Algorithme de clustering appliqué à la tarification d'un produit d'assurance MPCI

Mémoire d'actuariat

Amaury VATRY

2022

Superviseur Institut des Actuaires :

Olivier Lopez

Superviseur PartnerRe :

Caroline Chedal-Anglay

Institut des ACTUAIRES

PartnerRe

08/02/2023

Amaury

08/02/2023

PartnerRe

Algorithme de clustering appliqué à la tarification d'un produit d'assurance MPCI

Mémoire d'actuariat

Amaury VATRY

2022

Superviseur Institut des Actuaires :

Olivier Lopez

Superviseur PartnerRe :

Caroline Chedal-Anglay

Institut des ACTUAIRES

PartnerRe

Abstract

Key words: Agriculture, agroclimatic zones, burning costs, loss costs, clustering, unsupervised learning, k-means algorithm, adverse selection

Pricing an agricultural insurance product with a unique price at national level generates a risk of adverse selection especially when the studied country has different agroclimatic regions with very distinct results. Farmers in unfavorable regions are indeed more inclined to underwrite insurance than others. Dividing the territory into zones with homogeneous results is thus necessary to reduce this phenomenon. However, in this case, the burning cost method, commonly used in pricing, comes up against many difficulties: defining the right level of details, proposing a price for all regions even when data is insufficient, establishing an objective criterion to distinguish the zones, and last but not least, obtaining results which make sense in terms of geography. To meet these challenges, a specific pricing method has been developed. The protocol simultaneously capitalises on the burning cost method, the knowledge of experts and an unsupervised learning clustering method: the k-means algorithm. This methodology reconciles in a very satisfactory way the actuarial results and the available agroclimatic information, while limiting at most the impact of subjective decisions.

Résumé

Mots clés : Agriculture, zones agroclimatiques, burning costs, taux de pertes, segmentation, clustering, apprentissage non-supervisé, algorithme des k moyennes, antisélection

La tarification d'un produit d'assurance agricole avec un prix unique à l'échelle nationale engendre un risque d'antisélection, en particulier lorsque le pays étudié possède différentes régions agroclimatiques avec des résultats bien distincts. En effet, les agriculteurs des régions défavorables sont plus enclins à s'assurer que les autres. Pour limiter cet effet, il faut donc diviser le territoire en zones aux résultats homogènes. La méthode du burning cost, habituellement utilisée en tarification, se heurte alors à de nombreuses difficultés : le choix de l'échelle étudiée, l'obligation de proposer un tarif à des régions dont les données sont insuffisantes, la nécessité d'un critère objectif pour distinguer les zones, enfin et surtout, l'importance de la cohérence géographique qui en résulte. Pour répondre à ces enjeux, un protocole de tarification bien précis a été développé. Il exploite à la fois la méthode du burning cost, la connaissance des experts et une méthode de segmentation par apprentissage non-supervisé : l'algorithme des k moyennes. Cette méthodologie concilie de manière très satisfaisante les résultats actuariels et les informations agroclimatiques disponibles, tout en limitant au maximum l'impact de certaines décisions subjectives.

Remerciements

Je tiens à exprimer toute ma gratitude à ma manager Caroline Chedal-Anglay, qui m'a accueilli dans son équipe, formé à la tarification puis soutenu dans ce projet et a supervisé mon mémoire.

Je souhaite aussi remercier très chaleureusement mon collègue et mentor Patrice Fourgassie avec lequel j'ai conçu le modèle exposé dans ce mémoire et qui, jour après jour, depuis trois ans me fait bénéficier de sa grande expérience et de son savoir-faire infiniment précieux.

Un grand merci à Nicolas Chatelain qui a, lui aussi, grandement contribué aux travaux exposés ici et m'a fait profiter de sa grande expérience en réassurance de l'agriculture.

Je tiens à remercier sincèrement Sylvain Jarrier qui a soutenu mon projet et a généreusement relu mon mémoire.

Merci à Isabelle, Sili, Rafael, Cindy et Kathrin, mes collègues directs qui m'ont beaucoup appris également et avec qui c'est un réel bonheur de collaborer.

Je remercie aussi vivement M. Olivier Lopez, mon second superviseur qui a pu apporter un regard extérieur et ses conseils pour la rédaction du mémoire et la présentation orale.

Enfin, mes remerciements s'adressent à l'équipe de direction du Collège des Ingénieurs, qui a construit le cadre nécessaire à l'obtention du diplôme d'actuaire et m'a encouragé dans ce projet.

Pour terminer, je tiens à rendre hommage à Gilles Maret avec lequel j'ai fait mes premiers pas dans le monde de la réassurance. Gilles était brillant, avenant et très investi dans sa profession. Il était d'ailleurs un membre particulièrement actif de l'Institut des actuaires et présidait notamment la Commission d'agrément depuis de nombreuses années. Gilles est malheureusement parti trop tôt le 16 juillet 2022.

Table des matières

Confidentialité et anonymisation

Ce mémoire d'actuariat a été réalisé en se basant sur des données réelles. Dans le but de préserver l'anonymat du client, de ses données et de celles de PartnerRe, de nombreux chiffres et noms ont été modifiés intentionnellement. Notamment, l'assureur sera dénommé CéréAssure et le pays qu'il couvre, le Céréland.

Introduction

Le réassureur est souvent défini de manière simpliste comme l'assureur des assureurs. Son cœur de métier consiste effectivement à prendre en charge les risques cédés par les assureurs mais son rôle d'expert du risque et de conseiller n'est pas négligeable. En réalité, il participe souvent lui-même à la conception et la tarification des produits d'assurance. Ce mémoire s'inscrit dans ce cadre et porte sur la tarification d'un produit d'assurance réalisée par PartnerRe pour son client CéréAssure.

CéréAssure est un assureur dont le produit phare est une couverture multi-périls sur les récoltes au Céréland. Dans ce pays, les agriculteurs cultivent des céréales essentiellement arrosées par l'eau de pluie, ce qui génère des résultats fortement volatils dépendant des années plus ou moins sèches. De plus, le Céréland a la particularité de posséder des zones agroclimatiques très variées à l'origine de résultats très différents en fonction des régions.

En 2021, après avoir pratiqué un tarif unique à l'échelle nationale de 2013 à 2020, CéréAssure contacte PartnerRe pour définir une nouvelle grille de tarifs en se basant sur les 8 années de données récoltées. PartnerRe se fixe alors l'objectif de proposer des tarifs plus proches de la réalité de chaque zone agroclimatique dans le but de réduire le risque d'antisélection (cf. partie I-A-0).

Cependant la tâche s'avère très rapidement plus difficile que prévu. La méthode de tarification en agriculture essentiellement basée sur la méthode du burning cost se heurte à différents problèmes. Le choix de l'échelle est difficile car les régions sont très vastes tandis que de nombreux départements ne possèdent pas de résultats historiques suffisamment significatifs. En particulier, on peine à proposer une segmentation qui soit à la fois cohérente sur le plan actuariel et sur le plan géographique (cf. V-C-3).

De fil en aiguille, après différentes tentatives, une méthodologie de tarification originale a vu le jour. Celle-ci a donné des résultats très satisfaisants en surmontant les difficultés exposées ci-dessus. De plus, cette méthode a le mérite de pouvoir se généraliser à d'autres problèmes d'ordre géographique. Il paraissait donc pertinent de relater son élaboration dans un mémoire.

Les trois premières parties de celui-ci établissent le contexte, présentent le produit et décrivent les enjeux et objectifs qui en découlent. La partie I expose d'abord le cadre de l'assurance en agriculture (enjeux, antisélection, intervention de l'État, points de vue de l'assureur et du réassureur…) ainsi que le produit d'assurance, son histoire et les objectifs des travaux réalisés ici. La partie II présente le fichier ; elle étudie la qualité des données et explique le prétraitement de celles-ci (transformation en données as-if) pour la suite de l'analyse. Puis, la partie III rappelle les fondements théoriques et mathématiques de la méthode du burning cost ainsi que ses limites sur le plan pratique et les enjeux qui en découlent : choix d'une échelle suffisamment significative (niveau de granularité), recherche de simplicité et recherche de cohérence géographique.

Viennent ensuite les parties IV, V et VI qui décrivent trois différentes méthodes de tarification, de la plus simple à la plus aboutie. À chaque fois les avantages et les inconvénients seront étudiés. La partie IV décrit la méthodologie initiale adoptée par PartnerRe. Cette méthode, dite méthode des régions, n'est pas optimale car elle utilise des informations d'ordre administratif pour palier ses défauts. La partie V marque l'intégration de l'algorithme des k moyennes pour réaliser une segmentation automatique et objective en groupes homogènes. Elle aboutit à une méthodologie intermédiaire bien plus prometteuse mais ne gère pas encore parfaitement les imprécisions induites par les départements non significatifs. La partie VI expose la méthodologie finale. Elle montre comment appréhender judicieusement ces départements tout en conciliant résultats actuariels et connaissances des experts de manière optimale.

Enfin, ce mémoire s'achève avec la tarification de la réassurance. PartnerRe réassure CéréAssure avec un traité en excédent de perte (cf. I-B-2). La partie VII montre comment la segmentation utilisée pour le produit d'assurance peut aussi être intégrée en conséquence dans l'analyse du réassureur pour optimiser le prix. Celui-ci peut proposer des tarifs sur mesure et par groupe qui incitent l'assureur à optimiser son portefeuille. À la clé, un accord gagnant-gagnant entre les deux acteurs.

I. Contexte et objectifs

A. L'assurance en agriculture

1. Marché de l'assurance agriculture au sens large

Le marché mondial de l'assurance en agriculture représente environ 33 milliards de dollars de prime (selon l'Association Internationale des Assureurs Grêle (AIAG)). Le marché a connu un essor fantastique dans les années 2000 et notamment a doublé en l'espace de 5 ans (2005 – 2009).

L'assurance en agriculture est un champ très vaste qui englobe de nombreuses branches :

- les grandes cultures (céréales et légumineuses principalement),
- les cultures dites « spéciales » dont
  - l'arboriculture,
  - la viticulture,
  - l'horticulture qui inclut les cultures sous serre
- la sylviculture,
- les élevages bovins, porcins, caprins, ovins,
- les autres élevages plus spécifiques tels que l'aviculture, l'aquaculture, l'apiculture…

Ce mémoire va se focaliser sur les grandes cultures que l'on désigne plus simplement sous le terme « récolte » (ou en anglais « crop »). Cette catégorie représente d'ailleurs la plus grosse part du marché, environ 90%.

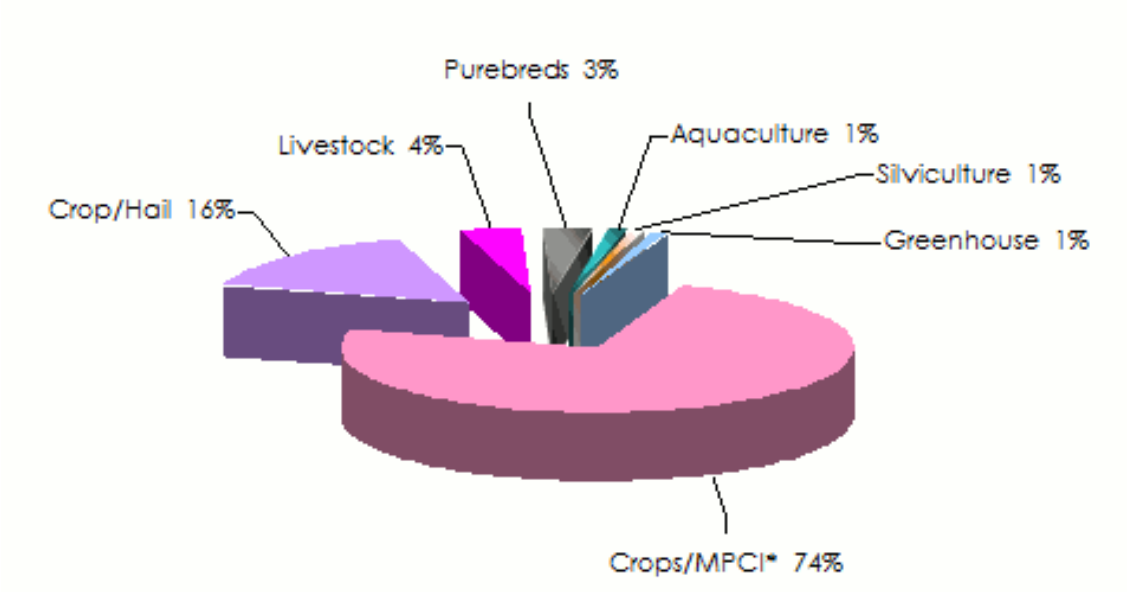

Figure 1 : Répartition des différentes lignes de business en agriculture selon SwissRe en 2014 (1)

2. Assurance récolte : périls et couverture

De nombreux aléas peuvent impacter les récoltes :

- des évènements relativement fréquents, très aléatoires, d'intensité locale, qui peuvent détruire une culture de façon imprévisible, soudaine et irrépressible : typiquement la grêle,
- des phénomènes d'ampleur plus vaste tels que la sécheresse, la surabondance de pluie, le gel ou la tempête, dont l'impact s'étend sur de larges régions et qui nuisent au rendement des cultures généralement de façon plus progressive
- des accidents plus sporadiques de portée moindre comme l'incendie ou la chute d'aéronef.

La couverture grêle

Parmi tous ces périls, seule la grêle représente un risque assurable en l'état. En effet, il s'agit d'un évènement aléatoire et soudain qui concerne tous les agriculteurs sans distinction et contre lequel il est techniquement quasiment impossible de se prémunir : les mesures de prévention sont coûteuses (filets) et pour certaines leur efficacité ne peut être démontrée (fusées antigrêles).

Les assureurs possèdent des statistiques remontant jusqu'à la fin du 19ème siècle (pour les pays les plus développés). Ils proposent donc une couverture grêle basée sur une technique actuarielle et une méthodologie d'expertise éprouvées.

La couverture multi-périls sur récoltes

La couverture des autres périls est en revanche plus complexe. D'une part, l'aléa est parfois prévisible par l'assuré à moyen terme (c'est le cas notamment de la sécheresse et de la surabondance de pluie). D'autre part, le risque dépend de nombreux facteurs tels que le relief et le type de sol qui sont intrinsèques aux parcelles de l'assuré. De plus, l'ampleur géographique des évènements rend la mutualisation des risques plus difficile et le risque d'antisélection est donc important (cf. partie I-3).

Pour tenter de remédier à ces difficultés, les assureurs proposent le plus souvent une couverture dite « multi-périls » qui regroupe tous les aléas (avec ou sans grêle) pour forcer la mutualisation. (En anglais MPCI, pour multi-peril crop insurance). C'est sans conteste la forme d'assurance récolte la plus commune.

En cas de sinistre, l'indemnisation prend la forme suivante :

$$\text{Indemnisation} = \text{valeur assurée} * \text{pourcentage de perte déterminé par expertise} - \text{franchise}$$

La valeur assurée est définie à la signature du contrat entre l'assureur et l'assuré (elle ne dépend pas des marchés financiers). Le pourcentage de perte est estimé par des expertises de terrains, obligatoires et contractuelles, organisées par la société

d'assurance. Ceci représente des coûts de gestion importants mais malheureusement encore incontournables. En effet, la corrélation entre les statistiques climatiques disponibles et les dommages effectifs aux rendements est souvent difficile à établir.

Couvertures indicielles (ou paramétriques)

Dans de plus rares cas, lorsqu'il existe un indicateur objectif corrélé aux dommages, l'assureur peut développer des produits d'assurance dits « paramétriques ». Dans cette couverture, l'indemnisation se déclenche lorsque l'indicateur atteint un certain seuil. L'indice peut être de plusieurs types :

- indice climatique : indice de précipitation, températures sur une période déterminée, etc.;
- indice de biomasse : indice de végétation déterminée sur une image satellite (mesure des rayonnements rouges et infrarouges des végétaux) ;
- indice de rendement : par exemple, rendement moyen régional officiellement annoncé par l'État ou un organisme indépendant.

L'assurance paramétrique est la forme d'assurance agricole la plus moderne mais également la plus vulnérable aux spécificités locales de terrain. Elle s'applique lorsque :

- l'assurance traditionnelle n'est pas économique (frais d'acquisition élevés, statistiques individuelles non disponibles, coûts d'expertise importants) ;
- l'aléa couvert est de large ampleur géographique ;
- les dommages sont relativement homogènes ;
- l'historique est suffisamment long et fiable sur cet indice pour permettre l'établissement d'un tarif.

3. Risque d'antisélection

Chaque agriculteur possède un risque qui lui est propre, selon la localisation de son terrain, la qualité de son sol et surtout selon la zone agroclimatique dans laquelle il se trouve. Malheureusement, l'assureur est souvent contraint de raisonner à grande échelle pour pouvoir appliquer la loi des grands nombres sur le plan actuariel (cf. partie III-A) mais également pour faciliter sa gestion. Il propose des tarifs très peu différenciés, parfois même uniques à l'échelle d'un pays (cf. partie II-D).

Le tarif est donc attractif pour la partie de la population qui se considère très souvent, à juste titre, comme étant la plus exposée. À l'inverse, les populations qui estiment leur risque faible ont tendance à moins s'assurer, voire à ne pas s'assurer du tout. Le portefeuille se retrouve alors déséquilibré avec une exposition aux risques en moyenne plus élevée que celle prévue au départ. C'est ce que l'on appelle le risque d'antisélection (ou de sélection adverse).

Le risque d'antisélection est une situation très redoutée par l'assureur car aucune issue n'est techniquement favorable. Si on décide d'augmenter le tarif pour compenser le déséquilibre, on aggrave alors le phénomène d'antisélection et un

cercle vicieux s'installe. Si au contraire on décide de baisser le tarif, le portefeuille est alors fatalement sous-tarifé : la prime reçue globalement est inférieure au sinistre attendu.

4. Partenariats public-privé

Une solution pour limiter le risque d'antisélection est de collaborer avec les instances gouvernementales.

D'un côté, l'État a tout intérêt à ce que les agriculteurs s'assurent car il veut garantir la stabilité économique et sociale en cas d'évènements catastrophiques de grande ampleur. En effet, l'assurance permet de lisser les résultats des exploitations agricoles. Elle porte une partie du risque et mutualise le reste à l'échelle internationale via la réassurance. En cas d'accident climatique, elle permet à l'agriculteur de réimplanter ses cultures l'année suivante sans s'appauvrir. Au quotidien, l'assurance peut aussi contribuer au développement durable de l'agriculture dans le pays par la mise en place de garanties incitatives prenant en compte les mesures de prévention et des formations de terrain via le « service après-vente » que représentent les expertises sinistres.

De l'autre, l'assureur a besoin d'un portefeuille homogène qui mutualise les risques. La solution pour l'État consiste donc à subventionner les primes d'assurance dans une certaine proportion. Tous les individus, même les moins à risque, se retrouvent alors avec un tarif très attrayant et souscrivent volontiers. Ainsi, le risque porté par le portefeuille s'équilibre : la prime calculée par l'assureur sur la population potentielle ciblée correspond bien à l'exposition de l'échantillon effectivement assuré.

En moyenne, dans le monde, on estime que le niveau de subvention de la prime d'assurance par les États est de 50%.

B. La réassurance en agriculture

Pour supporter ses propres engagements, l'assureur fait souvent appel aux réassureurs pour partager l'intégralité de ses risques (réassurance proportionnelle) ou bien céder ses risques de pointes (réassurance non proportionnelle).

En assurance récolte, la réassurance est quasiment systématique. En effet, les portefeuilles d'assurance agricole sont relativement petits et nationaux. Ils ne constituent pas à eux seuls une assiette de prime suffisante pour encaisser des chocs agroclimatiques rares mais sévères. Même les gros assureurs internationaux font appel à la réassurance pour sécuriser leur bilan.

