Mémoire présenté devant l'ENSAE Paris

pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires le 10/11/2023

Par : **Youssouf BANCE**

Titre : **Solution d'assurance indicielle beau temps contre les aléas climatiques en camping vacance : Tarification et impact du changement climatique**

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

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

Membres présents du jury de la filière

Entreprise : GALEA & Associés

Nom : Nicolas baradel

Signature :

Directeur du mémoire en entreprise :

Nom : Sergio OROZCO

Signature :

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

Signature du responsable entreprise

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Face aux événements climatiques de plus en plus extrêmes, les clubs de campings sont confrontés à de nombreux défis pour adapter leur offre et maintenir leur attractivité. L'assurance paramétrique apparaît comme une solution idéale parmi celles qui permettent aux clubs de campings de conserver leur niveau d'attractivité. Cependant, les produits d'assurance pour les campings en France métropolitaine restent, à ce jour, limités.

L'objectif de ce mémoire est de présenter un régime fictif nommé « beau temps » qui proposera une solution d'assurance indicielle contre les aléas climatiques (température, pluie et vent) permettant d'indemniser une victime du mauvais temps en camping. Il s'agira également dans ce mémoire d'évaluer l'impact du changement climatique sur le régime à l'horizon 2100. Pour atteint cet objectif ce mémoire exploite, dans la phase de conception, les données de l'INSEE sur les campings vacances en France métropolitaine ainsi que les données historiques de Météo France pour définir dans un premier temps les indices climatiques des trois risques (température, pluie et vent), faire la tarification du produit et mettre en place une grille tarifaire. Le suivi du ratio de rentabilité S/P sur la phase de déroulement à permis de s'assurer de la fiabilité du régime.

L'analyse de la sinistralité du régime « beau temps » à l'aide d'outils statistiques et de la Théorie des valeurs extrêmes révèle un caractère extrême de cette sinistralité. Par conséquent, ce mémoire propose une modélisation spécifique à travers l'arbre de régression pareto généralisée, permettant de prendre en considération ce caractère extrême lors de la projection du régime « beau temps ».

Afin d'évaluer l'impact du changement climatique sur le régime à l'horizon 2100, en fonction des scénarios du GIEC (RCP 2.6 et 8.5), les indicateurs du régime (ratio S/P, réserves) sont projetés en utilisant deux approches distinctes : l'approche déterministe et l'approche par modélisation. La première repose uniquement sur les données de projection du DRIAS et ne fait appel à aucun modèle. En revanche, la seconde intègre le modèle de l'arbre de régression Pareto pour la modélisation des sinistres extrêmes. Les résultats de la projection révèlent une forte volatilité des ratios S/P projetés. Les résultats suggèrent une situation généralement défavorable pour les données climatiques basées sur le scénario RCP 8.5 (pessimiste), et cette tendance est encore plus prononcée avec l'approche de modélisation.

Mots-clés : *Camping vacance, assurance paramétrique, « beau temps », température, précipitation, vitesse du vent, Théorie des valeurs extrêmes, Apprentissage automatique, Arbre de décision Pareto généralisée, GIEC, changement climatique.*

Abstract

In the face of increasingly extreme weather events, camping clubs are confronted with numerous challenges in adapting their offerings and maintaining their attractiveness. Parametric insurance appears as an ideal solution among those that allow camping clubs to preserve their level of attractiveness. However, insurance products for campsites in metropolitan France are still limited to date.

The objective of this thesis is to present a fictional scheme called "fair weather" that will offer an index-based insurance solution against climatic risks (temperature, rain, and wind), providing compensation to victims of bad weather during camping. This thesis also aims to evaluate the impact of climate change on the scheme by the year 2100. To achieve this goal, the thesis utilizes, in the design phase, data from INSEE on vacation campsites in metropolitan France, as well as historical data from Météo France. Initially, this data is used to define climatic indices for the three risks (temperature, rain, and wind), set the pricing for the product, and establish a pricing grid. Monitoring the S/P profitability ratio during the implementation phase ensures the reliability of the scheme.

