Mémoire présenté pour l'obtention du diplôme de
Master 2 de Droit, Economie et Gestion Mention Actuariat
et l'admission à l'Institut des Actuaires le 23/11/2021

Par : Ali GOUMAR

Titre : Validation du modèle de tarification et du modèle interne du besoin en capital de l'assurance agricole de Groupama.

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

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

Présidente :
Sandrine LEMERY
Membres présents du jury de l'Institut des Actuaires :
Edith BOCQUAIRE
Stéphane JASSON
Pierre PETAUTON
Membres présents du jury de la filière du CNAM :
Nathanaël ABECERA
Olivier DESMETTRE
David FAURE
François WEISS

Entreprise :
Nom : Groupama Assurances Mutuelles

Directeur de mémoire en entreprise :

Nom : Renaud BRUNEL
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

Signature du candidat

Secrétariat

Bibliothèque :

Table des matières

Remerciements

Je tiens d'abord à remercier Renaud BRUNEL, pour son suivi et son aide indispensable.

Je remercie Zakaria MOULIM, Valentine NAPOLEON, Youva MANSOUR et Lisa DIT THOME pour leurs aides tout au long de la rédaction de ce mémoire.

Je remercie Xavier AUBOUY pour sa relecture précieuse.

Et finalement, je remercie ma femme et mes enfants pour leurs soutiens.

J'ai une pensée toute particulière pour mon professeur au CNAM et au CEA, Monsieur Michel FROMENTEAU qui nous a quitté en 2017.

Résumé

Ce mémoire s'inscrit dans le cadre des travaux de validation des modèles à Groupama. Il porte notamment sur la revue du modèle de tarification et du modèle interne du besoin en capital du risque émanant de l'assurance agricole. Ce risque représente les pertes de rendements agricoles liées aux aléas climatiques que peuvent subir nos assurés. Que ce soit pour sa conception, son utilisation ou encore sa validation, le modèle interne du capital économique est soumis aux exigences réglementaires de Solvabilité 2. En revanche le modèle tarifaire est exempté de celles-ci sauf dans notre cas. En effet le modèle tarifaire fournit des données et des hypothèses en entrée du modèle interne et finalement, il est en adhérence avec ce dernier. Ces modèles font aussi l'objet de contrôles spécifiques et ponctuels par l'ACPR pour vérifier leurs pertinence et prudence.

Au vu de l'importance de cette activité dans la stratégie du Groupe Groupama, nous avons décidé de réaliser une validation indépendante des deux modèles y compris les changements opérés dans le cadre du plan d'action pour répondre aux observations de l'ACPR. Cette revue interne utilisera comme socle les normes et les exigences de Solvabilité 2 en termes de la validation du modèle interne. Nous avons aussi proposé une extension vers des modèles alternatifs pour challenger et s'assurer de l'adéquation entre le modèle et son objectif.

Puisque ce risque est particulier et nécessite des modèles spécifiques, plusieurs défis sont à relever pour l'accomplissement de ces travaux de validation. En effet, ce risque est particulier et nécessite des modèles spécifiques. L'enjeu est de sécuriser nos dirigeants sur la confiance que nous pouvons avoir sur l'utilisation de ces modèles et de mettre en œuvre un processus vertueux d'amélioration continue de ces derniers.

Mots clés : assurance agricole, modèle interne, GLM, tests statistiques, prime pure, SCR, risque de prime, validation, DLM

Abstract

This thesis is about the validation of the models used at Groupama, in particular the review of the pricing model, and the internal model for the risk capital requirement for agricultural insurance. This risk represents the loss of agricultural yields due to climatic hazards that policyholders may experience. The internal economic capital model is subject to the Solvency 2 regulatory requirements for its design, use and above all its validation. The pricing model is usually exempt from these requirements, but not in this case. The pricing model provides data and assumptions as input to the internal model and so it is in line with this model. These models are also subject to specific and punctual controls by the ACPR to verify their relevance and prudence.

Given the importance of this activity in the Groupama Group's strategy, it was decided to carry out an independent validation of the two models, including the changes made as part of the action plan to respond to the ACPR's observations. This internal review uses the Solvency 2 standards and requirements as a basis for the internal model validation. Also proposed are alternative models to challenge and ensure the adequacy between the model and its objective.

There are several challenges in carrying out this validation work, to reassure managers about the confidence in the use of these models and to implement a process of continuous improvement of them.

Keywords : agricultural insurance, internal model, GLM, statistical tests, pure premium, SCR, premium risk, validation, DLM

Introduction

Toutes les entreprises d'assurance ou de réassurance ayant opté pour l'utilisation d'un modèle interne à des fins d'évaluation de leur SCR dans la cadre de la réglementation Solvabilité 2 doivent décrire dans une politique les modalités de gouvernance du dit modèle. En application de cette exigence réglementaire, Groupama a mis en place un dispositif garantissant de manière continue le bon fonctionnement et la bonne utilisation de son modèle interne partiel non-vie. Conformément aux dispositions de l'article 115 de la Directive 2009/138/EC, celui-ci décrit en particulier l'ensemble des processus liés aux modifications apportées au modèle interne partiel, en distinguant précisément les modifications mineures et majeures. Il définit également les processus et le cadre de gouvernance nécessaires à la validation indépendante du modèle.

L'assurance agricole, et plus particulièrement sa composante multirisques climatiques (MRC) lancée en 2005, constitue une part importante du portefeuille et du profil du risque de Groupama : elle est au cœur de l'identité du Groupe et de sa stratégie.

À ce titre, ce segment constitue une des principales spécificités du modèle interne partiel de Groupama. L'assurance MRC protège nos sociétaires agriculteurs contre la perte de rendement due aux aléas climatiques (grêle, gel, sécheresse, etc…). Groupama a réalisé un nouveau modèle en 2018 pour la tarification de ce risque atypique marqué par une sinistralité assez dégradée ces dernières années, avec notamment une fréquence importante d'aléas climatiques majeurs.

Les contrôles réalisés par l'ACPR[1], respectivement sur le modèle interne en 2019, et sur la modélisation tarifaire MRC, ont donné lieu à des observations. Dans ce mémoire nous proposons une revue de validation approfondie des modélisations en MRC (tarifs et risques), en phase avec les normes et les exigences en termes de validation de modèle attendues dans le cadre de solvabilité 2.

Le processus de validation proposé est global : il intègre la revue des données, hypothèses, méthodes et résultats. Dans ce mémoire, la partie revue de la gouvernance n'est pas traitée et la revue de la qualité des données est limitée aux données utilisées pour le calibrage. C'est à l'issue de ce processus que les constats et recommandations de la validation sont résumés. Ils sont ensuite communiqués au comité de gouvernance des modèles afin de prendre des décisions quant aux futures évolutions de la modélisation et de l'utilisation de ces modèles.

Après un rappel du contexte de notre étude et un partage de quelques chiffres clés de l'assurance agricole en France, nous présenterons les deux modèles de Groupama dans une première partie. De fait, le modèle interne et le modèle tarifaire seront exposés -en portant une attention particulière à la modélisation du risque multirisque climatique sur récoltes-avant de décrire notre problématique puis d'y apporter des éléments de réponses.

S'en suivront, une troisième et une quatrième partie dédiée à chaque modèle Groupama où seront présentés : la modélisation du risque de primes pour le risque MRC, le plan de validation proposé ainsi que leur exécution avant d'aboutir à la synthèse des constats et recommandations visant à les améliorer respectivement.

En conclusion, nous donnerons des perspectives quant aux deux modèles, expliquerons la nécessité du renfoncement de leur interaction ainsi que leur cohérence par rapport à leurs objectifs respectifs.

[1] ACPR : Autorité de Contrôle Prudentiel et de Résolution

1. Contexte de l'étude

L'assurance agricole permet de protéger les agriculteurs contre les pertes de rendement dues aux événements climatiques. Les dernières années ont été marquées par des événements climatiques plus fréquents, liés à des périls différents : sécheresse en 2015, 2018, 2019 et 2020, inondations en 2016, gel en 2017, gel et intempéries en 2020, orages de grêle localisés et gel en 2021.

Elle permet de couvrir l'ensemble des risques climatiques. Deux types de contrats sont proposées par les assureurs : le contrat par culture et le contrat à l'exploitation. Le premier garantit une indemnisation pour chaque nature de récolte assurée (exemple : « blé d'hiver ») dès que la perte de rendement constatée suite à un sinistre pour cette nature de récolte est supérieure au seuil de déclenchement. Le deuxième garantit une indemnisation seulement si le total des pertes constatées pour l'ensemble des natures de récolte assurées, suite à un sinistre, est supérieur au seuil de déclenchement. Il y a mutualisation, au sein de l'exploitation, entre les différentes natures de récolte assurées, les gains sur une nature de récolte pouvant compenser les pertes sur une autre nature de récolte. Ces contrats sont moins onéreux que les contrats « par culture ».

Afin de favoriser le développement de l'assurance récolte, la France a décidé de mobiliser le Fonds européen agricole pour le développement rural (FEADER) pour financer jusqu'à 65 % du montant de la cotisation d'assurance correspondant au 1er niveau de garantie (niveau socle) et jusqu'à 45 % du montant de la cotisation correspondant au 2ème niveau (garanties complémentaires optionnelles). L'agriculteur supporte ainsi une partie limitée du montant du contrat d'assurance qu'il souscrit.

La Fédération Française de l'Assurance a publié dans son rapport « l'assurance agricole en 2019 » l'évolution du marché de l'assurance MRC sur cultures en France.

Exercice	Nombre de contrats (milliers)	Cotisations		Prime moyenne (€)	Montant des indemnités (K€)	Rapport S/P (%)
		Montant (K€)	*Variation (%)*			
2005	57.2	81 176	///	1 420	65 753	81
2006	63.9	90 776	*11.8*	1 420	88 053	97
2007	61.4	107 571	*18.5*	1 753	139 842	130
2008	62.9	160 244	*49*	2 549	91 339	57
2009	65.3	153 304	*-4.3*	2 347	133 374	87
2010	73.6	158 656	*3.5*	2 155	126 192	80
2011	76.7	212 099	*33.7*	2 764	219 523	104
2012	75.9	235 030	*10.8*	3 096	211 523	90
2013	75.2	267 771	*13.9*	3 563	341 401	127
2014	74.8	280 601	*4.8*	3 750	174 774	62
2015	67.1	249 639	*-11*	3 720	171 220	69
2016	64.8	249 799	*0.1*	3 855	576 555	231
2017	70.3	285 984	*14.5*	4 065	308 185	108
2018	68.5	320 743	*12.2*	4 680	291 176	91
2019	**69.1**	**340 870**	***6.3***	**4 935**	**406 364**	**119**

Figure 1 - Evolution du marché de l'assurance MRC sur cultures en France (Source : Rapport FFA "L'assurance agricole en 2019"

Les cotisations de l'assurance MRC proviennent essentiellement de contrats « par culture » (95 % des cotisations en 2019). Les contrats « à l'exploitation » représentent 3 999 contrats pour environ 16,6 millions d'euros de cotisations. Ces dernières concernent pour la majorité les vignes (60 % des cotisations en 2019) et les prairies (25 %). La grêle représente 14 % des indemnités MRC de 2019, soit 59 millions d'euros.

Groupama est le principal partenaire des exploitants agricoles en cas d'événement climatique :

- Plus de 60 000 exploitants agricoles assurés,
- 60 % de part de marché.

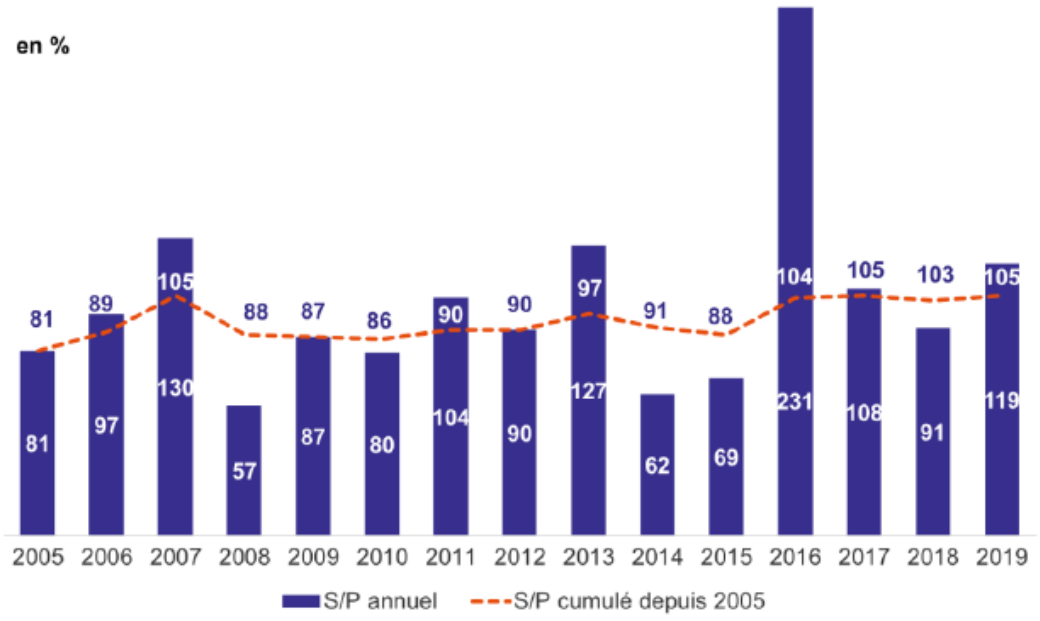


Figure 2 - Ratio sinistres à primes de l'assurance MRC (Source : Rapport FFA « L'assurance agricole en 2019 »)

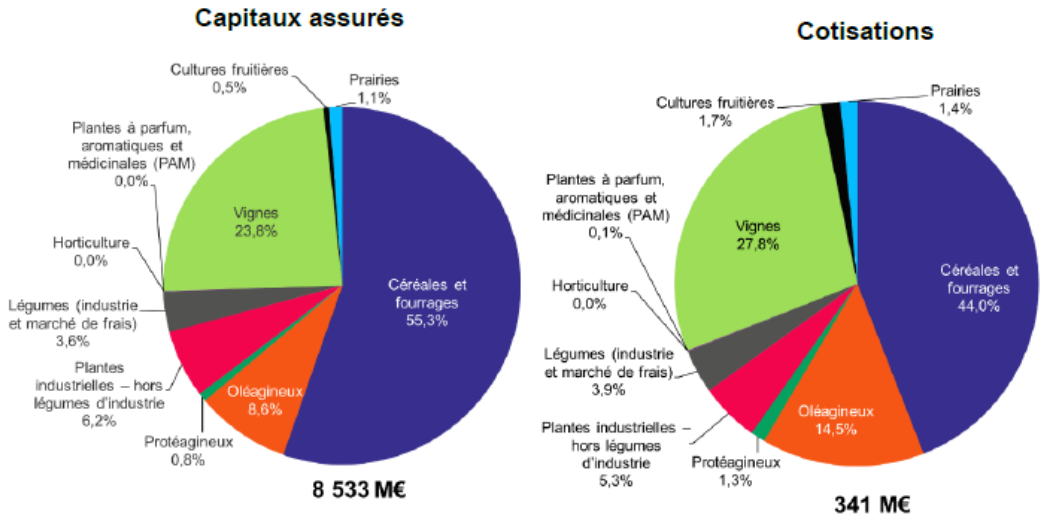


Figure 3 - Répartition des capitaux assurés et des cotisations en 2019 (Source : Rapport FFA « L'assurance agricole en 2019 »)

Depuis le lancement du contrat MRC en 2005, la France a connu des conditions climatiques très hétérogènes et surtout éloignées de la « normale ».

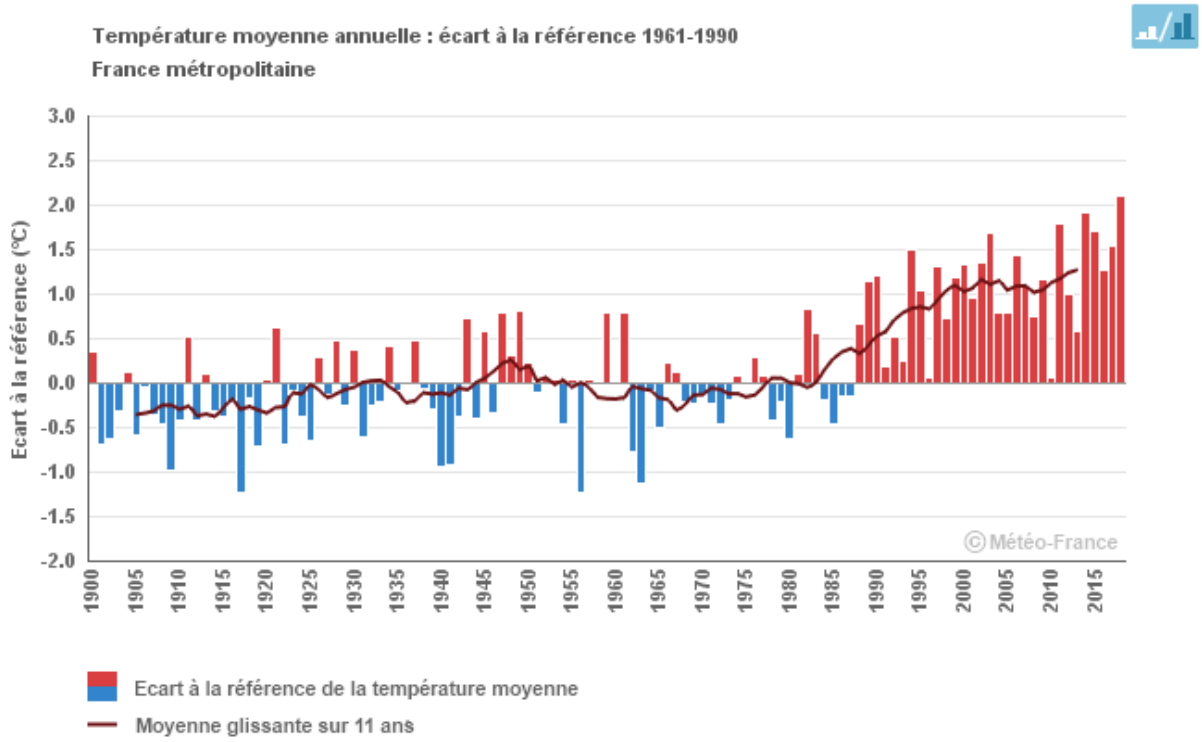


Figure 4 - Evolution des températures moyennes annuelles en France métropolitaine (Source : Météo France)

Nous remarquons sur le graphique ci-dessus une augmentation de la température moyenne annuelle ces dernières années dont les plus élevées ont été observées en 2011, 2014 et 2018.

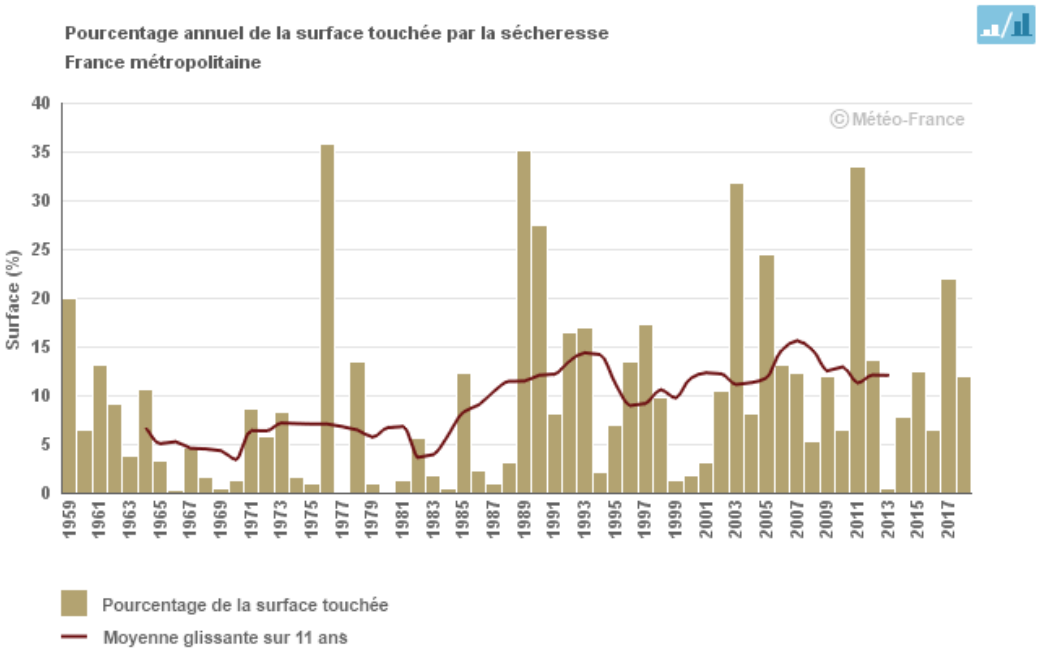
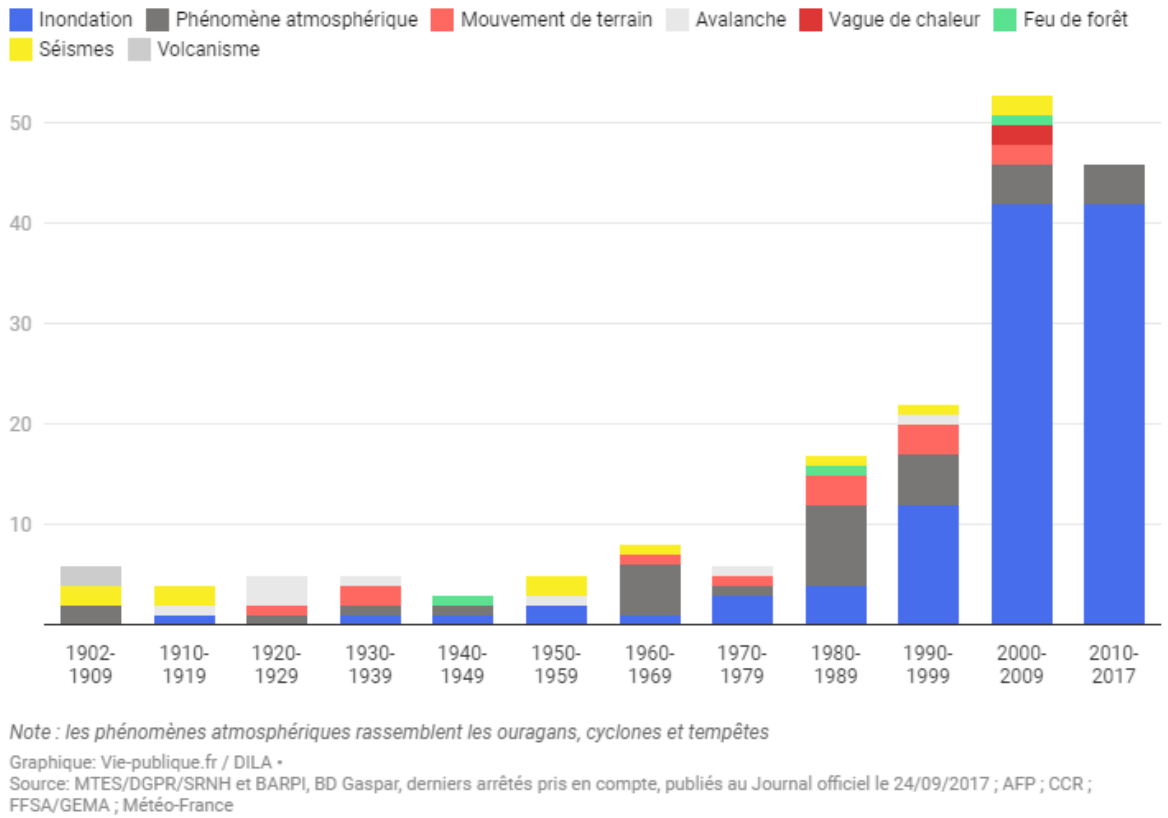


Figure 5 - Analyse du pourcentage annuel de la surface touchée par la sécheresse des sols

L’analyse du pourcentage annuel de la surface touchée par la sécheresse des sols depuis 1959 permet d’identifier les années ayant connu les événements les plus sévères comme 1976, 1989, 2003 et 2011.



Note : les phénomènes atmosphériques rassemblent les ouragans, cyclones et tempêtes

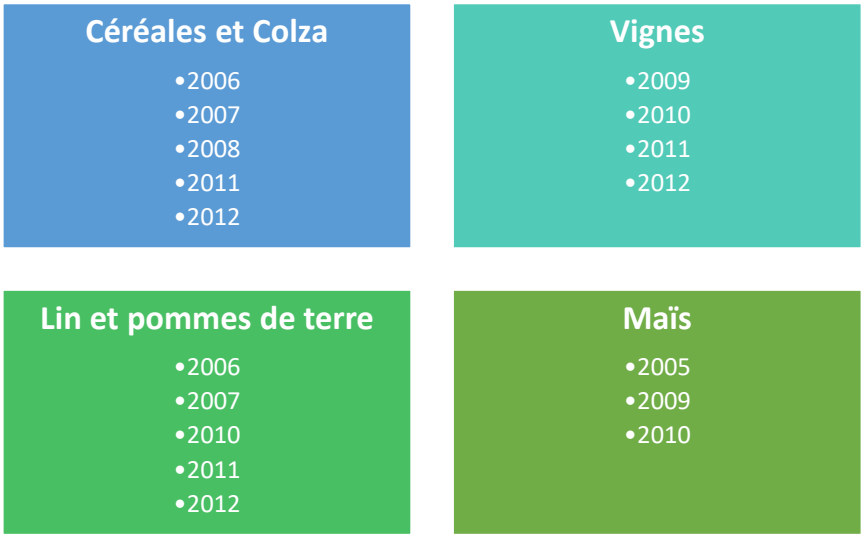
Graphique: Vie-publique.fr / DILA •
Source: MTES/DGPR/SRNH et BARPI, BD Gaspar, derniers arrêtés pris en compte, publiés au Journal officiel le 24/09/2017 ; AFP ; CCR ; FFSA/GEMA ; Météo-France

Figure 6 - Nombre d'événements dommageables recensés en France

Nous remarquons que le nombre d’inondations a fortement augmenté au cours des années. La cause de cette augmentation est dû à l’accroissement de l’urbanisation dans les zones inondables.

D’après le GIEC (Groupe d’Experts Intergouvernemental sur l’évolution du Climat), le réchauffement des océans et de l’atmosphère pourrait faire augmenter le nombre et l’étendu des événements climatiques extrêmes (tempêtes, inondations, sécheresses, …)

Avant 2016, la succession des autres aléas a engendré des années de pertes de rendement qui se résument pour les principales cultures comme suit :



L'année 2016 était une année de sinistralité exceptionnelle pour notre Groupe, visiblement sur une évolution dans la nature du risque, non seulement par l'augmentation des occurrences de paramètres climatiques « anormaux » affectant le développement des cultures, mais également sur des événements d'ampleur mais très localisés due à plusieurs facteurs :

- L'excès d'eau : les conditions de météo pluvieuses sur l'ensemble du mois de mai, et plus particulièrement les fortes précipitations cumulées entre le 27 et 30 Mai, ont eu des conséquences catastrophiques sur les céréales d'hiver (blé tendre, orge).

- Le manque de rayonnement solaire a eu une incidence sur la croissance des cultures d'hiver qui étaient au stade floraison et/ou au stade de fécondation : baisse de fertilité de l'épi, verse ou encore difficulté d'absorption de l'azote par la plante et avortement des grains. La charge totale des provisions liée aux excès d'eau et au manque de rayonnement est de 212,23 millions d'euros

- Le gel, observé au mois d'Avril, a entraîné des pertes dans le nord de la France sur les vignobles avec des dégâts à des stades très jeunes (bourgeons) et sur grandes cultures alors en méiose. Le cout total provisionné au titre de cet évènement est de 26,76 millions d'euros

- Deux événements grêle majeurs sont ensuite survenus en mai : Dans les départements de l'Yonne, la Charente, du Gers et des Pyrénées Atlantiques. Ce deuxième épisode a causé des dégâts dans les vignobles de Chablis, du Beaujolais, du Madiran, de Pacherenc, de Cognac, mais aussi sur les grandes cultures d'hiver (blé, orge et colza). Le coût total provisionné au titre de cet évènement à mi-août est de 69,26 millions d'euros.

1.1. Présentation du modèle interne de Groupama

1.1.1. Présentation du modèle interne MRC

L'objectif du modèle interne en MRC est de déterminer le capital économique (perte bicentenaire) correspondant au risque de la perte de rendements agricoles due aux autres aléas climatiques. En effet après un calibrage d'une loi de probabilités à partir des données historiques des rendements et des primes individuelles, le modèle simule les pertes liées aux aléas climatiques des agriculteurs assurés. Cela permet de déduire une distribution empirique des indemnisations à payer par Groupama dans un horizon d'un an.

L'indemnisation d'un assuré dépend de plusieurs critères/facteurs : du capital assuré, du rendement de référence et du prix moyen déclarés dans le contrat d'assurance. Elle tient aussi compte de la franchise et du rendement réel constaté à la fin de la saison après l'événement climatique.

Les rendements du secteur agricole sont impactés par les progrès technologiques (mécanisation, chimie, sélection génétique des espèces végétales, …) qui se manifestent par une tendance en termes d'amélioration des rendements jusqu'à une certaine date, où ces progrès marquent le pas et conduisent à une stabilisation des rendements. Cette deuxième lorsqu'elle existe est appelée plateau de rendements et l'année (date) de séparation est marquée comme une année de rupture entre ces deux phases.

La perte de rendement annuelle est modélisée par :

$$\tilde{Y} = R^{Max} - \tilde{R}$$

Où R^{Max} et $\tilde{R}$ expriment respectivement le rendement maximum théorique et la variable aléatoire du rendement.

Ce modèle s'appuie sur les hypothèses suivantes :

- $\tilde{Y}$ suit une distribution log-normale de moyenne μ et d'écart-type σ ;

L'estimation des paramètres (μ, σ) par région et par culture des lois de perte de rendement s'appuie sur la base AGRESTE[2], comme suit :

- Les rendements moyens des couples (caisse régionale, type de culture), déterminés comme la moyenne des rendements pondérés par les surfaces exposées depuis l'année de rupture retenue (s'il existe un plateau)
- Les coefficients de variation des couples (caisse régionale, type de culture), obtenus par le calcul des résidus résultant de la régression linéaire effectuée de chaque côté de l'année de rupture.

Contrairement au rendement moyen qui se base uniquement sur le jeu de données disponibles après l'année de stabilisation du rendement, le coefficient de variation se base sur l'ensemble du jeu de données.

Par ailleurs, la dépendance entre les variables de perte de rendement régionales est modélisée par une copule gaussienne. Celle-ci est entièrement définie par la matrice de corrélations des couples région x culture. Les historiques AGRESTE des rendements avec correction de tendance permettent de calibrer cette matrice.

[2] Base AGRESTE : données des rendements historiques de l'agriculture en France préparées par le service statistique ministériel de l'agriculture (https://agreste.agriculture.gouv.fr/)

1.1.1.1. Calibrage au niveau individuel

Nous définissons la variable de perte de rendement $\tilde{Y}_i$ au niveau d'un agriculteur i pour une culture c dans une région r. Pour un agriculteur i dans une région r pour la culture c, la modélisation de la perte se fait par $\tilde{Y}_i = \tilde{Y} \cdot \tilde{\varepsilon}_i$, où $\tilde{\varepsilon}_i$ est supposée indépendante de $\tilde{Y}$.

Il est fait l'hypothèse que les variables $\tilde{Y}_i$ et $\tilde{\varepsilon}_i$ ont la même distribution pour tous les agriculteurs pour un couple région × culture donné.

$$\mathbb{E}\left[\tilde{Y}_i \middle| \tilde{Y}\right] = \tilde{Y} \Rightarrow \mathbb{E}[\tilde{Y}_i] = \mathbb{E}[\tilde{Y}] = \mu \Rightarrow \mathbb{E}[\tilde{\varepsilon}_i] = 1$$

$$\mathbb{V}ar[\tilde{Y}_i] = \mathbb{V}ar[\tilde{Y} \cdot \tilde{\varepsilon}_i] = \mathbb{E}[\tilde{Y}^2] \cdot \mathbb{E}\left[{\tilde{\varepsilon}_i}^2\right] - \mathbb{E}^2[\tilde{Y} \cdot \tilde{\varepsilon}_i]$$

$$\Rightarrow \sigma_i^2 = (\mu^2 + \sigma^2) \cdot \mathbb{E}\left[{\tilde{\varepsilon}_i}^2\right] - \mu^2$$

$$\Rightarrow \mathbb{E}\left[{\tilde{\varepsilon}_i}^2\right] = \frac{\mu^2 + \sigma_i^2}{\mu^2 + \sigma^2}$$

Et

$$\mathbb{V}ar[\tilde{\varepsilon}_i] = \frac{1 + \frac{{\sigma_i}^2}{\mu^2}}{1 + \frac{\sigma^2}{\mu^2}} - 1 = \frac{1 + \left(CV_i \times \frac{\mathbb{E}[\tilde{R}_i]}{{R_i}^{Max} - \mathbb{E}[\tilde{R}_i]}\right)^2}{1 + \left(CV_{cr} \times \frac{\mathbb{E}[\tilde{R}_i]}{{R_i}^{Max} - \mathbb{E}[\tilde{R}_i]}\right)^2} - 1$$

Où CV_{cr} et CV_i expriment respectivement le coefficient de variation par culture c × région r et le coefficient de variation d'un agriculteur i.

La quantité $V[\tilde{\varepsilon}_i]$ est entièrement déterminée par :

- La moyenne et l'écart-type (μ, σ) de la perte de rendement régional,
- La moyenne et l'écart-type (μ, σ_i), de la perte de rendement individuel.

Ainsi :

- Le paramètre de moyenne μ est déterminé par la moyenne des rendements observés sur la base AGRESTE pour la région et la culture pertinente ;
- L'écart-type σ_i est alors déterminé par une notion d'inertie intra-classe, de sorte à retrouver le coefficient de variation individuel qui permet d'obtenir la prime pure issue du modèle de tarification.

1.1.1.2. Calcul de la prime pure

Pour un individu dont la loi de perte de rendement individuel est $(\tilde{Y}_i)$, la sinistralité espérée à charge de l'assureur pour chaque assuré s'exprime comme :

$$Prime\ Pure = \mathbb{E}(\tilde{S}_i) = P \cdot V_i \cdot \mathbb{E}\left[\left(\tilde{Y}_i - K_i\right)^+\right]$$

Avec

$$K_i = R_i^{Max} - (1 - \alpha_i) \cdot R_i^{\mathrm{Ref}}$$

Où P et V_i sont respectivement le prix de référence de la culture et la surface cultivée, et R_i^{Ref} le rendement de référence de chaque agriculteur i.

On a également :

$$\mathbb{E}\left[\left(\tilde{Y}_i - K_i\right)^+\right] = P\left(\tilde{Y}_i > K_i\right) \cdot \mathbb{E}\left(\tilde{Y}_i - K_i \middle| \tilde{Y}_i > K_i\right)$$

Le terme d'espérance conditionnelle dans le membre de droite est le *Mean Excess Loss* d'une loi log-normale de paramètres *m* et *s²* et qui peut s'écrire[3] :

$$e(x) \equiv \mathbb{E}\left(\tilde{Y}_i - x \middle| \tilde{Y}_i > x\right) = \frac{exp\left(m + \frac{s^2}{2}\right) \cdot \left\{1 - \Phi\left(\frac{\ln x - m - s^2}{s}\right)\right\}}{\left\{1 - \Phi\left(\frac{\ln x - m}{s}\right)\right\}} - x$$

En remarquant que $exp\left(m + \frac{s^2}{2}\right) = \mathbb{E}\left(\tilde{Y}_i\right)$, $1 - \Phi\left(\frac{\ln x - m}{s}\right) = P\left(\tilde{Y}_i > K_i\right)$, on obtient :

$$\mathbb{E}\left[\left(\tilde{Y}_i - K_i\right)^+\right] = \mathbb{E}\left(\tilde{Y}_i\right) \cdot \left[1 - \Phi\left(\frac{\ln(K_i) - m_i - s_i^2}{s_i}\right)\right] - K_i \cdot P\left(\tilde{Y}_i > K_i\right)$$

Où Φ est la fonction de répartition d'une loi normale centrée réduite et (m, s_i) sont les paramètres de la loi de perte de rendement individuel $(\tilde{Y}_i)$ reliés à la moyenne μ et à l'écart-type σ_i par les relations usuelles :

$$m_i = \ln \mu - \frac{1}{2}\ln\left(1 + \frac{s_i^2}{\mu^2}\right) \qquad \text{et} \qquad s_i^2 = \ln\left(1 + \frac{\sigma_i^2}{\mu^2}\right)$$

Ce montant théorique de la prime pure sera utilisé pour calibrer le paramètre de volatilité comme cela a été expliqué précédemment. Pour résoudre cette équation, un algorithme itératif incrémente le coefficient Rmax/Rmoy (avec un pas de 0,05) jusqu'à permettre de retrouver une solution à l'équation.

