (R3) découpé en 7 classes
- Dettes fiscales et sociales / passif (R4) découpé en 6 classes
- (Fonds de roulement – BFR) /CA (R9) découpé en 6 classes
- Marge brute d'autofinancement / Chiffre d'affaires (R13) découpé en 5 classes
- (Valeurs mobilières de placement + disponibilités) / dettes court terme (R20), découpé en 4 classes
- Charges de personnel / Valeur Ajoutée (R27) découpé en 5 classes

Les variables retenues sont celles qui permettent d'avoir une AUC optimum, mais également d'avoir un modèle interprétable par les analystes financiers.

A titre d'exemple, pour le ratio R2 – Fonds propres / Passif, nous retrouvons bien une relation décroissante entre la probabilité de défaut et les tranches du ratio. En effet, plus ce ratio est élevé, plus l'entreprise considérée a une structure financière solide et plus la probabilité de défaut est faible.

Les estimations des coefficients des variables explicatives pour le premier secteur sont résumées dans le tableau suivant (le détail est en annexe 6.2.1.2) :

(Intercept)	-2,31

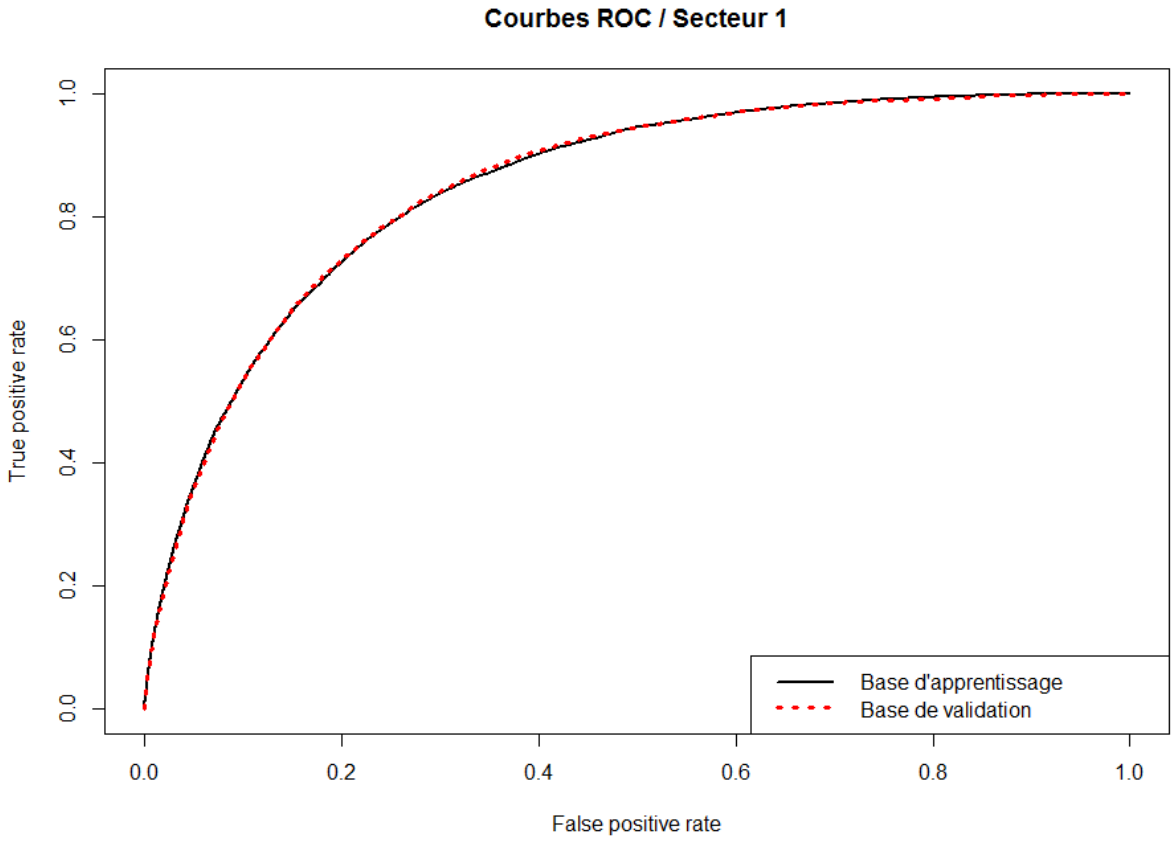
R2_t(-0.2,0]	-0,07	R3_t(0.05,0.1]	0,12
R2_t(0,0.1]	-0,18	R3_t(0.1,0.15]	0,19
R2_t(0.1,0.2]	-0,49	R3_t(0.15,0.2]	0,24
R2_t(0.2,0.3]	-0,81	R3_t(0.2,0.3]	0,29
R2_t(0.3,0.4]	-1,02	R3_t(0.3,0.35]	0,40
R2_t(0.4,0.5]	-1,23	R3_t(0.35,0.45]	0,50
R2_t(0.5,1]	-1,35	R3_t(0.45,2]	0,52

R4_t(0.1,0.15]	0,16	R9_t(-10,1]	-0,40
R4_t(0.15,0.2]	0,30	R9_t(1,15]	-0,58
R4_t(0.2,0.3]	0,50	R9_t(15,30]	-0,82
R4_t(0.3,0.4]	0,75	R9_t(30,50]	-1,15
R4_t(0.4,0.5]	0,90	R9_t(50,100]	-1,45
R4_t(0.5,2]	1,04	R9_t(100,2e+03]	-2,16

R13_t(-0.15,-0.05	-0,22	R27_t(0.8,0.9]	0,18
R13_t(-0.05,0]	-0,36	R27_t(0.9,1]	0,37
R13_t(0,0.025]	-0,52	R27_t(1,1.1]	0,46
R13_t(0.025,0.05	-0,64	R27_t(1.1,1.2]	0,58
R13_t(0.05,2]	-0,75	R27_t(1.2,5]	0,66

R20_t(0.2,0.25]	-0,27
R20_t(0.25,0.55]	-0,31
R20_t(0.55,1.5]	-0,35
R20_t(1.5,50]	-0,74

Les courbes ROC sont dessinées sur la base d'apprentissage (deux tiers de l'échantillon) et sur la base de validation (un tiers de l'échantillon).



L'aire sous la courbe (AUC) est comparable avec 0,848 pour la base d'apprentissage et 0,849 pour la base de validation.

Nous avons ainsi un score financier calibré sur trois années de publication.

Les mêmes étapes sont déroulées sur les trois autres secteurs, les ratios et estimations figurent en annexe (6.2.1). Ces modèles permettent ainsi de calculer une probabilité de défaut sur n'importe quelle société française à partir de ratios financiers. Ce score financier est la première étape du score final.

2.3 Le score final

2.3.1 Méthodologie et données

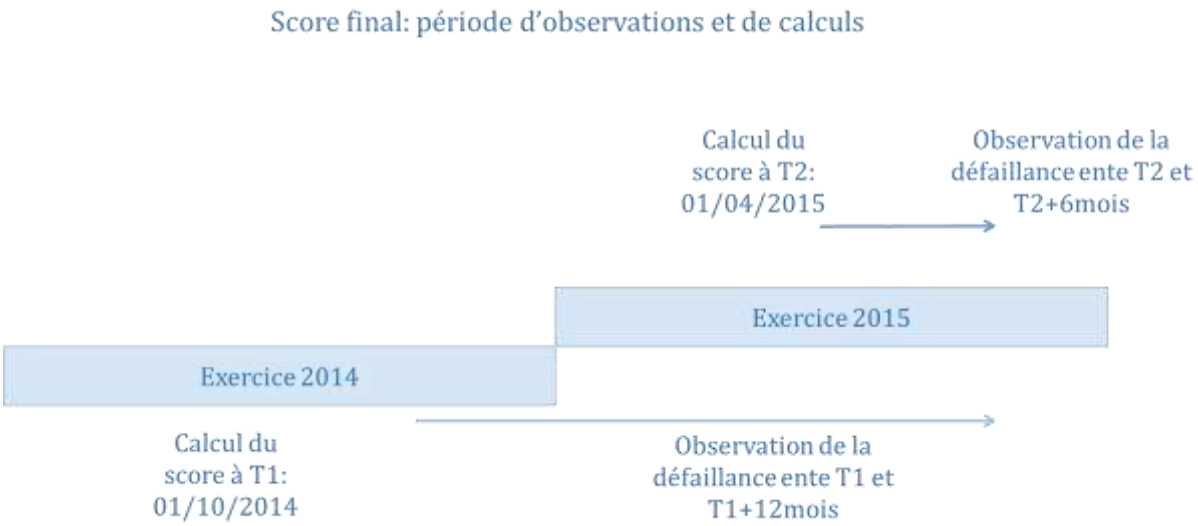
Le modèle de score final va s'appuyer sur le score financier et l'intégrer. Nous allons élargir les variables explicatives à des critères non financiers qui sont les suivants :

- Âge du dirigeant
- Âge de l'entreprise

- Capital social
- Département de l'entreprise
- Effectif de l'entreprise
- Forme juridique de l'entreprise
- Secteur d'activité

Par ailleurs le fonctionnement du score final n'est pas le même que celui du score financier. En effet, la méthodologie, quant à l'étude de la dynamique, est différente. Pour le score financier, nous avons étudié la défaillance dans les 24 mois suivant la clôture des comptes de chaque entreprise considérée pour les exercices de 2009 à 2011. La période d'observation est donc la même pour l'ensemble de notre échantillon. En revanche la date de début de l'observation peut être différente d'une entreprise à une autre car elle dépend de sa date de clôture des comptes.

Pour le score final, la date de départ du calcul du score sera la même pour l'ensemble des entreprises.



Ces périodes sont ainsi choisies pour deux raisons :

- Deux dates pour tenir compte de l'information publiée et disponible entre ces deux dates
- Nous ne disposons, pour calibrer le modèle, que des défaillances jusqu'au mois d'octobre 2015, ce qui borne donc cette étude.

La population de sociétés étudiées comprend 1 729 315 entreprises commerciales « saines » au 1er octobre 2014. Elles sont saines car nous excluons celles pour qui une procédure de redressement ou de liquidation judiciaire a déjà été déclenchée. Il serait en effet paradoxal de calculer une probabilité de défaut sur une société qui est déjà en défaillance.

Nous constatons sur notre échantillon 41 910 défaillances sur une année, ce qui représente 2,4% de l'échantillon.

Pour ce score également, la phase de gestion de données est lourde et consommatrice de ressources. Nous disposons d'un échantillon de plus de 2 millions d'entreprises.

Le cœur d'activité d'AXA Assurcrédit étant les PME, nous retirons les sociétés ayant un effectif supérieur à 2 000 salariés.

Nous retirons également les sociétés dont le chiffre d'affaires est inférieur à 12k€. Ce choix arbitraire est issu du constat que ce groupe de sociétés connait une très faible défaillance et ont un comportement non standard. Lorsque ce modèle passera en production, il sera possible de gérer ces profils.

Ces deux opérations de tri nous permettent de retirer les extrêmes (environ 5% de l'échantillon).

A la date du calcul du score, le bilan d'une entreprise n'est pas forcément disponible. Le choix fait est alors de retenir le plus récent publié. Si aucun bilan n'est publié, nous calibrons alors le *scoring* final sans *scoring* financier, avec les autres variables explicatives.

Au final, nous disposons de 7 variables non financières, d'une variable financière (score de 1 à 18) qui nous permettra de calculer une probabilité de défaut. Ce score final permet de couvrir un spectre plus large que le score financier car il prend également en compte les entreprises pour lesquelles les comptes ne sont pas disponibles.

2.3.2 Calibrage du score final

Préalablement au calibrage du modèle, nous regroupons les variables par tranche ou par classe. Nous faisons en sorte d'avoir une homogénéité de chaque classe en termes de taux de défaillance. Cette étape permet également d'apprécier la relation entre la variable expliquée et les différentes variables explicatives utilisées.

Regroupement de l'âge du dirigeant

Âge du dirigeant	Sociétés saines	Sociétés défaillantes	Répartition	Taux de défaillance
18 à 24 ans	14 276	927	0,8%	6,5%
25 à 29 ans	57 343	2 724	3,4%	4,8%
30 à 34 ans	118 021	4 359	7,0%	3,7%
35 à 50 ans	621 522	16 871	36,8%	2,7%
50 à 65 ans	508 780	10 472	30,2%	2,1%
non renseigné	218 860	3 977	13,0%	1,8%
65 ans et plus	115 077	1 674	6,8%	1,5%
Total	1 653 879	41 004	100,0%	2,5%

Nous constatons une relation décroissante entre le taux de défaillance et l'âge du dirigeant.

Nous procédons de même pour l'ensemble des variables utilisées, y compris le score financier qui est regroupé en 18 classes. L'ensemble de ces regroupements figurent en annexe 6.2.2.1.

Comme pour le score financier, nous sélectionnons les variables via différentes méthodes (procédure *stepwise* de minimisation de l'AIC et étude des relations entre chacune des variables explicatives avec le taux de défaillance notamment).

Les liens entre les variables explicatives et la défaillance sont étudiés via un V de Cramer. Les résultats via le V de Cramer sont dans le tableau suivant.

Score final: V de Cramer / base de calibrage

	CA	ANC	Effectif	APE	CJ	Age	Cap	Dép	Score fi	Anc Bil	def
CA	100,0%	19,5%	38,7%	8,9%	16,1%	7,2%	29,2%	10,1%	37,8%	49,2%	4,6%
ANC	19,5%	100,0%	30,0%	6,5%	17,2%	17,9%	31,7%	2,2%	16,3%	18,5%	7,2%
Effectif	38,7%	30,0%	100,0%	9,9%	16,9%	11,0%	33,4%	5,3%	17,4%	20,9%	2,9%
APE	8,9%	6,5%	9,9%	100,0%	25,8%	9,0%	10,0%	7,9%	6,2%	6,1%	8,0%
CJ	16,1%	17,2%	16,9%	25,8%	100,0%	10,7%	24,6%	6,8%	7,0%	7,3%	3,6%
Age	7,2%	17,9%	11,0%	9,0%	10,7%	100,0%	13,6%	4,5%	5,9%	7,0%	5,1%
Cap	29,2%	31,7%	33,4%	10,0%	24,6%	13,6%	100,0%	2,6%	14,5%	14,3%	4,1%
Dép	10,1%	2,2%	5,3%	7,9%	6,8%	4,5%	2,6%	100,0%	8,7%	12,2%	4,2%
Score fi	37,8%	16,3%	17,4%	6,2%	7,0%	5,9%	14,5%	8,7%	100,0%	53,5%	15,1%
Anc Bil	49,2%	18,5%	20,9%	6,1%	7,3%	7,0%	14,3%	12,2%	53,5%	100,0%	7,1%
def	4,6%	7,2%	2,9%	8,0%	3,6%	5,1%	4,1%	4,2%	15,1%	7,1%	100,0%

Avec :

- CA : dernier chiffre d'affaires connu
- ANC : ancienneté de la société
- Effectif : effectif de la société
- APE : secteur de la société
- CJ : catégorie juridique de la société
- Âge : âge du dirigeant de la société

- Cap : capital social de la société
- Dép : département du siège social de la société
- Score fi : score financier
- Anc Bil : ancienneté du dernier bilan connu
- Def : défaillance (variable dichotomique).

Les estimations des coefficients des variables explicatives pour le *scoring* sont résumées dans le tableau suivant (le détail des variables et des estimations figure en annexe 6.2.2) :

Constante	-9,93

Score Financier 02	1,15
Score Financier 03	1,61
Score Financier 04	2,21
Score Financier 05	2,42
Score Financier 06	2,6
Score Financier 07	2,81
Score Financier 08	3,25
Score Financier 09	3,37
Score Financier 10	3,75
Score Financier 11	3,88
Score Financier 12	4,08
Score Financier 13	4,19
Score Financier 14	4,35
Score Financier 15	4,57
Score Financier 16	4,66
Score Financier 17	4,78
Score Financier 18	4,88

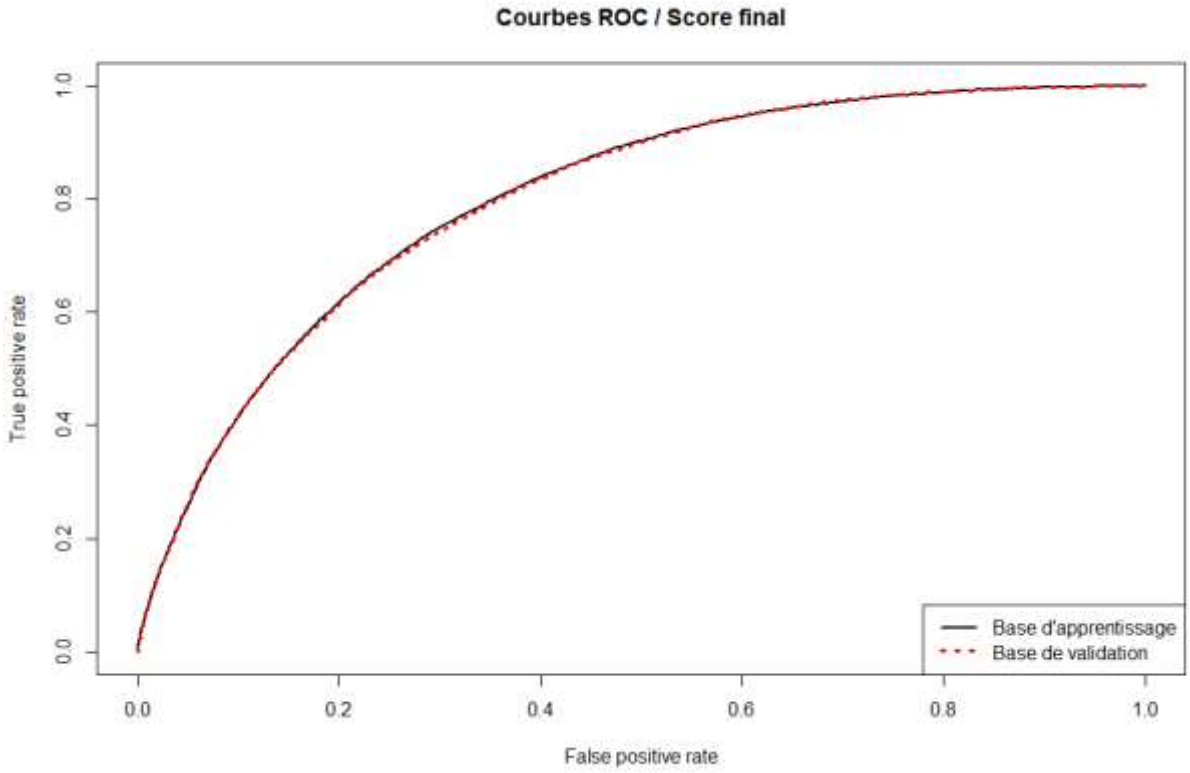
Département 02	0,32
Département 03	0,48
Département 04	0,59
Département 05	0,65
Département 06	0,67
Département 07	0,94
Département 08	1,08

Secteur 02	1,2
Secteur 03	1,54
Secteur 04	1,79
Secteur 05	2,09
Secteur 06	0,21
Secteur 07	2,31
Secteur 08	2,39
Secteur 09	2,61
Secteur 10	2,73

Age du dirigeant 02	-0,23
Age du dirigeant 03	0,08
Age du dirigeant 04	0,13
Age du dirigeant 05	0,21
Age du dirigeant 06	0,36
Age du dirigeant 07	0,54

Ancienneté entreprise 02	1,15
Ancienneté entreprise 03	0,35
Ancienneté entreprise 04	0,1
Ancienneté entreprise 05	0,66
Ancienneté entreprise 06	0,83

La régression logistique nous conduit à retenir le score financier, le secteur, l'ancienneté de l'entreprise, le département et l'âge du dirigeant comme variables explicatives (sortie R en annexe 6.2.2.2). Nous testons la robustesse de notre modélisation via une courbe ROC et le calcul de l'AUC sur la base de calibrage (deux tiers de notre échantillon) et la base de validation.



Les AUC sont de 0,799 pour la base d'apprentissage et 0,798 pour celle de validation.

2.3.3 Grille de *scoring*

Ce score ainsi développé nous permet, pour l'ensemble des entreprises pour lesquelles nous disposons d'une information (publication financière ou non), de calculer une probabilité de défaillance. Nous procédons également à un regroupement par classes afin de mettre en place une notation utilisable par les automates ou les arbitres. A ces notes sont associées des probabilités de défaillance que nous pouvons calculer sur plusieurs horizons.

Grille de notation issue du modèle de *scoring*

Note	Probabilité de défaut à 3 mois	Probabilité de défaut à 6 mois	Probabilité de défaut à 9 mois	Probabilité de défaut à 12 mois
1	5,7%	11,6%	16,0%	18,9%
2	4,2%	8,5%	11,3%	13,7%
3	2,6%	5,5%	7,5%	9,3%
4	1,7%	3,8%	5,4%	6,7%
5	1,0%	2,4%	3,5%	4,5%
6	0,8%	1,8%	2,6%	3,3%
7	0,4%	1,1%	1,6%	2,2%
8	0,3%	0,7%	1,0%	1,3%
9	0,1%	0,3%	0,4%	0,5%
10	0,0%	0,0%	0,1%	0,1%

Ce sont les observations des défaillances au sein d'une classe qui nous permettent de calculer les probabilités de défaut sur des horizons différents. La probabilité de défaut à trois mois des entreprises notées 1 sera donc le rapport du nombre de défauts sur les trois premiers mois d'observation et du nombre total d'entreprises notées 1.

2.4 Critiques, prolongement et conclusion du modèle de *scoring*

Nous avons voulu faire en sorte que le *scoring* soit utilisable et interprétable par le plus grand nombre et notamment par les analystes crédit. La première critique va donc concerner les relations utilisées pour modéliser les variables explicatives. Nous avons pris le parti de découper nos variables explicatives par classes (intervalles) alors que nous aurions pu chercher à affiner via des relations non linéaires. Ce choix a été fait par pure volonté de simplicité et d'efficacité du modèle, les analystes devant justifier un refus automatique d'agrément par le score auprès des assurés. Ils doivent alors être en mesure d'expliquer au client pourquoi un tel refus ainsi que les données de l'acheteur qui déterminent son score.

Nous souhaitions tester, dans le cadre de cette problématique, les nouveaux algorithmes d'apprentissage statistiques issus de la *data science*. Nous avons testé la robustesse du modèle en l'estimant par du *gradient boosting* (Tufféry 2015). Les résultats sont synthétisés dans le tableau de la page suivante.

Méthode	Paramètres	AUC
Grandient boosting	(5,100,0.1)	79,96%
Grandient boosting	(10,100,0.1)	80,28%
Grandient boosting	(15,100,0.1)	80,44%
Grandient boosting	(5,100,0.01)	75,79%
Grandient boosting	(10,100,0.01)	77,86%
Grandient boosting	(15,100,0.01)	78,52%

Paramètres : (profondeur, nombre d'arbres, pénalisation)

Ces méthodes ne nous permettent pas d'améliorer significativement la robustesse du modèle (en terme d'AUC , nous passons de 0,798 à 0.804 au mieux). Elles présentent l'inconvénient d'être une boite noire difficilement utilisable par les arbitres.

Le modèle de *scoring* ainsi mis en place permet d'apporter aux arbitres un outil d'aide à la décision pour leurs analyses crédit. Il peut également alimenter un automate pour accepter ou refuser les demandes d'agréments en fonction de la notation d'une société.

D'autre part, ce modèle de *scoring* nous permet d'extraire des probabilités de défaut que nous utilisons par la suite pour la tarification.

3 Tarification : Approche par les probabilités de défaut

Le modèle de *scoring* nous permet d'avoir, pour chaque entreprise française (hors artisans et professions libérales), une classe d'appartenance ainsi qu'une probabilité de défaut associée (issue du taux de défaillance dans chacune des classes).

Pour ce premier modèle de tarification, nous partons de l'hypothèse qu'un acheteur suit une loi de Bernoulli de probabilité P_i, son taux de défaut estimé par le *scoring* (nous avons donc la probabilité de défaut d'un acheteur de la classe i donné par P_i, le défaut de cet acheteur sera la variable aléatoire Y_i, on aura $Y_i \sim \mathfrak{B}(P_i)$).

Il est à noter en préambule les deux principales limites du modèle de tarification proposé :

- D'une part, nous introduisons un biais en utilisant le *scoring* tel que décrit en section 2. En effet, les probabilités de défaut du modèle de *scoring* que nous avons construit sont celles d'une procédure collective (redressement ou liquidation judiciaire) alors que les défauts assurés dans le cadre des contrats d'Axa Assurcrédit sont un défaut de paiement au sens large, qui correspond à une insolvabilité constatée par procédure collective, ou une insolvabilité présumée par simple carence ou retard de paiement constaté. Il serait éventuellement possible de chercher à développer un *scoring* intégrant également les évènements d'insolvabilité présumée. Mais la difficulté consiste alors à construire la base de ces évènements, qui nous limiterait de facto à un historique des seuls acheteurs garantis par AXA Assurcrédit, ce qui introduirait dès lors un autre biais puisqu'il s'agirait alors de restreindre la base à une sélection d'acheteurs déjà arbitrés par les équipes d'arbitres d'AXA Assurcrédit.
- D'autre part, nous considérons que le portefeuille d'acheteurs est inchangé pendant toute la durée de la période d'assurance comme si les limites de crédit accordées étaient non-annulables, alors même que les arbitres chercheront à anticiper les détériorations de solvabilité lorsqu'elles sont prévisibles et réduiront ou même résilieront les limites de garantie accordées en cours de période d'assurance (le monitoring) comme il est possible de faire au terme du contrat d'assurance (limites annulables en cours de période d'assurance).

Ces deux principales limites du modèle de tarification proposé ont à priori des influences inverses : la première limite sous évalue le risque de défaut et donc la tarification du modèle, la seconde limite sur-évalue au contraire le risque du portefeuille en omettant le monitoring et l'ajustement des garanties en cours de période, et surestime donc la tarification du modèle.

3.1 Définition des paramètres

Certains paramètres sont explicites, d'autres nécessitent des hypothèses que nous détaillons dans cette partie (Caja 2014).

Nous nous plaçons dans le cas où un assuré déclare l'ensemble de ses acheteurs à son assureur et demande une limite (agrément ou NDS). Il n'y a donc pas de « non dénommés ». Ce modèle pourra donc être parfaitement utilisé pour les assurés existants du portefeuille qui n'ont que des garanties déclaratives (pas d'acheteurs garantis en « non dénommés »).

3.1.1 L'« Exposure At Default » (EAD)

Au moment de la défaillance de l'acheteur, l'EAD permet de déterminer le montant en risque sur cet acheteur. Nous avons cependant des complexités supplémentaires dans le calcul de l'EAD résultant des caractéristiques du produit d'assurance-crédit

Utilisation des limites demandées :

- Un assuré peut utiliser son assurance comme un outil de prospection, il l'interroge alors pour demander une limite sur un prospect pour savoir si celui-ci serait assuré ou non sans forcément qu'une réelle transaction aboutisse avec cet acheteur potentiel. Cela va essentiellement concerner le « non dénommés surveillés » qui fait l'objet d'un accord de l'assureur crédit pour un montant fixé contractuellement (5 000 € dans le produit Global Kup), dans la mesure où le NDS n'entraine pas de frais supplémentaires pour l'assuré contrairement à l'agrément (gratuité du NDS en général dans les polices). Dans ce cas, l'assureur va voir apparaitre des autorisations dans ses systèmes qui ne correspondent pas à des expositions en risques réelles.

- Un assuré peut demander à son assureur une limite supérieure à ce qu'il facture réellement afin d'avoir une certaine flexibilité, en cas de développement de son chiffre d'affaires. Dans ce cas, s'il y a défaillance de l'acheteur, le montant réellement assuré sera plus faible que le montant garanti. Des phénomènes de saisonnalité peuvent également expliquer la non utilisation totale des encours pendant l'année d'assurance.

Nous résumons ces deux cas dans une variable que nous appellerons le « Use Factor » ou taux d'utilisation de la police (Becue 2013). Nous reviendrons par la suite sur la manière dont nous avons estimé ce « use factor ».

La somme des demandes d'agréments ne représente pas forcément le chiffre d'affaires assurable de l'assuré. En effet, durant l'année, l'assuré va émettre plusieurs factures à un même acheteur. Le chiffre d'affaires réalisé avec cet acheteur sera donc la somme des factures établies dans l'année. Nous introduisons la cyclicité de facturation qui va être le nombre de rotations des montants demandés dans une année. Si un assuré facture quatre fois à son acheteur dans l'année, alors cette cyclicité de 4 entraine un chiffre d'affaires réalisé avec cet acheteur de quatre fois l'agrément demandé (toutes choses égales par ailleurs).

L'ensemble des éléments (Chiffre d'affaires, agréments et factures présentées en cas de défaut) est retenu hors taxes.

Lors de ses transactions avec l'assuré, un acheteur va bénéficier de délais de paiement de la part de l'assuré, ceux-ci peuvent varier sensiblement d'un secteur à l'autre. Nous introduisons une variable « délai de paiement » qui va le prendre en compte. Ce délai sera exprimé en mois.

L'EAD va dépendre de ces trois éléments « Use Factor », « Cyclicité » et « Délai de paiement ». A tout instant, l'EAD ne peut être supérieur à l'agrément. En revanche, il peut être inférieur en fonction de ces paramètres.

Pour un acheteur i, nous aurons donc :

$$EAD_i = Agrément_i \times Min\left(1\,, Use\ Factor \times Max\left(1\,,\ \frac{Délai\ de\ paiement_i}{\frac{12}{Cyclicité_i}}\right)\right)$$

Le minimum est contractuel. En effet, il permet de spécifier que l'assureur ne s'engage qu'à hauteur de l'agrément accordé et pas au-delà. Le maximum est une hypothèse de travail. Il permet de prendre en compte le fait qu'un assuré peut avoir émis plusieurs factures pour un même acheteur et que ces factures ont un montant unitaire inférieur à l'agrément.

Si le délai de paiement est de 6 mois, la cyclicité de 4 (une facturation tous les 3 mois) et le « use factor » de 2/3, nous aurons alors l'exposition au défaut suivante :

$$EAD_i = Agrément_i \times Min\left(1\,, \frac{2}{3} \times Max\left(1\,,\ \frac{6}{\frac{12}{4}}\right)\right) = Agrément_i$$

En revanche, si le délai de paiement est de 3 mois, la cyclicité de 4 et le « *use factor* » de 2/3, on aura :

$$EAD_i = Agrément_i \times Min\left(1\,, \frac{2}{3} \times Max\left(1\,,\ \frac{3}{\frac{12}{4}}\right)\right) = Agrément_i \times \frac{2}{3}$$

Ainsi, l'exposition au défaut peut être inférieure à l'agrément accordé.

Il est à noter que, dans la pratique, la cyclicité et le délai de paiement ne sont pas des données actuellement disponibles dans les bases de données d'AXA Assurcrédit, et ce, encore moins au niveau de chaque acheteur. Nous retiendrons donc une hypothèse de cyclicité et de délai de paiement au niveau de l'assuré. Nous laissons cependant la possibilité de pouvoir modifier ces paramètres par la suite.

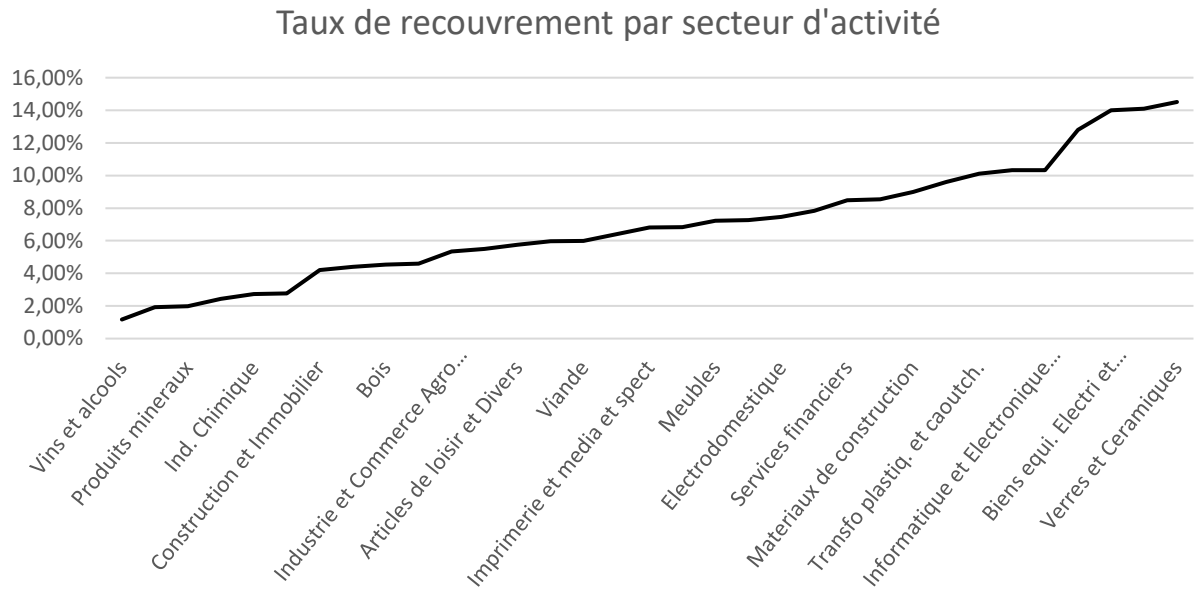
3.1.2 La perte en cas de défaut ("*Loss Given Default*" ou LGD)

En cas de défaut de l'acheteur, l'assureur reprend la créance et procède à son recouvrement. Si la compagnie arrive à recouvrer une partie de la créance, elle se rembourse à hauteur de son engagement (agrément corrigé de la quotité garantie), puis reverse le solde à l'assuré.

Nous utilisons la base sinistre existante pour estimer et segmenter un taux de recouvrement sur les sinistres passés. Etant donné que les probabilités de défaut issues du modèle de *scoring* sont calibrées sur des procédures collectives, nous ne retenons que les sinistres issus de procédures collectives et non ceux issus de carence (retard ou défaut de paiement n'entrainant pas de procédures collectives).

Le taux de recouvrement sur les sinistres liés à une procédure collective est de 6,5%, soit une « *Loss Given Default* » de 93,5%.

Nous observons cependant des disparités par secteur :



Nous utiliserons cette distinction sectorielle pour différencier la LGD par acheteur en fonction de son secteur d'appartenance.

3.1.3 Estimation du « *use factor* »

Le « use factor » regroupe deux composantes :

1- Le fait que l'assuré demande pour un acheteur une limite qui lui donnera de la « marge » pour traiter ensuite avec son client. Autrement dit, la limite demandée ne correspond pas au risque réel auquel l'assureur est exposé.

2- Le fait que l'assuré puisse utiliser son assureur crédit pour « tester » la qualité de ses prospects. C'est d'autant plus vrai lorsque le produit intègre des « agréments express » ou NDS. Nous pouvons ainsi nous trouver face à un assuré qui demande énormément d'agréments express alors qu'il n'y a pas de montant en risque.

Pour estimer le « use factor » lié à sa première composante ($Use\ Factor_1$), nous travaillons sur la base sinistres historiques. Sur cette dernière, nous comparons, pour chaque sinistre, la ou les facture(s) sinistrée(s) par acheteur avec l'agrément disponible pour cet acheteur. Le « use factor » lié au taux d'utilisation d'une limite demandée sera égal au rapport *Factures sinistrées / Agréments*. Le tout plafonné à 1 puisque l'assureur couvre le sinistre au maximum à hauteur de son agrément. Cette estimation nous fait retenir un « use factor » lié à l'utilisation de la limite de 53% pour le produit « Globale » et 44% pour le produit « KUP ».

L'estimation de la seconde composante du « use factor » ($Use\ Factor_2$) est faite assuré par assuré afin de distinguer ceux qui interrogent abondamment l'assureur des autres. En outre, cela permet de borner la masse assurable au chiffre d'affaires réel et donc de ne pas calculer une prime complètement déconnectée du chiffre d'affaires effectif de l'assuré.

Nous reconstituons d'abord un chiffre d'affaires attendu ($CA_{attendu}$) avec les éléments estimés précédemment (cyclicité et $Use\ Factor_1$) ainsi que les montants agréés. Le $CA_{attendu}$ sera ainsi le produit des montants agréés, de la cyclicité et du $Use\ Factor_1$.

Pour l'assuré j : $CA_{attendu\,j} = Use\ Factor_1 \times \left(Cyclicité_j \times \sum_i Agrément_{i,j}\right)$, i étant le nombre d'acheteurs de l'assuré j.

Le chiffre d'affaires attendu de l'assuré est donc la somme des agréments demandés pour chacun de ses acheteurs, corrigé du taux d'utilisation des agréments et multiplié par le nombre de facturations dans l'année (la cyclicité). Dans la pratique, nous fixons la cyclicité pour l'ensemble des acheteurs d'une police, cette donnée n'étant pas disponible pour l'instant.

Nous comparons ensuite le chiffre d'affaires attendu de l'assuré avec son chiffre d'affaires réel. La comparaison des deux quantités permet de calculer un « use factor » qui sera par la suite utilisé.

Nous avons donc :

$$CA_{réel} = CA_{attendu} \times Use\ Factor_2$$

Nous en déduisons : $\widehat{Use\ Factor_2} = \frac{CA_{Réel}}{CA_{attendu}}$

Là encore, nous plafonnons ce chiffre à 1.

Le « Use Factor » total est ainsi le produit des deux « use factors »1 et 2.

3.2 Principe de prime retenu

Nous retenons un principe de prime en quatre parties :

- Une prime pure, issue de la perte espérée ;
- La refacturation à l'assuré de la mobilisation des fonds propres solvabilité 2 à hauteur du coût du capital, net de réassurance ;
- Le coût du traité de réassurance ;
- Les frais.

Nous supposons dans un premier temps qu'il y a indépendance du défaut des acheteurs. Cette hypothèse est revue en section 3.5.2

3.2.1 La prime pure

La prime pure va être définie par la perte espérée (*Expected Loss* ou EL). Celle-ci est directement calculée de l'exposition au défaut (EAD) et de la probabilité de défaillance issue du *scoring*. Ainsi pour un portefeuille j de n acheteurs, nous aurons une prime pure égale à l'EL suivant :

$$EL_j = \sum_{i=1}^{n} P_i \times LGD_i \times QG_i \times EAD_i$$

Avec :

- j l'indice du portefeuille de l'assuré j,
- i l'indice d'un acheteur,
- LGD_i la *loss given default* de l'acheteur i (fonction de son secteur d'appartenance),
- QG_i la quotité garantie de l'acheteur i, exprimée en pourcentage. Celle-ci est propre à l'acheteur car elle peut différer selon que la limite accordée est un agrément express (NDS) ou un agrément,
- P_i la probabilité de défaut de l'acheteur i,
- EAD_i L'exposition au défaut de l'acheteur i.