1. Spécificités de l'agriculture en réassurance

L'agriculture est une spécialité intéressante en réassurance pour plusieurs raisons. Avant tout, elle constitue un fort élément de diversification : la grêle ou la sécheresse saisonnière ne se cumulent pas avec les expositions classiques. Elle est donc assez

décorrélée de l'état mondial de l'économie (contrairement à d'autres lignes de business telles que l'assurance sur les risques financiers, la construction de grands ouvrages ou de grandes usines, les commerces maritimes et aériens…). De même, elle n'est que très peu sensible aux cycles du marché de la réassurance. Enfin, l'intervention des États (cf. partie A-4) permet aux assureurs d'opérer dans un cadre réglementaire connu et limité qui en général garantit un certain contrôle du risque ainsi que la mutualisation des portefeuilles via des volumes importants de primes.

En contrepartie, l'agriculture est une ligne dont les résultats techniques sont très volatils. À titre d'illustration, les taux de pertes varient entre 40% et 150% pour l'assurance grêle et de 30% à 500% pour l'assurance multi-périls. Ceci force le réassureur à bloquer une part importante de capital pour anticiper les chocs. Un autre inconvénient spécifique à la réassurance proportionnelle est la quasi-absence de cash-flow. La majorité des primes sont payées en fin de campagne et le réassureur ne peut donc pas les placer pour les faire fructifier car les sinistres viennent compenser les primes à la fin du contrat.

Enfin, la réassurance en agriculture est impactée par le dérèglement climatique. Celui-ci augmente la fréquence et l'intensité des sécheresses, des fortes pluies et des incendies. La sinistralité va donc indéniablement s'aggraver dans le futur. Les primes d'assurances risquent de monter à un niveau tel que les États devront intervenir de plus en plus (cf partie I-A-4). Cependant, sur le plan actuariel, ceci reste encore très difficile à quantifier faute de méthode et de recul statistique.

2. Les structures de réassurance

Deux structures de réassurance dominent largement en assurance récolte. Pour la réassurance proportionnelle on utilise une structure quote-part tandis que pour la structure non proportionnelle on utilise l'excédent de perte.

Traité quote-part

La réassurance en quote-part (quota share en anglais) est la forme la plus simple de réassurance. Elle consiste à partager de manière proportionnelle à la fois les primes et les sinistres : assureur et réassureur partagent donc le même sort. Généralement le réassureur paye une commission à l'assureur pour participer aux coûts d'acquisition de l'assureur.

La réassurance quote-part est intéressante pour les petites compagnies, souvent jeunes, avec quelques lignes de business voire une seule et par conséquent une diversification très limitée. Ces petites compagnies qui développent un portefeuille et accroissent capital et réserves, n'ont pas encore la maturité nécessaire pour assumer l'intégralité des risques qu'elles souscrivent. Elles partagent entre 10% et 90% de leur portefeuille avec un ou des réassureur(s).

On trouve également des réassurances en quote-part pour les couvertures indicielles (cf. partie A-2) ou plus généralement tous les produits nouveaux et expérimentaux qui n'ont pas encore fait leurs preuves sur le long terme.

Traité excédent de perte (stop loss)

La réassurance en excédent de perte est une forme de réassurance non proportionnelle qui protège l'assureur contre les années catastrophiques. Elle prend en charge tout ou partie des sinistres lorsque l'assuré a déjà subi un certain taux de perte. Son nom anglais « *stop loss* » est d'ailleurs très parlant.

Le seuil à partir duquel la réassurance démarre se nomme *la priorité annuelle agrégée* (en anglais : *AAD* pour *Annual Agregate Deductible*). Elle est toujours accompagnée d'une *limite annuelle agrégée* qui correspond à la couverture maximale que le réassureur est prêt à proposer (en anglais : *AAL* pour *Annual Agregate Limit*). En agriculture, la priorité et la limite ne sont jamais exprimées en valeur absolue. Ils sont définis :

- très souvent, comme un pourcentage de la prime d'assurance effectivement souscrite par l'assureur. Dans ce cas on exprime également les sinistres en pourcentage de la prime. On parle alors de ratio de perte. (Loss ratio en anglais)
- plus rarement, et ce sera le cas dans ce mémoire (cf. partie VII), comme un pourcentage de la somme effectivement assurée. Dans ce cas on exprime les sinistres en pourcentage de la somme assurée et on parle de taux de perte (loss cost en anglais). Les valeurs ne dépassent alors jamais 100% car il n'est pas possible de détruire au-delà de 100% des récoltes.

Le tableau ci-dessous résume avec des exemples le fonctionnement d'un excédent de perte « 80% en excès de 120% »

Ratio de perte de l'assureur avant réassurance	Ratio de perte pris en charge par le réassureur 80% xs 120%	Ratio de perte de l'assureur après réassurance
100%	0% (priorité non atteinte)	100%
160%	40%	120%
200%	80%	120%
220%	80% (limite atteinte)	140%

Tableau 1 : Exemple de fonctionnement d'un stop loss 80% xs 120%

En échange d'une protection non proportionnelle, le réassureur demande une prime de réassurance qui s'exprime elle aussi comme un certain pourcentage de la prime (ou de la somme assurée) souscrite par l'assureur. Ce tarif est calculé en fonction de la volatilité du portefeuille et de la probabilité que le réassureur intervienne. En excédent de perte, le réassureur ne partage donc pas le même sort que son client

l'assureur. La compagnie d'assurance (autrement nommée « la cédante ») cède sa volatilité au réassureur et ne garde que les risques courants (on parle de risque attritionnel). Le réassureur, quant à lui, est exposé aux risques exceptionnels et ses résultats sont très volatils.

Généralement, les gros contrats de réassurance non proportionnelle sont divisés en plusieurs tranches (« layers » en anglais). Par exemple, l'excédent de perte ci-dessus pourrait de manière équivalente être divisé en deux tranches :

- Tranche 1 : 40% xs 120% (couverture entre 120% et 160%)
- Tranche 2 : 40% xs 160% (couverture entre 160% et 200%)

Un réassureur qui couvrirait ces deux tranches à parts égales obtiendrait exactement les mêmes résultats qu'en couvrant une unique tranche 80% xs 120%. Mais en réalité, l'intérêt majeur du système de tranches réside dans le fait que plusieurs réassureurs peuvent prendre des parts différentes dans chaque tranche selon leurs appétits respectifs. Par exemple, un réassureur qui cherche simplement à diversifier son portefeuille en limitant les risques prendra une part plus importante de la tranche 1 tandis que celui qui est à la recherche de business risqué mais plus profitable privilégiera la tranche 2.

Pour l'assureur, multiplier les tranches demande certes un plus gros travail de gestion mais en contrepartie il augmente ses chances de trouver des réassureurs en proposant plus de flexibilité. Ceci lui offre une plus grande marge de négociation et un plus grand réseau. De plus, l'assureur peut lui aussi décider de garder des parts plus ou moins grandes dans les différentes tranches. (La part conservée par l'assureur est appelée « la rétention ».)

Ce mémoire va maintenant développer principalement la tarification d'un produit d'assurance. Néanmoins, la tarification de la réassurance sera abordée en dernière partie (partie VII)

C. Produit d'assurance

En 2013, CéréAssure a développé avec l'aide de PartnerRe un nouveau produit d'assurance MPCI (Multi-Peril crop insurance) pour les agriculteurs du CéréLand. Cette couverture protège bien sûr l'assuré contre tous les types de périls envisageables : pluies excessives, gel, inondations, incendies, etc... Et notamment la sécheresse qui constitue le principal danger au CéréLand car une majorité des agriculteurs comptent uniquement sur l'eau de pluie pour irriguer leur culture.

Plus précisément, il s'agit ici en réalité d'une variante de la couverture classique car le produit étudié couvre la volatilité sur les rendements. Tant que l'agriculteur obtient un rendement proche du rendement de référence, il ne touche aucune indemnité. En revanche, si le rendement est inférieur à 75% du rendement de référence, l'agriculteur touche une indemnité en suivant une loi affine. À rendement nul, il touche 100% de la somme assurée. L'indemnité peut se traduire par l'équation suivante :

$$\text{Indemnités} = \left[\text{Somme assurée} * \left(1 - \frac{\text{rendement}}{\text{rendement ref} * 75\%}\right)\right]_+$$

Et graphiquement par la Figure 2:

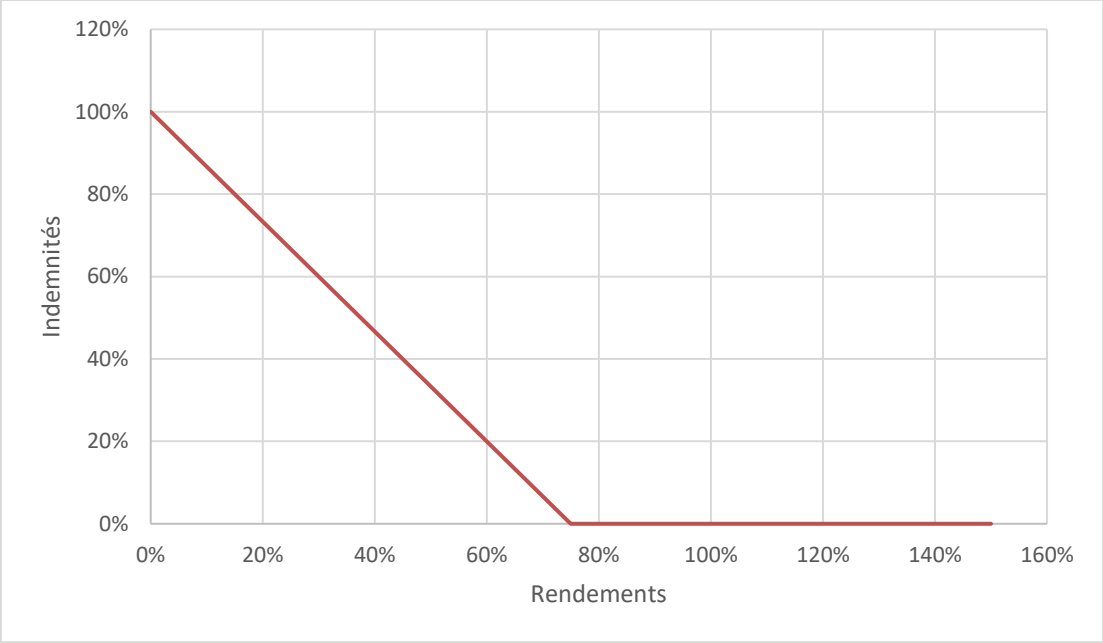

Figure 2 : Indemnités octroyées par le produit CéréAssure

On définit le rendement de l'assuré comme étant le poids moyen de la récolte par hectare cultivé. Si, sous certains aspects, la couverture peut ressembler à une couverture indicielle, il n'en est rien. Le rendement de référence est en fait défini chaque année et dans chaque commune par le ministère de l'agriculture sur la base de la moyenne des rendements communaux des 10 dernières années. Il s'agit d'un paramètre fixe (absolument nécessaire pour définir les indemnités). Tandis que les résultats des agriculteurs sont bel et bien la variable qui déclenchent ou non l'indemnisation.

PartnerRe en tant que réassureur a une double mission auprès de sa cédante. D'une part, évidemment prendre en charge les risques pour lesquels son client n'a pas d'appétence. Mais surtout, en tant qu'expert du risque, PartnerRe doit conseiller l'assureur lors de la tarification directe du produit. C'est en effet le réassureur qui, fort de son expérience internationale et de ses nombreuses relations avec différentes cédantes, est le plus à même d'évaluer le risque et le prix qui en découle.

D. Histoire du produit

À la création du produit en 2013, CéréAssure et PartnerRe n'avaient pas de données solides et concrètes pour établir la grille tarifaire du produit. À défaut, les informations du ministère de l'agriculture ont donc été utilisées. Le ministère avait identifié trois zones dans le pays concernant la culture de céréales irriguées à l'eau de pluie :

- la zone 1, défavorable, assez sèche avec des pluies aléatoires parfois absentes. (120€/Ha) ;
- la zone 2, intermédiaire (180€/Ha) ;
- la zone 3, favorable avec des pluies assez régulières et certaines (3.1 : 290€/Ha, 3.2 : 580€/Ha, 3.3 : 870€/Ha).

Faute de données, CéréAssure et PartnerRe avaient, à cette époque, opté pour un tarif unique mais en imposant des limites sur les sommes assurées. Ainsi, en zone 1, un assuré ne pouvait assurer que 120 euros par hectare et 180 euros en zone 2 tandis que les assurés de la zone 3 avaient accès à 3 différents niveaux de garantie 3.1, 3.2, 3.3.

Aujourd'hui, 8 ans plus tard, le produit a permis de récolter suffisamment de données historiques pour être exploitées. Indéniablement, certains départements apparaissent plus profitables que d'autres et l'on constate que la segmentation en zones du ministère n'est pas représentative de la réalité du portefeuille.

Le tarif unique est donc sans aucun doute source d'antisélection. Les agriculteurs des départements au climat défavorable perçoivent le prix unique proposé comme un tarif très intéressant et sont nombreux à souscrire. Inversement les agriculteurs des zones favorables trouvent le produit cher et non indispensable. Malgré la subvention du ministère, ils sont moins nombreux à souscrire. En conséquence, le portefeuille de la cédante risque de se retrouver déséquilibré.

Ce constat n'a rien d'étonnant car les risques liés aux excès ou manque d'eau sont sujets à une sélection adverse chronique. En effet, il est rare qu'à l'échelle d'un pays, les populations soient exposées à ces périls de manière homogène.

Il paraît alors naturel d'exploiter les 8 années de données pour mettre en place une nouvelle grille de tarification qui limiterait cet effet d'antisélection. Ceci ouvre la porte à de nombreuses questions. Comment peut-on établir des zones au comportement similaire ? Comment exploiter les données géographiques au mieux pour proposer une solution cohérente et simple à l'échelle nationale ? Quel tarif doit-on leur attribuer ?

L'objet de cette étude est de répondre à toutes ces questions et de proposer *in fine* une méthodologie pour segmenter un territoire en zones, qui soit à la fois cohérente sur le plan actuariel mais également sur le plan géographique.

II. Analyse et traitement préalable des données

A. Présentation des données

Le fichier principal utilisé dans cette étude est le fichier recensant toutes les polices souscrites entre 2013 et 2020 avec les montants indemnisés le cas échéant. Il possède une quinzaine de champs :

- le numéro de police,
- l'année de souscription,
- la région,
- le département,
- la commune,Nature
- l'identifiant de la commune,
- le type d'irrigation,
- le type de culture,
- la superficie,
- la zone de souscription (1,2 ou 3, comme définie par le ministère en 2013, cf. partie I-D),
- le niveau de garantie accordé : 120, 180, 290, 580, 870 €/Ha (cf. partie I-D),
- le capital assuré qui correspond au niveau de garantie multiplié par la superficie,
- le montant indemnisé le cas échéant.

Le fichier possède 842 000 lignes, une ligne par couple polices/années de souscription. Grâce à ce fichier à plat (flat file en anglais) et grâce aux différents champs indiqués ci-dessus, nous pouvons facilement générer des données à des échelles différentes : la commune, le département, la région et le pays tout entier.

Le portefeuille se compose à 90% de cultures de céréales et le reste de légumineuses. 95% de celles-ci ne sont pas irriguées mais arrosées naturellement par l'eau de pluie. Dans ce rapport, nous nous focaliserons seulement sur cette partie, la plus représentative du portefeuille.

B. Vérification sur les données

Avant de commencer l'étude actuarielle, il faut procéder à plusieurs traitements et vérifications. On commence par un nettoyage des données car certains champs, notamment les noms de communes et de départements, comportent parfois des fautes de frappe. Il convient d'homogénéiser tout cela.

On analyse également la cohérence des différents champs entre eux. Par exemple, on vérifie que les niveaux de garanties enregistrés correspondent bien aux zones de souscription, que le capital assuré est correctement calculé, que chaque commune est assignée à un seul département, etc… En résumé, on essaye de vérifier que

toutes les hypothèses implicites et évidentes que l'on imagine *a priori* sont effectivement valides. Ce travail préliminaire fastidieux est indispensable car à défaut on travaillerait sur des hypothèses erronées et on s'exposerait à des erreurs invisibles parfois non négligeables.

On procède ensuite à différentes analyses rapides pour dégrossir le sujet. Par exemple, on examine l'évolution des volumes sur la période étudiée. On observe ainsi des volumes de primes, de superficies et de capitaux plutôt stables à l'échelle nationale. On analyse également l'évolution des cultures au cours du temps et on vérifie que les proportions sont stables elles aussi : environ 65% de blé, 30% d'avoine et 5 % d'orge chaque année.

Enfin et surtout, en matière de résultats économiques, on vérifie que l'historique dont on dispose est suffisamment long et représentatif. Par chance, parmi les 8 années disponibles deux d'entre elles ont été extrêmement désastreuses et deux d'entre elles extrêmement favorables. Ce qui, au dire des experts locaux, est assez caractéristique des résultats d'assurance dans le pays.

C. Transformation en données « *As If* »

Dans la vie d'un produit d'assurance, plusieurs modifications peuvent intervenir sur les prix, sur la politique de souscription, voire sur les paramètres du produit lui-même. Les données étudiées ont la particularité de porter en elle l'histoire de ces modifications de contrat. Pour comparer ce qui est comparable et également pour étudier de potentielles modifications en 2021, on procède à la transformation des données en données « *As If* » (« comme si » en français). Il s'agit de simuler ce que seraient les données si elles subissaient les conditions de 2021.

Par exemple, on peut vouloir changer les niveaux de garanties accordés en fonction des zones. Le changement de ces niveaux va avoir un impact sur les sommes assurées dans chaque zone et donc, sur le résultat final. Pour étudier cet aspect, il est plus facile de modifier directement la colonne « niveau de garantie » du fichier en amont de l'analyse plutôt que de mener des raisonnements mathématiques en aval.

Dans certains cas, il est même impossible de corriger le résultat en aval de l'analyse. Par exemple, lors de cette étude, à la demande du client, nous avons étudié l'impact d'un changement de seuil pour le rendement de référence. Plus précisément, il fallait répondre à la question : « Que se passerait-il si le 75% de la partie I-C était remplacé par 65% ? ». Pour y répondre, il n'y a pas d'autres moyens que de recalculer directement les montants indemnisés dans le fichier initial et de relancer la même analyse ensuite.

L'utilisation d'un code (en langage R dans le cas présent) est donc fortement recommandée pour mener confortablement ce genre d'analyse sophistiquée. Les modifications étudiées, tels que les niveaux de garantie ou le seuil de référence, peuvent être définies comme des variables d'entrée du code que l'on peut modifier une fois l'analyse totalement scriptée. (Voir Annexe).

D. Conclusion sur les données

Le fichier possède de nombreux atouts. Les différentes vérifications menées en partie B ont permis de vérifier la bonne qualité des données : on relève très peu de lignes dont les champs se contredisent et, le cas échéant, elles sont facilement corrigeables. Sur le plan statistique les données mettent en évidence l'évolution d'un portefeuille stable (quant à la composition et au volume) qui a traversé à la fois de bonnes et de mauvaises années. La méthode du *burning cost* (voir partie III) est donc applicable.

Le fait de posséder différents niveaux administratifs est également un point très positif qui nous laisse une grande marge de manœuvre dans la façon de construire l'étude. De plus, le nombre important de lignes permet d'obtenir des statistiques significatives même à l'échelle de certaines communes. Cependant l'activité de souscription de la cédante est répartie de manière très hétérogène et des départements pourtant grands se retrouvent parfois sous-représentés. Dans la suite du mémoire, il y aura donc une réflexion à mener sur le niveau de finesse de l'analyse et également de grandes interrogations sur la façon de tarifer ces zones non significatives.

Si le fichier est grand en nombre de lignes, il ne l'est pas, en revanche, en nombre de champs. Les données sont statistiquement fiables mais on sait déjà que le modèle proposé sera limité s'agissant du nombre de paramètres pris en compte. Ceci nous dirige une fois de plus vers la méthode du *burning cost* qui a le mérite d'être simple et robuste.