The analysis of the claims experience for the "fair weather" scheme using statistical tools and Extreme Value Theory reveals an extreme nature of these claims. Consequently, this thesis proposes a specific modeling approach through the generalized Pareto regression tree, allowing for the consideration of this extreme nature when projecting the "fair weather" scheme.

To assess the impact of climate change on the scheme by the year 2100, considering the IPCC scenarios (RCP 2.6 and 8.5), the regime's indicators (S/P ratio, reserves) are projected using two distinct approaches : the deterministic approach and the modeling approach. The first relies solely on DRIAS projection data and does not involve any modeling. In contrast, the second incorporates the Pareto regression tree model for extreme loss modeling. The projection results reveal significant volatility in the projected S/P ratios. The findings suggest a generally unfavorable situation for climate data based on the RCP 8.5 scenario (pessimistic), and this trend is even more pronounced with the modeling approach.

Keywords : *Camping holliday, parametric insurance, « good weather », temperature, precipitation, wind speed, Extreme Value Theory, Machine Learning, Generalized Pareto Decision Tree, IPCC, climate change.*

Remerciements

« *Ressentir de la gratitude et ne pas l'exprimer, c'est comme emballer un cadeau et ne pas le donner.* »

William Arthur Ward

Je tiens à remercier le cabinet de conseil en actuariat Galéa et associés pour l'opportunité qui m'a été donnée de travailler dans un cadre agréable et motivant.

Plus particulièrement, je souhaite exprimer ma gratitude envers Norbert Gautron, président de GALEA, qui m'a permis de réaliser mon stage de fin d'étude au sein de cabinet de conseil.

Je suis également reconnaissant envers Nicolas Baradel pour son accompagnement depuis l'ENSAE Paris.

Je souhaite également exprimer ma gratitude à l'ensemble des stagiaires, alternants, consultants, managers et associés pour leur disponibilité et leur bonne humeur. Je remercie l'ensemble de l'équipe pédagogique de l'ENSAE Paris pour la qualité de la formation et de l'enseignement dispensés.

Enfin, un grand merci à mes parents, mes frères et soeurs, pour leur soutien infaillible et leurs encouragements tout au long de ma scolarité.

Note de synthèse

Introduction

Le tourisme est essentiel pour l'économie française, contribuant à 7.13% du PIB en 2016, avec 4.85% de la consommation des visiteurs nationaux et 2.28% de la consommation des visiteurs étrangers. Le tourisme englobe diverses activités lors de voyages à des endroits hors de l'environnement habituel, pour des motifs de loisirs, d'affaires, etc. Les campings sont prisés pour les vacances en famille.

Le changement climatique, marqué par une augmentation des phénomènes météorologiques extrêmes, a un impact sur le tourisme. Le GIEC prévoit plus de jours anormalement chauds dans un avenir proche. Les campeurs souffrent de ces conditions météorologiques extrêmes. Certains campings offrent des "garanties Soleil" pour contrer les vacances pluvieuses. Cependant, ces garanties sont limitées, ne couvrant que certains risques météorologiques, principalement la pluie. Les risques tels que les vagues de chaleur et les vents forts peuvent également affecter l'attrait des campings.

L'objectif de ce mémoire est de présenter un régime que l'on nommera régime « **beau temps** » qui proposera une solution d'assurance indicielle contre les aléas climatiques (température, pluie et vent) permettant d'indemniser une victime du mauvais temps en camping. Ce mémoire aura également pour objectif d'évaluer l'impact du changement climatique sur ce régime.

Pour atteindre cet objectif, les études de ce mémoire seront subdivisées en deux parties : tout d'abord, la mise en place du régime fictif, notamment la description de son fonctionnement et la création d'une grille tarifaire. Par la suite, l'implémentation de deux approches distinctes permettront d'évaluer l'impact du changement climatique sur les indicateurs clés du régime (ratio S/P et réserves).