La méthodologie de Groupama est synthétisée dans l'organigramme ci-dessous :

[3] Voir par exemple http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/tutorials/stfhtmlnode85.html

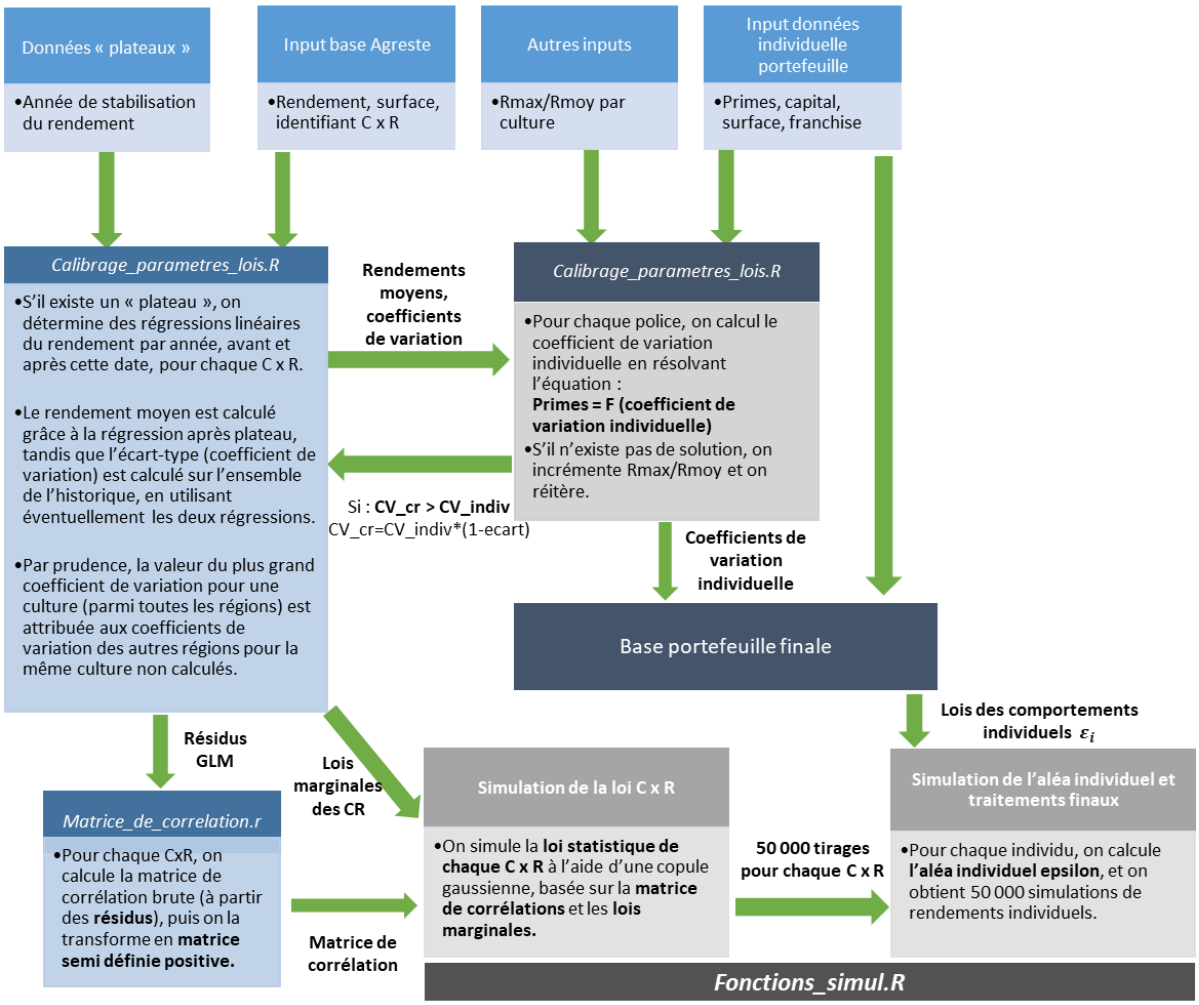


Figure 7 - Méthodologie de Groupama

1.1.2. Présentation du modèle tarifaire MRC

L'objectif d'un modèle de tarification est de segmenter les risques en catégories tarifables reflétant bien la prime pure. Cette dernière est déterminée pour atteindre l'équilibre technique sur le produit modulo les chargements appropriés. Nous cherchons une vision fine en espérance de la sinistralité et du risque de chaque segment dans notre portefeuille.

Pour tarifier ses produits MRC, Groupama utilise un modèle statistique basé sur les modèles GLM (Generalized Linear Model) qui est une généralisation de la Régression Linéaire. Ce type de modèle (GLM), permet d'analyser une variable cible à travers des variables explicatives et une loi adaptée. A la sortie du GLM on retrouve des coefficients correspondants à chaque modalité de chaque variable. Ces coefficients-là représentent une estimation de l'impact de chaque modalité sur la variable cible.

Le modèle utilisé par Groupama pour la modélisation de la prime pure est basé sur la formule suivante :

$$\boldsymbol{Prime_{Culture\backslash commune} = P0_{Culture} \times Coef_{Culture\backslash Commune} \times Capital_{assuré}}$$

Exemple pour la culture blé pour la commune 01001 :

$$\boldsymbol{Prime_{blé\backslash 01001} = P0_{blé} \times Coef_{blé\backslash 01001} \times Capital_{assuré}}$$

Avec :

$\boldsymbol{Prime_{blé\backslash 01001}}$: La prime pour un assuré qui a une culture blé dans la commune 01001.

$\boldsymbol{P0_{blé}}$ **:** Le taux de prime pure de référence pour la culture blé permettant l'équilibre primes et indemnisations sur toute la période de l'étude.

$\boldsymbol{Capital_{assuré}}$ **:** Le capital assuré pour l'agriculteur.

$\boldsymbol{Coef_{blé\backslash 01001}}$ **:** Le coefficient GLM pour la culture blé dans la commune 01001.

Le calcul de ce coefficient GLM est réalisé par la formule ci-dessous :

$$\boldsymbol{Coef_{culture\backslash commune} = \frac{p_\beta(x)}{1 - p_\beta(x)} = EXP(X'\beta)}$$

Avec :

$\boldsymbol{X'}$: Les variables explicatives

$\boldsymbol{\beta}$ **:** L'estimateur du modèle

$\boldsymbol{p_\beta(x)}$**: La** probabilité de sinistres pour une culture donnée dans une commune donnée modélisée par un modèle de régression logistique

Ce coefficient représente l'odds ratio (OR) ou le rapport des cotes des probabilités d'avoir le sinistre pour ceux qui ont les caractéristiques X d'une part et de la probabilité de ne pas avoir le sinistre pour ces mêmes individus d'autre part. Si :

• OR =1, la sinistralité est indépendante des caractéristiques X

• OR>1, la sinistralité est plus fréquente pour les individus qui ont les caractéristiques X.

• OR <1, la sinistralité est moins fréquente pour les individus qui ont les caractéristiques X.

Nous allons détailler le calcul de ce coefficient et le modèle tarifaire de Groupama pour les contrats MRC dans la partie 4 de ce mémoire.

1.2. Présentation de la problématique

L'assurance agricole est toujours en recherche de l'équilibre dans un environnement où les aléas climatiques sont de plus en plus fréquents et intenses : quand ce ne sont pas les précipitations fortes ce sont les sécheresses ou les gels qui prennent le relais au niveau de la sinistralité. Cette incertitude constitue un vrai défi pour les assureurs pour la tarification et le calcul du besoin en capital de cette assurance spécifique.

Le taux de pénétration de cette assurance reste modéré malgré les aides et subventions de l'Etat français ou de l'Union européenne. La structure ainsi que les frontières entre les couvertures des assureurs et de l'Etat sont en réflexion, elles pourront impacter l'équilibre et la rentabilité des assureurs sur ce marché. La réassurance est aussi un enjeu majeur pour atténuer ce risque, de fait elle constitue un pilier important dans sa couverture.

Dans ce contexte incertain, Groupama a refondu en 2018 le modèle utilisé pour la tarification de l'assurance multirisques climatiques autres aléas (hors grêle). Les résultats issus de ce modèle de tarification sont utilisés pour alimenter le modèle interne partiel, afin de calculer le besoin en capital sur ce même risque. Au sein du modèle tarifaire nous cherchons une vision fine du comportement de chaque segment en espérance : dans le modèle interne nous voulons plutôt capter le comportement de l'ensemble du portefeuille dans les extrêmes de la distribution. Un modèle de tarification ne peut pas être simplement un cas d'utilisation d'un modèle de capital économique.

Dans le cadre de Solvabilité 2 et vu l'importance stratégique en tant que leader sur le marché de l'assurance agricole, Groupama a décidé de valider avec le principe des 4 yeux les deux modèles de tarification et du besoin en capital. Malgré la différence de leurs objectifs, la démarche de validation peut être similaire, et c'est bien l'enjeu de ce mémoire.

	Objectifs	Hypothèses	Périmètre	Inputs	Outputs
Modèle Tarifaire	Calcul des primes pures	1) La sinistralité du portefeuille Groupama est la même que celle observée nationalement pour le retraitement de la P0. 2) Les variations de rendement observées à la maille nationale dans la base Agreste ne sont dues qu'à des aléas climatiques différents de la grêle. 3) Les pertes de rendement suivent une loi lognormale. 4) Le modèle économique fonctionne à la maille commune. 5) 28 cultures sont modélisées. 6) Les contrats multi culture par exploitation ne sont pas modélisés, mais ceux-ci sont démembrés par culture et intégrés aux contrats mono culture.	National	Base AGRESTE (depuis 1950 pour le blé et 1989 pour les autres cultures). Sinistralité Groupama depuis 2008. Données climatiques depuis 1979.	Primes pures par cultures/commune
Modèle Interne Partiel	Simuler une distribution de charge des sinistres brute à la maille Groupe (Caisse régionale métropolitaine, CR/Culture) et à la maille contrat/culture. La prise en compte de la réassurance est faite au niveau de Rex avec les autres garanties.	1) La dépendance entre les variables de pertes de rendement régionales est modélisée par une copule gaussienne. 2) Au lieu de travailler par commune comme dans la tarification, le modèle économique fonctionne au niveau de la région (plus précisément de la caisse régionale). 3) Seuls les rendements régionaux corrèlent les agriculteurs entre eux. 4) 28 cultures sont modélisées ; 5) Les contrats multi culture par exploitation ne sont pas modélisés, mais ceux-ci sont démembrés par culture et intégrés aux contrats mono culture. 6) les distributions de pertes suivent la loi lognormale. 7) Les rendements de référence utilisés dans le modèle économique correspondent aux espérances des rendements régionaux calibrées .	Caisses régionales	Primes pures, Franchises, Capitaux assurés, Coeffs Rmax/Rmoy, historique des rendements régionaux de la base AGRESTE depuis 1989, la base DISAR	50000 simulations de la charge des sinistres, à la maille (Globale, Caisses régionales et Caisses/Cultures)

Figure 8 : Tableau comparatif des modèles tarification/Capital Economique

Notre problématique est de définir un processus de validation adéquat et conforme aux exigences réglementaires. L'objectif est d'élaborer un plan de tests cohérent afin de vérifier la pertinence des méthodes et modèles utilisés, la qualité des données, mais aussi la mise en œuvre pratique ainsi que la gouvernance de ces deux modèles.

Tout au long de ce mémoire, nous allons proposer des solutions pour répondre à notre problématique et nous présenterons les résultats de l'exécution de notre plan de tests aux différentes composantes des deux modèles. Nous en déduirons ensuite des constats et des recommandations qui seront communiqués et partagés dans des instances dirigeantes de Groupama.

Enfin, ces travaux permettront de sécuriser nos dirigeants sur la confiance que nous pouvons avoir sur l'utilisation de ces modèles et appuiera la mise en œuvre d'un processus vertueux d'amélioration continue de ces derniers.

2. Outils mathématiques

Dans ce chapitre nous allons présenter les outils actuariels utilisés pour la réalisation des travaux de notre validation des modèles interne et tarifaire de la multirisque climatique. Pour plus de détails des démonstrations sont disponibles en annexes 1 à 4.

2.1. Présentation du modèle de régression linéaire

La régression linéaire est une méthode statistique visant à analyser la relation (association) entre une variable dite **variable dépendante** qui est la variable étudiée (à expliquer) et une (régression linéaire simple) ou plusieurs variables (régression linéaire multiple) dites **variables indépendantes (explicatives).** Dans notre cas, nous allons s'intéresser à une régression linéaire simple.

L'équation du modèle de régression linéaire s'écrit comme suit :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Où :

- Y_i : La variable dépendante (variable aléatoire)
- β_0, β_1 : Les coefficients (ordonnée à l'origine et pente)
- X_i : La variable indépendante (variable explicative), où i = 1, ..., n, les indices des observations contenues dans l'échantillon
- ε_i : Une erreur aléatoire

Le but est d'estimer les paramètres β_0 et β_1. Dans notre cas, Y est le rendement et X est l'année.

2.1.1. Hypothèses de modèle

Les hypothèses retenues sont les suivantes :

- Le modèle est linéaire en X_i
- Les valeurs de X_i sont observées sans erreur (X_i non aléatoire)
- $E(\varepsilon_i) = 0$, le modèle est bien spécifié donc l'erreur moyenne est nulle
- $Var(\varepsilon_i) = \sigma_\varepsilon^2$, la variance de l'erreur est constante.
- $Cov(\varepsilon_i, \varepsilon_j) = 0$, les erreurs sont non corrélées (ou indépendantes). Une erreur à l'instant t n'a pas d'influence sur les erreurs suivantes.
- $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.
- $Cov(\varepsilon_i, X_i) = 0$

2.1.2. Estimation des paramètres de modèle

Pour estimer les paramètres β_0 et β_1[4], nous appliquons la méthode des Moindres Carrés Ordinaires qui consiste à minimiser la somme des carrées des erreurs.

$$Min \sum_{i=1}^{n} {\varepsilon_i}^2 = Min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

[4] Voir annexe 1.

2.1.3. Estimation de la variance de l'erreur

L'estimateur de la variance de l'erreur (σ_ε^2) noté $\widehat{\sigma 2_\varepsilon}$ est donc égal à :

$\widehat{\sigma 2_\varepsilon} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$ avec $e_i = y_i - \hat{y}_i$

2.1.4. Qualité d'ajustement des modèles

Nous définissons[5] le R² par :

$$R^2 = \frac{SCE}{SCT}$$

Ce dernier donne la part de la variabilité totale de Y expliquée par X.

2.1.5. Test de significativité de modèle

Le test de significativité consiste à vérifier si la variable explicative est pertinente dans le modèle ou non. L'hypothèse nulle correspond à la situation où la variable X n'explique pas la variable d'intérêt (Y) Le test s'écrit :

$$H0 : \beta_i = 0 \qquad VS \qquad H1 : \beta_i \neq 0$$

Pour vérifier la significativité de la variable, nous utilisons les résultats de la régression effectuée dans R et nous comparons la p-value de la variable avec le seuil de risque 5%. Si la p-value est supérieure à 5%, nous acceptons le test (la variable n'est pas significative), sinon la variable est significative.

2.1.6. Vérification des hypothèses de modèle de régression

Dans cette partie, nous allons expliquer le principe des tests des trois principales hypothèses de la régression, à savoir les hypothèses de :

- Normalité des résidus,
- Autocorrélation des résidus
- Hétéroscédasticité des résidus.

2.1.6.1. Test de normalité des résidus

Le test de Shapiro-Wilk[6], permet de savoir si une série de données suit une loi normale. Le test s'écrit :

H0 : La série suit une loi normale. VS H1 : La série ne suit pas une loi normale.

[5] Voir démonstration annexe 2

[6] Les hypothèses de normalité seront vérifiées à l'aide du test de Shapiro-Wilk. Si nombre d'observation inférieur à 50 (N<=50), sinon nous utilisons le test de Kolmogorov-Smirnov (N>50).

- Si la p-value du test est inférieure au niveau alpha choisi (5% par exemple), alors l'hypothèse nulle est rejetée (les résidus ne suivent pas la loi normale).
- Si la p-value est supérieure au niveau alpha choisi (5% par exemple), alors on ne doit pas rejeter l'hypothèse nulle (les résidus suivent bien la loi normale).

2.1.6.2. Test d'hétéroscédasticité

Nous utilisons le test de BREUSCH-PAGAN à l'aide du logiciel R. Le test s'écrit comme suit :

H0 : Homoscédasticité (les résidus ont tous la même variance)

VS

H1 : Hétéroscédasticité (les résidus n'ont pas tous la même variance).

- Si la p-value est inférieure au niveau alpha choisi (5% par exemple), alors l'hypothèse nulle est rejetée (les résidus n'ont pas tous la même variance).
- Si la p-value est supérieure au niveau alpha choisi (5% par exemple), alors on ne doit pas rejeter l'hypothèse nulle (les résidus ont tous la même variance).

2.1.6.3. Analyse de l'autocorrélation des résidus

Nous utilisons le test de DURBIN WATSON pour analyser l'autocorrélation des résidus. Il s'agit du test suivant :

H0 : Absence d'autocorrélation. VS H1 : Présence d'autocorrélation.

L'interprétation est identique aux tests précédents.

2.2. Modèle linéaire généralisé (GLM)

2.2.1. Préambule

Les modèles linéaires généralisés ou **Generalized Linear Models (GLM)** sont une extension des modèles linéaires classiques permettant d'étudier la liaison entre une variable dépendante Y et un ensemble de variables explicatives X_1, …, X_P. Parmi ces modèles, nous pouvons citer :

- Le modèle linéaire général (la régression multiple, analyse de la variance, …).
- La régression logistique.
- La régression de Poisson.
- Le modèle Gaussien.
- Le modèle Gamma.

Dans le modèle GLM, on suppose que $(Y/X=x_i)$ suit une loi de la famille exponentielle. Comme toutes les lois de la famille exponentielle, sa densité s'écrit sous la forme suivante :

$$f_{X/x_i}(\theta, \emptyset, y) = exp\left(\frac{y\theta - b(\theta)}{a(\emptyset)} + c(y, \emptyset)\right)$$

Avec :

- a(.), b(.) et c(.) des fonctions dérivables,
- b(.) deux fois dérivable,
- b'(.) inversible.

Nous disposons alors de deux égalités sur les premiers moments :

$$\begin{cases} \boldsymbol{E(Y) = b'(\theta)} \\ \boldsymbol{Var(Y) = b''(\theta) \times a(\emptyset)} \end{cases}$$

Nous remarquons que la moyenne est en fonction de θ et la variance du couple $(\theta, \emptyset)$. Notons que dans les modèles de régression simple, nous étudions $E\left({}^{Y}/_{X = x_i}\right) = X_i\beta$. Cependant, dans les modèles linéaires généralisés, nous ajoutons une transformation de l'espérance conditionnelle via une « fonction de lien ». Nous obtenons alors :

$$\boldsymbol{g(E\left({}^{Y}/_{X = x_i}\right)) = X_i\beta.}$$

Où g(.) est la fonction de lien, inversible ($g^{-1}existe$) et X_i le vecteur ligne des variables explicatives.

Dans notre cas, nous effectuerons notre étude en utilisant le modèle LOGIT que nous allons détailler dans la partie suivante.

2.2.2. Principaux modèles GLM

2.2.2.1. Le modèle logistique

Le modèle logistique est utilisé lorsqu'on veut expliquer une variable Y prenant les valeurs 0 si « échec » ou 1 si « succès » par une ou plusieurs variables $X_1, X_2, \ldots X_P$.

Par exemple :

- Dans le cas de la modélisation du risque de défaut, Y vaut 1 si un emprunteur fait défaut, 0 sinon. Les variables explicatives peuvent être l'âge, la profession, le statut matrimonial, le fait d'être ou non propriétaire…
- Pour la modélisation de l'occurrence de sinistres, Y vaut 1 si l'assuré va faire un sinistre 0 sinon. La variable X donne par exemple l'âge, le sexe, la région, …

Lorsque la variable dépendante est dichotomique l'application de la régression linéaire n'est pas appropriée. En effet, dans le cas où la variable dépendante ne prend que la modalité 0 ou 1 l'application de la régression linéaire pose les problèmes suivants :

Soit $\boldsymbol{Y_i = X_i'\beta + U_i}$ le modèle de la régression linéaire.

Comme $\boldsymbol{Y_i}$ ne peut prendre que deux modalités 0 et 1, il en est de même de la perturbation $\boldsymbol{U_i}$. En effet, $\boldsymbol{U_i}$ prendra la valeur $\boldsymbol{1 - X_i'\beta}$ (si $\boldsymbol{Y_i = 1}$) ou la valeur $\boldsymbol{-X_i'\beta}$ (si $\boldsymbol{Y_i = 0}$). Donc nous ne pouvons pas dire que les résidus suivent une loi normale.

Nous avons alors : $\boldsymbol{U_i} = \begin{cases} \boldsymbol{1 - X_i'\beta} \text{ avec une probabilité } \boldsymbol{P_i = P(Y_i = 1)}. \\ \boldsymbol{-X_i'\beta} \text{ avec une probabilité } \boldsymbol{1 - P_i = P(Y_i = 0)}. \end{cases}$

Si nous nous posons dans le cadre de la régression linéaire, l'espérance de U_i devrait être nulle telle que :

$$\boldsymbol{E(U_i) = 0}$$

$$\Leftrightarrow \boldsymbol{E(U_i) = P_i\, x\, (1 - X_i'\beta) + (1 - P_i)\, x(-X_i'\beta) = 0}$$

$$\Leftrightarrow \boldsymbol{P_i = X_i'\beta}$$

Donc $\boldsymbol{X_i'\beta}$ correspond à une probabilité et doit satisfaire certaines propriétés : $\boldsymbol{X_i'\beta}$ doit être compris entre 0 et 1.

Mais nous ne sommes pas sûrs de trouver un estimateur *B* qui puisse satisfaire cette propriété.

C'est pour toutes ces raisons que l'utilisation de la régression linéaire ; dans le cas d'une variable dépendante dichotomique, n'est pas appropriée et que nous utilisons à la place la régression logistique. Pour mieux constater inadéquation pour modéliser correctement la variable endogène dichotomique, considérons un modèle linéaire avec une seule variable explicative notée $\boldsymbol{x}$, et une constante.

Posons β = (β_0 β_1), et considérons le modèle linéaire suivant :

$$\boldsymbol{y_i = \beta_0 + x_i\beta_1 + \varepsilon_i \quad \forall i = 1, \dots, N}$$

Il suffit de se placer dans un repère (x, y) et de reproduire les N différents couples $(x_i, y_i), \forall i = 1, \dots, N$.

Du fait du statut dichotomique de la variable endogène, le nuage de points ainsi obtenu se situe soit sur la droite $y = 0$, soit sur la parallèle $y = 1$:

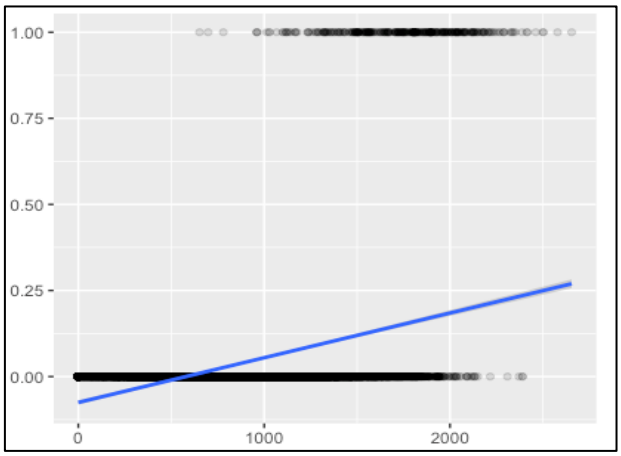


Figure 9 - Graphe d'une modélisation par régression linéaire simple pour une variable dichotomique expliquée par une seule variable

Il est impossible d'ajuster de façon satisfaisante, par une seule droite, le nuage de points, associé à une variable dichotomique qui, par nature, est réparti sur deux droites parallèles.

<u>Spécificités du modèle logistique :</u>

Dans cette modélisation, nous cherchons à modéliser la probabilité d'avoir un « succès ». Autrement dit :

$$\boldsymbol{P(Y = 1 | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p).}$$

Notons :

$$\boldsymbol{\pi(x) = P(Y = 1|X_1 = x_1, \ldots, X_p = x_p)}$$

Et supposons que :

$$\boldsymbol{\pi(x) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}$$

Où F est une fonction de répartition inversible, tel que :

- $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_p}$ des paramètres inconnus à estimer ;
- $\boldsymbol{F(t) = \frac{exp(t)}{1+exp(t)}}$

Le modèle logistique dans le cas d'une seule variable explicative, avec constante, s'écrit comme suit :

$$\boldsymbol{\pi(x_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}}$$

Graphiquement, cette fonction prend la forme suivante :

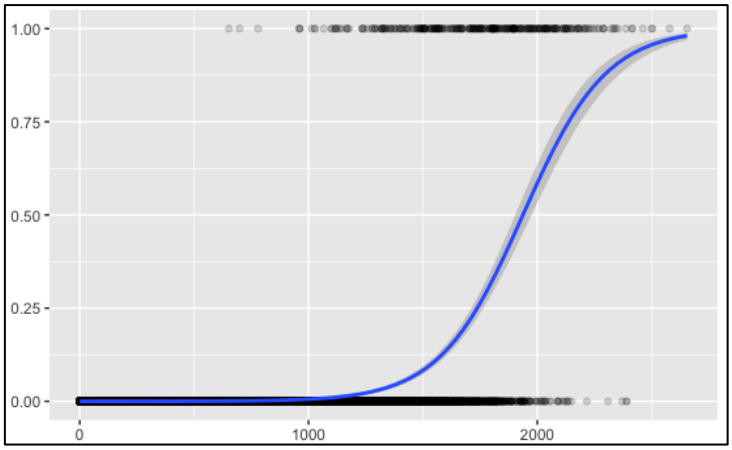


Figure 10 - Graphe d'une modélisation par régression logistique pour une variable dichotomique expliquée par une seule variable

Si nous avons plusieurs variables, l'équation de la fonction logistique est :

$$\boldsymbol{\pi(x_i) = \frac{exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p})}{(1 + exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p}))}}$$

Interprétation des coefficients :

Une fois le modèle estimé (la méthode d'estimation sera présentée plus tard dans le chapitre), nous obtenons des valeurs pour les paramètres β, qu'il faut interpréter.
En effet, l'effet marginal de la j[ème] variable explicative est défini par :

$$\boldsymbol{\frac{\delta F(X_i \beta)}{\delta X_{ij}} = \beta_j * f(X_i \beta)}$$

Avec $\boldsymbol{f(X_i\beta)}$ une densité de probabilité qui est positive.

De ce fait, le signe de l'effet marginal de la j[ème] variable pour l'individu i X_{ij} dépend du paramètre $\boldsymbol{\beta_j}$:

- Si $\boldsymbol{\beta_j > 0}$, $\boldsymbol{X_{ij}}$ a un effet positif sur l'évènement considéré.
- Si $\boldsymbol{\beta_j < 0}$, $\boldsymbol{X_{ij}}$ a un effet négatif sur l'évènement considéré.

2.2.2.2. Le modèle Gaussien

Le modèle Gaussien est utilisé dans le cas où la variable à expliquer est continue et appartient à $\mathbb{R}$. La densité de la famille des lois $\mathrm{N}(\mu_i, \sigma^2)$ s'écrit :

$$\boldsymbol{f(\mu_i, y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\}}$$

La loi gaussienne, de moyenne μ_i et de variance σ^2, appartient à la famille exponentielle, avec :

$$\boldsymbol{\theta = \mu_i, \emptyset = \sigma^2, a(\emptyset) = \emptyset, b(\theta) = \frac{\theta^2}{2} et\ c(y, \emptyset) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + 2\log(2\pi\sigma^2)\right), y\ \epsilon\mathbb{R}}$$

Exemple :

Modélisation du taux de chômage en fonction des variables explicatives macroéconomique tels que le taux de croissance, taux d'inflation, …

Dans cette modélisation nous sommes face à des nombres réels (positifs et inferieurs à 1).

Nous pouvons donc penser à utiliser le modèle gaussien.

2.2.2.3. Le modèle Gamma

Le modèle Gamma est utilisé lorsque la variable à expliquer appartient à $\mathbb{R}^+$. La densité d'une loi Gamma de paramètres υ et λ s'écrit :

$$\boldsymbol{f(y) = \frac{\lambda}{\Gamma(\upsilon)}(\lambda y)^{\upsilon-1} e^{-\lambda y}, y \geq 0}$$

Cette loi appartient à la famille exponentielle, avec :

$\boldsymbol{\theta = \frac{-1}{\mu}\ et\ \mu = \frac{\upsilon}{\lambda}, \emptyset = \frac{1}{\upsilon}}$,

$\boldsymbol{a(\emptyset) = \emptyset = \frac{1}{\upsilon}, b(\theta) = \ln(\mu)\ et\ c(y, \emptyset) = \upsilon\ln(\upsilon y) - \ln(y) - \ln\left(\Gamma(\upsilon)\right), y\ \epsilon\mathbb{R}^+}$.

Exemple :

Modélisation de la dépense pour la consommation d'un bien

Dans cette modélisation, nous sommes face à une variable dépendante appartenant à $\mathbb{R}^+$.

2.2.3. Estimation des paramètres dans le cadre du GLM

Pour estimer les paramètres d'une régression GLM, on utilise la méthode du maximum de vraisemblance. La vraisemblance s'écrit :

$$\boldsymbol{L(\beta) = \prod_{i=1}^{i=N} f(yi(\beta), \theta i, \emptyset i)}$$

Et la log-vraisemblance vaut :

$$\boldsymbol{log\ (L(\beta)) = \sum_{i=1}^{i=N} log(f(yi(\beta), \theta i, \emptyset i))}$$

Nous dérivons ensuite la log-vraisemblance par rapport à β_j pour obtenir les équations suivantes :

$$\boldsymbol{\frac{\partial \log\ (L(\beta))}{\partial \beta_j} = 0 \ \ \forall\, j = 1 \ldots . N}$$

Ces équations ne sont pas toujours linéaires nous n'avons donc pas d'estimateur explicite de $\boldsymbol{\beta}$. Pour résoudre ces équations, nous faisons appels à des méthodes d'estimations numériques et notamment celle de Newton-Raphson.

L'estimateur $\boldsymbol{\widehat{\beta}}$ de $\boldsymbol{\beta}$ est un estimateur du maximum de vraisemblance ayant les qualités suivantes :

- Il est asymptotiquement sans biais.
- Il a une variance minimale.
- Il converge asymptotiquement en loi vers la loi normale.

2.2.4. Test dans le cadre du GLM

Quatre tests sont généralement utilisés pour tester la significativité des variables explicatives et l'adéquation du modèle :

- Le test de Wald (significativité des variables).
- Le test du rapport de vraisemblance (significativité des variables et adéquation du modèle).
- Le test de la déviance (adéquation du modèle).
- Le test d'adéquation du modèle par diagramme Quantile-Quantile.

2.2.4.1. Test de Wald

Le test de Wald sert spécifiquement à tester la nullité d'un ou plusieurs coefficients sauf la constante. Dans le cas où nous voulons tester la significativité d'un seul coefficient c'est-à-dire :

$$\boldsymbol{H0: \beta_j = 0 \quad VS \quad H1: \beta_j \neq 0,}$$

Le test de Wald s'écrit :

$$\boldsymbol{W = \frac{\beta^2}{V(\beta)} \rightarrow \chi_1^2.}$$

Par ailleurs, si nous souhaitons tester la significativité de plusieurs coefficients :

$$\boldsymbol{H0 : \beta_1 = \cdots = \beta_k = 0 \quad VS \quad H1 : \exists\, i\, tel\ que\ \beta_i \neq 0},$$

Le test s'écrit comme suit :

$$\boldsymbol{W = \hat{\beta}\, V(\hat{\beta})\hat{\beta} \rightarrow \chi_k^2}$$

Avec **k** le degré de liberté :

k = nombre d'observations – le nombre des paramètres à estimer.

2.2.4.2. Test du rapport de vraisemblance

La statistique de test est basée sur la différence des rapports de vraisemblance entre le modèle complet et le modèle sous **H0 :**

$$\boldsymbol{H0 : \beta_j = 0 \quad ou \quad H0 : \beta_1 = \cdots = \beta_k = 0}$$

Notons $\boldsymbol{\hat{\beta}_{H0}}$ l'estimateur du maximum de vraisemblance contraint par H0.

Nous avons alors, sous H0 :

$$\boldsymbol{2\left(\log\left(L(\hat{\beta})\right) - \log\left(L(\hat{\beta}_{H0})\right)\right) \rightarrow \chi_p^2}$$

2.2.4.3. Test de la déviance

Pour tester l'adéquation du modèle, nous pouvons utiliser le test de déviance défini par :

$$\boldsymbol{D = 2\emptyset(l_{sature} - l_{ajusté})}$$

Avec $l = \log\left(L(\hat{\beta})\right)$:

- l_{sature} est le log-vraisemblance du modèle saturé.
- $l_{ajusté}$ est le log-vraisemblance du modèle ajusté.

La déviance est un écart en termes de log-vraisemblance entre le modèle saturé d'ajustement maximum et le modèle considéré. La déviance standardisée est définie par :

$$\boldsymbol{D^* = \frac{D}{\emptyset} = \left[2\left(l_{sature} - l_{ajusté}\right)\right] \geq 0}$$

Elle suit une $\boldsymbol{\chi_p^2}$, avec :

p **=** nombre d'observation – nombre de paramètres à estimer.

Plus $\boldsymbol{D^*}$ est petite, plus le modèle ajusté est bon.

Le test de la déviance s'écrit comme suit :

$$H_0 : le\ modèle\ ajusté\ est\ adéquat. \quad VS \quad H_1 : le\ modèle\ ajusté\ n'est\ pas\ adéquat$$

Nous rejetons H_0 si : $D^*_{obs} > \chi^2_p$.

2.2.4.4. Test d'adéquation du modèle par le diagramme quantile-quantile

En statistiques, le diagramme Quantile-Quantile, ou Q-Q plot est un outil graphique permettant d'évaluer la pertinence de l'ajustement d'une distribution donnée à un modèle théorique.

Le terme de quantile-quantile provient du fait de comparer la position de certains quantiles dans la population observée avec leur position dans la population théorique.

Le diagramme quantile-quantile permet également de comparer deux distributions que nous estimons semblables.

2.3. Forêts aléatoires et arbre de décision

Les **forêts aléatoires** sont des méthodes qui permettent d'obtenir des modèles prédictifs pour la classification et la régression. La méthode met en œuvre des arbres de décisions binaires, notamment des arbres CART proposés par *Breiman et al. (1984).*

L'idée générale derrière la méthode est la suivante : au lieu d'essayer d'obtenir une méthode optimisée en une fois, nous générons plusieurs prédicteurs avant de mettre en commun leurs différentes prédictions.

L'objectif d'un arbre de décision est donc de partitionner les observations en n classes distinctes. Pour cela, l'arbre sépare en 2 parties la base initiale, puis chaque partie est également séparée en 2.

Cette opération se poursuit jusqu'à ce que nous ayons atteint un critère d'arrêt.

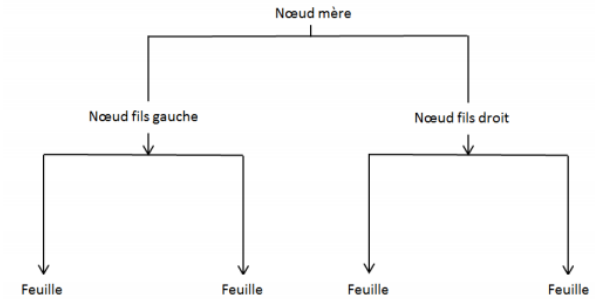


Figure 11 - Exemple d'arbre de décision

Les séparations sont identifiées par une valeur seuil de la variable sélectionnée. Cette séparation est effectuée selon le type de la variable :

- Si la variable sélectionnée pour effectuer la séparation est quantitative, nous distinguons la partie où la variable est inférieure à un seuil et la partie où la variable est supérieure à ce même seuil.

- Si la variable sélectionnée pour effectuer la séparation est qualitative alors la séparation est identifiée par l'appartenance à un groupe de modalités de la variable séparatrice, nous distinguons la partie où la variable appartient à un premier groupe de modalités et la partie où la variable appartient au groupe complémentaire.

Pour chaque séparation, nous cherchons la variable qui permet de séparer les individus en 2 groupes distincts et où chaque groupe est le plus homogène possible pour la variable séparatrice. Nous mesurons cette homogénéité via une fonction d'hétérogénéité qui dépend du type de variable à prédire, quantitative ou qualitative. Nous nous arrêtons quand le nœud est homogène, ou quand il n'est plus possible de séparer ou quand le nombre d'observations dans la feuille devient trop faible.

Soit X = [X_1, X_2, …, X_p,Y] notre échantillon statistique de taille $n \times (p+1)$.