III. Tarification : méthode générale et enjeux

Fort de ces 8 années d'expérience et de ses données importantes, PartnerRe et sa cédante souhaitent tous deux établir une tarification qui reflète au mieux la réalité du portefeuille. La méthode du *burning cost* sur les taux de perte est alors tout indiquée. Cependant, en théorie très simple à appliquer, cette méthode se heurte en pratique à de nombreuses difficultés.

Cette partie rappelle brièvement la méthode théorique du *burning cost* sur les taux de perte avant d'exposer les différents enjeux de l'étude sur le plan pratique.

A. Méthode théorique du *burning cost* sur les taux de pertes

La tarification d'un produit d'assurance consiste à établir la prime technique que l'assuré devra verser à l'assureur pour obtenir la couverture d'un certain capital assuré contre certains risques bien définis dans le contrat. Cette prime technique se décompose en une prime pure et des chargements.

La prime pure est la partie de la prime reflétant le risque sous-jacent. L'assuré doit payer 100 euros de prime pure si, en espérance, l'assureur s'attend à l'indemniser de 100 euros de sinistres.

A cette prime pure, l'assureur ajoutera les différents chargements qui servent :

- à payer les frais de gestion (dont la rémunération des agents),
- à protéger l'assureur contre la volatilité des sinistres et l'imprécision sur la tarification,
- à payer les taxes etc…,
- éventuellement à générer du profit selon la politique et le statut de l'acteur qui assure (assurance, mutuelle, État…).

Le calcul des chargements est presque toujours géré par l'assureur. Le calcul de la prime pure en revanche peut être réalisé par le réassureur qui apporte son expertise du risque.

Plus précisément, nous parlerons ici du taux pur (ou taux de prime pure) qui permet d'exprimer cette prime pure en pourcentage du capital assuré :

$$\text{taux pur} = \frac{\text{prime pure}}{\text{capital assuré}}$$

Comme expliqué plus haut, la prime pure doit être égale au montant de sinistre attendu. Ainsi :

$$\text{taux pur} = \frac{\text{prime pure}}{\text{capital assuré}} = \frac{\text{sinistres attendus}}{\text{capital assuré}}$$

D’autre part, on définit le taux de perte comme étant le rapport entre les sinistres et le capital assuré historique :

$$\text{taux de perte} = \frac{\text{sinistres}}{\text{capital assuré}}$$

Par exemple, lorsqu’un agriculteur perd la moitié de ses récoltes à cause d’une mauvaise météo, son taux de perte est de 50%.

Par conséquent, le taux pur est en fait égal au taux de perte attendu :

$$\text{taux pur} = \frac{\text{sinistres attendus}}{\text{capital assuré}} = \text{taux de perte attendu}$$

Estimer le taux de perte attendu est l’objet central de ce rapport. Lorsque l’histoire d’un portefeuille est suffisamment longue, une manière simple de l’estimer est de calculer le taux de perte historique qui est un bon indicateur du taux de perte attendu. En effet, par la loi des grands nombres, plus le nombre d’années est élevé, plus le taux de perte historique converge vers le taux de perte attendu.

$$\text{taux de perte historique} = \frac{\sum_{\text{années}} \text{sinistres}}{\sum_{\text{années}} \text{capital assuré}}$$

$$\text{taux de perte attendu} = \lim_{\text{années} \to \infty} \text{taux de perte historique}$$

La méthode du *burning cost* consiste simplement à décréter que le taux de perte attendu est égal au taux de perte historique.

Dans la suite du rapport, on emploiera souvent le terme « taux de perte » pour désigner indifféremment le taux de perte historique ou le taux de perte attendu car ceux-ci sont égaux.

B. Limites de la théorie

Bien sûr, tout ceci suppose que chaque année suive la même loi de probabilité et donc, qu’aucune tendance n’existe : le climat n’évolue pas au cours du temps, le comportement des agriculteurs non plus (sélection, antisélection, aléa moral, fraude…), ni même le comportement de l’assureur (marketing, contrôle des sinistres). Nous choisirons cette hypothèse de travail et l’étude de ces tendances ne fait pas l’objet du mémoire.

En assurance agricole, il est également très important de vérifier que la proportion de chaque culture (céréales, légumineuses, fruits…) reste constante au cours du temps. Si tel n’est pas le cas, un calcul de taux de perte culture par culture est bien plus adapté voire obligatoire. Par chance, dans le cas pratique exposé ici, le portefeuille est très majoritairement composé de blé et d’orge aux comportements similaires (cf. partie II-B).

En revanche, les considérations géographiques posent un sérieux problème. En effet, le CéréLand est un pays aux climats variés et un rapide calcul de taux de pertes historiques par région et par département montre à quels points les résultats peuvent varier en fonction de la localisation.

	2013	2014	2015	2016	2017	2018	2019	2020	Total
Région A	2%	18%	1%	78%	0%	0%	56%	85%	**29%**
Région B	0%	43%	0%	80%	0%	0%	39%	87%	**38%**
Région C	24%	61%	1%	70%	0%	22%	35%	94%	**34%**
Région D	0%	0%	1%	67%	0%	0%	0%	63%	**23%**
Région E	7%	58%	0%	58%	6%	15%	71%	100%	**26%**
Région F	16%	46%	0%	73%	11%	0%	76%	88%	**47%**
Région G	11%	47%	30%	83%	5%	31%	73%	54%	**40%**
Région H	4%	8%	3%	75%	0%	0%	1%	48%	**21%**
Région I	30%	80%	5%	87%	0%	15%	75%	100%	**42%**
Région J	5%	1%	0%	11%	0%	0%	1%	0%	**3%**
Pays	**8%**	**35%**	**8%**	**76%**	**4%**	**4%**	**43%**	**76%**	**35%**

Figure 3 : Taux de perte historique (2013-2020) du CéréLand par région administrative

On pourrait être tenté de simplement calculer et appliquer des taux purs par région ou par département mais trois problèmes majeurs se posent alors :

- le choix de la granularité ;
- la nécessité d'une solution commerciale simple ;
- la nécessité d'une solution géographiquement cohérente.

C. Enjeux

1. Le choix de la granularité

Les données dont nous disposons nous permettent d'analyser le portefeuille à différentes échelles : nationale, régionale, départementale et communale (cf. partie II).

Une échelle trop grande ne permet pas de refléter l'hétérogénéité réelle du portefeuille. Se limiter au taux de perte national, c'est occulter complètement les disparités géographiques qui existent bel et bien et le risque d'antisélection qui en découle. (cf. partie I-A-3)

A l'inverse, utiliser une échelle trop petite nous expose à des résultats volatils et donc instables. Comme expliqué plus haut, la théorie du burning cost est basée sur la loi des grands nombres. Cela nécessite de nombreuses polices et de nombreux sinistres. Dans le cas précis de CéréAssure, l'échelle communale ne répond pas à

ces critères car il existe plus d'une centaine de communes dans lesquelles l'information est basée sur deux ou trois polices seulement comme le montre les histogrammes ci-dessous.

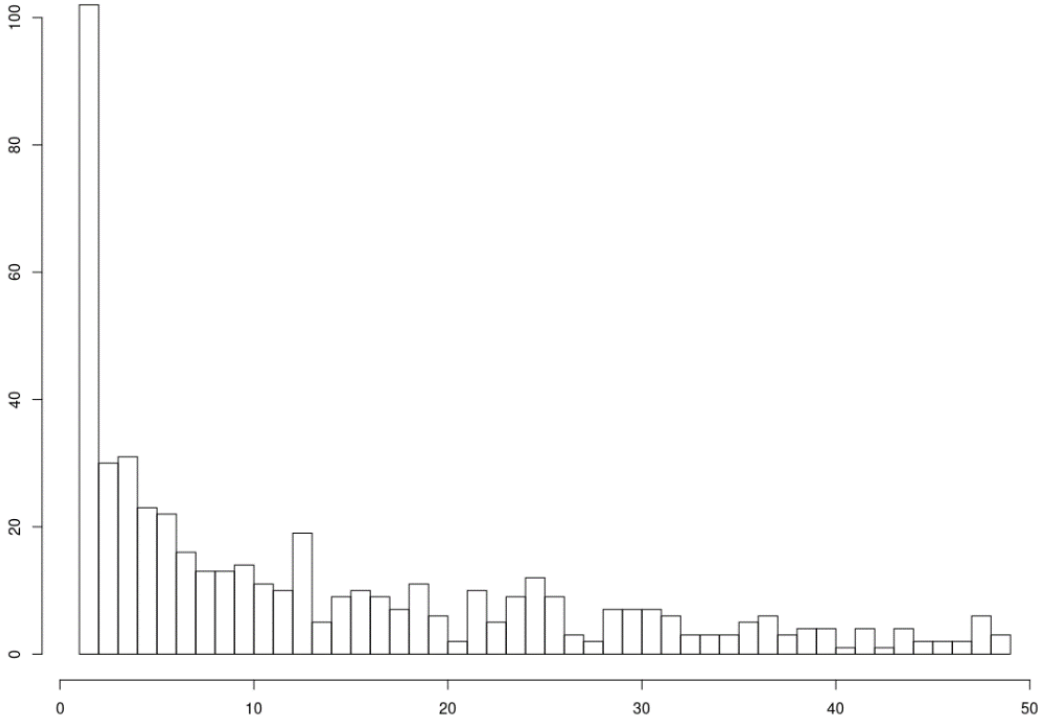

Figure 4 : Histogramme du nombre de communes en fonction du nombre de polices. Communes possédant de 1 à 50 polices.

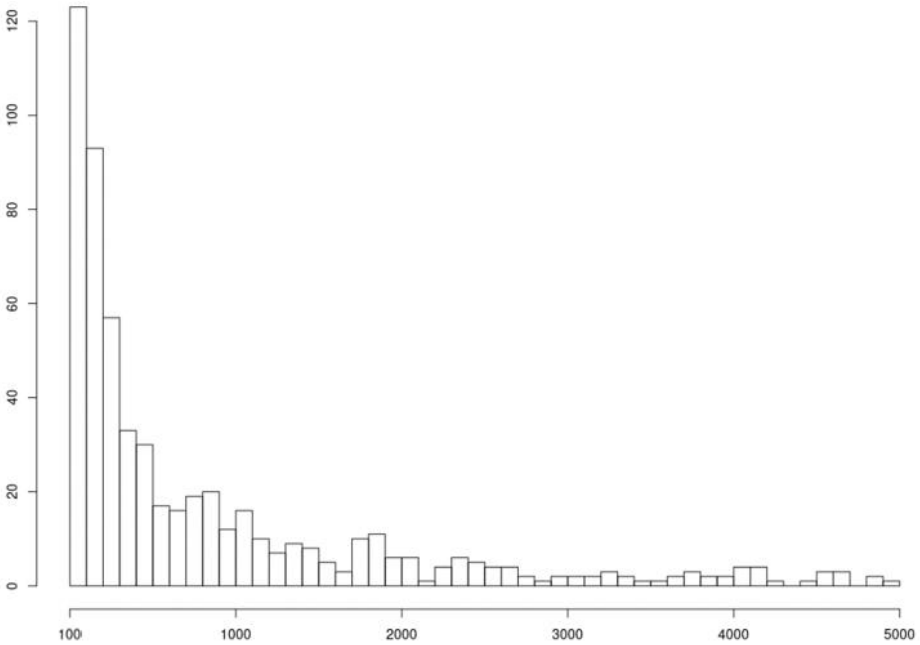

Figure 5 : Histogramme du nombre de communes en fonction du nombre de polices. Communes possédant de 100 à 5000 polices (tranches de 100 polices)

Avec 17 régions, l'échelle régionale semble un peu trop grossière pour modéliser les différents « climats ». Il faudra donc se focaliser sur l'échelle départementale bien que celle-ci ait aussi un défaut important : son hétérogénéité.

À cette échelle, le portefeuille de CéréAssure est loin d'être uniformément réparti. Ainsi, à l'instar de certaines communes, certains départements ne possèdent pas de données suffisamment significatives pour établir un taux de perte fiable. Il faudra néanmoins trouver une solution pour traiter ces départements.

2. La nécessité d'une solution commerciale simple

La tarification doit reposer bien évidement sur la technique et sur les données mais elle doit aussi prendre en compte les contraintes d'ordre commercial. La méthodologie employée ne peut donc aboutir à une multitude de tarifs. À des fins commerciales, on ne peut proposer que 6 ou 7 tarifs tout au plus ou alors un système de majoration et de minoration avec les règles les plus simples possible.

Cette contrainte est encore plus importante si les agriculteurs bénéficient d'une forte subvention de l'État car le ministère de l'agriculture est alors impliqué dans les négociations. Selon sa politique, le ministère pourra éventuellement défendre une égalité maximale entre les assurés.

3. La nécessité d'une solution géographiquement cohérente

On s'attend à ce que les résultats d'un département soient fortement liés à sa localisation et au climat qui y règne mais d'autres facteurs peuvent influencer le rendement. Notamment des facteurs socioéconomiques et culturels très locaux. Par exemple, les types de culture (orge, blé, etc...), les pratiques de la région (mécanisation ou non), le niveau de professionnalisation (activité principale, job secondaire ou production en masse).

Il est donc possible que la réalité actuarielle (les résultats historiques de l'assureur) ne coïncide pas parfaitement à la géographie du pays ou à ce que l'on connaît de son climat. Cependant, à la fin, toujours dans un but commercial, la tarification proposée devra proposer une certaine cohérence sur le plan géographique.

La suite de ce rapport va présenter trois méthodologies (de la plus simple à la plus sophistiquée) qui essayent de répondre au mieux à tous ces enjeux.

IV. Méthodologie initiale : La méthode des régions

Cette première partie expose la méthode initialement utilisée par PartnerRe les années précédant cette étude. Elle permettait à PartnerRe de vérifier que le tarif unique proposé par CéréAssure était raisonnable. Notamment, elle servait à vérifier quels départements déviaient beaucoup de la moyenne nationale dans le but d'établir des limites de souscriptions dans le contrat de réassurance.

A. Description

Cette méthode, dite « méthode des régions » est la solution la plus simple et la plus rapide à réaliser. Elle consiste à prendre le taux de perte historique à l'échelle du département lorsque celui-ci est suffisamment significatif et, à défaut, à lui appliquer le taux de perte historique de la région administrative auquel il appartient. Ainsi chaque département se voit attribuer un taux de perte attendu qui peut être agrégé à l'échelle nationale. La manière de définir les seuils de significativité ainsi que la façon d'agréger les résultats sont détaillées dans les deux sous-parties suivantes.

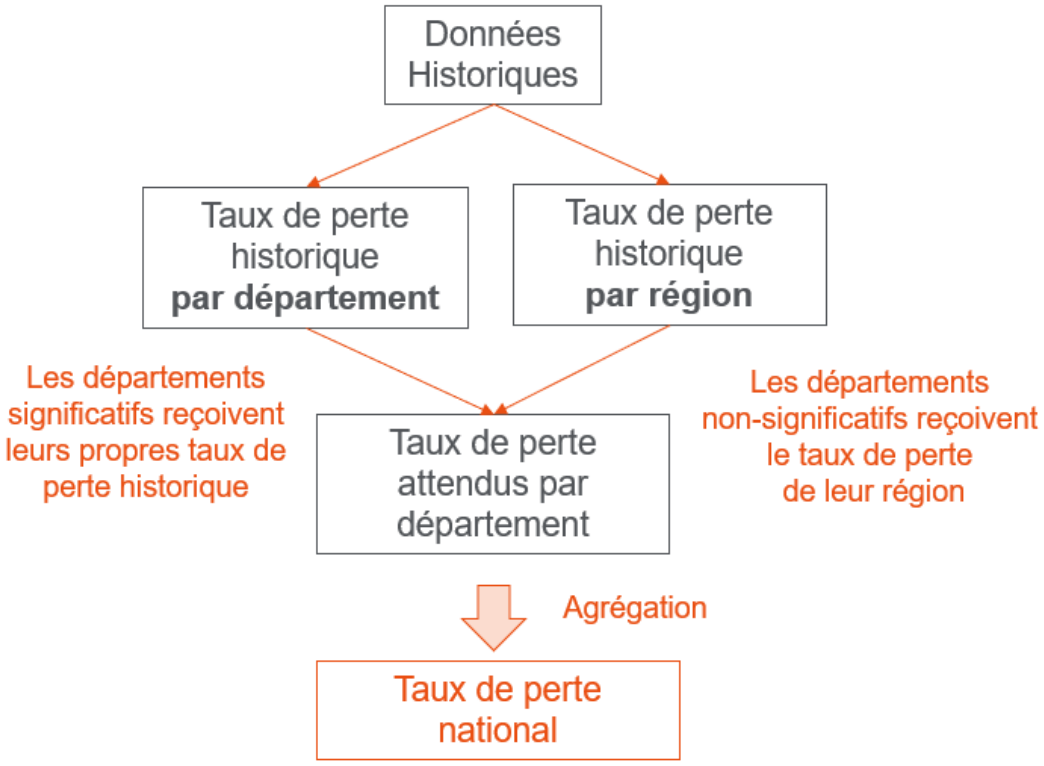

Figure 6 : Méthode des régions

1. Établir des seuils de significativité par département

La méthode des régions permet de ne pas apporter de crédit aux départements dont les données sont insuffisantes et pour lesquels la méthode du *burning cost* serait faussée. Mais pour ce faire, un critère précis doit être utilisé pour distinguer le significatif du non significatif.

Superficie assurée, nombre d'années dans les données, capitaux assurés sont autant de données sur lesquelles des seuils de significativité peuvent être définis. Établir ces seuils est une opération délicate et toujours emprunte de subjectivité. Elle se base à la fois sur l'expérience de l'assureur, l'expérience du marché, l'expérience de l'actuaire etc.. Néanmoins, des analyses de données factuelles peuvent guider la réflexion.

Par exemple, dans le cas de la superficie, nous pouvons étudier les coefficients de variations de chaque département. Le coefficient de variation (ou CV) d'une série de données est le rapport entre l'écart type et la moyenne.

$$\mathrm{CV} = \frac{\text{écart type}}{\text{moyenne}}$$

Le CV est une métrique très utile en assurance et réassurance car, étant normalisé par la moyenne, elle est sans unité et permet d'universaliser le concept de volatilité. Grâce au CV on peut comparer différentes affaires entre elles par exemple. Un CV haut indique une grande volatilité dans les résultats observés. Cette volatilité peut être constatée quand la loi de distribution à l'origine de l'échantillon est effectivement volatile mais elle peut aussi être observée lorsque la quantité de données n'est pas suffisamment significative et que la loi des grands nombres ne s'applique pas. Ce dernier phénomène est facilement observable en traçant le CV des taux de perte historiques de chaque département en regard de la superficie. Les CV sont globalement plus hauts lorsque la superficie assurée est petite.