Fonctionnement du régime, tarification et analyses

Le régime « beau temps » permet de se couvrir contre trois risques climatiques que sont : la température, la pluie et le vent.

Pour le risque de température, le régime adopte deux indicateurs de température : une à la baisse (Tb) et l'autre à la hausse (Th). Le premier est relatif aux vagues de chaleur et le deuxième aux périodes de vagues de froid. Ces indicateur pour le jour i sont :

$$Th(i) = max(T_{moyenne}(i) - T_{saison}(i), 0) \quad et \quad Tb(i) = -min(T_{moyenne}(i) - T_{saison}(i), 0) \qquad (1)$$

avec $T_{moyenne}(i)$: Température moyenne journalière du jour i ; $T_{saison}(i)$: Température moyenne de la saison du jour i sur la décennie précédente. Par exemple, pour les jours du mois de janvier 2020, $T_{saison}(i)$ représentera la moyenne des température en période hivernale entre 2001 et 2011.

Pour le risque vent, l'indicateur associé est noté Ph est donnée pour le jour i par :

$$Vh(i) = V(i) - V_{saison}(i) \tag{2}$$

avec $V(i)$: Vitesse de vent moyenne (m/s) qu'il a fait je jour i ; $V_{saison}(i)$: Vitesse de vent moyenne de la saison du jour i sur la décennie précédente.

Quant au risque de pluie l'indicateur associé est noté Ph et est donnée pour le jour i par :

$$Ph(i) = P(i) - P_{saison}(i) \tag{3}$$

avec $P(i)$: Précipitation moyenne du jour i ; $P_{saison}(i)$: Précipitation moyenne de la saison du jour i sur la décennie précédente.

Sur le modèle d'une assurance paramétrique, le régime proposerait un remboursement de 50 €, en cas de dépassement d'un ou plusieurs indices associés aux risques de température, vent et pluie pour un individu, et par nuitée, ayant souscrit au contrat. Le seuil pour un indice particulier est son quantile 95% observé sur la période décennale passée. Pour souscrire au contrat de ce régime, l'individu paiera une prime qui couvrira une partie ou la totalité de son séjour en camping.

Si on désigne par $\mathbb{1}_T(i)$ la fonction indicatrice de dépassement (1 en cas de dépassement 0 sinon) associée au risque température, $\mathbb{1}_P(i)$ celle associée au risque pluie et $\mathbb{1}_V(i)$ celle associée au risque vent, la fonction d'indemnisation totale pour pour un individu, et par nuitée, ayant souscrit au contrat s'écrit de la manière suivante :

$$I_{totale}(i) = 50 \times max(\mathbb{1}_T(i), \mathbb{1}_V(i), \mathbb{1}_P(i)) \tag{4}$$

La mise en place d'une grille tarifaire adaptée à la description du fonctionnement du régime « beau temps » a nécessité l'utilisation des données sur les nuitées de campings issues des enquêtes sur la fréquentation dans l'hôtellerie de plein air réalisées par l'INSEE et les données historiques de Météo France (SYNOP). Plusieurs étapes de retraitement ont permis d'obtenir des données journalières pour chaque région, comprenant les variables climatiques d'intérêt et le nombre de nuitées. Les données SYNOP couvrent les années 2001 à 2021, tandis que les données sur le nombre de nuitées couvrent les années 2011 à 2021.

La grille tarifaire, définie par région et par mois, a été obtenue en utilisant l'approche du coût moyen attendu pondéré par les nuitées sur la période 2011 à 2015. Il s'agit de faire la moyenne des indemnités historiques potentielles qui auraient été versées par la structure du

contrat au cours de la période 2011 à 2015, par région et par mois pondéré par les nuitées (Nu). La prime pure pour le mois m et la région est donnée par la formule suivante :

$$\pi_{mois,region} = \frac{\sum_j I(m,r,j) * Nu(m,r,j)}{\sum_j Nu(m,r,j)} \quad (5)$$

La figure 1 présente la grille tarifaire du régime « beau temps ».