3.2.2 Le coût des fonds propres

Les réglementations « Solvabilité 2 » imposent de mobiliser des fonds propres pour couvrir l'évènement bicentenaire (le SCR, *Solvency Capital Requirement*). Le coût de ces fonds propres est une donnée externe paramétrable, fixée à 10%. Nous allons intégrer ce coût dans la prime à hauteur des capitaux propres que mobilisent ce nouveau risque. L'objectif est de refacturer à l'assuré le coût des fonds propres que sa police nécessite de mobiliser.

Cependant, AXA Assurcredit bénéficie d'un traité de réassurance qui lui permet de limiter le ratio *Sinistres / Prime* à 0,8. Nous devons donc en tenir compte dans le cadre du calcul du coût de la mobilisation de SCR pour ne pas compter deux fois la même chose (le risque d'une part et le coût de la réassurance d'autre part).

Nous optons pour un calcul explicite en tenant compte du traité de réassurance. C'est-à-dire que nous allons exprimer la mobilisation de SCR par rapport à la prime finale chargée des frais généraux. Chaque assuré va donc mobiliser un SCR au niveau de 0,8 fois la prime moins la prime pure (représentée par *l'expected loss*). Nous faisons donc l'hypothèse que l'événement bicentenaire est au-delà de 0,8 fois la prime totale.

3.2.3 Les charges

Nous affectons les différents frais que connait l'assureur pour charger également la prime des frais et commissions. Cela va inclure :

- La rémunération du traité de réassurance. Axa Assurcrédit bénéficie d'un traité de réassurance en excédent de pertes qui prévoit une couverture au-delà de 80% de ratio *Sinistre / Prime*. Ce traité de réassurance est facturé 4,95% de la totalité des primes.
- Les commissions payées aux courtiers ou agents qui représentent 12% de la prime.
- Le coût de gestion des sinistres, évalué à 3,8% de la prime.
- Le coût de gestion et d'arbitrage qui représente 3,6% de la prime.
- Le coût de souscription d'un nouveau contrat, évalué à 8,3% de la prime, mais amorti sur 10 ans. L'amortissement sur 10 ans correspond à la durée de vie moyenne d'une police dans l'entreprise.
- L'amortissement des frais fixes (coûts de structure, comptabilité, informatique, système d'information, management). Ces frais fixes représentent, dans la structure actuelle, environ 25% de la prime. L'objectif est de ramener la structure de coût globale, hors réassurance, à 40% de ratio *coûts / primes*. Ces 40% correspondent également aux conditions auxquelles Axa Assurcredit cède une partie de ses primes dans le cadre d'un second traité de réassurance en quote-part. Nous retenons donc 20% de frais fixes comme hypothèse principale, correspondant à un ratio *coûts / primes* de 40%.

Nous avons ainsi un total de 45,18% de frais, avec réassurance, tous exprimés en pourcentage de la prime finale.

3.2.4 La prime finale

La prime finale est considérée hors marge de l'entreprise. Nous raisonnons, en effet, sur une prime qui permet de construire un ratio combiné de 100%.

L'approche d'AXA est de calculer un ratio combiné économique (ECR) qui prend mieux en compte le risque que le ratio combiné traditionnel (AXA - Comité technique RI 2014) :

- Prime pure ;
- Frais et charges ;
- Prise en compte de la réassurance ;
- Coût du capital.

Un ECR possède une référence d'équilibre à 100%. Le portefeuille est rentable si l'ECR est inférieur à 100%. Nos travaux vont donc consister à calculer une prime correspondant à l'ECR d'équilibre (hors marge).

Pour notre exercice, la prime sera la somme de l'*expected loss*, du coût des fonds propres liés au risque de l'assuré au-delà de la prime pure, du coût de la réassurance et des charges.

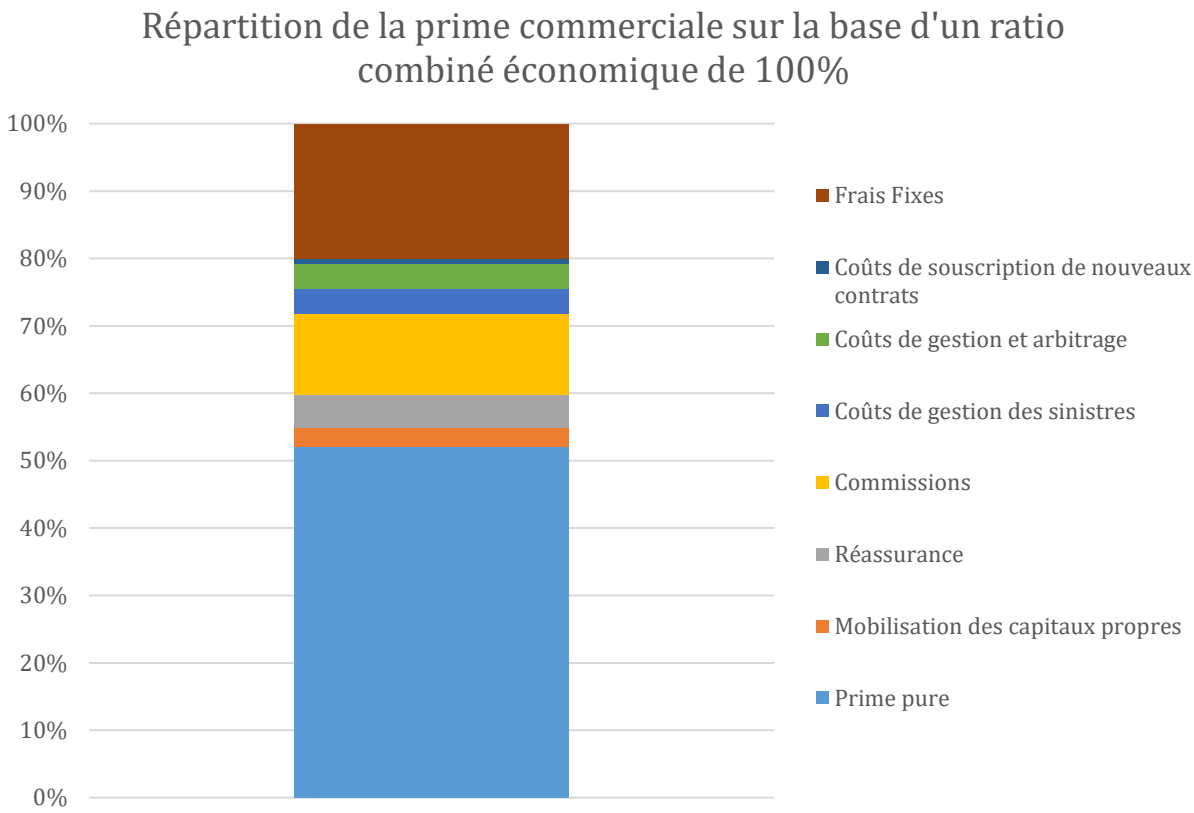
Nous avons donc un ensemble de métriques qui s'exprime en fonction de la prime finale. En effet, avec P la prime finale et EL l'*expected loss* (qui peut être assimilée à une prime pure), en prenant en compte notre hypothèse centrale de 45,18% de frais et le traité de réassurance, nous avons la décomposition suivante de la prime totale :

$$P = EL + 10\% \times (80\% \times P - EL) + 45{,}18\% \times P$$

$$\Rightarrow P = 1{,}922 \times EL$$

Nous pouvons ainsi calculer une prime commerciale qui représente le minimum à tarifer pour couvrir les coûts et la mobilisation de capitaux propres au sens du SCR après traité de réassurance.

Sur la base d'un ratio combiné économique de 100%, la prime se décompose ainsi :



3.3 Application au contrat « Globale »

Nous allons appliquer notre modèle de tarification sur les données du contrat « Globale ».

Nous disposons pour cela des 921 polices de ce contrat ainsi que la prime perçue (soit le montant forfaitaire, soit le chiffre d'affaires assurable et le tarif appliqué en pourcentage de ce dernier), du détail par police des acheteurs (agrément hors « non dénommés »), du modèle de *scoring* calibré dans la partie 2 et des probabilités de défaut.

Lorsqu'un acheteur n'est pas présent dans la base de *scoring*, nous lui attribuons un score de 4 par défaut. Ce choix arbitraire est fait pour intégrer l'effet de l'arbitrage qui élimine généralement les trois niveaux de score les plus faibles.

Nous fixons le délai de paiement à 3 mois, la cyclicité à 4 facturations par an. Ce délai de paiement est au-delà de la moyenne nationale de 57 jours (Observatoire des délais de paiement 2016).

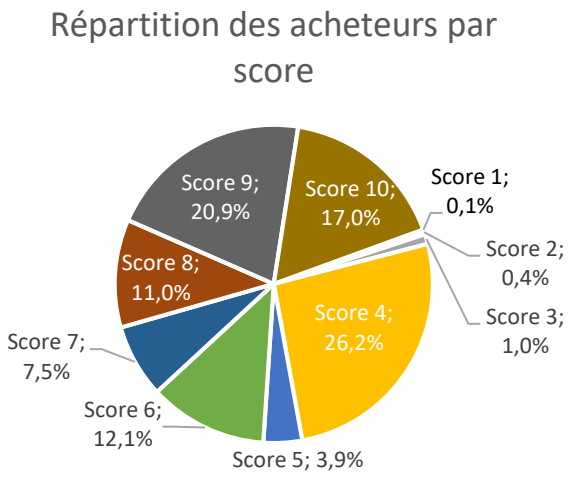
Le taux de recouvrement dépend du secteur. La « loss given default » a été estimée à l'aide de la base sinistres comme indiqué dans la partie 3.1.2.

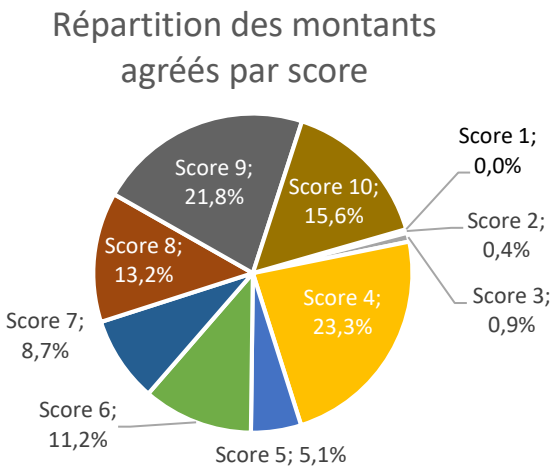
3.3.1 Caractéristiques du contrat « Globale »

Le contrat comporte 921 polices avec une garantie moyenne demandée par acheteur de 15,1k€. Il y a en moyenne 339 acheteurs par police (minimum 1 et maximum 7 938). L'encours total agréé est de 4 731,4 M€ avec un taux d'acceptation des demandes de limite par l'arbitrage (agréments et agréments express) de 87%.

761 polices sont tarifées sur la base d'un pourcentage de la masse assurable tandis que 160 sont à prime forfaitaire.

La prime cumulée du produit sur les polices observées est de 11,127M€ sur une base annuelle.





Nous utilisons les montants agréés par l'arbitrage actuellement en place. Celui-ci est réalisé sur la base du modèle de *scoring* de Coface.

3.3.2 Tarification du contrat « Globale »

Pour les données utilisables (c'est-à-dire pour lesquelles nous disposons de l'ensemble des informations nécessaires), nous calculons les primes sur 744 polices en portefeuille. Celles-ci représentent un montant de prime de 9, 403 M€.

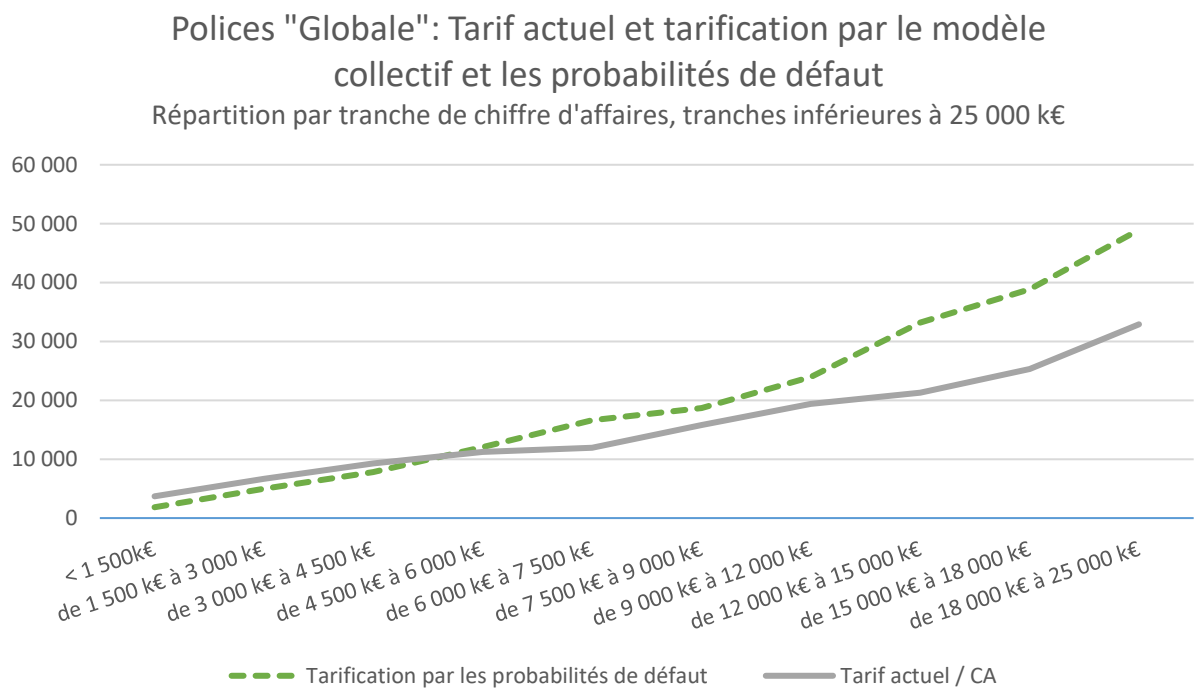
Sur la base de coûts à 40% de la prime, nous tarifons le produit « Globale » à 14,06 M€ (hors marge de l'assureur), soit presque 50% au-dessus du montant actuel.

Nous faisons des classes de chiffre d'affaires pour comparer les primes tarifées par les probabilités de défaut des primes actuelles selon ce critère de différenciation. Pour cela nous décomposons notre échantillon en 12 classes. Les résultats sont résumés dans le tableau de la page suivante.

Tranche de chiffre d'affaires de l'assuré	Tarif actuel	Tarification par les probabilités de défaut	Effectif en pourcentage de l'effectif total
< 1 500k€	3 694	1 839	15%
de 1 500 k€ à 3 000 k€	6 671	4 992	26%
de 3 000 k€ à 4 500 k€	9 284	7 803	17%
de 4 500 k€ à 6 000 k€	11 238	12 041	10%
de 6 000 k€ à 7 500 k€	11 954	16 622	6%
de 7 500 k€ à 9 000 k€	15 807	18 684	6%
de 9 000 k€ à 12 000 k€	19 386	23 978	6%
de 12 000 k€ à 15 000 k€	21 302	33 222	3%
de 15 000 k€ à 18 000 k€	25 312	38 840	2%
de 18 000 k€ à 25 000 k€	32 901	48 942	4%
de 25 000 k€ à 50 000 k€	43 151	87 238	2%
> 50 000 k€	69 211	211 343	2%

Nous constatons une forte divergence entre la tarification par les probabilités de défaut et le tarif actuel sur les assurés à chiffre d'affaires élevé. Cette divergence devient très forte sur les deux dernières classes de chiffre d'affaires (34 polices sont concernées, soit 4,5% de notre échantillon, mais 29% des montants de primes actuelles), nous avons extrait ces polices qui nécessitent en effet un traitement plus spécifique et plus détaillé.

En excluant ces 34 polices dont les clients présentent les chiffres d'affaires les plus élevés, nous obtenons la dynamique suivante :



Nous constatons le même phénomène de forte croissance des tarifs sur les tranches plus élevées de chiffre d'affaires. En excluant ces 34 polices, le total des primes par application de la tarification par les probabilités de défaut est de 7,3M€ contre 6,7M€ pour les primes actuelles, soit 8,8% au-dessus. Nous reviendrons sur cette tarification dans la synthèse (partie 5.1).

3.4 Application au contrat « Kup »

Nous travaillons maintenant sur les données du contrat « Kup » et nous allons appliquer notre modèle de tarification sur l'ensemble du portefeuille d'assurés.

Pour cela nous disposons :

- Des 544 polices composant ce produit avec la prime forfaitaire appliquée ;
- Du détail par police des acheteurs avec leur Siren, le montant d'agrément demandé par l'assuré (ou s'il s'agit de NDS de la limite de 5 000 € prévue par la grille tarifaire actuelle du produit ou 10 000 € prévue dans l'ancien contrat KUP s'il est resté en vigueur) ;
- Du chiffre d'affaires des assurés ;
- Du modèle de *scoring* ;
- De la grille de probabilité de défaut en fonction du score.

Lorsqu'un acheteur n'est pas présent dans la base de *scoring*, nous lui attribuons un score de 4 par défaut. Ce choix arbitraire est fait pour intégrer l'effet de l'arbitrage qui élimine généralement les trois niveaux de score les plus faibles.

La quotité garantie appliquée est contractuelle, à savoir 80% pour les agréments express et 90% pour les agréments.

Nous fixons le délai de paiement à 3 mois, la cyclicité à 4 facturations par an. Ce délai de paiement est au-delà de la moyenne nationale de 57 jours (Observatoire des délais de paiement 2016).

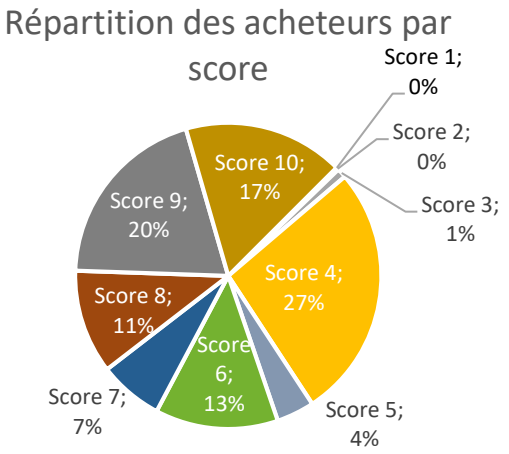
Le taux de recouvrement dépend du secteur. La « loss given default » a été estimée à l'aide de la base sinistres comme indiqué dans la partie 3.1.2.

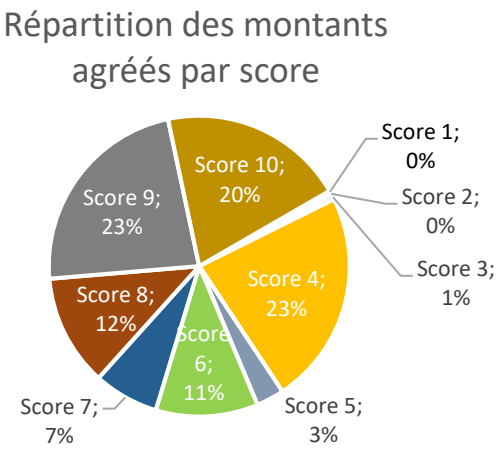
3.4.1 Caractéristiques du contrat Global Kup

Le contrat comporte 544 polices avec une garantie moyenne demandée par acheteur de 8 800 €. Il y a en moyenne 225 acheteurs par police (minimum 1 et maximum 4 822). L'encours total agréé est de 929,97 M€ avec un taux d'acceptation des demandes de limite (agréments et agréments express) de 86%.

La prime cumulée du produit est de 2,5M€ par année.

Les graphiques suivants indiquent la répartition des acheteurs par score, en nombre d'acheteurs et en montant par acheteur.





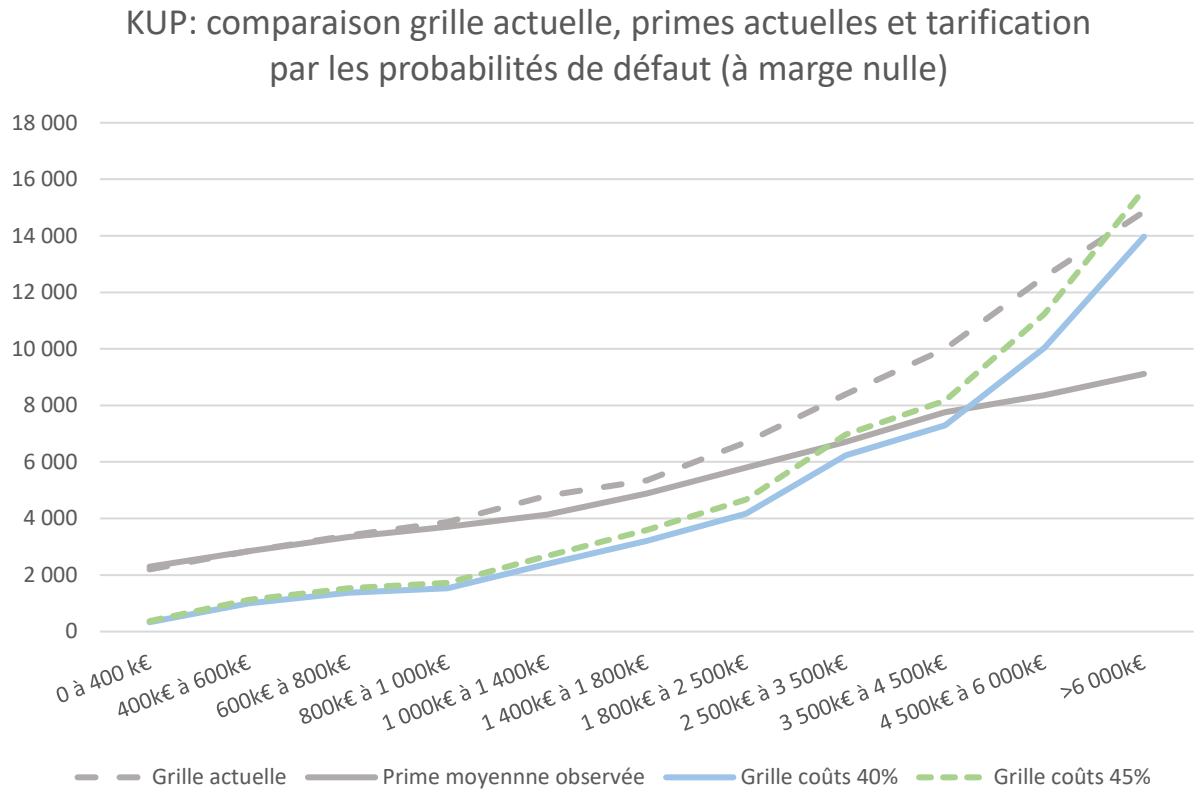
Nous utilisons les montants agréés par l'arbitrage actuellement en place. Celui-ci est réalisé sur la base du modèle de *scoring* de Coface.

3.4.2 Tarification du produit Global KUP

Pour les données utilisables (c'est-à-dire pour lesquelles nous disposons de l'ensemble des informations nécessaires), nous calculons les primes sur 448 polices en portefeuille. Celles-ci représentent un montant de primes de 2,204 M€.

Sur la base de coûts à 40% de la prime, nous tarifons le produit KUP à 1,668 M€ (hors marge de l'assureur), soit 24% en dessous du montant actuel. Sur la base de coûts à 45% de la prime, le tarif hors marge de l'assureur s'élève à 1 867 k€, soit 15,3% en dessous de ce qui est actuellement perçu.

L'outil nous permet de re-tarifer l'ensemble des polices puis de recomposer la prime unitaire par tranche de chiffre d'affaires comme indiqué sur le graphique suivant :



Nous constatons que ce modèle estime des primes faibles sur les tranches basses de chiffre d'affaires. Celles-ci sont en effet dépendantes du nombre d'acheteurs par assuré qui est faible pour les tranches basses de chiffre d'affaires.

Nous constatons également que les primes pratiquées ne correspondent pas à la grille tarifaire du produit. Ceci est d'autant plus vrai que le chiffre d'affaires est élevé. En effet, nous indiquons en introduction de ce rapport que le marché est très concurrentiel et que la variable d'ajustement est le prix, ceci en est une illustration.

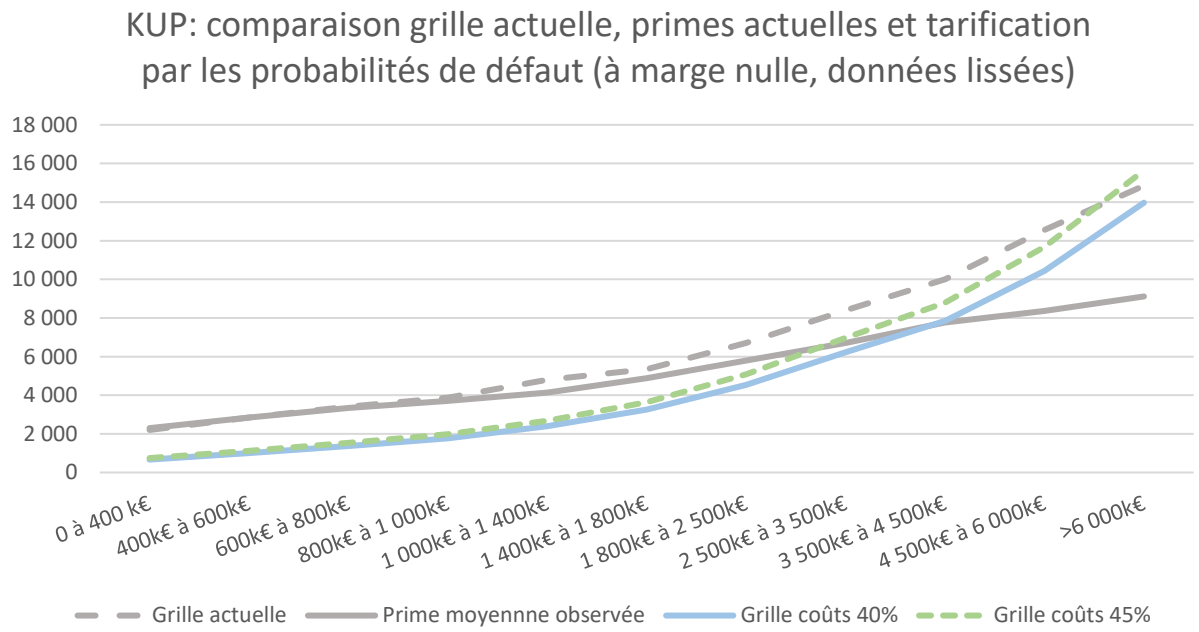
Ce premier outil de tarification permet de répartir la prime par tranche de chiffre d'affaires de l'assuré et ainsi recomposer une grille tarifaire.

Répartition de la prime calculée par tranche de chiffre d'affaires Global Kup (€)

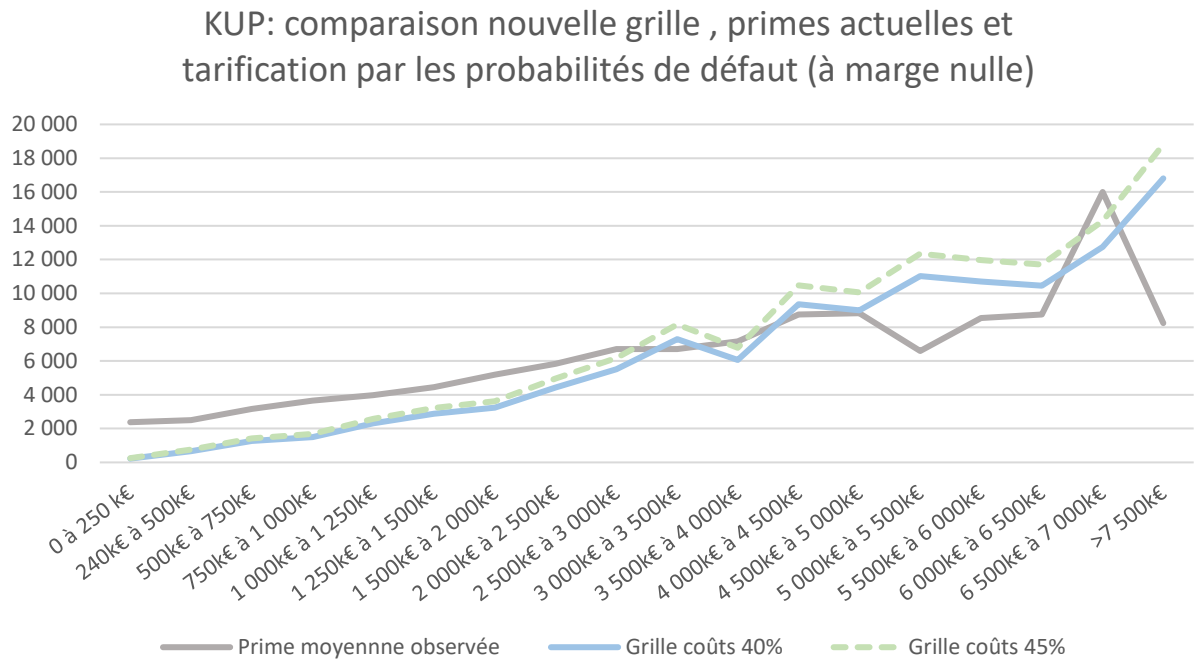
			Données brutes		Grille avec lissage (*)		
Tranches de chiffre d'affaires	Grille actuelle	Prime moyenne observée	Grille coûts 40%	Grille coûts 45%	Grille coûts 40%	Grille coûts 45%	Effectif en pourcentage de l'effectif total
0 à 400 k€	2 200	2 297	331	371	667	747	6%
400k€ à 600k€	2 850	2 842	1 003	1 123	1 003	1 123	8%
600k€ à 800k€	3 400	3 345	1 372	1 536	1 372	1 536	5%
800k€ à 1 000k€	3 880	3 707	1 536	1 720	1 766	1 977	8%
1 000k€ à 1 400k€	4 800	4 134	2 389	2 674	2 389	2 674	14%
1 400k€ à 1 800k€	5 350	4 884	3 206	3 589	3 256	3 645	17%
1 800k€ à 2 500k€	6 700	5 801	4 173	4 672	4 536	5 078	11%
2 500k€ à 3 500k€	8 400	6 710	6 228	6 973	6 228	6 973	13%
3 500k€ à 4 500k€	10 000	7 762	7 301	8 174	7 854	8 793	5%
4 500k€ à 6 000k€	12 550	8 359	10 033	11 232	10 435	11 682	6%
>6 000k€	14 850	9 114	13 970	15 640	13 970	15 640	7%

(*) : lissage des primes et nous imposons une croissance des primes sur les tranches croissantes de chiffre d'affaires

Le lissage permet d'avoir moins de variation et surtout nous imposons aux primes de suivre la croissance du chiffre d'affaires d'une tranche à une autre.



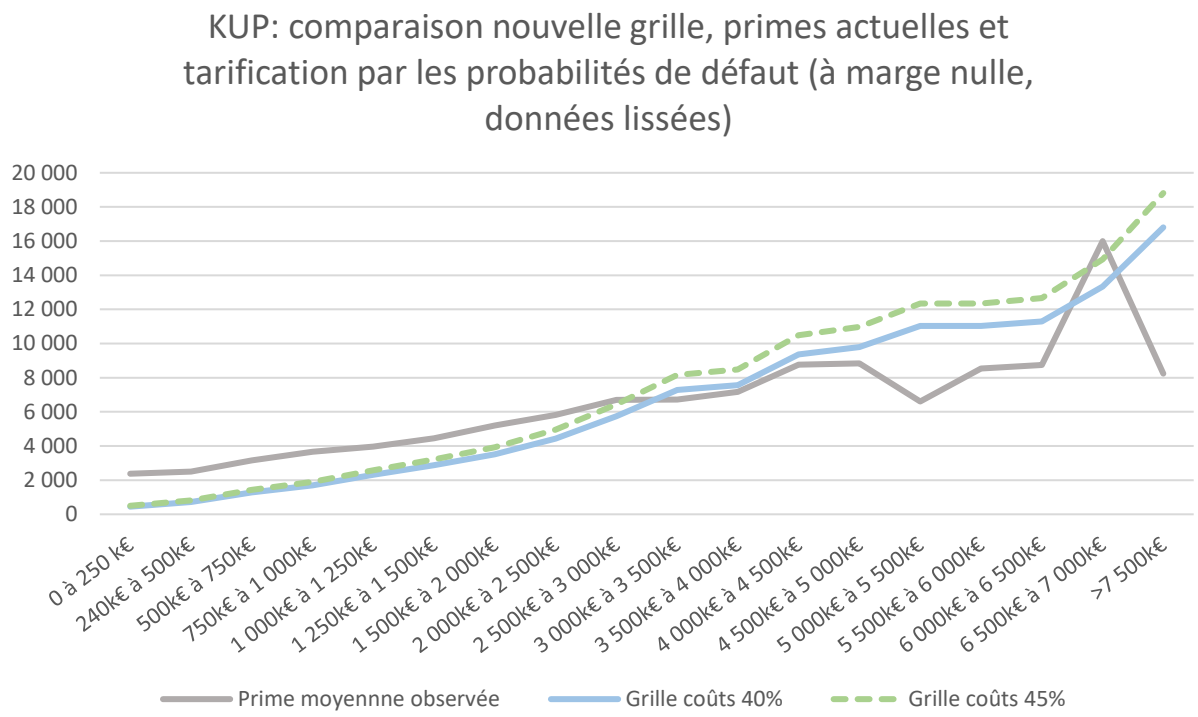
AXA Assurcrédit souhaite également étudier l'implémentation d'une nouvelle grille plus fine en termes de tranches de chiffre d'affaires (18 tranches au lieu de 11 actuellement). Nous faisons donc ce découpage :



Répartition de la prime calculée par tranche de chiffre d'affaires nouvelle : grille Kup (€)

		Données brutes		Grille avec lissage		
Tranches de chiffre d'affaires	Prime moyenne observée	Grille coûts 40%	Grille coûts 45%	Grille coûts 40%	Grille coûts 45%	Effectif en pourcentage de l'effectif total
0 à 250 k€	2 377	224	251	447	500	3%
240k€ à 500k€	2 498	670	750	723	809	7%
500k€ à 750k€	3 154	1 275	1 427	1 275	1 427	9%
750k€ à 1000k€	3 662	1 502	1 682	1 694	1 897	9%
1 000k€ à 1 250k€	3 972	2 305	2 581	2 305	2 581	9%
1 250k€ à 1 500k€	4 448	2 872	3 215	2 872	3 215	10%
1 500k€ à 2 000k€	5 197	3 233	3 620	3 512	3 932	14%
2 000k€ à 2 500k€	5 827	4 432	4 962	4 432	4 962	9%
2 500k€ à 3 000k€	6 709	5 521	6 181	5 747	6 434	8%
3 000k€ à 3 500k€	6 713	7 289	8 160	7 289	8 160	5%
3 500k€ à 4 000k€	7 164	6 063	6 788	7 572	8 477	3%
4 000k€ à 4 500k€	8 758	9 364	10 484	9 364	10 484	2%
4 500k€ à 5 000k€	8 833	8 992	10 067	9 796	10 968	2%
5 000k€ à 5 500k€	6 600	11 033	12 352	11 033	12 352	1%
5 500k€ à 6 000k€	8 546	10 699	11 978	11 033	12 352	2%
6 000k€ à 6 500k€	8 753	10 461	11 711	11 305	12 656	3%
6 500k€ à 7 000k€	16 000	12 754	14 279	13 340	14 935	1%
>7 500k€	8 236	16 805	18 814	16 805	18 814	4%

Un lissage des primes permet d'avoir un comportement mois erratique de la grille :



Nous reverrons ces résultats avec les données issues de la seconde modélisation testée.

3.5 Faiblesses et évolutions du modèle

3.5.1 Hypothèses relatives à la construction du modèle

Le modèle, tel que décrit, suppose que l'ensemble des acheteurs d'un assuré soit connu à la date de la tarification, ce qui exclut le « non dénommés » (ND) ainsi que la tarification de prospects pour lesquels nous n'avons pas d'information exhaustive du portefeuille d'acheteurs.

Pour les assurés en portefeuille, qui auraient également des garanties en « non-dénommés », il serait envisageable, mais bien sûr plus long, d'estimer le nombre d'acheteurs correspondant au « non-dénommés », leur volume d'encours annuel garanti ainsi que leur score moyen et sa dispersion. En effet, l'assuré déclare son chiffre d'affaires annuellement avec son nombre d'acheteurs garantis en « dénommés » ou en « non-dénommés », et compte-tenu de son secteur d'activité, nous pourrons faire une estimation réaliste du niveau de score de défaillance de ses acheteurs en « non-dénommés ». Cette

approche permettrait donc d'appliquer également le modèle de tarification à des assurés disposant de garanties en « non-dénommés ».

Pour les prospects qui sont en cours de négociation et qui nécessitent une tarification, celle-ci ne peut en général s'appuyer que sur l'échantillon d'acheteurs communiqué par le prospect à l'assuré (5% à 50% du nombre total d'acheteurs en général). En effet, cet échantillon fait l'objet d'une Etude de Clientèle (EDC) communiquée par l'assureur au prospect indiquant quelles garanties pourraient être accordées en totalité de la demande ou partiellement. Cet échantillon est donc censé être un bon « proxi » de l'ensemble des acheteurs du prospect. L'assureur n'ayant par ailleurs que l'information du secteur, du chiffre d'affaires total assurable et du nombre d'acheteurs assurables à garantir, il faudrait donc adapter l'approche du modèle de tarification en faisant une extrapolation à partir de l'échantillon de l'EDC pour estimer la donnée des scores de défaut du reste des acheteurs.

3.5.2 Non-indépendance des événements de défaut

Nous avons fait l'hypothèse implicite de l'indépendance des événements de défaillance. Cette hypothèse a été prise dans un souci de simplicité. Il est néanmoins intuitif de penser que, dans une économie cyclique, les défauts des entreprises ont tendance à survenir plus fréquemment lorsque la situation de l'économie globale ou d'un secteur spécifique se dégrade.