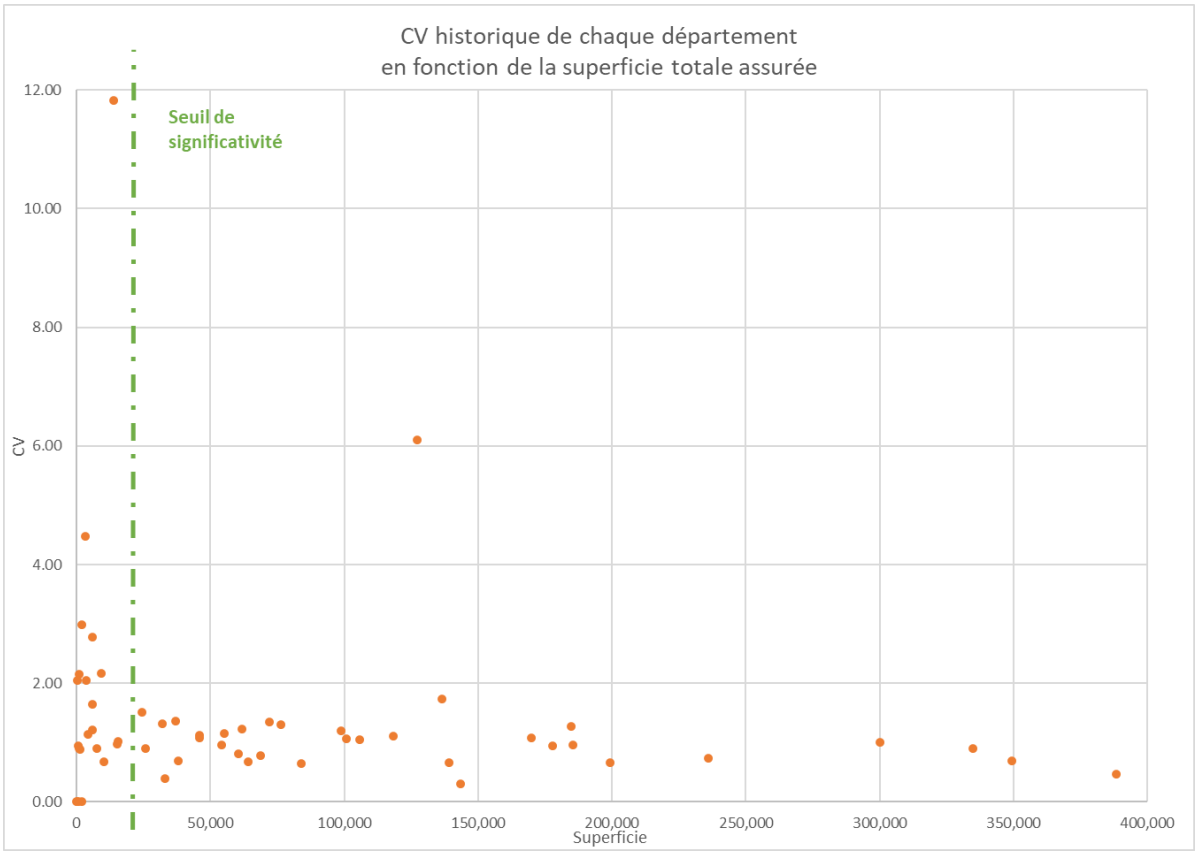

Figure 7 : Établir des seuils de significativité.
Exemple avec la superficie d'un département

Le seuil de significativité (ici en vert) est alors choisi là où la volatilité semble drastiquement s'atténuer. Il convient de choisir un seuil suffisamment grand pour éviter un maximum de volatilité superflue. Inversement, il convient de garder un seuil suffisamment bas pour préserver la majorité de l'information. Dans cet exemple nous fixons le seuil à 20 000 hectares. Ce qui n'exclut en réalité que 0.4% de la superficie totale (somme des superficies à gauche du seuil divisée par la superficie totale).

De manière similaire, on peut étudier la volatilité en fonction des capitaux assurés et fixer un seuil à 10 millions d'euros assurés. Et, s'agissant du nombre d'années, afin de réaliser un traitement homogène, on décide que l'intégralité des années est nécessaire pour que le département soit qualifié de significatif.

En résumé, les départements significatifs ont donc été définis comme étant les départements :

- souscrits entre 2013 et 2020 sans interruptions,
- dont la superficie assurée totale cumulée sur 8 ans dépasse les 20 000 hectares,
- dont les capitaux assurés cumulés sur 8 ans dépassent les 10 millions d'euros.

2. Agréger les départements

La méthode pour agréger les résultats départementaux en un résultat national est courante. Il s'agit tout simplement de moyenner les taux de pertes départementaux (TP_{dep}) en les pondérant par les capitaux assurés dans chaque département (CA_{dep}).

$$\mathrm{TP}_{\mathrm{national}} = \frac{\sum_{\mathrm{dep}} \mathrm{TP}_{\mathrm{dep}} * \mathrm{CA}_{\mathrm{dep}}}{\sum_{\mathrm{dep}} \mathrm{CA}_{\mathrm{dep}}}$$

On peut remarquer à ce stade que, dénuée de tout autre traitement, la méthode expliquée jusqu'ici n'a absolument aucun intérêt. En effet, pourquoi ne pas calculer le taux de perte national directement en sommant tous les sinistres et tous les capitaux assurés ? On obtiendrait à peu près le même résultat et ce, sans se soucier de distinguer les départements significatifs et non significatifs.

En réalité, le fait de disposer des taux de pertes à une échelle inférieure permet de réaliser des scénarios dit « *As If* » (« comme si » en français) sur une grande diversité de paramètres.

L'exemple le plus évident et aussi le plus important ici est la prise en compte de la répartition actuelle des capitaux assurés. Si on cherche à établir le taux de perte national en 2022 et qu'on dispose des capitaux assurés par départements en 2021, le calcul suivant est sûrement beaucoup plus précis :

$$\mathrm{TP}_{\mathrm{national},\ as\ if\ 2021} = \frac{\sum_{\mathrm{dep}} \mathrm{TP}_{\mathrm{dep}} * \mathrm{CA}_{\mathrm{dep},2021}}{\sum_{\mathrm{dep}} \mathrm{CA}_{\mathrm{dep},2021}}$$

De la même manière, on peut profiter d'autres traitements *as if* réalisés à l'échelle inférieure. Par exemple, nous avons vu en partie I-D que les garanties accordées différaient selon les zones. Si on voulait étudier l'impact d'un changement de garantie par zone, on obtiendrait des taux de pertes différents pour chaque département car ceux-ci ne couvrent pas les mêmes zones dans les mêmes proportions. L'agrégation des taux de pertes par département est alors, une fois encore, plus précis ici.

Enfin, l'analyse par département permet d'identifier les départements à risque et éventuellement de réfléchir à l'ajout de limites de souscription sur ces départements en particulier. On peut simuler facilement l'impact de ces limites sur le résultat final en modifiant la formule comme suit :

$$\mathrm{TP}_{\mathrm{national},\ \text{as if 2021,avec limites}} = \frac{\sum_{\mathrm{dep}} \mathrm{TP}_{\mathrm{dep}} * \mathrm{Min}(\mathrm{CA}_{\mathrm{dep},2021}, \mathrm{Limite}_{\mathrm{dep}})}{\sum_{\mathrm{dep}} \mathrm{Min}(\mathrm{CA}_{\mathrm{dep},2021}, \mathrm{Limite}_{\mathrm{dep}})}$$

B. Avantages et inconvénients

La méthode des régions utilisée jusqu'ici chez PartnerRe a le mérite d'être simple et les calculs sont assez faciles avec un tableur et des tableaux croisés dynamiques judicieux. Le fichier de données, à plat, police par police, fourni par la cédante avec ses champs (départements, régions, sommes assurées, sinistres, etc…) se prête particulièrement bien à ces calculs (cf partie II-A).

Elle permet d'attribuer un taux de perte à chaque département, donnant ainsi une bonne idée des bons et des mauvais, et de choisir pour l'avenir, lesquels développer le plus dans le portefeuille et lesquels limiter, réformer ou surveiller.

Cependant, la méthode des régions souffre d'une « approximation administrative » très grossière pour les départements non significatifs. Le fait d'utiliser le résultat de la région administrative n'est évidemment pas satisfaisant car les frontières administratives n'ont rien à voir avec la réalité agroclimatique. Un petit département sur le littoral pourrait se voir attribuer le même taux de perte qu'un autre éloigné de plusieurs centaines de kilomètres pour des raisons administratives uniquement et ceci n'a pas de sens.

De plus, la méthode des régions n'instaure aucune frontière particulière et ne permet pas de construire des groupes de tarifications à elle seule. Sans plus de traitement, la solution commerciale reste le tarif unique, solution simple mais peu réaliste lorsque les résultats sont très hétérogènes comme dans le cas du Céréland. Limiter l'antisélection n'est possible que par l'instauration de limites sur les sommes assurées.

La méthode des régions en résumé :

- Définition des départements significatifs
- Taux de perte départemental pour les départements significatifs
- Taux de perte régional pour les départements non significatifs
- Agrégation as if au niveau national

Avantages :

+ Calculs simples
+ Analyse au niveau départemental (Instauration de limites possible)

Inconvénients :

– « Approximation administrative » loin de la réalité agroclimatique
– Tarif national unique

Pour créer des tarifs différents, il faudrait avant tout créer des groupes de départements différents. Pour ce faire, nous pourrions classer les départements par taux de perte croissant et déterminer arbitrairement des frontières entre ces taux de pertes. Cependant des milliers de segmentations sont possibles et cette méthode serait très subjective, discutable et conflictuelle. N'oublions pas que sur le plan commercial, il faudra savoir justifier cette segmentation et expliquer pourquoi tel assuré aura un tarif plus élevé que tel autre dans le département voisin. Il faut donc proposer une méthode suffisamment objective et rationnelle pour établir ces groupes de départements. C'est l'objet de la méthode suivante.

V. Méthodologie intermédiaire : Tarification par clustering

Le terme clustering est un anglicisme qui pourrait se traduire par regroupement. Il est très utilisé en statistiques et en intelligence artificielle pour l'apprentissage non supervisé et n'a pas réellement d'équivalent français.

Le clustering consiste à classer chaque observation/échantillon en un certain nombre de groupes et selon leur similarité. L'algorithme réalisant le clustering ne possède aucune information préalable sur la nature des groupes à constituer. (C'est ce qu'on appelle l'apprentissage non supervisé). Dans cette étude, nous utilisons l'algorithme des k moyennes pour créer des groupes de départements similaires.

A. L'algorithme des k moyennes

Cette partie formalise le problème des k moyennes avant d'exposer l'algorithme le plus simple (et le plus célèbre) pour la résoudre.

1. Le problème des k moyennes

Le problème des k moyennes consiste à regrouper des échantillons en un nombre prédéfini de groupes k aux caractéristiques les plus similaires possibles. Le nombre de caractéristiques étudiées dépend du problème considéré, il définit la dimension d du problème.

En termes plus mathématiques, il s'agit de créer des groupes tels que la somme des variances dans chaque groupe soit la plus petite possible. On peut le formuler comme suit.

Etant donné un entier k et étant donné X, un ensemble de n échantillons à d variables

$$X = \{x_1, \dots, x_n\} \text{ dans } \mathbb{R}^d$$

Trouver une partition $G = \{G_1, \dots, G_k\}$ de X qui minimise la fonction :

$$\delta(G) = \sum_{i=1}^{k} \sum_{x_j \in G_i} \|x_j - c_i\|^2$$

Où les c_i représentent les barycentres de chaque groupe :

$$c_i = \sum_{x_j \in G_i} \frac{x_j}{|G_i|}$$

Dans le cas précis de cette étude, nous nous placerons en dimension 1. X représente l'ensemble des taux de pertes pour chaque département. L'entier k est fixé par

l'équipe commerciale et il représente le nombre de groupes tarifaires désirés. Le problème des k moyennes consiste ici à segmenter les départements en groupes tel que chaque taux de perte soit assez proche du taux de perte moyen de son groupe.

2. L'algorithme des k moyennes

Dans sa forme la plus générale (décrite ci-dessus), le problème des k moyennes est un problème difficile à résoudre à la fois de manière exacte et efficace. L'algorithme des k moyennes est un algorithme qui a fait ses preuves dans de nombreuses applications. Il est considéré comme simple, rapide et efficace bien que ne garantissant pas l'exactitude sur des cas complexes.

L'algorithme sous sa forme la plus classique (celle de Stuart Lloyd en 1957) se compose de quatre étapes (2):

1. choisir k centres aléatoirement $C = \{c_1, \dots, c_k\}$,
2. pour chaque c_i, établir l'ensemble C_i des points de X qui sont plus proches de c_i que de n'importe quel autre c_j,
3. pour chaque C_i, écraser les c_i par les barycentres des C_i : $c_i = \sum_{x_j \in C_i} \frac{x_j}{|C_i|}$,
4. réitérer les étapes 2 et 3 jusqu'à ce que C ne change plus.

L'image suivante illustre très bien ces différentes étapes.

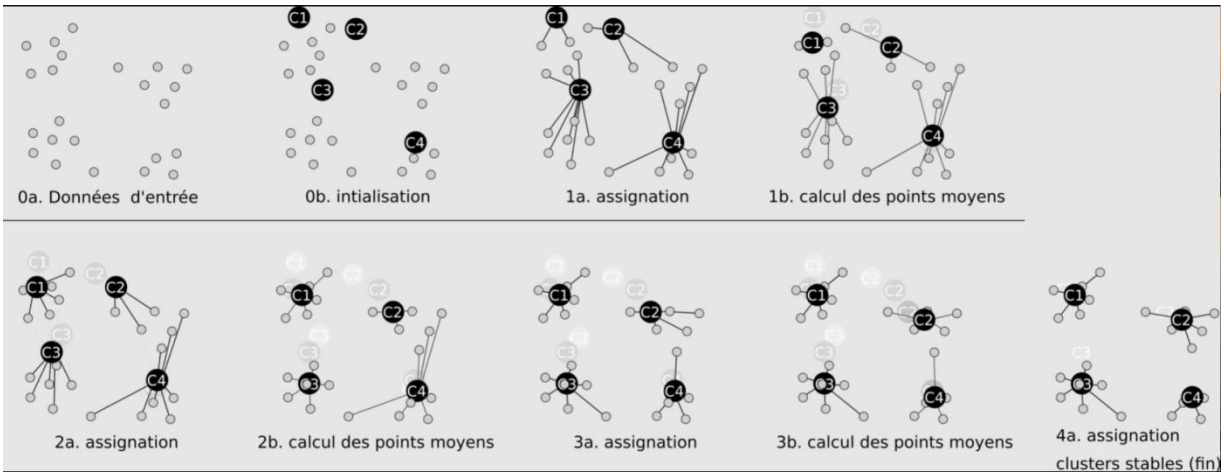

Figure 8 : Evolution étape par étape de l'algorithme des k moyennes en deux dimensions et avec k = 4 (3)

Variantes :

L'algorithme des k moyennes a connu de nombreuses variantes au cours de l'histoire :

- La méthode de Forgy en 1965 est l'équivalent continu de la méthode discrète de Lloyd (4). La méthode de Forgy et Lloyd souffre d'un problème majeur : le résultat final est très influencé par l'initialisation aléatoire des centres. Une des

raisons principales est qu'à chaque étape les centres sont actualisés simultanément.

- En 1967, MacQueen pallie ce défaut : il propose de réaliser une boucle, non pas sur les centres, mais sur chaque point de X. Pour chaque x_i, on regarde quel centre est le plus proche et on lui assigne le groupe de ce centre. Dans la foulée, on actualise les coordonnées du centre avec l'arrivée de x_i. Cette méthode demande donc davantage de calculs mais le résultat est amélioré : il est désormais moins influencé par l'initialisation (5)
- En 1979, Hartigan et Wong profitent de la puissance croissante des processeurs et proposent une méthode plus sophistiquée. Là encore, la boucle porte sur les x_i mais cette fois-ci dans une sous-boucle on regarde pour chaque c_j ce qu'il adviendrait de la fonction $\delta(G)$ définie en partie V-A-1 si le point x_i lui était assigné. On réalise alors l'assignation qui diminue le plus la fonction $\delta(G)$. (6)
- On ne compte plus les variantes existantes. La recherche dans ce domaine est foisonnante et toujours d'actualité. Le clustering est en effet un outil fondamental en statistiques et en intelligence artificielle.

Initialisation des centres :

Malgré des algorithmes de plus en plus sophistiqués, la solution obtenue dépend toujours des centres de départ tirés aléatoirement. Elle n'est qu'un optimum local. En général on réitère l'algorithme des milliers de fois avec des centres initiaux différents afin de tester différents optimums et de retourner le meilleur résultat trouvé.

Il existe d'ailleurs de nombreuses méthodes pour initialiser les centres :

- l'initialisation, la plus simple, celle de Forgy, consiste à sélectionner aléatoirement, simultanément et de manière uniforme k points de l'ensemble X. Cette méthode conduit souvent l'algorithme dans des impasses avec des optimums locaux peu satisfaisants. Il faut donc tester de nombreuses configurations initiales, les meilleurs étant celles où les points initiaux sont assez bien répartis dans l'espace X ;
- l'initialisation par les partitions aléatoires consiste à attribuer aléatoirement un groupe à chaque point de X avant de calculer les centres correspondants. En général, les centres initiaux se trouvent donc très proches les uns des autres mais l'algorithme les fait diverger petit à petit. La méthode Hartigan et Wong est initialisée de cette manière;
- l'initialisation kmeans++, plus moderne, est aussi très populaire et efficace. Elle consiste à sélectionner aléatoirement k points directement dans le set de données et suffisamment éloignés les uns des autres. Pour ce faire, on sélectionne successivement les centres et on utilise une loi de probabilité qui favorise la distance par rapport aux centres déjà sélectionnés. (7)

Implémentation dans R :

Dans le cas précis de cette étude, nous utilisons l'algorithme k-means de la librairie stats de R (Voir Annexe). Par défaut, il s'agit de la version de Hartigan et Wong.

Comme nous travaillons ici en une seule dimension, nous pouvons nous permettre d'élever le nombre de tentatives à 100 000 tout en préservant une durée de calcul raisonnable (quelques secondes seulement). Ceci nous garantit un résultat unique et optimal.

Employer cette méthode pour résoudre ce problème d'optimisation en une dimension peut sembler exagéré. Cependant l'avantage de l'avoir implémenté ainsi permet de prévoir d'éventuelles améliorations ou bien des applications à d'autres portefeuilles, qui utiliseraient davantage de dimensions. Avec un fichier de données plus grand (plus d'années) et avec un nombre de champs plus important, on pourrait imaginer ajouter d'autres paramètres que le taux de perte pour établir la segmentation.

B. Description

La méthode par clustering débute de la même manière que la méthode des régions (cf. IV-A) : on établit des critères de significativité pour les départements et les départements non significatifs se voient attribuer le taux de perte de leur région.

Ensuite, on considère l'ensemble des taux de pertes de chaque département et on applique l'algorithme des k moyennes pour segmenter les départements en plusieurs groupes de départements plus ou moins favorables.

A titre d'illustration, le graphique suivant représente les taux de pertes de chaque département. Les bulles colorées représentent les différents groupes obtenus lorsque l'algorithme est lancé avec k = 6.

Figure 9 : Algorithme des k moyennes appliqué au taux de perte des départements pour 6 groupes tarifaires

On observe que l'algorithme a détecté un département aberrant au taux de pertes extrêmement grand ; il l'isole dans son propre groupe. Il détecte aussi 7 bons départements aux taux de pertes allant de 6% à 16% tandis que le suivant est loin à 19%. Il crée donc un groupe de 6% à 16%. On distingue également une frontière nette au niveau de 35%. Jusqu'ici, un œil humain aurait pu proposer la même chose. Cependant placer les 2 dernières frontières manquantes serait très subjectif. L'algorithme des k moyennes permet d'apporter une solution automatisable et objective : celle qui minimise le total des variances au sein de chaque groupe.

Disposant de cette nouvelle segmentation, on peut maintenant oublier les taux de pertes de l'étape 1 (méthode des régions) et repartir des données sources pour dresser des taux de pertes par groupe de département.