Grille tarifaire													
	Ile-de-France	Centre-Val de Loire	Bourgogne-Franche-Comte	Normandie	Hauts-de-France	Grand Est	Pays de la Loire	Bretagne	Nouvelle-Aquitaine	Occitanie	Auvergne-Rhone-Alpes	Provence-Alpes-Cote d Azur	Corse
Janvier	3.32 €	2.68 €	10.31 €	6.03 €	12.35 €	8.74 €	5.11 €	11.61 €	2.29 €	2.26 €	7.43 €	5.06 €	4.91 €
Février	4.93 €	7.19 €	11.22 €	7.25 €	7.88 €	9.30 €	6.93 €	10.82 €	6.85 €	5.40 €	9.00 €	5.25 €	4.67 €
Mars	11.95 €	13.14 €	13.14 €	20.47 €	21.66 €	23.11 €	10.64 €	13.93 €	6.89 €	8.16 €	11.42 €	10.97 €	6.48 €
Avril	4.13 €	6.83 €	5.08 €	7.17 €	5.55 €	2.92 €	5.65 €	9.80 €	3.46 €	4.83 €	4.45 €	10.39 €	2.55 €
Mai	2.99 €	4.46 €	4.84 €	2.92 €	6.25 €	3.59 €	4.60 €	8.04 €	2.16 €	3.07 €	4.39 €	7.58 €	4.00 €
Juin	9.84 €	4.35 €	7.65 €	9.78 €	9.10 €	3.11 €	4.66 €	12.22 €	6.55 €	5.69 €	3.84 €	8.43 €	1.94 €
Juillet	8.05 €	6.30 €	10.03 €	9.41 €	10.43 €	5.84 €	5.12 €	10.50 €	4.02 €	6.48 €	8.60 €	18.69 €	11.87 €
Août	6.79 €	4.86 €	6.46 €	7.38 €	7.78 €	4.22 €	5.18 €	10.12 €	6.66 €	5.70 €	6.38 €	13.17 €	8.39 €
Septembre	2.36 €	2.47 €	3.62 €	3.52 €	4.53 €	0.94 €	2.35 €	4.55 €	2.35 €	1.69 €	4.87 €	5.59 €	3.05 €
Octobre	3.38 €	3.49 €	7.52 €	8.81 €	5.13 €	6.33 €	4.42 €	7.96 €	3.07 €	3.50 €	7.40 €	8.91 €	5.84 €
Novembre	8.36 €	7.72 €	16.81 €	15.34 €	12.53 €	14.22 €	7.10 €	15.64 €	7.99 €	9.54 €	11.42 €	14.46 €	5.02 €
Décembre	3.85 €	3.40 €	6.89 €	12.07 €	10.03 €	5.92 €	7.04 €	14.22 €	4.81 €	1.25 €	2.71 €	2.93 €	1.55 €

Figure 1 – Prime pure du régime « **beau temps** ».

A l'aide de cette grille tarifaire, et en supposant que le nombre d'assurées qui achèterait le produit (part de marché) est une proportion variable du nombre de nuitées on obtient les ratios S/P sur les années 2016 à 2021 (figure ci-dessous).