Nous reprenons ici, pour explorer une façon de traiter ce problème de corrélations, une partie de l'article de Messieurs J.F. Decroocq, F. Planchet et F. Magnin concernant la modélisation du risque systématique en assurance-crédit (Planchet, Decroocq et Magnin 2009) ainsi que les travaux de Monsieur A. Dupont (Dupont 2010). L'objectif est d'étudier des premières pistes pour mettre en place une structure de corrélations.

Nous considérons que chaque acheteur i fait défaut lorsque sa capacité à payer, notée Z_i passe sous un seul noté d_i. Nous nous plaçons dans un modèle à un facteur et nous supposons qu'une variable Y sera le facteur économique (ou état de l'économie).

Nous pouvons alors écrire Z_i de la manière suivante :

$$Z_i = \sqrt{\rho}Y + \sqrt{1-\rho}\varepsilon_i$$

On supposera $Y \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{N}(0,1)$ et Y indépendant de ε. ε est le facteur spécifique ou facteur de risque idiosyncratique.

Le coefficient ρ est la part d'exposition de la capacité à payer Z au risque systémique, représenté par le facteur macroéconomique Y.

Conditionnellement à Y, les variables $Z_i, i \in \{1, \dots, n\}$ sont indépendantes entre elles et vérifient :

$$\mathbb{P}(Z_i < d_i|Y) = \Phi\left(\frac{d_i - \sqrt{\rho}Y}{\sqrt{1-\rho}}\right)$$

Avec Φ la fonction de répartition de la loi normale centrée réduite. Cette dernière étant croissance, en posant 0 la valeur d'équilibre du facteur macroéconomique (car il suit une loi normale centrée réduite), nous constatons que la probabilité de défaut conditionnelle augmente lorsque Y est sous sa moyenne.

Deux difficultés seront propres à l'étude de PME :

- La variable Z_i, généralement le rendement de l'actif, n'est pas disponible. Il conviendra donc de rechercher des proxy (entreprises d'un même secteur par exemple).
- Le calcul du coefficient ρ, qui indique la plus ou moins grande sensibilité d'une entreprise au facteur systémique.

En supposant que $\mathbb{V}(Z_i) = \mathbb{V}(Y) = 1$, alors $\sqrt{\rho} = Cov(Z_i, Y)$.

Monsieur A. Dupont (Dupont 2010) indique dans son rapport que ce coefficient de sensibilité est relativement élevé pour les grandes entreprises cotées en bourse. En revanche, l'étude de cette corrélation est impossible pour les petites entreprises non cotées. L'auteur indique cependant que ces corrélations seront beaucoup plus faibles pour les petites entreprises que pour les grandes, les PME semblant moins sensibles aux petites variations économiques cycliques qui sont captées par ρ (mais plus fragiles aux crises).

Il est à noter que l'article de Messieurs JF Decroocq, F Planchet et F Magnin (Planchet, Decroocq et Magnin 2009) indique en substance que si les assureurs crédit couvrent des risques liés aux grandes entreprises, l'essentiel de leur exposition est sur des entreprises de petites et moyennes tailles. Les auteurs indiquent qu'un certain nombre d'analyses montre que la corrélation des défauts des PME est très faible. D'autre part, l'atténuation des risques via une politique d'arbitrage plus restrictive est également plus efficace sur les petites entreprises, en particulier en cas de ralentissement de l'activité où les agréments peuvent être plus facilement réduits sur ce type de société conduisant à réduire de manière significative les pertes de l'assureur.

Par conséquent, le risque global lorsque les acheteurs sont des PME est plus lié à son risque spécifique. Ceci nous permet d'admettre, en première approche, notre hypothèse simplificatrice d'indépendance des événements de défaillance, AXA Assurcrédit s'étant concentré sur une clientèle de petites et moyennes entreprises.

4 Modélisation d'une prime par le modèle collectif

4.1 Cadre théorique

4.1.1 Les modèles linéaires généralisés

Historiquement, les statisticiens ainsi que les actuaires utilisaient principalement des régressions linéaires (simples ou multiples). L'utilisation de modèles plus appropriés, notamment en tarification, s'est effectuée à la fin du 20ème siècle par les actuaires londoniens de la City University ; ils ont utilisé des modèles linéaires généralisés (GLM, pour Generalized Linear Models).

Les GLM (Wajnberg 2011) permettent de s'affranchir de l'hypothèse de normalité des résidus du modèle, en traitant les lois faisant partie de la famille exponentielle (loi Binomiale, loi Normale, loi de Poisson, loi Gamma et loi Inverse Gaussienne).

Dans la famille des lois exponentielles, les lois de probabilité ont deux paramètres θ et ϕ dont la densité (discrète ou continue) peut s'écrire :

$$f(y/\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right), y \in \mathrm{S}$$

Où S est un sous-ensemble de $\mathbb{N}$ ou de $\mathbb{R}$. θ est alors appelé le paramètre naturel et ϕ le paramètre de dispersion.

Nous pouvons parfois être amenés à utiliser une pondération (modélisation de la fréquence par exemple), nous remplaçons alors ϕ par ϕ/ω où ω est le poids.

Nous souhaitons alors garder la relation linéaire qui lie $Y \in \mathbb{R}^{n\times 1}$ (la variable à expliquer) et $X \in \mathbb{R}^{n\times p+1}$ (les p variables explicatives). Nous passons alors à un modèle de régression du type $Y \sim \mathcal{L}(\mu)$ où $\mu = \mathbb{E}[Y] = g^{-1}(X\beta)$. On appelle alors g la fonction de lien.

Chacune des lois de probabilité de la famille exponentielle possède une fonction de lien « spécifique », dite fonction de lien canonique, définie par θ, le paramètre naturel. Il est défini par $\theta = g(\mu)$ où $\mu = g^{-1}(\theta)$, c'est-à-dire le lien qui nous permet de revenir à $\theta = X\beta$.

Il est possible de choisir, en théorie, n'importe quelle fonction de lien g (bijective). Le tableau ci-après indique le lien canonique possible sur le logiciel utilisé R pour les différentes lois de probabilité.

Loi de probabilité	μ	μ^{-1}	$\sqrt{\mu}$	$log(\mu)$	μ^{-2}	$logit(\mu)$	$\varphi^{-1}(\mu)$
Normale	*	*		*			
Poisson	*		*	*			
Gamma	*	*		*			
Inverse Gaussienne	*	*		*	*		
Binomiale					*	*	*

4.1.2 Le modèle collectif

Les assureurs vendent un service via un contrat dans lequel ils s'engagent à indemniser leurs assurés, en cas de sinistre, durant la période du contrat (généralement un an). La particularité de cette activité est donc de vendre un service avant d'en connaître le coût, d'où la nécessité pour l'assureur d'estimer la prime avec la plus grande fiabilité possible. Logiquement, la prime pure (celle qui permet de ne couvrir que les sinistres à venir hors frais et marge) se calcule par l'espérance mathématique des indemnisations à venir. Le modèle collectif s'appuie sur deux variables distinctes, l'une de fréquence des sinistres N et l'autre de sévérité des sinistres Y (Charpentier et Denuit, Mathématiques de l'assurance non-vie - Tome I: Principes fondamentaux de théorie du risque 2004).

Soient S le montant de sinistre total subi par l'entreprise, N le nombre de sinistres (la fréquence) et Y le montant des sinistres individuels (la sévérité), le modèle sera de la forme :

$S = \sum_{i=1}^{N} Y_i$, avec S(ω)=0 si N(ω)=0.

Nous sommes en présence de deux variables aléatoires, N et Y.

L'une des propriétés de ce modèle est que l'espérance mathématique des sinistres est égale au produit des espérances mathématiques de la fréquence et de la sévérité : $\mathbb{E}[S] = \mathbb{E}[N]\times\mathbb{E}[Y]$, sous réserve que les variables soient indépendantes et identiquement distribuées.

L'objectif est alors d'essayer d'expliquer ces variables aléatoires N et Y à l'aide des données disponibles (regroupées dans un vecteur appelé X) afin de segmenter la population d'assurés en fonction de leurs caractéristiques, on a ainsi $\mathbb{E}[S] = \mathbb{E}[N/X]\times\mathbb{E}[Y/X]$.

La méthode d'estimation utilisée sera le modèle linéaire généralisé (GLM). La tarification va consister à :

- Calibrer un modèle sur la fréquence : $\mathbb{E}(N) = g_1^{-1}(X^T\beta_1)$
- Calibrer un modèle sur la sévérité : $\mathbb{E}(Y) = g_2^{-1}(X^T\beta_2)$

Où g est la fonction lien utilisée dans le cadre du GLM, X le vecteur de données explicatives, β le vecteur des paramètres permettant de différencier les tarifs.

Les variables explicatives X peuvent être numériques ou catégorielles. Les variables numériques (ex : Chiffre d'affaires) sont, soit regroupées par classes (tranche de chiffre d'affaires par exemple), soit modélisées selon leurs caractéristiques propres (relation linéaire, polynomiale…).

4.2 Construction des bases

Nous disposons des données suivantes :

- Une base « sinistres » regroupant l'ensemble des sinistres subis depuis 10 ans ;
- Le portefeuille des assurés à des intervalles réguliers depuis début 2012 ;
- Le portefeuille d'acheteurs des assurés aux mêmes intervalles depuis début 2012.

La base sinistres regroupe l'ensemble des sinistres sur 10 ans avec leurs caractéristiques (la police concernée, le montant du sinistre avant et après frais, avant et après recouvrement, le montant facturé par l'assuré auprès de son acheteur et le sinistre effectivement réglé).

Le portefeuille « assurés » regroupe les caractéristiques de la police et de l'assuré (secteur, adresse, SIREN, quotité garantie par type de couverture, type de contrat, nombre d'acheteurs en agréments ou « non dénommés surveillés », date de résiliation de la police si elle est résiliée et la date de souscription de la police). Il y a entre 1 600 et 2 000 assurés en portefeuille selon l'année.

Le portefeuille d'acheteurs regroupe quelques caractéristiques de chaque acheteur (raison sociale, pays, ville, code postal, score Coface, police de rattachement, type de limite - agrément ou NDS -, encours agréé). Il comprend environ 500 000 acheteurs. Nous conservons ici le score Coface car c'est celui qui a servi à réaliser la sélection des risques en portefeuille et nous en avons l'historique.

Nous souhaitons pouvoir calculer une prime pure en prenant en compte les caractéristiques des bases existantes. L'enjeu est de pouvoir améliorer la tarification actuelle, mais également de pouvoir déléguer une partie des propositions de tarification des souscripteurs aux commerciaux ou intermédiaires agents et courtiers.

A la souscription d'un contrat d'assurance-crédit, l'assureur effectue une étude de clientèle basée sur un questionnaire type. Celui-ci est peu ou prou équivalent à ce que demande la concurrence, c'est une norme de marché. La mise en œuvre du modèle de tarification va essayer de modéliser les données disponibles au travers de ce questionnaire pour les confronter à la fréquence et l'intensité des sinistres.

Le questionnaire de souscription comprend les éléments suivants :

- Les éléments qualitatifs de l'assuré (SIREN, raison sociale, adresse, secteur d'activité…) ;
- Le chiffre d'affaires de l'assuré (exercices en cours et passés) ;

- La partie export (UE et Autres pays) ;
- La composition du chiffre d'affaires (sociétés apparentées, particuliers, commerçants et artisans, administrations et collectivités publiques, entreprises industrielles et commerciales, autres). Malheureusement ces données, servant à calculer le chiffre d'affaires assurable, ne sont pas stockées dans les bases de données et ne sont donc pas disponibles (en revanche, d'autres bases nous permettent d'accéder aux déclarations annuelles de chiffre d'affaires assurés annuelles qui ont été reçues par l'assureur de la part de l'assuré conformément aux dispositions contractuelles, elles ne sont cependant pas centralisées et non exportables en l'état) ;
- Le nombre de clients par tranche d'encours (inférieur à 5k€, entre 5 et 10k€, entre 10 et 20k€, entre 20 et 50k€, entre 50 et 150 k€, entre 150 et 450 k€ puis au-delà de 450 k€). Cette donnée sera approximée par le nombre de demandes d'agréments et d'agréments express en portefeuille sur chacune de ces tranches ;
- L'étude des principaux clients de l'assuré (de 10 à plusieurs dizaines selon les cas) avec le nom du client, son SIREN et le montant de l'encours ;
- Les créances irrécouvrables ou douteuses et les principales défaillances (malheureusement ces données restent indisponibles informatiquement dans les bases) ;
- Des éléments qualitatifs sur la gestion du risque de crédit clients mis en place par le prospect.

Nous reconstituons une partie de ces éléments avec les bases disponibles.

L'étape de gestion de données et de constitution de la base de données préalable à l'étude a nécessité plusieurs étapes.

Les portefeuilles acheteurs et assurés sont disponibles à des dates régulières. Nous les avons considérés à chaque début d'année civile. Nous avons donc reconstitué des années-polices sur une base calendaire et non sur la base de chacune des polices qui se renouvellent tout au long de l'année pour une durée d'un an.

Pour cela, nous prenons la base en début d'année, nous y intégrons les souscriptions de l'année en cours ainsi que les résiliations. Cela nous permet de calculer l'exposition de notre base assurés dans l'année afin de prendre en compte les assurés qui ne sont pas présents tout au long de la période. Ce travail nous permet de générer une base regroupant les caractéristiques des assurés sur une référence calendaire uniformisée.

Sur la base acheteurs, nous effectuons le même travail de regroupement des acheteurs par police à une date donnée, puis nous calculons ou extrayons les variables synthétiques ou qualitatives que nous souhaitons retenir (nombre d'acheteurs, montant et moyenne des agréments etc…). La liste des variables générées est en annexe 6.3.1.

Une fois la base acheteurs générée pour chacune des années disponibles (de début 2012 à mi 2016), nous rattachons les informations acheteurs calculées ou extraites à notre base assurés pour chacune des années concernées.

La base sinistres fournit des informations sur chaque sinistre subit dans le portefeuille d'AXA Assurcrédit. Dans sa comptabilité, AXA Assurcrédit rattache un sinistre à l'année

correspondant à la première facture concernée par l'incident. C'est la règle que nous appliquons pour construire notre base de données. Nous rattachons ainsi à chaque assuré sinistré et pour chaque année le nombre de sinistres qu'il a connu dans l'année calendaire ainsi que le montant cumulé de ces sinistres. Ces informations sont ensuite rattachées à notre base assurés.

Nous agrégeons ensuite les cinq années de données disponibles afin de procéder aux modélisations.

Nous disposons ainsi de 9 193 observations de polices (tous produits confondus) sur les cinq années considérées. Nous distinguons les polices d'un même assuré sur chacune des années de présence en portefeuille.

En pratique, nous n'utilisons pas les six premiers mois de 2016 pour calibrer notre modèle. En effet, compte-tenu des délais de déclaration des sinistres et de la méthode de rattachement de ceux-ci à une police, nous risquons d'avoir des sinistres manquants pour le début de l'année en cours, du fait du délai de développement complet de la sinistralité en général.

4.3 Méthodologie générale

La méthodologie va être identique quel que soit le produit étudié (Bentahar et Vercherin 2013) :

Nous mettons en place une modélisation de la prime pure selon le modèle collectif fréquence / intensité. Nous n'intégrons pas à ce stade de modélisation d'événements rares via la théorie des valeurs extrêmes. En effet, le montant maximum d'un sinistre est limité de toutes façons par l'agrément d'une part, puis par un montant maximum d'indemnité sur une police d'autre part (généralement 20 à 30 fois la prime payée). Il est à noter que nous sommes ainsi plus face à un risque de fréquence (multiplication du nombre de sinistres) que d'intensité.

Nous utilisons principalement des modèles linéaires généralisés pour estimer les paramètres.

Nous calibrons nos modèles sur une base de test (deux tiers des données) puis nous testons la robustesse sur une base de validation (un tiers des données).

Nous déterminons ensuite une prime en réintégrant les frais et le coût de la mobilisation des capitaux propres réglementaires générés par chaque contrat (selon le principe de prime présenté dans la partie 3.2).

4.4 Produit « Globale »

Le contrat « Globale » fait l'objet d'une tarification « sur mesure », c'est-à-dire qu'il n'existe pas de grille de tarification. L'assuré est étudié via un questionnaire qu'il remplit

préalablement à la souscription, de cette étude découlera le tarif le plus souvent exprimé en pourcentage du chiffre d'affaires.

L'objet de la tarification va donc être de calibrer un modèle fréquence / coût en utilisant les variables à disposition. L'objectif est d'avoir un outil d'aide à la décision issu de la sinistralité constatée sur le portefeuille existant.

Nous disposons d'un portefeuille de 4 années polices (2012 à 2015) représentant 4 319 assurés. Nous n'utilisons pas l'année 2016 pour calibrer le modèle, la sinistralité étant faible à fin juin (date de disponibilité des données) en raison des délais de déclarations et de développement des sinistres.

4.4.1 Modélisation de la fréquence des sinistres

4.4.1.1 Analyse des données

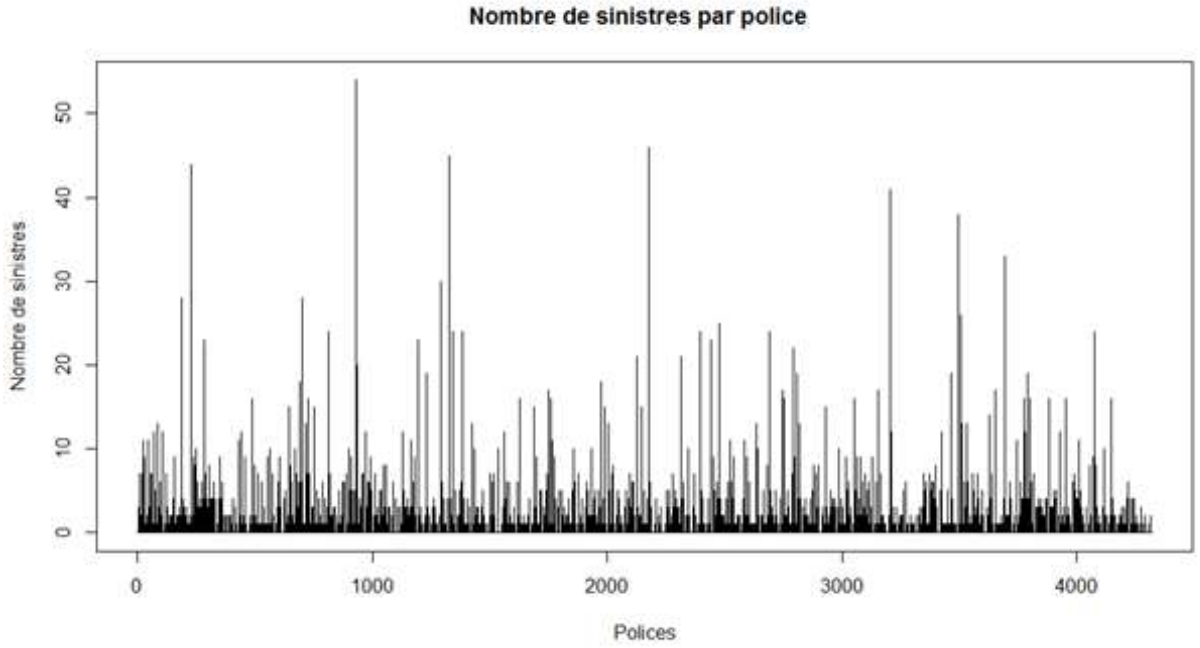
La sinistralité moyenne est de 1,19 sinistre par assuré. 36% des assurés ont au moins un sinistre sur la période 2012-2015.

Nous observons d'abord les répartitions du nombre de sinistres selon les assurés :

	Nombre de sinistres / base totale	Nombre de sinistres / assurés sinistrés
Minimum	0	1
Quantile 25%	0	1
Médiane	0	2
Moyenne	1,19	3,284
Quantile 75%	1	3
Quantile 90%	3	7
Quantile 95%	6	11
Quantile 99%	16	24
Max	54	54
Ecart type	3,21	4,64

Même si le risque maximum est limité par un montant maximum d'indemnité, nous observons que nous sommes face à un risque de fréquence de sinistres élevés, un assuré ayant totalisé, par exemple, 54 sinistres sur une année dans son contrat.

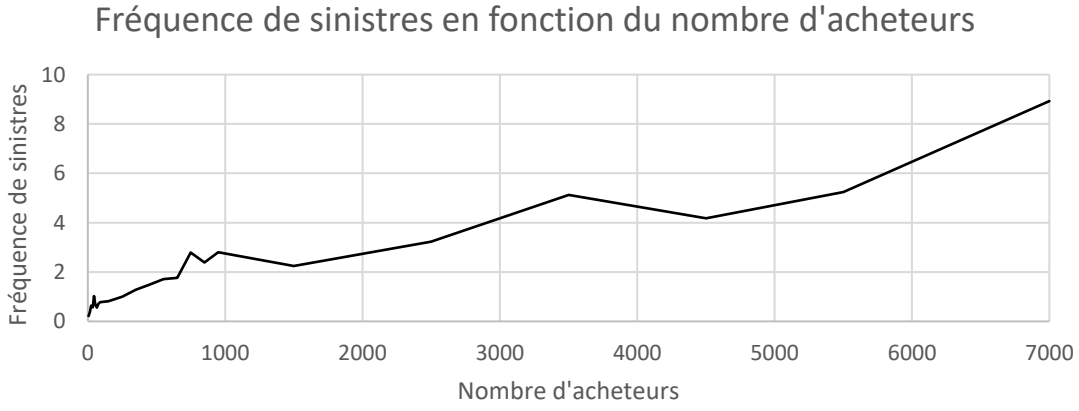
Nous observons également le nombre de sinistres par police :



Quelques contrats enregistrent ainsi une forte sinistralité en fréquence mais seront plafonnés en montant.

Nous étudions d'abord le comportement entre chaque variable explicative disponible et la fréquence des sinistres afin d'apprécier la relation qu'il peut y avoir. Deux outils sont utilisés pour cette étape, l'analyse graphique et le V de Cramer.

A titre d'exemple, la variable explicative « nombre d'acheteurs » qui regroupe le nombre d'acheteurs d'une police ayant bénéficié d'un agrément ou d'un agrément express



L'observation graphique nous laisse supposer qu'il y a une relation croissante entre le nombre d'acheteurs et la fréquence des sinistres (ce qui est assez intuitif).

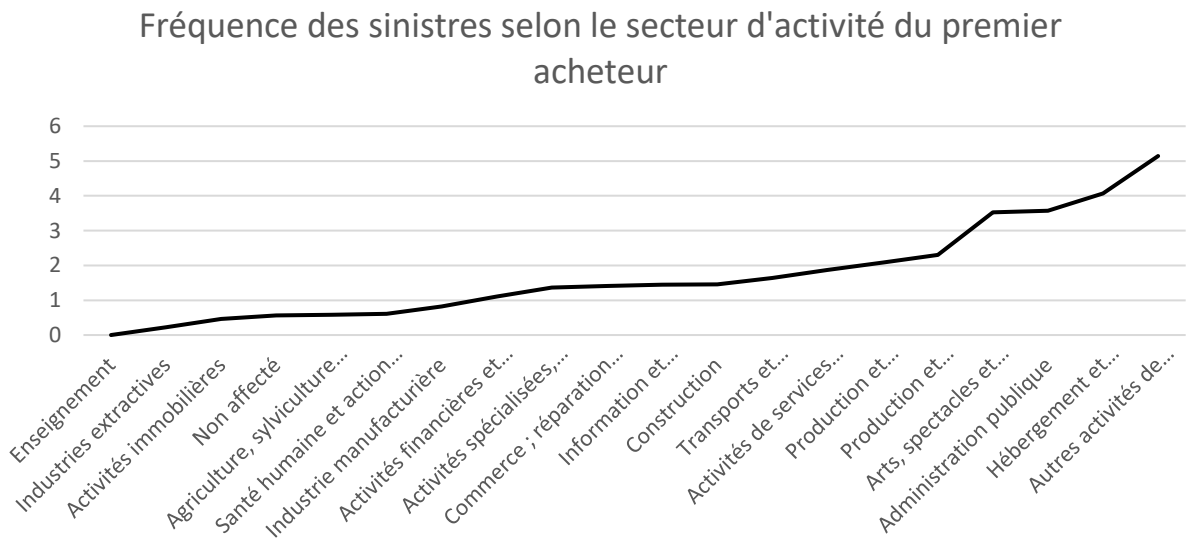
Nous calculons également les V de Cramer entre les variables nombre d'acheteurs et fréquence de sinistres, mais également avec des variables nombre d'agréments express et nombre d'agréments. Ces statistiques nous confirment le lien entre ces variables.

V de Cramer	Nombre de sinistres	Nombre d'acheteurs	Nombre d'agréments	Nombre d'agréments Express
Nombre de sinistres	100%	76%	74%	68%
Nombre d'acheteurs	76%	100%	82%	84%
Nombre d'agréments	74%	82%	100%	79%
Nombre d'agréments Express	68%	84%	79%	100%

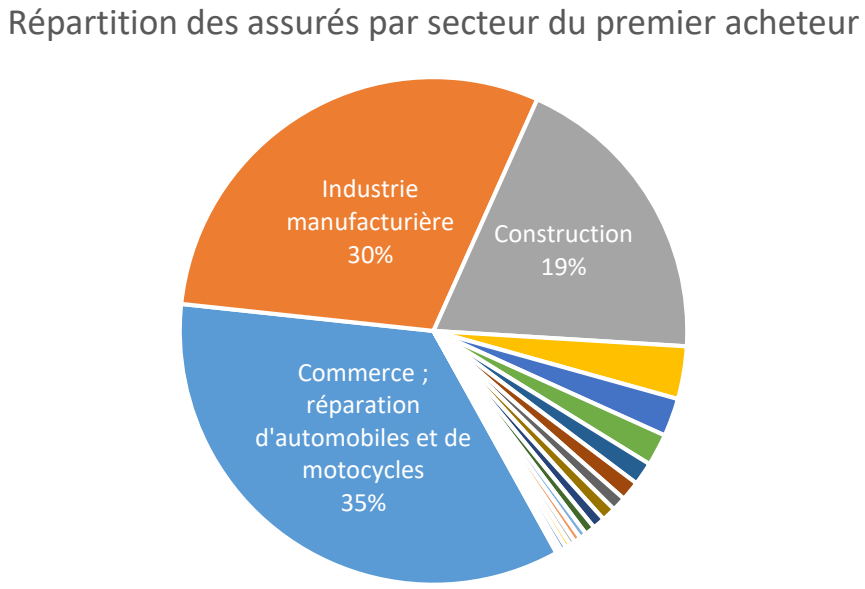
Nous ne retiendrons au final que le nombre d'acheteurs dans nos estimations car c'est la variable qui a la plus forte corrélation avec le nombre de sinistres d'une part, et c'est également celle qui se révélera la plus significative dans nos estimations d'autre part.

Nous retenons le secteur d'activité du premier acheteur plutôt que celui de l'assuré. En effet, le risque assuré est celui du défaut des acheteurs et non pas, bien évidemment, de l'assuré lui-même.

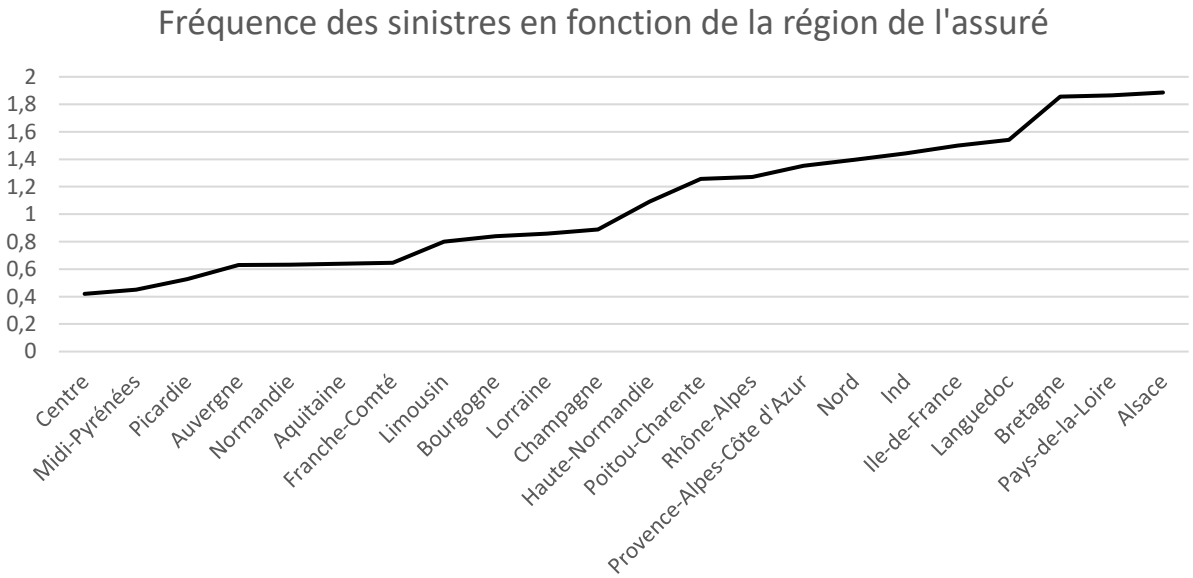
Le produit d'assurance-crédit assure le risque des acheteurs et non celui de l'assuré. De plus cette variable apporte une meilleure corrélation avec la fréquence des sinistres.



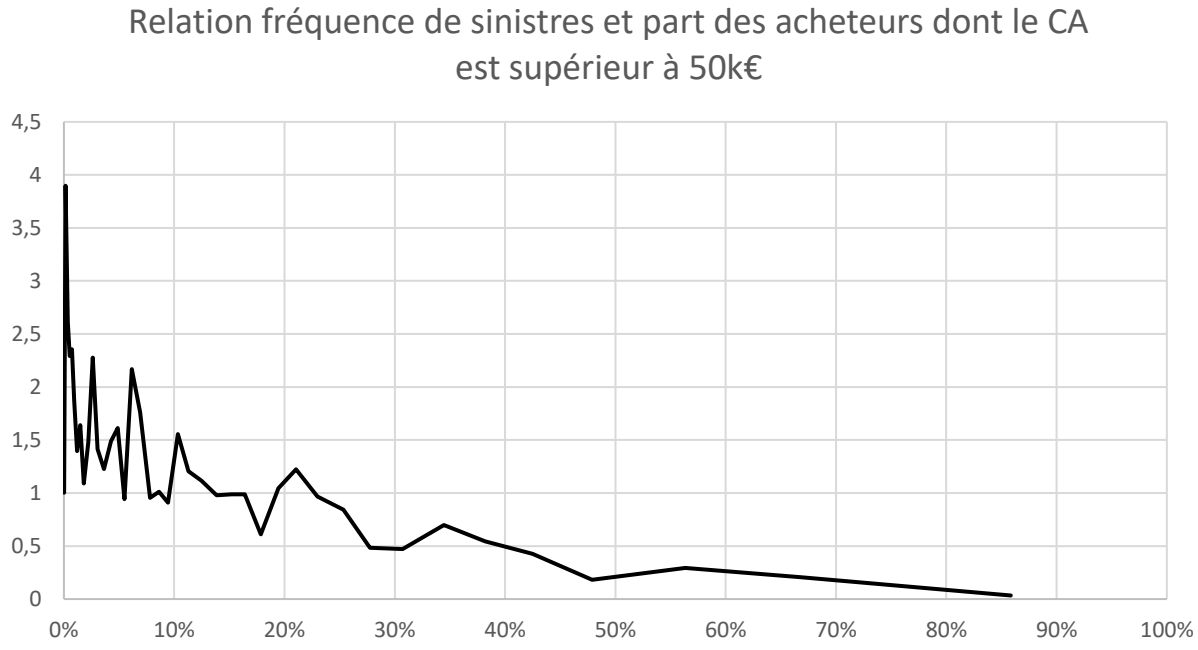
Il est à noter que trois secteurs regroupent une grande partie des acheteurs (graphique suivant), nous ne retiendrons donc que ces trois secteurs.

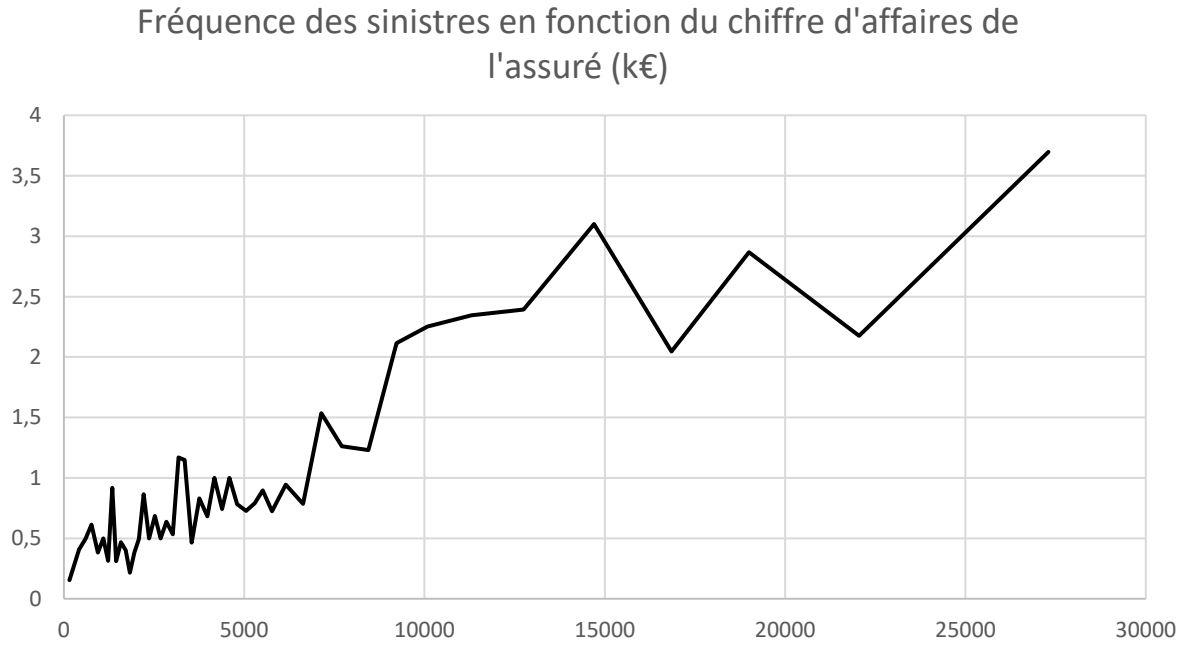


Nous retenons également la région de l'assuré.



D’autres variables sont retenues comme la part des acheteurs de plus de 50k€ de CA ou le chiffre d’affaires de l’assuré.





En fonction de la forme des courbes, nous procédons à des regroupements (essentiellement par classes). Les modélisations ou regroupements de variables explicatives ne sont conservés que s’ils permettent d’améliorer les paramètres du modèle.

Lorsque c'est nécessaire, les regroupements en catégories des variables explicatives sont faits soit de manière supervisée (choix arbitraires de classes), soit via des découpages endogènes (nous estimons un arbre de régression puis nous découpons notre variable en intervalles selon les critères de séparations effectuées à chaque nœud de l'arbre).

4.4.1.2 Estimation et sélection du modèle

Nous testons donc la significativité des coefficients estimés ainsi que la fonction lien du GLM. Pour cette dernière, nous testons les lois de Poisson, quasi-Poisson, Tweedie et binomiale négative.

- Loi de Poisson

Soit N, notre variable de comptage. Si N suit une loi de Poisson, nous aurons alors :

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

$$\text{Et } \mathbb{E}(N) = \mathbb{V}(N) = \lambda$$

Nous avons donc, dans le cadre de la loi de Poisson, égalité de l'espérance et de la variance (équi-dispersion).

- Loi quasi-Poisson

La loi de quasi-Poisson permet d'introduire une sur-dispersion, c'est-à-dire qu'elle permet de fixer un paramètre supplémentaire autorisant de ne pas avoir égalité entre l'espérance et la variance :

$$\mathbb{V}(N) = \varphi\mu$$

- Loi de Tweedie

La loi de Tweedie est une loi de Poisson composée avec des sauts suivant une loi Gamma.

- Loi binomiale négative

La loi binomiale négative permet également la sur-dispersion. C'est une loi « mélange » d'une loi de Poisson et d'une loi Gamma :

$$\mathbb{P}(N = k) = \frac{\Gamma(k + \alpha^{-1})}{\Gamma(k+1)\Gamma(\alpha^{-1})}\left(\frac{1}{1 + {\lambda}/{\alpha}}\right)^{\alpha^{-1}}\left(1 - \frac{1}{1 + {\lambda}/{\alpha}}\right)^{k}, \forall k \in \mathbb{N}$$

Nous avons deux types de lois binomiales négatives :

Type 1 : $\mathbb{E}(N) = \lambda$ et $\mathbb{V}(N) = \lambda + \alpha\lambda$

Type 2 : $\mathbb{E}(N) = \lambda$ et $\mathbb{V}(N) = \lambda + \alpha\lambda^2$

Les modèles ont tous été calibrés sur le même échantillon comprenant deux tiers de la base de données initiale. Les comparaisons de modèles se font avec le tiers de données restantes (base de validation). Nous faisons des prévisions de sinistralité avec la base de validation que nous comparons avec la réalisation. Le modèle retenu est celui qui minimise la somme du carré des erreurs (MSE) ainsi que l'erreur moyenne quadratique (RMSE).

Ces deux critères nous conduisent à retenir une loi de Poisson ou quasi-Poisson. Cependant, un test de sur-dispersion (test d'égalité entre l'espérance et la variance) nous conduit à sélectionner la quasi-Poisson plutôt que la loi de Poisson. Le coefficient de dispersion est en effet de 3,2 ; nous en concluons que la variance est significativement supérieure à l'espérance.