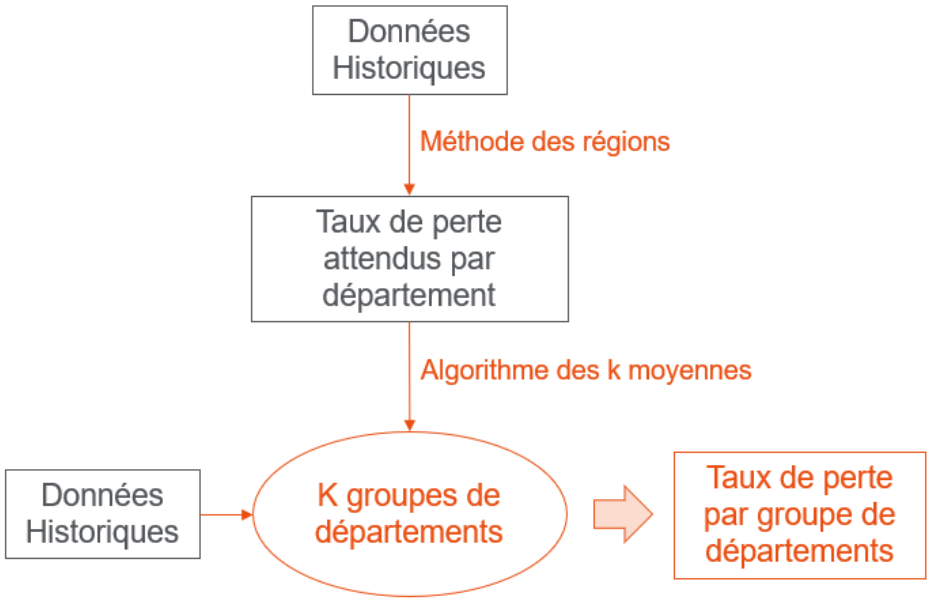

Figure 10 : Méthode de la tarification par clustering

C. Résultats

De la même manière que nous pouvons établir un taux de perte par groupe, nous pouvons également le faire en fonction des zones définies par le ministère de l'agriculture (cf. I-D). Pour rappel, ces zones avaient été utilisées pour définir des limites de garanties variables. La zone 1, la plus défavorable avait une limite basse pour limiter les mauvais résultats et inversement.

Il est intéressant à ce stade de vérifier la pertinence de ces zones en regard de notre toute nouvelle classification en groupes.

Taux de perte Groupes \ Zones	Zone 1	Zone 2	Zone 3	Total
Groupe 1	38%	15%	13%	**14%**
Groupe 2	15%	23%	24%	**23%**
Groupe 3	31%	27%	33%	**30%**
Groupe 4	38%	40%	39%	**39%**
Groupe 5	40%	48%		**47%**
Groupe 6 (aberrant)		70%	30%	**68%**
Pays	**36%**	**41%**	**28%**	**34%**

Tableau 2 : Résultats de la méthodologie intermédiaire (partie V)

Un premier constat frappant et indépendant de la nouvelle classification porte sur le total de chaque zone. La zone 1 prétendument défavorable s'avère en fait avoir un taux de perte historique inférieur à la zone 2. De plus, on remarque que les trois taux de pertes sont tous très proches de la moyenne générale.

On observe ensuite que la segmentation en groupes est bien plus satisfaisante. En effet, alors que celle-ci possède 6 groupes au lieu de 3, les intervalles entre chaque taux de perte sont plus grands et couvrent l'intégralité du spectre des résultats [6% ;68%]. On constate aussi que même à une échelle plus fine la nouvelle segmentation reste cohérente : chaque colonne est strictement croissante. On note seulement deux exceptions. La première (groupe 6 zone 3) concerne le groupe 6 qui est justement utilisé pour isoler un département au comportement aberrant (cf partie précédente) et la deuxième (groupe 1 zone 1) est expliquée en partie C-1.

Au contraire, la segmentation en zones n'est pas robuste. Par exemple à la ligne 3, c'est la zone 2 la plus favorable et la zone 3 la moins favorable, tout le contraire du total. La nouvelle segmentation discrédite donc le système de zones et montre que des segmentations bien plus réalistes sont possibles.

D. Avantages et inconvénients

La méthode présentée ici nous permet donc de créer des groupes de tarification là où la méthode précédente ne fournissait qu'un taux national. Ces groupes sont créés de manière objective avec une métrique claire, celle de la variance du taux de perte intra-groupe. De plus, le souscripteur peut fixer lui-même le nombre k de groupes désiré en fonction des besoins.

Cependant la méthode souffre encore de trois défauts majeurs et nous allons les développer dans les sous parties-suivantes :

1. elle perpétue la même « approximation administrative », très grossière, observée dans la méthode initiale (IV-B) : un département non significatif se voit attribuer le taux de perte de sa région administrative. Ce qui n'a pas de réel sens sur le plan agroclimatique ;
2. les départements non significatifs ont un impact non négligeable sur le clustering ;

3. la carte obtenue est très hétérogène, ce qui est difficile à défendre sur le plan commercial.

1. La gestion des départements non significatifs encore trop grossière

Les effets de « l'approximation administrative » s'observent très bien dans les résultats exposés précédemment (cf. Tableau 2). Pour groupe 1 zone 1, nous observons un taux de perte de 38% qui détonne avec le reste des résultats ; en voici la raison.

Le groupe 1 étant un groupe très favorable, il coïncide plutôt bien avec la segmentation du ministère dans le sens où la quasi-totalité des départements se trouvant dans le groupe 1 sont dans la zone « intermédiaire » ou dans la zone 3 « favorable ». En regardant les données de plus près, on s'aperçoit qu'en réalité un seul département se retrouve dans le groupe 1 et la zone 1. Or ce département est non significatif sur la zone 1, son taux de perte réel de 3% n'a donc pas été considéré. C'est le taux de perte de 38% provenant de la région administrative correspondante qui a été pris en compte à la place.

2. Un clustering à manier avec précaution

L'approximation décrite plus haut a lieu avant le clustering et ceci a un fort impact sur le résultat. En effet, en analysant plus précisément les données qui génèrent la Figure 10, on s'aperçoit que de nombreux points sont confondus car ils possèdent le même taux perte. Ceci arrive quand une même région possède plusieurs départements non significatifs et transfère à chacun le taux de perte régional.

Malheureusement, le clustering considère chaque point avec une importance égale. Il ne prend pas en compte le volume du département. Le résultat se retrouve donc fortement influencé par des amas de départements non significatifs possédant le même taux de perte. La segmentation des départements en groupes devient alors très sensible. Ainsi de petites variations de taux de pertes dans les données historiques peuvent entrainer un chamboulement de la segmentation.

3. Le manque de cohérence géographique

Nous avons observé précédemment que l'approximation administrative entraîne des aberrations visibles dans les résultats à l'échelle macroscopique. A fortiori de telles aberrations apparaissent également à des échelles inférieures : une bonne vingtaine de départements non significatifs se retrouvent à tort dans certains groupes favorables alors que le souscripteur sait pertinemment que le climat ne l'est pas (et inversement). Le souscripteur doit alors procéder à de nombreuses corrections subjectives en déplaçant un département d'un groupe vers un autre. Ce qui montre la défaillance de la méthodologie. Dans la figure suivante nous pouvons observer la segmentation proposée par le programme avant et après correction du souscripteur.

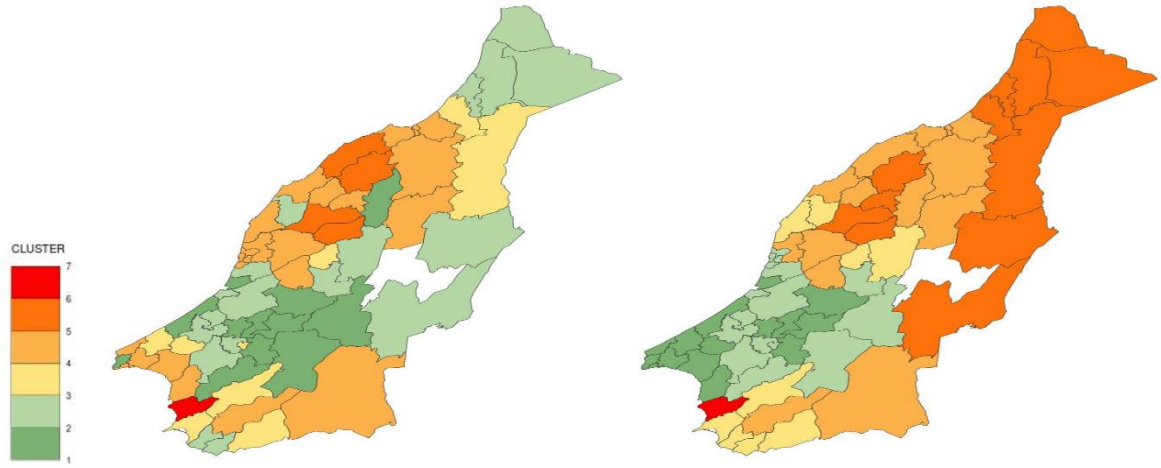

Figure 11 : Segmentation des départements en 6 groupes (méthode intermédiaire - partie V) avant et après correction du souscripteur

Un deuxième problème important saute aux yeux : il s'agit de l'hétérogénéité de la segmentation proposée à gauche. Même en admettant que cette solution soit actuariellement correcte, il n'est pas crédible ni envisageable sur le plan commercial, de proposer une tarification aussi morcelée.

C'est à partir de ces 3 constats que la méthode finale du clustering en région agroclimatique a été imaginée. La partie suivante résout simultanément les 3 problèmes : elle permet de gérer les départements significatifs de manière différente, de rendre le clustering moins sensible et de proposer un résultat cohérent sur le plan géographique.

La tarification par clustering en résumé :

- Définition des départements significatifs
- Taux de perte départemental pour les départements significatifs
- Taux de perte régional pour les départements non significatifs
- Segmentation de tous les départements en k groupes homogènes (algorithme k-means)
- Correction manuelle pour certains départements non significatifs
- Calcul des taux de pertes de chaque groupe

Avantages :

+ Tarifs par groupe
+ Nombre de groupes flexible
+ Analyse au niveau départemental (Instauration de limites possible)

Inconvénients :

– « Approximation administrative » atténuée mais toujours présente
– Clustering sensible et biaisé par les départements non significatifs
– Manque de cohérence géographique

VI. Méthodologie finale : Tarification par clustering sur département significatif

A. Description

La tarification par clustering sur département significatif est la méthodologie la plus satisfaisante à laquelle nous avons finalement abouti. Elle combine des principes vus dans les deux parties précédentes mais se scinde cette fois en trois étapes distinctes pour bien mener la gestion des départements non significatifs :

- étape 1 : Clustering des départements significatifs basé sur l'expérience,
- étape 2 : Extrapolation manuelle pour les départements non significatifs basée sur les connaissances des experts,
- étape 3 : Fusion de l'expérience et de l'extrapolation.

1. Clustering des départements significatifs

Comme son nom l'indique, cette étape consiste à appliquer le clustering vu en partie V-B en se limitant cette fois-ci à liste des départements significatifs établie en partie IV-A-1 i.e les départements :

- souscrits entre 2013 et 2020 sans interruptions,
- dont la superficie assurée totale cumulée sur 8 ans dépasse les 20 000 hectares,
- dont les capitaux assurés cumulés sur 8 ans dépassent les 10 millions d'euros.

Le fait de se restreindre aux départements significatifs avant la phase de clustering est très important pour deux raisons.

D'abord cela permet de supprimer le problème des doublons. Pour rappel, dans la méthode précédente le clustering était pollué par les départements non significatifs possédant exactement le même taux de perte (du fait de l'approximation par la méthode des régions) et ceci était absurde car les départements non significatifs avaient finalement plus de poids que les autres.

Rappelons également que le clustering est une étape sensible dans le sens où une petite variation dans les taux de pertes considérés peut générer une segmentation très différente. En particulier, un département aberrant peut tout à fait s'octroyer un groupe à lui tout seul. Il est donc indispensable de baser le clustering sur des données fiables qui reflètent vraiment l'expérience de la cédante.

La figure ci-dessous montre le clustering sur départements significatifs comparé au clustering sur tous les départements.

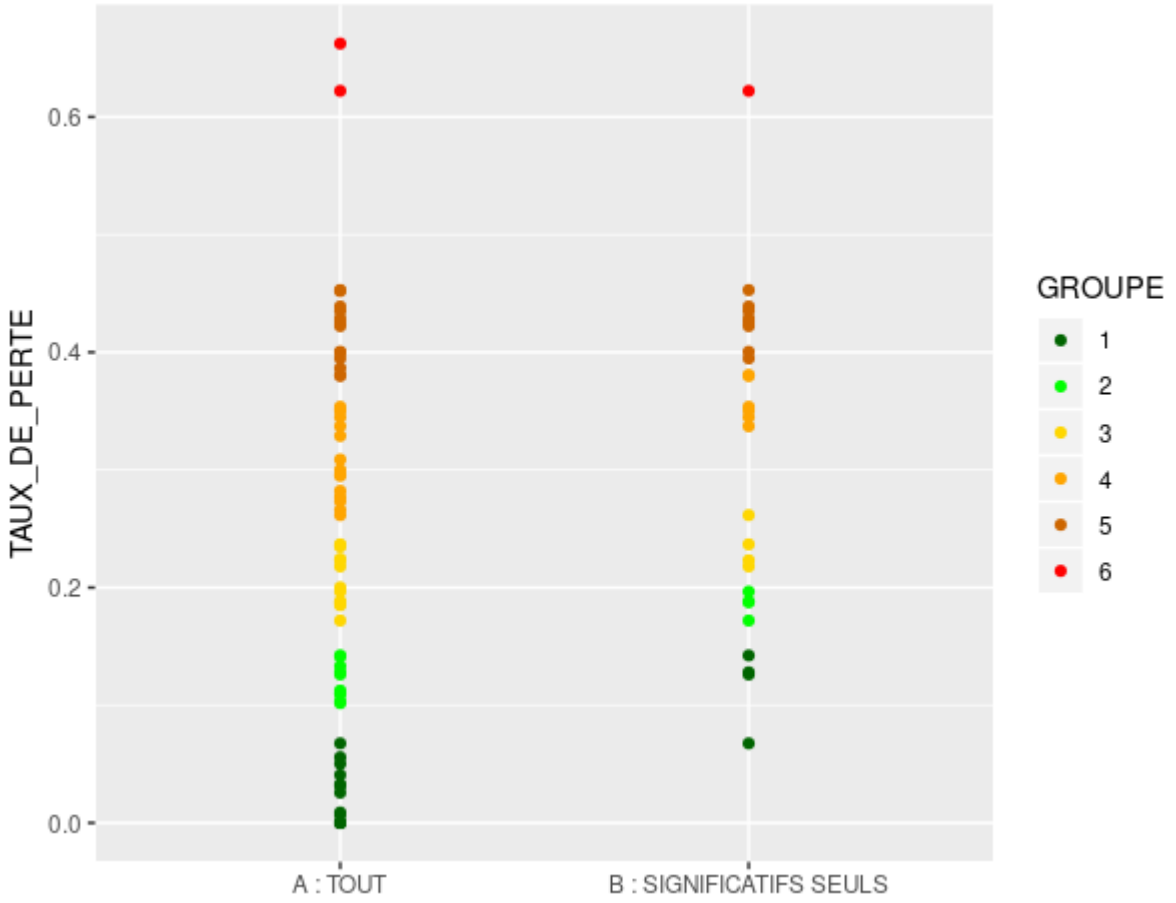

Figure 12 : Clustering sans et avec filtrage des départements non significatifs

On observe à gauche de nombreux départements non significatifs, au taux de perte faible. Il s'agit de départements avec une histoire tellement courte qu'ils n'ont jamais eu de sinistres. La suppression de ceux-ci avant l'étape de clustering a un fort impact sur les groupes 1 et 2. Avant l'algorithme constatait qu'il était facile et courant d'avoir un taux de perte inférieur à 10%. Maintenant l'algorithme intègre le fait que 15% est en fait un excellent score également.

2. Extrapolation manuelle pour les départements non significatifs

Nous disposons désormais d'une segmentation fiable et nous pouvons déjà dresser une carte des départements significatifs.

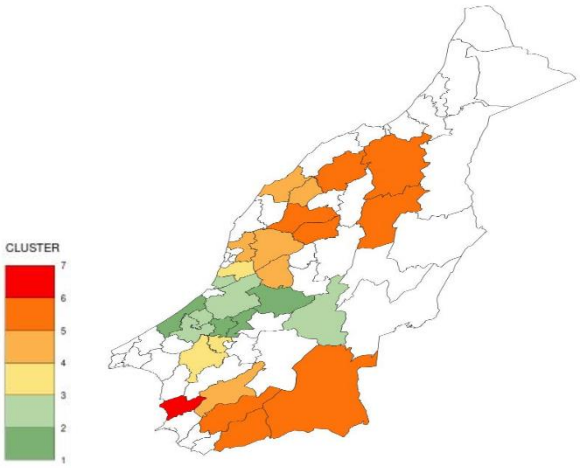

Figure 13 : Segmentation automatique des départements significatifs en 6 groupes (méthode finale – partie VI)

On voit apparaître de grandes zones de couleurs connexes ainsi que des dégradés relativement cohérents. L'étape de clustering n'a pourtant pas utilisé d'informations sur la disposition géographique. C'est donc une preuve de son bon fonctionnement. Ceci augmente également la légitimité d'un modèle construit en zone agroclimatique.

Il reste maintenant à attribuer un groupe aux départements non significatifs. Par définition, il n'y pas assez d'expérience sur ces départements pour en tirer des conclusions solides. Nous avons déjà vu en Figure 11 à quel point il était vain de s'efforcer de proposer une segmentation pour ces départements-ci car le souscripteur doit de toute façon intervenir au cas par cas dans la classification de ceux-ci.

Contrairement à la méthodologie précédente, ce fait est ici complétement assumé : il n'est pas possible de traiter ces départements en se basant sur les chiffres, l'opération manuelle est donc nécessaire et nous pouvons d'ailleurs en profiter pour améliorer la cohérence géographique. C'est le souscripteur qui se charge de cette étape. Il assigne les départements non significatifs aux groupes déjà établis en se basant sur plusieurs critères.

1) Il utilise principalement les cartes dressées par les experts agroclimatiques du pays. Ces experts renseignent le souscripteur sur les différents climats du pays et leurs différents impacts sur les cultures. La zone est-elle propice à l'agriculture ou non ? Les résultats sont-ils volatils ? Le souscripteur fait en sorte de reproduire approximativement la même carte que l'expert tout en respectant les résultats du clustering significatif : il ne doit jamais changer un département significatif de groupe.
2) En cas de doute, il peut croiser son analyse en utilisant les statistiques du ministère et observer la volatilité des rendements. Pour chaque groupe établi dans la partie précédente on peut se faire une idée du niveau de volatilité correspondant dans les statistiques du ministère. Un département non

significatif a lui aussi un certain niveau de volatilité dans les données du ministère et on peut donc établir une correspondance.

3) Il essaye également de rétablir une certaine continuité spatiale entre les départements du même groupe en vue de dresser une carte des tarifs la moins morcelée possible. Si un département non significatif se situe entre deux départements significatifs du même groupe, on lui attribuera de préférence le même groupe dans la mesure où il n'y a pas de contradictions avec les points précédents.

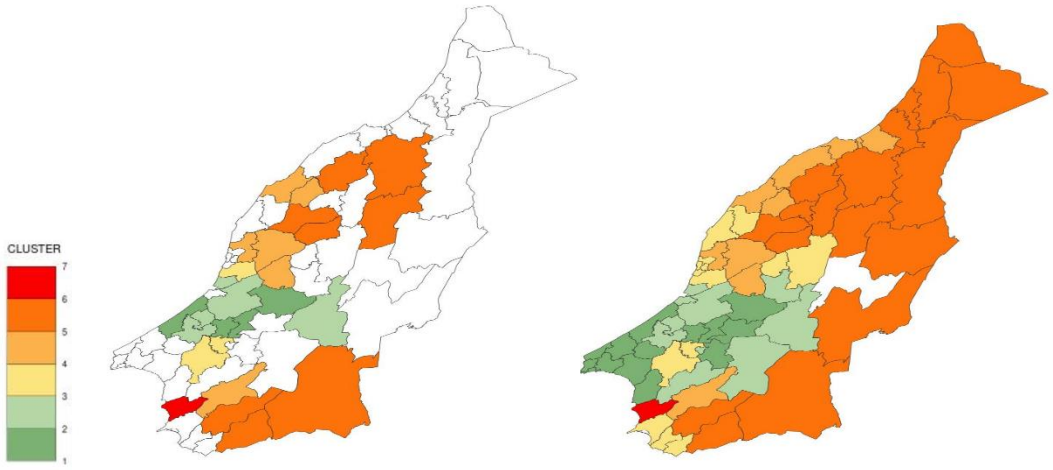

Figure 14 : Extrapolation manuelle opérée par le souscripteur sur les départements non significatifs en se basant sur la carte significative (méthode finale – partie VI)

3. Fusion de l'expérience et de l'extrapolation

Les départements non significatifs possèdent donc désormais un groupe. Cependant ils ne possèdent pas encore un taux de perte. Nous allons donc réappliquer le même principe vu dans la méthode des régions mais cette fois-ci avec une différence notable. Chaque département non significatif se verra attribuer le taux de perte du groupe agroclimatique auquel il se rapporte (et non plus le taux de perte de la région administrative).