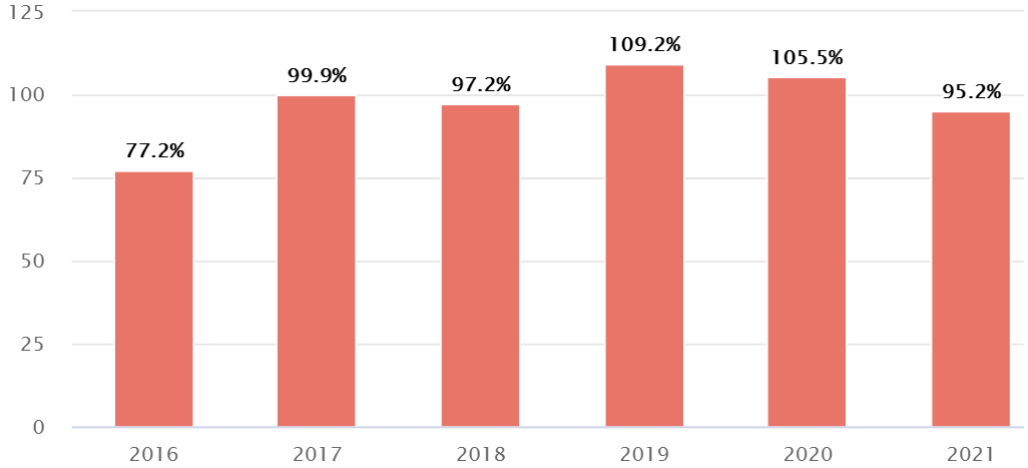


Figure 2 – Évolution du ratio S/P du régime (2016 - 2021).

Le ratio S/P du régime « beau temps » reste satisfaisant au cours des trois premières années (2016, 2017, 2018), demeurant inférieur à 100%. Cependant, une détérioration du ratio S/P est observée pour les années 2019 et 2020, affichant respectivement des ratios S/P de 109.2% et 105.5%. Cette évolution nous incite à envisager une révision tarifaire afin de prendre en compte la sinistralité des années plus récentes. Cette révision tarifaire a conduit à une augmentation des tarifs de 10% à partir de l'année 2021.

L'analyse de la sinistralité découlant de la mise en oeuvre du régime, à l'aide d'outils de la théorie des valeurs extrêmes et des statistiques, met en évidence une sinistralité extrême appartenant au domaine d'attraction de Fréchet. Le tableau ci-dessous présentes les résultats des différents méthodes de détermination du seuil des extrêmes u.

	Mean excess plot	Hill Plot	Gestengarbe plot	Minimisation AMSE
Seuil u	125 000€ 180 000€	125 000€ 170 000€ 260 000€	132 200€	131 000€

Table 1 – Synthèse des seuils par méthode de détermination

la plupart des seuils obtenus sont proches de 130 000€ . En nous basant sur l'opinion d'experts, nous avons conclu que le seuil de **130 000€** apparaît de manière consistante dans la plupart des résultats obtenus, ce qui en fait un choix approprié pour distinguer les sinistres extrêmes des sinistres non-extrêmes. Avec ce seuil, les sinistres extrêmes représentent 5.7% de l'ensemble de la sinistralité du régime « beau temps » y compris les valeurs nulles.

Au vu du caractère extrême de la sinistralité du régime, le modèle de l'arbre de régression pareto généralisée apparaît comme un modèle intéressant pour concevoir des classes de sinistres plus homogènes en fonction de variables explicatives intéressante pour une meilleure projection de la sinistralité extrême selon un horizon donné. La figure ci-dessous montre l'arbre obtenu à partir de la procédure de régression pareto géné- ralisée (GP CART).