Le modèle permettant de minimiser MSE et RMSE retenu sera donc celui de quasi-Poisson pour cette partie. Les coefficients obtenus sont résumés dans le tableau suivant :

	Estimate	Std. Error	t value	Pr(>|t|)	Significativité
(Intercept)	-3,074	2,93E-01	-1,05E+01	2,18E-25	***
RégionAquitaine	-9,6E-01	3,25E-01	-2,95E+00	3,21E-03	**
RégionR1	-0,87353	2,46E-01	-3,55E+00	3,93E-04	***
RégionBasse-Normandie	1,03840	2,37E-01	4,38E+00	1,23E-05	***
RégionR2	-0,09736	1,96E-01	-4,98E-01	6,19E-01	
RégionR5	0,01503	1,62E-01	9,29E-02	9,26E-01	
RégionChampagne	-0,30191	2,45E-01	-1,23E+00	2,18E-01	
RégionFranche-Comté	-0,77453	2,63E-01	-2,94E+00	3,30E-03	**
RégionHaute-Normandie	-0,38382	2,37E-01	-1,62E+00	1,05E-01	
RégionR4	-0,12267	1,34E-01	-9,18E-01	3,59E-01	
RégionR3	0,09497	1,51E-01	6,30E-01	5,29E-01	
RégionLanguegoc	0,16171	2,06E-01	7,83E-01	4,33E-01	
RégionPicardie	-0,35492	3,73E-01	-9,52E-01	3,41E-01	
RégionPoitou-Charente	-0,09171	2,25E-01	-4,07E-01	6,84E-01	
secteur.ach1Commerce_rep_auto	-0,18905	9,44E-02	-2,00E+00	4,54E-02	*
secteur.ach1Construction	0,03533	1,07E-01	3,31E-01	7,41E-01	
secteur.ach1Manufacturier	-0,44519	1,07E-01	-4,15E+00	3,49E-05	***
QG_ND	0,01727	1,19E-03	1,45E+01	7,39E-46	***
limite_NDS	0,00002	1,33E-05	1,80E+00	7,26E-02	.
score_mini_10_acheteurs	-0,02875	2,69E-02	-1,07E+00	2,86E-01	.
part_acheteur_sup_50	-2,02425	5,75E-01	-3,52E+00	4,42E-04	***
log(total_agrement_et_NDS)	0,52308	2,84E-02	1,84E+01	4,96E-72	***
log(moyenne_agrement_et_NDS)	-0,31459	9,24E-02	-3,40E+00	6,71E-04	***

Le secteur « secteur.ach1Commerce_rep_auto » est la catégorie INSEE (classification NAF 2008) « Commerce ; réparation d'automobiles et de motocycles ». Le secteur « secteur.ach1Manufacturier » est la catégorie INSEE « Industrie manufacturière ».

Les regroupements sont les suivants pour les régions :

- R1 : Auvergne, Centre, Midi-Pyrénées.
- R2 : Bourgogne, Limousin, Lorraine et Normandie.
- R3 : Rhône-Alpes et indéterminée.
- R4 : Ile de France, Nord et Provence-Alpes-Côtes d'Azur.
- R5 : Bretagne et Pays-de-la-Loire.

Nous choisissons de conserver les regroupements de régions en l'état malgré la faible significativité de certains coefficients. En effet, les fréquences de sinistralité nous paraissent trop éloignées les unes des autres pour faire plus de regroupements.

Les coefficients obtenus (colonne « Estimate ») sont interprétables par rapport à la variable expliquée. A titre d'exemple, la part des acheteurs ayant un chiffre d'affaires supérieur à 50k€ (« part_acheteur_sup_50 ») a un coefficient négatif, donc plus l'assuré aura dans son portefeuille des acheteurs dont le CA est supérieur à 50k€, plus sa fréquence de sinistre sera faible.

4.4.2 Modélisation du coût des sinistres

Pour cette deuxième partie de la modélisation, nous nous intéressons aux montants de sinistres. Le sinistre moyen est de 4 636 €.

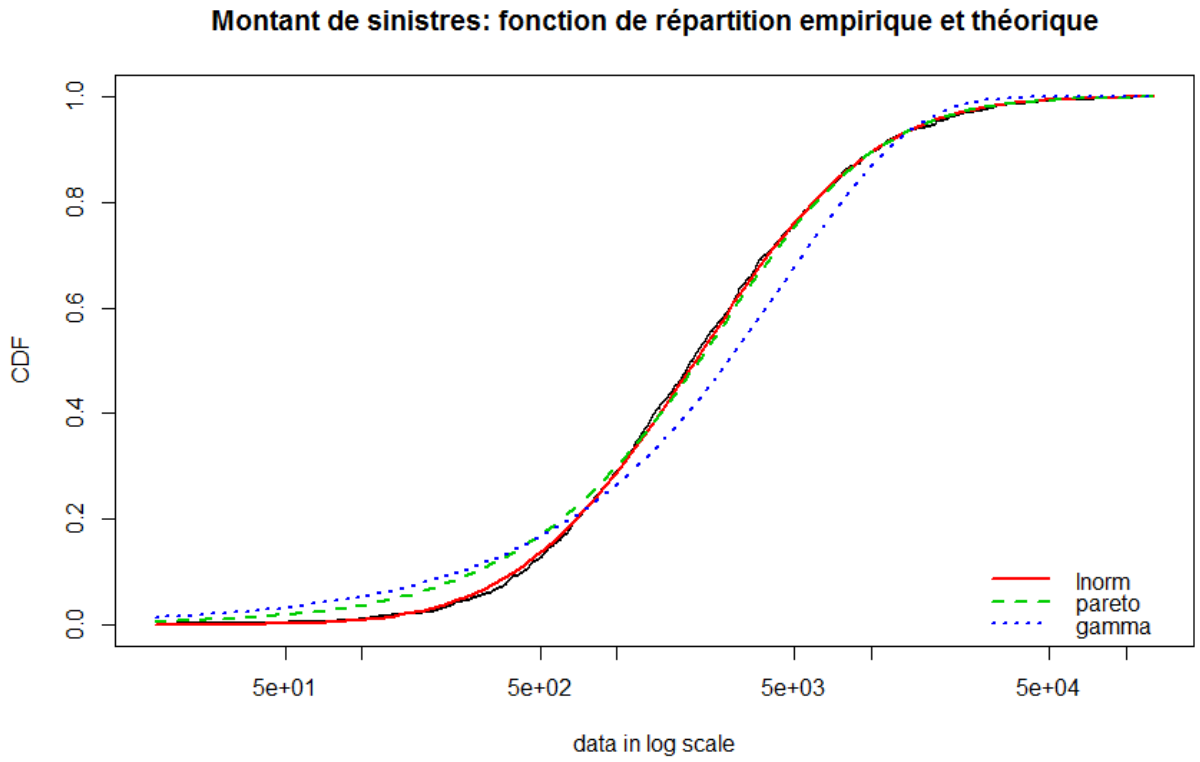
La première étape consiste à déterminer la loi qui convient le mieux à nos observations. Nous modélisons ensuite selon cette loi, en testant les différentes variables explicatives à notre disposition.

4.4.2.1 Détermination de la loi des sinistres

Nous disposons de 4 319 polices sur quatre ans. 1 564 polices ont subi un ou plusieurs sinistres sur la période.

Il nous faut d'abord déterminer la loi statistique des montants de sinistre. Nous estimons les paramètres de loi log normale, gamma et Pareto avec nos données. Nous traçons ensuite

la fonction de répartition théorique de ces trois lois que nous comparons avec la fonction de répartition empirique de nos données. Le résultat figure dans le graphique suivant.



La fonction de répartition empirique se confond avec la fonction de répartition théorique de la loi Log-Normale, c'est donc celle qui sera retenue pour calibrer le modèle.

4.4.2.2 Calibrage du GLM

Il est à noter que la loi Log-Normale ne faisant pas partie de la famille des lois exponentielles, il n'est à priori pas possible de calibrer un GLM Log-Normale. Cependant nous avons la relation suivante :

$$Y \sim \mathcal{LN}(\mu, \sigma^2) \Leftrightarrow X = log(Y) \sim \mathcal{N}(\mu, \sigma^2)$$

Avec μ la moyenne et σ l'écart-type

Nous pouvons donc calibrer un GLM gaussien sur le log de nos sinistres (qui est bien défini sur $\mathbb{R}$). Nous devons également prendre en compte, lorsque nous effectuons des prédictions, les relations suivantes :

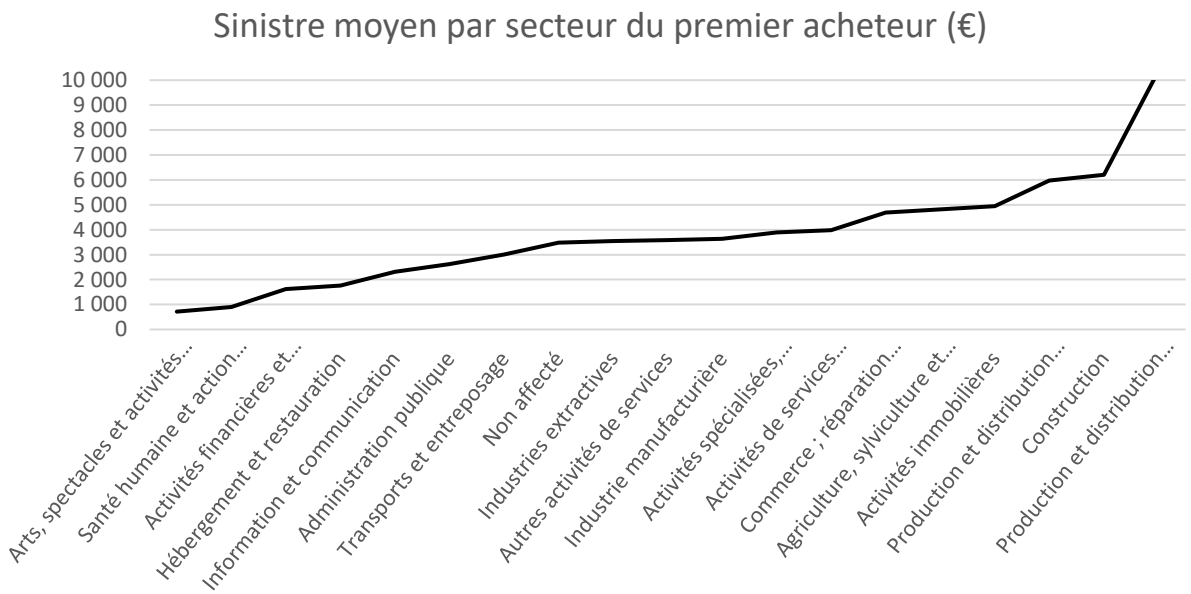
$$\mathbb{E}(Y) = \mathbb{E}(e^X) = e^{\mu + \frac{\sigma^2}{2}} \neq exp(\mathbb{E}[X]) = e^{\mu}$$

Nous devons donc ajouter le terme correctif $e^{\frac{\sigma^2}{2}}$ à notre e^{μ} prédit par le GLM.

Nous utilisons le critère de minimisation de *l'Akaïke* pour sélectionner les variables explicatives. Idem lorsque nous les modélisons ou regroupons. Les regroupements des variables catégorielles sont le plus souvent faites en regroupant celles qui sont proches en termes de montant de sinistre moyen constaté (notamment pour les regroupements des secteurs et des régions).

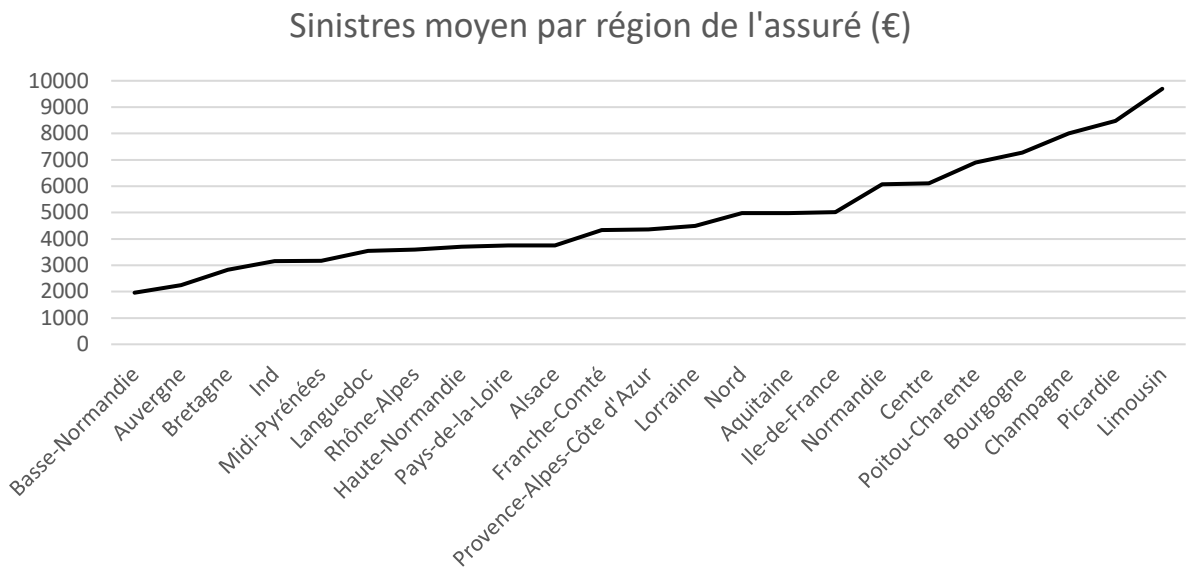
Nous procédons à l'analyse des variables explicatives et à leur segmentation. Là encore nous procédons à des regroupements qui peuvent être différents de ceux faits pour les fréquences de sinistres.

A titre d'exemple, nous avons ci-après le coût moyen d'un sinistre par secteur du premier acheteur :



Nous ne retrouvons pas la même hiérarchie que pour les fréquences. Les regroupements par classe seront donc différents.

Il en est de même pour l'analyse des régions de l'assuré :



Les paramètres du modèle sont résumés dans le tableau suivant :

	Estimate	Std. Error	t value	Pr(>|t|)	Significativité
(Intercept)	5,780	2,73E-01	2,11E+01	2,44E-83	***
secteur.ach1Commerce_rep_auto	1,2E-02	1,05E-01	1,13E-01	9,10E-01	
secteur.ach1Construction	0,35254	1,15E-01	3,07E+00	2,22E-03	**
secteur.ach1Manufacturier	-0,12641	1,11E-01	-1,14E+00	2,55E-01	
QG_agrement	0,00512	2,39E-03	2,15E+00	3,21E-02	*
QG_ND	-0,01258	1,22E-03	-1,03E+01	7,20E-24	***
limite_NDS	0,00005	1,26E-05	4,01E+00	6,61E-05	***
nb_NDS	-0,00011	4,78E-05	-2,30E+00	2,17E-02	*
log(moyenne_agrement_et_NDS)	0,55015	7,24E-02	7,60E+00	6,58E-14	***
ecart_type_agrement_et_NDS	-0,00395	1,71E-03	-2,31E+00	2,12E-02	*

Nous avons une cohérence des coefficients : le montant de sinistre est croissant avec le montant d'agréments et d'agréments express accordés, avec une clientèle liée au commerce et réparation automobile ou la construction.

Plus la limite de « non dénommés surveillés » (agréments express) est élevée, plus le sinistre moyen est élevé.

Le sinistre augmente avec la quotité garantie des agréments. La quotité garantie du « non dénommés » (ND) permet de prendre en compte le fait que l'assuré puisse bénéficier ou non d'autorisation en « non dénommés ». Il est logique alors que le sinistre moyen soit plus faible si l'assuré bénéficie de cette option, la limite de « non dénommés » étant généralement faible.

L'écart-type peut être interprété comme un indicateur de diversification du portefeuille d'acheteur.

4.4.3 Calcul de la prime pure

Un fois les deux modèles estimés, nous estimons la prime pure sur la base de validation (un tiers de nos données) et nous la comparons aux montants de sinistres constatés.

Pour chacun des assurés de la base de validation, nous estimons une fréquence de sinistre puis un montant moyen. Le produit des deux permet de calculer la prime pure par assuré. Nous observons également la somme sur l'ensemble des données que nous comparons.

Sur les données de validation, nous avons un montant total de sinistres de 5 062 404 €. Notre modélisation estime une prime pure de 6 433 788 €, soit un écart de 27% avec la sinistralité effective.

4.5 Produit Global KUP

Le produit Global KUP fait l'objet d'une grille tarifaire dépendant du chiffre d'affaires de l'assuré.

L'objet de la tarification va donc consister à calibrer un modèle fréquence / coût, comme pour le produit « Globale », mais en décomposant ensuite selon le chiffre d'affaires des assurés.

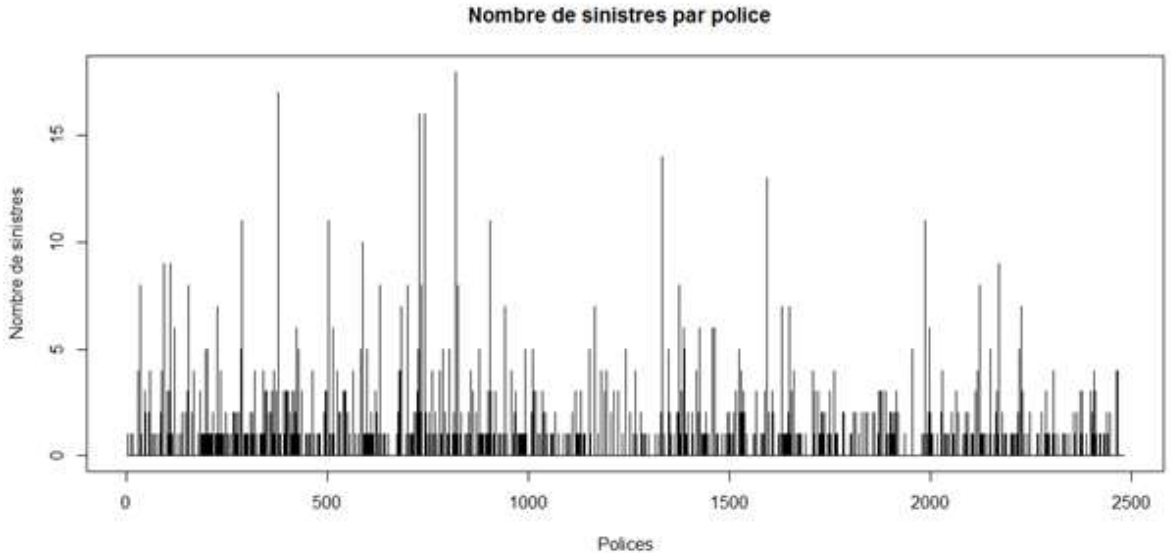
Nous étudions également la pertinence d'introduire d'autres variables explicatives, l'objectif étant de fournir une certaine flexibilité par rapport à la grille initiale dans une négociation commerciale. Nous calibrons donc notre modèle avec les variables disponibles, à l'instar de ce qui a été réalisé précédemment sur le produit « Globale ».

Nous disposons d'un portefeuille de 4 années polices (2012 à 2015) représentant 2 481 assurés. Sur la période, 1 375 sinistres ont été déclarés pour ce produit.

4.5.1 Modélisation de la fréquence des sinistres

4.5.1.1 Données

Nous disposons de 2 481 polices sur les quatre années d'observations. 659 assurés connaissent au moins un sinistre sur la période.

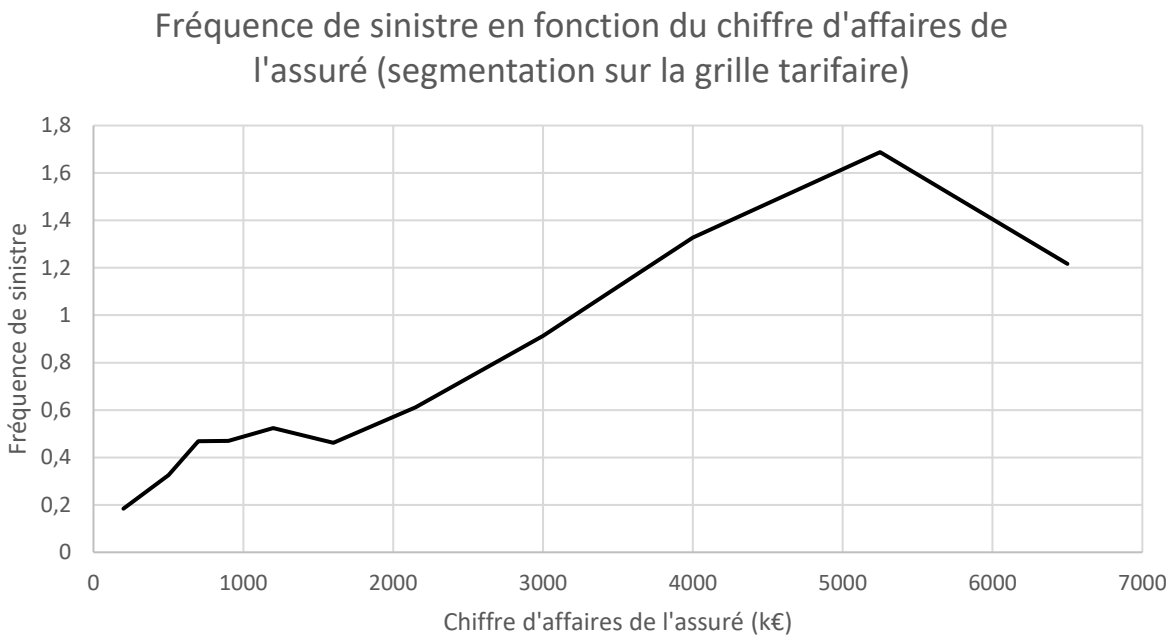


Nous observons d'abord les répartitions du nombre de sinistres selon les assurés

	Nombre de sinistres / base totale	Nombre de sinistres / assurés sinistrés
Minimum	0	1
Quantile 25%	0	1
Médiane	0	1
Moyenne	0,55	2,09
Quantile 75%	1	2
Quantile 90%	2	4
Quantile 95%	3	6
Quantile 99%	7	11
Max	18	18
Ecart type	1,43	2,15

La fréquence moyenne est faible et un nombre conséquent d'assurés n'ont pas de sinistres (78%).

Le produit Global Kup fonctionne par tranche de chiffre d'affaires de l'assuré. Nous observons donc notre portefeuille sur les tranches tarifaires de la grille existante.



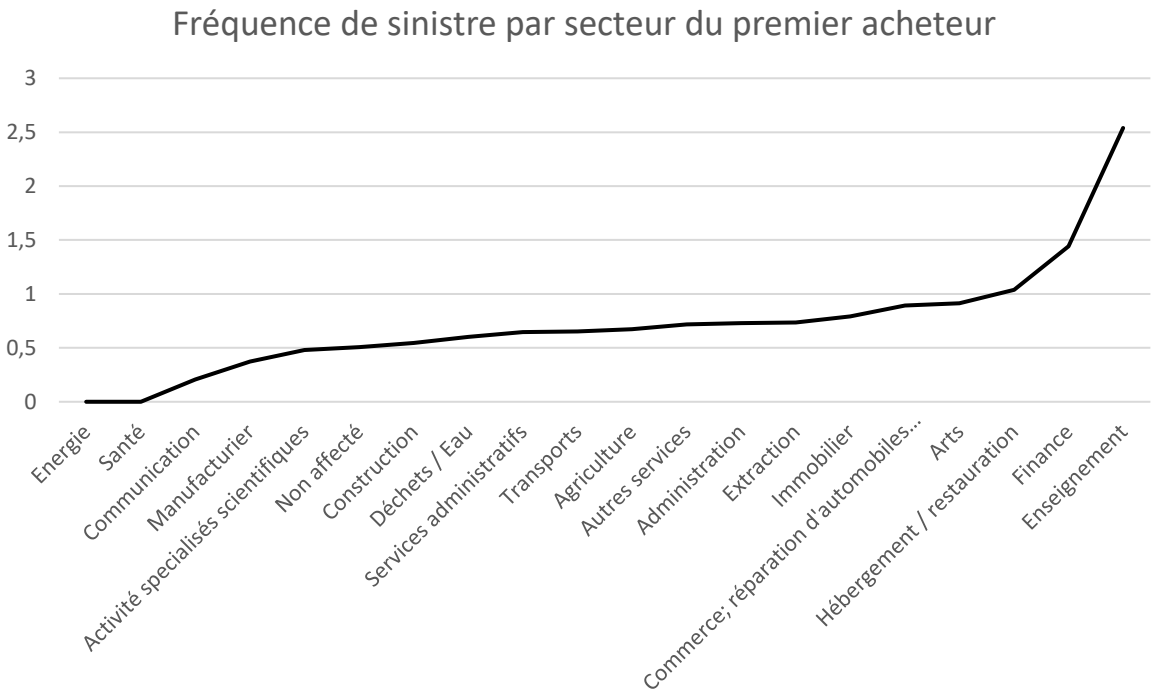
La relation est globalement croissante en fonction du CA pour les « petits CA », il est à noter cependant que les effectifs d'assurés sont concentrés sur les tranches petites et moyennes, comme le résume le tableau suivant :

Tranches de Chiffre d'affaires (k€)	Effectifs	Proportion
0 - 400	267	11%
400-600	231	9%
600-800	226	9%
800-1 000	224	9%
1 000-1 400	359	14%
1 400-1 800	317	13%
1 800-2 500	299	12%
2 500-3 500	253	10%
3 500-4 500	106	4%
4 500-6 000	105	4%
>6 000	94	4%

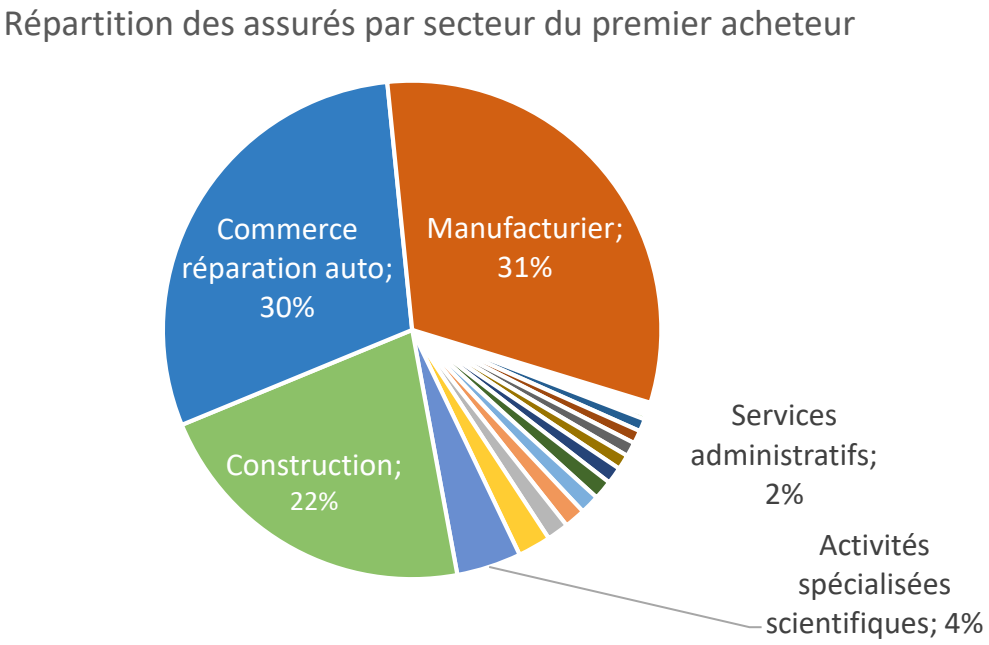
Cette répartition est conforme à l'objet de ce produit qui est de s'adresser à des petites entreprises.

Nous testons également d'autres variables explicatives que le chiffre d'affaires, de manière à proposer de la flexibilité aux commerciaux dans la négociation tarifaire.

Les secteurs du premier acheteur seront testés, celui-ci reflétant mieux le risque (qui est celui du défaut des acheteurs et non de l'assuré).

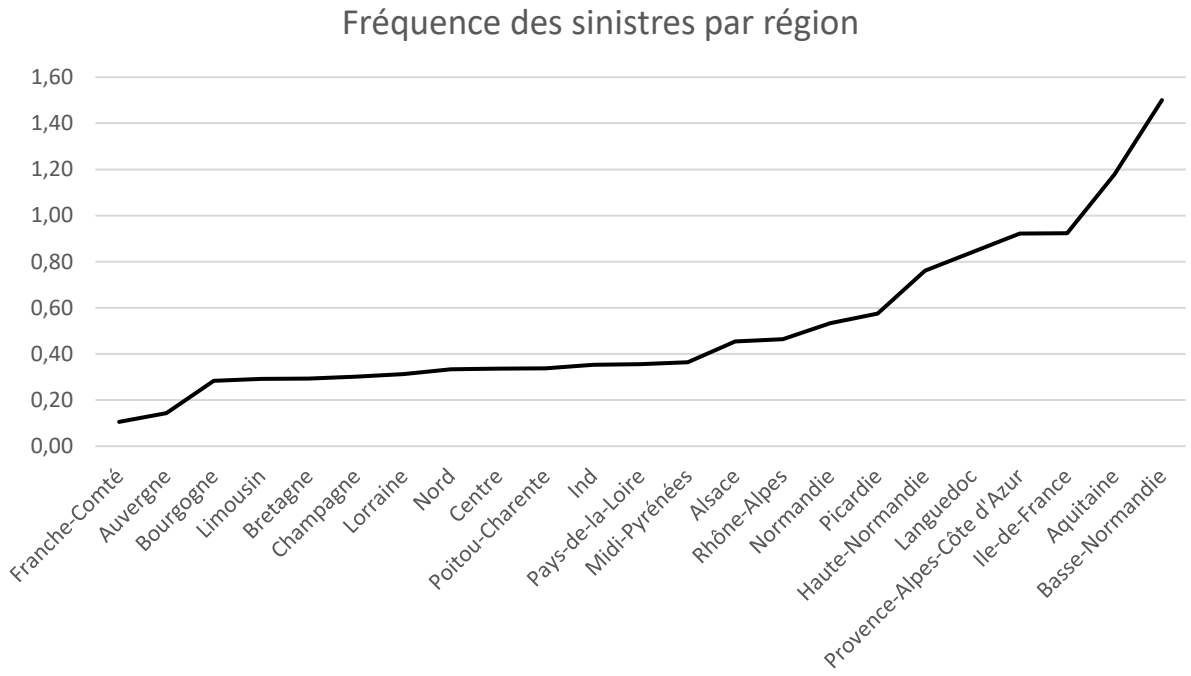


Il est à noter que le portefeuille d'assurés est fortement concentré sur trois secteurs qui représentent plus de 80% du portefeuille.



Nous regrouperons l'ensemble des autres secteurs pour ne distinguer que ces trois gros.

De même, nous observons les relations entre la fréquence des sinistres et l'appartenance des assurés aux différentes régions.



Nous n'avons pas le même phénomène de sur-représentation d'une région par rapport aux autres dans notre échantillon. Nous regroupons donc les régions, lorsque c'est nécessaire, selon leur fréquence de sinistres.

4.5.1.2 Calibrage du GLM

Nous conservons le chiffre d'affaires comme variable explicative en le gardant en variable numérique (nous n'imposons pas une distinction par classe de chiffre d'affaires correspondant aux intervalles de la grille tarifaire).

Nous testons donc la significativité des coefficients estimés ainsi que la fonction lien du GLM. Pour cette dernière, nous testons une loi de Poisson, quasi-Poisson, Tweedie et binomiale négative.

Les modèles ont tous été calibrés sur le même échantillon comprenant deux tiers de la base de données initiale. Les comparaisons de modèles se font avec le tiers de données restantes (base de validation). Nous faisons des prévisions de fréquence de sinistres avec la base de validation que nous comparons avec la réalisation. Le modèle retenu est celui qui minimise la somme du carré des erreurs (MSE) ainsi que l'erreur moyenne quadratique (RMSE).

Nos estimations nous conduisent à retenir une loi de Poisson sur-dispersée (quasi-Poisson).

Les coefficients de l'estimation par GLM figurent dans le tableau suivant :

	Estimate	Std. Error	t value	Pr(>|t|)	Significativité
(Intercept)	-5,834	6,83E-01	-8,54E+00	3,00E-17	***
log(Assure_CA_Global)	0,22189	8,43E-02	2,63E+00	8,54E-03	**
RégionAquitaine	0,95425	2,75E-01	3,47E+00	5,43E-04	***
RégionBasse-Normandie	0,83415	7,16E-01	1,16E+00	2,44E-01	
RégionR2	0,64705	2,31E-01	2,80E+00	5,10E-03	**
RégionIle-de-France	0,74115	1,89E-01	3,92E+00	9,34E-05	***
RégionProvence-Alpes-Côte d'Azur	0,86115	2,13E-01	4,05E+00	5,45E-05	***
RégionRhône-Alpes	0,33542	2,29E-01	1,47E+00	1,43E-01	
secteur.ach1Commerce_rep_auto	0,14856	1,80E-01	8,27E-01	4,08E-01	
secteur.ach1Construction	-0,01485	2,12E-01	-7,00E-02	9,44E-01	
secteur.ach1Manufacturier	-0,16212	2,09E-01	-7,75E-01	4,38E-01	
score_mini_10_acheteurs	-0,00918	5,42E-02	-1,69E-01	8,66E-01	
part_acheteur_sup_50	-8,66654	2,34E+00	-3,71E+00	2,14E-04	***
log(total_agrement_et_NDS)	0,46535	5,72E-02	8,13E+00	7,95E-16	***

Avec :

- Assure_CA_Global : le chiffre d'affaires total de l'assuré
- Région : la région de l'assuré
- Secteur.ach1 : le secteur du plus gros acheteur de l'assuré
- Score_mini_10_acheteurs : le score le plus faible parmi les 10 premiers acheteurs de l'assuré.
- Part_acheteur_sup_50 : la part d'acheteurs de l'assuré dont l'encours est supérieur à 50k€
- Total_agrement_et_NDS : le montant cumulé des agréments et agréments express.

Et avec les regroupements suivants pour les régions :

- R1 : Alsace, Auvergne, Bourgogne, Bretagne, Centre, Champagne, Franche-Comté, Limousin, Lorraine, Midi-Pyrénées, Nord, Pays-de-la-Loire, Poitou-Charentes.
- R2 : Haute-Normandie, Normandie et Picardie.

Il est à noter qu'il y n'y a pas égalité entre « Total_agrément_et_NDS » et le chiffre d'affaires. En effet, nous avons la différence entre la limite de l'agrément et l'exposition réelle ainsi que les agréments demandés qui ne sont que des interrogations de la base qui entrent en jeu (le « use factor » précédemment mentionné).

Nous constatons donc une croissance de la sinistralité avec la taille de l'assuré ainsi qu'avec le montant d'agréments et agréments express demandés.

Le coefficient du score du plus mauvais acheteur est négatif, indiquant que lorsque la qualité du portefeuille s'améliore, la fréquence de sinistres baisse.

De même, plus les acheteurs de l'assuré ont un chiffre d'affaires élevé, plus la fréquence de sinistres baisse.

Nous retrouvons des coefficients logiques par rapport à la sinistralité constatée au niveau des secteurs et des régions.

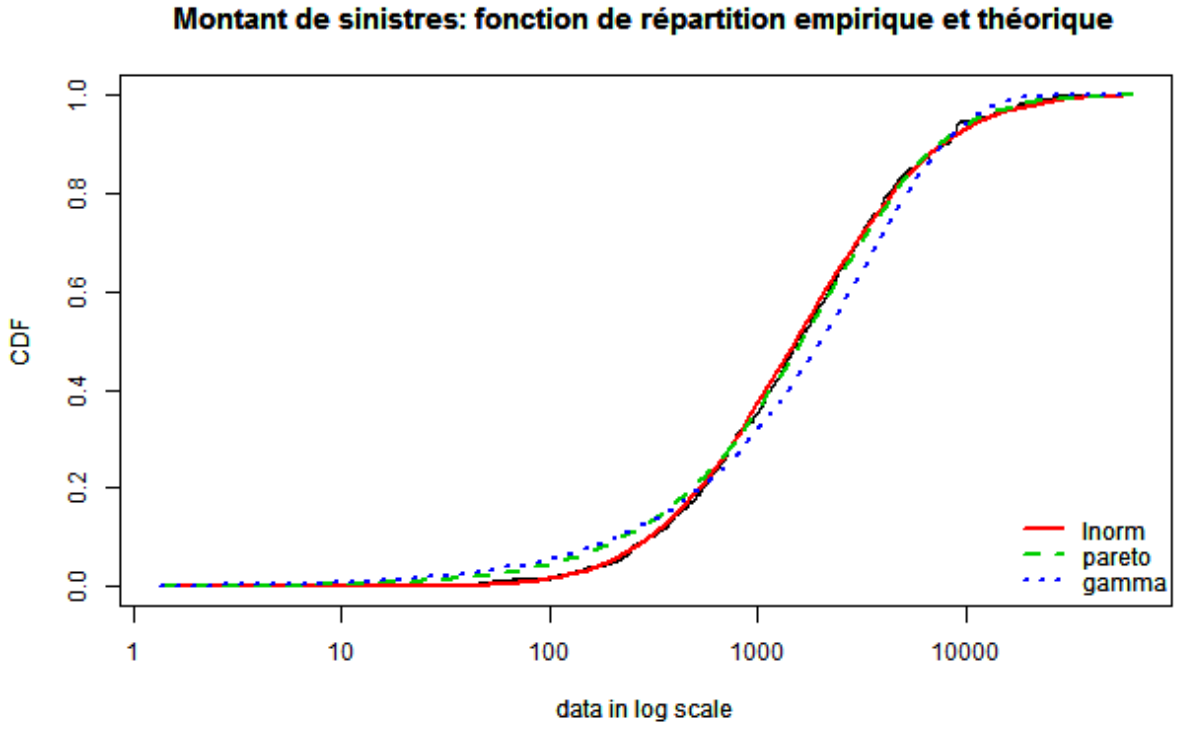
4.5.2 Modélisation du coût des sinistres

4.5.2.1 Données

Nous disposons de 659 sinistres sur notre échantillon. Le sinistre moyen est de 3 234 €.

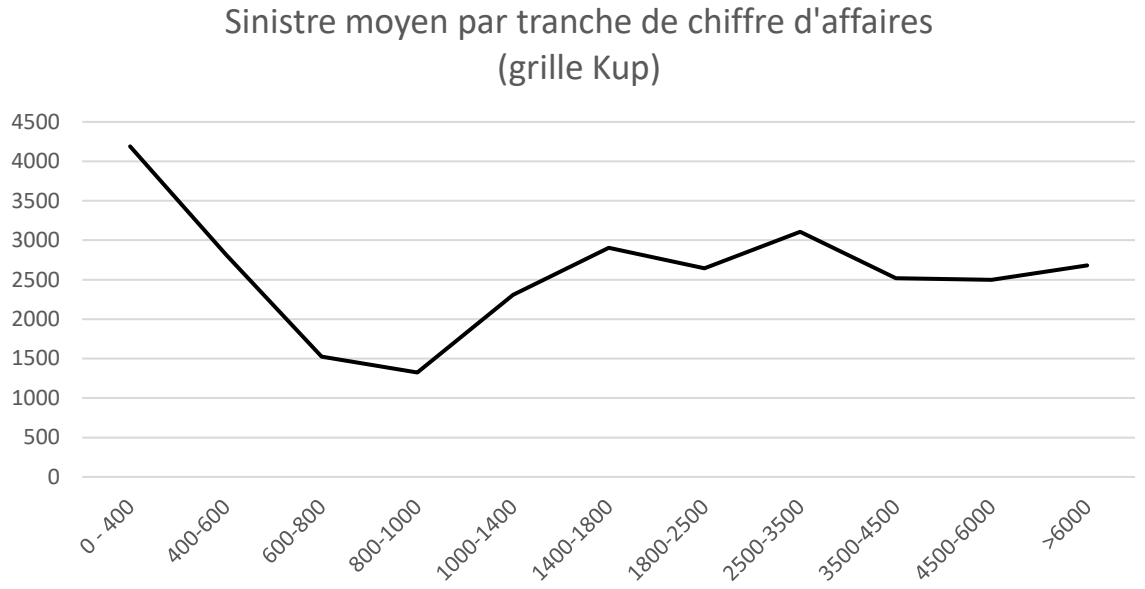
Nous nous heurtons donc à un manque de données qui nous oblige à regrouper le plus possible les modalités de certaines variables qualitatives (les régions ou les secteurs notamment).