Pour calculer ces taux de pertes par groupe, on additionne tous les sinistres historiques et toutes les sommes assurées de ce groupe. Attention, à cette étape, les départements non significatifs sont eux aussi intégrés dans le calcul car il s'agit d'obtenir des taux de perte représentatifs des régions agroclimatiques dans leur ensemble. (Pour rappel, dans ce rapport le terme « non significatif » signifie seulement : « trop petit pour constituer un chiffre fiable et utilisable à l'échelle du département »). L'impact de cette intégration est toujours faible puisqu'il s'agit de petits volumes mais il n'est pas complètement négligeable. Par exemple, au moment de l'extrapolation manuelle plusieurs départements sont ajoutés au groupe 5 mais si ceux-ci se situent majoritairement sous la moyenne du groupe 5 avant extrapolation, la moyenne finale sera alors légèrement décalée vers le bas.

Nous disposons donc maintenant des taux de pertes par groupe et des taux de pertes par département. Et lorsqu'un département n'est pas significatif, on écrase son taux de perte historique par le taux de perte de son groupe.

Le fait de posséder les taux de pertes attendus dans tous les départements offre ensuite à l'actuaire la possibilité de réaliser tous les traitements nécessaires et imaginables. En particulier, on peut procéder à des simulations en fonction des sommes assurées attendues pour chaque département (calcul expliqué en IV-A-2). Une fois les traitements effectués à cette petite échelle, on peut, bien sûr, par agrégation, générer les résultats à toutes les échelles supérieures : région administrative, pays et surtout groupe agroclimatique.

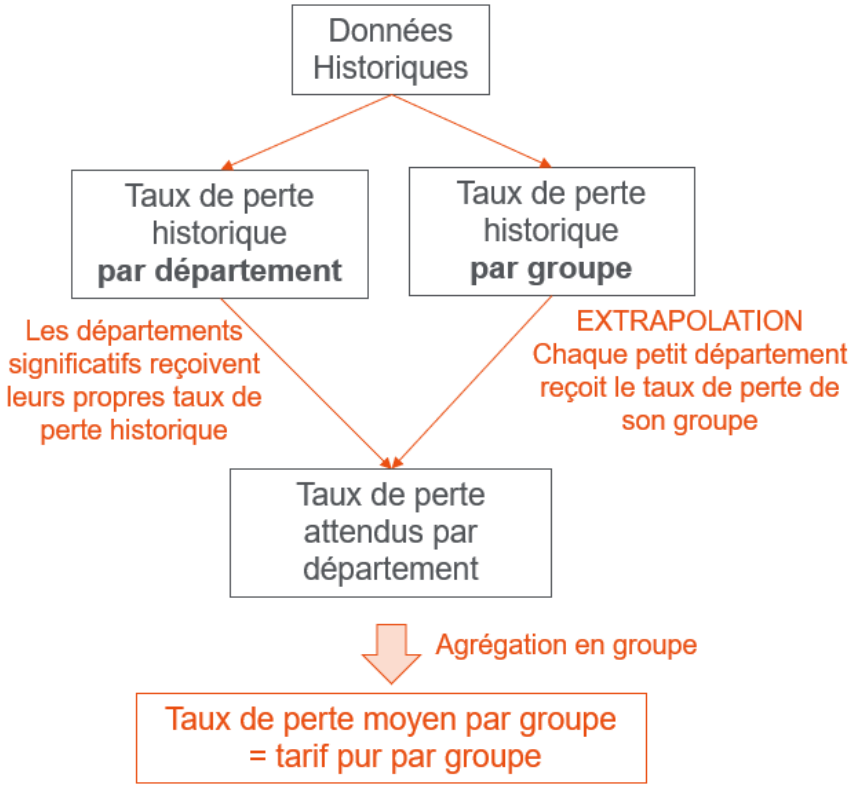

Figure 15 : Étape finale de la tarification par clustering sur départements significatifs

B. Résultats

De même que pour la méthode précédente, nous pouvons dresser le tableau des taux pertes par groupe et zone pour comparer la méthode finale et la méthode historique en zones. Dans un premier temps, observons les résultats des départements significatifs seulement ; c'est-à-dire les résultats avant l'extrapolation manuelle du souscripteur.

Taux de pertes Groupes \ Zones	Zone 1	Zone 2	Zone 3	Total
Groupe 1	3%	13%	12%	**12%**
Groupe 2	19%	23%	19%	**20%**
Groupe 3		26%	25%	**25%**
Groupe 4	37%	34%	36%	**36%**
Groupe 5	41%	45%	50%	**44%**
Groupe 6 (aberrant)		64%	27%	**62%**
Pays	**39%**	**40%**	**26%**	**33%**

Tableau 3 : Résultats de la méthodologie finale (partie VI) restreints aux départements significatifs

On constate que tous les avantages acquis par la méthode intermédiaire sont préservés dans la méthode finale :

- la segmentation en groupes couvre le spectre des taux de pertes de manière beaucoup plus satisfaisante que celle en zones ;
- cette segmentation est robuste même à des échelles inférieures ;
- le groupe 6 sert toujours à isoler le département aberrant.

De plus l'erreur grossière en groupe 1 zone 1 qui était due à « l'approximation administrative » a disparu comme prévu. Les résultats sont donc beaucoup plus logiques et fiables. Ajoutons maintenant les départements non significatifs pour une vision globale et finale.

Taux de pertes Groupes \ Zones	Zone 1	Zone 2	Zone 3	Total
Groupe 1	3%	12%	12%	**12%**
Groupe 2	19%	23%	19%	**20%**
Groupe 3	31%	26%	25%	**25%**
Groupe 4	37%	35%	36%	**36%**
Groupe 5	39%	45%	50%	**43%**
Groupe 6 (aberrant)		64%	27%	**62%**
Pays	**37%**	**39%**	**26%**	**32%**

Tableau 4 : Résultats de la méthodologie finale (partie VI)

Comme attendu, les modifications ne sont que très marginales. Ceci est un point extrêmement positif car c'est la preuve que l'étape manuelle et subjective de l'extrapolation n'est pas critique dans le processus.

C. Discussion

Cette méthodologie finale corrige donc les trois inconvénients majeurs de la méthode précédente. Le clustering n'est maintenant plus aussi sensible car il se focalise sur les départements significatifs uniquement. Les départements non significatifs sont utilisés astucieusement pour rétablir la cohérence géographique recherchée. Ils n'interviennent dans les calculs qu'au dernier moment et leur impact est de toute façon marginal.

Le principe même du modèle reflète très bien la réalité du produit, c'est-à-dire un produit dont les résultats dépendent avant tout de la zone agroclimatique où il est souscrit. Il permet de mettre en lumière les dangers de l'antisélection en exposant les énormes écarts de taux de pertes entre les zones.

Cette méthode remplit avec succès les objectifs fixés au départ car le souscripteur peut demander le nombre n de groupes qu'il souhaite et obtenir en guise de résultats n zones agroclimatiques cohérentes et les n tarifs correspondants. Proposer ces n tarifs « sur mesure » à la place d'un tarif national unique est indéniablement le moyen le plus simple et le plus efficace de réduire l'effet d'antisélection auprès des agriculteurs.

En tant que réassureur, il reste maintenant à tirer parti de ce nouveau modèle en n groupes et à tarifer l'excédent de perte en conséquence.

La tarification par clustering sur départements significatifs en résumé :

- Taux de perte départemental pour les départements significatifs
- Segmentation des départements significatifs en k groupes homogènes (algorithme k-means)
- Assignation manuelle des départements non significatifs en fonction de :
  - la connaissance des experts
  - la cohérence géographique
  - données additionnelles extérieures (e.g. ministère)
- Calcul des taux de perte par groupe (en incluant les non significatifs)
- Taux de perte du groupe pour les départements non significatifs
- Agrégation as if au niveau groupe

Avantages :

+ Tarifs par groupe
+ Nombre de groupes flexible
+ Cohérence géographique et agroclimatique
+ Analyse au niveau départemental (Instauration de limites possible)

Inconvénients :

– Méthode sophistiquée
– Code recommandé
– Intervention manuelle pour les départements non significatifs

VII. Réassurance

Les parties précédentes de ce mémoire ont montré comment la tarification du produit d'assurance a été remaniée en profondeur avec la mise en place d'une grille de tarifs basée sur le regroupement de départements similaires. Nous souhaitons, pour conclure ce mémoire, expliquer comment on peut adapter la méthode classique de tarification de la réassurance en y intégrant la segmentation en groupes fraichement instaurée. Le but du réassureur étant d'encourager l'assureur à optimiser son portefeuille.

A. Programme de réassurance et enjeux sur la nouvelle tarification

Dès la création du produit d'assurance en 2013 (cf. partie I-C), PartnerRe et CéréAssure avaient convenu d'un contrat de réassurance en excédent de pertes exprimées, non pas en pourcentage de la prime, mais en pourcentage de la somme assurée. Ceci dans le but de suivre la même logique que le produit d'assurance dont l'indemnité et les résultats sont exprimés en taux de perte.

La structure se divisait en trois tranches et reste à ce jour inchangée :

- Tranche 1 : 30% xs 10%
- Tranche 2 : 40% xs 40%
- Tranche 3 : 20% xs 80%

(Voir partie I-B-2-2 pour un rappel sur le fonctionnement de l'excédent de perte et le système de tranche)

Depuis 2013, seule la prime de réassurance demandée a pu changer au cours du temps en fonction de l'évolution des résultats historiques, de la conjoncture ainsi que quelques conditions sur les limites par départements. La méthode de tarification de la réassurance était une méthode tout à fait classique (voir partie B ci-dessous).

En 2021, avec l'amélioration du produit d'assurance et la conception des groupes agroclimatiques, nous savons maintenant que certaines régions ont des résultats bien meilleurs que les autres. En tant que réassureur, intégrer cette même segmentation dans la tarification de la réassurance est à la fois un devoir mais aussi une superbe opportunité. Le fait d'intégrer cette segmentation dans notre tarification permet d'inciter l'assureur à optimiser son portefeuille en vue d'une meilleure diversification et d'une volatilité atténuée. C'est aussi le moyen de s'aligner avec le client et de lui garantir la crédibilité de notre étude.

Intégrer la segmentation n'est cependant pas sans enjeux. Tout d'abord, cela complexifie grandement la grille tarifaire qui passe d'une seule dimension à deux. En réassurance, il y a déjà effectivement un tarif par tranche. Ajouter la notion de groupe multiplie donc le nombre de tarifs : un par groupe et par tranche. Dans notre exemple avec 6 groupes, nous passons donc de 3 tarifs à 18 tarifs.

De plus, il est important de rester cohérent avec la tarification passée. En supposant toutes choses égales par ailleurs, c'est-à-dire que le portefeuille n'évolue pas entre 2020 et 2021, que l'appréciation du risque ne change pas (i.e les résultats de la nouvelle année ne remettent pas en question l'analyse de l'année précédente), etc… le tarif total ne devrait pas changer avec le nouveau système de tarification.

Avant de proposer en partie C une méthode pour intégrer les groupes agroclimatiques dans la grille tarifaire, nous exposons la méthode de tarification classique en réassurance. Nous essayerons de rester le plus synthétique possible pour ne pas s'éloigner du sujet principal : la segmentation.

B. Méthode de tarification classique

L'agriculture fait partie des branches courtes (Short tail en anglais). Les engagements durent souvent un an tandis que le développement des sinistres s'étale de quelques mois à deux ans selon les pays. Le réassureur reçoit l'information sur le taux de perte de l'année N, lorsque la récolte est finie depuis un moment déjà : à la fin de l'année N ou bien de l'année N+1 pour les développements les plus longs. Il peut donc rapidement considérer les taux de pertes historiques comme définitifs (quitte à exclure la dernière année si nécessaire).

(Ce qui n'est pas le cas dans les branches longues i.e à long développement comme par exemple en responsabilité civile où la valeur finale du sinistre peut mettre plusieurs années avant d'être fixée. On utilise dans ce cas des techniques telles que la méthode Chain Ladder ou la méthode Bornhuetter–Ferguson)

En conséquence, les données minimales nécessaires pour tarifer la réassurance d'un produit d'assurance en agriculture sont les taux de pertes par année. Dans cette partie nous utiliserons comme exemple l'historique suivant :

2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
30%	10%	50%	20%	40%	20%	30%	40%	30%	80%

Il existe principalement deux méthodes à la disposition du réassureur :

- **La méthode empirique** applicable lorsque le volume de données est suffisamment important, en particulier lorsque l'assureur a déjà subi un certain nombre d'années suffisamment catastrophiques pour impacter les tranches étudiées. Elle se cantonne aux données historiques et calcule des résultats historiques et « as if » par tranche.
- **La méthode analytique** utilisable avec un plus faible nombre d'années et l'ajout d'hypothèses si nécessaire. Elle permet d'extrapoler les données avec une courbe de distribution.

Notons qu'il est toujours bon, lorsque la situation le permet, d'appliquer les deux méthodes et de confronter les résultats pour s'assurer de la cohérence de l'ensemble.

1. Méthode empirique

La méthode empirique consiste à calculer les résultats de la structure étudiée sur les données historiques. Cependant, il faut souvent procéder à des transformations « as if » comme en assurance (cf II-C), ceci dans le but de travailler sur une liste de taux de pertes homogènes qui présentent les mêmes risques et expositions ainsi que les mêmes conditions tarifaires qu'aujourd'hui.

Avec les tranches et les données énoncées plus haut, la méthode empirique donnerait les résultats suivants.

	Taux de perte brut assureur	T1 30% xs 10%	T2 40% xs 40%	T3 20% xs 80%
2011	**30%**	20%	0%	0%
2012	**10%**	0%	0%	0%
2013	**50%**	30%	10%	0%
2014	**20%**	10%	0%	0%
2015	**40%**	30%	0%	0%
2016	**20%**	10%	0%	0%
2017	**30%**	20%	0%	0%
2018	**40%**	30%	0%	0%
2019	**30%**	20%	0%	0%
2020	**80%**	30%	40%	0%
Moyenne historique	**35%**	**20%**	**5%**	**0%**

Tableau 5 : Exemple du fonctionnement de la méthode de tarification empirique

On remarque ici toute l'importance d'avoir un historique suffisamment long et représentatif avec de bonnes années et des années extrêmement mauvaises. La tranche 1 est presque toujours touchée et souvent traversée. On dit qu'elle « travaille ». La moyenne historique est donc très fiable. La tranche 2 est touchée deux fois dont une fois traversée. Sa moyenne de 5% reste un indicateur crédible. En revanche, la tranche 3 n'est absolument pas touchée. La méthode empirique ne permet donc pas de tarifer cette tranche et on doit recourir à des hypothèses ou utiliser la méthode analytique à la place.

La méthode empirique est certes limitée mais elle a le mérite de se baser sur la réalité et permet d'établir des points de repères concrets et fiables sur lesquels s'appuie la réflexion.

2. Méthode analytique

La méthode analytique consiste à construire un modèle mathématique, plus précisément, une courbe de distribution sur les montants de perte qui respecte au mieux les données existantes et permet d'extrapoler les événements rares et extrêmes.

Pour ce faire, de nombreuses courbes existent dans la grande famille mathématique des lois de probabilités à distribution continue. Trois types de courbes prédominent : la loi log normale, la loi pareto et la loi gamma (Figure 16). On sélectionne l'une ou l'autre de ces courbes en fonction de la ligne de business et du risque couvert.

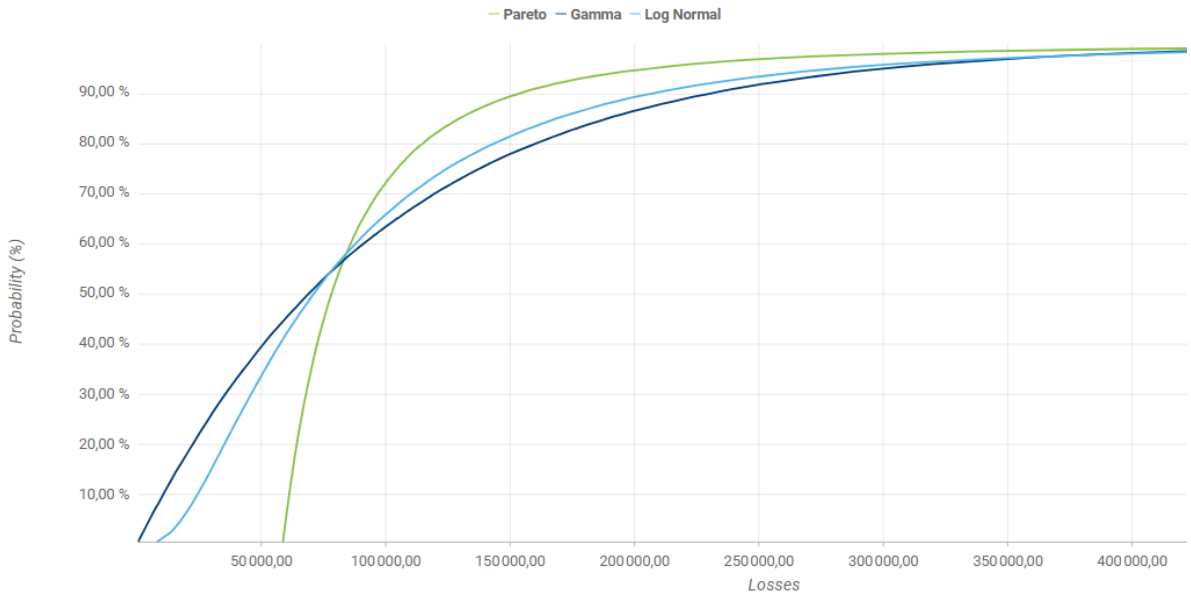

Figure 16 : Lois de distribution cumulée Pareto, Gamma et Log Normale avec une espérance et un écart type à 100 000 (CV = 1). En abscisse, les montants de sinistres potentiels. En ordonnée, la probabilité d'obtenir le montant ou moins, e.g. la courbe pareto en vert indique 9 chances sur 10 d'obtenir un montant de sinistres inférieur à 150 000.

Ces trois types de courbes ont en commun d'être régies par deux paramètres. À chaque fois ceux-ci se déduisent entièrement à partir de l'espérance et de la variance que l'on se fixe. Par exemple, pour la loi log normale, on déduit les paramètres mu et sigma selon ces deux formules :

$$\begin{cases} \sigma^2 = \ln\left(1 + \dfrac{Var(X)}{\big(E(X)\big)^2}\right) \\ \mu = \ln\big(E(X)\big) - \dfrac{\sigma^2}{2} \end{cases}$$

Estimation de la moyenne (espérance)

On détermine tout d'abord l'espérance en s'appuyant sur la méthode du burning cost (cf. III-A). En fonction de l'échantillon, cette étape peut s'avérer plus ou moins délicate. Lorsqu'une année catastrophique se glisse dans les données, il faut parfois la lisser sur plusieurs années. Si un taux de perte de 80% se glisse dans un historique de 10 ans alors que nous savons par ailleurs que ce genre d'évènement extrême ne survient que tous les 20 ans, on divisera cet impact catastrophique par deux. Reprenons le même échantillon que précédemment pour illustrer.