Notre variable à prédire Y est qualitative de modalités {m_1,m_2, ….., m_k} en fonction de X_1, …, X_p.

Dans le cas d'une variable à prédire qualitative, la fonction d'hétérogénéité D_N pour le noeud N est la suivante, appelée fonction d'impureté de Gini :

$$D_N = \sum_{i=1}^{k} p_i^N (1 - p_i^N)$$

Où :

- p_i^N la proportion d'individus dans le nœud N pour lesquels Y = m_i.
- D_{N_G} et D_{N_D} le nœud fils gauche et le nœud fils droit.

Nous cherchons la variable X_j et la séparation en 2 nœuds fils minimisant l'hétérogénéité des nœuds fils, ce qui revient à maximiser la quantité :

$$Max_{\text{divisions de Xi,i=1….p}}\ D_N - D_{N_G} + D_{N_D}$$

Nous séparons notre échantillon successivement avec cette règle jusqu'à avoir des feuilles ne contenant plus qu'une observation ou jusqu'à ce que l'hétérogénéité des feuilles soit nulle.

3. Validation de la modélisation du risque de prime de la MRC

3.1. Présentation de la modélisation du risque de prime de la MRC

3.1.1. Introduction

Le portefeuille agricole de Groupama est composé de 2 types de contrats :

- Les contrats Grêle (modélisés directement sous REX[7] en modèle interne).
- Les contrats Climats (commercialisés depuis 2005) qui permettent de souscrire au produit Multi-Risques Climatiques (MRC) dont la modélisation de la garantie « Autres Aléas » est le sujet de ce mémoire.

Suite aux intempéries de 2016 ayant conduit à des indemnisations record pour Groupama dans le cadre des contrats Climats, le modèle tarifaire a été revu en 2017.

Pour modéliser la sinistralité spécifique des contrats MRC, Groupama a choisi, dans le cadre de son modèle interne, de s'appuyer sur son modèle tarifaire afin d'assurer une cohérence des inputs et hypothèses retenues dans les deux modèles.

L'objectif dans cette partie est de rappeler d'abord la méthodologie de la modélisation de la MRC, pour ensuite valider les modifications proposées par l'équipe modélisation pour modéliser le risque de primes MRC dans le MINV [8]

3.1.2. Rappel de la méthodologie

3.1.2.1. Introduction

Le contrat MRC couvre les agriculteurs assurés contre la baisse de leurs rendements suite à la survenance d'événements climatiques.

Lors de la signature du contrat, l'assureur définit des variables tarifaires comme :

- Le rendement de référence (en fonction du rendement historique de l'agriculteur),
- Le prix de référence (défini à l'avance dans le contrat : ce contrat ne couvre pas la variation des prix).
- Le taux de franchise (choisi à l'avance par l'agriculteur).

Le modèle MRC du MINV a pour objectif de créer une distribution de sinistralité totale annuelle (actuellement basée sur 50 000 simulations) qui intervient ensuite en entrée du modèle de risque de prime de la ligne d'activité FIRE. Pour ce faire, le modèle s'appuie sur les données de la base AGRESTE et les données des portefeuilles du modèle tarifaire de l'équipe tarification (Primes, franchises, surfaces et capitaux assurés, où une ligne correspond à un contrat pour un agriculteur et pour une culture donnée).

3.1.2.2. Présentation du modèle

[7] Risk Explorer (REX) : Outil de gestion des risques
[8] MINV : Modèle Interne Non-Vie

Dans le modèle MINV de la MRC[9], nous définissons la sinistralité d'une culture pour un agriculteur par la formule suivante :

$$\tilde{S}_i = P \cdot V_i \cdot Max\left((1-\alpha_i)\cdot R_i^{\text{Ref}} - \tilde{R}_i\,;\,0\right) = P \cdot V_i \cdot Max\left(\tilde{Y}_i - K_i\,;\,0\right)$$

Où :

- $K_i = R_i^{Max} - (1-\alpha_i)\cdot R_i^{\text{Ref}}$
- P le prix de référence
- V_i la surface cultivée
- α_i la franchise
- R^{Max} le rendement maximal atteint historiquement
- R^{Ref} le rendement de référence.

La figure ci-dessous explique le principe de l'indemnisation dans le cas du modèle interne non-vie de Groupama.

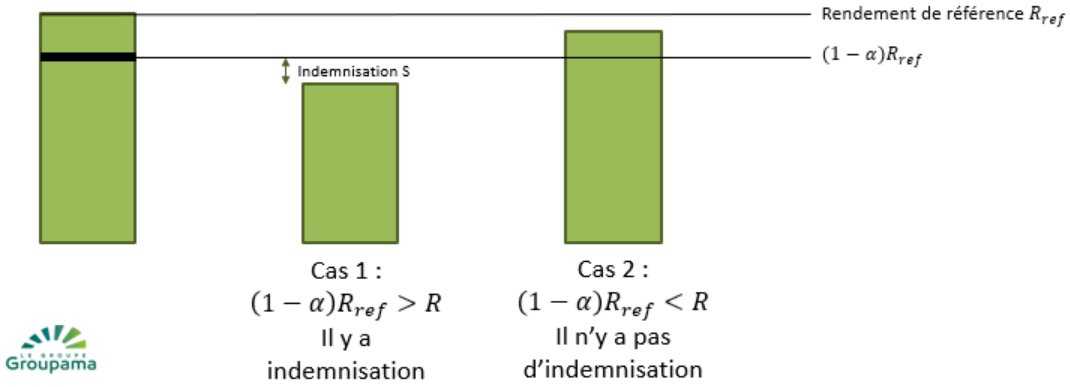


Figure 12 - Principe de l'indemnisation d'un sinistre autres aléas dans le cas du MINV de Groupama

3.1.2.3. Calibrage de la sinistralité individuelle

Notre objectif est l'estimation de la sinistralité individuelle pour un agriculteur S_i par la formule suivante :

$$S_i = P \times V_i \times \max\left(\tilde{Y}_i - K_i,\ 0\right)$$

Avec : $K_i = R_i^{Max} - (1-\alpha_i)\cdot R_i^{Ref}$, $\tilde{Y}_i = \tilde{Y}_{rc}\cdot\tilde{\varepsilon}_i$ [10]avec $i \in rc\ (région, culture)$

Où :

[9] Source interne : 18 02 20 RCalibrage du modèle climat_vdef
[10] Démonstration en annexe 3

- $\tilde{Y}_{rc} \sim lognormale(\mu, \sigma^2)$, $\tilde{Y}_{rc}$ est la variable de perte de rendement d'un couple CR[11]/culture (suivant une loi Log-Normale).
- $\tilde{\varepsilon}_i \sim lognormale(1, \sigma_{\varepsilon_i}^2)$ ε_i le risque spécifique de chaque agriculteur.
- $\tilde{Y}_i \sim lognormale(\mu, \sigma_i^2)$ la perte de rendement individuel.

Le schéma suivant explique les étapes de calcul de l'indemnisation dans le modèle interne non-vie pour les contrats MRC :

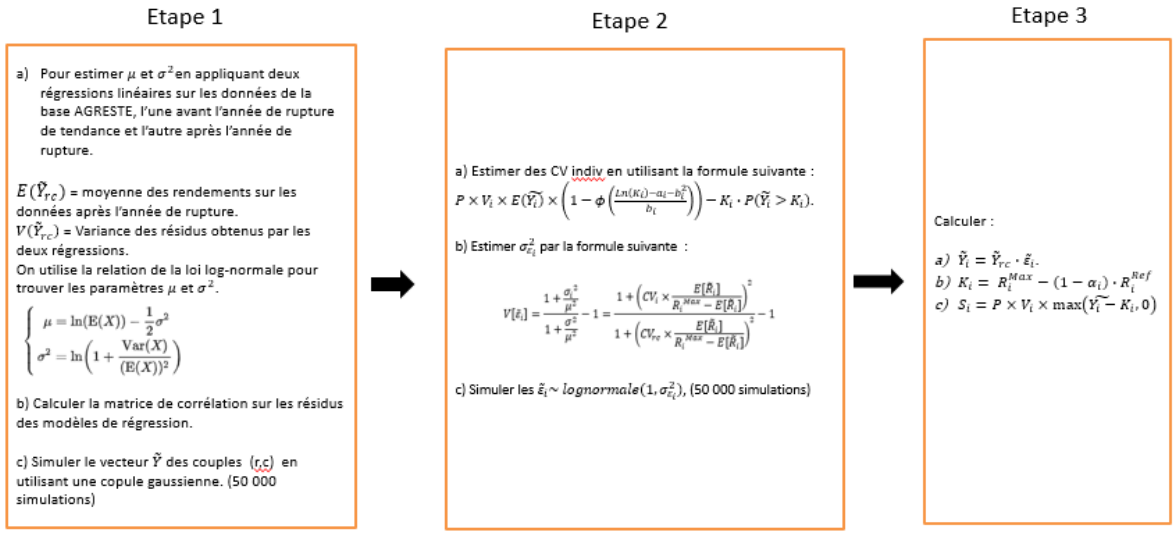


Figure 13 - Etapes de calcul de l'indemnisation dans le MINV pour les contrats MRC

Nous présentons ci-dessous un exemple de calcul de l'indemnisation individuelle :

		Ytield	eps_i	Y_i		S_i	S_i_Prog_R	Ecart		
ID_contrat	20187501032936408201 90NA4	1,679918	1,094767553	1,83912		0,00	0,00	0,00%		
		2,120987	0,309094538	0,655585		0,00	0,00	0,00%	MAX	2592,71
prix_ref	139,5932943	0,6625489	0,759719766	0,503351		0,00	0,00	0,00%	MIN	0,00
surfaces_expo	2	1,66811	0,882192031	1,471593		0,00	0,00	0,00%		
R_max	9,077799952	2,595348	0,989055847	2,566944		0,00	0,00	0,00%		
franchises	0,25	1,013846	0,394638006	0,400102		0,00	0,00	0,00%		
rdts_ref	7,564833293	2,073062	0,990592596	2,05356		0,00	0,00	0,00%		
		0,8119354	1,308263983	1,062226		0,00	0,00	0,00%		
		1,52606	1,799874093	2,746716		0,00	0,00	0,00%		
K_i	3,404174982	1,236355	0,598551022	0,740022		0,00	0,00	0,00%		
		1,013973	1,781814939	1,806712		0,00	0,00	0,00%		
		2,294272	0,996162718	2,285468		0,00	0,00	0,00%		
Kapital_assuré	2112	1,46055	1,091599932	1,594336		0,00	0,00	0,00%		
maximum_indem	1584	2,080359	0,660117771	1,373282		0,00	0,00	0,00%		
		1,602579	1,318773703	2,113439		0,00	0,00	0,00%		
		0,8600125	0,738045715	0,634729		0,00	0,00	0,00%		
		0,6090026	1,160516882	0,706758		0,00	0,00	0,00%		
		1,314291	0,514056285	0,67562		0,00	0,00	0,00%		
		0,7850955	0,691195228	0,542654		0,00	0,00	0,00%		

Figure 14 - Exemple de calcul de l'indemnisation individuelle

[11] Caisse régionale de Groupama

3.1.2.4. La distribution de la sinistralité totale

Pour calculer la distribution de la sinistralité totale (actuellement basée sur 50 000 simulations), nous calculons la somme des sinistralités individuelles S_i (50 000 pour chaque individus) de chaque contrat.

Nous obtenons par la suite une distribution agrégée comme présenté ci-dessous :

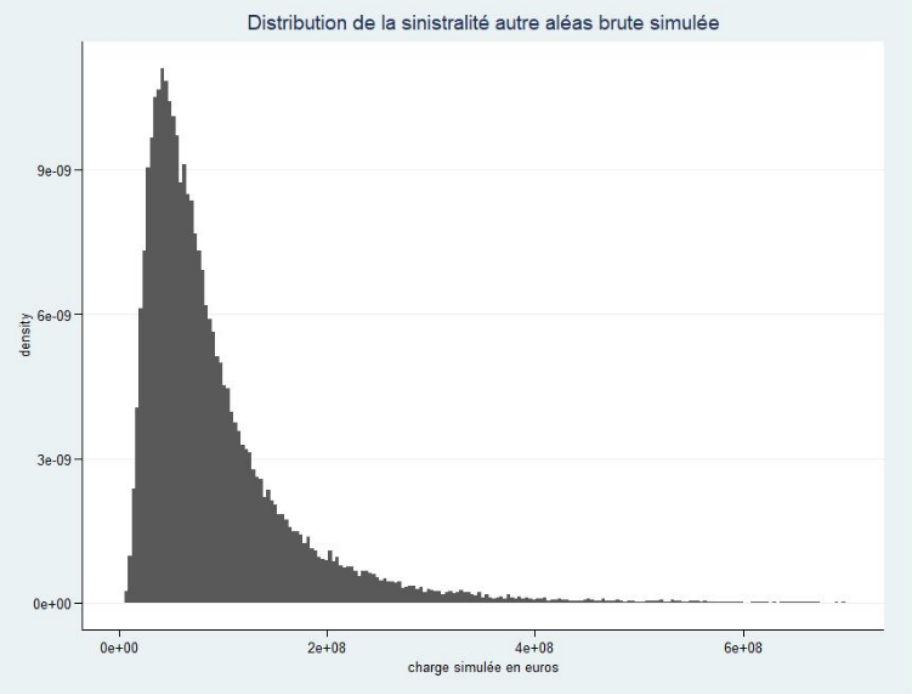


Figure 15 - Distribution de la charge totale

3.1.2.5. Implémentation sous Risk Explorer

Nous trouvons la sortie de la fonction de distribution cumulative de la sinistralité totale dans un fichier *txt* (voir *figure 16* ci-dessous) :

Probabilité	Charge Cumulée
0,000%	-
0,002%	5 592 454
0,004%	5 710 054
0,006%	6 057 553
0,008%	6 083 479
0,010%	6 469 621
0,012%	6 769 201
0,014%	6 797 203
0,016%	6 808 983
0,018%	6 960 093
0,020%	7 147 103
0,022%	7 204 896
0,024%	7 206 338
0,026%	7 370 814
0,028%	7 476 506
0,030%	7 499 686
0,032%	7 547 459
0,034%	7 579 965
0,036%	7 635 039
0,038%	7 715 840
0,040%	7 757 480
0,042%	7 803 186

Figure 16- Distribution cumulative de la sinistralité totale

La 1ère colonne représente la probabilité d'obtenir une valeur inférieure à x ($F_X(x) = P(X \leq x)$) et la 2ème colonne représente la valeur x associée à cette probabilité.

L'implémentation sous REX de ce fichier est réalisée de la maniéré suivante :

La 1ère étape consiste à faire un copier/coller du fichier de la sinistralité totale dans la partie 5 du graphique suivant :

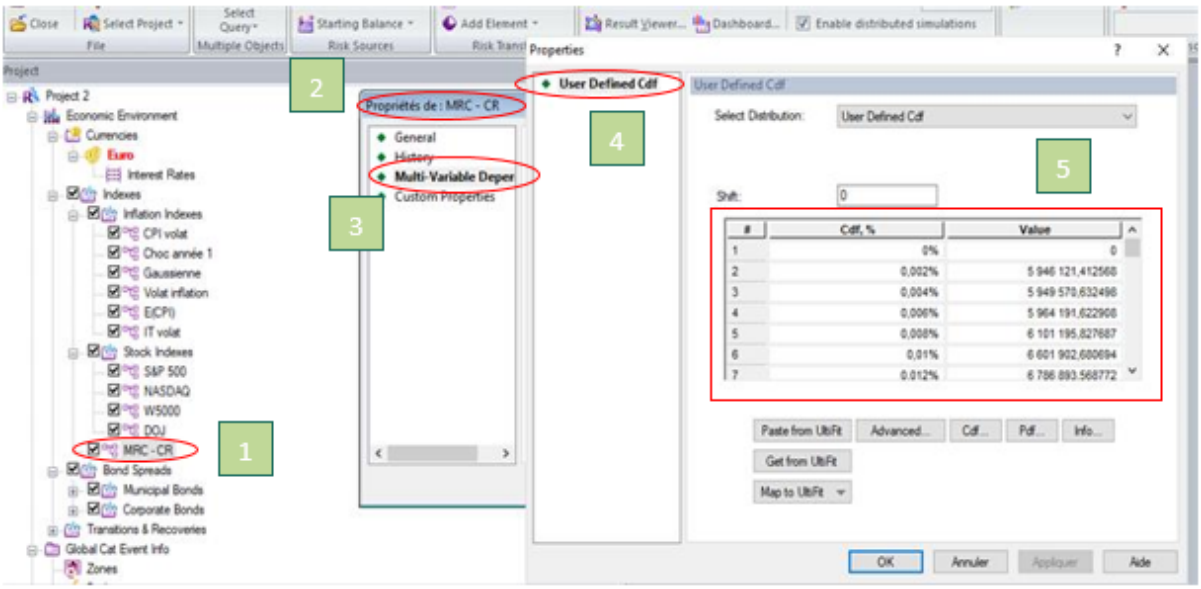


Figure 17 - Première étape de l'implémentation sous REX

Dans la 2ème étape, nous devons définir le poids[12] de la sinistralité totale pour chaque caisse (voir le graphe suivant) :

[12] Source : SUIVI CTP BRANCHE RECOLTES v2020.xlsx (Fichier interne)

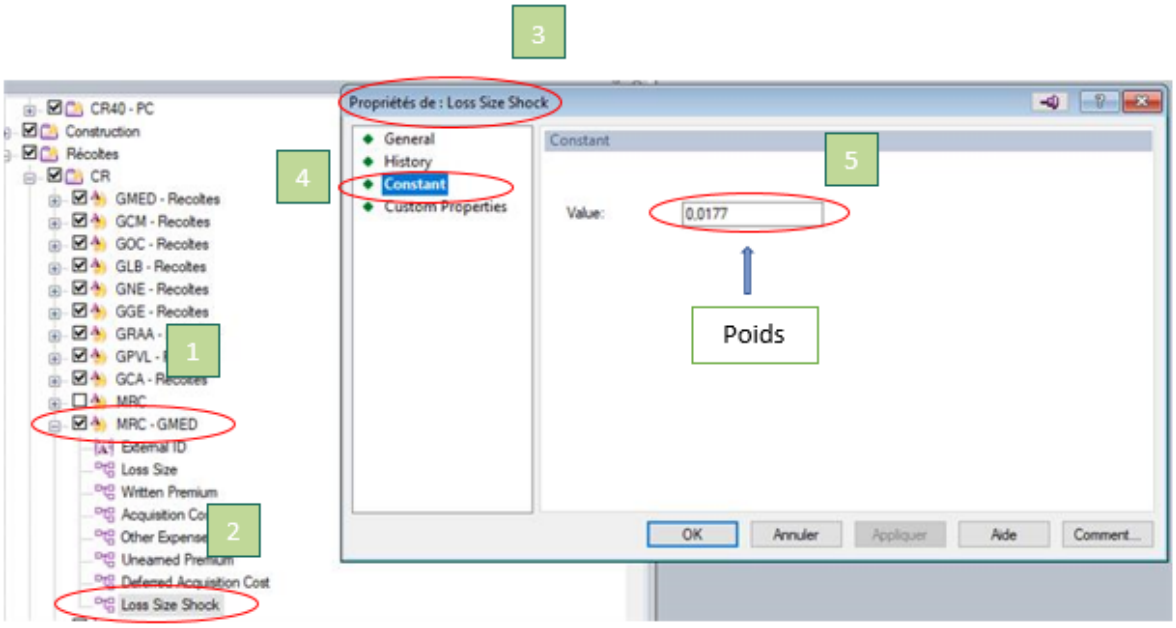


Figure 18 - Deuxième étape de l'implémentation sous REX

3.2. Analyse des travaux de l'équipe de modélisation

En s'appuyant sur une étude de 2010[13] montrant que pour certaines cultures, une tendance croissante due au progrès technologique était observée sur les rendements avant une phase de stabilisation, des années de rupture ont été retenues dans la modélisation du risque MRC par Groupama.

Pour les données présentant une année de rupture, une régression est effectuée avant le seuil et une après le seuil pour obtenir les résidus.

Les graphiques ci-dessous sont des exemples significatifs des comportements observables sur les rendements des cultures avec une année de rupture de tendance :

- La ligne verticale rouge indique l'année de rupture pour la culture étudiée
- Les deux droites bleues représentent les deux régressions actuellement réalisées avant et après l'année de rupture lorsqu'il y en a une.

[13] BRISSON, Nadine, GATE, Philippe, GOUACHE, David, *et al*. « Why are wheat yields stagnating in Europe? A comprehensive data analysis for France ». *Field Crops Research*, 2010, vol. 119, no 1, p. 201-212

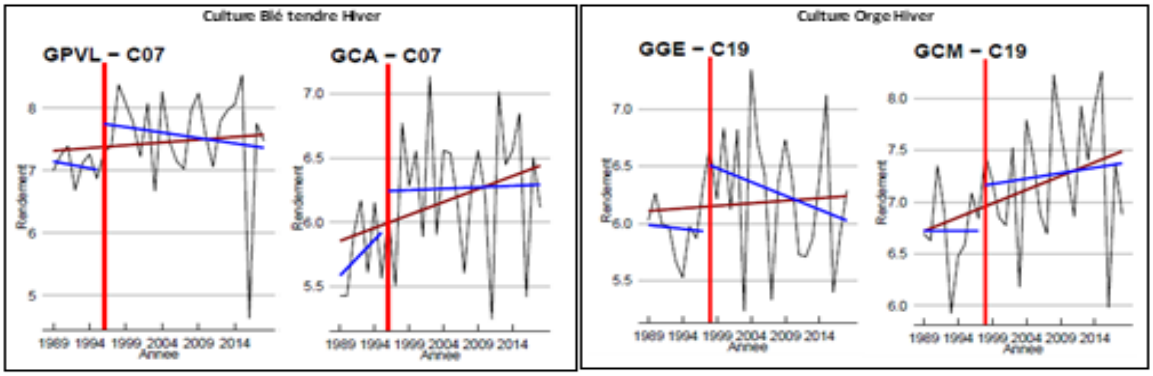


Figure 19 – Illustration de la différence entre les régressions linéaires sur tout l'historique ou en distinguant deux périodes avant et après l'année de rupture

Concernant l'utilisation d'une année de rupture, nous constatons les limites suivantes:

- Année de rupture déterminée au niveau national mais calibrage des paramètres réalisé au niveau Caisse Régionale.
- Pertinence de l'utilisation d'une année de rupture commune à toutes les Caisses Régionales injustifiée.
- Discontinuité observée au niveau de l'année de rupture.
- Données insuffisantes pour établir des tendances fiables dans les régressions linéaires.
- Réalisation d'une seconde régression linéaire post-année de rupture pour « modéliser » un plateau.
- Incohérence des données utilisées pour calibrer la moyenne et la volatilité des rendements.

3.2.1. Impact de la prise en compte d'un plateau

Pour traiter ces limites concernant l'année de rupture, l'équipe modélisation a proposé de ne plus faire intervenir la notion de plateaux dans les rendements des couples Caisses x Cultures. Deux possibilités ont été choisi pour estimer les rendements moyens et leurs volatilités :

1) Considérer qu'il existe bien une tendance sur les rendements et réaliser une régression linéaire (sur la variable année) appliquée sur la totalité de l'historique.

2) Calculer le rendement moyen et estimer la volatilité à partir des écarts à la moyenne, c'est-à-dire retenir les écarts entre chaque point de rendement historique et le rendement moyen pour calibrer la volatilité.

	En k€	Moyenne	Quantile 99,5%
Méthode régression linéaire plateau	Résultat	93 637	472 090
Régression linéaire tout historique	Résultat	93 574	509 086
	Variation	-0,07%	7,84%
Ecart à la moyenne tout	Résultat	94 500	611 925
	Variation	0,92%	29,62%

Figure 20 – Résultats des différentes méthodes

Remarque de la validation :

Le retrait de la rupture de tendance implique une hausse du quantile 99,5%, qui est beaucoup plus marquée lorsqu'on retient une méthode moyenne et écart à la moyenne (+29.6%) plutôt qu'une régression linéaire unique hausse de 7.8%)

3.2.2. Etude de la significativité des régressions

Afin de connaître la meilleure de ces deux solutions proposées pour estimer la moyenne et la volatilité, l'équipe modélisation a réalisé des tests de significativité globale (tests de Fisher) sur les régressions linéaires simples des rendements de tous les couples CR[14] x Cultures.

Dans le cas de régressions linéaires simples, si le modèle n'est globalement pas significatif cela indique qu'aucune tendance n'est observée. Cela indique également que l'ajout de cette variable explicative (ici les années) ne permet pas de modéliser le phénomène plus précisément car les points sont répartis sans structure linéaire significative.

Une non-significativité globale des régressions linéaires simples indiquerait qu'il faudrait utiliser l'écart à la moyenne pour obtenir les résidus des rendements.

Sur les rendements des cultures de références AGRESTE, les taux de significativité globale sur l'ensemble des couples CR x Cultures existants sont les suivant :

	Régression significative (en %)	Régression non significative (en %)	Total
Risque 5%	63,80%	36,20%	100%
Risque 10%	69,33%	30,67%	100%

Figure 21 - Nombre de régressions significatives - Portefeuille Groupama

	Régression significative (en %)	Régression non significative (en %)	Total
Risque 5%	61,96%	38,04%	100%
Risque 10%	67,39%	32,61%	100%

Figure 22 - Nombre de régressions significatives - Tous couples CR x Culture possibles

Remarques et préconisations de la validation :

- Nous avons remarqué que deux tiers des régressions sont significatifs pour un niveau de risque de 10%. Cela signifie qu'aucune tendance n'est observée pour 1/3 des régressions.
- Nous remarquons aussi que la méthode la plus adaptée pour l'étude de l'évolution des rendements est la régression linéaire (sur la variable année) appliquée sur tout l'historique.

[14] CR : Caisse Régionale

- Nous proposons donc d'analyser la significativité des régressions en fonction du capital assuré afin d'expliquer la non-significativité de ces régressions et de vérifier l'importance de ces régressions.

3.3. Travaux complémentaires sur la méthode régression linéaire

3.3.1. Significativité des régressions (Capital assuré)

Pour étudier la non-significativité des régressions nous avons réalisé des travaux complémentaires sur la méthode de régression linéaire. Le tableau suivant présente la significativité des régressions en fonction du capital assuré pour chaque couple CR x culture.

Meta_Gpt	Étiquettes	CR13	CR28	CR31	CR35	CR51	CR67	CR69	CR75	CR79	Total général
ARBO / PEPINIERES_Printemps	C01	8,38	0,23	0,22				2,40		0,10	11,33
AUTRES CEREALES_Hiver	C02	1,12	7,34	7,72	0,87	1,93	6,049	9,03	11,06	7,99	53,10
AUTRES CEREALES_Printemps	C03	0,27	0,60	1,18	0,02	1,05	0,743	1,20	2,04	1,03	8,12
BETTERAVE_Printemps	C04		30,35		0,10	101,95	7,984		93,99	0,00	234,38
BLE DUR_Hiver	C05	2,33	7,74	5,30	0,43	0,07	0,074	0,75	22,53	7,27	46,48
BLE DUR_Printemps	C06					0,01					0,01
BLE TENDRE_Hiver	C07	2,25	266,51	28,37	19,37	281,24	125,845	57,39	406,65	85,09	1272,72
BLE TENDRE_Printemps	C08		0,28			0,90	0,692	0,13			1,99
COLZA_Hiver	C09	0,56	76,47	5,45	4,02	55,84	40,715	11,51	93,56	13,78	301,90
COLZA_Printemps	C10		0,04			0,29	0,061	0,00			0,38
LEGUMES FEUILLES - FLEURS - PETITS FRUITS_Printemps	C11	0,44	0,53		0,03	7,71	0,448	0,02	2,87	0,18	12,23
LEGUMES RACINES ET BULBES_Printemps	C12					2,29	0,290		2,51		5,09
LEGUMINEUSES_Hiver	C13					0,05		0,24			0,29
LEGUMINEUSES_Printemps	C14	0,44	2,96	24,18	0,56	9,83	6,979	7,35	12,80	5,55	70,64
MAIS_Printemps	C17	2,88	131,34	137,11	19,91	48,66	151,498	50,09	89,49	65,94	696,92
MAIS SEMENCE_Printemps	C18	5,44	12,55	65,99	24,10		3,462	9,44	2,29	11,41	134,69
ORGE_Hiver	C19	0,96	54,72	10,97	3,28	47,21	39,524	19,35	86,16	16,25	278,42
ORGE_Printemps	C20		6,86			67,63	15,892	3,31	81,70	2,63	178,02
POIS PROTEAGINEUX_Hiver	C23		0,03			1,63		0,52			2,18
POIS PROTEAGINEUX_Printemps	C24	0,10	3,63	0,36	0,09	3,74	2,578	0,50	14,66	1,61	27,26
POMME DE TERRE_Printemps	C25	0,99	30,35	0,77	0,04	72,31	3,906	0,57	56,85	0,03	165,82
TOURNESOL_Printemps	C26	6,43	1,87	20,19	0,64	2,90	6,557	7,21	17,60	24,41	87,80
VITI AOC_Printemps	C27	104,80			13,07	25,78	131,293	97,98	40,10	51,88	464,91
VITI IG_Printemps	C28	261,89		65,64	3,70		0,015	0,95	1,31	8,63	342,14
	Total général	399,28	634,38	373,44	90,22	733,02	544,604	279,93	1038,17	303,79	4396,85

Figure 23 - Significativité des régressions en fonction du capital assuré pour chaque couple CR x culture (Données en M€, Année 2019)

5% <= P-value < 10%

P-value < 5%

Nous remarquons que pour certains couples CR x Culture, le capital assuré pour les régressions non significatives est très important et dépasse dans certains cas les 200 M€, notamment pour CR51/C07 (281) et CR75/C07 (407).

Le tableau suivant présente le pourcentage du capital assuré des régressions non significatives pour chaque culture.

		Somme significativité		Somme non significativité		Pourcentage non significativité		
Meta_Gpt	Étiquettes	Risque 5%	Risque 10%	Risque 5%	Risque 10%	Risque 5%	Risque 10%	Total
ARBO / PEPINIERES_Printemps	C01	10,88	10,88	0,45	0,45	4,01%	4,01%	11,331701
AUTRES CEREALES_Hiver	C02	25,23	27,16	27,88	25,95	52,49%	48,86%	53,104051
AUTRES CEREALES_Printemps	C03	2,65	3,70	5,47	4,42	67,34%	54,43%	8,1239244
BETTERAVE_Printemps	C04	234,28	234,28	0,10	0,10	0,04%	0,04%	234,38387
BLE DUR_Hiver	C05	7,63	14,90	38,85	31,58	83,59%	67,94%	46,483755
BLE DUR_Printemps	C06	0,00	0,01	0,01	0,00	100,00%	0,00%	0,0058787
BLE TENDRE_Hiver	C07	135,08	401,60	1137,64	871,12	89,39%	68,45%	1272,7201
BLE TENDRE_Printemps	C08	1,87	1,87	0,13	0,13	6,29%	6,29%	1,9946386
COLZA_Hiver	C09	235,89	235,89	66,01	66,01	21,86%	21,86%	301,8997
COLZA_Printemps	C10	0,32	0,32	0,06	0,06	15,79%	15,79%	0,3842315
LEGUMES FEUILLES - FLEURS - PETITS FRUITS_Printemps	C11	3,63	3,63	8,60	8,60	70,33%	70,33%	12,228835
LEGUMES RACINES ET BULBES_Printemps	C12	5,09	5,09	0,00	0,00	0,00%	0,00%	5,0937335
LEGUMINEUSES_Hiver	C13	0,29	0,29	0,00	0,00	0,00%	0,00%	0,2884157
LEGUMINEUSES_Printemps	C14	46,46	46,46	24,18	24,18	34,23%	34,23%	70,635865
MAIS_Printemps	C17	630,98	696,92	65,94	0,00	9,46%	0,00%	696,92227
MAIS SEMENCE_Printemps	C18	101,68	101,68	33,01	33,01	24,51%	24,51%	134,69317
ORGE_Hiver	C19	105,52	152,73	172,90	125,69	62,10%	45,14%	278,4167
ORGE_Printemps	C20	162,13	162,13	15,89	15,89	8,93%	8,93%	178,02439
POIS PROTEAGINEUX_Hiver	C23	2,18	2,18	0,00	0,00	0,00%	0,00%	2,1750984
POIS PROTEAGINEUX_Printemps	C24	26,90	26,90	0,36	0,36	1,32%	1,32%	27,261141
POMME DE TERRE_Printemps	C25	165,17	165,17	0,65	0,65	0,39%	0,39%	165,82033
TOURNESOL_Printemps	C26	35,68	60,09	52,12	27,71	59,37%	31,56%	87,802837
VITI AOC_Printemps	C27	385,96	385,96	78,95	78,95	16,98%	16,98%	464,91005
VITI IG_Printemps	C28	264,17	267,88	77,97	74,26	22,79%	21,71%	342,14032

Figure 24 - Pourcentage du capital assuré des régressions non significatives pour chaque culture

Nous remarquons que pour certaines cultures, le pourcentage du capital assuré pour les régressions non significatives dépasse 70% (à l'exemple du blé tendre hiver (83%)).

Afin d'améliorer l'interprétation de ces résultats, nous les avons représentés graphiquement :

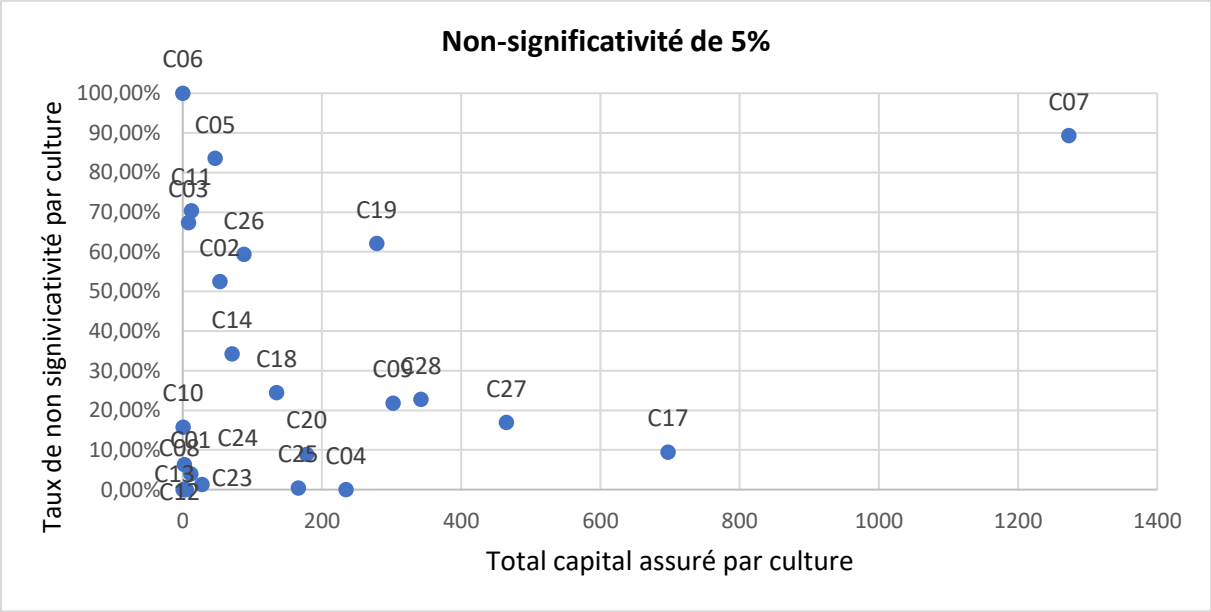


Figure 25- Pourcentage du capital assuré des régressions non significatives en fonction du total du capital assuré pour chaque culture pour le niveau de risque 5%

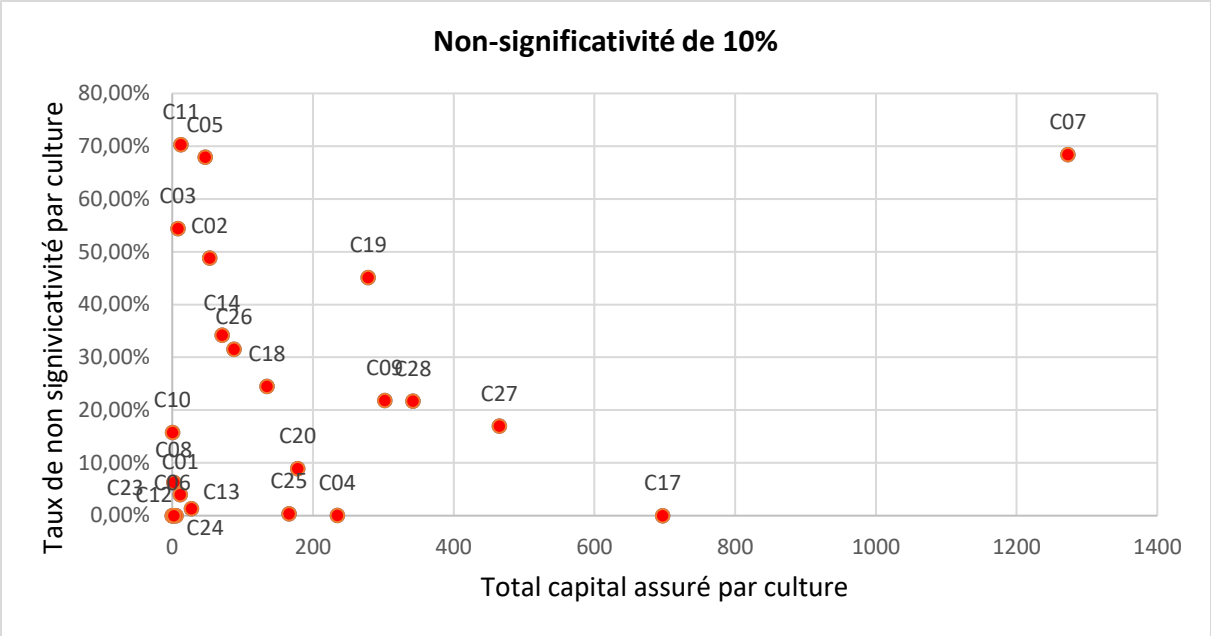


Figure 26 - Pourcentage du capital assuré des régressions non significatives en fonction du total du capital assuré pour chaque culture pour le niveau de risque 10%

Nous observons sur ces deux graphiques un taux de non-significativité des régressions très élevé pour les cultures qui ont un capital assuré très important (notamment C07, C19).