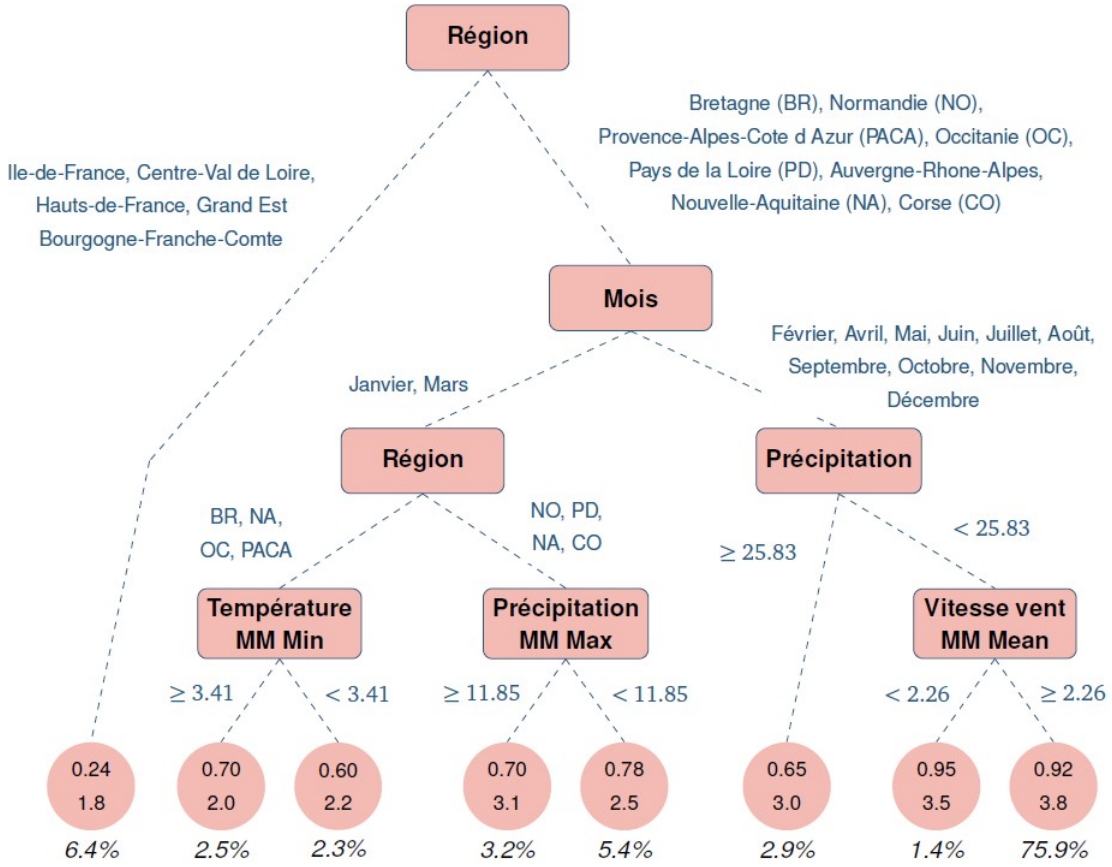


Figure 3 – Arbre de régression pareto généralisée (GP CART). Pour chaque feuille, la valeur du paramètre de forme ξ (première ligne) et le paramètre d'échelle σ à 10^{-5} (deuxième ligne) sont donnés.

L'arbre est composé (*leafs*) de 8 feuilles avec quatre (4) *splits* selon seulement 6 variables : la région, le mois, la précipitation journalière, le minimum de la température journalière sur un glissement de 7 jours (*MM Min*), le maximum de la précipitation journalière sur un glissement de 7 jours (*MM Max*) et la moyenne de la vitesse du vent en glissement de 7 jours (*MM Mean*).

Données, approches et résultats de la projection

Pour la projection du régime jusqu'en 2100, nous avons utilisé les données fournies par DRIAS. DRIAS met à disposition les résultats de projections issus de modèles climatiques développés dans différents laboratoires de modélisation du climat. Il est important de noter que l'ensemble des données DRIAS, y compris celles disponibles pour les périodes passées, ne sont pas des données météorologiques observées, mais plutôt des données simulées par ces modèles climatiques. Pour le traitement des données, nous avons agrégé les données journalières de la base DRIAS sur la période de 2023 à 2100, en utilisant les scénarios RCP 2.6 et 8.5, afin qu'elles correspondent à la résolution spatiale et temporelle choisie pour notre étude.

Ce mémoire adopte deux approches pour la projection du régime jusqu'en 2100 : *une approche déterministe* et *une approche basée sur la modélisation*. Dans les deux approches, il est essentiel de formuler un certain nombre d'hypothèses réalistes. Ces hypothèses portent principalement sur l'évolution de la population assurée et la grille tarifaire. En ce qui concerne l'hypothèse concernant l'évolution de la population assurée, elle est supposée représenter 25% des nuitées totales, dont on suppose qu'elle suit une croissance géométrique. Quant à l'hypothèse sur la grille tarifaire, on suppose qu'elle reste constante sur la période de projection et correspond à la dernière grille tarifaire ayant subi un redressement.