2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
30%	10%	50%	20%	40%	20%	30%	40%	30%	**80%**

On peut calculer le taux de perte attritionnel, c'est-à-dire le taux moyen hors catastrophes. Il correspond à la moyenne entre 2011 et 2019 : 30%. En 2020, la partie catastrophique s'élève donc à 50% et c'est cette partie qui sera lissée sur 20 ans au lieu de 10 : 50%/20 = 2,5%. Le taux de perte attendu est finalement égal à la somme de la partie attrionnelle et de la partie catastrophique : 32.5%

Inversement, si la période ne comporte pas d'années extrêmes, il faudra faire une hypothèse sur la fréquence et la sévérité d'un événement catastrophique. C'est généralement le souscripteur qui la formule, fort de son expérience, de sa connaissance du terrain, du marché et de ses discussions avec d'autres spécialistes. Prenons pour illustration l'échantillon suivant.

2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
30%	10%	50%	20%	40%	20%	30%	40%	30%	30%

Si nous formulons par ailleurs l'hypothèse d'une année catastrophique à 80% tous les 20 ans, nous obtenons alors le même résultat : une moyenne attritionnelle de 30% et un chargement pour couvrir les catastrophes futures (80% - 30%)/20 = 2.5%.

Estimation du CV

Au lieu d'estimer la variance, le réassureur préfère en réalité manipuler le CV. Pour rappel, le coefficient de variation (l'écart type divisé par la moyenne) est une façon de visualiser la volatilité de manière adimensionnée pour pouvoir comparer différentes affaires (voir partie IV-A-1). Cette notion est capitale pour estimer les risques de pointe.

Le réassureur établit régulièrement des études comparatives par zone géographique et par produit d'assurance qui lui servent généralement de référentiel. De plus, il peut aussi utiliser une ou deux hypothèses en guise de repères. Par exemple, on peut supposer un taux de perte de 50% tous les 30 ans et un taux de perte de 90% tous les 200 ans. Alternativement, si l'historique est suffisamment long on peut prendre un

point extrême directement disponible dans les données. Par exemple, sur un historique de 20 ans prendre le pire taux de perte et fixer la fréquence à 1/20. Dans tous les cas, là encore, on discute ces points de repères avec le souscripteur pour profiter de son expérience et de ses connaissances. Le CV est ainsi déterminé de manière à rapprocher la courbe de ces points de repères autant que faire se peut.

Simulation de Monte Carlo

Grâce à la courbe analytique ainsi établie, nous avons une définition absolument précise et continue des montants de pertes cumulées et de leurs probabilités de survenance. On procède alors ensuite à une simulation de Monte Carlo, pour calculer le résultat dans chaque tranche.

La simulation de Monte Carlo exploite la loi des grands nombres. Elle simule un nombre suffisamment grand d'années, en général 100 000, de telle manière que la moyenne observée tende vers la moyenne théorique (voir III-A pour un rappel sur la loi des grands nombres). Concrètement, pour chaque simulation on choisit un nombre aléatoire entre 0 et 1, et on se réfère à la courbe pour obtenir le montant de perte correspondant. Ce montant de perte est alors réparti entre les tranches conformément au traité de réassurance. L'algorithme garde en mémoire les montants dans chaque tranche. Une fois les 100 000 simulations réalisées, il suffit ensuite de calculer la moyenne dans chaque tranche pour obtenir le résultat escompté.

C. Méthode de tarification avec prise en compte des groupes

En suivant la méthode de tarification vue précédemment nous sommes capables de dresser autant de tarifs que de tranches. Nous prenons maintenant l'exemple de la première tranche et montrons comment on peut répartir ce tarif en tarif par segment. À des fins pédagogiques et dans le but de limiter la taille des tableaux, on se limitera cette fois-ci à trois groupes agroclimatiques avec les données suivantes :

Capital assuré($)	Groupe 1	Groupe 2	Groupe 3	Total
2013	41 121	108 999	235 327	**385 448**
2014	26 098	147 249	352 336	**525 682**
2015	34 978	116 540	356 897	**508 414**
2016	68 459	275 917	495 112	**839 489**
2017	45 453	208 942	434 890	**689 285**
2018	68 217	339 349	439 275	**846 842**
2019	39 224	217 658	446 957	**703 840**
2020	37 413	252 450	544 951	**834 814**
2021	32 027	180 538	442 124	**654 690**
Total	**392 990**	**1 847 643**	**3 747 870**	**5 988 503**

Taux de perte	Groupe 1	Groupe 2	Groupe 3	Total
2013	1,3%	2,8%	13,3%	**9,1%**
2014	1,9%	8,3%	54,7%	**39,1%**
2015	2,6%	2,5%	11,1%	**8,5%**
2016	52,5%	88,7%	88,0%	**85,4%**
2017	0,0%	0,6%	6,3%	**4,2%**
2018	0,0%	0,2%	7,6%	**4,0%**
2019	2,2%	7,4%	73,4%	**49,0%**
2020	35,8%	68,0%	98,4%	**86,4%**
2021	0,0%	3,4%	11,7%	**8,8%**
Total	13,3%	24,8%	44,7%	**36,52%**

Tableau 6 : Exemple de données historiques segmentées en 3 groupes (Capitaux assurés et taux de pertes)

On remarque encore une fois ici l'efficacité de la segmentation car elle permet par exemple de mettre en évidence que les mauvais résultats de 2014 et 2019 sont essentiellement liés au groupe agroclimatique 3. On en déduit que la sécheresse intense n'a pas affecté l'ensemble du pays ces années-là mais seulement une partie, celle qui, justement, était repérée comme défavorable. Les tarifs de réassurance doivent donc, eux aussi, être plus élevés sur le groupe 3.

1. Méthode

La méthode employée est similaire à la méthode empirique. La première étape consiste à appliquer la tranche de réassurance 30% xs 10% à chaque groupe distinctement. Ce qui conduit aux résultats suivants :

Taux de perte	Groupe 1	Groupe 2	Groupe 3	Total
2013	0,0%	0,0%	3,3%	**0,0%**
2014	0,0%	0,0%	30,0%	**29,1%**
2015	0,0%	0,0%	1,1%	**0,0%**
2016	30,0%	30,0%	30,0%	**30,0%**
2017	0,0%	0,0%	0,0%	**0,0%**
2018	0,0%	0,0%	0,0%	**0,0%**
2019	0,0%	0,0%	30,0%	**30,0%**
2020	25,8%	30,0%	30,0%	**30,0%**
2021	0,0%	0,0%	1,7%	**0,0%**
Moyenne pondérée	**7,7%**	**8,6%**	**15,2%**	**14,5%**

Tableau 7 : Taux de perte après réassurance (tranche 30% xs 10%)

Dans ce tableau, la colonne Total représente les taux de perte historique après réassurance (Le 39.1% de 2014 du tableau précédent devient bien 29.1% dans ce tableau). La ligne de total en revanche est obtenue en pondérant les taux de pertes par les capitaux assurés dans le but d'obtenir le taux de perte historique moyen de chaque groupe. On peut estimer que le portefeuille total génère, dans la tranche 30% xs 10%, un taux de perte empirique moyen de 14.46%. Il faudrait donc encaisser environ 866 000 dollars de primes pures (14.5% * 5 988 503) pour couvrir les pertes attendues.

À ce stade, nous ne nous pouvons pas simplement utiliser les taux indiqués en dernière ligne comme tarifs par groupe pour trois raisons.

Premièrement, si on appliquait ces taux tels quels à l'échelle de chaque segment, on n'encaisserait pas la prime attendue, comme le montre le tableau ci-dessous. En effet, la structure de réassurance avec ses limites maximales et minimales brise la propriété de linéarité de la moyenne pondérée.

	Groupe 1	Groupe 2	Groupe 3	Total
Taux pur = taux de perte	**7,7%**	**8,6%**	**15,2%**	**14,5%**
Capital assuré	392 990	1 847 643	3 747 870	5 988 503
Prime unique				**866 222**
Prime par groupe	30 194	158 510	571 080	**759 784**
Taux pur moyen				**12.7%**

Tableau 8 : Comparaison prime unique et prime par groupe sans correction

Deuxièmement, le réassureur a besoin d'une méthode flexible pour conserver la liberté de s'adapter à d'autres arguments extérieurs. Par exemple, avant 2021, PartnerRe avait fixé avec CéréAssure un taux de 15% sur la tranche 1 en partant du constat qu'historiquement la tranche avait été complètement traversée une fois sur deux (15% = 30%/2). L'hypothèse n'étant pas remise en question, si PartnerRe veut garder ce chiffre en 2022, les taux de chaque groupe doivent alors être revus légèrement à la hausse.

Enfin, troisièmement, il est plus judicieux de pondérer avec les capitaux les plus récents ceux de 2021, plutôt qu'avec les capitaux totaux. C'est effectivement la proportion de chaque groupe dans le portefeuille de 2022 qui nous intéresse et les proportions de 2021 en sont la meilleure estimation.

Pour résoudre ces trois points à la fois, on applique une correction proportionnelle afin de corriger les taux par groupe. D'abord on calcule la moyenne pondérée des taux par les capitaux de 2021 :

$$\frac{\sum_{\text{groupes}} \text{TP}_{\text{groupe}} * \text{CA}_{\text{groupe,2021}}}{\sum_{\text{dep}} \text{CA}_{\text{dep,2021}}} = 13.0\% = TP_{avant}$$

Puis on corrige chaque taux de perte par $\frac{TP_{cible}}{TP_{avant}} = \frac{15\%}{13.0\%}$ dans le but d'obtenir, cette fois, le taux pur moyen désiré.

	Groupe 1	Groupe 2	Groupe 3	Total
Taux pur = taux de perte	**7,7%**	**8,6%**	**15,2%**	
Taux pur corrigé	**8,8%**	**9,9%**	**17,5%**	
Capital assure 2021	32 027	180 538	442 124	654 690
Prime par groupe	2 832	17 827	77 543	**98 203**
Taux pur moyen				**15 %**

Tableau 9 : Prime par groupe après correction

Le réassureur peut maintenant proposer de manière équivalente un taux unique à 15% ou bien les 3 taux 8,8%, 9,9% et 17,5%.

2. Vérification As If

La méthode présentée dans la partie précédente est une méthode assez efficace mais néanmoins approximative. Il importe donc de procéder à une vérification. Une bonne manière consiste à raisonner en sens inverse avec des analyses as if. Nous pouvons étudier les résultats que nous aurions obtenus si la grille de tarifs avait été proposée depuis le début.

	Primes As IF				Taux pur AS IF	Taux de perte AS IF	Ratio
	Groupe 1	Groupe 2	Groupe 3	Total			
2013	3 636 539	10 763 441	41 273 677	55 673 656	14,4%	0,0%	**0,0%**
2014	2 307 935	14 540 467	61 795 544	78 643 947	15,0%	29,1%	**194,3%**
2015	3 093 222	11 508 022	62 595 483	77 196 727	15,2%	0,0%	**0,0%**
2016	6 054 160	27 246 227	86 836 888	120 137 276	14,3%	30,0%	**209,6%**
2017	4 019 595	20 632 602	76 274 652	100 926 850	14,6%	0,0%	**0,0%**
2018	6 032 762	33 509 961	77 043 760	116 586 483	13,8%	0,0%	**0,0%**
2019	3 468 776	21 493 292	78 391 111	103 353 179	14,7%	30,0%	**204,3%**
2020	3 308 563	24 928 874	95 578 074	123 815 512	14,8%	30,0%	**202,3%**
2021	2 832 329	17 827 761	77 543 336	98 203 426	15,0%	0,0%	**0,0%**
Total	**34 753 881**	**182 450 648**	**657 332 526**	**874 537 055**	**14,6%**	**14,5%**	**99,4%**

Tableau 10 : Vérification As If avec les capitaux historiques

Le taux de prime pur est alors comparé à celui que nous aurions perçu avec le taux de perte historiquement subi. Si la tarification est correcte, le ratio doit être proche de

100% car il indique la bonne adéquation de la prime avec les sinistres. C'est bien le cas ici avec 99,4%.

3. Avantages et Inconvénients de la réassurance en segment

La combinaison de la méthode empirique avec la solution de la correction proportionnelle permet donc de dresser des tarifs de réassurance par groupe de manière simple et efficace. Elle permet de mettre en exergue les disparités entre les groupes : on a pu montrer que le groupe 3 a un fort impact négatif sur les résultats de la tranche une par exemple. Et pourtant, la proportion de ce groupe (en capital assuré) n'a fait qu'augmenter depuis 4 ans (cf. Tableau 11). Il faut pouvoir traduire ce fait dans la grille de tarifs et la solution proposée ici remplit parfaitement l'objectif. Avec un taux deux fois plus bas sur le groupe 1 que sur le groupe 3, l'assureur est fortement incité à souscrire davantage dans les meilleurs départements pour réduire le coût de la réassurance (en plus d'améliorer ses propres résultats au niveau assurance). À l'inverse, si l'assureur persiste à augmenter sa proportion de départements défavorables, le tarif augmentera automatiquement et de manière continue.

Capital assuré (%)	Groupe 1	Groupe 2	Groupe 3	Total
2013	11%	28%	61%	100%
2014	5%	28%	67%	100%
2015	7%	23%	70%	100%
2016	8%	33%	59%	100%
2017	7%	30%	63%	100%
2018	8%	40%	52%	100%
2019	6%	31%	64%	100%
2020	4%	30%	65%	100%
2021	5%	28%	68%	100%
Total	**7%**	**31%**	**63%**	**100%**

Tableau 11 : Évolution des proportions de chaque groupe en % du capital assuré

La méthode présente néanmoins quelques inconvénients. Avant tout, comme toute méthode empirique, elle nécessite un historique suffisamment long et représentatif avec de bonnes et de mauvaises années où la tranche est tantôt touchée, tantôt traversée. Il faut également chaque année un volume significatif dans chaque groupe. La réassurance segmentée en groupes n'est donc pas toujours praticable.

Par exemple, le groupe 1 ne touche pas la troisième tranche 20% xs 80% et il est donc impossible d'appliquer la règle de proportionnalité sur un 0% historique. Deux options sont alors possibles. Soit on essaye de formuler une hypothèse extrême, pertinente, comme dans la méthode empirique (partie VII-B-1). Ceci est compliqué

lorsqu'on se restreint à une partie de l'ensemble car on peut difficilement faire appel aux connaissances des experts sur un groupe très particulier de départements. Soit on renonce à la segmentation sur cette tranche. Ce qui, de toutes façons, n'a pas beaucoup d'impact sur le tarif global, car la troisième tranche, peu touchée, est beaucoup moins chère que les deux autres. Dans notre cas précis, nous avons opté pour cette dernière solution.

Enfin, il faut toujours garder à l'esprit que la méthode empirique est par nature très sensible à l'historique. Vu sous un angle statistique, les 9 années enregistrées ne sont qu'un échantillon de 9 observations. Ici, les observations nous montrent des résultats très proches entre les groupes 1 et 2 pour la tranche 1 mais un échantillonnage différent aurait peut-être généré un écart de tarif plus grand entre les deux groupes. Pour cette raison, toutes les vérifications possibles sont bienvenues. Les tarifs doivent être maniés avec prudence et remis en question chaque année lorsqu'un nouveau chiffre est disponible. Ceci ne diffère en rien de la tarification classique de la réassurance en agriculture mais l'analyse à l'échelle des groupes requiert simplement un degré de prudence supplémentaire.

Conclusion

La méthode du burning cost as if est dans toutes les lignes de business la méthode de tarification la plus robuste, la plus efficace et la plus utilisée tant en assurance résultats passés est en première approximation le meilleur indicateur des résultats futurs (cf. partie III). Cependant cette simplicité apparente soulève en réalité de nombreuses questions lorsque, dans la pratique, on essaye d'analyser un portefeuille à des échelles inférieures. Jusqu'à quel point peut-on l'appliquer ? Peut-on légitimement comparer les résultats d'un sous-ensemble qui possède beaucoup de données avec un autre qui en possède moins ? Peut-on segmenter un portefeuille à partir des résultats du burning cost ?

Ce mémoire s'est consacré à la question très particulière en agriculture de la géographie, en étudiant un portefeuille hétérogène dont les résultats dépendent grandement de zones agroclimatiques peu ou mal définies au départ. Nous avons proposé une méthodologie qui s'appuie fortement sur la méthode du burning cost tout en palliant ses défauts. Il a fallu tâtonner pour perfectionner la méthodologie petit à petit car les étapes et l'ordre dans lequel elles interviennent ont un impact sur le résultat et sa signification.

Parmi les très nombreuses versions de ce projet, nous avons choisi de nous attarder sur la méthodologie dite « intermédiaire », (Partie V) une étape clé dans ce travail de recherche qui a marqué l'intégration d'une méthode de segmentation automatique basée sur une méthode d'apprentissage non supervisée. Nous avons choisi l'algorithme des k moyennes pour sa simplicité et également pour sa transparence. Sur le plan commercial, c'est une méthode facile à défendre car son objectif est clair et parlant pour le client assureur : la minimisation de la variance dans chaque sous-groupe. Une solution très pragmatique et réaliste étant donné l'information disponible. Pour aller plus loin on pourrait imaginer utiliser d'autres méthodes de clustering en intégrant à chaque fois l'aspect géographique (8). Par exemple, appliquer la méthode density based partioning en adaptant les règles de voisinage au problème étudié. Ou encore utiliser un algorithme de clustering hiérarchique qui n'autoriserait l'association de deux groupes de départements qu'à la condition que ceux-ci soient connexes géographiquement. Autant de sujets qui mériteraient un rapport entier.

La méthodologie finale décrite en partie VI est la méthode la plus aboutie et donne des résultats très satisfaisants. Elle règle le problème important des départements non significatifs et de la cohérence géographique. Notons ici que les progrès obtenus entre la version intermédiaire et la version finale ne proviennent pas d'un outil technique comme le burning cost ou le clustering. Ils proviennent plutôt d'un bon agencement des étapes, d'une bonne « recette » établie à force de tâtonnements, de bon sens et de vérifications rigoureuses. Au-delà de la théorie, c'est aussi cela la réalité du métier d'actuaire.

Cette méthode finale apporte de nombreux avantages comparés à la méthode initiale (cf. partie IV). D'abord, elle fournit une meilleure compréhension du portefeuille à différentes échelles et délimite les zones risquées de manière plus précise en collant

à la réalité du risque sous-jacent. Surtout, elle propose différentes segmentations (et les tarifs correspondants) en vue de réduire l'effet d'antisélection. Cette même segmentation peut être intégrée aux tarifs de réassurance grâce à une méthode empirique (cf. partie VII). Ceci permet au réassureur de s'aligner avec son client mais aussi de l'encourager à optimiser son portefeuille. Enfin, soulignons le caractère général de cette méthode qui apporte une solution accessible et intéressante à de nombreux sujets, en agriculture bien sûr, mais aussi dans n'importe quelle ligne de business où la dimension géographique joue un rôle.

En contrepartie, elle est évidemment un peu plus complexe. Du temps a été nécessaire pour la développer mais également pour la présenter et la justifier notamment auprès du client et indirectement auprès du ministère qui doit lui aussi approuver ce changement de méthode ainsi que la multiplication des tarifs ; ce qui apporte une dimension politique au problème et rend la pédagogie d'autant plus importante.

Bibliographie

1. Agricultural insurance: products and schemes. *Atlas Magazine Insurance nexs around the world.* [Online] 06 2017. https://www.atlas-mag.net/en/article/agricultural-insurance-products-and-schemes.

2. *Least square quantization in PCM.* Lloyd, S. P. 2, 1957, IEEE Transactions on Information Theory, Vol. 28, pp. 129-137.

3. K-moyennes. *Wikipedia.* [Online] 02 01, 2022. https://fr.wikipedia.org/wiki/K-moyennes#Algorithme_classique.

4. *Cluster analysis of multivariate data: efficiency versus interpretability of classifications.* Forgy, E.W. 1965, Biometrics, Vol. 21, pp. 768-769. JSTOR 2528559.