Maintenant nous présentons dans le tableau suivant le pourcentage des capitaux assurés pour les régressions non-significatives pour l'ensemble des couples CR x Cultures.

	Montant	Pourcentage
somme non significative 5%	1807,16	41,10%
somme non significative 10%	1389,13	31,59%
Total	4396,844963	

Figure 27 - Pourcentage des capitaux assurés pour les régressions non-significatives pour l'ensemble des couples CR x Cultures

La partie du capital assuré pour les régressions non-significatives reste importante pour les deux niveaux de risque (5% et 10%) :

- Pour un niveau de risque de 5%, les régressions non-significatives représentent presque 41% du capital assuré.
- Pour un niveau de risque de 10%, les régressions non significatives représentent presque 32% du capital assuré.

Nous proposons donc d'utiliser la méthode écart à la moyenne pour les régressions non-significatives. Autrement dit, nous utilisons une combinaison des deux méthodes pour calculer les résidus : si la régression est significative nous utilisons une régression simple, sinon nous utilisons la méthode écart à la moyenne.

3.3.2. La qualité d'ajustement des modèles

L'étude de la qualité d'ajustement des modèles de régression a montré que les régressions linéaires effectuées sur les rendements au niveau CR x Culture sont globalement assez peu représentatives. Le tableau ci-dessous présente le nombre des régressions en fonction des R^2.

R_carre	>10%	>20%	>30%	>40%	>50%
Nombre	127	99	70	45	24
Pourcentage	67,91%	52,94%	37,43%	24,06%	12,83%

Nombre des Regressions	187

Figure 28 - Nombre des régressions en fonction des R^2

D'après ce tableau, nous remarquons que presque 67% des régressions ont un R^2 supérieur à 10%, et seulement 12,8% des régressions ont un R^2 supérieur à 50%.

3.3.3. Vérification des hypothèses des régressions

Dans cette partie nous allons tester les hypothèses de la régression linéaire, à savoir, homoscédasticité, normalité des résidus et absence d'autocorrélation des résidus pour l'ensembles des couples Caisses/Cultures. Les résultats obtenus sont les suivants :

Test d'acceptation de la normalité des résidus	5%	10%
Ensemble des couples CR/Culture	78,61%	70,59%

Test d'acceptation d'homoscédasticité des résidus	5%	10%
Ensemble des couples CR/Culture	89,30%	81,82%

Test d'acceptation d'absence d'autocorrélation des résidus	5%	10%
Ensemble des couples CR/Culture	75,40%	70,59%

La majorité des régressions linéaires vérifient les trois hypothèses de la régression linéaire (75% pour un niveau d'erreur de 5%), cela confirme la pertinence de cette méthode.

3.3.4. Méthode alternative (Méthode IF)

Pour éviter le problème de la non-significativité de certaines régressions (1/3 des régressions), nous avons proposé une méthode alternative qui combine les deux méthodes proposées par l'équipe modélisation, comme expliqué ci-dessous :

- **SI** la régression est significative, les résidus seront calculés à l'aide d'une régression simple
- **SINON**, l'écart à la moyenne sera utilisé.

Nous présentons ci-dessous une comparaison entre la méthode de régression linéaire plateau et les autres méthodes, à savoir, la méthode moyenne et la méthode IF (avec les deux niveaux de risque 5% et 10%).

	En K euro	Moyenne	Quantile 99,5%
Méthode régression linéaire plateau	Résultat	93 637	472 090
Regression linéaire tout historique	Résultat	93 574	509 086
	Variation	-0,07%	7,84%
Ecart à la moyenne tout historique	Résultat	94 500	611 925
	Variation	0,92%	29,62%
Méthode IF Regression_Moyenne Risque 5%	Résultat	93 633	499 496
	Variation	0,00%	5,81%
Méthode IF Regression_Moyenne Risque 10%	Résultat	93 639	496 498
	Variation	0,00%	5,17%

Figure 29 - Comparaison entre la méthode de régression linéaire plateau et la méthode moyenne et IF

Dans le tableau suivant, nous comparons la méthode IF (avec les deux niveaux de risque 5% et 10%) et la méthode de régression linéaire tout historique.

	En K euro	Moyenne	Quantile 99,5%
Regression linéaire tout historique	Résultat	93 574	509 086
Méthode IF Regression_Moyenne Risque 5%	Résultat	93 633	499 496
	Variation	0,06%	-1,88%
Méthode IF Regression_Moyenne Risque 10%	Résultat	93 639	496 498
	Variation	0,07%	-2,47%

Figure 30 - Comparaison entre la méthode IF et la méthode de régression linéaire tout historique

La méthode IF Régression Moyenne permet d'éviter le problème de la non-significativité de certaines régressions.

Nous observons des écarts négligeables entre la régression linéaire tout historique et la méthode IF.

3.3.5. Impact de la méthode IF sur le SCR du Groupe

Dans ce tableau, nous avons mesuré l'impact de la méthode IF (avec les deux niveaux de risque 5% et 10%) sur le SCR globale du groupe.

	En M euro	SCR Groupe MIP
Regression linéaire tout historique	Résultat	436,7
Méthode IF Regression_Moyenne Risque 5%	Résultat	437,7
	Variation	0,23%
Méthode IF Regression_Moyenne Risque 10%	Résultat	436,5
	Variation	-0,05%

Figure 31 - Impact de la méthode IF sur le SCR globale du groupe

En termes du SCR, nous observons des écarts très négligeables entre la régression linéaire tout historique et la méthode IF.

3.3.6. Backtesting

Nous avons réalisé un backtesting en comparant la distribution théorique de la charge brute avec les sinistralités observées sur chacune des années d'exercice.

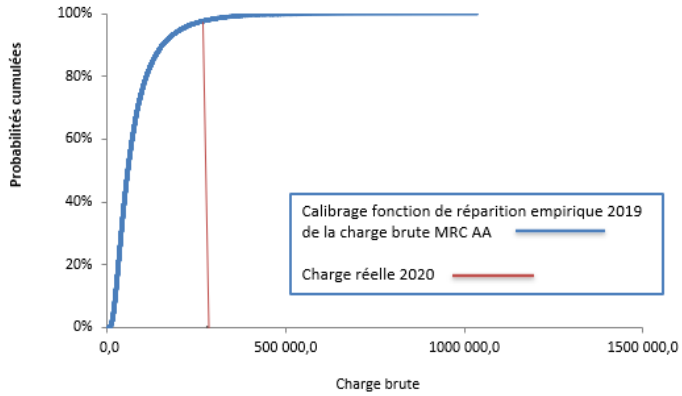


Figure 32 – Fonction de répartition de la charge brut

D'après la distribution théorique des charges obtenues en 2019, la période de retour de 2020 est de 48 ans. Groupama dispose 15 ans d'historique sur le risque de MRC, une période de retour de 48 ans indique donc que la sinistralité de 2020 est importante. En effet, comme le montre le tableau ci-dessous, l'année 2020 est la deuxième année la plus sinistrée après l'année 2016 qui est exceptionnelle.

Année	Charge de la sinistralité	Probabilité	Période de retour
2005	21 278 659,20	0,08	1,09
2006	37 509 174,52	0,28	1,39
2007	51 409 743,54	0,44	1,79
2008	40 386 287,22	0,32	1,46
2009	25 143 425,76	0,13	1,14
2010	68 835 253,39	0,59	2,46
2011	134 958 772,34	0,86	7,40
2012	123 995 042,21	0,84	6,27
2013	146 531 096,29	0,89	8,81
2014	51 222 895,83	0,44	1,78
2015	105 255 827,41	0,79	4,66
2016	345 382 386,46	0,99	101,63
2017	113 297 282,65	0,81	5,32
2018	104 545 053,48	0,78	4,61
2019	163 213 671,39	0,91	11,27
2020	280 077 966,89	0,98	48,03

Figure 33 - Charges totales des sinistres et période de retour par année

3.4. Sinistralité maximale par couple CR x Culture

Nous avons remarqué que le modèle utilisé actuellement conduit à des **sinistralités simulées pouvant être supérieures aux capitaux assurés** par Groupama pour certains couples CR/Cultures. Ci-dessous les résultats obtenus lors du Closing 2019 :

	C01		C02		C03		C04		C05	
	Capital assuré	Perte max simulée	Capital assuré	Perte max simulée	Capital assuré	Perte max simulée	Capital assuré	Perte max simulée	Capital assuré	Perte max simulée
CR13	9 077 472	31 737 403	1 105 907	4 449 563	164 091	538 783			3 789 335	5 919 738
CR28			7 046 947	3 303 962	481 449	933 153	33 873 902	6 086 333	8 899 192	20 384 537
CR31	191 566	1 735 502	7 763 988	4 269 070	979 520	18 176 492			9 059 886	13 364 138
CR35			836 090	537 166	19 389	67 533	51 614	41 149	762 366	1 241 273
CR51			1 821 103	1 713 535	684 532	5 548 191	107 380 038	15 889 528	92 132	343 787
CR67			5 282 259	3 835 393	466 222	1 790 889	10 039 057	11 613 224	44 722	110 568
CR69	1 562 752	14 168 187	8 176 922	5 236 156	541 325	1 157 034			987 448	1 599 119
CR75			8 481 907	7 768 959	931 519	2 315 530	100 198 851	19 743 753	27 568 441	65 552 422
CR79			6 817 617	7 655 103	624 359	700 662	3 780	4 594	11 442 241	10 579 097

Figure 34 - Comparaison entre les capitaux assurés et la perte maximale

Le tableau ci-dessus présente un exemple de la comparaison entre les capitaux assurés et la perte maximale mais dans la réalité la **perte maximale que peut subir Groupama** sera égale aux **capitaux assurés moins la franchise** choisie par l'assuré au moment de la souscription de son contrat.

Sur les années de souscription 2017 et 2018 :

- **88% des contrats avaient une franchise de 25%,**
- 5% une franchise de 30%,
- 3% une franchise de 20%,
- 3% une franchise de 15%,
- 2% une franchise de 10%.

Limitation des pertes maximales :

Pour limiter les pertes simulées au niveau des montants des capitaux assurés, nous avons proposé deux solutions :

- Une première solution consiste à caper pour chaque contrat la perte de rendement simulée par le rendement moyen de façon à avoir pour chaque contrat une perte maximale égale aux capitaux assurés.

- Une deuxième possibilité est de limiter au niveau CR x Culture la somme des pertes individuelles simulées au montant des capitaux totaux assurés pour ce couple.

En K euro		Moyenne	Quantile 99,5%
Regression linéaire tout historique Sans cap	Résultat	93 574	509 086
Regression linéaire tout historique Cap CR/Cult	Résultat	93 046	493 533
	Variation	-0,56%	-3,06%
Regression linéaire tout historique Cap Sinistralité individuelle	Résultat	73 975	420 430
	Variation	-20,94%	-17,41%

En K euro		Moyenne	Quantile 99,5%
Méthode IF Regression_Moyenne Sans cap 5%	Résultat	93 633	499 496
Méthode IF Regression_Moyenne Cap CR/Cult 5%	Résultat	93 084	489 561
	Variation	-0,59%	-1,99%
Méthode IF Regression_Moyenne Cap Sinistralité individuelle 5%	Résultat	74 013	422 583
	Variation	-20,95%	-15,40%

En K euro		Moyenne	Quantile 99,5%
Méthode IF Regression_Moyenne Sans cap 10%	Résultat	93 639	496 498
Méthode IF Regression_Moyenne Cap CR/Cult 10%	Résultat	93 097	485 382
	Variation	-0,58%	-2,24%
Méthode IF Regression_Moyenne Cap Sinistralité individuelle 10%	Résultat	74 023	421 336
	Variation	-20,95%	-15,14%

Figure 35 - Comparaison de chaque méthode avec les 2 tests de Cap de la sinistralité

Dans les tableaux ci-dessus, nous comparons chaque méthode avec la même méthode en appliquant la limitation de la sinistralité soit au niveau individuel (Cap sinistralité individuelle), soit au niveau du couple CR x Culture (Cap CR x Culture).

Remarques et préconisations de la validation :

- La limitation de la sinistralité au montant des capitaux assurés par couple CR x Culture est prudente par rapport à la première solution proposée (Cap individuel).
- La limitation de la sinistralité au montant des capitaux assurés par couple CR x Culture permet de conserver un quantile du même ordre de grandeur que sans retraitement.
- Nous préconisons à l'équipe modèle interne de limiter la sinistralité au montant des capitaux assurés par couple CR x Culture.
- Nous recommandons d'utiliser la méthode « IF » pour éviter le problème de la non-significativité des régressions.

3.5. Autres travaux réalisés

3.5.1. Modélisation des rendements de culture

La validation a réalisé une étude approfondie (basée sur une étude réalisée par l'INRA[15]) pour déterminer la meilleure méthode pour la modélisation de la série temporelle du rendement de culture.

[15] Etude réalise par L'INRA en 2013 *« Evolution des rendements de culture à différentes échelles : Estimation des tendances passées et futures en tenant compte des incertitudes »*

Cette étude regroupe les méthodes les plus utilisées sur le marché pour la modélisation de la série des rendements :

- Régression Linéaire (L)
- Régression Quadratique (Q)
- Régression Cubique (C)
- Lissage exponentiel (méthode Holt-Winters : HW)
- Lissage exponentiel simple (LES).
- Linéaire dynamique (DLM)

L'objectif est de comparer ces méthodes et de déterminer la ou les meilleures méthodes pour analyser l'évolution des rendements.

Données utilisées :

La série de l'évolution de rendement de la culture blé depuis 1950 (à l'échelle nationale).

3.5.1.1. Régression linéaire (L)

Le modèle vérifie l'équation suivante :

$$Rendement = a + b * année + \varepsilon \,; avec \quad \varepsilon \sim N(0, \sigma^2)$$

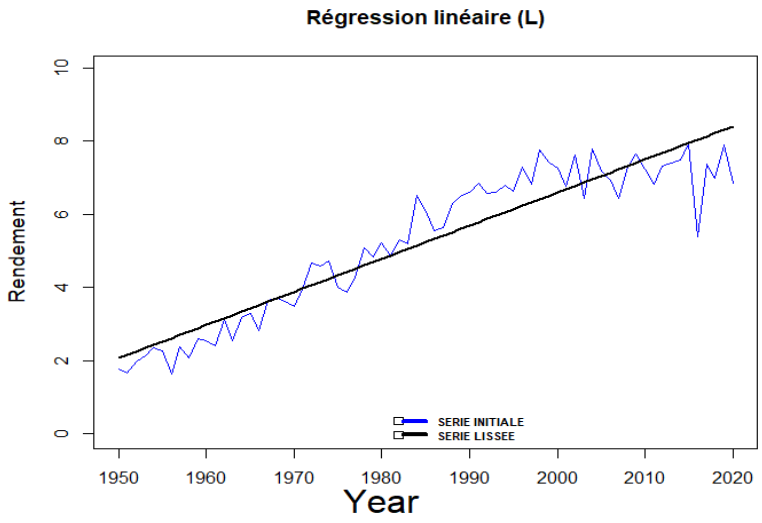


Figure 36 - Résultats de la régression linéaire

Nous observons que le modèle de régression linéaire s'ajuste bien avec les données.

3.5.1.2. Régression quadratique (Q)

Le modèle vérifie l'équation suivante :

$$Rendement = a + b * année + c * année^2 + \varepsilon \,; avec\ \varepsilon \sim N(0, \sigma^2)$$

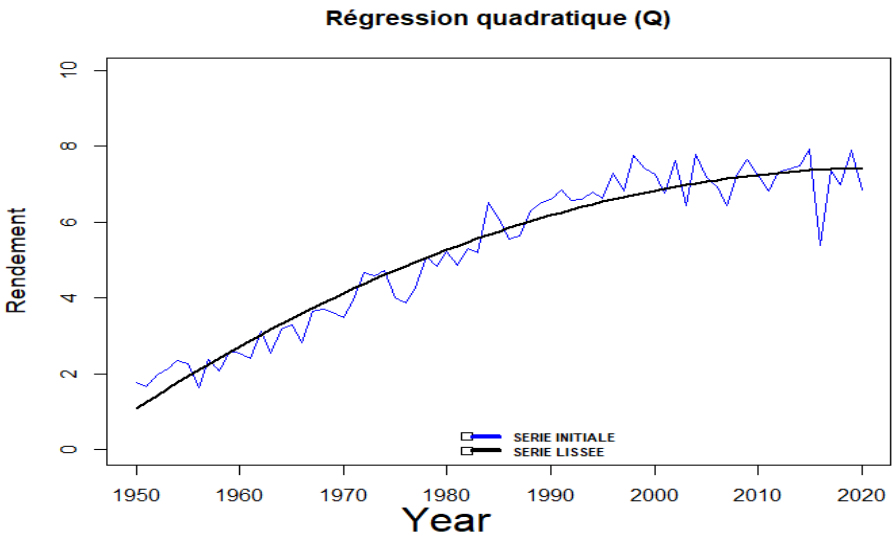


Figure 37 - Résultats de la régression quadratique

L'ajout d'une variable année au carré permet d'améliorer la qualité d'ajustement de modèle en passant d'un RMSE de 0.68 à 0.50 (voir le tableau de comparaison dans la partie 3.5.1.9).

3.5.1.3. Régression cubique (C)

Le modèle vérifie l'équation suivante :

$$Rendement = a + b * année + c * année^2 + d * année^3 + \varepsilon \; ; avec \; \varepsilon \sim N(0, \sigma^2)$$

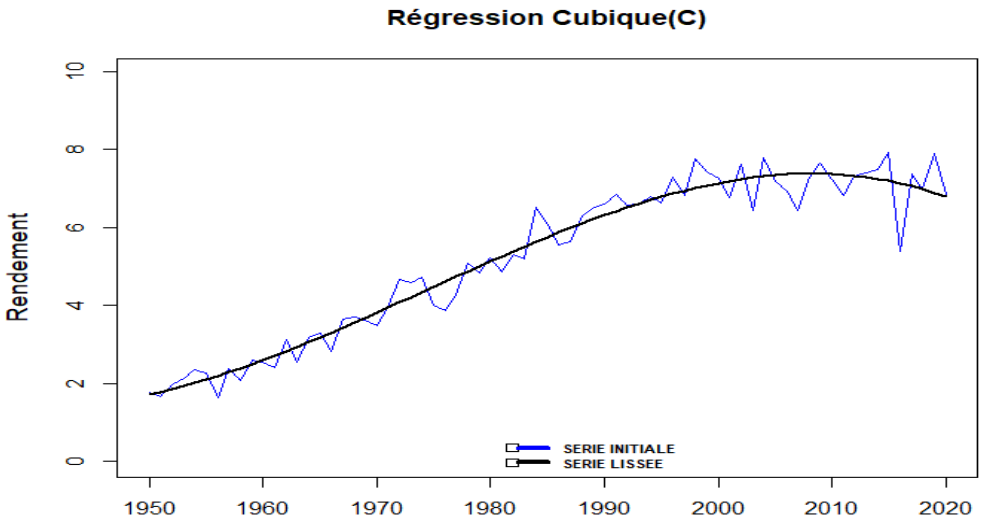


Figure 38 - Résultats de la régression cubique

Le modèle de régression cubique s'ajuste bien avec les données de la série des rendements.

3.5.1.4. Comparaison des régressions

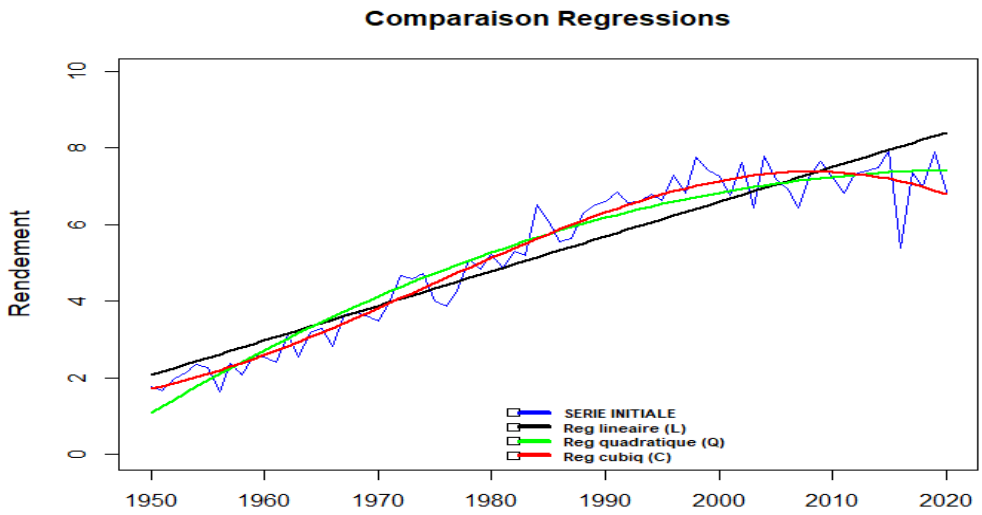


Figure 39 - Comparaison des 3 modèles précédents

Une comparaison entre ces trois modèles de régression a montré qu'ils sont pertinents, mais le modèle de régression cubique s'ajuste plus avec la série des données.

3.5.1.5. Lissage exponentiel (Méthode Holt-Winters : HW)

Le modèle vérifie l'équation suivante :

$$Y_{t+\Delta k} = m_t + b_t * \Delta t$$

Avec :

$m_t = \alpha_0 \times y_t + (1 - \alpha_0) \times Y_t$

y_t: Valeur réellement observée à l'instant t.

Y_t : Valeur prédite par le modèle à l'instant t.

Δt : écart de temps entre 2 rendements successifs.

$b_t = \alpha_1 \times (m_t - m_{t-1}) + (1 - \alpha_1) \times b_{t-1}$

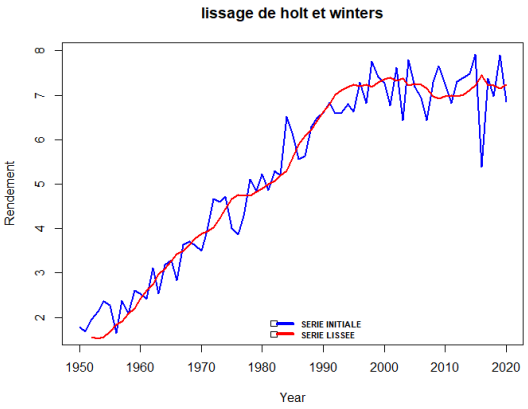
Le choix des paramètres de lissage $\alpha_0 et\ \alpha_1$ est en général effectué par l'utilisateur en fonction de la connaissance qu'il a de la série.

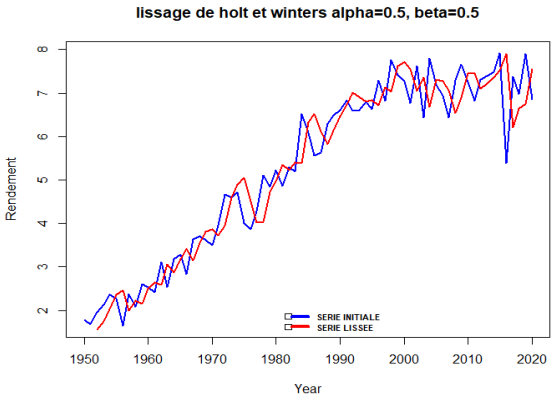
Nous pouvons également choisir les valeurs de $\alpha_0 et\ \alpha_1$qui minimisent la somme des carrés des erreurs de prévision en utilisant la fonction dans R *HoltWinters().*

1er cas : Les paramètres du modèle sont estimés par le modèle en utilisant la minimisation de la somme des erreurs de prédiction : α_0= 0,08 et $\alpha_1 = 1$.

2ème cas : Les paramètres du modèles sont fixés à :

α_0= 0,5 et $\alpha_1 = 0{,}5$.





Nous observons que le modèle de Holt-Winters suit l'évolution de la série initiale ce qui démontre la pertinence de ce modèle.

3.5.1.6. Lissage exponentiel simple (LES)

Le lissage exponentiel simple permet d'effectuer des prévisions pour des séries chronologiques :

$$\widehat{X_{T+1}} = \alpha \sum_{j=0}^{T-1} (1-\alpha)^j X_{T-j}$$

$\alpha \in [0,1]$ s'appelle la constante de lissage. L'initialisation est $\widehat{X_1} = X_1$.

À partir de la définition ci-dessus, nous obtenons directement :

$$\widehat{X_{T+1}} = \alpha X_T + (1-\alpha)\widehat{X_T}$$

Un problème important en pratique est celui du choix de la constante de lissage α qui est en général effectué par l'utilisateur en fonction de la connaissance qu'il a de la série. Nous pouvons également choisir la valeur de α qui minimise la somme des carrés des erreurs de prévision.

La fonction *ets*() sur R permet de modéliser les méthodes de lissages exponentiels avec une possibilité de choix de critère d'optimisation (maximisation de vraisemblance, MAE, RMSE, MASE ….).

1er cas : Le paramètre de modèle est estimé par le modèle en utilisant la minimisation de la somme des erreurs de prédiction : $\alpha = 0{,}44$

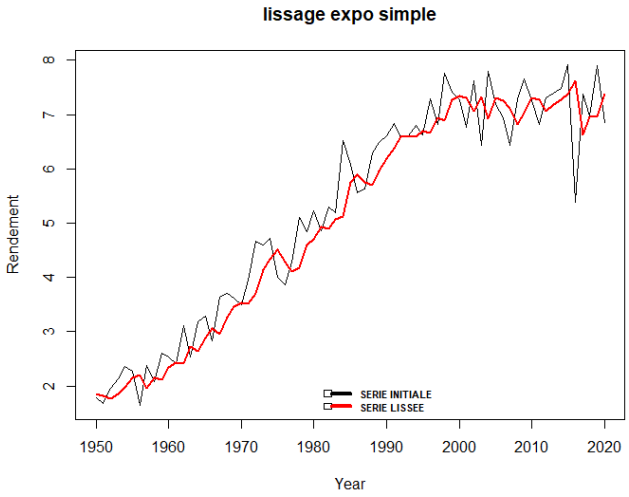


Figure 40 - Lissage exponentiel simple 1er cas

2ème cas : Fixer le paramètre de modèle à $\alpha = 0{,}5$

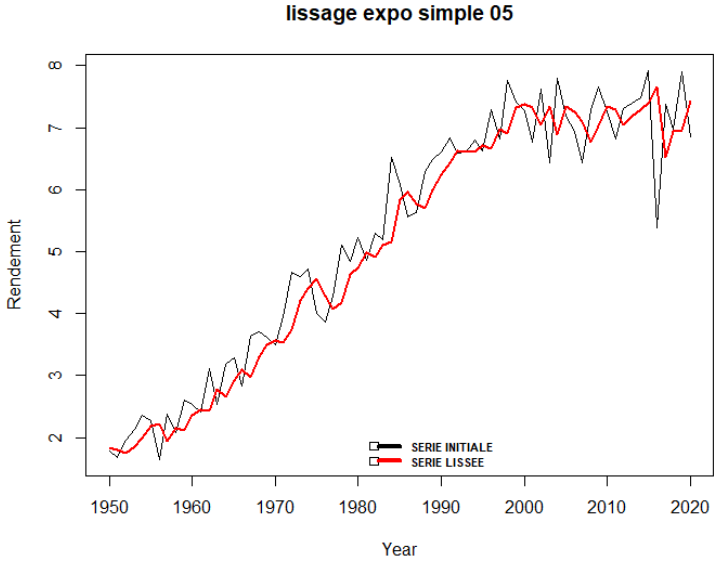


Figure 41 - Lissage exponentiel simple 2ème cas

Les observations sont les mêmes que celles des modèles de Holt-Winters.

3.5.1.7. Lissage exponentiel simple (LES) vs Holt-Winters

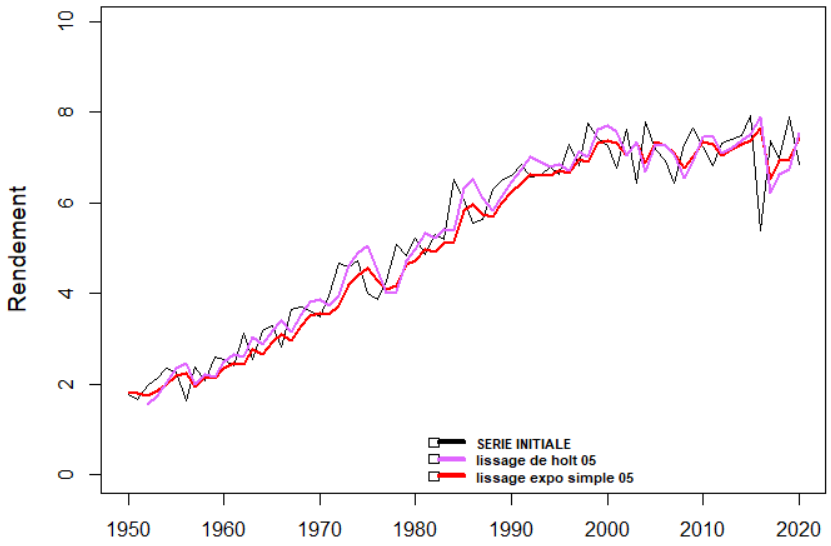


Figure 42 - Comparaison LES vs HW

Les deux modèles sont pertinents pour estimer l'évolution de la série étudiée.

3.5.1.8. Linéaire dynamique (DLM)

Les modèles linéaires dynamiques (ou modèle de filtre de Kalman) à espace d'état offrent un cadre puissant et flexible pour modéliser et analyser une très large gamme de phénomènes dynamiques. Ils sont très utilisés pour la modélisation des séries temporelles (*Harvey, 1990 ; Bentoglio et al., 2001*). En effet, ils constituent l'une des classes les plus simples des modèles à espace d'état.

Formellement, un modèle à espace d'état est constitué de deux équations : une équation d'état et une équation d'observation (ou de mesure). L'équation d'état formule la dynamique des variables d'état et l'équation d'observation relie les variables observées au vecteur d'état non observé.

Ces équations sont de la forme suivante :

- **L'équation d'état :**

$$\alpha_t = F\alpha_{t-1} + W_{t-1} \qquad W_{t-1} \sim N(0, Q),\ t = 1, \ldots, T$$

Où les W_t sont des vecteurs aléatoires indépendants de dimension p représentant le bruit d'état, de moyenne nulle et de matrice de covariance Q. F est une matrice de dimension (p×p) dite matrice d'état ou matrice de transition. Nous considérons généralement une distribution normale a priori sur le vecteur d'état à l'instant t = 0, de moyenne m_0 et de matrice de covariance C_0 : $\alpha_0 \sim$ N (m_0, C_0).

- **L'équation d'observation ou de mesure :**

x_t = A α_t + v_t, $\quad v_t \sim N(0, R)$, t = 1..T

Où v_t est un vecteur aléatoire de dimension d représentant le bruit de mesure, de moyenne nulle et de matrice de covariance R. A est une matrice de dimension (d × p) dite matrice d'observation.

3.5.1.8.1. Filtre de Kalman

Le filtre de Kalman fait partie des innovations les plus remarquables du 20e siècle. C'est un ensemble d'équations récursives qui ont pour objectif de trouver des estimateurs linéaires optimaux du vecteur d'état en fonction de toutes les observations, en minimisant l'erreur quadratique moyenne.

3.5.1.8.2. Présentation de l'algorithme

Dans le cadre des modèles à espace d'état, l'une des utilisations principales du filtre de Kalman est l'estimation des **variables d'état**. Avant de détailler le calcul du filtre de Kalman, il convient de définir les notations suivantes :

$$\alpha_t^{t-1} = E(\alpha_t | x_{1:t-1}),$$

$$P_t^{t-1} = V(\alpha_t | x_{1:t-1}),$$

Où α_t^{t-1} représente l'espérance de l'état αt sachant toutes les données connues jusqu'à la date t − 1, notées $x_{1:t-1}$, et P_t^{t-1} t désigne la matrice de covariance de l'état α_t sachant $x_{1:t-1}$.

Partant des valeurs initiales α_0 et P_0, le filtre de Kalman est défini par les deux phases suivantes :

- Prédiction (forecasting) : Cette étape consiste à rechercher la meilleure prédiction de l'état courant sachant toutes les informations précédentes.

- Filtrage (filtering) : Lorsque l'observation à l'instant t est disponible, une mise à jour de la prédiction et de la covariance précédente est effectuée en utilisant cette observation.

3.5.1.8.3. Algorithme de Filtre de Kalman

Algorithme 3: Filtre de Kalman

Initialisation : $\boldsymbol{\alpha}_0, P_0$

pour $t = 1, \ldots, T$ **faire**

Prédiction :

$$\begin{aligned} \boldsymbol{\alpha}_t^{t-1} &= F\,\boldsymbol{\alpha}_{t-1}^{t-1}, \\ P_t^{t-1} &= F\,P_{t-1}^{t-1}\,F' + Q. \end{aligned}$$

Calcul du gain de Kalman :

$$K_t = P_t^{t-1} A' (A\,P_t^{t-1} A' + R)^{-1}.$$

Mise à jour :

$$\begin{aligned} \boldsymbol{\alpha}_t^t &= \boldsymbol{\alpha}_t^{t-1} + K_t \sum_{i=1}^{n_t} (\mathbf{x}_{ti} - A\,\boldsymbol{\alpha}_t^{t-1}), \\ P_t^t &= [\mathbf{I} - K_t A] P_t^{t-1}. \end{aligned}$$

Sortie : $\{(\boldsymbol{\alpha}_t^t, P_t^t); t = 1, \ldots, T\}$.

3.5.1.8.4. Modèle avec tendance dynamique

Dans notre étude, nous allons tester un modèle linéaire dynamique avec tendance dynamique (basé sur le filtre de Kalman) pour modéliser la tendance de la série de rendement.

- L'équation d'observation :

  X_t = A Z_t + ε_t Avec A = (1, 0) $Z_t = \begin{pmatrix} a_t \\ b_t \end{pmatrix}$, $\varepsilon_t \sim N\,(0, \sigma_\varepsilon^2)$

  Donc :

  $$X_t = a_t + \varepsilon_t$$

- L'équation d'état :

  $Z_t = FZ_{t-1} + W_{t-1}$ Avec $W_{t-1} \sim N\,(0, \Sigma)$ $F = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ $\Sigma = \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix}$

  Donc :

  $a_t = a_{t-1} + b_{t-1} + W_{t-1}^a$ $W_{t-1}^a \sim N\,(0, \sigma_a^2)$

  $b_t = b_{t-1} + W_{t-1}^b$ $W_{t-1}^b \sim N\,(0, \sigma_b^2)$

Nous avons donc testé ce type de modèle sur les données de rendement des blés depuis 1950.