Nous cherchons d'abord à déterminer la loi des montants de sinistres. Pour cela, comme précédemment, nous estimons les paramètres selon plusieurs lois et nous comparons les fonctions de répartition théorique et empirique. Le résultat figure dans le graphique suivant :

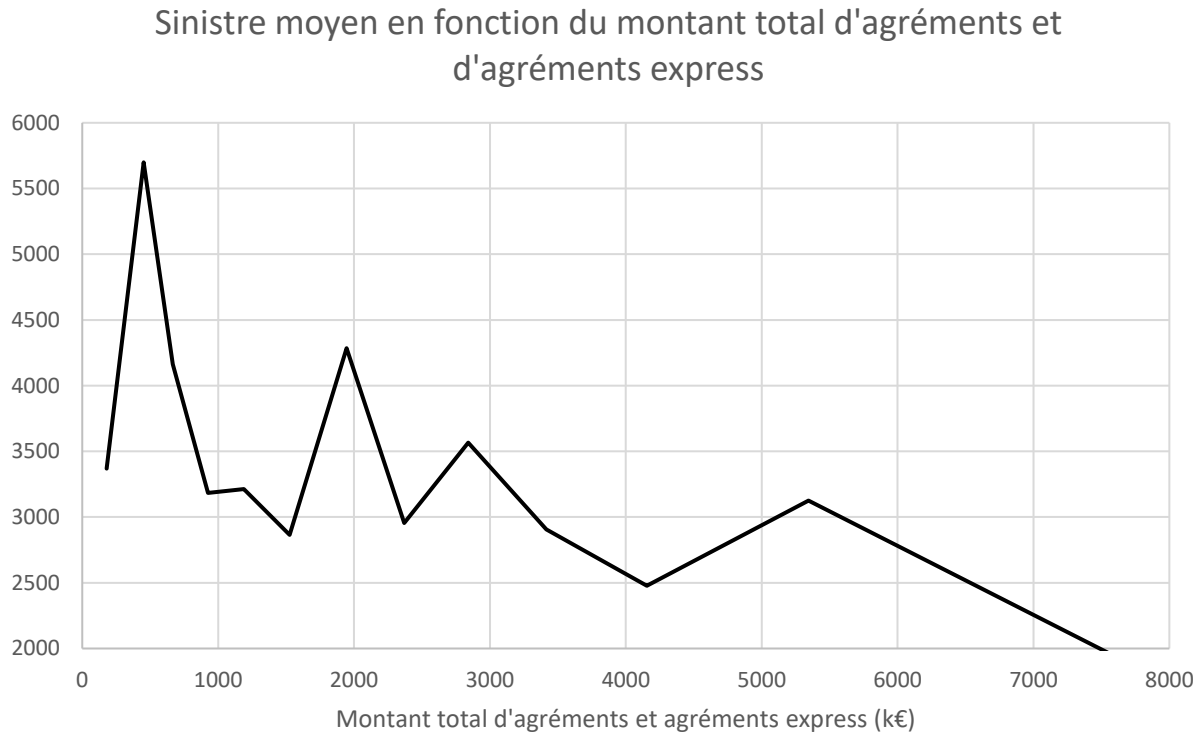


Nous retenons la loi Log Normale. La méthode d'estimation sera la même que pour la police Globale.

Nous procédons à l'analyse des variables explicatives et leur segmentation. Tout d'abord la répartition par tranche de chiffre d'affaires se révèle difficile à interpréter car elle ne fait pas ressortir de relation claire. Cette variable ne sera donc pas retenue pour modéliser le coût.



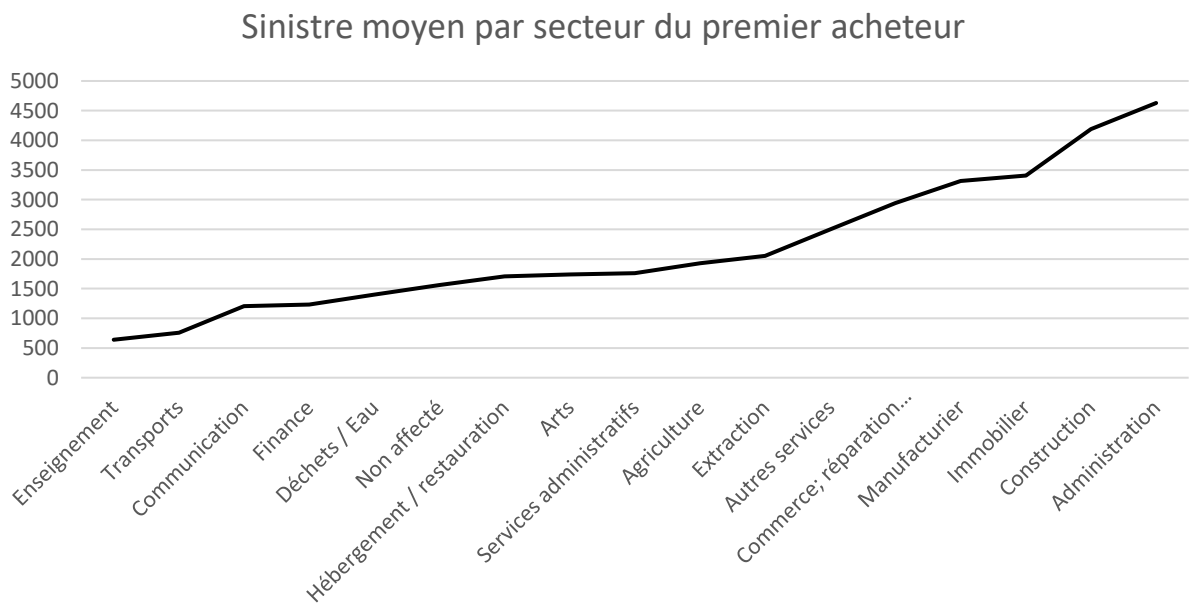
En revanche, le montant total des agréments et agréments express fait ressortir une relation plus interprétable, même si elle demeure erratique.



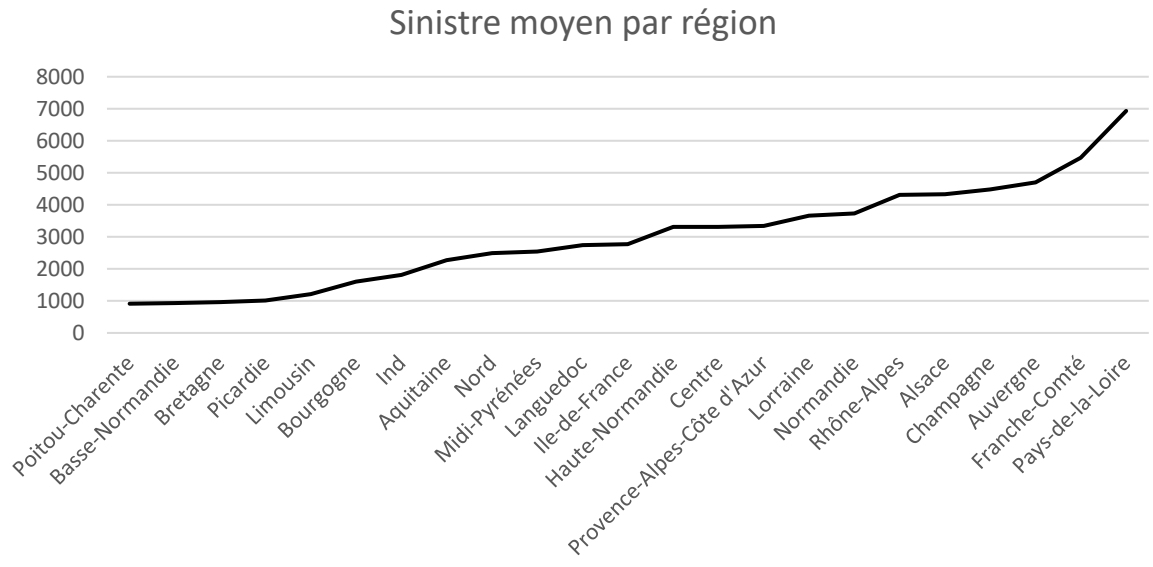
C'est cette variable que nous retiendrons. Il est normal d'avoir une différence entre le chiffre d'affaires de l'assuré et le nombre total d'agrément et agrément express. Le « use

factor » entre en effet en jeu entre les deux (soit la limite est potentiellement plus importante que le montant facturé sous-jacent, soit l'assuré entre beaucoup de demandes en base pour tester ses prospects).

Par secteur du premier acheteur, nous constatons que 83% des sinistres et 89% du montant de sinistres sont concentrés sur seulement trois secteurs (Commerce et réparation automobile, Construction et secteur manufacturier). Nous regroupons l'ensemble des autres secteurs pour ne distinguer que ces trois dans nos estimations.



Nous procédons également à des regroupements par région en fonction des plus proches montants de sinistre moyen.



4.5.2.2 Calibrage du GLM

Nous testons les différentes variables explicatives que nous avons en base.

Nos critères de sélection de modèle sont la minimisation de l'*Akaïké* et le test de significativité du T de Student.

Le modèle final figure ci-après :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8,014	3,83E-01	2,09E+01	7,08E-69	***
secteur.ach1Commerce_rep_auto	0,16584	1,68E-01	9,87E-01	3,24E-01	
secteur.ach1Construction	0,62967	1,85E-01	3,41E+00	7,19E-04	***
secteur.ach1Manufacturier	0,29448	1,78E-01	1,65E+00	9,93E-02	.
log(total_agrement_et_NDS)	-0,13721	4,73E-02	-2,90E+00	3,86E-03	**
part_acheteur_sup_50	5,21671	1,10E+00	4,75E+00	2,74E-06	***

Nous conservons les trois secteurs fortement représentés dans notre échantillon, les coefficients estimés étant cohérents avec l'analyse faite. Les régions ne s'avèrent pas discriminantes lors de nos estimations.

Plus le montant d'agréments et agréments express est élevé, moins le montant de sinistres est élevé. Cela peut s'interpréter par le fait que plus un assuré a demandé d'agréments et d'agréments express, plus le montant facturé à chacun de ses acheteurs est faible alors que la limite est au minimum 5 000 € pour les agréments express.

En revanche, la part des acheteurs dont le chiffre d'affaires est supérieur à 50k€ joue un rôle positif sur le coût de sinistres. Les assurés vont avoir tendance à traiter des montants plus importants avec les grandes entreprises qu'avec les petites.

4.5.3 Calcul de la prime pure

Tout comme pour la police Globale, nous appliquons notre modèle sur la base de validation pour calculer une prime pure et la comparer avec les sinistres constatés.

Pour chacun des assurés de la base de validation, nous estimons une fréquence de sinistre puis un montant moyen. Le produit des deux permet de calculer la prime pure par assuré. Nous observons également la somme sur l'ensemble des données que nous comparons.

Sur les données de validation, nous avons un montant total de sinistre de 1 069 631€. Notre modélisation estime une prime pure pour l'ensemble des polices de 1 294 067 €, soit un écart de 21% avec la sinistralité effective. Le modèle sur-estime donc le montant des sinistres et conduit à une sur-tarification dans les mêmes proportions.

4.6 Prime nette.

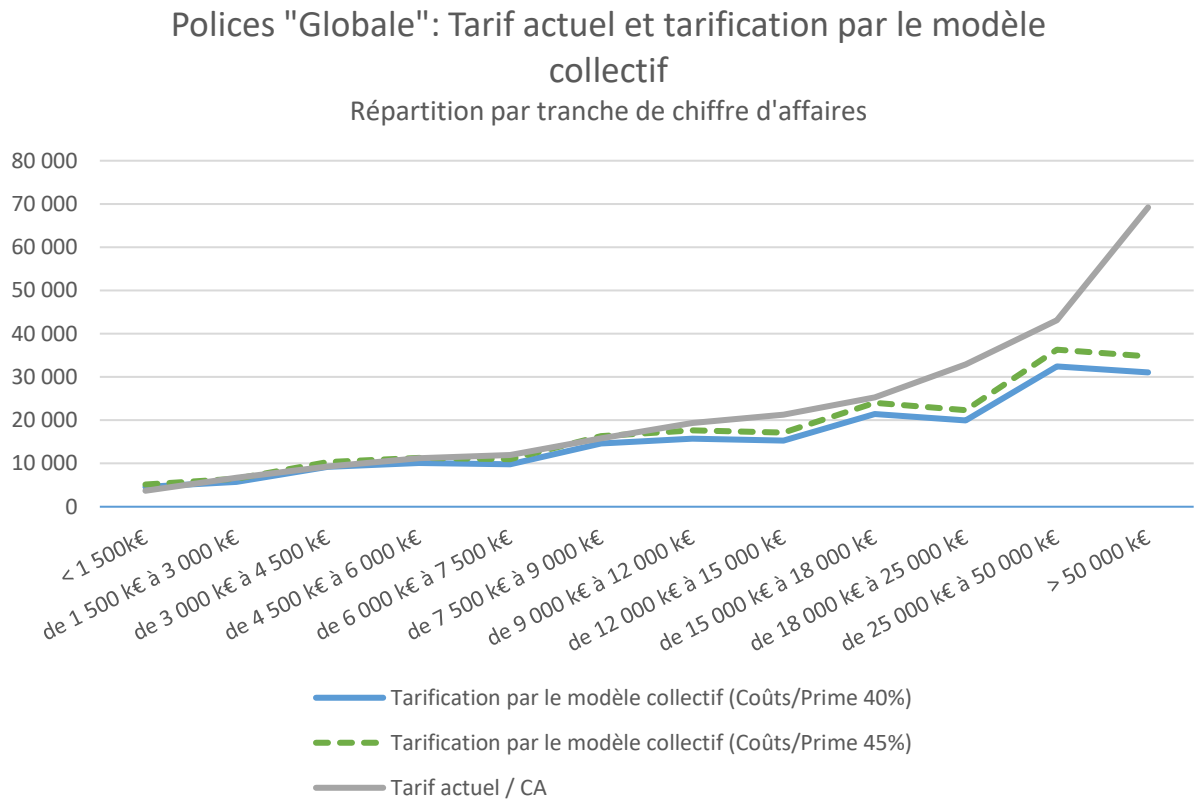
La prime pure n'est pas suffisante car elle ne tient pas compte du risque. La théorie de la ruine démontre en effet que si c'est ce principe de prime qui est retenu, il conduit la compagnie d'assurance à la ruine de façon quasi-certaine (Baradel 2016).

Le passage de la prime pure à la prime nette va donc consister, notamment, à déterminer un coefficient de chargement qui tienne compte du risque.

Nous utilisons la même méthode que celle mise en place sur la tarification par les probabilités de défaut (décrite au paragraphe 3.2.4)

4.6.1 Polices « Globale »

A l'instar de ce que nous avons présenté sur le modèle de tarification par les probabilités de défaut, nous classons notre tarification par tranche de chiffre d'affaires, en retenant les mêmes que précédemment.



La tarification par le modèle collectif s'éloigne de la tarification actuelle sur les tranches les plus élevées, elle se révèle en effet plus faible que le tarif actuel.

Tranches de chiffre d'affaires de l'assuré	Tarif actuel	Tarification par le modèle collectif (Coûts/Prime 40%)	Tarification par le modèle collectif (Coûts/Prime 45%)	Effectif en pourcentage de l'effectif total
< 1 500k€	3 694	4 591	5 140	15%
de 1 500 k€ à 3 000 k€	6 671	5 775	6 466	26%
de 3 000 k€ à 4 500 k€	9 284	9 227	10 330	17%
de 4 500 k€ à 6 000 k€	11 238	10 095	11 302	10%
de 6 000 k€ à 7 500 k€	11 954	9 805	10 978	6%
de 7 500 k€ à 9 000 k€	15 807	14 610	16 356	6%
de 9 000 k€ à 12 000 k€	19 386	15 740	17 622	6%
de 12 000 k€ à 15 000 k€	21 302	15 325	17 158	3%
de 15 000 k€ à 18 000 k€	25 312	21 461	24 027	2%
de 18 000 k€ à 25 000 k€	32 901	19 938	22 322	4%
de 25 000 k€ à 50 000 k€	43 151	32 435	36 313	2%
> 50 000 k€	69 211	31 036	34 747	2%

La tranche la plus élevée regroupe 16 polices sur les 744 (soit 2.2%) mais près de 10% des primes totales.

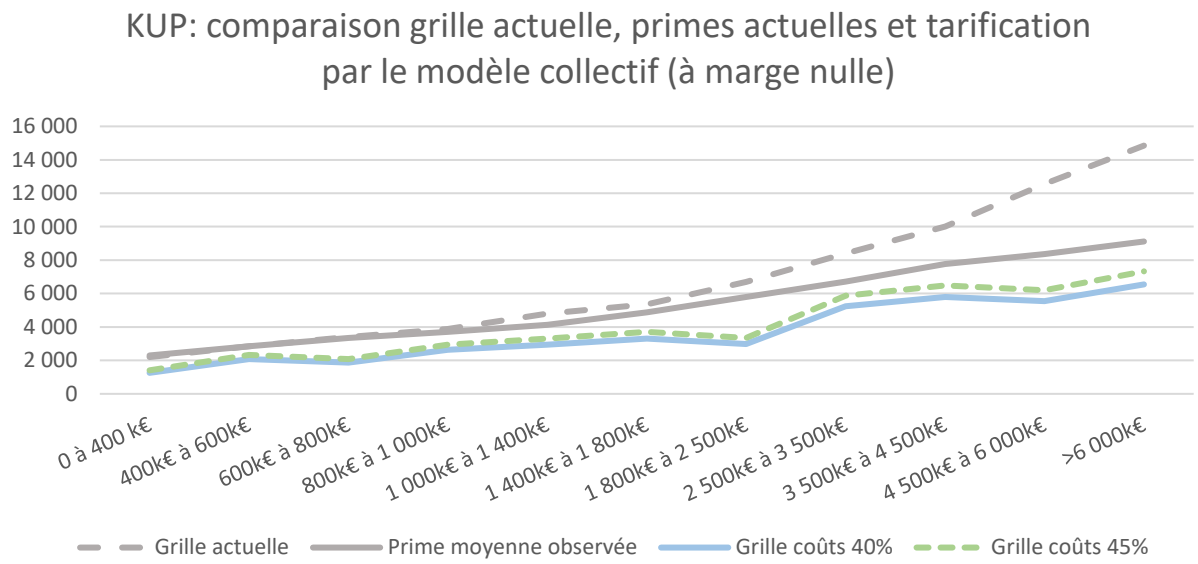
Le modèle collectif que nous avons ainsi calibré laisse entrevoir, soit une marge de manœuvre sur les sociétés ayant un chiffre d'affaires élevé, soit la nécessité de l'améliorer sur ces tranches de CA.

4.6.2 Polices « KUP »

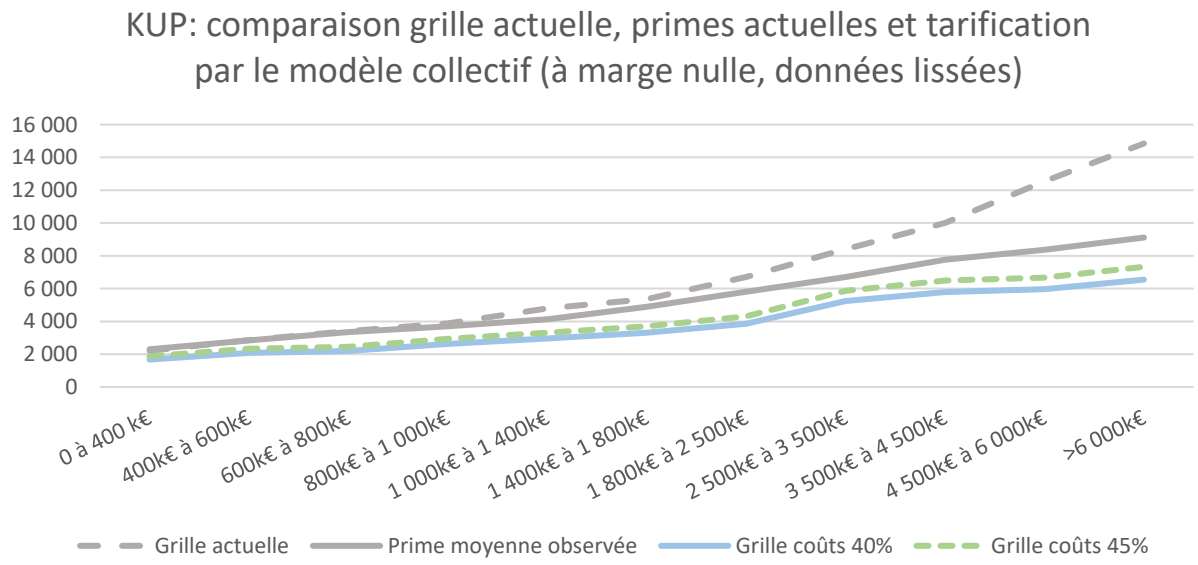
Comme avec le premier modèle, nous ré-estimons le produit KUP selon la grille actuelle et la nouvelle grille plus fine qu'AXA Assurcrédit souhaiterait mettre en place, en appliquant le modèle que nous avons préalablement calibré.

Nous tarifons les polices avec une structure de coûts à 40% des primes et une seconde à 45% des primes. Nous n'incluons pas de marge pour l'assureur.

Sur la grille actuelle, nous obtenons la tarification suivante :



Afin d'avoir un comportement moins erratique de la prime selon la tranche de chiffre d'affaires, nous lissons les primes par tranche (moyenne mobile centrée sur la tranche actuelle) et nous imposons également que le tarif de la tranche supérieure soit supérieur ou égal à celui de la tranche précédente. Cela nous permet d'obtenir la tarification du graphe ci-après :

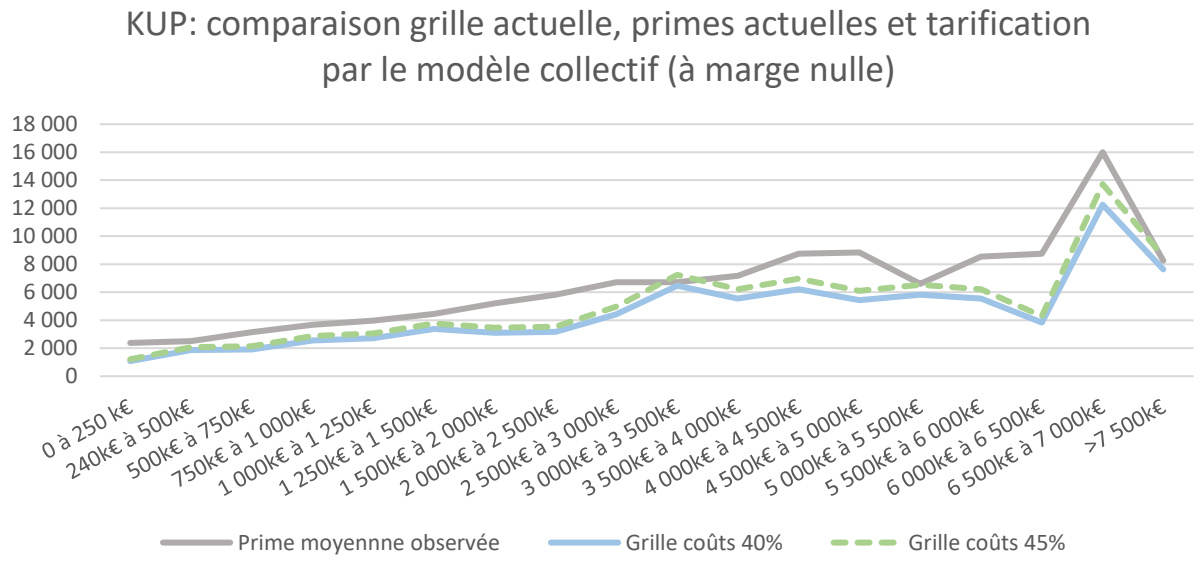


Soit la grille suivante :

Tranches de chiffre d'affaires	Grille actuelle	Prime moyenne observée	Prime calculée		Grille avec lissage		Effectif en pourcentage de l'effectif total
			Grille coûts 40%	Grille coûts 45%	Grille coûts 40%	Grille coûts 45%	
0 à 400 k€	2 200	2 297	1 257	1 407	1 670	1 869	6%
400k€ à 600k€	2 850	2 842	2 083	2 332	2 083	2 332	8%
600k€ à 800k€	3 400	3 345	1 862	2 084	2 190	2 452	5%
800k€ à 1 000k€	3 880	3 707	2 626	2 939	2 626	2 939	8%
1 000k€ à 1 400k€	4 800	4 134	2 946	3 298	2 960	3 314	14%
1 400k€ à 1 800k€	5 350	4 884	3 308	3 704	3 308	3 704	17%
1 800k€ à 2 500k€	6 700	5 801	2 981	3 338	3 845	4 305	11%
2 500k€ à 3 500k€	8 400	6 710	5 245	5 872	5 245	5 872	13%
3 500k€ à 4 500k€	10 000	7 762	5 794	6 486	5 794	6 486	5%
4 500k€ à 6 000k€	12 550	8 359	5 541	6 204	5 960	6 673	6%
>6 000k€	14 850	9 114	6 546	7 329	6 546	7 329	7%

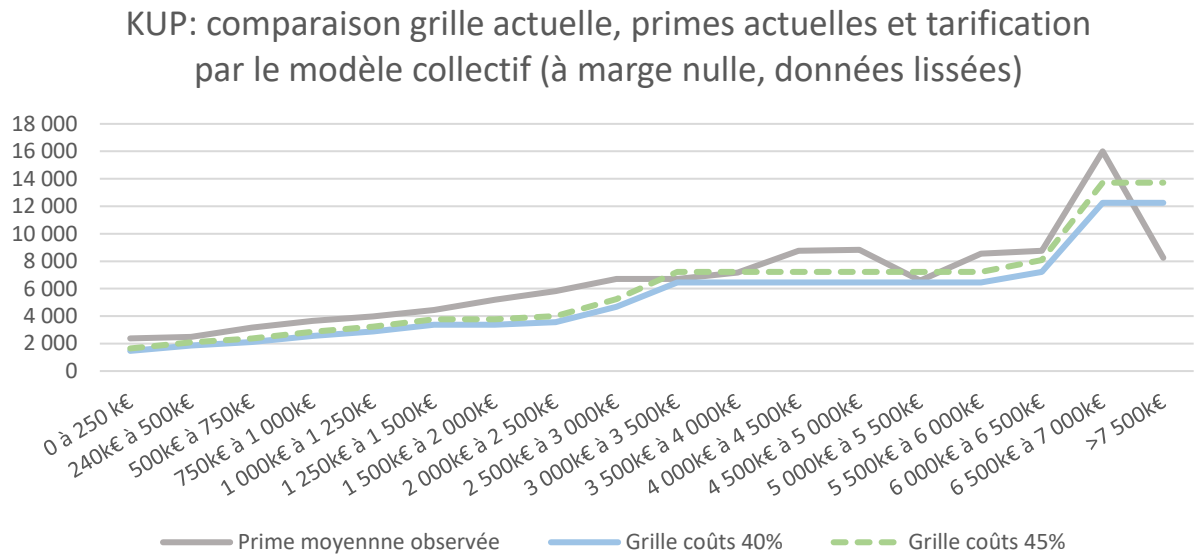
Cette tarification par le modèle collectif permet d'obtenir une grille qui est plus en phase avec les prix actuellement pratiqués. Il est à noter que ces prix ne prennent pas en compte de marge pour l'assureur.

Nous calibrons également une tarification selon la nouvelle grille plus fine qu'AXA Assurcrédit souhaite étudier avec les contrats existants :



Le comportement est nettement plus erratique sur les tranches élevées de chiffre d'affaires car moins de contrats sont représentés. Le fait de lisser et d'imposer un tarif croissant

lorsque nous passons dans une tranche plus élevée permet de corriger ce problème sur les dernières tranches :



Ainsi nous avons la grille suivante, correspondant aux nouvelles tranches de chiffre d'affaires :

		Données brutes		Grille avec lissage		
Tranches de chiffre d'affaires	Prime moyenne observée	Grille coûts 40%	Grille coûts 45%	Grille coûts 40%	Grille coûts 45%	Effectif en pourcentage de l'effectif total
0 à 250 k€	2 377	1 081	1 211	1 471	1 646	3%
240k€ à 500k€	2 498	1 860	2 082	1 860	2 082	7%
500k€ à 750k€	3 154	1 919	2 148	2 113	2 366	9%
750k€ à 1 000k€	3 662	2 560	2 866	2 560	2 866	9%
1 000k€ à 1 250k€	3 972	2 723	3 048	2 883	3 228	9%
1 250k€ à 1 500k€	4 448	3 367	3 769	3 367	3 769	10%
1 500k€ à 2 000k€	5 197	3 104	3 475	3 367	3 769	14%
2 000k€ à 2 500k€	5 827	3 167	3 546	3 569	3 996	9%
2 500k€ à 3 000k€	6 709	4 436	4 967	4 687	5 247	8%
3 000k€ à 3 500k€	6 713	6 458	7 230	6 458	7 230	5%
3 500k€ à 4 000k€	7 164	5 537	6 200	6 458	7 230	3%
4 000k€ à 4 500k€	8 758	6 221	6 965	6 458	7 230	2%
4 500k€ à 5 000k€	8 833	5 437	6 086	6 458	7 230	2%
5 000k€ à 5 500k€	6 600	5 830	6 527	6 458	7 230	1%
5 500k€ à 6 000k€	8 546	5 538	6 201	6 458	7 230	2%
6 000k€ à 6 500k€	8 753	3 840	4 299	7 211	8 073	3%
6 500k€ à 7 000k€	16 000	12 253	13 718	12 253	13 718	1%
>7 500k€	8 236	7 625	8 536	12 253	13 718	4%

5 Synthèse et comparaison des deux modèles de tarification

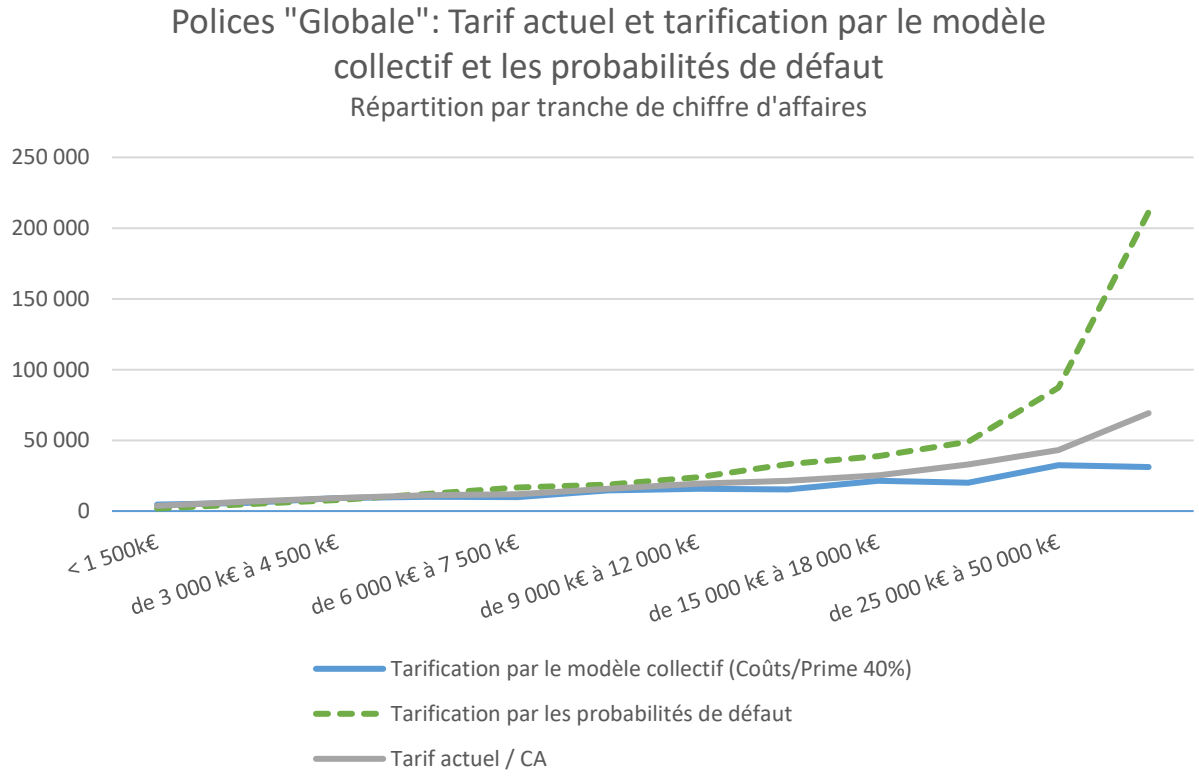
5.1 Polices « Globale »

Nous avons calibré deux modèles de tarification sur les polices « Globale » qui permettent de recalculer un tarif pour chacune d'entre elles.

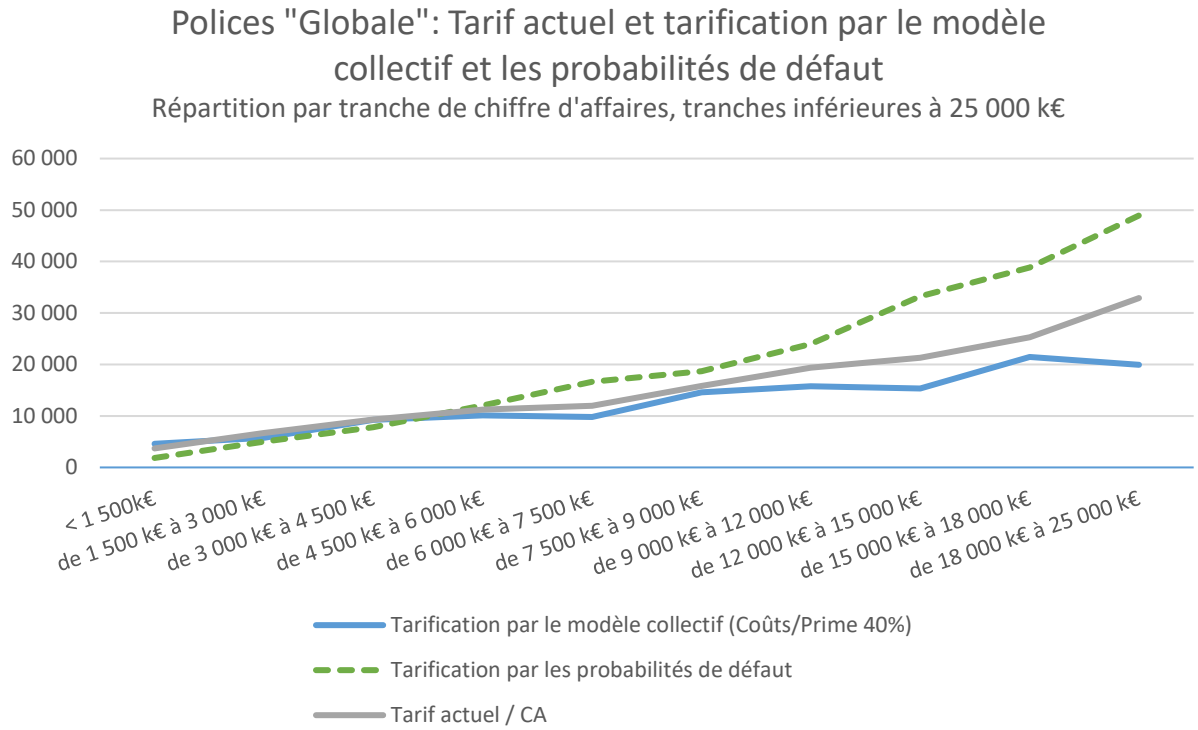
Le modèle collectif présente l'avantage de ne pas nécessiter l'intégralité de l'information des acheteurs, contrairement au modèle basé sur les probabilités de défaut, qui lui, intègre la totalité de l'information sur tous les acheteurs en portefeuille.

Le modèle collectif permet ainsi en particulier d'être directement applicable pour la tarification des dossiers prospects pour lesquels nous n'avons pas l'information de l'ensemble des acheteurs à garantir.

Nous avons ainsi tarifé les polices du portefeuille « Globale », comme le montre le graphique suivant :



Nous enlevons les deux tranches les plus élevées pour observer le comportement des deux modèles de tarification :

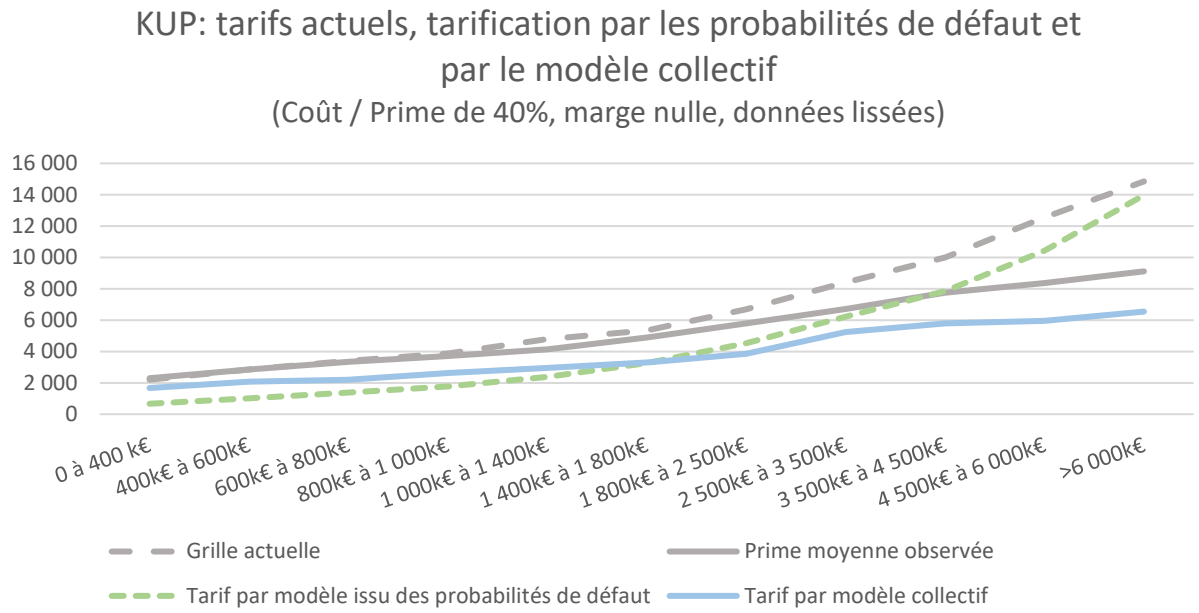


Nous obtenons, avec le modèle collectif, un tarif au-dessous de la tarification actuelle pour les tranches élevées de chiffre d'affaires et au-dessus pour les tranches les plus faibles. La tarification par les probabilités de défaut est en sens inverse, mais produit des tarifs nettement plus élevés sur les tranches les plus hautes.