5. *Some methods for classification and analysis of multivariate observations.* MacQueen, J. Berkeley : s.n., 1967, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297.

6. *Algorithm AS 136: A K-means clustering algorithm.* Hartigan, J. A. and Wong, M. A. 1, 1979, Applied Statistics, Vol. 28, pp. 100-108.

7. *k means++: The Advantages of Careful Seeding.* Arthur, David and Sergei, Vassililvitskii. Philadephia : s.n., 2007, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027-1035.

8. *Survey of Clustering Data Mining Techniques.* Berkhin, Pavel. 2006, Grouping Multidimensional Data. Springer.

9. Allied Market Research. *Crop Insurance Market by Coverage and Distribution Channel : Global Opportunity Analysis and Industry Forecast, 2020–2027.* 2020. 5976955.

Table des figures

Tableaux

Annexe : code R utilisé pour l'analyse

Par soucis de confidentialité, le code n'est pas divulgué dans son intégralité. Seules les parties essentielles à la compréhension du mémoire sont exposées.

```r
# libraries and packages ----------------------------------------------------
library(tidyverse)
library(cartography)
library(ggplot2)
library(readxl)
rm(list=ls())

# Parameters ----------------------------------------------------------------

performance_test <- FALSE
show_details <- TRUE
respect_excel <- FALSE
cartography <- TRUE
significative_only <- FALSE
cereland <- sf::st_read("~/rdata/avatry/actuarial-thesis2/cereland.shp")%>% select(REGION = name_1, DEPARTEMENT = name_2)

first_year <- 2013
last_year  <- 2020
year_list <- first_year:last_year
ref_year <- 2020
old_deductible <- 0.25 #Rendement de référence X 75%
new_deductible <- 0.35 #Rendement de référence X 65%

threshold_years <- 8
if(threshold_years > length(year_list)){stop("WARNING threshold_years is too big")}
threshold_superficie <- 20000
threshold_capital <- 10e6

cluster_number <- 6

#Stop Loss structure
layer_number <- 3
limits <- c(0.3, 0.3, 0.1)
rets <- c(0.1, 0.4, 0.8)
```

```r
# Read and clean data ------------------------------------------------------------
data1220 <- readRDS("dataset_data1220.rds")
data21 <- readRDS("dataset_data21.rds")
data <- bind_rows(data1220,data21)
rm(list=c("data1220","data21"))

data <- rename(data, MONTANT_INDEMNISE = 'MONTANT INDEMNISE')
data <- rename(data, CAPITAL_ASSURE = 'CAPITAL ASSURE')
data$DEPARTEMENT <- str_replace(data$DEPARTEMENT,"-"," ")

data <- mutate(data, DEPARTEMENT = str_to_upper(DEPARTEMENT))
data <- mutate(data, CULTURE = str_to_upper(CULTURE))
data <- mutate(data, COMMUNE = str_to_upper(COMMUNE))
data <- mutate(data, ZONE_NIV = ifelse((ZONE == 3), paste(ZONE, NIVEAU, sep = "_"), ZONE))
#data <- mutate(data, ZONE_NIV = ZONE) #forget zone niv
data <- mutate(data, ZONE_NIV = factor(ZONE_NIV))

# Step0 As If Data ---------------------------------------------------------

#Deductible correction
k <- (1-old_deductible)/(1-new_deductible)
k2 <- 1-k
data <- mutate(data, MONTANT_CORRIGE = k*MONTANT_CORRIGE + k2*CAPITAL_ASSURE)
data <- mutate(data, MONTANT_CORRIGE = ifelse(MONTANT_CORRIGE>0,MONTANT_CORRIGE,0))
```

```r
# Step1 Clustering ----------------------------------------------------------

#LC per departement and year
if (show_details){
  LC_departement_year <- data %>%
    select(DEPARTEMENT, YEAR, SUPERFICIE, MONTANT_CORRIGE, CAPITAL_ASSURE) %>%
    group_by(DEPARTEMENT, YEAR) %>%
    summarise( SUPERFICIE_TOTALE = sum(SUPERFICIE, na.rm = TRUE), LOSS = sum(MONTANT_CORRIGE, na.rm = TRUE), SI = sum(CAPITAL_ASSURE, na.rm = TRUE)) %>%
    mutate(LC = LOSS/SI)
  LC_departement_year_display <- LC_departement_year %>%
    select(YEAR, DEPARTEMENT, LC) %>%
    pivot_wider(names_from = YEAR, values_from = LC) %>%
    select(DEPARTEMENT, num_range("", first_year:last_year))
}

#LC per departement
LC_departement <- data %>%
  select(DEPARTEMENT, YEAR, SUPERFICIE, MONTANT_CORRIGE, CAPITAL_ASSURE) %>%
  group_by(DEPARTEMENT) %>%
  summarise(SUPERFICIE_TOTALE = sum(SUPERFICIE, na.rm = TRUE), LOSS = sum(MONTANT_CORRIGE, na.rm = TRUE), SI = sum(CAPITAL_ASSURE, na.rm = TRUE)) %>%
  mutate(LC = LOSS/SI)
years_departement <- data %>% select(DEPARTEMENT, YEAR) %>% distinct() %>% group_by(DEPARTEMENT) %>% summarise(YEARS = n())
LC_departement <- full_join(LC_departement, years_departement, by = "DEPARTEMENT")

#Defining significative departement
LC_departement <- mutate(LC_departement, SIGNIFICATIVE = (YEARS >= seuil_years)&(SUPERFICIE_TOTALE >= seuil_superficie)&(SI >= seuil_capital))
LC_departement_significatif <- LC_departement %>% filter(SIGNIFICATIVE == TRUE)
#Fill data
data <- left_join(data, select(LC_departement, DEPARTEMENT, SIGNIFICATIVE), by = "DEPARTEMENT")

#K means
LC_departement_significatif <- arrange(LC_departement_significatif, LC)
clustering <- kmeans(LC_departement_significatif$LC, cluster_number, iter.max = 1000, nstart = 100000)
#Reordering clusters
cluster_reordering <- data.frame(original = clustering$cluster) %>% distinct() %>% bind_cols(data.frame(CLUSTER = 1:cluster_number))
LC_departement_significatif <- mutate(LC_departement_significatif, original = clustering$cluster)
LC_departement_significatif <- left_join(LC_departement_significatif, cluster_reordering, by ="original")
# Summary
cluster_departement_significatif <- select(LC_departement_significatif, DEPARTEMENT, CLUSTER)
```

```r
# Step 2 Cluster Extrapolation ---------------------------------------------

#read and clean data
cluster_extrapolation <- read_xlsx("~/rdata/avatry/actuarial-thesis2/Extrapolation.xlsx",sheet = 1) #filled manually
cluster_extrapolation <- mutate(cluster_extrapolation, DEPARTEMENT = str_to_upper(DEPARTEMENT))
cluster_extrapolation$DEPARTEMENT <- str_replace(cluster_extrapolation$DEPARTEMENT,"-"," ")

# comparing with original clustering
cluster_extrapolation <- full_join(cluster_departement_significatif, cluster_extrapolation, by= "DEPARTEMENT")
cluster_extrapolation <- rename(cluster_extrapolation, ORIGINAL_CLUSTER = CLUSTER)
cluster_extrapolation <- mutate(cluster_extrapolation, CHECK = (ORIGINAL_CLUSTER == CLUSTERING_EXTRAPOLATED))
# checking that the extrapolation respects the original proposition
check <- nrow(filter(cluster_extrapolation, CHECK == FALSE)) == 0
if (!check) {warning("WARNING : The proposition of extrapolation doesn't respect the original clustering for significative departements")}
# Combining both information
cluster_extrapolation <- mutate(cluster_extrapolation, SIGNIFICATIVE = !is.na(ORIGINAL_CLUSTER))
cluster_extrapolation <- mutate(cluster_extrapolation, CLUSTER = ifelse(SIGNIFICATIVE, ORIGINAL_CLUSTER, CLUSTERING_EXTRAPOLATED))

# Fill data
cluster_departement <- select(cluster_extrapolation, DEPARTEMENT, CLUSTER)
data <- left_join(data, cluster_departement, by = "DEPARTEMENT")

#Cartography

if(cartography){
  #map selection cereland
  map <- cereland
  departements <- data %>% select(DEPARTEMENT) %>% distinct
  map <- right_join(map, departements, by = "DEPARTEMENT")

  #cleaning plot
  if(!is.null(dev.list())) {dev.off()}

  #significative departements
  jpeg(file="clusters_significative_departements.jpeg",width = 1200, 1200)
  map1 <- full_join(map, mutate(cluster_departement_significatif, ORIGINAL_CLUSTER = CLUSTER), by ="DEPARTEMENT")
  par(mar = c(0,0,0,0))
  mypal <- mypal <- carto.pal(pal1 = "green.pal", n1 = 2, pal2 = "orange.pal", n2 = 4)
  choroLayer(x = map1, var ="ORIGINAL_CLUSTER", breaks = (1:(cluster_number+1)), col = mypal)
  #labelLayer(x = map1, txt = "DEPARTEMENT", cex = 0.7, halo = "TRUE", bg = "white")
  dev.off()

  #all departements
  jpeg(file="clusters_departements.jpeg",width = 1200, 1200)
  map2 <- full_join(map, cluster_extrapolation, by ="DEPARTEMENT")
  par(mar = c(0,0,0,0))
  mypal <- mypal <- carto.pal(pal1 = "green.pal", n1 = 2, pal2 = "orange.pal", n2 = 4)
  choroLayer(x = map2, var ="CLUSTER", breaks = (1:(cluster_number+1)), col = mypal)
  #labelLayer(x = map2, txt = "DEPARTEMENT", cex = 0.7, halo = "TRUE", bg = "white")
  dev.off()
}
```

```r
# Step 3 Computing pure rates per cluster and zone ------------------------

# Step 3_1 LC per Cluster and zone_niv ----------------------------------------

#LC per cluster and ZONE_NIV and year
if (show_details){
  LC_cluster_zone_niv_year <- data %>%
    select(ZONE_NIV, CLUSTER, YEAR, SUPERFICIE, MONTANT_CORRIGE, CAPITAL_ASSURE) %>%
    group_by(ZONE_NIV, CLUSTER, YEAR) %>%
    summarise( SUPERFICIE_TOTALE = sum(SUPERFICIE, na.rm = TRUE), LOSS = sum(MONTANT_CORRIGE, na.rm = TRUE), SI = sum(CAPITAL_ASSURE, na.rm = TRUE)) %>%
    mutate(LC_CLUSTER = LOSS/SI)
  LC_cluster_zone_niv_year_display <- LC_cluster_zone_niv_year %>%
    select(YEAR, ZONE_NIV, CLUSTER, LC_CLUSTER) %>%
    pivot_wider(names_from = YEAR, values_from = LC_CLUSTER) %>%
    select(ZONE_NIV, CLUSTER, num_range("", first_year:last_year))
}

#LC per cluster and ZONE_NIV
LC_cluster_zone_niv <- data %>%
  select(ZONE_NIV, CLUSTER, YEAR, SUPERFICIE, MONTANT_CORRIGE, CAPITAL_ASSURE) %>%
  group_by(ZONE_NIV, CLUSTER) %>%
  summarise(SUPERFICIE_TOTALE = sum(SUPERFICIE, na.rm = TRUE), LOSS = sum(MONTANT_CORRIGE, na.rm = TRUE), SI = sum(CAPITAL_ASSURE, na.rm = TRUE)) %>%
  mutate(LC_CLUSTER = LOSS/SI)
years_cluster_zone_niv <- data %>% select(ZONE_NIV, CLUSTER, YEAR) %>% distinct() %>% group_by(ZONE_NIV, CLUSTER) %>% summarise(YEARS = n())
LC_cluster_zone_niv <- full_join(LC_cluster_zone_niv, years_cluster_zone_niv, by = c("CLUSTER","ZONE_NIV"))
```

```r
# Step 3_2 LC per Departement and zone_niv -------------------------------------

#LC per departement and ZONE_NIV and year
if (show_details){
  LC_departement_zone_niv_year <- data %>%
    select(ZONE_NIV, DEPARTEMENT, YEAR, SUPERFICIE, MONTANT_CORRIGE, CAPITAL_ASSURE) %>%
    group_by(ZONE_NIV, DEPARTEMENT, YEAR) %>%
    summarise( SUPERFICIE_TOTALE = sum(SUPERFICIE, na.rm = TRUE), LOSS = sum(MONTANT_CORRIGE, na.rm = TRUE), SI = sum(CAPITAL_ASSURE, na.rm = TRUE)) %>%
    mutate(LC_COMP = LOSS/SI)
  LC_departement_zone_niv_year_display <- LC_departement_zone_niv_year %>%
    select(YEAR, ZONE_NIV, DEPARTEMENT, LC_COMP) %>%
    pivot_wider(names_from = YEAR, values_from = LC_COMP) %>%
    select(ZONE_NIV, DEPARTEMENT, num_range("", first_year:last_year))
}

#LC per departement and ZONE_NIV
#LC_COMP
LC_departement_zone_niv <- data %>%
  select(ZONE_NIV, DEPARTEMENT, YEAR, SUPERFICIE, MONTANT_CORRIGE, CAPITAL_ASSURE) %>%
  group_by(ZONE_NIV, DEPARTEMENT) %>%
  summarise(SUPERFICIE_TOTALE = sum(SUPERFICIE, na.rm = TRUE), LOSS = sum(MONTANT_CORRIGE, na.rm = TRUE), SI = sum(CAPITAL_ASSURE, na.rm = TRUE)) %>%
  mutate(LC_COMP = LOSS/SI)
years_departement_zone_niv <- data %>% select(ZONE_NIV, DEPARTEMENT, YEAR) %>% distinct() %>% group_by(ZONE_NIV, DEPARTEMENT) %>% summarise(YEARS = n())
LC_departement_zone_niv <- full_join(LC_departement_zone_niv, years_departement_zone_niv, by = c("DEPARTEMENT","ZONE_NIV"))
#LC_EST
LC_departement_zone_niv <- mutate(LC_departement_zone_niv, SIGNIFICATIVE = (YEARS >= seuil_years)&(SUPERFICIE_TOTALE >= seuil_superficie))
LC_departement_zone_niv <- left_join(LC_departement_zone_niv, cluster_departement, by ="DEPARTEMENT")
LC_departement_zone_niv <- left_join(LC_departement_zone_niv, select(LC_cluster_zone_niv, ZONE_NIV, CLUSTER, LC_CLUSTER), by = c("ZONE_NIV","CLUSTER"))
LC_departement_zone_niv <- mutate(LC_departement_zone_niv, LC_EST = ifelse(SIGNIFICATIVE, LC_COMP, LC_CLUSTER))
```

```r
# Step 3_3 Computing rates ------------------------------------------------

#Add LC_EST et EST_LOSS to the data
data <- left_join(data, select(LC_departement_zone_niv, DEPARTEMENT, ZONE_NIV, LC_EST), by = c("DEPARTEMENT","ZONE_NIV"))
data <- mutate(data, EST_LOSS = LC_EST*CAPITAL_ASSURE)

#Establishing rates and si
rates <- data %>%
  select(CLUSTER, ZONE_NIV,EST_LOSS, CAPITAL_ASSURE) %>%
  group_by(CLUSTER, ZONE_NIV) %>%
  summarise(LOSS = new_capital_coef*sum(EST_LOSS, na.rm = TRUE), SI = new_capital_coef*sum(CAPITAL_ASSURE, na.rm = TRUE)) %>%
  mutate(LC = LOSS/SI)
rates_total <- sum(rates$LOSS, na.rm = TRUE)/sum(rates$SI, na.rm = TRUE)
rates_cluster <- rates %>%
  group_by(CLUSTER) %>%
  summarise(LOSS = sum(LOSS, na.rm = TRUE), SI = sum(SI, na.rm = TRUE)) %>%
  mutate(LC = LOSS/SI)
rates_zone_niv <- rates %>%
  group_by(ZONE_NIV) %>%
  summarise(LOSS = sum(LOSS, na.rm = TRUE), SI = sum(SI, na.rm = TRUE)) %>%
  mutate(LC = LOSS/SI) %>%
  select(ZONE_NIV, LC) %>%
  pivot_wider(names_from = ZONE_NIV, values_from = LC) %>%
  bind_cols(data.frame(CLUSTER = sum(1:cluster_number*10^rev(1:cluster_number-1)), TOTAL = rates_total))
rates_display <- rates %>%
  select(CLUSTER, ZONE_NIV, LC) %>%
  pivot_wider(names_from = ZONE_NIV, values_from = LC) %>%
  bind_cols(select(rates_cluster, TOTAL = LC)) %>%
  bind_rows(rates_zone_niv)
si_zone_niv <- rates %>%
  group_by(ZONE_NIV) %>%
  summarise(SI = sum(SI), na.rm = TRUE) %>%
  select(ZONE_NIV, SI) %>%
  pivot_wider(names_from = ZONE_NIV, values_from = SI) %>%
  bind_cols(data.frame(CLUSTER = sum(1:cluster_number*10^rev(1:cluster_number-1)), TOTAL = sum(rates$SI)))
si_display <- rates %>%
  select(CLUSTER, ZONE_NIV, SI) %>%
  pivot_wider(names_from = ZONE_NIV, values_from = SI) %>%
  bind_cols(select(rates_cluster, TOTAL = SI)) %>%
  bind_rows(si_zone_niv)
```

```r
#Establishing rates and si (significative departement only)
if (show_details){
  rates_significative <- data %>%
    filter(SIGNIFICATIVE == TRUE) %>%
    select(CLUSTER, ZONE_NIV,EST_LOSS, CAPITAL_ASSURE) %>%
    group_by(CLUSTER, ZONE_NIV) %>%
    summarise(LOSS = new_capital_coef*sum(EST_LOSS, na.rm = TRUE), SI = new_capital_coef*sum(CAPITAL_ASSURE, na.rm = TRUE)) %>%
    mutate(LC = LOSS/SI)
  rates_total_significative <- sum(rates_significative$LOSS)/sum(rates_significative$SI)
  rates_cluster_significative <- rates_significative %>%
    group_by(CLUSTER) %>%
    summarise(LOSS = sum(LOSS, na.rm = TRUE), SI = sum(SI, na.rm = TRUE)) %>%
    mutate(LC = LOSS/SI)
  rates_zone_niv_significative <- rates_significative %>%
    group_by(ZONE_NIV) %>%
    summarise(LOSS = sum(LOSS, na.rm = TRUE), SI = sum(SI, na.rm = TRUE)) %>%
    mutate(LC = LOSS/SI) %>%
    select(ZONE_NIV, LC) %>%
    pivot_wider(names_from = ZONE_NIV, values_from = LC) %>%
    bind_cols(data.frame(CLUSTER = sum(1:cluster_number*10^rev(1:cluster_number-1)), TOTAL = rates_total_significative))
  rates_significative_display <- rates_significative %>%
    select(CLUSTER, ZONE_NIV, LC) %>%
    pivot_wider(names_from = ZONE_NIV, values_from = LC) %>%
    bind_cols(select(rates_cluster_significative, TOTAL = LC)) %>%
    bind_rows(rates_zone_niv_significative)
  si_zone_niv_significative <- rates_significative %>%
    group_by(ZONE_NIV) %>%
    summarise(SI = sum(SI), na.rm = TRUE) %>%
    select(ZONE_NIV, SI) %>%
    pivot_wider(names_from = ZONE_NIV, values_from = SI) %>%
    bind_cols(data.frame(CLUSTER = sum(1:cluster_number*10^rev(1:cluster_number-1)), TOTAL = sum(rates_significative$SI)))
  si_significative_display <- rates_significative %>%
    select(CLUSTER, ZONE_NIV, SI) %>%
    pivot_wider(names_from = ZONE_NIV, values_from = SI) %>%
    bind_cols(select(rates_cluster_significative, TOTAL = SI)) %>%
    bind_rows(si_zone_niv_significative)
}
```