La première approche, appelée *déterministe*, consiste simplement à appliquer les règles du mécanisme du régime « **beau temps** » aux données de la base DRIAS en fonction des scénarios RCP 2.6 et RCP 8.5. Par exemple, en utilisant les données futures du RCP 2.6, nous recalculons les indices de température (Th), de précipitation (Ph) et de vitesse du vent (Vh). Si l'un de ces indices dépasse son seuil défini, le régime paiera une indemnité de 50 euros aux assurés. La prime totale et le sinistre total projetés pour une année donnée sont alors déterminés en utilisant les hypothèses formulées concernant l'évolution de la population assurée et la grille tarifaire.

L'approche par la *modélisation*, en revanche, suppose que le processus qui régit le fonctionnement du régime (déclenchement des paiements, indemnité,...) est aléatoire. Soit S la distribution du paiement (soit 0 en cas de non déclenchement ou Y positive strictement).

Alors l'espérance mathématique de S s'écrit :

$$E(S) = P(S > 0)\,[P(Y \geq u)E(Y|Y \geq u) + (1 - P(Y \geq u))E(Y|Y < u)] \tag{6}$$

Au vu de cette espérance de la distribution de S, nous considérons que la distribution de S est une variable de mélange de même distribution que $\tau(x)*[\delta(x)Z_1+(1-\delta(x))Z_2]$, avec :

- $\tau(x)$ et $\delta(x)$ sont des variables aléatoires de Bernoulli dépendant de X. $p_i(x) = P(\tau(x) = 1)$ et $m_i(x) = P(\delta(x) = 1)$ représentent respectivement la probabilité de déclenchement ($S \neq 0$) et la probabilité de dépassement du seuil des extrêmes u ;
- $Z_1 = Y|Y \geq u$, a une distribution qui est estimée à l'aide *l'arbre de pareto generalisée* (GP CART) de la figure 3 avec $u = 130\,000$€ ;
- $Z_2 = Y|Y < u$, sera modélisé à l'aide d'un GLM classique pour la sévérité. On considère $\tau \perp\!\!\!\perp (\delta, Z_1, Z_2)$ et $\delta \perp\!\!\!\perp (Z_1, Z_2)$.

τ est calibré à l'aide d'un modèle de forêt aléatoire (RF) avec une profondeur maximale de l'arbre 15 entraîné sur les données équilibrés à l'aide de la technique de rééchantillonnage (SMOTE+ENN). La technique du SMOTE+ENN vise à améliorer la capacité de généralisation d'un modèle tout en réduisant le déséquilibre de classe. Il commence par appliquer SMOTE pour générer des exemples synthétiques de la classe minoritaire, puis il applique ENN pour éliminer les exemples mal classés de la classe majoritaire. δ est calibré à l'aide d'un modèle de forêt aléatoire d'un approche que nous adopté dans le cadre de ce mémoire. Cette procédure consiste à estimer $\hat{y}$ par un modèle de foret aléatoire. Ensuite, les probabilités de dépassements p_i sont estimés en utilisant une distribution normale de moyenne $\hat{y}$ et d'écart type σ_y : $\mathcal{N}(\hat{y}, \sigma_y)$, avec σ_y égal à l'écart type de y sur les données d'entraînement.

$$p_i = 1 - CDF_{\mathcal{N}\left(\hat{y}_i, \sigma_{\hat{y}}^2\right)}(u) \tag{7}$$

Quant à la sinistralité attritionnelle $Y|Y < u$, elle est calibrée à l'aide d'un modèle de GLM Gamma avec fonction de lien logarithme. Enfin, la sinistralité extrême est projetée à l'aide de l'arbre de pareto généralisée (GP CART) et d'arbre de régression avec fonction de perte qu'on calibre sur chacune des 8 feuilles du GP CART. Cette procédure permet un gain significatif en terme de RMSE d'environ 15% par rapport à la procédure qui consisterait à appliquer un arbre de régression avec fonction de perte quadratique directement sur les données extrêmes.