La comparaison entre les données initiales et celles issues du modèle testé donne les résultats suivants :

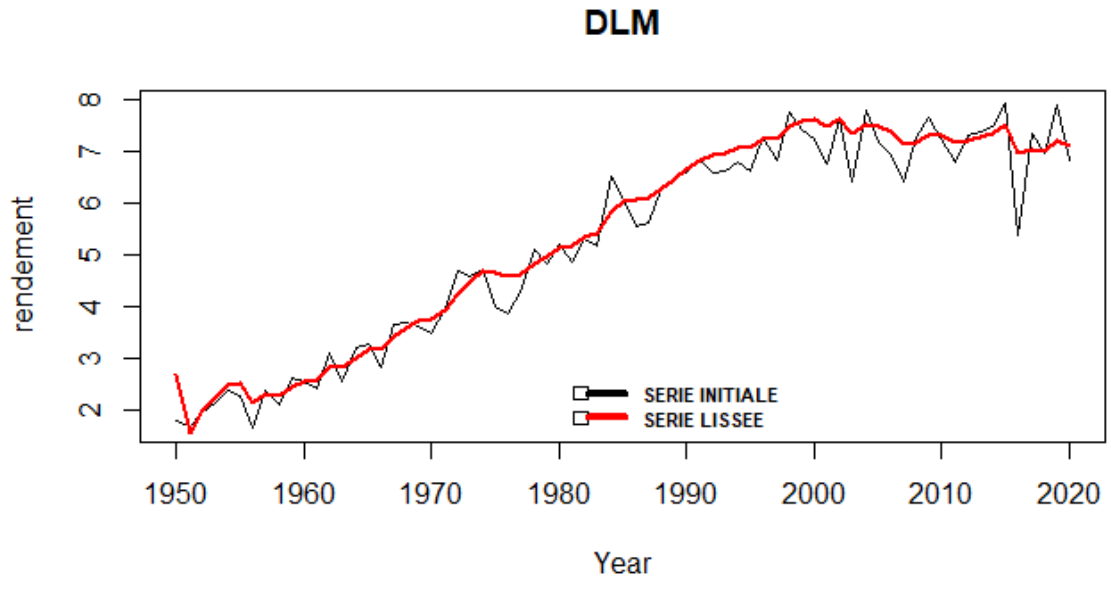


Figure 43 - Comparaison entre les données initiales et celles issues du modèle testé

À l'analyse de ce graphique, nous observons que le modèle linéaire dynamique a une bonne qualité d'ajustement par rapport aux autres modèles testés.

3.5.1.9. Comparaison des modèles

Pour comparer les modèles, nous avons utilisé l'indice RMSE qui représente la racine de la moyenne arithmétique des carrés des écarts entre les prévisions du modèle et les observations.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^{pred})^2}$$

modèle	RMSE
Régression Linéaire (L)	0.68
Régression Quadratique (Q)	0.50
Régression Cubique (C)	0.43
méthode Holt-Winters: HW (alpha =0.08, beta=1)	0.50
méthode Holt-Winters: HW (alpha =0.5, beta=0.5)	0.61
Lissage exponentiel simple (LES) (alpha = 0.44)	0.54
Lissage exponentiel simple (LES) (alpha = 0.5)	0.55
Linéaire dynamique (DLM)	0.39

Figure 44 - Comparaison des RMSE de chaque méthode

Le meilleur modèle permettant d'estimer la tendance de la série nationale des rendements [du blé tendre de 1950 à 2020] est le Modèle Linéaire dynamique (DLM).

3.5.2. Modélisation de la MRC (Modèle Linéaire Dynamique vs Régression Linéaire)

Comme expliqué précédemment, nous avons réalisé une étude pour choisir le meilleur modèle qui estime la tendance de la série des rendements. Nous avons donc choisi le modèle linéaire dynamique qui a une bonne qualité d'ajustement par rapport aux autres modèles testés.

L'objectif de cette partie est d'intégrer ce modèle (modèle linéaire dynamique) dans la modélisation de la MRC pour le comparer avec le modèle de régression linéaire (choisi par l'équipe modélisation). Cela va nous aider à démontrer la pertinence du choix de l'équipe modélisation et par la suite renforcer la modélisation du risque de la MRC.

Nos objectifs sont de :

- Appliquer le modèle de régression linéaire et le modèle linéaire dynamique sur les rendements de chaque maille CR x Culture.
- Comparer ces deux méthodes en utilisant des indices statistiques.
- Choisir le meilleur modèle qui représente la tendance des rendements au niveau de la CR x Culture.

Pour comparer les modèles, nous avons utilisé le même indice RMSE présenté ci-dessous qui représente la racine de la moyenne arithmétique des carrés des écarts entre les prévisions du modèle et les observations.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^{pred})^2}$$

Les tableaux suivants représentent des statistiques sur le choix de meilleur modèle pour chaque couple CR x Culture en se basant sur l'indice RMSE :

En nombre	RMSE
Modèle dynamique linéaire	55
Régression linéaire	132
Total	187

En %	RMSE
Modèle dynamique linéaire	29%
Régression linéaire	71%
Total	100%

Le tableau ci-dessous montre les résultats des simulations pour la modélisation de MRC pour ces deux méthodes, pour chaque couple Caisse x Culture soit 187 couples au total.

	Moyenne	Quantile 99,5
Modèle linéaire dynamique	105 811	903 247
Régression linéaire	93 046	493 533
Ecart	-12,06%	-45,36%

Le modèle linéaire dynamique est plus prudent que la régression linéaire mais cette dernière est plus pertinente pour estimer la tendance des rendements au niveau de la maille CR x Culture.

La régression linéaire est meilleure que le modèle linéaire dynamique pour 71% des couples CR x Culture. La non-pertinence des modèles linéaires dynamiques dans les 71% des cas est justifiée par la complexité de ces modèles et l'impossibilité d'avoir un modèle avec les mêmes paramètres (même degrés de modèle et même point de départ) qui est valide pour tous les couples Caisse x Culture.

De plus, le modèle linéaire dynamique est une méthode d'estimation intéressante pour la modélisation des séries temporelles, mais qui n'est utilisable que lorsque l'on peut décrire assez précisément notre système. En effet, le modèle linéaire dynamique comme nous avons présenté ci-dessus est constitué par un système des deux équations (équation d'état et équation d'observation) et le développeur a besoin de modéliser le système assez précisément pour chaque type de série afin de désigner un modèle efficace. S'il est impossible de trouver une modélisation correcte du système, il est alors préférable de se tourner vers d'autres méthodes de prédiction tel que la régression linéaire.

Dans notre cas nous avons presque 187 séries de rendements pour chaque couple Caisse x Culture, donc il est difficile et même impossible d'avoir un système d'équation efficace sur toutes ces séries vu la diversité des comportements de ces dernières (voir un exemple des séries ci-dessous).

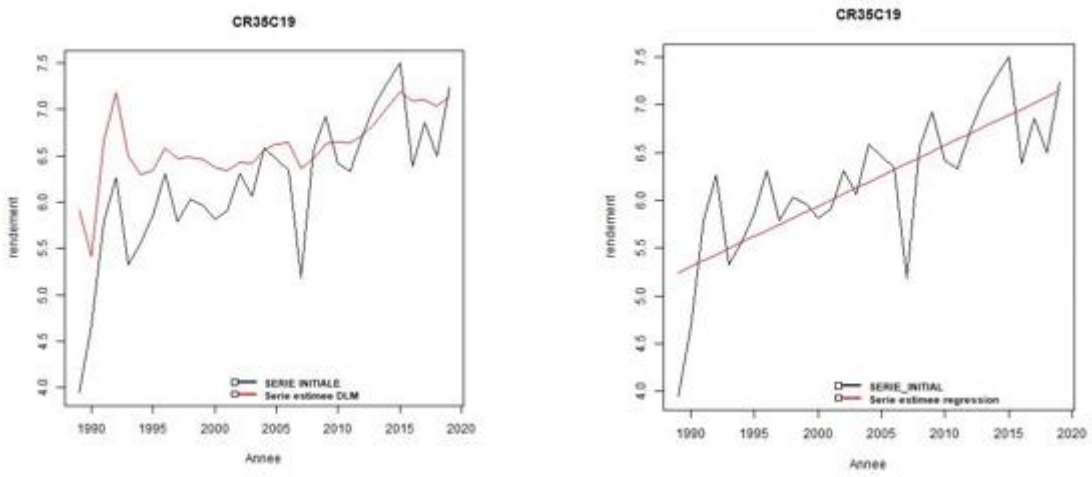


Figure 45 – Exemple de régression DLM vs régression linéaire

Dans l'exemple ci-dessus, nous remarquons que le modèle linéaire dynamique surestime toujours la série modélisée contrairement à la méthode de la régression linéaire qui s'intéresse à trouver une droite d'ajustement passant au milieu des points de la série modélisée et donc minimise les distances qui la séparent des points.

Nous concluons donc que la régression linéaire est meilleure que le modèle linéaire dynamique en termes de simplicité et d'ajustement pour estimer la tendance des rendements à la maille CR x Culture. Ceci confirme donc la pertinence du choix de l'équipe modélisation.

3.6. Conclusion

3.6.1. Synthèse de la validation

Sur la base de ces différents travaux, les principales conclusions et préconisations de la validation sont les suivantes :

- Deux tiers des régressions sont significatifs pour un niveau de risque de 10%. Cela montre qu'aucune tendance n'est observée pour 1/3 des régressions.

- Nous proposons donc de présenter la significativité des régressions en fonction du capital assuré afin d'expliquer la non-significativité de ces régressions et de vérifier l'importance de ces régressions en termes des capitaux assurés.

- La partie du capital assuré pour les régressions non significatives reste importante pour les deux niveaux de risque (5%, 10%) :
  - Pour un niveau de risque de 5%, les régressions non significatives représentent presque 41% du capital assuré.
  - Pour un niveau de risque de 10%, les régressions non significatives représentent presque 32% du capital assuré.

- Nous proposons donc d'utiliser la méthode écart à la moyenne pour les régressions non significatives. Autrement dit, nous utilisons une combinaison des deux méthodes : si la

régression est significative nous utilisons une régression simple pour calculer les résidus, sinon nous utilisons la méthode écart à la moyenne pour calculer les résidus.

- **La majorité des régressions linéaires vérifient les trois hypothèses de la régression linéaire** (75% pour un niveau d'erreur de 5%). Cela confirme la pertinence de cette méthode.

- La méthode IF Régression Moyenne permet d'éviter le problème de la non-significativité de certaines régressions.

- Nous observons des écarts négligeables entre la régression linéaire tout historique et la méthode IF.

- La limitation de la sinistralité au montant des capitaux assurés par couple CR/Culture est prudente par rapport à la première solution proposée (Cap individuel).

- Le meilleur modèle permettant d'estimer la tendance de la série des rendements à **l'échelle nationale** est le modèle Linéaire dynamique (DLM).

- La régression linéaire est meilleure que le modèle linéaire dynamique pour estimer la tendance des rendements à la maille **CR/Culture**

- Nous confirmons donc la pertinence du choix de l'équipe modélisation.

4. Validation de la tarification de la MRC

4.1. Contexte de l'étude

Dans le cadre du renforcement des modèles de groupe, nous avons réalisé une revue du modèle tarifaire Multi Risques Climatiques. L'objectif principal de cette validation est de revoir les parties méthodologiques et pratiques de ce modèle.

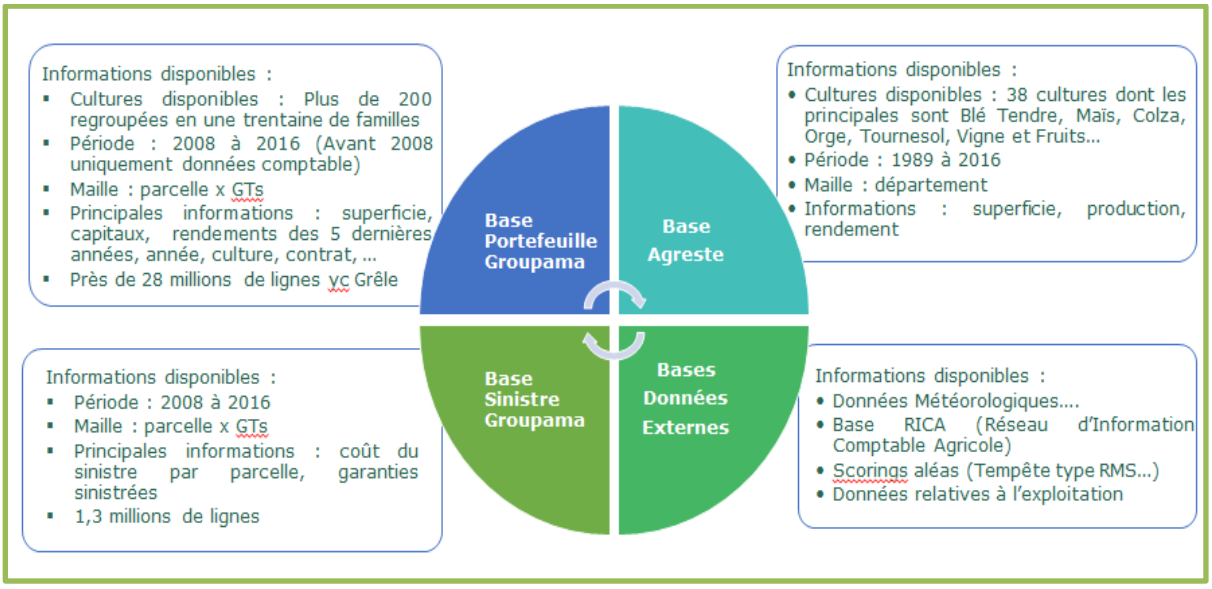
Notre approche de revue pour le modèle tarifaire a consisté à réaliser les travaux suivants :

- Revue de la constitution de la base de données.
- Revue de la méthodologie de la sélection des variables et les variables retenues.
- Revue du calcul de la prime pure.

4.2. Base de données

4.2.1. Constitution de la base

4.2.1.1. Inputs



4.2.1.2. Périmètres

- Séparation des risques grêle et autres aléas

Le produit MRC permet de couvrir deux types de risques : le risque grêle et le risque autres aléas. La grêle frappe de manière très concentrée géographiquement. Ainsi il se peut qu'une parcelle soit touchée et que celle juste à côté ne le soit pas. Les agriculteurs assurent donc le plus souvent leurs cultures avec une franchise à la parcelle. À l'inverse les autres aléas comme la sécheresse ou l'inondation touchent une zone large. La franchise des garanties couvrant ce type de risque est à la nature de récolte. Ces deux risques Grêle et Autres Aléas vont donc être étudiés de manière différente

avec deux modèles différents. C'est au risque Autres Aléas que nous allons nous intéresser dans ce rapport.

- Contrats Seuil 25 Franchise 25

Le modèle de tarification est construit uniquement avec les données des contrats seuil 25 franchise 25. En effet, c'est celui qui est le plus vendu pour le risque autres aléas. Les agriculteurs bénéficient d'une subvention de l'Etat sur une partie de leur prime lorsqu'ils souscrivent ce type de contrat, ce qui les incite à le choisir. Les autres niveaux de franchise ne sont toutefois pas oubliés. Une fois le tarif pour les contrats seuil 25 franchise 25 établi, des coefficients de franchise sont construits pour chacun des autres niveaux de franchise pour obtenir les primes adaptées.

4.2.2. Retraitement des données

L'équipe tarification commence par créer la base qui servira à l'étude, elle est constituée :

- Des données portefeuille
- Des données sinistres
- Des données externes climatiques

Nous commençons avec les bases sinistres et portefeuille 2008-2015 et 2016. Les tables sont concaténées pour avoir une base sinistre et une base portefeuille 2008-2016, en enlevant les variables qui étaient disponibles uniquement entre 2008 et 2015 mais pas après et inversement.

Les bases portefeuille et sinistres sont filtrées pour rentrer dans le périmètre d'étude en enlevant les sinistres grêles et en retirant les contrats à un niveau de franchise 'Autres Aléas' différent de seuil 25 franchise 25.

Ces deux bases sont fusionnées à l'aide d'une clé : Assuré x Année x Culture, en enlevant les doublons. Du côté de la validation, nous nous sommes intéressés à la raison de la présence de ces doublons et ce qu'impliquait le fait de les supprimer. Ils sont en fait dus à certaines variables : « *identifiant technique GT souscrite* », « *identifiant catalogue module* », et « *identifiant catalogue GAR tarifable* ». Un contrat donné est constitué dans la base informatique d'autant de lignes qu'il possède de garanties élémentaires tarifables. C'est la somme de ces garanties élémentaires, elles-mêmes regroupées par modules, qui constituent le contrat. Or la base de données servira uniquement à l'analyse de la fréquence des sinistres. Avoir plusieurs lignes pour chaque contrat n'est donc pas nécessaire, une seule est suffisante pour indiquer l'occurrence d'un sinistre.

Les contrats sans sinistres sont conservés : une variable « *TOP SIN »* est créée égale à 0 ou 1 qui indique s'il y a eu des coûts d'indemnisation ou non. Une variable « *métagroupement »* est formée à l'aide du type de culture et de la saison ainsi : culture_saison.

Le taux de rapprochement des sinistres est d'environ 92 ,1%. 21961 sinistrés n'ont pas été rapprochés à un contrat ce qui correspond à 92,7% des coûts d'indemnisation, ce qui est un bon chiffre pour continuer l'étude. Il faudrait cependant s'intéresser à la cause de ces non-rapprochements qui peuvent biaiser le tarif s'ils sont trop fréquents pour certaines cultures. La base 'contrats' ayant été filtrée pour

ne conserver que ceux à seuil et franchise 25%, ces sinistres n'ont pas été rapprochés notamment parce qu'ils correspondent à des contrats souscrits à d'autres niveaux de franchise. Le cas d'un sinistre total précoce au semis (pas de levée des semences), qui implique de ressemer une autre culture sur les parcelles concernées et donc un numéro de version de contrat différent, pourrait être une autre cause de non-rapprochement, qu'il conviendrait de quantifier.

Les données climatiques par commune x année sont ajoutées, puis les données vents. Une base de données avec des statistiques sur les vents dans la période février-juillet est créée. C'est la période durant laquelle les cultures sont exposées aux vents. En dehors de cette période, les cultures sont soit encore sous terre ou à l'état de graine (cultures semées en automne), soit non encore semées (cultures semées au printemps) soit ne portent pas de potentiel de récolte à cette période (vigne, arbres fruitiers). En tout état de cause, elles ne sont pas vulnérables à ce risque en dehors de la période sélectionnée.

Remarques de la validation :

Il est recommandé de s'intéresser aux sinistres non rapprochés et de détailler précisément l'hypothèse des changements de cultures en cours d'année.

4.2.3. Etudes des sinistres non rapprochés

4.2.3.1. Etude générale toutes cultures

Lors du rapprochement de la base sinistre sur la base portefeuille pour créer la base d'étude, certains sinistres ne sont rapprochés à aucun contrat. Comme expliqué précédemment, cela peut être dû à des changements de culture en cours d'année. Mais il est important de vérifier que ces sinistres non rapprochés ne biaisent pas la fréquence de sinistres pour certaines cultures. En effet, la base d'étude va être utilisée pour le modèle GLM qui va donner des indicateurs de risque de sinistre pour chaque assuré mais aucune information sur le montant de ces sinistres. Nous ne nous intéresserons donc pas dans cette étude aux couts des indemnisations des sinistres non rapprochés, seulement à leur nombre.

Cette étude va être effectuée à l'aide d'un code R. Nous commençons par récupérer la base des sinistres non rapprochés. Nous gardons qu'une seule garantie par contrat pour ne pas compter plusieurs fois le sinistre d'un assuré avec plusieurs garanties, seule l'occurrence d'un sinistre nous intéresse. Ensuite nous les comptons et regardons la proportion par culture parmi tous les sinistres. C'est-à-dire, pour une culture donnée quelle proportion de sinistre est non rapprochée.

Nous obtenons le tableau suivant pour les capitaux les plus importants :

Culture	Capitaux (M€)	Nombre de sinistres non rapprochés	Proportion de sinistres non rapprochés	Nombre de sinistres total
BLE TENDRE Hiver	1 300	2 575	4%	57 581
VITI	786	6 375	35%	18 241
MAIS Printemps	689	2 377	5%	50 345
COLZA Hiver	413	2 215	6%	34 228
ORGE Hiver	317	2 149	8%	26 038

Nous remarquons que certaines cultures ont un pourcentage élevé de sinistres non rapprochés. Pour la culture VITI (vignes) par exemple, 35% de ses sinistres ne sont pas rapprochés à un contrat. Les cultures dans la base sinistres non rapprochés sont celles réellement sinistrées, et celles dans la base portefeuille sont les cultures de remplacement. Donc ces 35% de sinistres sur la culture VITI non rapprochés à un contrat du portefeuille représentent bien des sinistres VITI. Il est important de s'y intéresser afin d'identifier la cause du non-rapprochement.

4.2.3.2. Etude pour la culture VITI

Nous commençons par déterminer la cause du non-rapprochement de ces 6375 sinistres. Le rapprochement est fait à l'aide d'une clé contenant les variables « *identifiant contrat* », « *année* » et « *code culture* ». Comme expliqué précédemment, le non-rapprochement peut être causé par des changements de culture en cours d'année et donc être causé par la variable *« code culture »*. Or les vignes sont une culture pérenne, c'est-à-dire qu'elles sont permanentes et ne peuvent pas être remplacées. Cette variable ne peut donc normalement pas en être la cause. À l'aide d'un code R, nous nous apercevons que sur les 6375 sinistres non rapprochés, 6368 ne sont pas rapprochés car leur identifiant contrat ne se trouve pas dans la base portefeuille. C'est donc la variable *« identifiant contrat »* qui est responsable du non-rapprochement des sinistres VITI. Cela est peut-être dû à certaines pratiques des caisses régionales lors de la souscription et de l'enregistrement des sinistres qui provoquent des identifiants contrats différents. Nous allons donc nous intéresser à la localisation de ces sinistres non rapprochés pour peut-être identifier une ou plusieurs caisses régionales.

Dans la base des sinistres, nous avons comme indice de localisation le code commune INSEE. La maille commune est trop petite pour cette étude, nous allons donc créer une colonne région à l'aide d'une table de correspondance « code commune INSEE / Région ». Une fois la variable « *Région* » créée, on va pouvoir s'intéresser comme précédemment au nombre de sinistres VITI par région, le nombre de sinistres VITI non rapprochés par région, ainsi que la proportion de non rapprochés par région. Nous obtenons le tableau suivant pour les régions les plus importantes :

Région	Nombre de sinistres VITI non rapprochés	Proportion de sinistres non rapprochés VITI	Nombre de sinistres totaux VITI
BOURGOGNE	3 354	72%	4 632
RHONE-ALPES	2 216	53%	4 179
LANGUEDOC-ROUSSILLON	279	7%	4 146
MIDI-PYRENEES	55	3%	1 694
PROVENCE-ALPES-COTE D'AZUR	54	15%	369

Nous constatons que les régions Bourgogne et Rhône-Alpes ont un pourcentage de sinistres non rapprochés élevé (72% et 53 %) pour un nombre de sinistres totaux élevé (4632 et 4179). Ces deux régions sont donc principalement responsables du non-rapprochement des sinistres VITI.

Nous pouvons aussi nous intéresser directement aux statistiques des caisses régionales à l'aide de la variable *« Numéro caisse régionale »* dans la base sinistre. Nous obtenons ce tableau de résultats :

Code caisse régionale	Nombre de sinistres VITI non rapprochés	Proportion de sinistres non rapprochés VITI	Nombre de sinistres totaux VITI
Caisse 1	5 030	75%	6 663
Caisse 2	629	11%	5 953
Caisse 3	65	3%	2 039
Caisse 4	29	2%	1 305
Caisse 5	129	13%	958
Caisse 6	379	55%	694
Caisse 7	10	3%	334
Caisse 8	104	35%	295

Nous retrouvons bien la caisse régionale numéro 1 avec 75% de sinistres VITI non rapprochés pour 6663 sinistres VITI. Ainsi que la caisse 6 avec 55% de sinistres VITI non rapprochés pour 694 sinistres VITI.

Après échanges avec l'équipe modèle tarifaire, ces sinistres sembleraient ne pas appartenir à la franchise 25%.

Remarques de la validation :

Nous recommandons d'approfondir l'analyse des sinistres non rapprochés afin de s'assurer que la modélisation n'est pas biaisée. C'est à dire de quantifier la part des sinistres écartés liés au changement de cultures et qui peuvent être dans le périmètre de l'étude versus les contrats qui ne sont pas souscrites au seuil/franchise 25%.

4.2.4. Sélection des variables

4.2.4.1. Méthode Random Forest

Dans le cadre de la tarification, l'équipe tarification a effectué une étude de la fréquence des sinistres. L'objectif étant de déterminer le risque de sinistre pour un contrat donné, en fonction de ses variables internes (variables portefeuille) et externes (climatiques). Pour cela, il a d'abord fallu sélectionner les variables explicatives utiles à cette étude. C'est la méthode du Random Forest ou Foret Aléatoire qui a été choisie afin de repérer les variables les plus importantes dans la prévision de l'occurrence d'un sinistre.

4.2.4.1.1. Principe des Random Forest

Les « forêts aléatoires » ou « Random Forest » sont un outil de Machine Learning permettant de prédire une certaine variable appelée variable cible, en fonction de variables explicatives. Pour ce faire, l'algorithme va classer les différents enregistrements de notre base de données à l'aide des variables explicatives de manière optimale c'est-à-dire en créant des groupes d'individus les plus homogènes possibles, des groupes dans lesquels la majorité des individus ont la même valeur de variable cible. En réalisant cette classification, certaines variables vont se révéler être plus efficaces que d'autres dans l'homogénéisation des groupes et par conséquent dans la prédiction de la variable cible. L'algorithme va pouvoir mesurer cette efficacité pour chacune des variables explicatives et ainsi donner un indice sur la significativité de ces dernières, appelé « importance ». Les variables avec les importances les plus élevées seront celles qui auront le plus d'impact dans la détermination de la variable cible, celles à garder dans le modèle.

4.2.4.1.2. Création de la base pour la modélisation

Le fichier d'origine est le fichier « AUTRES_ALEA_M_E.rds ». Il contient les données portefeuille et sinistres des assurés avec les données météorologiques auxquelles nous allons ajouter les données vents du fichier « DATA_VENT_STATS.csv ». Chaque ligne du tableau correspond à un type de culture pour un assuré pour une année, sur une commune donnée (le contrat est supposé souscrit entièrement sur cette commune, cf § Détermination de MAXINSEE du chapitre 5). Les variables inutiles à la prévision du sinistre sont éliminées, ce sont principalement des informations sur les contrats des assurés comme *identifiant_technique_gt_souscrite* ou *kapital.*

Nous avons donc un tableau d'un peu plus d'un million d'observations et de 119 variables qui se présente ainsi :

Avec des données climatiques :

Commune	Amplitude thermique Journalière automne	Amplitude thermique Journalière été	Amplitude thermique Journalière hiver
01001	7,81	9,82	5,28
01002	8,19	10,48	6,09
01004	7,94	10,16	5,71
01005	7,80	9,88	5,31
01006	7,77	10,11	5,73
01007	8,05	10,29	5,85
01008	7,94	10,16	5,71
01009	7,88	10,17	5,74
01010	7,48	9,69	5,44
01011	7,31	9,46	5,48

Des données portefeuille :

Identifiant contrat	Surface exploitation	Nombre de parcelles	Saison	Type de culture
1	10	1	Hiver	BLE TENDRE
2	20	1	Hiver	BLE TENDRE
3	40	2	Hiver	BLE TENDRE
4	38	3	Hiver	BLE TENDRE
5	11	2	Hiver	BLE TENDRE
6	16	2	Hiver	BLE TENDRE
7	20	2	Hiver	BLE TENDRE
8	22	2	Hiver	BLE TENDRE
9	10	1	Hiver	BLE TENDRE

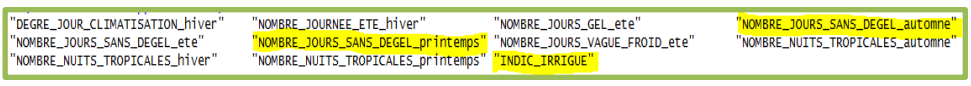
Les quatre variables VENT :

Commune	Quantile 90% des rafales moyennes	Moyenne des rafales max et second max	Nombre de rafales > 70	Nombre de rafales > 100
01001	56	69	75	7
01002	52	63	52	0
01004	54	65	49	3
01005	58	71	83	6
01006	46	58	30	3
01007	55	66	65	2
01008	53	64	47	3
01009	45	57	33	2
01010	52	63	60	2
01011	51	61	48	1

Tous les types de cultures sont distingués en fonction de leur période de récoltes (hiver ou printemps). Nous pouvons voir les trois premières variables climatiques, elles sont identiques pour ces 6 contrats car ils se situent tous dans la même commune (la maille utilisée pour les données climatiques).

Ensuite, les variables qui ont une variance proche de zéro sont jugées inutiles car elles n'apportent pas assez d'information. Elles sont donc retirées à l'aide d'une fonction R *nearZeroVar.* Il peut être dangereux de supprimer des variables avec de faibles variances. En effet une variable avec de faibles valeurs aura forcément une faible variance. Elle pourrait alors être supprimée même si elle est importante dans la prévision des sinistres. Heureusement, cette fonction ne calcule pas une variance classique mais va plutôt s'intéresser à la fréquence d'apparition de la valeur la plus représentée. Si celle-ci est trop élevée alors la variable va être supprimée. Aucune variable n'est donc supprimée à tort. Seulement celles qui n'apportent pas d'information car presque tous les contrats ont la même valeur.

Les variables supprimées sont :



Cependant, les variables en jaune sont tout de même conservées. Ce sont celles que l'équipe tarification et la Direction Métier ont jugées logiquement influentes dans notre étude malgré leur manque de variabilité. En effet, le gel est une des causes de sinistres autres aléas, le nombre de jour sans dégel devrait donc avoir une importance dans le modèle. L'indice d'irrigation, égal à 0 ou 1 indique si la culture est irriguée ou non par l'agriculteur. C'est une protection contre la sécheresse et donc une variable importante dans cette étude.

Les variables quantitatives sont découpées en 5 classes pour le Random Forest, à l'aide d'une fonction sur R qui permet de choisir ces classes afin qu'elles soient toutes également représentées par les contrats. Le nombre 5 est choisi car il permet de limiter le nombre de modalités à étudier toute en gardant une précision raisonnable. Ce nombre de classes pourra être ensuite modifié pour certaines variables par choix d'expert. Par exemple, deux classes juxtaposées présentant le même risque de sinistre vont être regroupées en une.

Les contrats avec NA pour une de leurs variables sont supprimés. Cela revient à supprimer 30 000 lignes sur les 1 160 000 originales (2% des observations), ce qui reste largement suffisant pour effectuer le Random Forest.

Ce n'est pas le climat de l'année entière qui va impacter la qualité des récoltes. Uniquement certaines saisons sont pertinentes en fonction de la date de la récolte. Si c'est une culture de printemps, on aura besoin des données climatiques des saisons de printemps et d'été. Si c'est une culture d'hiver ce seront celles des saisons de printemps et d'automne. Les variables climatiques relatives à la saison d'hiver sont donc supprimées car inutiles dans l'étude de la sinistralité des récoltes.

4.2.4.1.3. Significativités des variables

L'objectif est d'analyser le risque de sinistre de chaque contrat à l'aide de ses caractéristiques (culture, surface, ...) et des données climatiques de sa commune. Nous voulons donc nous intéresser à l'importance de chacune des variables dans la prédiction de le variable *TOP_SIN*, égale à 1 s'il y a eu sinistre pour cet assuré cette année pour ce type de culture, et 0 sinon. Cette variable sera donc la variable cible, les autres seront les variables explicatives.

Deux groupes de cultures sont distingués : les cultures de printemps, et les cultures d'hiver. En effet ces deux types de cultures sont récoltés à des périodes différentes et auront donc des causes de sinistres climatiques différentes. Les significativités des variables sont donc différentes pour chacune des deux périodes de récoltes.

- Importances globales

Les importances globales par groupe de culture sont d'abord étudiées. Pour un groupe de culture donné (hiver ou printemps), un Random Forest est effectué pour chacune des cultures du groupe en question, avec les variables climatiques prises sur la bonne période. Nous obtenons alors une importance pour chaque variable pour chaque culture du groupe que nous allons sommer puis diviser par le nombre de cultures dans ce groupe pour obtenir finalement une importance de chaque variable pour le groupe de culture donné.

Par exemple, pour les cultures d'hiver, sont prises dans le Random Forest uniquement les cultures d'hiver : blé tendre hiver, orge hiver, colza hiver avec les variables climatiques du printemps et de l'automne. Les importances sont sommées pour chaque variable pour toutes les cultures et nous les divisons par 3 (soit le nombre de cultures dans ce groupe des cultures d'hiver) pour obtenir l'importance de chaque variable sur la saison hiver.

L'équipe tarification a fourni dans son rapport un graphe des importances mais sans les variables VENT qui sont arrivées plus tard que les autres données climatiques au cours de leur étude. Elles n'ont toutefois pas été oubliées dans l'étude. Nous avons donc exécuté le code R en ajoutant les données VENT au Random Forest. Nous obtenons alors ce graphe ci-dessous :

Nous retrouvons les mêmes variables importantes que l'équipe tarification avec les 4 variables VENT intercalées.

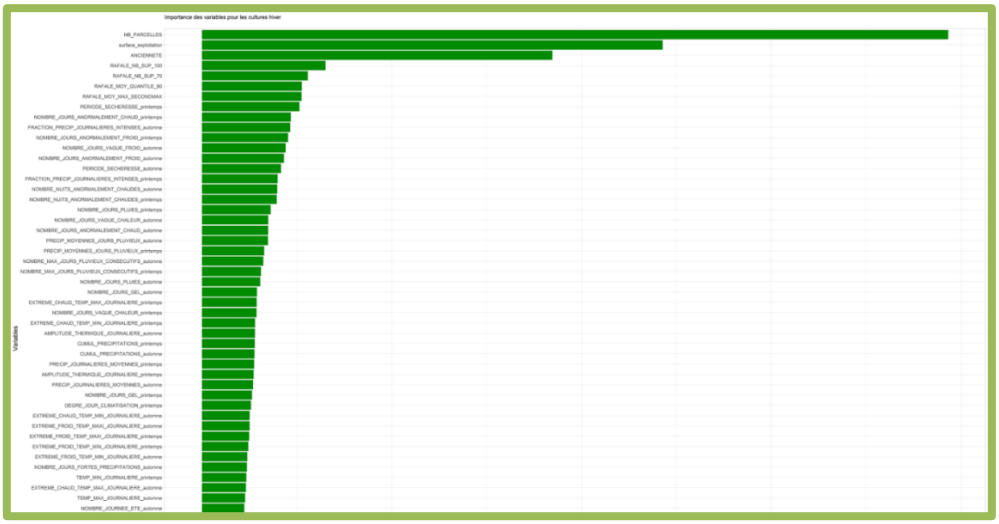


Figure 46 - Importance des variables pour les cultures d'hivers

Si nous zoomons sur les plus importantes on a :

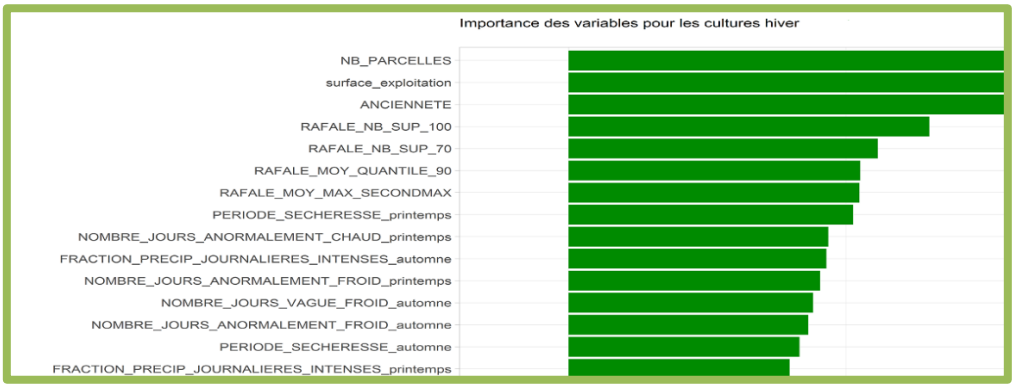


Figure 47 - Zoom sur les variables les plus importantes pour les cultures d'hivers

- Importances par cultures

Les importances sont ensuite affichées non pas par groupe de culture mais par cultures. Le Random Forest est effectué sur le même tableau de données que précédemment. La culture Blé tendre hiver étant celle avec les

capitaux assurés autres aléas les plus importants (1 300 M€), nous allons nous en servir comme exemple dans toute la suite du rapport, en commençant par le rejeu du random forest par culture :

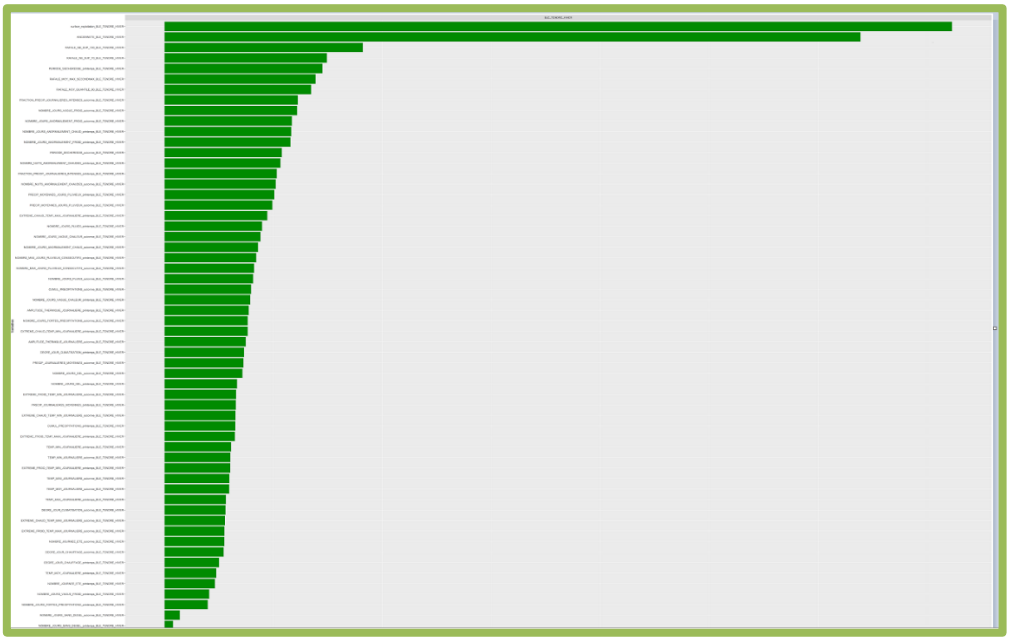


Figure 48 - Importance des variables par culture

Nous constatons que l'ordre d'importance est modifié mais les variables les plus importantes restent les mêmes.