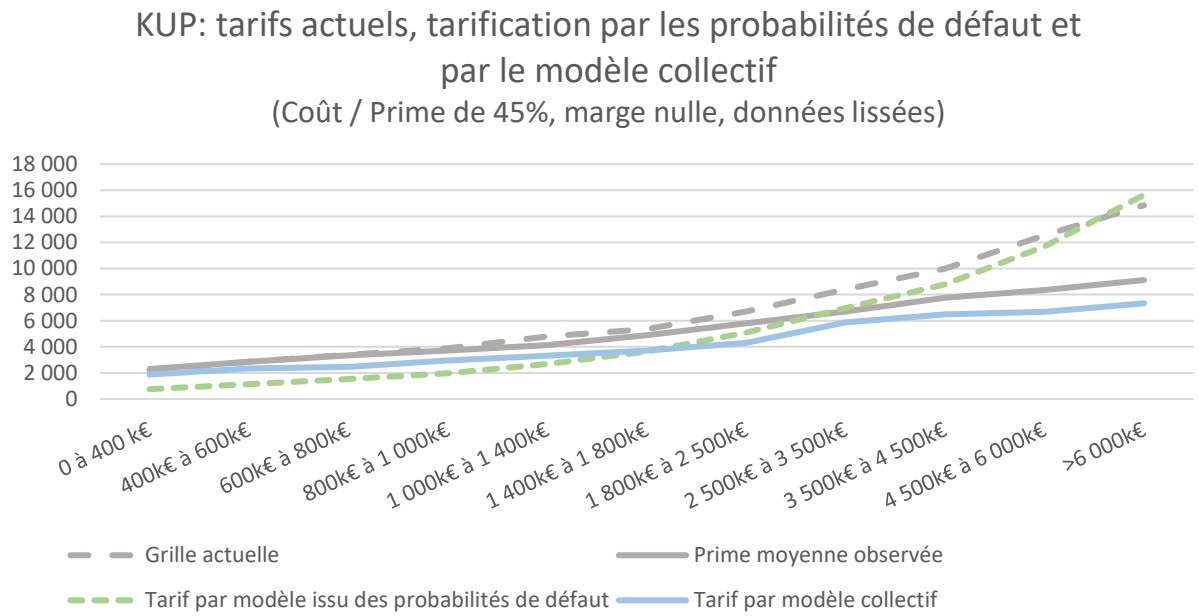
L'outil de tarification par le modèle collectif diverge nettement moins de la tarification actuelle que celui par les probabilités de défaut.

5.2 KUP

Nous avons calibré deux modèles de tarification que nous avons utilisés pour tarifer les polices « KUP ». Cela nous permet de comparer la tarification par les deux méthodes par rapport à l'existant, comme le montre le graphique suivant.



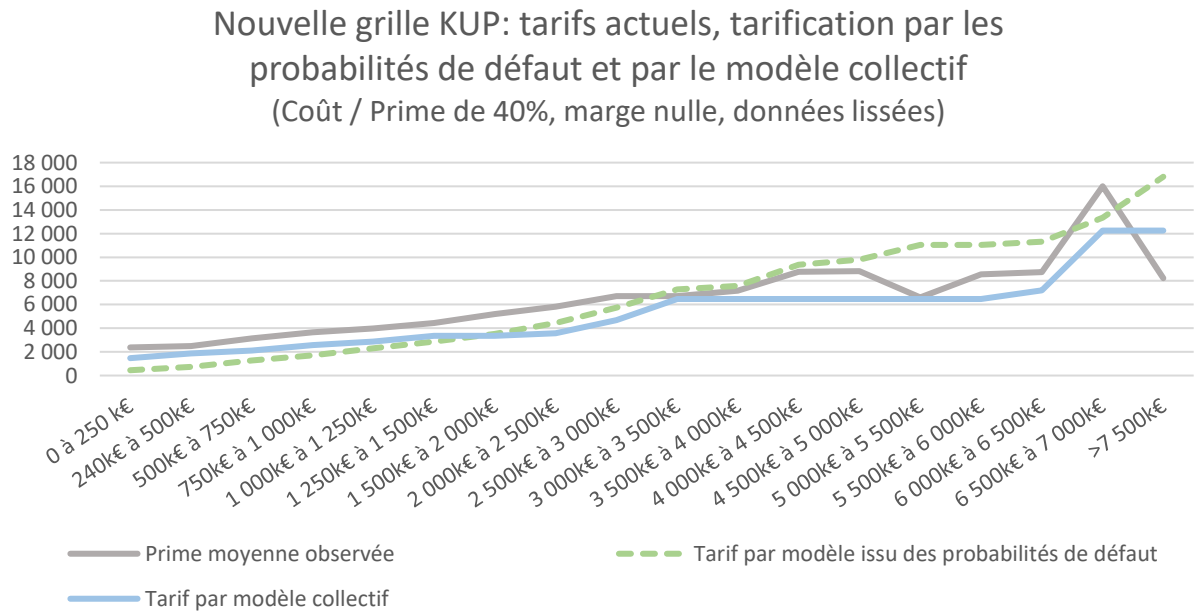
La structure de Coûts/Prime à 40% est la cible de la société, ce ratio est plutôt aujourd'hui autour de 45%, ce qui modifie la tarification comme indiqué dans le graphique suivant.



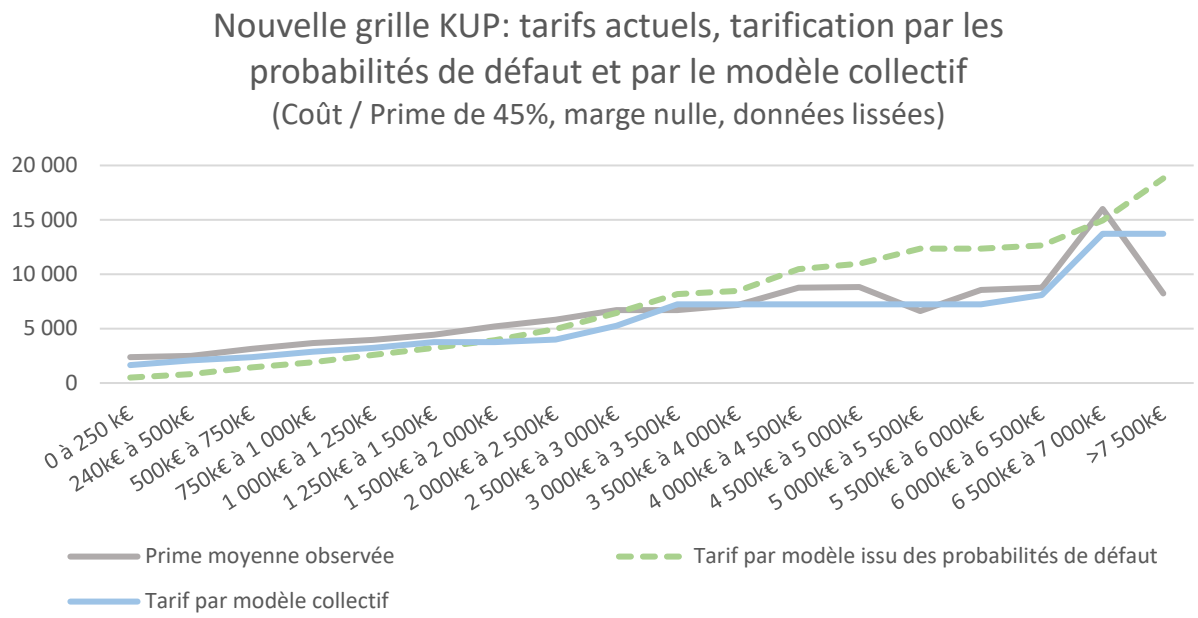
La grille correspondant aux données lissées est alors la suivante :

Tranches de chiffre d'affaires	Grille actuelle	Prime moyenne observée	Coûts 40% de la prime		Coûts 45% de la prime	
			Tarif par modèle issu des probabilités de défaut	Tarif par modèle collectif	Tarif par modèle issu des probabilités de défaut	Tarif par modèle collectif
0 à 400 k€	2 200	2 297	667	1 670	747	1 869
400k€ à 600k€	2 850	2 842	1 003	2 083	1 123	2 332
600k€ à 800k€	3 400	3 345	1 372	2 190	1 536	2 452
800k€ à 1 000k€	3 880	3 707	1 766	2 626	1 977	2 939
1 000k€ à 1 400k€	4 800	4 134	2 389	2 960	2 674	3 314
1 400k€ à 1 800k€	5 350	4 884	3 256	3 308	3 645	3 704
1 800k€ à 2 500k€	6 700	5 801	4 536	3 845	5 078	4 305
2 500k€ à 3 500k€	8 400	6 710	6 228	5 245	6 973	5 872
3 500k€ à 4 500k€	10 000	7 762	7 854	5 794	8 793	6 486
4 500k€ à 6000k€	12 550	8 359	10 435	5 960	11 682	6 673
>6 000k€	14 850	9 114	13 970	6 546	15 640	7 329

De la même façon, nous calibrons la tarification sur la nouvelle grille :



Ces tarifs sont également calculables avec un ratio Coûts/Prime de 45%



Soit les grilles suivantes :

		Coûts 40% de la prime		Coûts 45% de la prime	
Tranches de chiffre d'affaires	Prime moyenne observée	Tarif par modèle issu des probabilités de défaut	Tarif par modèle collectif	Tarif par modèle issu des probabilités de défaut	Tarif par modèle collectif
0 à 250 k€	2 377	447	1 471	500	1 646
240k€ à 500k€	2 498	723	1 860	809	2 082
500k€ à 750k€	3 154	1 275	2 113	1 427	2 366
750k€ à 1000k€	3 662	1 694	2 560	1 897	2 866
1000k€ à 1250k€	3 972	2 305	2 883	2 581	3 228
1250k€ à 1500k€	4 448	2 872	3 367	3 215	3 769
1500k€ à 2000k€	5 197	3 512	3 367	3 932	3 769
2000k€ à 2500k€	5 827	4 432	3 569	4 962	3 996
2500k€ à 3000k€	6 709	5 747	4 687	6 434	5 247
3000k€ à 3500k€	6 713	7 289	6 458	8 160	7 230
3500k€ à 4000k€	7 164	7 572	6 458	8 477	7 230
4000k€ à 4500k€	8 758	9 364	6 458	10 484	7 230
4500k€ à 5000k€	8 833	9 796	6 458	10 968	7 230
5000k€ à 5500k€	6 600	11 033	6 458	12 352	7 230
5500k€ à 6000k€	8 546	11 033	6 458	12 352	7 230
6000k€ à 6500k€	8 753	11 305	7 211	12 656	8 073
6500k€ à 7000k€	16 000	13 340	12 253	14 935	13 718
>7500k€	8 236	16 805	12 253	18 814	13 718

Nous constatons que la tarification par les probabilités de défaut est systématiquement inférieure à celle par le modèle collectif sur les sociétés à faible chiffre d'affaires. Ce phénomène s'inverse sur les tranches élevées de tarification.

Conclusion

Le modèle de *scoring* développé permet de noter les entreprises françaises faisant l'objet de demande d'agrément. La classification ainsi faite par l'assureur est nécessaire pour sélectionner les risques qu'il accepte de couvrir et pour mettre ensuite en place sa politique d'arbitrage en lien avec sa sinistralité.

Comme nous avons pu l'observer, le modèle de tarification issu des probabilités de défaut donne des résultats parfois éloignés de la tarification actuelle et du marché sur certaines tranches de chiffre d'affaires. Celui-ci est issu d'hypothèses prises en amont difficilement vérifiables sans une étude approfondie auprès de la clientèle de l'assureur (notamment en ce qui concerne la cyclicité et les délais de paiements). Nous avons également, dans un souci de simplification de nos premières estimations, choisi de supposer l'indépendance des probabilités de défaillance et le calcul de celles-ci à partir des défaillances légales uniquement. L'une des premières possibilités d'enrichissement serait de calibrer également des probabilités de défaut à partir de la base sinistres pour la tarification, cela permettrait d'inclure l'effet de l'arbitrage et d'utiliser des probabilités de défaillance plus proches de la réalité de l'assureur. L'hypothèse de non indépendance des probabilités de défaillance pourrait être également étudiée afin de prendre en compte les potentielles corrélations entre le défaut des acheteurs.

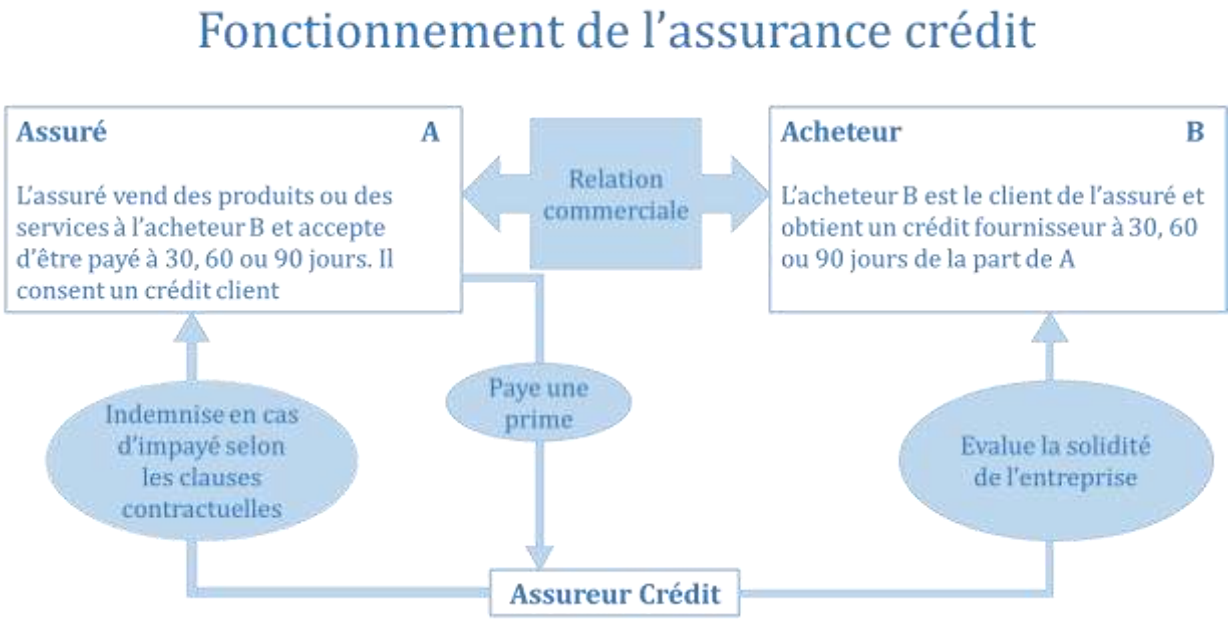
La modélisation par le modèle collectif présente l'avantage d'être simple à mettre en œuvre et utilise un ensemble de données déjà disponibles dans les bases de la compagnie. Il permet également de prendre en compte le profil de clientèle de l'assureur pour tarifer les prospects potentiels avec une information restreinte sur leur profil de risque. Nous constatons également qu'il apporte des informations sur les marges de manœuvres que peut avoir l'assureur sur ses tarifs actuels dans un marché où la concurrence se fait essentiellement par le prix. Il est ainsi rapidement applicable et répond au besoin d'AXA Assurcrédit de pouvoir plus sous-traiter les souscriptions au réseau. Cependant, il ne prend en compte qu'une information partielle sur les acheteurs de l'assuré et peut se révéler moins précis qu'une tarification par les probabilités de défaut pour la base de contrats existants, d'où la nécessité de retravailler notre premier modèle pour en améliorer la précision.

Note de synthèse.

L'assurance-crédit est un marché relativement oligopolistique (les trois premiers acteurs totalisent plus de 80% de parts de marché), mature et extrêmement concurrentiel. AXA Assurcrédit, acteur spécialisé sur le segment des PME, souhaite refondre sa tarification pour être à la fois plus sélectif sur les risques, mais également plus agressif sur son cœur de clientèle.

L'assurance-crédit permet à une entreprise de se couvrir contre le risque d'impayés de ses acheteurs. Ce produit a donc la particularité de porter, non pas sur le risque du client de l'assureur, mais sur le défaut des acheteurs de l'assuré.

L'assureur va avoir un double rôle : en premier lieu, celui d'apporter une couverture contre le risque d'impayés, mais plus indirectement de fournir à son assuré une information sur la qualité de ses acheteurs et prospects à travers les agréments délivrés ou refusés.



Les prospects mettent à disposition de la compagnie d'assurance-crédit une partie de leur portefeuille d'acheteurs afin d'en déterminer le risque et de proposer une tarification qui servira de référence pour la première année.

Une fois le contrat engagé, l'assuré va devoir déclarer l'ensemble de ses acheteurs à son assureur. Ce dernier va alors déterminer les risques qu'il accepte de couvrir ou non et accorder des agréments pour les acheteurs qui vont bénéficier du contrat d'assurance. Il réexamine régulièrement l'ensemble des acheteurs agréés afin de prendre en compte les

dernières informations disponibles sur leur situation financière et revoir éventuellement l'autorisation de couverture.

Le premier outil indispensable à l'assureur est donc un modèle de *scoring* permettant de sélectionner les risques qu'il accepte de couvrir. Cette sélection est le plus souvent automatique et systématique en fonction du score d'appartenance de l'acheteur concerné.

Nous calibrons un modèle de *scoring* sur les entreprises françaises qui doit à la fois être robuste et interprétable par les arbitres de l'assureur. Pour cela, nous disposons des bilans et comptes de résultats d'entreprises françaises pour la période 2005-2015, des défaillances légales de ces entreprises sur la même période ainsi que des données de profils des entreprises françaises à deux dates.

La variable explicative étant dichotomique (elle prend la valeur de un en cas de défaillance et zéro sinon), nous utilisons la régression logistique. Le *scoring* est développé en deux étapes, tout d'abord un sous-score financier, uniquement basé sur les éléments financiers des entreprises, puis le modèle de *scoring* final.

Pour calibrer le sous-score financier, nous observons sur trois exercices distincts (2009 à 2011) les défaillances qui interviennent dans les vingt-quatre mois suivant la date de clôture des comptes. Nous définissons des ratios financiers qui seront utilisés comme variables explicatives de la survenance ou non d'une défaillance. Cette première étape nous permet de calibrer un score financier avec six à dix variables discriminantes sur quatre macro-secteurs.

Le modèle de *scoring* final intègre le sous-score financier ainsi que des éléments non financiers. Il est calibré à partir d'une date fixe sur une période d'un an. L'échantillon étudié regroupe un peu plus de 1,7 million d'entreprises commerciales. Nous travaillons, comme pour le score financier, à rechercher les variables les plus discriminantes permettant d'expliquer la défaillance.

Le modèle de *scoring* ainsi calibré nous permet de calculer des probabilités de défaillance pour les entreprises françaises. Ces entreprises sont regroupées par classe permettant aux arbitres de sélectionner les risques que l'assureur accepte de couvrir.

Note	1	2	3	4	5	6	7	8	9	10
Probabilité de défaut à 12 mois	18,9%	13,7%	9,3%	6,7%	4,5%	3,3%	2,2%	1,3%	0,5%	0,1%

Cette première étape est ainsi un préalable à l'élaboration d'une tarification basée sur les probabilités de défaut.

La prime pure étant l'espérance mathématique de pertes, nous devons fixer les hypothèses permettant de calculer l'exposition en cas de perte (EAD ou *Exposure At Default*), puis la perte espérée via les probabilités de défaut.

Nous disposons, pour estimer notre EAD, des agréments délivrés par l'assureur à son client ainsi que de son chiffre d'affaires.

Les agréments peuvent ne pas refléter le risque effectivement porté par la compagnie. En effet :

- Certains clients demandent des agréments pour des prospects avec lesquels ils ne concrétisent pas de transaction, l'assureur crédit est alors utilisé comme un outil de prospection commerciale.
- Le montant demandé peut se révéler supérieur au risque effectivement couvert, l'assuré voulant garder une certaine flexibilité pour traiter avec son client.

Ces deux éléments nous conduisent à déterminer un taux d'utilisation des agréments ou « use factor » qui sera calculé par assuré. La partie du « use factor » liée au montant autorisé est estimée via la base sinistres en comparant les agréments des montants de sinistres effectivement déclarés. La partie du « use factor » liée aux prospects est calculée par assuré en comparant le chiffre d'affaires réel d'un chiffre d'affaires reconstitué via les agréments demandés.

Nous avons ainsi la possibilité de calculer l'exposition au défaut par acheteur (indicé i) puis par assuré (indicé j) :

$$EAD_i = Agrément_i \times Min\left(1\,, Use\ Factor \times Max\left(1\,,\ \frac{Délai\ de\ paiement_i}{\frac{12}{Cyclicité_i}}\right)\right)$$

La cyclicité permet de fixer le nombre de factures émises dans l'année et est utilisée pour reconstituer le chiffre d'affaires.

La prime pure va être calculée en utilisant les probabilités de défaillance du modèle de *scoring*, nous aurons donc pour un portefeuille d'acheteurs d'un assuré j une perte espérée qui sera :

$$EL_j = \sum_{i=1}^{n} P_i \times LGD_i \times QG_i \times EAD_i$$

Avec :

- j l'indice du portefeuille de l'assuré j.
- i l'indice d'un acheteur.
- LGD_i la « loss given default » de l'acheteur i (fonction de son secteur d'appartenance).
- QG_i la quotité garantie de l'acheteur i, c'est le pourcentage de l'agrément effectivement couvert.
- P_i la probabilité de défaut de l'acheteur i.
- EAD_i L'exposition au défaut de l'acheteur i.

L'une des faiblesses de ce premier modèle est qu'il suppose que l'assureur ait la connaissance complète du portefeuille d'acheteurs de son client. Si c'est le cas pour les clients existants, les propositions commerciales pour les prospects se font sur la base d'un échantillon d'acheteurs communiqué par le futur assuré (généralement les dix principaux clients).

Nous choisissons de mettre en place un second outil de tarification basé sur le modèle collectif en se restreignant aux données disponibles lors de la souscription d'un nouveau contrat.

En premier lieu, nous construisons une base de données reprenant les caractéristiques des clients ainsi que de leurs dix premiers acheteurs à partir des bases de données existantes sur quatre années (2012 à 2015). Nous intégrons ensuite la sinistralité de ces années (nombre de sinistres par police dans l'année et montant cumulé de ces sinistres). Une fois les bases de données construites, nous calibrons un modèle sur la fréquence puis un second sur l'intensité sur un partie de nos données. Deux modèles sont calibrés sur les deux principaux contrats de la société.

Nous obtenons un écart entre la prime pure et le montant cumulé de sinistres de 27% pour le contrat dit « Globale » et 21% sur le produit dit « Kup » nettement plus standardisé.

Le passage de la prime pure à la prime totale est réalisé sur la base d'un ratio combiné économique (ECR) de 100%, c'est-à-dire hors marge. La prime finale sera alors composée

de la prime pure, des frais et charges (exprimés en pourcentage de la prime totale), du coût du traité de réassurance et du coût du capital mobilisé dans le cadre de la réglementation prudentielle Solvabilité II.

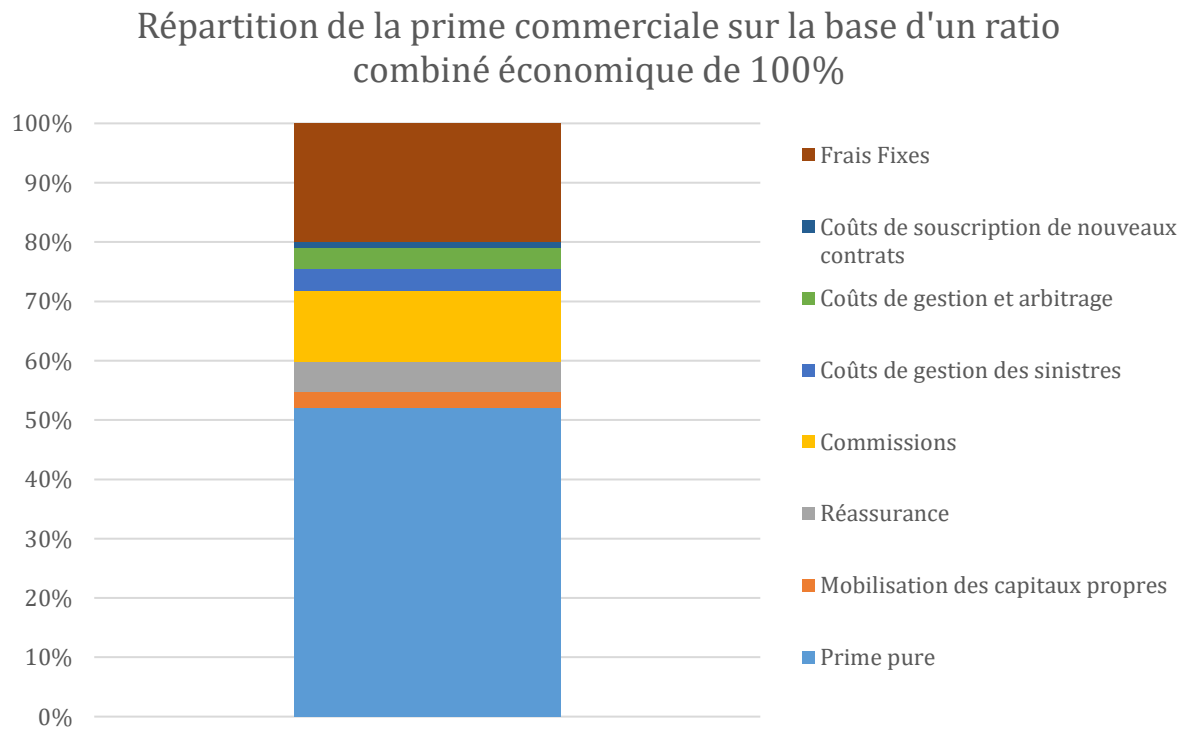
La société vise 40% de ratio coûts / primes. Le traité de réassurance est facturé 4,95% de la prime totale et permet à l'entreprise de limiter sa sinistralité à 80% des primes perçues. La sinistralité étant ainsi bornée, nous connaissons le risque de perte maximum (80% des primes totales) et pouvons en déduire le capital à mobiliser dans le cadre de Solvabilité II (le SCR sera de 80% de la prime totale moins la prime pure). La mobilisation de ce capital est ensuite refacturée à chaque assuré au prorata de sa prime.

La prime finale P peut ainsi s'écrire en fonction de P et de la prime pure (PP) :

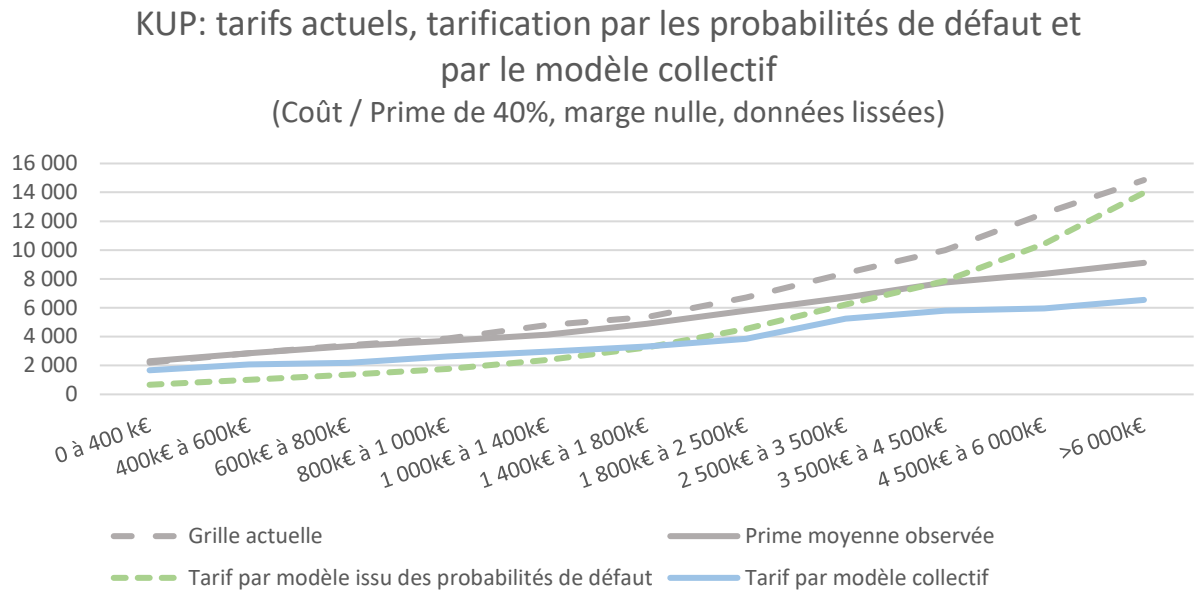
$$P = PP + 10\% \times (80\% \times P - PP) + 45{,}18\% \times P$$

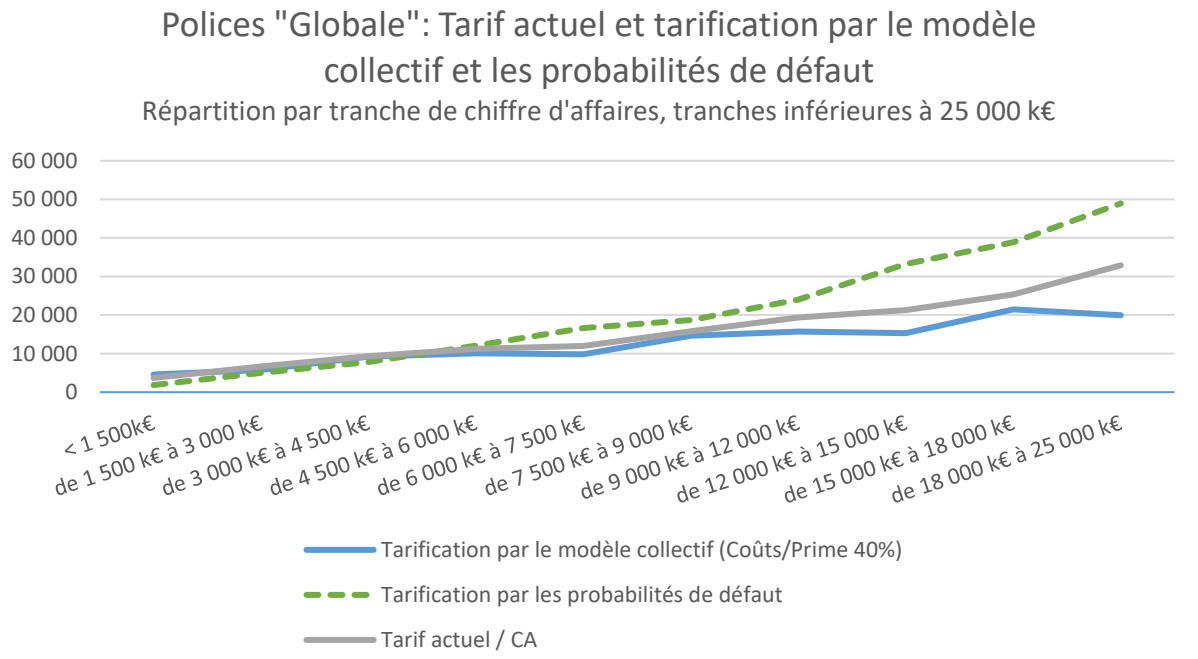
$$\Rightarrow P = 1{,}922 \times PP$$

Ce qui permet de mesurer et piloter la prime et les frais sur la base de ce ratio combiné économique comme présenté ci-dessous :



Nous disposons d'outils qui nous permettent de mettre en œuvre une tarification sur les deux principaux produits de la société par les deux modèles, et la comparer à la tarification actuelle :





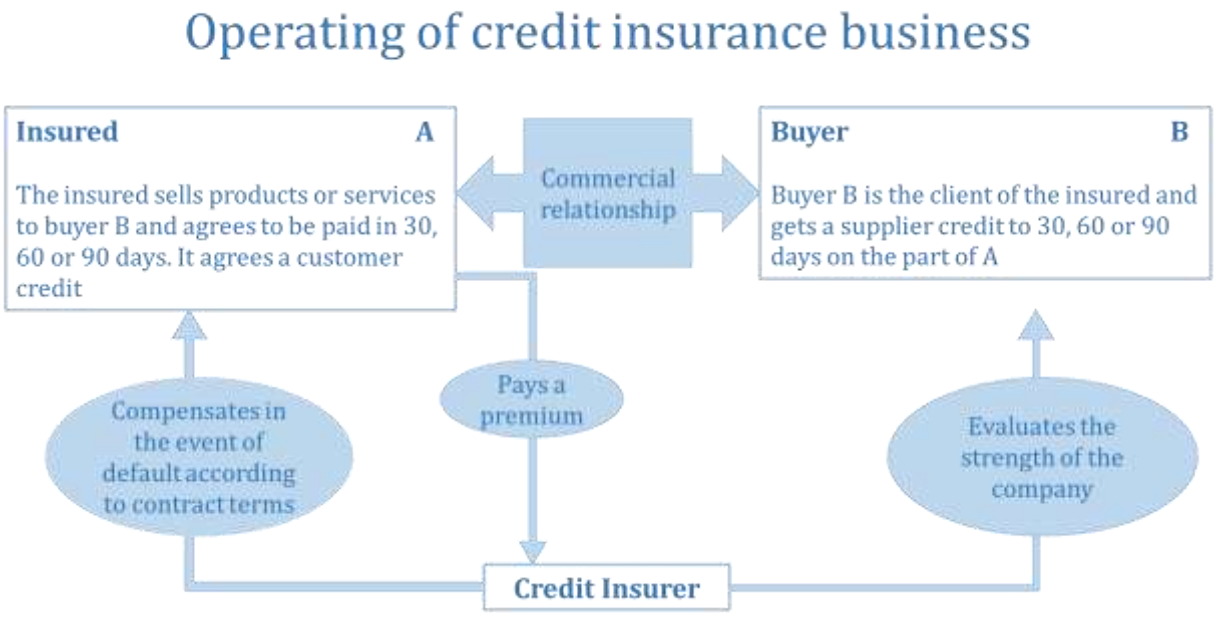
Le tarif calculé via une modélisation par les probabilités de défaut se révèle très éloigné de la réalité des tarifs pratiqués et nécessite encore des ajustements à la fois pour le calcul des probabilités de défaillances utilisées ainsi que pour l'agrégation des risques. En revanche, le tarif issu du modèle collectif permet d'avoir un outil plus synthétique de tarification utilisable pour les prospects. Il permet également d'identifier les contrats ou les tranches de CA pour lesquelles l'assureur a une certaine marge de manœuvre par rapport au tarif actuel.

Executive summary

Credit insurance is a market which is relatively oligopolistic (the three main players represent more than 80% of the market shares), mature and extremely competitive. AXA Assurcrédit, a company specialized in the SME segment, intends to overhaul its pricing to be both more selective regarding risk, but also more aggressive regarding its core customers.

Credit insurance enables a company to hedge against the risk of default in payment of its buyers. This product has therefore the particularity to focus not on the risk of the insurance's customer itself, but rather on the failure of insured's buyers.

Insurers will have a dual role: first to provide coverage against the risk of default, second, and more indirectly, to provide information on the quality of its buyers and prospects through approvals or denied agreements.



Prospects make available to the credit insurance company part of their buyers' portfolio in order to determine the risk and propose a pricing that it can use as reference for the first year.

Once the contract has been signed, the insured has to declare all its buyers to its insurer. The insurer will determine the risks that it agrees to cover or not and provide agreements for the buyers which will benefit from the insurance contract. It regularly reviews all

buyer's agreements to take into account the latest information on their financial situation and possibly review the authorization of coverage.

The first necessary tool for the insurer is therefore a scoring model to select the risks that it accepts to cover. This selection is often automatic and systematic according to the score of the buyer.

It is necessary for the calibration of the scoring model for French companies to be both robust and interpretable by the arbitrators of the insurer. To achieve this, we hold the balance sheets and income statements of French companies for the period 2005-2015, as well as the legal defaults of these companies over the same period and their data profiles on two dates.

The explanatory variable being dichotomous (it takes the value of one in case of failure and zero otherwise), we use logistic regression. The scoring is developed in two stages, first a financial score, solely based on the financial elements of the businesses, then the final scoring model.

To calibrate the financial score, we observe during three separate years (from 2009 to 2011) the failures occurring within 24 months of the date of the closing of the accounts. We define financial ratios to be used as predictors for the occurrence of a failure. This first step allows us to calibrate a financial score with six to ten discriminating variables on four macro-sectors.

The final scoring model incorporates the financial score and non-financial items. It is calibrated from a fixed date for a period of one year. The sample includes slightly more than 1.7 million commercial companies. As to the financial score, the calibration consists of searching the most discriminating variables to explain the failure.

Once the model is calibrated, it allows us to calculate probabilities of failure for French companies. These companies are grouped into classes allowing the arbitrators to select the risks the insurer will agree to cover

Score level	1	2	3	4	5	6	7	8	9	10
12 months' default probability	18,9%	13,7%	9,3%	6,7%	4,5%	3,3%	2,2%	1,3%	0,5%	0,1%

This first step is thus a prerequisite for the development of a pricing based on the default probabilities.

The pure premium being the mathematical expectation of losses, so as to calculate it, we must first fix the hypothesis used to determine the exposure in case of loss (EAD or Exposure At Default), then the loss expected via default probabilities.

To estimate our EAD, we have agreements given by the insurer to its customers and its turnover.

Agreements may not reflect the risk carried by the company. Indeed:

- Some clients ask agreements for prospects with which they can have no transaction at all, the credit insurer is then used as a prospecting tool.
- The limit requested may be above the effectively covered risk, the insured wanting to keep some flexibility to deal with his client.

Those two elements enable us to determine a rate of use for the agreements or "use factor" which will be calculated for each insured. The part of the "use factor" related to the authorized amount is estimated via the sinister base by comparing the limits with the claims amount. The part of the “use factor” related to prospects is calculated by comparing the actual turnover with an estimated turnover using agreements amounts.

We can calculate the exposure to default per buyer (indices i) and per insured (indices j):

$$EAD_i = Agreement_i \times Min\left(1\ , Use\ Factor \times Max\left(1\ ,\ \frac{Payment\ period_i}{\frac{12}{Cyclicality_i}}\right)\right)$$

The cyclicality allows to set the number of invoices issued in the year and is used to reconstitute the turnover.

The pure premium will be calculated using the default probabilities of the scoring model, so the expected loss will be for a portfolio of buyers of an insured j:

$$EL_j = \sum_{i=1}^{n} P_i \times LGD_i \times QG_i \times EAD_i$$

With:

- j the index of the insured portfolio j.
- i the index of a buyer.
- LGD_i the" loss given default" of buyer i (depending on its area of belonging).
- QG_i the percentage of cover buyer i, this is the percentage of approval actually covered.
- P_i the default probability of the buyer i.
- EAD_i the exposure to the default of the buyer i.

One of the weaknesses of this first pricing model is that it requires that the insurer has full knowledge of the portfolio of buyers of its client. If this is the case for existing customers, business proposals for prospects are based on a sample of buyers provided by the future insured (generally the ten major customers). We have chosen to set up a second pricing tool based on the collective model restricting data available during the subscription process of a new contract.

First, we develop a database with the characteristics of the customers as well as their first ten buyers from existing databases on four years (2012-2015). We then integrate claims for each of these years (cumulative number of claims by police in the year and amount of these claims). Once the databases have been built, we calibrate one model on the claim frequency then a second on claim intensity. Two models are calibrated for the two major contracts of the company.

Those pricing models have a discrepancy between pure premium and the cumulative amount of claims of 27% for the 'Globale' contract and 21% for the Kup product (which is much more standardized).

The bridge between the pure premium and the total premium is achieved on the basis of an economic combined ratio (ECR) of 100%, i.e. out of margin. The final premium will then consist of pure premium, fees and charges (as a percentage of the total premium), the cost of the reinsurance treaty and the cost of the capital needed under Solvency II Regulation.

The company targets a 40% ratio for costs / premiums. The reinsurance treaty charges 4.95% of the total premium and allows the company to limit its total claims to 80% of the

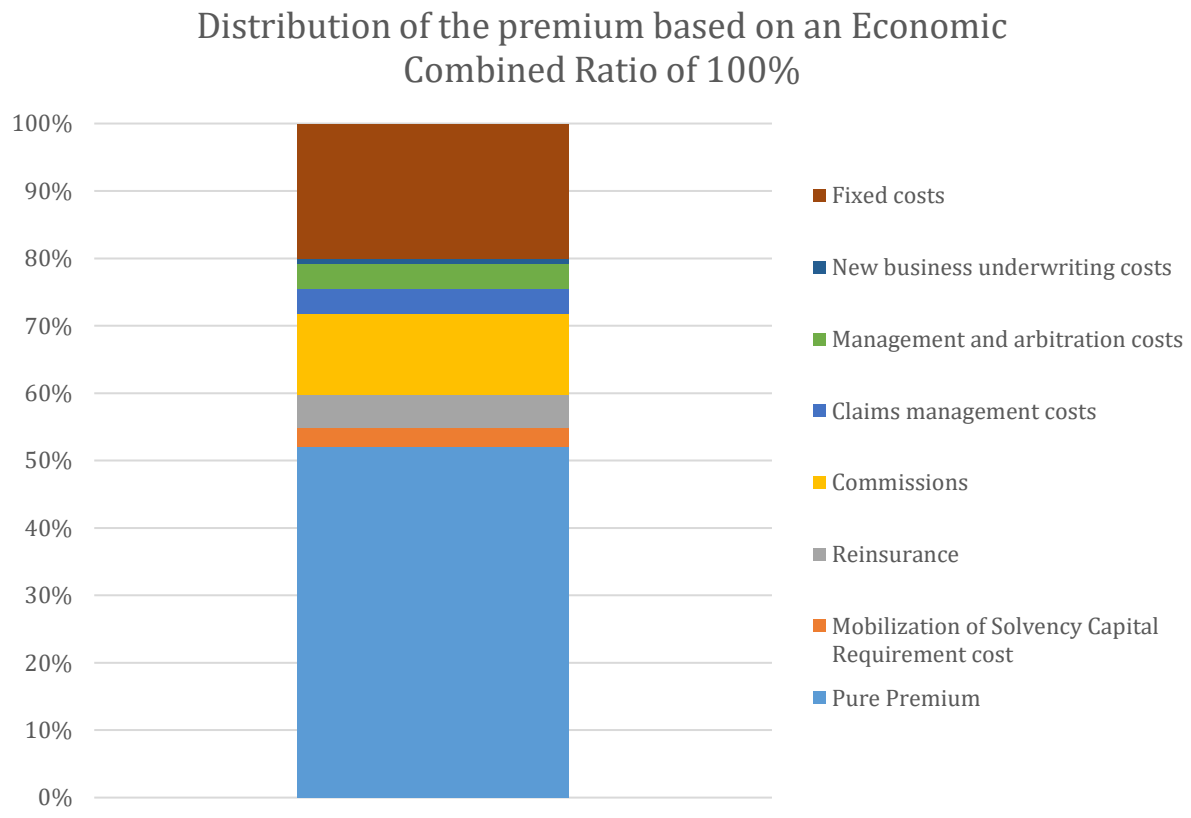
premiums collected. Claims being then limited, we know the risk of maximum loss (80% of total premiums) and can deduce the capital to mobilize under Solvency II (the SCR will be 80% of the total premium less the pure premium). The mobilization of this capital is then charged to each insured in proportion to his total premium.

The final premium P can thus be written according to P and the pure premium (PP) :

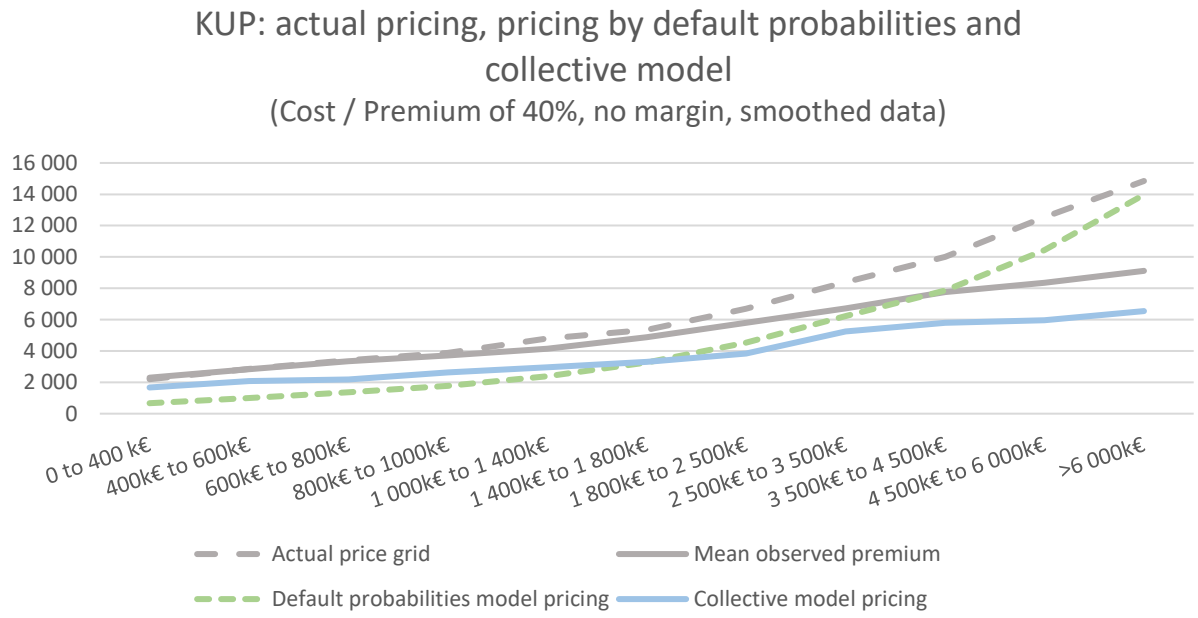
$$P = PP + 10\% \times (80\% \times P - PP) + 45.18\% \times P$$

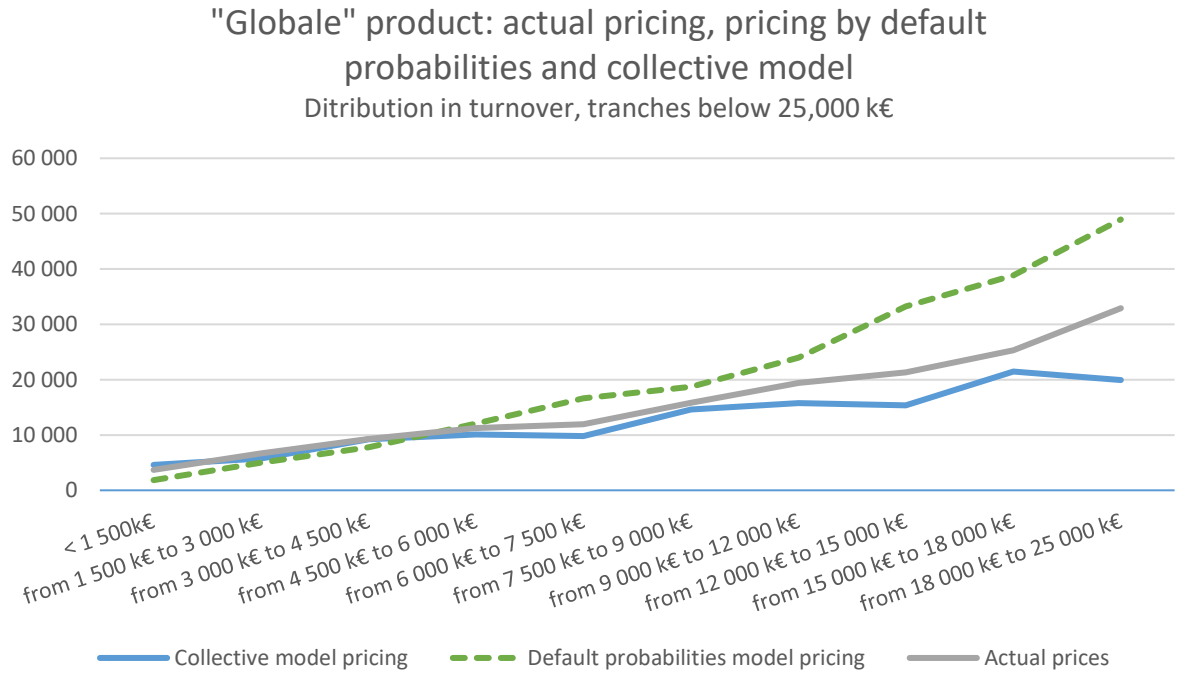
$$\Rightarrow P = 1.922 \times PP$$

This allows to measure and control the premium and fees on the basis of the economic combined ratio below:



We have tools that enable us to implement a pricing on the two main products of the company using both models and compare it to the current pricing:





The price calculated through modelling by default probabilities appears to be far from the price reality and still requires adjustments both for the calculation of the probabilities of failure used as well as for the aggregation of risk. However, prices computed with the collective model allows to have a more synthetic pricing tool usable for prospects. It also allows to identify contracts or turnover trenches on which the insurer has a certain flexibility regarding the current tariff.

Bibliographie

AXA - Comité technique RI. «Nouvelle approche segmentée des PLR.» 2014.

Banque de France. «Fiche N° 224 - Les sociétés d'assurance-crédit.» Direction des entreprises - Référentiel des Financements des Entreprises, 2012.

Banque de France. «Fiche N°701 - Les produits de l'assurance crédit.» Direction des entreprises - Référentiel des Financements des Entreprises, 2013.

Baradel, Nicolas. «Théorie du risque.» 2016.

Becue, Paul. *Assurance-crédit et assurance-cautionnement.* Kluwer, 2013.

Bentahar, Billel, et Antoine Vercherin. «Refondre un tarif dans un nouveau contexte réglementaire et concurrentiel.» Mémoire du Centre d'Etude Actuariel, AXA, 2013.

Boisselier, Patrick, et Dominique Dufour. «Scoring et anticipation de défaillance des entreprises: une approche par la régression logistique. Identification et maîtrise des risques: enjeux pour l'audit, la comptabilité et le contrôle de gestion.» 2003, éd. HAL.

Caja, Anisa. «Contribution à la mesure des engagements et du besoin en capital pour un assureur crédit.» *Thèse.* Universite Claude Bernard Lyon 1, 2014.

Charpentier, Arthur. «Actuariat de l'assurance non-vie.» 2015.

Charpentier, Arthur, et Michel Denuit. *Mathématiques de l'assurance non-vie - Tome I: Principes fondamentaux de théorie du risque.* Economica, 2004.

—. *Mathématiques de l'assurance non-vie - Tome II: Tarification et provisionnement.* Economica, 2004.

Dupont, Alexis. «Fair price en Assurance-Crédit dans un cadre de diversification.» Mémoire d'acuariat, ENSAE / Euler Hermès, 2010.

Observatoire des délais de paiement. «Rapport annuel de l'observatoire des délais de paiement 2014-2015.» 2016.

Planchet, Frédéric, Jean-François Decroocq, et Fabrice Magnin. «Systematic risk modelisation in credit risk insurance.» 2009.

Rakotomalala, Ricco. *Pratique de la régression logistique (Régression Logistique Binaire et Polytomique).* Université Lumière Lyon 2, 2015.

Tufféry, Stéphane. *Modélisation prédictive et apprentissage statistique.* Technip, 2015.

Wajnberg, Eric. «Introduction au Modèle Linéaire Généralisé.» 2011. http://www.unice.fr/coquillard/UE7/cours%20IV%20(GLM%20I).pdf.

6 Annexes

6.1 Courbe ROC et AUC

La courbe ROC : *Receiver Operating Characteristic* (Charpentier, Actuariat de l'assurance non-vie 2015).

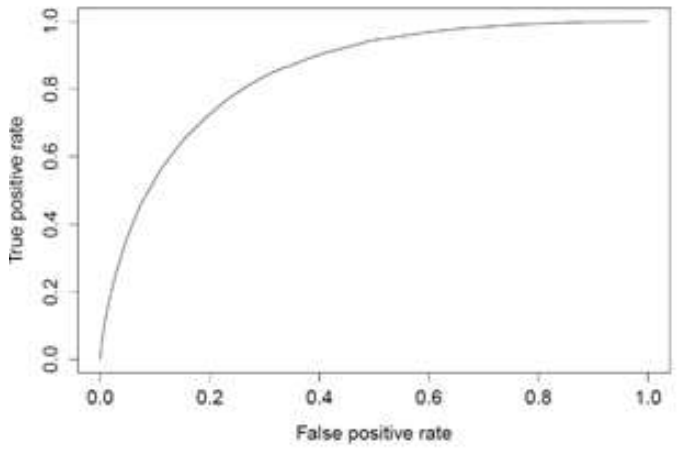
Nous disposons d'observations Y_i prenant les valeurs 0 (nous dirons négatives) ou 1 (nous dirons positives).

Nous construisons un modèle qui prédit π_i. En se fixant un seuil s, si $\widehat{\pi_i} \leq s$, nous prédisons $\widehat{Y_i} = 0$, et si $\widehat{\pi_i} < s$, nous prédisons $\widehat{Y_i} = 1$.

Le modèle sera performant si les positifs sont prédits positifs et les négatifs sont prédits négatifs. Le choix du seuil s permettra de minimiser soit les faux positifs (FP), soit les faux négatifs (FN).

	Y=1	*Y=0*
$\widehat{Y_i} = 1$	Vrais positifs	Faux positifs
$\widehat{Y_i} = 0$	Faux négatifs	Vrais négatifs

La courbe ROC est tracée en faisant varier le seuil de 0 à 1 et, pour chaque cas, nous calculons le taux de vrais positifs, le taux de faux positifs que l'on reporte dans un graphique avec, en abscisse, le taux de faux positifs et en ordonnée le taux de vrais positifs.



L'aire sous la courbe ROC (*Area Under the Curve* ou AUC) peut être interprétée comme la probabilité qu'une observation soit mieux prédite qu'un tirage purement aléatoire.

L'utilisation de la courbe ROC et de l'AUC va donc consister à maximiser l'AUC.

6.2 Modèle de *scoring*

6.2.1 *Scoring* financier

6.2.1.1 Ratios financiers étudiés

nom « R »	Ratio
R1	(actif-dettes-provisions pour risques et charges-actif immobilisé)/passif
R2	fonds propres/passif
R3	dettes fournisseurs/passif
R4	dettes fiscales et sociales/passif
R5	capitaux permanents/passif
R6	EBE/passif
R7	résultat net/passif
R8	FDR/CA*365
R9	(FDR-BFR)/CA*365
R10	résultat d'exploitation/CA
R11	intérêts financiers/CA
R12	RCAI/CA
R13	MBA/CA
R14	fonds propres/capital social
R15	dettes financières/fonds propres
R16	dettes financières long terme/fonds propres
R17	CAF/fonds propres
R18	actif circulant/dettes court terme
R19	valeurs réalisables court terme/dettes court terme
R20	(valeurs mobilières de placement + disponibilités)/dettes court terme
R21	EBE/dettes
R22	fond propres/capitaux permanents
R23	ressources stables/emplois stables
R24	fonds propres/ressources stables
R25	dettes financières/ressources stables
R26	(RCAI + intérêts financiers)/ressources stables
R27	charges de personnel/VA
R28	dettes financières/EBE
R29	intérêts financiers/EBE
R30	dettes financières/CAF
R31	FDR/BFR
R33	résultat d'exploitation/(actif immobilisé + BFR)

R35	(dettes financières nettes)/EBE
R36	dettes/CAF
R37	fonds propres/actif immobilisé net
R38	fonds propres/dettes financières
R39	fonds propres/dettes
R40	(charges fixes + résultat d'exploitation)/(charges fixes + intérêts financiers)

6.2.1.2 Score financier : estimations des paramètres sur le secteur 1

Extractions du logiciel R

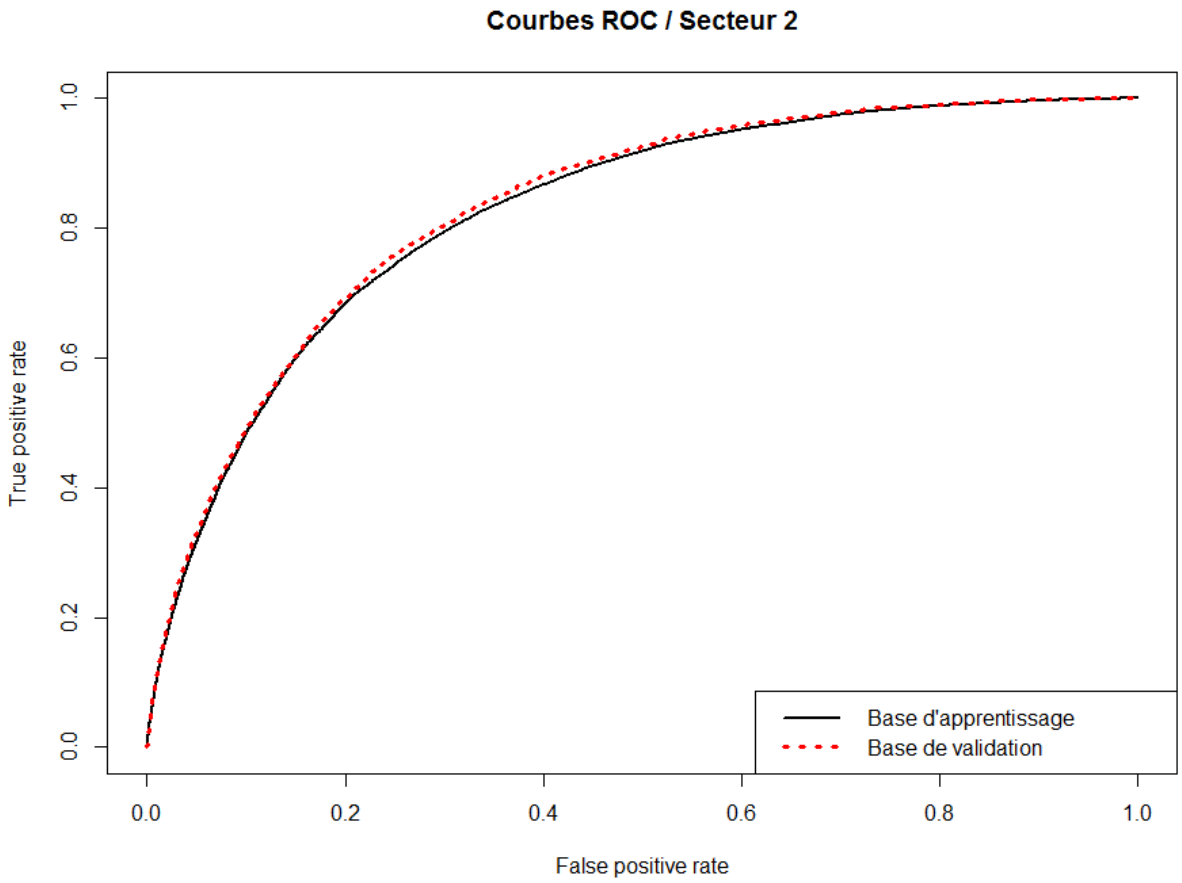
```
Call:
glm(formula = def24 ~ ., family = binomial(link = "logit"), data = subset(train,
    select = c(def24, R2_t, R3_t, R4_t, R9_t, R13_t, R20_t, R27_t)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1406  -0.2622  -0.1475  -0.0843   3.8246

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.30589    0.07874 -29.286  < 2e-16 ***
R2_t(-0.2,0]       -0.06525    0.03546  -1.840 0.065750 .
R2_t(0,0.1]        -0.17757    0.03459  -5.133 2.85e-07 ***
R2_t(0.1,0.2]      -0.48804    0.03554 -13.733  < 2e-16 ***
R2_t(0.2,0.3]      -0.81037    0.03940 -20.569  < 2e-16 ***
R2_t(0.3,0.4]      -1.01910    0.04453 -22.884  < 2e-16 ***
R2_t(0.4,0.5]      -1.22802    0.05293 -23.202  < 2e-16 ***
R2_t(0.5,1]        -1.35234    0.05675 -23.830  < 2e-16 ***
R3_t(0.05,0.1]      0.11528    0.05158   2.235 0.025431 *
R3_t(0.1,0.15]      0.18894    0.04954   3.814 0.000137 ***
R3_t(0.15,0.2]      0.24120    0.04923   4.900 9.59e-07 ***
R3_t(0.2,0.3]       0.28574    0.04672   6.116 9.61e-10 ***
R3_t(0.3,0.35]      0.39830    0.05260   7.573 3.65e-14 ***
R3_t(0.35,0.45]     0.50384    0.05014  10.049  < 2e-16 ***
R3_t(0.45,2]        0.52473    0.05043  10.406  < 2e-16 ***
R4_t(0.1,0.15]      0.16111    0.05275   3.054 0.002255 **
R4_t(0.15,0.2]      0.30473    0.04933   6.177 6.54e-10 ***
R4_t(0.2,0.3]       0.50024    0.04481  11.165  < 2e-16 ***
R4_t(0.3,0.4]       0.75349    0.04620  16.310  < 2e-16 ***
R4_t(0.4,0.5]       0.90374    0.04876  18.536  < 2e-16 ***
R4_t(0.5,2]         1.03592    0.04663  22.218  < 2e-16 ***
R9_t(-10,1]        -0.40109    0.02983 -13.444  < 2e-16 ***
R9_t(1,15]         -0.58339    0.02974 -19.615  < 2e-16 ***
R9_t(15,30]        -0.82247    0.04165 -19.748  < 2e-16 ***
R9_t(30,50]        -1.15328    0.05620 -20.520  < 2e-16 ***
R9_t(50,100]       -1.45256    0.06851 -21.203  < 2e-16 ***
R9_t(100,2e+03]    -2.16140    0.10703 -20.194  < 2e-16 ***
R13_t(-0.15,-0.05] -0.22366    0.03970  -5.634 1.76e-08 ***
R13_t(-0.05,0]     -0.35953    0.04661  -7.713 1.23e-14 ***
R13_t(0,0.025]     -0.52078    0.04985 -10.448  < 2e-16 ***
R13_t(0.025,0.05]  -0.63596    0.05197 -12.236  < 2e-16 ***
R13_t(0.05,2]      -0.74915    0.05151 -14.543  < 2e-16 ***
R20_t(0.2,0.25]    -0.26571    0.05304  -5.010 5.45e-07 ***
R20_t(0.25,0.55]   -0.30816    0.04278  -7.203 5.89e-13 ***
R20_t(0.55,1.5]    -0.35022    0.06199  -5.650 1.60e-08 ***
R20_t(1.5,50]      -0.74448    0.12466  -5.972 2.35e-09 ***
R27_t(0.8,0.9]      0.18134    0.03380   5.366 8.06e-08 ***
R27_t(0.9,1]        0.36830    0.03712   9.922  < 2e-16 ***
R27_t(1,1.1]        0.45833    0.04419  10.373  < 2e-16 ***
R27_t(1.1,1.2]      0.58122    0.05015  11.591  < 2e-16 ***
R27_t(1.2,5]        0.65792    0.04530  14.524  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118855  on 395717  degrees of freedom
Residual deviance:  94977  on 395677  degrees of freedom
AIC: 95059

Number of Fisher Scoring iterations: 9
```

6.2.1.3 Score financier : estimations des paramètres sur le secteur 2

Extractions du logiciel R

```
Call:
glm(formula = def24 ~ ., family = binomial(link = "logit"), data = subset(train,
    select = c(def24, R2_t, R3_t, R4_t, R5_t, R9_t, R12_t, R13_t,
        R20_t, R23_t, R27_t)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8290  -0.2106  -0.1283  -0.0794   3.7343

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.45338    0.06178 -39.711  < 2e-16 ***
R2_t(-0.2,0]       -0.06300    0.03263  -1.931  0.05348 .
R2_t(0,0.1]        -0.20986    0.03400  -6.173 6.72e-10 ***
R2_t(0.1,0.2]      -0.60628    0.03844 -15.772  < 2e-16 ***
R2_t(0.2,0.5]      -1.13522    0.03808 -29.813  < 2e-16 ***
R2_t(0.5,1]        -1.53431    0.05514 -27.824  < 2e-16 ***
R3_t(0.1,0.25]      0.37845    0.02802  13.509  < 2e-16 ***
R3_t(0.25,2]        0.58105    0.03252  17.866  < 2e-16 ***
R4_t(0.1,0.25]      0.20644    0.02785   7.413 1.23e-13 ***
R4_t(0.25,2]        0.33011    0.03375   9.781  < 2e-16 ***
R5_t(0.1,0.2]       0.25977    0.04343   5.981 2.21e-09 ***
R5_t(0.2,0.6]       0.30560    0.03582   8.532  < 2e-16 ***
R5_t(0.6,1.5]       0.52427    0.04900  10.699  < 2e-16 ***
R9_t(1,5]          -0.06273    0.03068  -2.045  0.04087 *
R9_t(5,10]         -0.11211    0.03868  -2.898  0.00375 **
R9_t(10,20]        -0.18020    0.04286  -4.204 2.62e-05 ***
R9_t(20,80]        -0.49390    0.05007  -9.864  < 2e-16 ***
R9_t(80,2e+03]     -0.87222    0.08415 -10.365  < 2e-16 ***
R12_t(-0.15,-0.1]  -0.05414    0.04429  -1.222  0.22153
R12_t(-0.1,-0.04]  -0.19063    0.04487  -4.249 2.15e-05 ***
R12_t(-0.04,0]     -0.44408    0.04735  -9.379  < 2e-16 ***
R12_t(0,0.02]      -0.58993    0.05107 -11.552  < 2e-16 ***
R12_t(0.02,0.03]   -0.66952    0.06286 -10.650  < 2e-16 ***
R12_t(0.03,0.07]   -0.74208    0.05480 -13.541  < 2e-16 ***
R12_t(0.07,2]      -0.79953    0.05965 -13.405  < 2e-16 ***
R13_t(-0.1,0]      -0.09793    0.04177  -2.344  0.01906 *
R13_t(0,0.025]     -0.32087    0.04837  -6.633 3.28e-11 ***
R13_t(0.025,0.075] -0.38071    0.04807  -7.921 2.36e-15 ***
R13_t(0.075,2]     -0.44531    0.05147  -8.653  < 2e-16 ***
R20_t(0.02,0.05]   -0.07695    0.03214  -2.394  0.01665 *
R20_t(0.05,0.1]    -0.27129    0.03715  -7.303 2.82e-13 ***
R20_t(0.1,0.2]     -0.38296    0.04290  -8.927  < 2e-16 ***
R20_t(0.2,0.4]     -0.63458    0.05168 -12.280  < 2e-16 ***
R20_t(0.4,0.75]    -0.84952    0.06331 -13.419  < 2e-16 ***
R20_t(0.75,50]     -1.09070    0.07565 -14.418  < 2e-16 ***
R23_t(0.6,0.8]     -0.06892    0.03695  -1.866  0.06211 .
R23_t(0.8,0.95]    -0.16416    0.03979  -4.126 3.69e-05 ***
R23_t(0.95,150]    -0.30040    0.03684  -8.154 3.52e-16 ***
R27_t(0.85,0.95]    0.02391    0.03158   0.757  0.44906
R27_t(0.95,1.15]    0.10137    0.03445   2.943  0.00325 **
R27_t(1.15,1.35]    0.25166    0.04335   5.805 6.44e-09 ***
R27_t(1.35,5]       0.34525    0.03937   8.769  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 113712  on 561285  degrees of freedom
Residual deviance:  96253  on 561244  degrees of freedom
AIC: 96337

Number of Fisher Scoring iterations: 8
```

6.2.1.4 Score financier, secteur 2 : courbes ROC et AUC



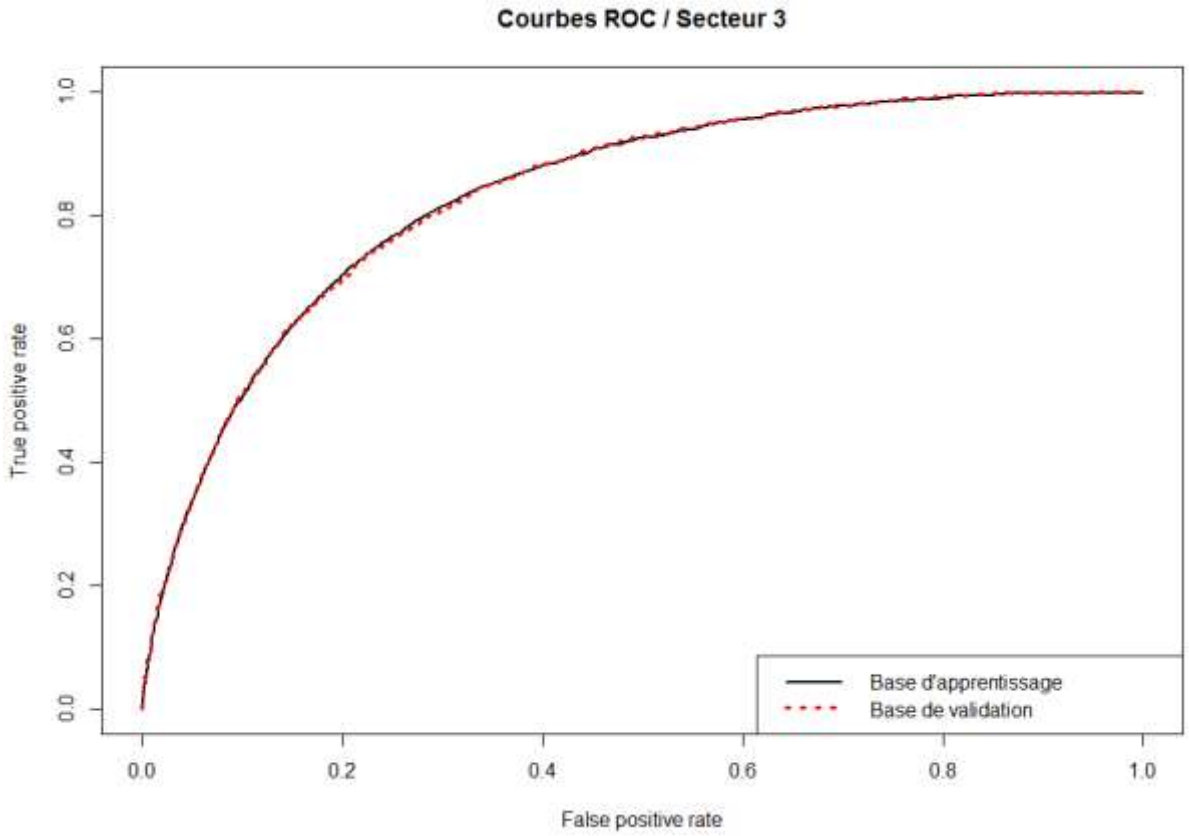

```
AUC : 0.8233099 (Base d'apprentissage)
AUC : 0.8306569 (Base de validation)
```

6.2.1.5 Score financier : estimations des paramètres sur le secteur 3

Extractions du logiciel R

```
Call:
glm(formula = def24 ~ ., family = binomial(link = "logit"), data = subset(train,
    select = c(def24, R2_t, R5_t, R9_t, R12_t, R19_t, R20_t)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6616  -0.1899  -0.1129  -0.0681   3.7073

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.40795    0.04536 -31.039  < 2e-16 ***
R2_t(-0.075,0.075] -0.23288    0.05808  -4.010 6.08e-05 ***
R2_t(0.075,0.15]   -0.38151    0.06501  -5.869 4.39e-09 ***
R2_t(0.15,0.25]    -0.60408    0.06507  -9.283  < 2e-16 ***
R2_t(0.25,1]       -0.93015    0.06325 -14.707  < 2e-16 ***
R5_t(-0.05,0.05]   -0.05466    0.07726  -0.707 0.479312
R5_t(0.05,0.35]    -0.26542    0.06256  -4.243 2.21e-05 ***
R5_t(0.35,0.55]    -0.38953    0.06874  -5.666 1.46e-08 ***
R5_t(0.55,1.5]     -0.75527    0.07308 -10.334  < 2e-16 ***
R9_t(1,10]         -0.20769    0.04595  -4.520 6.18e-06 ***
R9_t(10,20]        -0.29426    0.06708  -4.387 1.15e-05 ***
R9_t(20,30]        -0.44174    0.08120  -5.440 5.32e-08 ***
R9_t(30,100]       -0.62089    0.08129  -7.638 2.20e-14 ***
R9_t(100,2e+03]    -1.27775    0.11631 -10.985  < 2e-16 ***
R12_t(-0.1,0]      -0.38550    0.04339  -8.885  < 2e-16 ***
R12_t(0,0.03]      -0.73895    0.04837 -15.277  < 2e-16 ***
R12_t(0.03,0.09]   -0.90003    0.05252 -17.136  < 2e-16 ***
R12_t(0.09,0.15]   -1.17096    0.07795 -15.022  < 2e-16 ***
R12_t(0.15,2]      -1.38427    0.08160 -16.964  < 2e-16 ***
R19_t(0.9,1.1]     -0.06182    0.04863  -1.271 0.203610
R19_t(1.1,1.5]     -0.17716    0.04698  -3.771 0.000162 ***
R19_t(1.5,2]       -0.21747    0.05999  -3.625 0.000289 ***
R19_t(2,50]        -0.41482    0.07141  -5.809 6.30e-09 ***
R20_t(0.05,0.2]    -0.15543    0.05356  -2.902 0.003709 **
R20_t(0.2,0.4]     -0.37115    0.07756  -4.785 1.71e-06 ***
R20_t(0.4,0.7]     -0.44561    0.09296  -4.794 1.64e-06 ***
R20_t(0.7,1]       -0.55088    0.11608  -4.746 2.08e-06 ***
R20_t(1,50]        -0.70090    0.12341  -5.680 1.35e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45794  on 260236  degrees of freedom
Residual deviance: 38460  on 260209  degrees of freedom
AIC: 38516

Number of Fisher Scoring iterations: 8
```