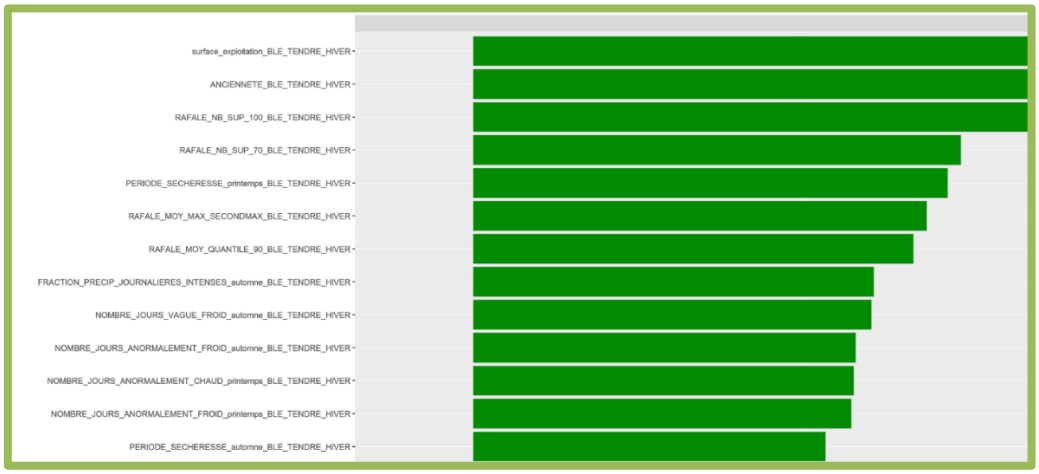


Figure 49- Extrait de l'importance des variables par culture

La précision d'un algorithme Random Forest dépend de certains paramètres comme le nombre d'arbres créés. Cela peut, peut-être modifier l'importance des variables. Il a été effectué avec 100 arbres par l'équipe tarification. Nous essayons donc avec une valeur 10 fois plus grande, 1000. En sachant que la valeur par défaut de la fonction est 500 arbres, 1000 devrait suffire :

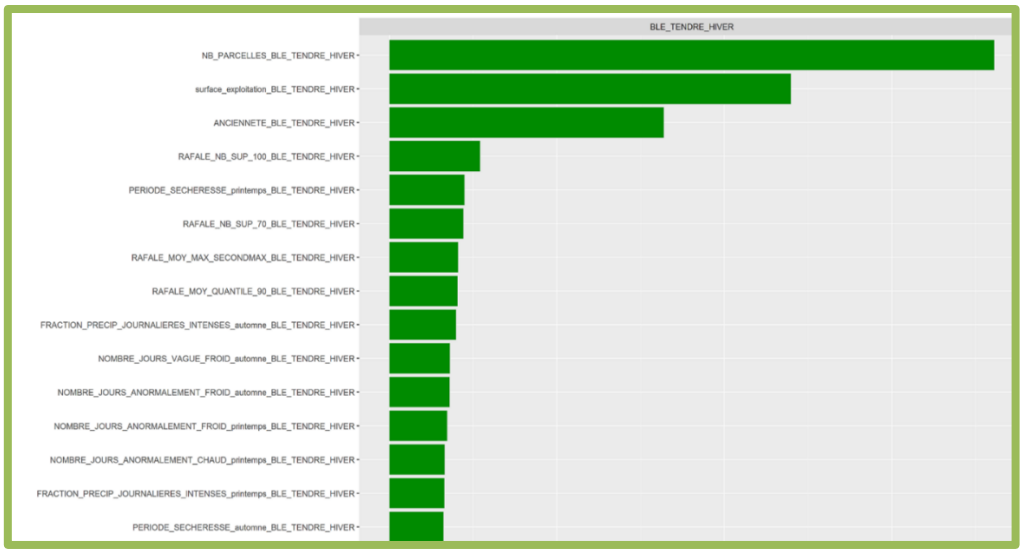


Figure 50 - Importance des variables Blé tendre hiver

L'ordre de certaines variables change mais les variables les plus importantes restent les mêmes : 100 arbres étaient donc bien suffisants pour déterminer les variables les plus importantes.

Dans la suite nous allons utiliser les 15 variables les plus importantes pour les cultures d'hiver soit celles présentes sur le dernier graphique.

Remarques de la validation :

- Nous recommandons d'actualiser le rapport fait par l'équipe tarification après ajout des données VENTS dans le modèle.
- A l'exception des données VENTS les variables les plus importantes obtenues concordent bien avec celles obtenues par l'équipe tarification. Afin de permettre de retrouver exactement les résultats lorsque le code est relancé, il est conseillé d'ajouter une 'graine' (pour rendre les tirages aléatoires du test toujours identiques).
- Nous recommandons de choisir un exemple explicite dans la documentation qui permettra de retracer les règles de décision pour la sélection des variables.

Nous recommandons de conclure plus explicitement sur les variables conservées pour la suite dans le rapport de l'équipe tarification : quelles ont été les variables retenues à la suite du Random Forrest, puis après l'étude des corrélations, pour chaque groupe de culture.

4.2.4.2. Corrélation

Si des variables sont trop corrélées entre elles, l'analyse de leur impact sur l'occurrence des sinistres peut être négligée. L'équipe tarification a donc décidé d'effectuer une étude de ces corrélations à l'aide du V de Cramer. Si des variables corrélées sont repérées, uniquement une d'entre elles sera conservée, par choix d'expert, par rapport aux qualités des données concernant ces variables ou par rapport à la

compréhension du tarif par les clients. Par exemple la variable nombre de jours anormalement chauds parait plus cohérente que nombre de nuits anormalement chaudes.

Pour effectuer cette analyse l'équipe tarification a itéré l'algorithme qui affiche les corrélations entre les variables en enlevant à chaque étape les variables jugées trop corrélées avec une autre. Et présente le résultat final dans son rapport.

Nous allons essayer de reproduire ce raisonnement en partant des 15 variables les plus importantes pour la saison hiver, et essayer de retrouver les résultats de l'équipe tarification.

Dans la suite nous allons nous intéresser à la culture BLE_TENDRE_HIVER qui va servir d'exemple. Nous commençons donc par regarder les corrélations entre les variables les plus importantes d'hiver.

Voici le résultat final des corrélations :

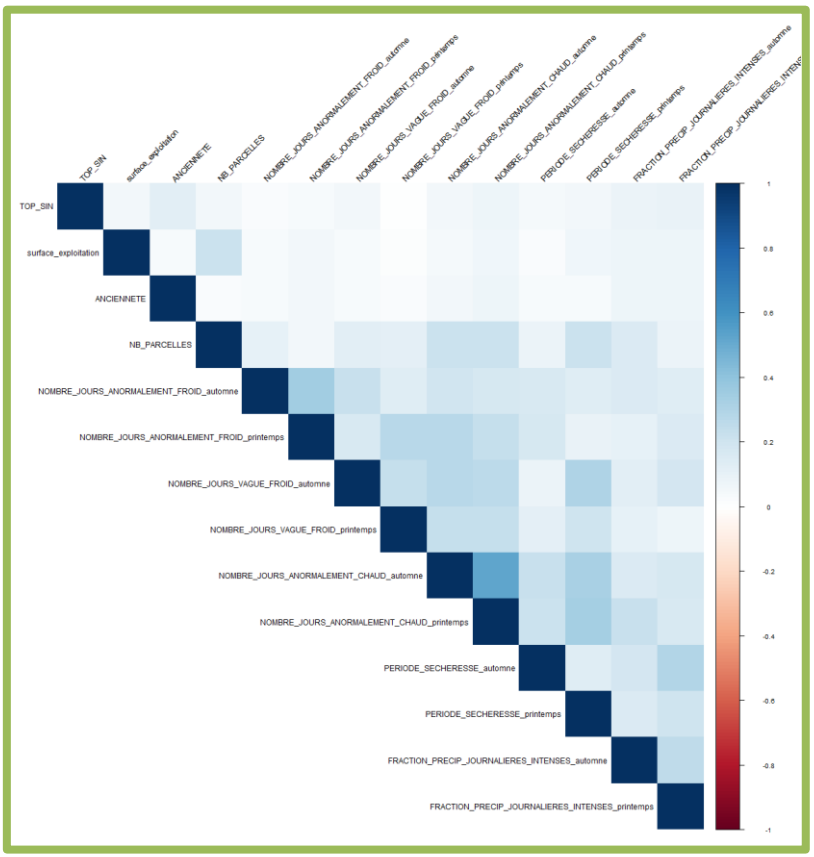


Figure 51 - Corrélation entre les variables

Plus la case est bleue plus la corrélation entre les deux variables est forte positivement. Si elle est blanche alors les variables sont décorrélées.

Nous commençons notre étude avec les 15 variables les plus importantes et nous obtenons :

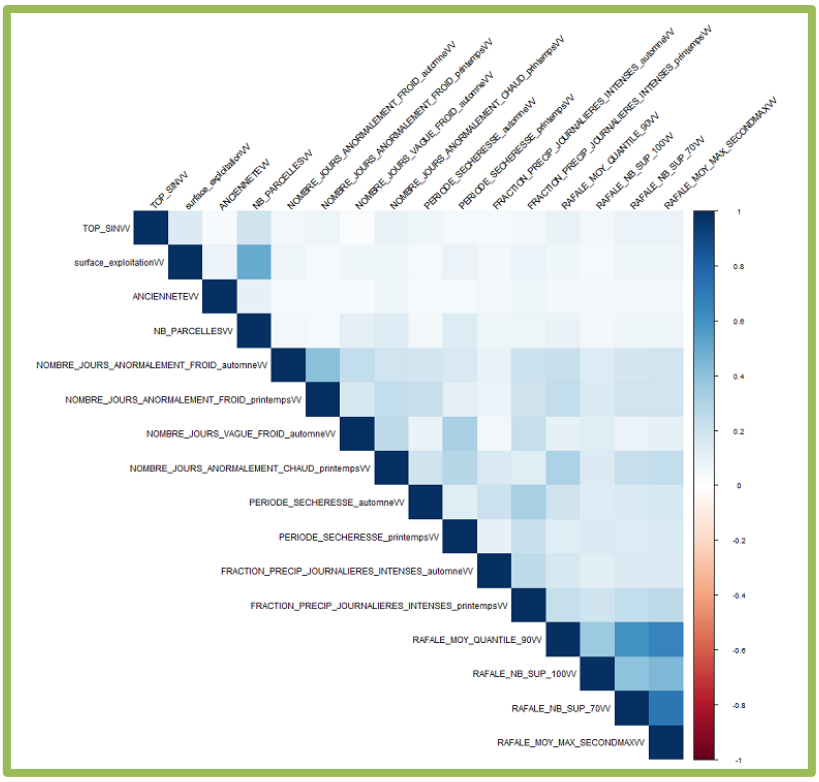


Figure 52 -Analyse de la corrélation réalisée par la validation

Nous remarquons que « **RAFALE_NB_SUP_70** » et « **RAFALE_MOY_MAX_SECONDMAX** » sont très corrélées avec « RAFALE_MOY_QUANTILE_90 ». Nous retirerons donc ces deux variables.

Dans la suite, ces 3 variables pour l'automne ne seront donc pas utilisées.

« NB_PARCELLES » est corrélée avec « surface_exploitation ». Nous décidons de garder surface exploitation qui parait plus cohérente dans la prévision d'un sinistre. Plus la surface de l'exploitation est grande, plus elle a de chance d'être touchée par un aléa climatique. Mais le nombre de parcelles ne donne pas d'indice sur la surface.

En enlevant ces trois variables nous obtenons un nouveau graphique :

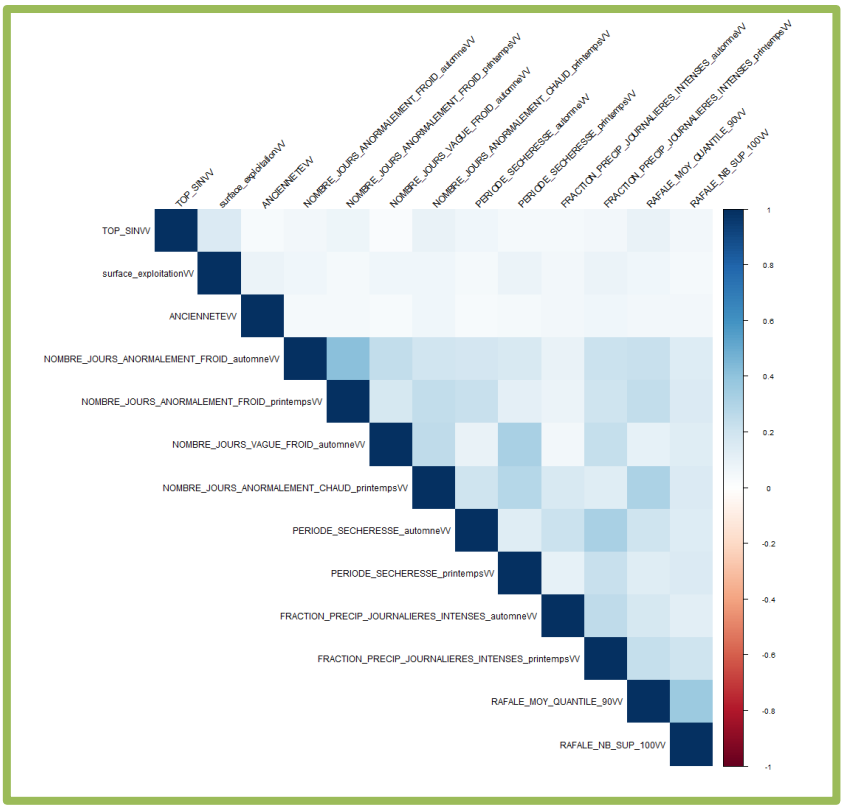


Figure 53 - Analyse de la corrélation réalisée par la validation

Nous obtenons un résultat légèrement différent de celui de l'équipe tarification dû à la présence des variables vents. De plus l'équipe tarification a choisi, après discussion avec la Direction Métier, d'ajouter les correspondances automne/printemps, ce qui explique les variables ajoutées dans le tableau de corrélations. Nous notons tout de même que les variables automne/printemps sont généralement corrélées.

La sélection des variables ne se fait pas forcément avec les corrélations tout juste affichées. Nous gardons ces résultats pour la suite et des variables seront supprimées par la suite à l'aide de l'analyse des coefficients GLM.

Remarques de la validation :

- Nous recommandons d'actualiser la documentation avec le travail effectué sur les variables VENT.
- Nous recommandons de reprendre l'exemple choisi dans la documentation pour le calcul de l'importance afin de pouvoir poursuivre la sélection des variables commencée avec le random forest.

4.3. Modélisation de la Prime Pure

4.3.1. Modèle GLM

4.3.1.1. Description du modèle

Le GLM (Generalized Linear Model) est une généralisation de la Régression Linéaire. Il permet d'analyser une variable cible à travers des variables explicatives et une loi adaptée. À la sortie du GLM nous retrouvons des coefficients correspondants à chaque modalité de chaque variable. Ces coefficients-là représentent une estimation de l'impact de chaque modalité sur la variable cible.

Dans le cas du modèle tarifaire, la variable cible est ''TOP_SIN''. Elle prend 1 comme valeur en cas de contrat sinistré et 0 sinon. De plus, les variables explicatives sont à choisir parmi celles à la sortie du Random Forest, à savoir :

- "surface_exploitation",
- "ANCIENNETE" ,
- "NB_PARCELLES",
- "NOMBRE_JOURS_ANORMALEMENT_FROID_automne",
- "NOMBRE_JOURS_ANORMALEMENT_FROID_printemps",
- "NOMBRE_JOURS_VAGUE_FROID_automne",
- "NOMBRE_JOURS_VAGUE_FROID_printemps",
- "NOMBRE_JOURS_ANORMALEMENT_CHAUD_automne",
- "NOMBRE_JOURS_ANORMALEMENT_CHAUD_printemps",
- "PERIODE_SECHERESSE_automne",
- "PERIODE_SECHERESSE_printemps",
- "RAFALE_MOY_QUANTILE_90",
- "RAFALE_NB_SUP_70",
- "RAFALE_NB_SUP_100",
- "RAFALE_MOY_MAX_SECONDMAX",
- "INDIC_IRRIGUE",
- "NOMBRE_JOURS_SANS_DEGEL_printemps",
- "NOMBRE_JOURS_SANS_DEGEL_automne",
- "FRACTION_PRECIP_JOURNALIERES_INTENSES_automne",
- "FRACTION_PRECIP_JOURNALIERES_INTENSES_printemps",

Variables internes

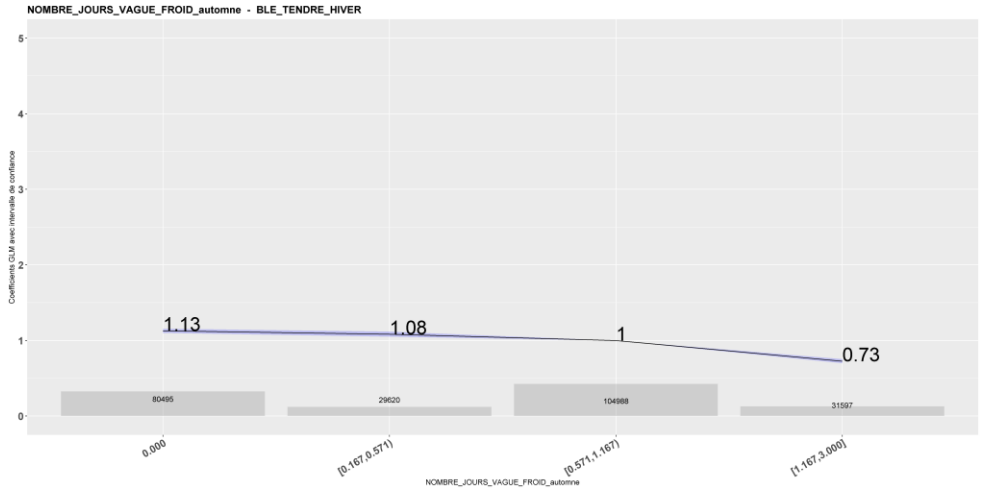
Variables climatiques

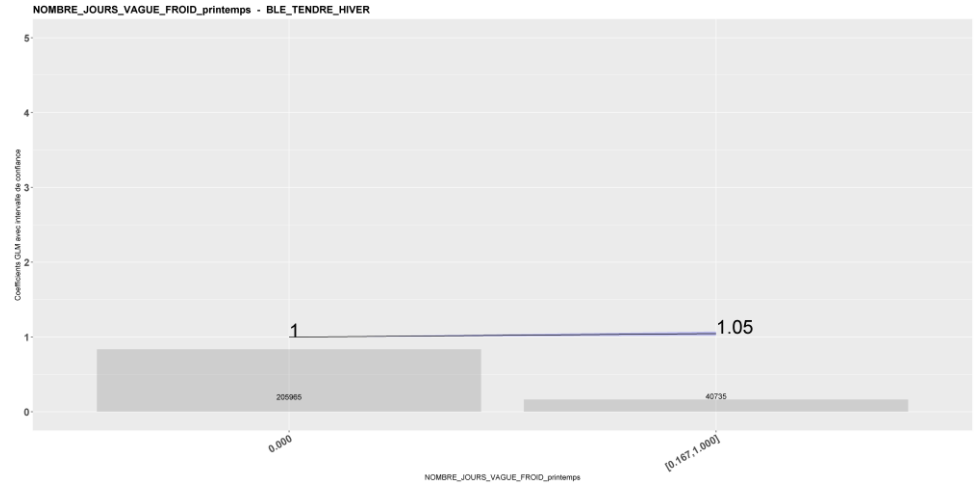
D'autre part, la loi la plus intuitive à utiliser pour modéliser la variable ''TOP_SIN'' est bien une loi de Bernoulli, d'où l'utilisation de cette loi dans le GLM. L'output de l'algorithme serait dès lors un coefficient par chaque modalité de variable. Ce coefficient évalue la probabilité d'un sinistre dans un contrat présentant les modalités en question.

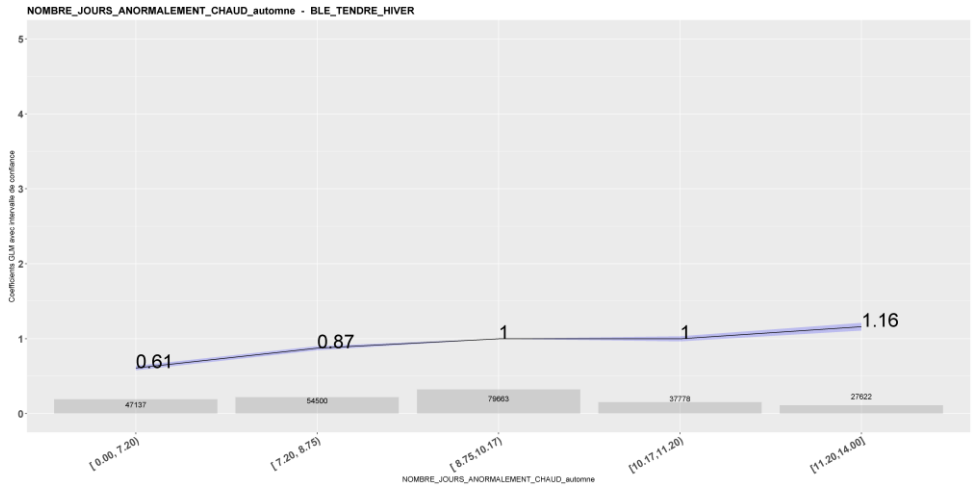
Nous avons réexécuté le code dans le cas de la culture Blé Tendre d'Hiver avec les mêmes variables que celles choisies par l'équipe tarification (le choix de ces variables étant fait par un expert de la direction métiers). Ces variables sont les suivantes :

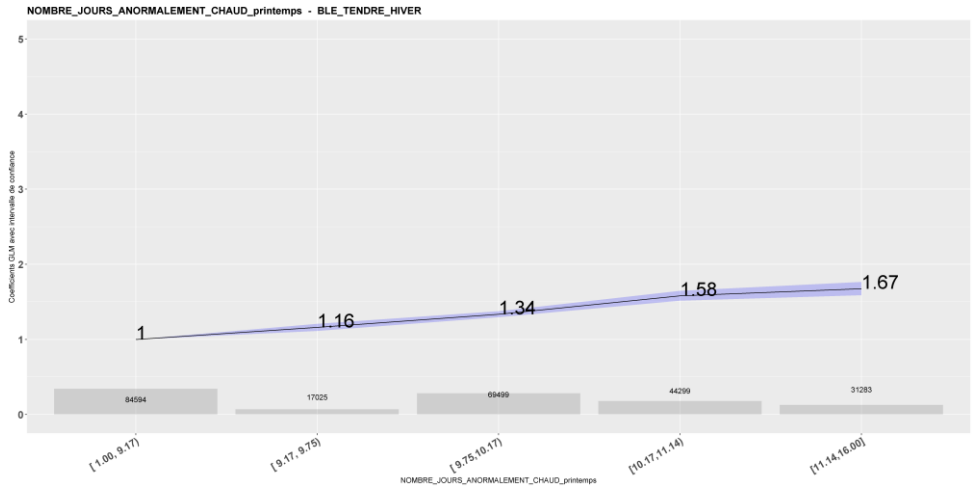
- "surface_exploitation",
- "ANCIENNETE",
- "NOMBRE_JOURS_VAGUE_FROID_automne",
- "NOMBRE_JOURS_VAGUE_FROID_printemps",
- "NOMBRE_JOURS_ANORMALEMENT_CHAUD_automne",
- "NOMBRE_JOURS_ANORMALEMENT_CHAUD_printemps",
- "NOMBRE_JOURS_ANORMALEMENT_FROID_automne",
- "NOMBRE_JOURS_ANORMALEMENT_FROID_printemps",
- "RAFALE_MOY_QUANTILE_90",
- "RAFALE_NB_SUP_100",
- "FRACTION_PRECIP_JOURNALIERES_INTENSES_automne",
- "FRACTION_PRECIP_JOURNALIERES_INTENSES_printemps",
- "PERIODE_SECHERESSE_automne",
- "PERIODE_SECHERESSE_printemps",

Nous obtenons les coefficients suivants (pour quelques variables, pour la culture Blé Tendre Hiver) :