6.2.1.6 Score financier, secteur 3 : courbes ROC et AUC



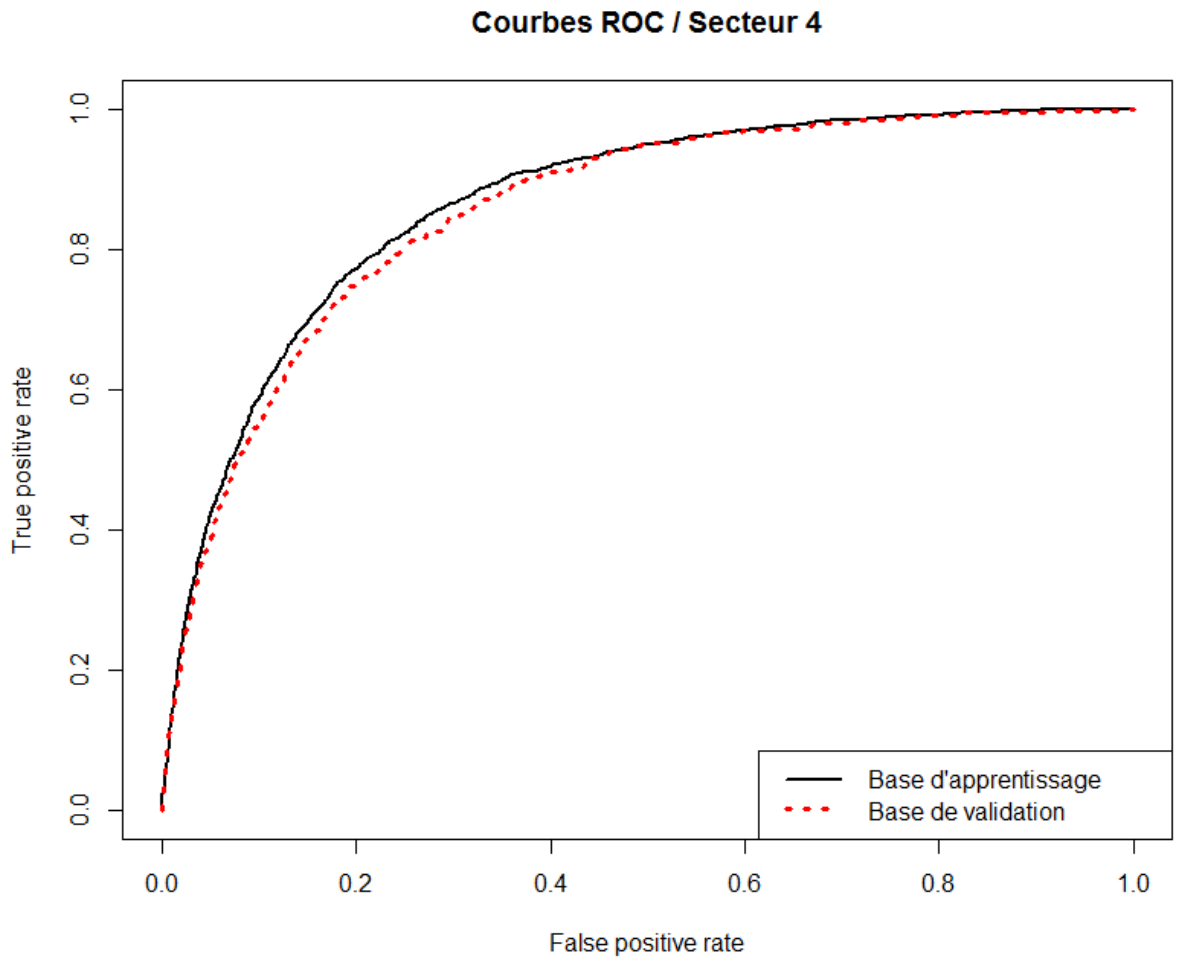

```
AUC : 0.8329844 (Base d'apprentissage)
AUC : 0.8332862 (Base de validation)
```

6.2.1.7 Score financier : estimations des paramètres sur le secteur 4

Extractions du logiciel R

```
Call:
glm(formula = def24 ~ ., family = binomial(link = "logit"), data = subset(train,
    select = c(def24, R2_t, R3_t, R4_t, R9_t, R20_t, R27_t, R40_t)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6997  -0.1424  -0.0806  -0.0469   4.1033

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.25504    0.14650 -29.045  < 2e-16 ***
R2_t(-0.2,0]     -0.36781    0.08885  -4.140 3.48e-05 ***
R2_t(0,0.1]      -0.60419    0.08665  -6.973 3.10e-12 ***
R2_t(0.1,0.3]    -1.10139    0.08454 -13.029  < 2e-16 ***
R2_t(0.3,0.45]   -1.59234    0.12283 -12.964  < 2e-16 ***
R2_t(0.45,1]     -1.88770    0.13969 -13.513  < 2e-16 ***
R3_t(0.05,0.48]   0.85353    0.08225  10.377  < 2e-16 ***
R3_t(0.48,2]      0.90952    0.11907   7.639 2.19e-14 ***
R4_t(0.075,0.15]  0.58617    0.14670   3.996 6.45e-05 ***
R4_t(0.15,0.3]    1.06877    0.12917   8.274  < 2e-16 ***
R4_t(0.3,2]       1.42044    0.12274  11.573  < 2e-16 ***
R9_t(20,40]      -0.23834    0.10280  -2.318 0.020425 *
R9_t(40,140]     -0.45987    0.11397  -4.035 5.46e-05 ***
R9_t(140,2e+03]  -1.36455    0.18023  -7.571 3.70e-14 ***
R20_t(0.1,0.3]   -0.24657    0.08408  -2.933 0.003362 **
R20_t(0.3,0.8]   -0.37940    0.11520  -3.293 0.000990 ***
R20_t(0.8,50]    -0.53745    0.15561  -3.454 0.000553 ***
R27_t(0.8,0.9]    0.31492    0.08910   3.534 0.000409 ***
R27_t(0.9,1]      0.37107    0.09285   3.996 6.43e-05 ***
R27_t(1,5]        0.64254    0.08147   7.886 3.11e-15 ***
R40_t(0.7,1]     -0.43846    0.10154  -4.318 1.57e-05 ***
R40_t(1,5]       -0.66895    0.11118  -6.017 1.78e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14948  on 117109  degrees of freedom
Residual deviance: 12101  on 117088  degrees of freedom
AIC: 12145

Number of Fisher Scoring iterations: 9
```

6.2.1.8 Score financier, secteur 4 : courbes ROC et AUC




```
AUC : 0.8642683 (Base d'apprentissage)
AUC : 0.8539737 (Base de validation)
```

6.2.2 Score final

6.2.2.1 Score final : regroupement des variables catégorielles

Code Variable	Âge du dirigeant	Sociétés saines	Sociétés défaillantes	Répartition	Taux de défaillance
age_cg07	18 à 24 ans	14 276	927	0,8%	6,5%
age_cg06	25 à 29 ans	57 343	2 724	3,4%	4,8%
age_cg05	30 à 34 ans	118 021	4 359	7,0%	3,7%
age_cg04	35 à 50 ans	621 522	16 871	36,8%	2,7%
age_cg03	50 à 65 ans	508 780	10 472	30,2%	2,1%
age_cg02	non renseigné	218 860	3 977	13,0%	1,8%
age_cg01	65 ans et plus	115 077	1 674	6,8%	1,5%
	Total	1 653 879	41 004	100,0%	2,5%

Code Variable	Capital Social	Sociétés saines	Sociétés défaillantes	Répartition	Taux de défaillance
capital_cg01	Moins de 7 k€	579 541	18 961	35,0%	3,3%
capital_cg02	De 7 k€ à 28 k€	722 274	16 425	43,7%	2,3%
capital_cg03	De 28 K€ à 280 k€	289 567	4 984	17,5%	1,7%
capital_cg04	280 k€ et plus	62 497	634	3,8%	1,0%
	Total	1 653 879	41 004	100,0%	2,5%

Code Variable	Département	Sociétés saines	Sociétés défaillantes	Répartition	Taux de défaillance
departement_cg08	Zone 08	31 881	1 268	1,9%	4,0%
departement_cg07	Zone 07	109 656	3 926	6,6%	3,6%
departement_cg06	Zone 06	155 398	5 051	9,4%	3,3%
departement_cg05	Zone 05	466 776	12 874	28,2%	2,8%
departement_cg04	Zone 04	363 881	9 028	22,0%	2,5%
departement_cg03	Zone 03	150 450	3 344	9,1%	2,2%
departement_cg02	Zone 02	29 022	510	1,8%	1,8%
departement_cg01	Zone 01	346 815	5 003	21,0%	1,4%
	Total	1 653 879	41 004	100,0%	2,5%

Code Variable	Ancienneté de l'entreprise	Sociétés saines	Sociétés défaillantes	Répartition	Taux de défaillance
anc_cg06	de 12 à 47 mois	396 266	16 503	24,0%	4,2%
anc_cg05	de 48 à 71 mois	218 828	6 688	13,2%	3,1%
anc_cg04	moins de 12 mois	96 834	2 544	5,9%	2,6%
anc_cg03	de 72 à 119 mois	329 090	7 040	19,9%	2,1%
anc_cg02	de 120 à 239 mois	350 339	5 304	21,2%	1,5%
anc_cg01	Plus de 240 mois	262 522	2 925	15,9%	1,1%
	Total	1 653 879	41 004	100,0%	2,5%

Code Variable	Effectif de l'entreprise	Sociétés saines	Sociétés défaillantes	Répartition	Taux de défaillance
eff_cg06	Aucun salarié	526 234	16 066	31,8%	3,1%
eff_cg05	1 salarié	199 565	4 915	12,1%	2,5%
eff_cg04	de 2 à 4 salariés	593 089	14 093	35,9%	2,4%
eff_cg03	de 5 à 19 salariés	275 399	4 943	16,7%	1,8%
eff_cg02	de 20 à 49 salariés	46 628	831	2,8%	1,8%
eff_cg01	de 50 à 2000 salarié	12 964	156	0,8%	1,2%
	Total	1 653 879	41 004	100,0%	2,5%

Code Variable	Catégorie juridique	Sociétés saines	Sociétés défaillantes	Répartition	Taux de défaillance
cj_cg01	Autre	75 496	253	4,6%	0,3%
cj_cg02	SA	22 651	269	1,4%	1,2%
cj_cg03	SAS	241 249	5 279	14,6%	2,2%
cj_cg04	SARL	1 314 483	35 203	79,5%	2,7%
	Total	1 653 879	41 004	100,0%	2,5%

Code Variable	Secteur de l'entreprise	Sociétés saines	Sociétés défaillantes	Répartition	Taux de défaillance
g01	Secteur 1	44 443	94	2,7%	0,2%
g02	Secteur 2	57 264	365	3,5%	0,6%
g03	Secteur 3	220 776	2 034	13,3%	0,9%
g04	Secteur 4	94 503	1 219	5,7%	1,3%
g05	Secteur 5	108 343	1 863	6,6%	1,7%
g06	Secteur 6	252 220	5 284	15,3%	2,1%
g07	Secteur 7	119 411	2 993	7,2%	2,5%
g08	Secteur 8	357 705	9 867	21,6%	2,8%
g09	Secteur 9	176 093	6 951	10,6%	3,9%
g10	Secteur 10	223 121	10 334	13,5%	4,6%
	Total	1 653 879	41 004	100,0%	2,5%

Code Variable	Notation financière la plus récente jusqu'à 2012	Sociétés saines	Sociétés défaillantes	Répartition	Taux de défaillance
scorefig01	01	273 428	314	16,5%	0,1%
scorefig02	02	197 646	871	12,0%	0,4%
scorefig03	03	79 384	631	4,8%	0,8%
scorefig04	04	110 847	1 276	6,7%	1,2%
scorefig05	05	70 932	1 131	4,3%	1,6%
scorefig06	06	62 675	1 311	3,8%	2,1%
scorefig07	07	34 215	910	2,1%	2,7%
scorefig08	08 et AB	663 800	22 742	40,1%	3,4%
scorefig09	09	64 917	3 010	3,9%	4,6%
scorefig10	10	25 081	1 582	1,5%	6,3%
scorefig11	11	22 242	1 658	1,3%	7,5%
scorefig12	12	11 081	953	0,7%	8,6%
scorefig13	13	18 751	1 839	1,1%	9,8%
scorefig14	14	6 801	859	0,4%	12,6%
scorefig15	15	5 612	789	0,3%	14,1%
scorefig16	16	4 430	718	0,3%	16,2%
scorefig17	17	1 314	250	0,1%	19,0%
scorefig18	18	723	160	0,0%	22,1%
	Total	1 653 879	41 004	100,0%	2,5%

AB: absence de bilan

6.2.2.2 Score final : estimation des paramètres

```
Call:
glm(formula = def ~ ., family = binomial(link = "logit"), data = subset(train,
    select = c(def24, scorefi, ape_c, anc_c, departement_c, age_c)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1706  -0.2564  -0.1622  -0.0790   4.0558

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -9.93277    0.14975 -66.329  < 2e-16 ***
scorefig02        1.15066    0.07830  14.695  < 2e-16 ***
scorefig03        1.61184    0.08276  19.477  < 2e-16 ***
scorefig04        2.20841    0.07429  29.727  < 2e-16 ***
scorefig05        2.41777    0.07547  32.037  < 2e-16 ***
scorefig06        2.59557    0.07447  34.855  < 2e-16 ***
scorefig07        2.81034    0.07809  35.990  < 2e-16 ***
scorefig08        3.25014    0.06663  48.779  < 2e-16 ***
scorefig09        3.37221    0.06987  48.266  < 2e-16 ***
scorefig10        3.74806    0.07325  51.169  < 2e-16 ***
scorefig11        3.87495    0.07316  52.963  < 2e-16 ***
scorefig12        4.08002    0.07830  52.105  < 2e-16 ***
scorefig13        4.19205    0.07272  57.645  < 2e-16 ***
scorefig14        4.34726    0.08059  53.946  < 2e-16 ***
scorefig15        4.57415    0.08122  56.319  < 2e-16 ***
scorefig16        4.66352    0.08269  56.394  < 2e-16 ***
scorefig17        4.77544    0.10995  43.434  < 2e-16 ***
scorefig18        4.87459    0.12974  37.572  < 2e-16 ***
ape_cg02          1.19724    0.14438   8.292  < 2e-16 ***
ape_cg03          1.53467    0.13231  11.599  < 2e-16 ***
ape_cg04          1.79098    0.13424  13.342  < 2e-16 ***
ape_cg05          2.08878    0.13256  15.758  < 2e-16 ***
ape_cg06          2.20491    0.13060  16.882  < 2e-16 ***
ape_cg07          2.31103    0.13150  17.575  < 2e-16 ***
ape_cg08          2.38474    0.13009  18.332  < 2e-16 ***
ape_cg09          2.60719    0.13037  19.998  < 2e-16 ***
ape_cg10          2.73721    0.13015  21.031  < 2e-16 ***
anc_cg02          0.14871    0.02904   5.120 3.05e-07 ***
anc_cg03          0.35311    0.02836  12.449  < 2e-16 ***
anc_cg04          0.09788    0.03581   2.733  0.00627 **
anc_cg05          0.65708    0.02883  22.792  < 2e-16 ***
anc_cg06          0.82793    0.02661  31.113  < 2e-16 ***
departement_cg02  0.31687    0.05835   5.431 5.61e-08 ***
departement_cg03  0.48051    0.02819  17.048  < 2e-16 ***
departement_cg04  0.59244    0.02255  26.272  < 2e-16 ***
departement_cg05  0.65006    0.02129  30.537  < 2e-16 ***
departement_cg06  0.66677    0.02555  26.099  < 2e-16 ***
departement_cg07  0.93978    0.02729  34.437  < 2e-16 ***
departement_cg08  1.08075    0.04066  26.578  < 2e-16 ***
age_cg02         -0.23030    0.03700  -6.225 4.83e-10 ***
age_cg03          0.07759    0.03301   2.350  0.01875 *
age_cg04          0.13307    0.03273   4.065 4.80e-05 ***
age_cg05          0.20548    0.03720   5.523 3.33e-08 ***
age_cg06          0.35917    0.04030   8.914  < 2e-16 ***
age_cg07          0.53724    0.05359  10.025  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 255564  on 1102585  degrees of freedom
Residual deviance: 222474  on 1102541  degrees of freedom
AIC: 222564
```

6.3 Modélisation de prime par le modèle collectif

6.3.1 Variables générées à partir des bases existantes

Nous disposons ou générons pour chaque année disponible des variables suivantes :

Variables relatives à la police de l'assuré :

Numéro de police, Siren de l'assuré, code et libellé NACE de l'assuré, code postal et région de l'assuré.

Date de prise d'effet de la police, de résiliation de la police, nous en déduisons l'exposition au contrat dans l'année pour prendre en compte une éventuelle censure.

Type de police, taux de la prime, limite de décaissement, prime actuelle.

Quotité garantie des agréments, des agréments express (NDS) et du « non dénommés ».

Limite de « non dénommés » et agrément express (NDS).

Chiffre d'affaires de l'assuré.

Variables relatives aux acheteurs :

Agrément maximum, taux d'acceptation des demandes d'agréments, nombre de demandes d'agréments refusées.

Montant cumulé, montant moyen, écart type et nombre d'agréments, d'agréments express (NDS) et des deux.

Pour les dix premiers acheteurs, le score, le secteur et le montant de l'agrément.

Le nombre d'acheteurs dont la limite est inférieure à 5k€, entre 5 et 10k€, entre 10 et 20k€, entre 20 et 50k€, entre 50 et 150k€, entre 150 et 450k€ et au-delà de 450k€ : nous en déduisons les proportions.

Variables relatives aux sinistres :

Le nombre de sinistre par police dans l'année et le montant associé.

6.3.2 Produit « Globale »

6.3.2.1 Modélisation de la fréquence

6.3.2.1.1 Régression de Poisson

```
glm(formula = NB_sinistres ~ +Région + secteur.ach1 + QG_ND +
    limite_NDS + score_mini_10_acheteurs + part_acheteur_sup_50 +
    log(total_agrement_et_NDS) + log(moyenne_agrement_et_NDS) +
    offset(log(censure)), family = poisson(link = "log"), data = dtrain_freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8274  -1.1493  -0.6832   0.0330  14.8813

Coefficients:
                              Estimate   Std. Error z value Pr(>|z|)
(Intercept)                  -3.074e+00  1.451e-01 -21.180  < 2e-16 ***
RégionAquitaine              -9.573e-01  1.610e-01  -5.944 2.78e-09 ***
RégionR1                     -8.735e-01  1.221e-01  -7.153 8.48e-13 ***
RégionBasse-Normandie         1.038e+00  1.176e-01   8.826  < 2e-16 ***
RégionR2                     -9.736e-02  9.701e-02  -1.004 0.315602
RégionR5                      1.503e-02  8.021e-02   0.187 0.851389
RégionChampagne              -3.019e-01  1.216e-01  -2.483 0.013029 *
RégionFranche-Comté          -7.745e-01  1.307e-01  -5.928 3.07e-09 ***
RégionHaute-Normandie        -3.838e-01  1.174e-01  -3.269 0.001078 **
RégionR4                     -1.227e-01  6.629e-02  -1.850 0.064251 .
RégionR3                      9.497e-02  7.477e-02   1.270 0.204013
RégionLanguegoc               1.617e-01  1.024e-01   1.579 0.114348
RégionPicardie               -3.549e-01  1.850e-01  -1.919 0.054995 .
RégionPoitou-Charente        -9.171e-02  1.119e-01  -0.820 0.412369
secteur.ach1Commerce_rep_auto -1.891e-01  4.686e-02  -4.035 5.47e-05 ***
secteur.ach1Construction      3.533e-02  5.299e-02   0.667 0.504906
secteur.ach1Manufacturier    -4.452e-01  5.328e-02  -8.355  < 2e-16 ***
QG_ND                         1.727e-02  5.927e-04  29.144  < 2e-16 ***
limite_NDS                    2.393e-05  6.612e-06   3.619 0.000295 ***
score_mini_10_acheteurs      -2.875e-02  1.337e-02  -2.151 0.031466 *
part_acheteur_sup_50         -2.024e+00  2.855e-01  -7.091 1.33e-12 ***
log(total_agrement_et_NDS)    5.231e-01  1.407e-02  37.173  < 2e-16 ***
log(moyenne_agrement_et_NDS) -3.146e-01  4.584e-02  -6.862 6.77e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10997.4  on 3023  degrees of freedom
Residual deviance:  7029.9  on 3001  degrees of freedom
AIC: 10016

Number of Fisher Scoring iterations: 6

       Overdispersion test

data:  regpoi1
z = 5.1015, p-value = 1.684e-07
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  4.057349
```

6.3.2.1.2 Régression de Quasi-Poisson

```
glm(formula = NB_sinistres ~ +Région + secteur.ach1 + QG_ND +
    limite_NDS + score_mini_10_acheteurs + part_acheteur_sup_50 +
    log(total_agrement_et_NDS) + log(moyenne_agrement_et_NDS) +
    offset(log(censure)), family = quasipoisson(link = "log"),
    data = dtrain_freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8274  -1.1493  -0.6832   0.0330  14.8813

Coefficients:
                              Estimate   Std. Error t value Pr(>|t|)
(Intercept)                  -3.074e+00  2.925e-01 -10.508  < 2e-16 ***
RégionAquitaine              -9.573e-01  3.246e-01  -2.949 0.003213 **
RégionR1                     -8.735e-01  2.461e-01  -3.549 0.000393 ***
RégionBasse-Normandie         1.038e+00  2.371e-01   4.379 1.23e-05 ***
RégionR2                     -9.736e-02  1.955e-01  -0.498 0.618608
RégionR5                      1.503e-02  1.617e-01   0.093 0.925952
RégionChampagne              -3.019e-01  2.451e-01  -1.232 0.218098
RégionFranche-Comté          -7.745e-01  2.634e-01  -2.941 0.003299 **
RégionHaute-Normandie        -3.838e-01  2.366e-01  -1.622 0.104919
RégionR4                     -1.227e-01  1.336e-01  -0.918 0.358672
RégionR3                      9.497e-02  1.507e-01   0.630 0.528629
RégionLanguegoc               1.617e-01  2.064e-01   0.783 0.433484
RégionPicardie               -3.549e-01  3.728e-01  -0.952 0.341166
RégionPoitou-Charente        -9.171e-02  2.255e-01  -0.407 0.684267
secteur.ach1Commerce_rep_auto -1.891e-01  9.445e-02  -2.002 0.045412 *
secteur.ach1Construction      3.533e-02  1.068e-01   0.331 0.740812
secteur.ach1Manufacturier    -4.452e-01  1.074e-01  -4.145 3.49e-05 ***
QG_ND                         1.727e-02  1.195e-03  14.459  < 2e-16 ***
limite_NDS                    2.393e-05  1.333e-05   1.796 0.072645 .
score_mini_10_acheteurs      -2.875e-02  2.694e-02  -1.067 0.285957
part_acheteur_sup_50         -2.024e+00  5.754e-01  -3.518 0.000442 ***
log(total_agrement_et_NDS)    5.231e-01  2.836e-02  18.442  < 2e-16 ***
log(moyenne_agrement_et_NDS) -3.146e-01  9.240e-02  -3.405 0.000671 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 4.062787)

    Null deviance: 10997.4  on 3023  degrees of freedom
Residual deviance:  7029.9  on 3001  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

6.3.2.1.3 Régression de Tweedie

```
glm(formula = NB_sinistres ~ +Région + secteur.ach1 + QG_ND +
    limite_NDS + score_mini_10_acheteurs + part_acheteur_sup_50 +
    log(total_agrement_et_NDS) + log(moyenne_agrement_et_NDS) +
    offset(log(censure)), family = tweedie(var.power = 1.5, link.power = 0),
    data = dtrain_freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7104  -1.7470  -1.2658   0.0347   9.2771

Coefficients:
                              Estimate   Std. Erro  t value Pr(>|t|)
(Intercept)                  -3.010e+00  3.051e-01  -9.866  < 2e-16 ***
RégionAquitaine              -1.248e+00  3.223e-01  -3.872 0.000110 ***
RégionR1                     -9.944e-01  2.280e-01  -4.360 1.34e-05 ***
RégionBasse-Normandie         8.150e-01  3.397e-01   2.400 0.016478 *
RégionR2                     -2.342e-01  2.073e-01  -1.130 0.258747
RégionR5                     -1.429e-01  1.928e-01  -0.741 0.458525
RégionChampagne              -3.087e-01  2.657e-01  -1.162 0.245462
RégionFranche-Comté          -7.983e-01  2.542e-01  -3.141 0.001701 **
RégionHaute-Normandie        -4.900e-01  2.523e-01  -1.942 0.052243 .
RégionR4                     -1.905e-01  1.536e-01  -1.240 0.215157
RégionR3                     -2.026e-01  1.776e-01  -1.140 0.254189
RégionLanguegoc              -8.143e-02  2.459e-01  -0.331 0.740556
RégionPicardie               -3.814e-01  3.554e-01  -1.073 0.283332
RégionPoitou-Charente        -5.209e-01  2.628e-01  -1.982 0.047614 *
secteur.ach1Commerce_rep_auto -1.850e-01  1.108e-01  -1.669 0.095177 .
secteur.ach1Construction      3.274e-02  1.236e-01   0.265 0.791047
secteur.ach1Manufacturier    -3.729e-01  1.180e-01  -3.160 0.001594 **
QG_ND                         2.105e-02  1.329e-03  15.840  < 2e-16 ***
limite_NDS                    4.225e-05  1.442e-05   2.929 0.003422 **
score_mini_10_acheteurs      -6.207e-03  3.057e-02  -0.203 0.839096
part_acheteur_sup_50         -1.739e+00  5.024e-01  -3.461 0.000546 ***
log(total_agrement_et_NDS)    5.223e-01  3.000e-02  17.408  < 2e-16 ***
log(moyenne_agrement_et_NDS) -4.260e-01  9.129e-02  -4.667 3.20e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 4.035168)

    Null deviance: 12146.5  on 3023  degrees of freedom
Residual deviance:  8486.1  on 3001  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

6.3.2.1.4 Régression Binomiale-Négative

```
glm(formula = NB_sinistres ~ +Région + secteur.ach1 + QG_ND +
    limite_NDS + score_mini_10_acheteurs + part_acheteur_sup_50 +
    log(total_agrement_et_NDS) + log(moyenne_agrement_et_NDS) +
    offset(log(censure)), family = negative.binomial(1), data = dtrain_freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2708  -0.9665  -0.5757   0.0181   5.4978

Coefficients:
                                Estimate   Std. Error t value Pr(>|t|)
(Intercept)                    -2.957e+00  3.026e-01  -9.770  < 2e-16 ***
RégionAquitaine                -1.275e+00  3.137e-01  -4.066 4.91e-05 ***
RégionR1                       -1.022e+00  2.200e-01  -4.648 3.50e-06 ***
RégionBasse-Normandie           7.814e-01  3.634e-01   2.150 0.031614 *
RégionR2                       -2.768e-01  1.968e-01  -1.407 0.159521
RégionR5                       -2.005e-01  1.871e-01  -1.071 0.284096
RégionChampagne                -4.226e-01  2.624e-01  -1.611 0.107385
RégionFranche-Comté            -8.483e-01  2.425e-01  -3.498 0.000476 ***
RégionHaute-Normandie          -5.437e-01  2.404e-01  -2.262 0.023771 *
RégionR4                       -2.296e-01  1.474e-01  -1.558 0.119328
RégionR3                       -3.082e-01  1.714e-01  -1.798 0.072295 .
RégionLanguegoc                -1.968e-01  2.363e-01  -0.833 0.404994
RégionPicardie                 -4.727e-01  3.492e-01  -1.354 0.175964
RégionPoitou-Charente          -6.070e-01  2.547e-01  -2.383 0.017216 *
secteur.ach1Commerce_rep_auto -1.739e-01  1.083e-01  -1.606 0.108482
secteur.ach1Construction        3.560e-02  1.197e-01   0.297 0.766170
secteur.ach1Manufacturier      -3.504e-01  1.148e-01  -3.054 0.002280 **
QG_ND                           2.166e-02  1.313e-03  16.497  < 2e-16 ***
limite_NDS                      3.847e-05  1.398e-05   2.752 0.005966 **
score_mini_10_acheteurs        -3.242e-03  2.967e-02  -0.109 0.912989
part_acheteur_sup_50           -1.672e+00  5.062e-01  -3.303 0.000968 ***
log(total_agrement_et_NDS)      5.200e-01  3.028e-02  17.175  < 2e-16 ***
log(moyenne_agrement_et_NDS)   -4.210e-01  8.970e-02  -4.693 2.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1) family taken to be 1.651929)

    Null deviance: 4652.7  on 3023  degrees of freedom
Residual deviance: 2956.6  on 3001  degrees of freedom
AIC: 7371.6

Number of Fisher Scoring iterations: 9
```

6.3.2.1.5 Comparaison des MSE et RMSE

Régression	MAE	MSE	RMSE
Poisson	1.254869	7.408319	2.721823
Quasi-Poisson	1.254869	7.408319	2.721823
Tweedie	1.280823	7.837418	2.799539
Binomiale Négative	1.292354	7.932429	2.816457

MAE : Mean Absolute Error (Erreur absolue moyenne), moyenne arithmétique des valeurs absolues des écarts.

MSE : Mean Square Error (Moyenne du carré des erreurs), c'est la moyenne arithmétique des carrés des écarts entre les prévisions et les observations.

RMSE : Racine carré du MSE, c'est l'erreur type.

6.3.2.2 Modélisation de l'intensité

```
lm(formula = LnMT ~ +secteur.ach1 + QG_agrement + QG_ND + limite_NDS +
    nb_NDS + log(moyenne_agrement_et_NDS) + ecart_type_agrement_et_NDS,
    data = dtrain_sin)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5145 -0.7129  0.0306  0.7680  3.6970

Coefficients:
                              Estimate   Std. Error t value Pr(>|t|)
(Intercept)                   5.780e+00  2.735e-01  21.136  < 2e-16 ***
secteur.ach1Commerce_rep_auto 1.186e-02  1.045e-01   0.113  0.90967
secteur.ach1Construction      3.525e-01  1.150e-01   3.066  0.00222 **
secteur.ach1Manufacturier     -1.264e-01  1.110e-01  -1.139  0.25511
QG_agrement                   5.123e-03  2.388e-03   2.145  0.03215 *
QG_ND                         -1.258e-02  1.220e-03 -10.316  < 2e-16 ***
limite_NDS                    5.062e-05  1.264e-05   4.006 6.61e-05 ***
nb_NDS                        -1.098e-04  4.779e-05  -2.298  0.02173 *
log(moyenne_agrement_et_NDS)  5.501e-01  7.243e-02   7.596 6.58e-14 ***
ecart_type_agrement_et_NDS    -3.946e-03  1.710e-03  -2.307  0.02123 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.156 on 1087 degrees of freedom
Multiple R-squared:  0.192,     Adjusted R-squared:  0.1853
F-statistic:  28.7 on 9 and 1087 DF,  p-value: < 2.2e-16

AIC(regln)
3443.529
```

6.3.3 Produit Global KUP

6.3.3.1 Modélisation de la fréquence

6.3.3.1.1 Régression de Poisson

```
glm(formula = NB_sinistres ~ log(Assure_CA_Global) + Région +
    secteur.ach1 + score_mini_10_acheteurs + part_acheteur_sup_50 +
    log(total_agrement_et_NDS) + offset(log(censure)), family = poisson(link = "log"),
    data = dtrain_freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8619  -0.8962  -0.5380  -0.0785   8.2203

Coefficients:
                                 Estimate  Std. Error z value Pr(>|z|)
(Intercept)                     -5.833807   0.346870 -16.818  < 2e-16 ***
log(Assure_CA_Global)            0.221894   0.042775   5.187 2.13e-07 ***
RégionAquitaine                  0.954246   0.139775   6.827 8.67e-12 ***
RégionBasse-Normandie            0.834147   0.363619   2.294  0.02179 *
RégionR2                         0.647050   0.117111   5.525 3.29e-08 ***
RégionIle-de-France              0.741145   0.096048   7.716 1.20e-14 ***
RégionProvence-Alpes-Côte d'Azur 0.861150   0.108033   7.971 1.57e-15 ***
RégionRhône-Alpes                0.335421   0.116052   2.890  0.00385 **
secteur.ach1Commerce_rep_auto    0.148558   0.091192   1.629  0.10330
secteur.ach1Construction        -0.014850   0.107650  -0.138  0.89028
secteur.ach1Manufacturier       -0.162121   0.106136  -1.527  0.12664
score_mini_10_acheteurs         -0.009178   0.027523  -0.333  0.73879
part_acheteur_sup_50            -8.666540   1.185803  -7.309 2.70e-13 ***
log(total_agrement_et_NDS)       0.465349   0.029042  16.023  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3120.0  on 1737  degrees of freedom
Residual deviance: 2252.6  on 1724  degrees of freedom
AIC: 3408.5

Number of Fisher Scoring iterations: 6

> plot(regpoi1$residuals)
> dispersiontest(regpoi1)

       Overdispersion test

data:  regpoi1
z = 4.3364, p-value = 7.241e-06
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
   2.50678
```

6.3.3.1.2 Régression de Quasi-Poisson

```
glm(formula = NB_sinistres ~ log(Assure_CA_Global) + Région +
    secteur.ach1 + score_mini_10_acheteurs + part_acheteur_sup_50 +
    log(total_agrement_et_NDS) + offset(log(censure)), family = quasipoisson(link =
"log"),
    data = dtrain_freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8619  -0.8962  -0.5380  -0.0785   8.2203

Coefficients:
                                 Estimate  Std. Error t value Pr(>|t|)
(Intercept)                      -5.833807   0.683425  -8.536  < 2e-16 ***
log(Assure_CA_Global)             0.221894   0.084278   2.633 0.008542 **
RégionAquitaine                   0.954246   0.275392   3.465 0.000543 ***
RégionBasse-Normandie             0.834147   0.716426   1.164 0.244456
RégionR2                          0.647050   0.230739   2.804 0.005100 **
RégionIle-de-France               0.741145   0.189240   3.916 9.34e-05 ***
RégionProvence-Alpes-Côte d'Azur  0.861150   0.212853   4.046 5.45e-05 ***
RégionRhône-Alpes                 0.335421   0.228654   1.467 0.142575
secteur.ach1Commerce_rep_auto     0.148558   0.179673   0.827 0.408452
secteur.ach1Construction         -0.014850   0.212099  -0.070 0.944191
secteur.ach1Manufacturier        -0.162121   0.209116  -0.775 0.438288
score_mini_10_acheteurs          -0.009178   0.054227  -0.169 0.865623
part_acheteur_sup_50             -8.666540   2.336344  -3.709 0.000214 ***
log(total_agrement_et_NDS)        0.465349   0.057221   8.132 7.95e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 3.881934)

    Null deviance: 3120.0  on 1737  degrees of freedom
Residual deviance: 2252.6  on 1724  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

6.3.3.1.3 Régression de Tweedie

```
glm(formula = NB_sinistres ~ log(Assure_CA_Global) + Région +
    secteur.ach1 + score_mini_10_acheteurs + part_acheteur_sup_50 +
    log(total_agrement_et_NDS) + offset(log(censure)), family = tweedie(var.power =
1.5,
    link.power = 0), data = dtrain_freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8200  -1.6333  -1.3028  -0.3048   7.5472

Coefficients:
                                 Estimate  Std. Error t value Pr(>|t|)
(Intercept)                      -4.80283    0.80144  -5.993 2.51e-09 ***
log(Assure_CA_Global)             0.15034    0.10506   1.431  0.15261
RégionAquitaine                   0.67711    0.41627   1.627  0.10400
RégionBasse-Normandie             0.74333    1.12473   0.661  0.50877
RégionR2                          0.56570    0.30196   1.873  0.06118 .
RégionIle-de-France               0.72896    0.23405   3.115  0.00187 **
RégionProvence-Alpes-Côte d'Azur  0.69012    0.28524   2.419  0.01565 *
RégionRhône-Alpes                 0.16313    0.27483   0.594  0.55288
secteur.ach1Commerce_rep_auto     0.16367    0.24332   0.673  0.50127
secteur.ach1Construction         -0.11605    0.27563  -0.421  0.67379
secteur.ach1Manufacturier        -0.37399    0.26623  -1.405  0.16028
score_mini_10_acheteurs           0.01834    0.07289   0.252  0.80136
part_acheteur_sup_50             -5.70320    1.42137  -4.012 6.27e-05 ***
log(total_agrement_et_NDS)        0.40787    0.07084   5.758 1.01e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 8.444973)

    Null deviance: 5456.8  on 1737  degrees of freedom
Residual deviance: 4377.9  on 1724  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 12
```

6.3.3.1.4 Régression binomiale négative

```
glm(formula = NB_sinistres ~ log(Assure_CA_Global) + Région +
    secteur.ach1 + score_mini_10_acheteurs + part_acheteur_sup_50 +
    log(total_agrement_et_NDS) + offset(log(censure)), family = negative.binomial(1),
    data = dtrain_freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8265  -0.8196  -0.5185  -0.0906   4.3623

Coefficients:
                                 Estimate  Std. Error t value Pr(>|t|)
(Intercept)                      -5.63454    0.73139  -7.704 2.21e-14 ***
log(Assure_CA_Global)             0.20536    0.09179   2.237 0.025399 *
RégionAquitaine                   0.73140    0.35381   2.067 0.038861 *
RégionBasse-Normandie             0.81866    0.89936   0.910 0.362808
RégionR2                          0.64862    0.25243   2.569 0.010269 *
RégionIle-de-France               0.75688    0.20245   3.739 0.000191 ***
RégionProvence-Alpes-Côte d'Azur  0.79070    0.23919   3.306 0.000967 ***
RégionRhône-Alpes                 0.24881    0.23779   1.046 0.295544
secteur.ach1Commerce_rep_auto     0.17910    0.20756   0.863 0.388317
secteur.ach1Construction         -0.04667    0.23571  -0.198 0.843088
secteur.ach1Manufacturier        -0.26404    0.22990  -1.148 0.250928
score_mini_10_acheteurs           0.01189    0.06280   0.189 0.849802
part_acheteur_sup_50             -7.91442    1.96914  -4.019 6.09e-05 ***
log(total_agrement_et_NDS)        0.45274    0.06395   7.080 2.09e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1) family taken to be 2.603311)

    Null deviance: 1894.4  on 1737  degrees of freedom
Residual deviance: 1376.3  on 1724  degrees of freedom
AIC: 3039.1

Number of Fisher Scoring iterations: 6
```

6.3.3.1.5 Comparaison des MSE et RMSE

Régression	MAE	MSE	RMSE
Poisson	0.634580	1.863336	1.365041
Quasi-Poisson	0.634580	1.863336	1.365041
Tweedie	0.665510	1.887164	1.373741
Binomiale Négative	0.642857	1.872643	1.369787

MAE : Mean Absolute Error (Erreur absolue moyenne), moyenne arithmétique des valeurs absolues des écarts.

MSE : Mean Square Error (Moyenne du carré des erreurs), c’est la moyenne arithmétique des carrés des écarts entre les prévisions et les observations.

RMSE : Racine carré du MSE, c’est l’erreur type.

6.3.3.2 Modélisation de l'intensité

```
lm(formula = LnMT ~ +secteur.ach1 + log(total_agrement_et_NDS) +
    part_acheteur_sup_50, data = dtrain_sin)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8623 -0.7467  0.0755  0.8177  3.7443

Coefficients:
                             Estimate  Std. Error t value Pr(>|t|)
(Intercept)                   8.01431    0.38303  20.924  < 2e-16 ***
secteur.ach1Commerce_rep_auto 0.16584    0.16808   0.987 0.324315
secteur.ach1Construction      0.62967    0.18492   3.405 0.000719 ***
secteur.ach1Manufacturier     0.29448    0.17828   1.652 0.099271 .
log(total_agrement_et_NDS)   -0.13721    0.04725  -2.904 0.003863 **
part_acheteur_sup_50          5.21671    1.09863   4.748 2.74e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.225 on 462 degrees of freedom
Multiple R-squared:  0.1015,     Adjusted R-squared:  0.09178
F-statistic: 10.44 on 5 and 462 DF,  p-value: 1.665e-09

AIC(regln)
1526.122
```