4.3.1.2. Test de niveau III

Pour s'assurer encore une fois des résultats du Random Forest précédent, nous avons effectué un test de niveau 3 après l'exécution du GLM appliqué à la culture Blé Tendre d'Hiver. Les résultants ont été les suivants :

```
Model:
TOP_SINVV ~ surface_exploitationVV + ANCIENNETEVV + NOMBRE_JOURS_ANORMALEMENT_FROID_automneVV +
    NOMBRE_JOURS_ANORMALEMENT_FROID_printempsVV + NOMBRE_JOURS_VAGUE_FROID_automneVV +
    NOMBRE_JOURS_VAGUE_FROID_printempsVV + NOMBRE_JOURS_ANORMALEMENT_CHAUD_automneVV +
    NOMBRE_JOURS_ANORMALEMENT_CHAUD_printempsVV + PERIODE_SECHERESSE_automneVV +
    PERIODE_SECHERESSE_printempsVV + FRACTION_PRECIP_JOURNALIERES_INTENSES_automneVV +
    FRACTION_PRECIP_JOURNALIERES_INTENSES_printempsVV + RAFALE_MOY_QUANTILE_90VV +
    RAFALE_NB_SUP_100VV
                                          Df Deviance    AIC     LRT  Pr(>Chi)
<none>                                       279476 279590
surface_exploitationVV                     4   282089 282195 2612.45 < 2.2e-16 ***
ANCIENNETEVV                               8   279830 279928  353.81 < 2.2e-16 ***
NOMBRE_JOURS_ANORMALEMENT_FROID_automneVV   4   279574 279680   97.10 < 2.2e-16 ***
NOMBRE_JOURS_ANORMALEMENT_FROID_printempsVV 4   279494 279600   17.54   0.00152 **
NOMBRE_JOURS_VAGUE_FROID_automneVV         3   279804 279912  327.27 < 2.2e-16 ***
NOMBRE_JOURS_VAGUE_FROID_printempsVV       1   279494 279606   17.46 2.931e-05 ***
NOMBRE_JOURS_ANORMALEMENT_CHAUD_automneVV  4   279888 279994  411.68 < 2.2e-16 ***
NOMBRE_JOURS_ANORMALEMENT_CHAUD_printempsVV 4   279997 280103  520.23 < 2.2e-16 ***
PERIODE_SECHERESSE_automneVV               4   279576 279682   99.53 < 2.2e-16 ***
PERIODE_SECHERESSE_printempsVV             4   279821 279927  344.50 < 2.2e-16 ***
FRACTION_PRECIP_JOURNALIERES_INTENSES_automneVV  4   279506 279612   29.16 7.248e-06 ***
FRACTION_PRECIP_JOURNALIERES_INTENSES_printempsVV 4   279511 279617   34.49 5.910e-07 ***
RAFALE_MOY_QUANTILE_90VV                   4   279518 279624   41.13 2.529e-08 ***
RAFALE_NB_SUP_100VV                        4   279508 279614   31.74 2.158e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La fonction *drop1* calcule le GLM initial (qui correspond à la ligne <none>), puis elle réitère le GLM en supprimant à chaque fois l'une des variables explicatives et calcule la déviance du nouveau modèle par

rapport au modèle initial. Ce qui donne une idée sur l'importance de chaque variable. Cette dernière est exprimée à la fin de chaque ligne par des étoiles : plus il y a d'étoiles plus la variable est significative.

Nous remarquons que toutes les variables ont des notations fortes quant à leur significativité dans le modèle.

4.3.1.3. Tests du pouvoir prédictif du modèle GLM

L'objectif principal du modèle GLM est de prédire la variable à expliquer, dans notre cas TOP_SIN, à l'aide des variables explicatives. Le type de GLM utilisé par l'équipe tarification est un GLM binomial, qui va non pas donner des valeurs égales à 0 ou 1, mais la probabilité d'obtenir un sinistre pour chaque contrat de la base de données, en combinant les coefficients appropriés aux variables de chaque assuré.

L'équipe tarification a effectué ce modèle GLM à partir de la base de données entière, il va donc bien fonctionner sur ces données-là. Il serait intéressant de voir s'il est capable de garder son pouvoir prédictif sur des données qui n'ont pas été utilisées pour le modèle, car l'objectif est aussi de pouvoir définir un risque de sinistre dans des villes qui ne se trouvent pas dans le portefeuille d'assurés de GROUPAMA.

Pour effectuer ce travail, comme nous disposons uniquement des données de sinistralité de GROUPAMA, nous avons décidé de partager le portefeuille en deux : 80% des contrats seront utilisés pour établir les coefficients GLM, et constitueront la base d'apprentissage ; les 20% constitueront la base test. Nous allons donc récupérer les coefficients GLM produits par la base d'apprentissage, les appliquer à la base test pour obtenir des probabilités de sinistre pour chacun des contrats de la base test pour obtenir leur probabilité de sinistre. Pour savoir si ces probabilités sont proches de la réalité, nous allons les comparer à la fréquence réelle de sinistre. Pour déterminer cette fréquence réelle de sinistre nous ne pouvons pas travailler à la maille contrat qui nous donnerait au maximum 9 données par assurés à cause la faible profondeur d'historique. Mais comme les coefficients GLM de chaque assuré dépendent uniquement des données climatiques de sa ville, nous pouvons déterminer une fréquence de sinistre par ville, qui sera celle de tous les assurés s'y situant.

Remarque : Le GLM est effectué sur les cultures de blé tendre hiver, avec les variables retenues dans le fichier Excel « GLM CULTURES HIVER » de l'équipe tarification.

Nous obtenons un tableau de résultats avec 6 colonnes :

- La commune de l'assuré
- La prédiction de la probabilité de sinistre (qui dépend uniquement de la commune)
- Frequence20 : la fréquence de sinistre pour une commune donnée dans la base de test
- Frequence100 : la fréquence de sinistre pour une commune dans toute la base (apprentissage + test)
- Ecart20 : écart entre la prédiction et frequence20
- Ecart100 : écart entre la prédiction et frequence100

Voici le résultat par commune :

Commune	Prédiction	Fréquence20	Fréquence100	Ecart20	Ecart100
1001	0,165098226	0,4	0,166666667	0,23490177	0,00156844
1004	0,215761959	0	0	0,21576196	0,21576196
1005	0,152905436	0,2	0,172413793	0,04709456	0,01950836
1007	0,193453591	0,333333333	0,35	0,13987974	0,15654641
1016	0,293568675	0	0	0,29356868	0,29356868
1021	0,196975878	0,333333333	0,060606061	0,13635746	0,13636982
1023	0,165098226	0	0,4	0,16509823	0,23490177
1024	0,236335704	0	0	0,2363357	0,2363357
1025	0,21571287	0,666666667	0,131578947	0,4509538	0,08413392
1027	0,167060599	0	0	0,1670606	0,1670606
1028	0,215301008	0	0,125	0,21530101	0,09030101
1029	0,214817747	0,2	0,1	0,01481775	0,11481775
1034	0,181506653	0	0	0,18150665	0,18150665
1038	0,233127063	1	0,214285714	0,76687294	0,01884135
1040	0,233639257	0	0,125	0,23363926	0,10863926

Ainsi que des statistiques sur les prédictions et les écarts à la fréquence

	Moyenne	Min	Max
Prédiction	0,22	0,08	0,39
Ecart20	0,18	0,00	0,90
Ecart100	0,11	0,00	0,88

Les prédictions de probabilité de sinistre se situent entre 0.08 et 0.39 avec une moyenne de 0.22. Mais les écarts à la fréquence, que ce soit sur la base test ou la base totale sont un peu élevés : nous avons des moyennes de 0.18 et 0.11 pour les écarts alors que la moyenne des prédictions et 0.22. Cela est dû au manque de données lors du calcul des fréquences. Nous le remarquons bien car les résultats s'améliorent quand nous passons de la fréquence sur la base test à la fréquence sur la base totale. Cependant la base totale n'est pas suffisamment grande (252 000 contrats) pour pouvoir obtenir des fréquences précises.

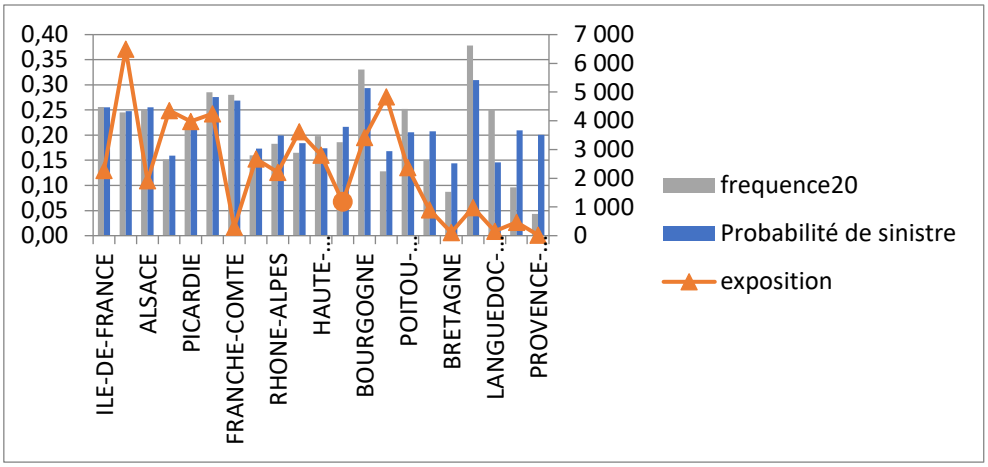
La maille commune ne permettant pas d'obtenir des résultats concluants, nous nous sommes intéressées à la maille région qui pourra nous permettre de limiter le problème du manque de données lors du calcul des fréquences historiques.

De la même manière que précédemment, des probabilités de sinistres par commune sont calculées sur la base test grâce au GLM effectué sur la base d'apprentissage. Pour agréger à la maille région nous effectuons une moyenne des probabilités sur tous les contrats d'une même région pour obtenir une probabilité moyenne de sinistre par région. Les fréquences de sinistres observés par région sont ensuite calculées et présentées dans le tableau de résultat ci-dessous :

Région	Probabilité moyenne de sinistre	Frequence100	Frequence20	Ecart100	Ecart20	Exposition
ILE-DE-FRANCE	0,26	0,26	0,26	0,01	0,00	2 277
CENTRE	0,25	0,24	0,25	0,01	0,00	6 494
ALSACE	0,26	0,25	0,25	0,01	0,00	1 920
MIDI-PYRENEES	0,16	0,15	0,15	0,01	0,01	4 358
PICARDIE	0,21	0,22	0,22	0,01	0,01	3 984
CHAMPAGNE-ARDENNE	0,28	0,28	0,29	0,01	0,01	4 252
FRANCHE-COMTE	0,27	0,32	0,28	0,05	0,01	300
BASSE-NORMANDIE	0,17	0,17	0,16	0,00	0,01	2 670
RHONE-ALPES	0,20	0,19	0,18	0,01	0,02	2 207
NORD-PAS-DE-CALAIS	0,18	0,16	0,16	0,02	0,02	3 619
HAUTE-NORMANDIE	0,17	0,20	0,20	0,02	0,02	2 808
AQUITAINE	0,22	0,16	0,19	0,05	0,03	1 173
BOURGOGNE	0,29	0,35	0,33	0,05	0,04	3 418
PAYS DE LA LOIRE	0,17	0,13	0,13	0,04	0,04	4 840
POITOU-CHARENTES	0,21	0,26	0,25	0,05	0,05	2 369
AUVERGNE	0,21	0,14	0,15	0,07	0,06	908
BRETAGNE	0,14	0,10	0,09	0,04	0,06	115
LORRAINE	0,31	0,37	0,38	0,06	0,07	986
LANGUEDOC-ROUSSILLON	0,15	0,23	0,25	0,08	0,11	151
LIMOUSIN	0,21	0,12	0,10	0,09	0,11	468
PROVENCE-ALPES-COTE D'AZUR	0,20	0,06	0,04	0,14	0,16	23

Nous retrouvons dans le tableau, pour chaque région, sa probabilité moyenne de sinistre, sa fréquence de sinistre dans la base totale (frequence100), sa fréquence de sinistre dans la base test (frequence20), les écarts entre la probabilité moyenne et les différentes fréquences. Nous avons ajouté une colonne exposition qui correspond au nombre de contrats présent dans la base test dans la région correspondante.

Nous visualisons les résultats sur le graphique ci-dessous :



Si les hauteurs de la barre grise et de la barre bleue sont proches alors la probabilité de sinistre pour la région donnée est bien prédite par GLM. Si ce n'est pas le cas, la courbe de l'exposition nous permet souvent de l'expliquer. En effet si cette dernière est faible, alors la fréquence est calculée avec très peu de données, ce qui explique le fait qu'elle soit différente de la probabilité théorique prédite par le GLM. Nous pouvons nous en rendre compte pour les régions sur la droite du graphique, les expositions sont très faibles et les hauteurs des barres vertes et bleues sont bien différentes.

Remarques de la validation :

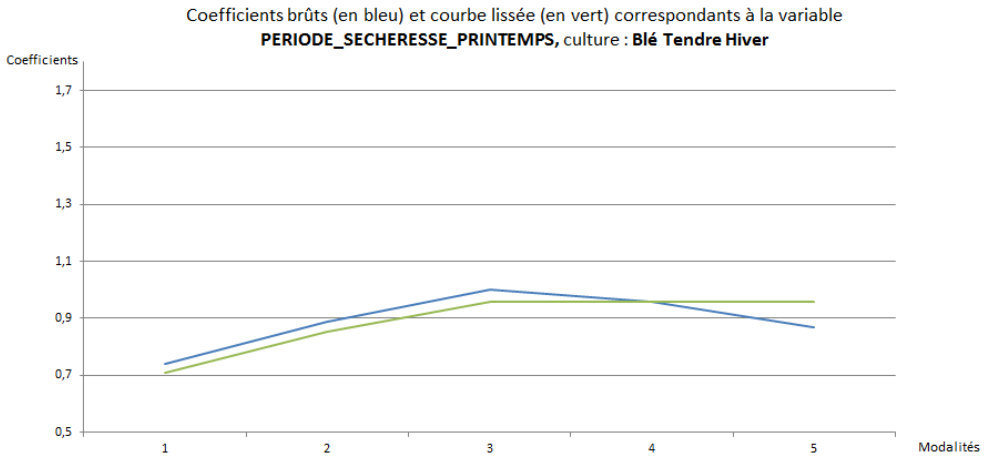
- La maille commune ne nous donnait pas assez d'historique pour obtenir des fréquences représentatives de la réalité. Mais la maille région nous permet d'obtenir des résultats globalement positifs qui témoignent de la bonne prédiction du risque de sinistre par le modèle GLM.
- Ce test pourra être effectué plus tard avec de nouvelles années d'historiques pour obtenir des fréquences plus précises.

4.3.2. Lissage des coefficients

Une fois les coefficients bruts du GLM obtenus pour chaque modalité de chaque variable. L'équipe tarification a procédé à la suppression de quelques variables dont les coefficients ne sont ni strictement croissants ni strictement décroissants. Par exemple, dans le cas du Blé Tendre d'Hiver, les variables "RAFALE_NB_SUP_100","NOMBRE_JOURS_ANORMALEMENT_FROID_printemps","NOMBRE_JOURS_ANORMALEMENT_CHAUD_automne", et "FRACTION_PRECIP_JOURNALIERES_INTENSES_automne" ont été supprimé. Les variables restantes sont donc :

term_	borne_inf	borne_sup	estimate	min	max	DETP
FRACTION_PRECIP_JOURNALIERES_INTENSES_printemps	40,4	52,2	1	1	1	1
FRACTION_PRECIP_JOURNALIERES_INTENSES_printemps	52,2	54,8	1,00693545	0,973723	1,04128074	1,02
FRACTION_PRECIP_JOURNALIERES_INTENSES_printemps	54,8	56,6	0,9400845	0,90695686	0,97442216	0,95
FRACTION_PRECIP_JOURNALIERES_INTENSES_printemps	56,6	59	0,94871243	0,91393706	0,98481099	0,96
FRACTION_PRECIP_JOURNALIERES_INTENSES_printemps	59	88,2	1,00063276	0,95600079	1,04734843	1,02
NOMBRE_JOURS_ANORMALEMENT_CHAUD_printemps	1	9,17	1	1	1	1,00
NOMBRE_JOURS_ANORMALEMENT_CHAUD_printemps	9,17	9,75	1,11398959	1,0689093	1,1609711	1,11
NOMBRE_JOURS_ANORMALEMENT_CHAUD_printemps	9,75	10,17	1,29487546	1,25636226	1,33456926	1,30
NOMBRE_JOURS_ANORMALEMENT_CHAUD_printemps	10,17	11,14	1,56726922	1,50322641	1,63404047	1,58
NOMBRE_JOURS_ANORMALEMENT_CHAUD_printemps	11,14	16	1,71639091	1,62725611	1,81040817	1,73
NOMBRE_JOURS_ANORMALEMENT_FROID_automne	7,4	8,17	1	1	1	1
NOMBRE_JOURS_ANORMALEMENT_FROID_automne	2	6,8	0,8989975	0,86673208	0,93246404	0,91
NOMBRE_JOURS_ANORMALEMENT_FROID_automne	6,8	7,4	0,91220256	0,88709364	0,93802219	0,92
NOMBRE_JOURS_ANORMALEMENT_FROID_automne	8,17	8,71	0,8991075	0,864166	0,93546182	0,9
NOMBRE_JOURS_ANORMALEMENT_FROID_automne	8,71	12	0,89243123	0,86546797	0,92023452	0,9
NOMBRE_JOURS_VAGUE_FROID_automne	0,571	1,167	1	1	1	1
NOMBRE_JOURS_VAGUE_FROID_automne	0	NA	1,10190084	1,0755583	1,12888856	1,12
NOMBRE_JOURS_VAGUE_FROID_automne	0,167	0,571	1,07426018	1,04172065	1,10781612	1,07
NOMBRE_JOURS_VAGUE_FROID_automne	1,167	3	0,75665094	0,72928555	0,78504317	0,77
NOMBRE_JOURS_VAGUE_FROID_printemps	0	NA	1	1	1	1
NOMBRE_JOURS_VAGUE_FROID_printemps	0,167	1	1,05955189	1,03122923	1,08865242	1,06
PERIODE_SECHERESSE_automne	14,2	15,2	1	1	1	1
PERIODE_SECHERESSE_automne	10	14,2	0,95033576	0,92215449	0,97937825	0,95
PERIODE_SECHERESSE_automne	15,2	15,4	1,03796761	0,99133113	1,08679807	1,01
PERIODE_SECHERESSE_automne	15,4	16,2	1,06808005	1,04167129	1,09515833	1,06
PERIODE_SECHERESSE_automne	16,2	25	1,18327369	1,13824902	1,23007937	1,18
PERIODE_SECHERESSE_printemps	15,2	16,2	1	1	1	1
PERIODE_SECHERESSE_printemps	8	14,2	0,72360471	0,69772539	0,75044391	0,71
PERIODE_SECHERESSE_printemps	14,2	15,2	0,9316099	0,90601121	0,95793186	0,93
PERIODE_SECHERESSE_printemps	16,2	16,8	0,97371037	0,94311655	1,00529663	0,96
PERIODE_SECHERESSE_printemps	16,8	26	0,89948576	0,87391608	0,92580359	0,91
RAFALE_MOY_QUANTILE_90	55,2	57,2	1	1	1	1
RAFALE_MOY_QUANTILE_90	33,7	52,3	1,01223951	0,97474451	1,0511768	1
RAFALE_MOY_QUANTILE_90	52,3	54	1,06196954	1,02810388	1,09695073	1
RAFALE_MOY_QUANTILE_90	54	55,2	1,04300289	1,0126426	1,07427343	1
RAFALE_MOY_QUANTILE_90	57,2	87,5	1,10881404	1,06699907	1,15226772	1

Ensuite, l'équipe tarification a procédé à un lissage fait à la main des coefficients des variables restantes, afin de donner une allure linéaire ascendante, descendante ou stable des coefficients.



Finalement, à partir des valeurs initiales de chaque variable par commune, une règle de trois est appliquée afin d'obtenir des coefficients par chaque commune. Nous avons refait ce travail et les coefficients retrouvés ont été proches.

Commune	Coefficient calculé par l'équipe de Validation	Coefficient calculé par la DETP	Différence entre les deux
01001	1,230691363	1,241969908	0,011278545
01002	1,198949499	1,209937149	0,01098765
01004	1,267288416	1,278902351	0,011613935
01005	1,20907602	1,220156474	0,011080454
01006	1,227117799	1,238363594	0,011245796
01007	1,184669786	1,195526571	0,010856785
01008	1,267288416	1,278902351	0,011613935
01009	0,957584068	0,966359749	0,008775681
01010	1,132336512	1,142713694	0,010377182
01011	1,261500132	1,273061021	0,011560889
01012	1,159056851	1,169678909	0,010622058
01013	1,264166806	1,275752133	0,011585327
01014	1,238362871	1,249711721	0,01134885
01015	1,098695639	1,108764523	0,010068884
01016	1,269046241	1,280676286	0,011630045
01017	1,214984235	1,226118834	0,011134599
01019	1,181947094	1,192778927	0,010831833
01021	1,230134283	1,241407723	0,01127344
01022	0,926470879	0,934961426	0,008490548
01023	1,250529162	1,261989509	0,011460347
01024	1,14205957	1,152525858	0,010466288
01025	1,244841616	1,256249839	0,011408224
01026	1,249258804	1,260707508	0,011448705
01027	1,129134028	1,139481861	0,010347833
01028	1,253348438	1,264834622	0,011486184
01029	1,196645633	1,20761217	0,010966537
01030	1,316939056	1,32900801	0,012068953
01031	1,261500132	1,273061021	0,011560889
01032	1,129134028	1,139481861	0,010347833
01033	1,260674701	1,272228025	0,011553324
01034	1,116389009	1,126620042	0,010231033
01035	1,112458934	1,12265395	0,010195016
01036	1,219977464	1,231157822	0,011180359
01037	1,045409554	1,054990103	0,009580549
01038	1,134223827	1,144618305	0,010394478
01039	0,926470879	0,934961426	0,008490548
01040	1,193696662	1,204636173	0,010939511

4.3.3. Constante de tarification

4.3.3.1. Calcul de la constante

Maintenant que les coefficients GLM sont lissés, on peut obtenir pour chaque commune pour chaque type de culture un coefficient indiquant le risque de sinistre pour une parcelle de ce type de culture dans cette commune, en multipliant les coefficients lissés des modalités correspondantes.

L'équipe tarification a modélisé la prime comme le produit des coefficients GLM multiplié par le capital et une constante de tarification.

Pour obtenir la prime, par exemple pour une culture de blé d'hiver dans la commune 01001 :

$$\text{Prime}_{\text{blé}\backslash 01001} = \text{P0}_{\text{blé}} \times \text{Coef}_{\text{blé}\backslash 01001} \times \text{Capital}_{\text{assuré}}$$

Nous multiplions le coefficient GLM de la culture blé dans la ville 01001 par une constante P0 propre à chaque culture et par le capital assuré.

C'est cette constante P0 qu'il faut déterminer. Elle va permettre d'équilibrer les primes avec les coûts des sinistres.

$$\text{Prestations}_{\text{blé}} = \text{P0}_{\text{blé}} \times \sum_{\text{commune}} \text{Coef}_{\text{blé}\backslash\text{commune}} \times \text{Capital}_{\text{blé}\backslash\text{commune}}$$

Nous avons alors :

$$P0_{blé} = \frac{\text{Prestations}_{\text{blé}}}{\sum_{\text{commune}} \text{Coef}_{\text{blé\commune}} \times \text{Capital}_{\text{blé\commune}}}$$

Ce calcul est effectué à l'aide du fichier CTE_TARIF.R.

Les inputs sont :

- Un tableau Excel pour les **coefficients GLM** avec deux colonnes : commune et coefficient. Composé de trois feuilles pour les cultures Blé hiver, Colza, et Orge.

- La base **total sinistre** avec les couts totaux des sinistres par ville/année/culture. Par exemple pour la commune 1001 :

Année	Culture	Code Commune	Total sinistre
2009	MAIS	1001	10 043
2011	BLE TENDRE	1001	0
2011	MAIS	1001	0
2013	BLE TENDRE	1001	167
2013	MAIS	1001	1 722
2014	BLE TENDRE	1001	934
2014	TOURNESOL	1001	326
2015	BLE TENDRE	1001	107
2015	LEGUMINEUSES	1001	3 480
2015	MAIS	1001	91 461
2015	TOURNESOL	1001	3 150
2016	BLE TENDRE	1001	1 276

- La base des **capitaux** avec l'identifiant contrat, le type de culture, l'année et le capital assuré pour chaque ville/type de culture/année. Par exemple pour la commune 1001 en 2009 et 2010 :

Année	Type de culture	Code Commune	Capital total assuré
2009	AUTRES CEREALES	1001	1092
2009	BLE TENDRE	1001	132 059
2009	COLZA	1001	18 074
2009	MAIS	1001	196 702
2009	ORGE	1001	10 721
2010	BLE TENDRE	1001	112 430
2010	COLZA	1001	11 318
2010	MAIS	1001	171 466
2010	ORGE	1001	21 559
2010	TOURNESOL	1001	12 722

Les sinistres totaux sont regroupés par culture ainsi (on est toujours sur l'exemple des cultures d'hiver) :

Type de culture	Total sinistre
BLE TENDRE	193 851 004
COLZA	49 179 672
ORGE	53 164 386

Les trois feuilles Excel de coefficients sont regroupées dans un même tableau. Par exemple pour les communes 1001 et 10022 :

Code Commune	Coefficient GLM	Type de culture
1001	1,241969908	BLE TENDRE
1001	0,885012761	COLZA
1001	0,727670448	ORGE
10022	1,306991868	BLE TENDRE
10022	1,013649775	COLZA
10022	0,803426931	ORGE

Les colonnes coefficients GLM et total sinistres sont ajoutées à la table des capitaux en fusionnant les tableaux. Par exemple pour la ville 1001 pour le blé tendre :

Type de culture	Code Commune	Année	Total capital assuré	Coefficient GLM	Total sinistres
BLE TENDRE	1001	2009	132 059	1,241969908	193 851 004
BLE TENDRE	1001	2010	112 430	1,241969908	193 851 004
BLE TENDRE	1001	2011	131 194	1,241969908	193 851 004
BLE TENDRE	1001	2012	108 664	1,241969908	193 851 004
BLE TENDRE	1001	2013	118 719	1,241969908	193 851 004
BLE TENDRE	1001	2014	153 118	1,241969908	193 851 004
BLE TENDRE	1001	2015	108 493	1,241969908	193 851 004
BLE TENDRE	1001	2016	92 381	1,241969908	193 851 004

À partir de ce tableau, nous pouvons calculer les P0 de chaque culture :

META_Regroupement2	CTE_TARIF
BLE TENDRE	0,014316426
COLZA	0,015151556
ORGE	0,021234175

Remarques de la validation :

- Nous retrouvions les mêmes résultats à périmètre identique. Dans le cadre de l'étude un teste avec/sans 2016 a été réalisé pour évaluer l'impact de la prise en compte ou non de cette année
- Faire attention à la traçabilité de l'utilisation de cette fonctionnalité avec/sans 2016 dans code R

4.3.3.2. Retraitement de la constante

Dans le calcul de la constante P0, l'année 2016 est prise en compte. Or la sinistralité de 2016 a été exceptionnelle et la considérer totalement amènerait une surestimation de la valeur de P0 et donc aussi des montants des primes pures. A l'inverse si nous la supprimons totalement, nous allons sous-estimer la valeur des primes. Il faut donc trouver quelle part de la sinistralité 2016 tiendra compte de son caractère exceptionnel et qui permettra d'ajuster la constante P0 précédemment calculée.

Pour cela, une étude a été faite par l'équipe tarification sur la période de retour d'une année avec la sinistralité de 2016 et sur les coûts d'indemnisation. Le but est de déterminer la distribution des coûts afin de comparer le coût moyen observé au coût moyen probable selon cette distribution. Nous saurons alors de quel pourcentage diminuer la constante P0.

Pour cette étude, les inputs sont :

- Un tableau provenant de la base agreste (données en libre accès fournies par le ministère de l'agriculture) avec les **rendements moyens nationaux** par année (1950-2020) pour la culture en question.

Année	Surface	Production	Rendement
2007	4758488	30 618 391	6,43
2008	5038071	36 715 437	7,29
2009	4698605	35 986 116	7,66
2010	4881909	35 382 247	7,25
2011	4975755	33 887 479	6,81
2012	4817908	35 237 671	7,31
2013	4957370	36 702 460	7,40
2014	4983553	37 296 180	7,48
2015	5144272	40 835 316	7,94
2016	5125057	27 550 502	5,38
2017	4948207	36 468 623	7,37
2018	4866337	33 955 378	6,98
2019	4983221	39 409 446	7,91
2020	4221660	28 906 578	6,85

- Un tableau avec les **coûts d'indemnisation** totaux pour chaque année provenant des données Groupama.

Année	Coût Réel des indemnisations
2008	2 384 447
2009	1 301 672
2010	3 677 837
2011	34 415 219
2012	19 408 494
2013	10 942 559
2014	11 344 720
2015	3 160 840
2016	175 090 919
2017	7 455 409
2018	9 534 140
2019	4 301 742
2020	33 237 851

- Période de retour de la sinistralité de 2016

L'objectif est de déterminer la probabilité d'avoir la sinistralité obtenue pendant la l'année 2016 qui va servir à construire la distribution des coûts d'indemnisation totaux par année.

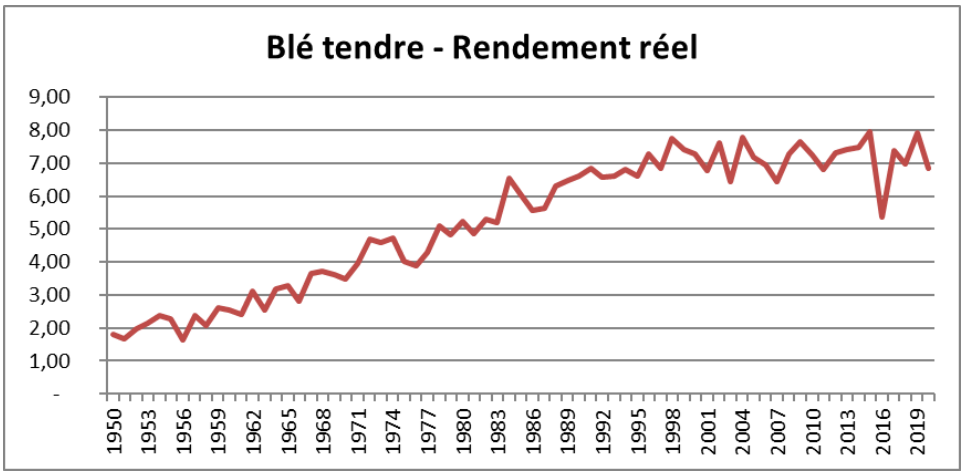
L'équipe tarification utilise les données nationales de la base agreste pour calculer la période de retour de 2016, certaines hypothèses ont été faites :

→ Les pertes de rendements nationaux ne sont dues qu'aux sinistres autres aléas. En effet les sinistres grêles sont très localisés et n'impactent pas les pertes de rendement nationaux. De même pour les sinistres liés aux insectes ou autre.

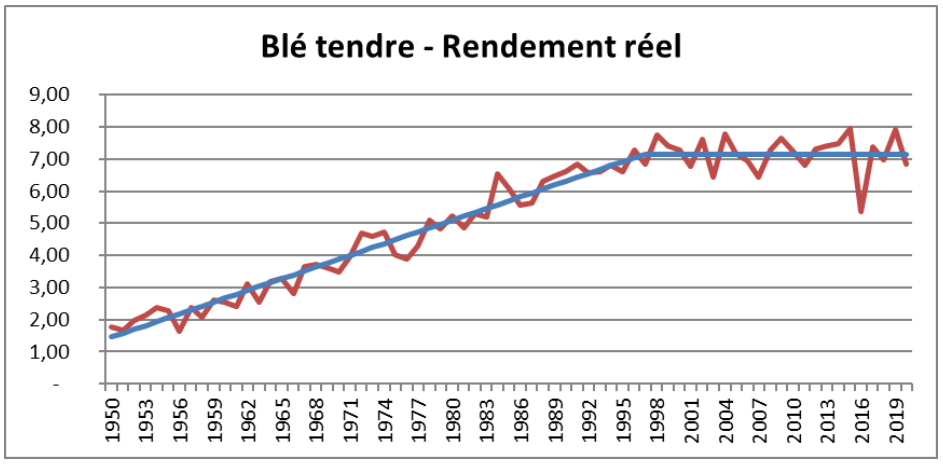
→ La sinistralité perçue par le ministère de l'agriculture est la même que celle du portefeuille Groupama. Cela est possible car les sinistres autres aléas frappent sur de larges zones géographiques. Si le rendement moyen national a baissé alors le rendement moyen du portefeuille Groupama a baissé aussi.

À partir de ces rendements moyens nationaux, nous allons essayer d'établir la distribution des rendements. Ainsi calculer la probabilité d'obtenir un rendement inférieur à celui de 2016 grâce à la fonction de répartition, et donc déterminer une période de retour de la sinistralité de 2016.

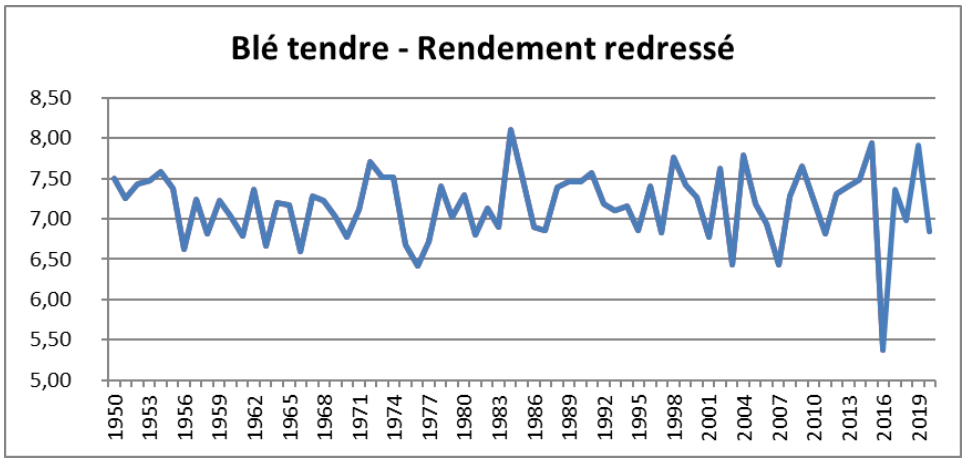
La première étape est d'observer les rendements nationaux de la base agreste. L'équipe tarification les représente graphiquement :



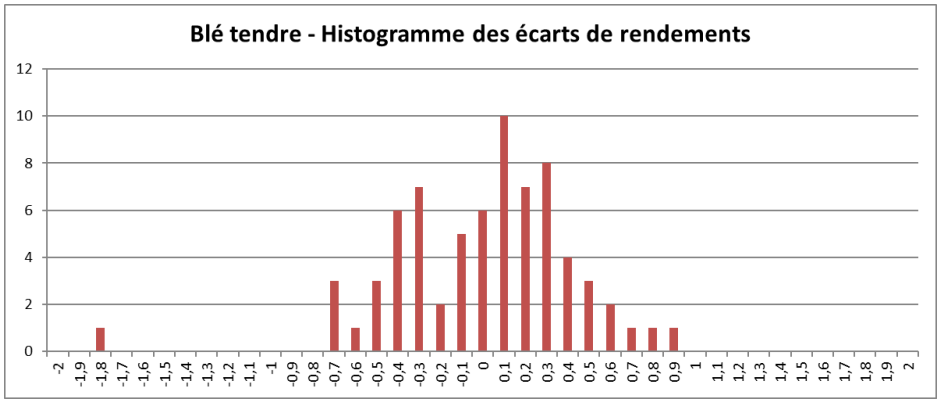
Nous remarquons que le rendement tend à augmenter entre les années 1950 et 1997 puis se stabilise ensuite jusqu'à 2015. Nous pouvons expliquer cette progression par les progrès technologiques en agriculture. Nous ne pouvons pas utiliser ces données ainsi car cela voudrait dire que par exemple 2016 est une bonne année comparée à 1950. Nous ne pouvons pas non plus comparer deux années entre elles si les rendements attendus ne sont pas les mêmes pour chacune d'elles. L'équipe tarification a donc d'abord modélisé la tendance avant la stabilisation, pour pouvoir la supprimer et ramener tous les rendements entre 1950 et 2020 à la même échelle :



La tendance est en rouge. Pour redresser la courbe avant 1997, il faut additionner les écarts à la tendance avec la valeur de rendement en 1997. Nous obtenons alors cette courbe de rendements redressés :

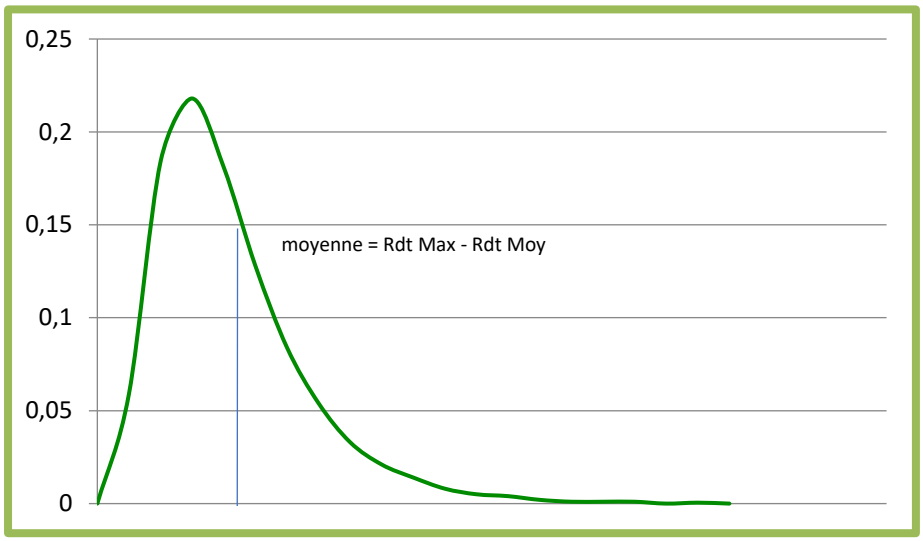


En regardant les écarts de rendements à la moyenne sur l'histogramme ci-dessous, nous reconnaissons la forme d'une loi log-normale inversée. Il est alors supposé que les rendements suivent une loi log-normale inverse.



Cette loi n'est pas une loi classique dont nous pouvons estimer les paramètres. L'équipe tarification utilise alors une astuce permettant de se ramener à une log-normale usuelle.

Ceci est la distribution supposée des rendements c'est à dire une log-normale inverse. Si nous prenons l'opposé des rendements et que nous leur ajoutons le rendement max, nous obtenons alors la distribution de rendement$_{max}$ – rendement qui se présente ainsi :



Nous avons alors une log-normale inverse comme loi de probabilité de l'écart entre le rendement et le rendement maximum.

Il est alors possible d'estimer ses paramètres à l'aide de la moyenne et l'écart type des R_{max}-R. Une fois cette distribution tracée, il suffit d'effectuer la transformation dans l'autre sens et retrouver la distribution des rendements.

Une fois la distribution obtenue, l'équipe tarification prend la probabilité d'avoir une perte de rendement supérieure ou égale à celle de 2016 pour obtenir la période de retour de l'année 2016.

Il est déterminé que la sinistralité de 2016 revient environ **une fois tous les 77 ans**, c'est à dire avec une probabilité de 0,013.

Remarques de la validation :

Lors du passage de la log-normale inverse à la log-normale classique, un rendement max doit être déterminé. Le rendement max pris dans le fichier de l'équipe tarification est le rendement maximum observé depuis 1950. Cela signifierait que théoriquement on ne pourrait pas obtenir dans le futur un rendement supérieur à celui-ci ce qui n'est pas vrai. Il faudrait, comme nous l'a expliqué l'équipe tarification lors de nos échanges prendre un rendement max supérieur à celui observé. Toutefois l'impact sur la période de retour est faible.

- Etude des coûts et coefficient de retraitement

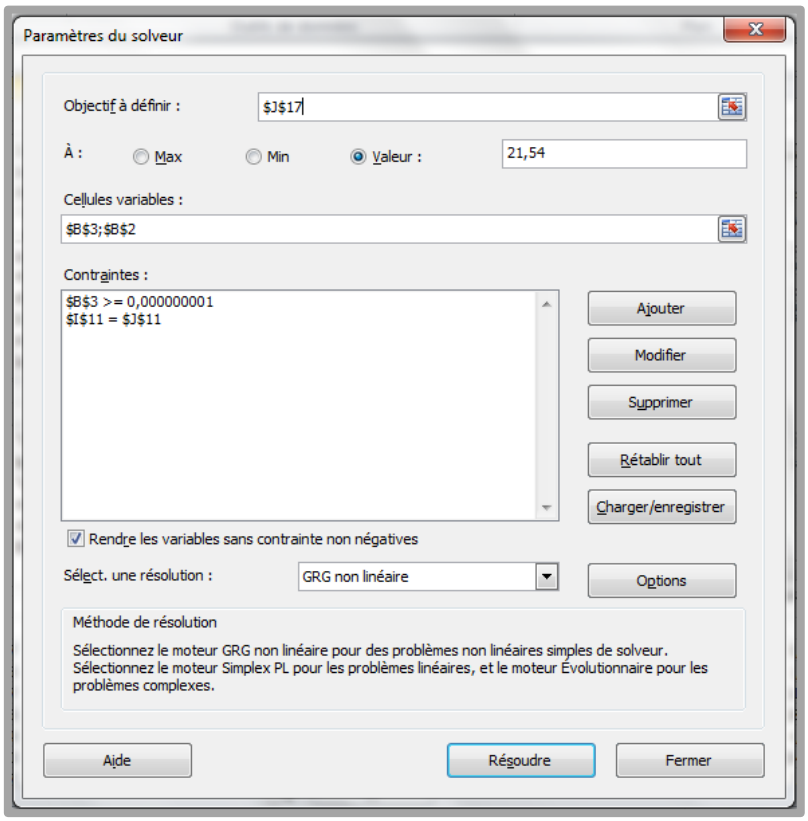
Maintenant, l'input utilisé est le tableau des coûts d'indemnisation pour chaque année de 2008 à 2020.

Année	Coût Réel des indemnisations
2008	2 384 447
2009	1 301 672
2010	3 677 837
2011	34 415 219
2012	19 408 494
2013	10 942 559
2014	11 344 720
2015	3 160 840
2016	175 090 919
2017	7 455 409
2018	9 534 140
2019	4 301 742
2020	33 237 851

L'objectif de l'équipe tarification ici est de déterminer la loi de probabilité des coûts d'indemnisation annuels pour obtenir l'espérance de ces coûts qui sera donc inférieure à la moyenne observée à cause de l'année 2016, et finalement calculer le ratio :

$$\frac{\text{Coût moyen probable}}{\text{Coût moyen observé}}$$

Ce sera le coefficient de retraitement de la constante de tarification. La difficulté de la démarche réside donc dans la détermination de la distribution des coûts. L'équipe tarification suppose que les



coûts suivent une loi log-normale car ils peuvent être assimilés à une perte rendement. En effet, comme vu lors du calcul de la période de retour de 2016, les écarts de rendement $R_{max} - R$ suivent une loi log-normale. Pour déterminer la bonne loi il faut choisir les bons paramètres. Le principe du choix de ces paramètres repose sur un solveur Excel qui va modifier les paramètres jusqu'à que certaines contraintes soient respectées.

Nous voyons que les cellules variables sont B3 et B2 qui correspondent aux paramètres de la loi log-normale.

En parallèle du solveur, le fichier Excel effectue 50 simulations de coûts d'indemnisation annuels sur 10 années. Chacune de ces 50 simulations permet de modéliser les coûts d'indemnisations obtenus au cours des 10 années d'historiques de Groupama. Pour ce faire, pour chacune des 10 années, une probabilité est choisie aléatoirement entre 0 et 1, et le coût pour cette année est déterminé en prenant le quantile correspondant pour la loi log-normale avec les paramètres situés en cellule B2 et B3. Cependant, pour une de ces 10 années, la probabilité n'est pas choisie aléatoirement, mais correspond à la probabilité de retour de l'année 2016 afin de se rapprocher au mieux de la réalité des données de Groupama. Nous obtenons alors 50 coûts d'indemnisations totaux dont nous calculons la moyenne dans la cellule J17.

Si nous regardons la fenêtre du solveur Excel, nous voyons en cellule J17 l'objectif du solveur. La valeur devant être atteinte par cette dernière est 21,54 soit la moyenne des coûts observés sur 10 années en millions d'euros. Le solveur va donc effectuer les étapes suivantes : changer les paramètres de la loi, ré-effectuer les 50 simulations, et ce jusqu'à ce que la moyenne des coûts observés soit le plus proche possible de la moyenne modélisée par les simulations.

Nous avons aussi deux contraintes dans le solveur :

- B3 > 0 pour que la variance des coûts soit non nulle
- I11 = J11, pour que le coût d'indemnisation de 2016 corresponde au quantile 0.86 de la loi des coûts. Pour avoir une cohérence entre la probabilité de retour de 2016 et la probabilité d'obtenir les coûts de 2016.

L'équipe tarification trouve qu'en moyenne les coûts devraient s'élever à 11,7 M par an, alors qu'on en a observé 21,54 à cause de 2016. Les coûts théoriques correspondent à 54% des coûts observés. Il faut donc prendre uniquement 54% de la constante de tarification.

Remarques de la validation :

- Nous recommandons de simuler uniquement 9 années de coûts d'indemnisation pour concorder avec l'historique Groupama.
- Dans les conditions du solveur, les coûts théoriques et observés de 2016 doivent être les mêmes. Cependant on observe deux nombres différents sur le tableur Excel. Après échange avec l'équipe tarification, la première résolution a été faite à la main sans solveur sur la culture Blé tendre Hiver, ce qui explique que les contraintes du solveur ne soient pas exactement respectées. Nous recommandons de recalculer le coefficient de retraitement pour le blé tendre hiver avec le solveur pour les prochaines utilisations.

- En exécutant le solveur avec uniquement 9 années de simulations et en s'assurant que la contrainte d'égalité des coûts de 2016 est respectée, on trouve un coefficient de 60%

4.3.3.3. Tests d'adéquation de la loi Log normale

- Tests d'adéquation de l'équipe tarification

Pour déterminer la bonne loi de distribution des rendements, des tests d'adéquation ont été effectué par l'équipe tarification.

Ces tests ont été effectués à l'aide d'un code R, sur une base de données agreste contenant des données de rendement par Culture/Département/Année.

Le test utilisé est le test de Kolmogorov Smirnov qui indique l'écart entre nos valeurs de rendement et une loi donnée. On va donc chercher la loi pour laquelle la valeur du test est la plus faible, c'est-à-dire la loi qui s'écarte le moins possible de nos données.

Ce test a été effectué par l'équipe tarification, pour chaque Culture/Département dans le cas où ces regroupements présentaient plus de 5 années d'historique. L'objectif étant d'avoir un échantillon de données suffisamment important pour qu'il soit un minimum pertinent. Les regroupements pour lesquels la loi log normale l'emporte sur la loi Weibull sont comptés et inversement, pour déterminer le type de loi de distribution de rendements qui va être utilisée dans le modèle.

Remarques de la validation :

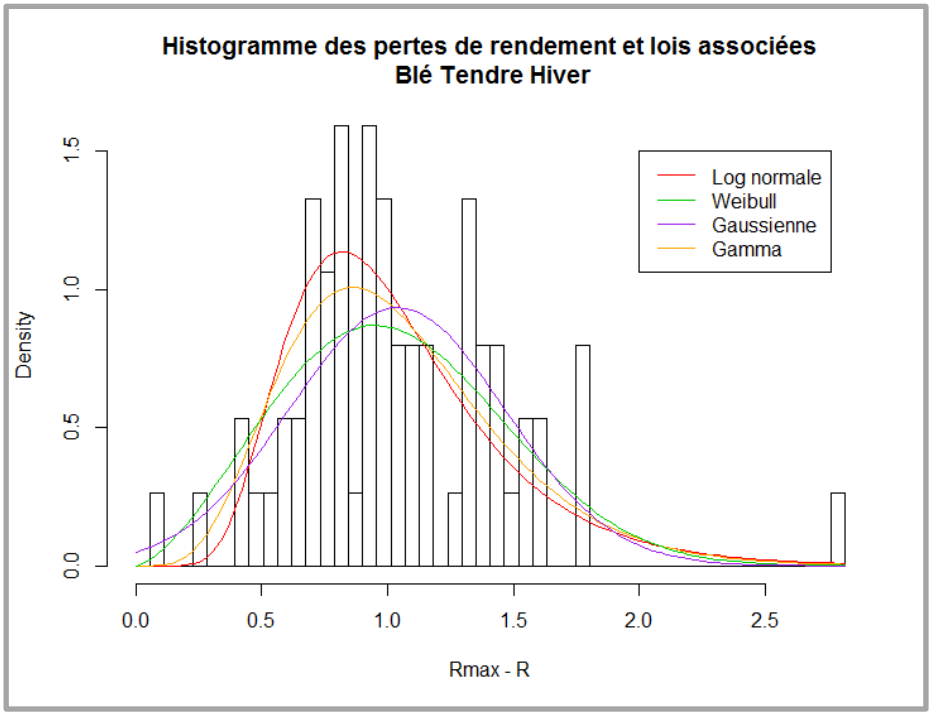
- Le code R fournis par l'équipe tarification effectue le test sur des données de rendements au lieu des écarts entre le rendement maximum théorique et le rendement : Rmax – R
- Nous ne savons pas si les rendements sont redressés comme cela a été fait dans le fichier Excel du retraitement de la constante.
- Le code présente un bug lors de son exécution au moment de la culture C17 dans le département 92, ce qui empêche le test d'être effectué sur l'intégralité de la base. Les résultats présentés en commentaires à la fin du code R sont les mêmes que ceux obtenu après exécution du code partielle à cause du bug.

- Tests d'adéquation de l'équipe de validation

Travail sur R

L'équipe de validation a effectué un travail sur les données agreste nationales du blé tendre hiver afin de tester l'adéquation de la loi log normale.

Nous avons utilisé les données de rendements nationaux redressés par l'équipe tarification lors du calcul de la période de retour de la sinistralité 2016.



Nous avons affiché ci-dessus l'histogramme des pertes de rendement par rapport au rendement maximum théorique de la culture blé tendre hiver.

Puis nous avons réalisé le test de Kolmogorov Smirnov pour chacune des quatre lois : log normale, normale, weibull, et gamma. Nous obtenons ces résultats :

Loi	Valeur du test
Log normale	0,103
Weibull	0,081
Gaussienne	0,100
Gamma	0,075

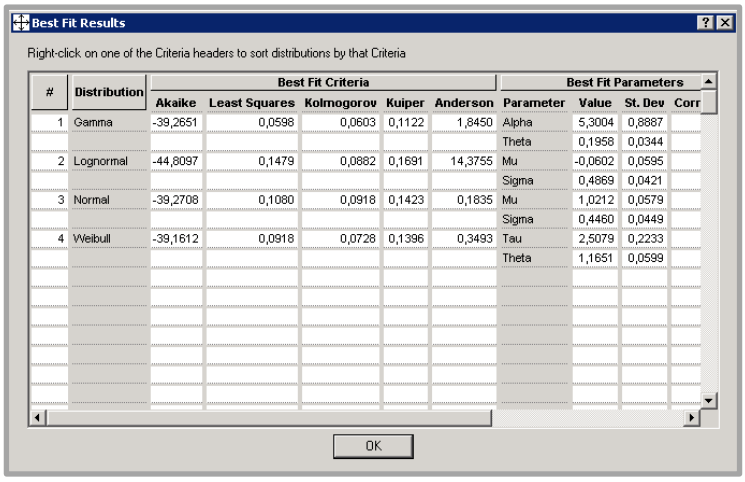
Figure 54 - Résultats du test de Kolmogorov Smirnov

D'après le test de Kolmogorov Smirnov, la loi Gamma ressort en première position.

Travail sur Risk Explorer

Nous avons également utilisé un outil auquel nous avons accès nommé « Risk Explorer ». Il permet de tester différents types de loi rapidement sur un jeu de données et de donner la valeur de différents tests d'adéquation.

Toujours sur les 67 années d'historiques de rendements nationaux de la base agreste pour le blé tendre hiver, nous exécutons l'outil. Nous obtenons ce tableau :



#	Distribution	Best Fit Criteria					Best Fit Parameters			
		Akaike	Least Squares	Kolmogorov	Kuiper	Anderson	Parameter	Value	St. Dev	Corr
1	Gamma	-39,2651	0,0598	0,0603	0,1122	1,8450	Alpha	5,3004	0,8887	
							Theta	0,1958	0,0344	
2	Lognormal	-44,8097	0,1479	0,0882	0,1691	14,3755	Mu	-0,0602	0,0595	
							Sigma	0,4869	0,0421	
3	Normal	-39,2708	0,1080	0,0918	0,1423	0,1835	Mu	1,0212	0,0579	
							Sigma	0,4460	0,0449	
4	Weibull	-39,1612	0,0918	0,0728	0,1396	0,3493	Tau	2,5079	0,2233	
							Theta	1,1651	0,0599	

Les tests « Kuiper » et « Least Squares » sont semblables au test de Kolmogorov, plus la valeur est faible, plus la loi colle bien aux données.

L'adéquation statistique de la loi Gamma sur ces valeurs de rendement est confirmée avec Risk Explorer.

- ⇨ Statistiquement, c'est la loi gamma qui ressort mais cela ne remet pas forcément en cause le choix de la log-normale qui était déjà utilisée dans le modèle précédent.
- ⇨ Cette étude a été faite sur uniquement 71 valeurs de rendement, nous recommandons de poursuivre le travail sur une base de données plus conséquente (avec la méthode utilisée par l'équipe tarification par exemple) et de l'étendre aux autres cultures.

4.3.4. Coefficients de franchise

4.3.4.1. Coefficients de franchise du modèle

Jusqu'ici, l'étude de l'équipe tarification a porté uniquement sur le tarif des produits seuil 25 franchise 25. Cependant différents niveaux de franchise sont vendus et leurs tarifs doivent être adaptés. En effet un contrat à franchise 10% a plus de chances d'être sinistré, et doit donc être plus cher qu'un contrat à franchise 25%. L'équipe tarification a donc effectué une étude afin d'appliquer un coefficient au tarif pour l'adapter à la franchise choisie par l'agriculteur. Ce coefficient sera égal à 1 pour la franchise 25, supérieur à 1 pour les franchises inferieures à 25, et inférieur à 1 pour les franchises supérieures à 25.

L'objectif de cette étude a été, à l'aide de la distribution des rendements, de déterminer la probabilité de déclencher une franchise et ainsi une perte de rendement moyenne probable. Cette dernière sera divisée par celle pour une franchise 25 pour obtenir le coefficient voulu.

Le coût d'indemnisation s'exprime ainsi : $\text{Coût} = \text{Perte}_{\text{rendement}} \times \text{Prix} \times \text{Surface}$

Les pertes de rendement sont donc les seules variables aléatoires du modèle déterminant les coûts d'indemnisation la surface de la culture et le prix de vente étant fixés au départ. Donc s'intéresser uniquement aux pertes de rendement est suffisant pour déterminer les variations de tarif entre les niveaux de franchises.

Toute l'étude a été effectuée sur Excel. L'input est un tableau avec pour chaque culture, le rendement moyen, le rendement maximum paramètre de loi log-normale inverse, l'écart type des rendements.

Rendement max	Rendement moy	Rmax - Rmoy	Rdt_Max / Rdt_Moy	Rdt min	Ecart-type (20 ans)
12,07	9,94	2,12	121%	7,82	1,29
5,34	4,97	0,37	107%	4,60	0,23
5,77	5,60	0,17	103%	5,43	0,10
96,93	80,92	16,00	120%	64,05	6,63
5,85	5,05	0,80	116%	3,57	0,50

Ce sont ces paramètres qui vont permettre de déterminer la distribution des rendements de la même manière que lors du retraitement de la constante de tarification.

Nous récupérons les variables utiles dans le tableau en input. Ensuite, on continue en prenant pour exemple le blé tendre Hiver :

Moyenne	7,16
Ecart-type	0,58
Culture_Saison	BLE-TENDRE_Hiver

La moyenne et l'écart type vont être utilisés pour construire la distribution des R_{max} – R, suivant une log-normale. Nous devons donc avoir ici R_{max}-R_{moy} comme moyenne. Or dans le fichier de l'équipe tarification, le 7.16 représente R_{moy}, ce n'est pas la valeur souhaitée. Nous avons corrigé la formule de la cellule pour aller chercher R_{max}-R_{moy} dans le tableau input et continuer l'étude :

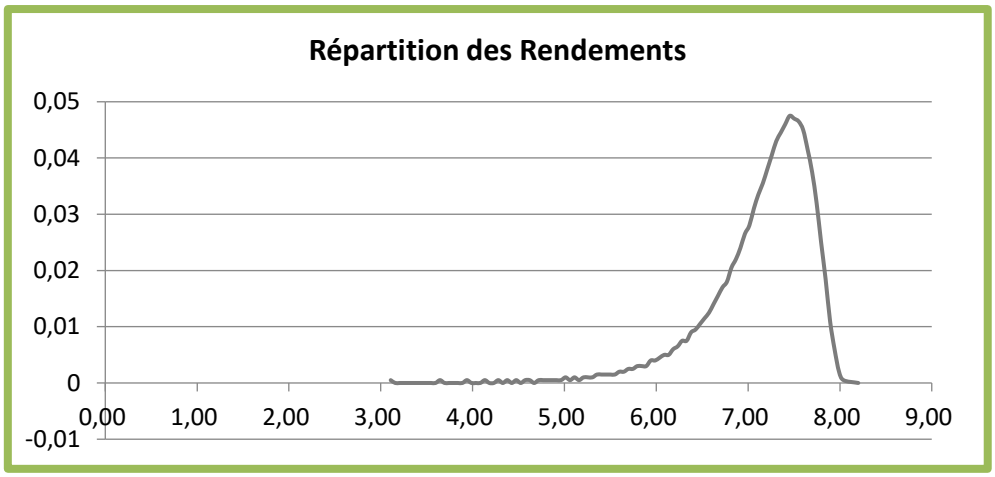
Moyenne	1,04
Ecart-type	0,58
Culture_Saison	BLE-TENDRE_Hiver

Ces paramètres sont utilisés pour obtenir la distribution log-normale des R_{max}-R, qui sera transformée à l'aide de la valeur de $R_{max.}$ Cependant dans le fichier de l'équipe tarification, le R_{max} utilisé est le maximum des R_{max}-R (5,04). Nous avons donc rajouté le bon R_{max} en entrée et corrigé la transformation de la distribution.

Moyenne	1,04
Ecart-type	0,58
Culture_Saison	BLE-TENDRE_Hiver

Rmax	8,2

Cela nous permet d'obtenir la bonne distribution des rendements pour le blé tendre hiver :



Un tableau est ensuite construit avec une ligne pour chaque niveau de franchise.

Franchise	Déclenchement	Fréquence	Coût moyen	Tarif
30%	5,02	0,80%	0,58	0,5%
29%	5,09	0,90%	0,58	0,5%
28%	5,16	1,05%	0,56	0,6%
27%	5,23	1,15%	0,58	0,7%
26%	5,30	1,30%	0,58	0,8%
25%	5,38	1,50%	0,56	0,8%

Le déclenchement indique pour quel rendement la franchise est activée, c'est à dire pour quel rendement une culture est sinistrée en fonction de son niveau de franchise. Le déclenchement s'effectue quand le rendement de l'agriculteur descend en dessous (1-Franchise)*R_{moy}. Par exemple pour une franchise 30%, le rendement de déclenchement va être 70% du rendement moyen, c'est celui que nous trouvons dans la deuxième colonne.

Ensuite pour chaque niveau de franchise une probabilité de déclenchement est calculée à l'aide du rendement de déclenchement et de la distribution des rendements. C'est la colonne « fréquence » du tableau. Nous voyons bien que plus la franchise est élevée, plus la probabilité de déclenchement est faible, et inversement.

Les écarts de rendement à indemniser sont calculés. Nous trouvons en colonne chacun des différents niveaux de franchise, et sur les lignes les pertes de rendement à indemniser pour 2000 rendements possibles correspondant aux quantiles déterminés. Pour une franchise donnée, pour chaque ligne, donc pour chaque rendement, si ce dernier est inférieur au rendement de déclenchement alors la

cellule sera égale à la différence entre le rendement de déclenchement et le rendement supposé obtenu sur cette ligne.

La moyenne de ces pertes de rendement est retranscrite pour chaque franchise dans le tableau de synthèse sous la colonne coût moyen.

Pour finalement obtenir une perte de rendement moyenne probable à indemniser dans la colonne « Tarif » du tableau de synthèse en multipliant les colonnes « coût moyen » et « fréquence ». On remarque que les tarifs seront plus élevés pour des faibles franchises car elles auront plus de chance de s'activer. À l'inverse ils seront faibles pour des franchises élevées qui s'activeront plus rarement.

Il reste à déterminer pour terminer, les coefficients de franchise. Comme l'équipe tarification a jusque-là bâti le modèle tarifaire sur les contrats à franchise 25, elle a décidé de mettre le coefficient de cette franchise à 1 et de calculer les autres par rapport à celui-ci. Les coefficients de franchises sont alors calculés en divisant les tarifs dans le tableau de synthèse par le tarif de la franchise 25. Nous avons donc un coefficient égal à 1 pour la franchise 25, des coefficients inferieurs à 1 pour des franchises élevées, et supérieurs à 1 pour des franchises plus faibles.

Comme dans le fichier de l'équipe tarification le R_{max} choisi n'était pas le bon (en ayant corrigé la moyenne de la feuille « Entrées »), les coefficients étaient différents de ce qu'ils devaient réellement être.

Dans le tableau ci-dessous on observe les résultats obtenus avant (rouge) et après (violet) la correction du R_{max} :

Franchise	Déclenchement	Fréquence	Coût moyen	Tarif	Coef Blé Rmax corrigé	Coef Blé Rmax d'origine	Ecart par ratio des coefficients
30%	5,02	0,80%	0,58	0,5%	0,55	0,70	79%
25%	5,38	1,50%	0,56	0,8%	1,00	1,00	100%
20%	5,73	2,75%	0,57	1,6%	1,84	1,44	128%
15%	6,09	5,25%	0,56	2,9%	3,45	2,07	167%
10%	6,45	10,30%	0,54	5,58%	6,59	3,01	219%

Nous remarquons que les coefficients pour la franchise 25 sont bien égaux à 1 pour les deux R_{max}. Les coefficients corrigés dans la colonne violette sont plus dispersés que ceux dans la colonne d'origine. En effet, pour les petites franchises ils sont plus élevés que ceux calculés avec le mauvais Rmax, et inversement, ils sont plus faibles pour les franchises plus élevées. Cela signifie que le tarif sera globalement plus juste avec le nouveau R_{max} : les contrats les moins risqués (franchise 30%) paieront moins et les contrats les plus risqués (franchise 10%) paieront plus.

Risque	Franchise	Capitaux assurés		Surfaces	Nombre de contrats	Primes portefeuille chargées 2017	
		M€	% du total	k ha		k€	% du total
Autres Aléas	0	0	0%	0		0	0%
Autres Aléas	5	0	0%	0		0	0%
Autres Aléas	10	87	2%	6	2 217	1 159	1%
Autres Aléas	15	165	4%	42	3 831	4 931	6%
Autres Aléas	20	120	3%	45	2 935	6 218	4%
Autres Aléas	25	3 954	87%	2 510	118 157	70 223	84%
Autres Aléas	30	203	4%	133	6 635	3 930	5%
Autres Aléas	Total	4 528	100%	2 737	133 775	83 461	100%

Figure 55 - Tableau de répartition en capitaux et en nombre de contrats du portefeuille » Autres Aléas » 2017 selon le niveau de franchise souscrite

Remarques de la validation :

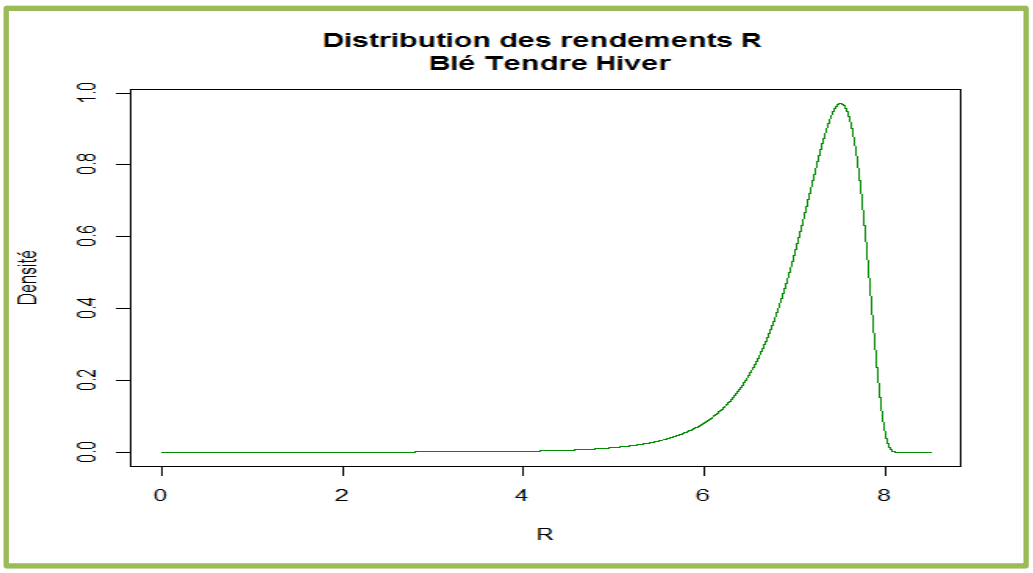
Nous recommandons de corriger les deux paramètres utilisés dans le calcul de la distribution des rendements : choisir R_{max}-R_{moy} dans la cellule moyenne en « Entrées », et choisir le bon rendement max dans la feuille « Loi ». La matérialité de cet impact est limitée principalement par la répartition des capitaux assurés par franchise qui sont très largement souscrits à franchise 25% (87%) et d'autre part par la compensation entre les écarts à franchise 30% (4% du portefeuille, 21% d'écart) et ceux à franchise 20% (3% du portefeuille, 28% d'écart)

4.3.4.2. Proposition de calcul alternative

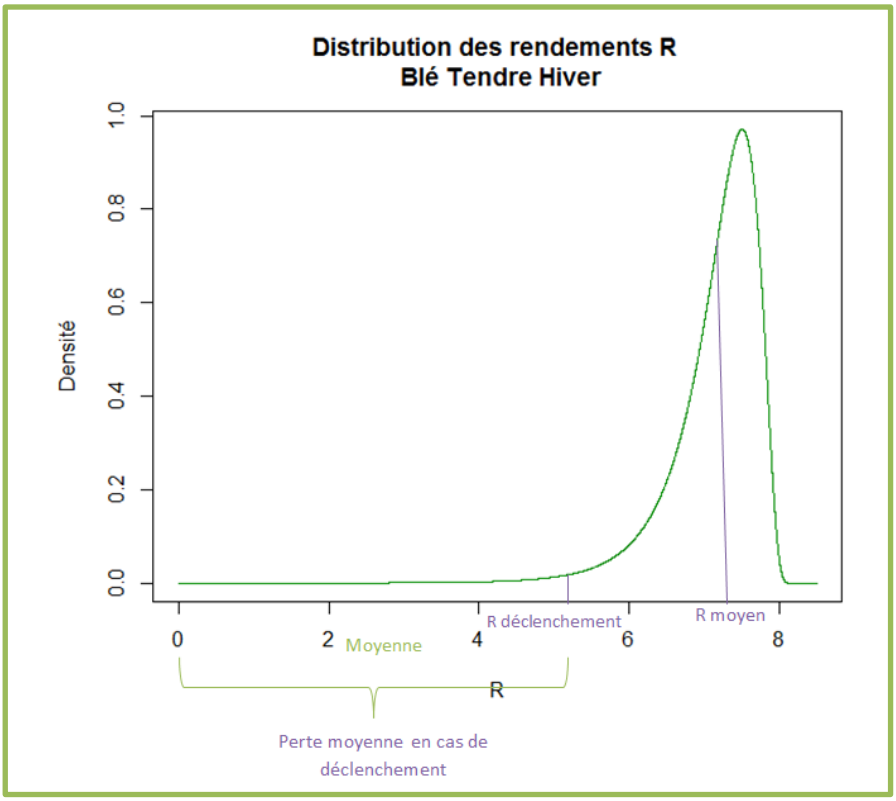
Nous avons essayé à la validation, une méthode de calcul des coefficients de franchise légèrement différente de celle utilisée par l'équipe tarification.

Méthode de l'équipe tarification :

Elle a la distribution des rendements qui se présente ainsi :



Pour un niveau de franchise donné (par exemple 30%) elle calcule le rendement de déclenchement de la franchise en prenant 70% du rendement moyen.



L'équipe tarification fait ensuite la moyenne des pertes de rendement pour les rendements inferieurs au rendement de déclenchement de la franchise. Ceci dans le butd' obtenir une perte de rendement moyenne en cas déclenchement, qui sera ensuite multipliée par la probabilité de déclenchement.

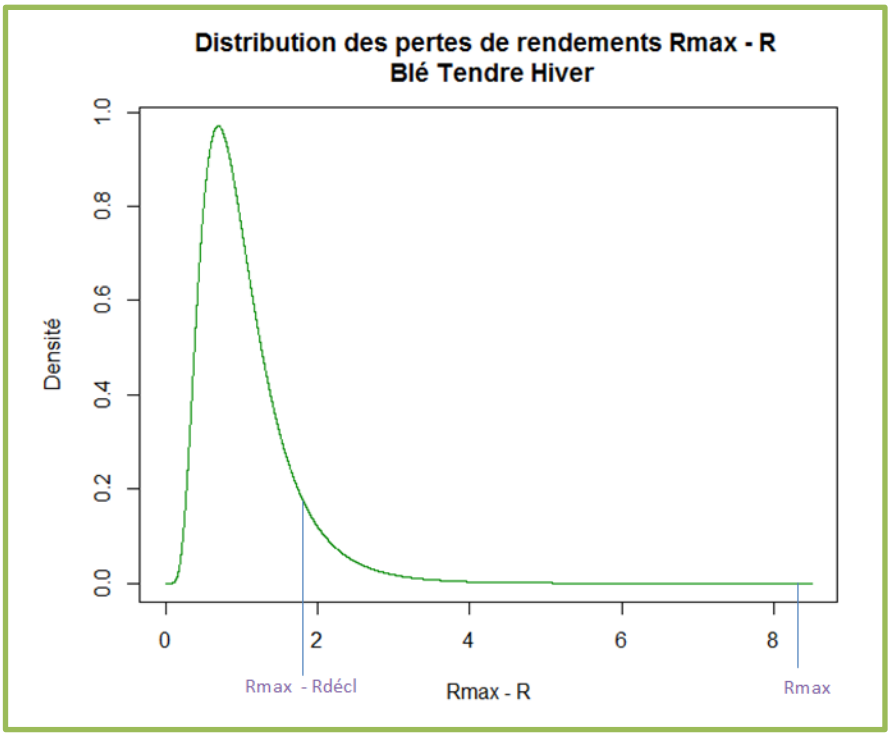
Idée de l'équipe de validation :

En multipliant la probabilité de déclenchement par la perte de rendement moyenne nous associons la probabilité de déclenchement à toutes les pertes supérieures à 30%. Le résultat voulu se rapproche d'une espérance de perte de rendement sachant qu'il y a indemnisation, mais chacune des pertes n'est pas pondérée par sa probabilité d'occurrence.

Nous avons donc essayé d'utiliser une méthode avec intégration à l'aide d'un code R. En partant de la loi log normale inverse modélisée par l'équipe tarification, nous intégrons les pertes de rendement (R déclenchement – R) quand le rendement est inférieur au rendement de déclenchement avec la densité de probabilité de la loi log normale inverse :

$$\int_0^{Rdécl} (Rdécl - R) \times dP\_lognorm\_inverse(R)$$

Pour vérifier nos résultats, nous effectuons le même type de calcul mais avec la loi log normale modélisant la distribution des pertes de rendement par rapport au rendement maximum théorique : Nous intégrons donc les pertes de rendements supérieures à Rmax – Rdéclenchement



$$\int_{Rmax-Rdécl}^{Rmax} [x - (Rmax - Rdécl)] \times dP\_lognrom(x)$$

Avec « x » les pertes Rmax – R.

Nous obtenons ces résultats :

Franchise	CoefDETP	Coef lognorm_inverse	Coef lognrom
30%	0,551	0,565	0,565
25%	1,000	1,000	1,000
20%	1,843	1,810	1,810
15%	3,453	3,356	3,356
10%	6,594	6,379	6,379
5%	13,061	12,384	12,384

Remarques de la validation :

Nous obtenons les mêmes coefficients avec nos deux méthodes, et ces derniers restent assez proches de ceux obtenus avec la méthode choisie par l'équipe tarification.

4.4. Le nouveau modèle de tarification

Suite à cette validation du modèle tarifaire, nous avons proposé une démarche tarifaire alternative à l'équipe de modélisation qui respecte ces deux critères :

- Modéliser des tarifs en adéquation avec la sinistralité constatée de Groupama et le changement climatique futur.
- Assurer une cohérence des hypothèses et méthodes entre le modèle tarifaire et le modèle interne de la MRC.

La démarche tarifaire proposée par l'équipe de la validation peut se résumer en 3 grandes étapes :

- Maillage : Nous suggérons de diviser le maillage du territoire en deux sous étapes :

  - Le lissage des données pour avoir de l'information sur les communes qu'on ne les a pas dans notre base de données pour couvrir toute la France.
  - La classification des données, consiste à utiliser les données lisser pour répartir les communes de la France en zones homogènes de risques, afin de fixer un tarif pour chaque zone homogène.

- Prime pure : Cette étape consiste à calculer la prime pure au sein de chaque zone. Pour ce faire nous proposons de tester plusieurs modèles statistiques :

  - Modèle avec décomposition fréquence coût

- Modèle sans décomposition fréquence coût :
    - Approche issue du modèle interne.
    - Approche basée sur le calibrage d'une loi sur les Sinistres/capital assuré (permettant de challenger les résultats du modèle interne)

- Chargement de sécurité : Nous pouvons ajouter par zone un chargement de sécurité.

  Prime pure d'une zone = E [perte] + alpha * écart-type des pertes

Conclusion

Les modèles sont de plus en plus utilisés chez les assureurs dans différents domaines notamment en ce qui concerne la tarification des risques et la détermination du capital économique. Ces modèles sont élaborés pour des objectifs précis. Ils s'appuient non seulement sur des hypothèses mais ils ont également des limites qui doivent être connues et comprises pour s'assurer de leur bonne utilisation dans les prises de décisions stratégiques et opérationnelles.

La validation est au cœur de cette bonne utilisation, puisqu' elle permet par le biais d'un processus et d'outils de validation de **s'assurer que ces modèles remplissent effectivement leurs objectifs. C'est aussi un enjeu fondamental en interne d'approfondissement de la connaissance et de la compréhension du modèle (notamment quant à ses limites).**

En ce qui concerne la validation du modèle interne de Groupama du risque MRC, les constats relevés ne remettent pas en question la pertinence avec le profil de risque. Les approches retenues par la modélisation répondent à son objectif d'estimation du capital économique. Ils sont basés sur les résultats des calculs miroirs et des analyses complémentaires des modélisations alternatives proposés dans ce mémoire

D'autres part, la validation du modèle tarifaire de Groupama du risque MRC a donné lieu à des constats qui justifient l'adéquation du modèle avec son objectif de détermination de la prime du risque MRC.

La validation propose dans ce mémoire à l'équipe de modélisation de tester des modèles alternatifs qui permettraient à la fois d'atteindre les deux objectifs de tarification et capital économique d'une part, mais également d'assurer une convergence des deux modèles d'autres part. Ces modèles seraient encore plus pertinents avec leur confrontation à la réalité du risque et aux avis des experts métier.

Le risque climatique sur récoltes est toujours croissant et fluctuant, par conséquent les modèles vont encore être amenés à évoluer pour mieux capter ce changement. La veille actuarielle constitue un élément clé pour les travaux de validation et ce mémoire en est bien la preuve.

Annexe 1 : Estimation des paramètres

Détermination de $\boldsymbol{\beta_0}$:

$$\frac{\partial \sum_{i=1}^{n} {\varepsilon_i}^2}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

D'où,

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

Avec $\sum_{i=1}^{n} \beta_0 = n\beta_0$, nous avons :

$$\frac{\sum_{i=1}^{n} Y_i}{n} = \beta_0 + \beta_1 \frac{\sum_{i=1}^{n} X_i}{n}$$

Et finalement nous trouvons, $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$

Détermination de $\boldsymbol{\beta_1}$:

$$\frac{\partial \sum_{i=1}^{n} {\varepsilon_i}^2}{\partial \beta_1} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$$

D'où,

$$\sum_{i=1}^{n} Y_i X_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} {X_i}^2$$

Nous remplaçons β_0 par $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$. Nous obtenons ainsi :

$$\sum_{i=1}^{n} Y_i X_i = \bar{Y} \sum_{i=1}^{n} X_i - \beta_1 \bar{X} \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} {X_i}^2$$

$$\sum_{i=1}^{n} Y_i X_i - \bar{Y} \sum_{i=1}^{n} X_i = \beta_1 \left(\sum_{i=1}^{n} {X_i}^2 - \bar{X} \sum_{i=1}^{n} X_i \right)$$

Donc

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} Y_i X_i - \bar{Y} \sum_{i=1}^{n} X_i}{\left(\sum_{i=1}^{n} {X_i}^2 - \bar{X} \sum_{i=1}^{n} X_i \right)} = \frac{\overline{YX} - \bar{Y}\bar{X}}{\overline{X^2} - \bar{X}^2}$$

Finalement, nous trouvons : $\widehat{\beta_1} = \frac{Cov(X,Y)}{V(X)}$

Les estimateurs des Moindres Carrés Ordinaires sont des estimateurs linéaires, non biaisés, convergents et à variance minimale c'est-à-dire efficaces (Best Linear Unbiased Estimators).

Annexe 2 : Démonstration R-carré.

Nous pouvons montrer que [16]:

$$\sum_{i=1}^{n} e_i = 0$$

Dans un premier temps, $\sum_{i=1}^{n} e_i = 0$, implique que $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$ donc $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$. Cela signifie qu'il y a égalité entre la moyenne de la série à expliquer et la moyenne de la série ajustée :

$$\bar{y}_i = \bar{\hat{y}}_i \quad (1)$$

D'autres part, nous avons : $y_i - \bar{y}_i = (\hat{y}_i - \bar{y}_i) + (y_i - \hat{y}_i)$

Nous en déduisons que :

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)(y_i - \hat{y}_i)$$

Donc :

$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)e_i$ (2) (Car $e_i = y_i - \hat{y}_i$)

Nous nous intéressons au dernier terme de cette équation :

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)e_i = \sum_{i=1}^{n}\hat{y}_i e_i - \bar{y}_i\sum_{i=1}^{n} e_i$$

$$= \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)e_i - \bar{y}_i\sum_{i=1}^{n} e_i$$

$$= \beta_0\sum_{i=1}^{n} e_i + \beta_1\sum_{i=1}^{n} x_i e_i - \bar{y}_i\sum_{i=1}^{n} e_i$$

Nous pouvons montrer aussi que [17]: $\sum_{i=1}^{n} x_i e_i = 0$

D'où :

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)e_i = 0$$

D'après l'équation (2), nous avons :

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^{n} e_i^2$$

De plus, avec l'équation (1), nous avons :

[16] Voir Annexe 3
[17] Voir Annexe 4

$$\sum_{i=1}^{n}(y_i - \overline{y_i})^2 = \sum_{i=1}^{n}(\widehat{y_i} - \overline{\widehat{y_i}}\ )^2 + \sum_{i=1}^{n} {e_i}^2$$

La somme des carrés totale (SCT) = la somme des carrés expliquées (SCE)+ la somme des carrés résiduels (SCR)

Cette équation va nous permettre de juger de la qualité de l'ajustement d'un modèle. En effet, plus la variance expliquée est proche de la variance totale, meilleur est l'ajustement du nuage de points par la droite des moindres carrés.

Nous définissons le R^2 par :

$$R^2 = \frac{SCE}{SCT}$$

Ce dernier donne la part de la variabilité totale de Y expliquée par X.

Annexe 3 : Démonstration de la somme $\sum e_i = 0$

Montrer que $\sum e_i = 0$:

$$e_i = Y_i - \widehat{Y_i} = Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i$$

On a : $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$

D'où :

$$e_i = Y_i - \widehat{Y_i} = Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i$$

$$e_i = Y_i - \bar{Y} + \widehat{\beta_1}\bar{X} - \widehat{\beta_1} X_i$$

En sommant sur les i, on a :

$$\sum e_i = \sum (Y_i - \bar{Y} + \widehat{\beta_1}\bar{X} - \widehat{\beta_1} X_i)$$

$$\sum e_i = \sum (Y_i - \bar{Y} + \widehat{\beta_1}\bar{X} - \widehat{\beta_1} X_i)$$

$$\sum e_i = \sum Y_i - \sum \bar{Y} + \sum \widehat{\beta_1}\bar{X} - \sum \widehat{\beta_1} X_i$$

$$\sum e_i = n\bar{Y} - n\bar{Y} + n\widehat{\beta_1}\bar{X} - n\widehat{\beta_1}\bar{X} = 0$$

Donc :

$$\boxed{\sum e_i = 0}$$

Annexe 4 : Démonstration de la somme $\sum X_i e_i = 0$

Montrer que $\sum X_i e_i = 0$:

$$\sum X_i e_i = \sum X_i (Y_i - \widehat{Y_i}) = \sum X_i Y_i - \sum X_i \widehat{Y_i}$$

$$= \sum X_i Y_i - \sum X_i (\widehat{\beta_0} + \widehat{\beta_1} X_i)$$

$$= \sum X_i Y_i - \widehat{\beta_0} \sum X_i - \widehat{\beta_1} \sum {X_i}^2$$

On a : $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$

$$= \sum X_i Y_i - (\bar{Y} - \widehat{\beta_1}\bar{X}) \sum X_i - \widehat{\beta_1} \sum {X_i}^2$$

$$= \sum X_i Y_i - \bar{Y} \sum X_i + \widehat{\beta_1}\bar{X} \sum X_i - \widehat{\beta_1} \sum {X_i}^2$$

$$= n\overline{XY} - n\bar{Y}\bar{X} + n\widehat{\beta_1}\bar{X}^2 - n\widehat{\beta_1}\overline{X^2}$$

$$= n\overline{XY} - n\bar{Y}\bar{X} + n\widehat{\beta_1}\bar{X}^2 - n\widehat{\beta_1}\overline{X^2}$$

$$= n[\overline{XY} - \bar{Y}\bar{X} - \widehat{\beta_1}(\overline{X^2} - \bar{X}^2)]$$

On a : $\widehat{\beta_1} = \frac{\overline{XY} - \bar{Y}\bar{X}}{\overline{X^2} - \bar{X}^2}$

$$= n[\overline{XY} - \bar{Y}\bar{X} - \overline{XY} + \bar{Y}\bar{X}]$$

Donc :

$$\sum X_i e_i = 0$$

Annexe 5 : Démonstration produit de lois log normales donne une loi log normale

Soient X et Y deux variables aléatoires positives suivants chacune une loi log normale.

On a : $XY = e^{\log(XY)} = e^{\log(X)} + e^{\log(Y)}$

Nous rappelons qu'une variable aléatoire Z suit une loi log normale si log(Z) suit une loi normale.

Autrement dit, comme X et Y suivent toutes les 2 une loi log-normale, XY suit une loi log-normale si et seulement si log(X) + log(Y) suit une loi normale.

D'où XY suit une loi log-normale si et seulement si (log(X), log(Y)) est un vecteur gaussien.

Or (log(X), log(Y)) est un vecteur gaussien.

Donc XY suit une loi log-normale.

Bibliographie

[1] Documentations internes de Groupama

[2] Régis BOURBONNAIS, Econométrie, *Cours et exercices corrigés 9e édition*, 2015

[3] Robert G. BROWN, *Exponential Smoothing for Predicting Demand*, Cambridge, Massachusetts (États-Unis), Arthur D. Little Inc., 1956

[4] Régis BOURBONNAIS et Michel TERRAZA, *Analyse des séries temporelles, application à l'économie et à la gestion*, 2010

[5] Laurent ROUVIERE, *Régression logistique avec R*, Université Rennes 2, UFR Sciences Sociales

[6] Laurent ROUVIERE, *Introduction aux GLM*, 2015 : https://perso.univ-rennes2.fr/system/files/users/rouviere_l/chapitre1_glm.pdf.

[7] Ministère de l'agriculture et de l'alimentation, *L'assurance multirisque climatique des récoltes*, Campagne 2021

[8] FFA

[9] BRISSON, Nadine, GATE, Philippe, GOUACHE, David, *et al*. « Why are wheat yields stagnating in Europe? A comprehensive data analysis for France *». Field Crops Research*, 2010, vol. 119, no 1, p. 201-212

Table des figures