Les sinistres sont ensuite estimés en calibrant chaque composante de ce mélange de modèles. Les primes, quant à elles, sont calculées en se basant sur les sinistres projetés et l'hypothèse relative à la grille tarifaire. L'intérêt de cette approche est de s'affranchir de l'hypothèse sur l'évolution de la population assurée. Un deuxième avantage est de prendre en compte, dans la projection, l'aspect extrême de la sinistralité du régime.

Les figures 4 et 5 présentent l'évolution des ratios S/P projetés selon les scénarios du GIEC (RCP 2.6 et RCP 8.5) respectivement selon l'approche déterministe et selon l'approche par modélisation.

De manière générale, nous avons constaté des conclusions assez similaires entre les deux approches, notamment en ce qui concerne les tendances. Dans les deux approches, les ra-

tios S/P projetés présentent une forte irrégularité, principalement due à la variabilité des données climatiques. On remarque une nette détérioration du ratio S/P en début de période de projection (2025 - 2028) dans l'approche déterministe, tandis que cette forte dégradation n'est pas observée dans l'approche basée sur la modélisation. Cette dégradation significative s'expliquerait principalement par un changement abrupt de la trajectoire climatique actuellement observée. De plus, au-delà de l'année 2063, dans le scénario 8.5, la situation devient très défavorable, avec des ratios qui dépassent souvent 100% et qui sont plus élevés que ceux du scénario optimiste 2.6 dans le cas de l'approche par modélisation. Enfin, les résultats indiquent une situation globalement défavorable pour les données climatiques issues du scénario RCP 8.5 (pessimiste) et est beaucoup plus marquée avec l'approche par modélisation.

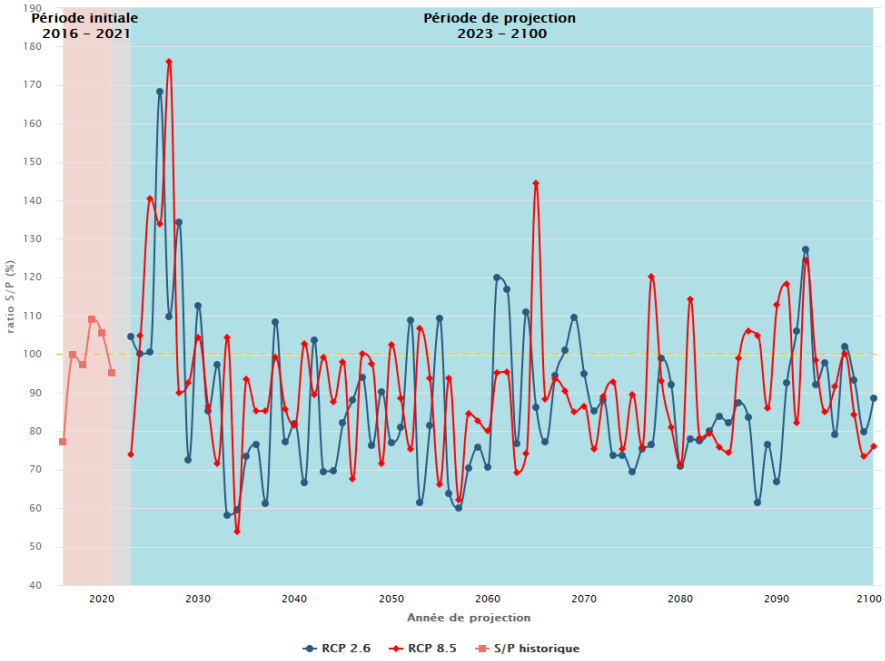


Figure 4 – S/P projetés selon les scénarios du GIEC et selon l'approche déterministe.

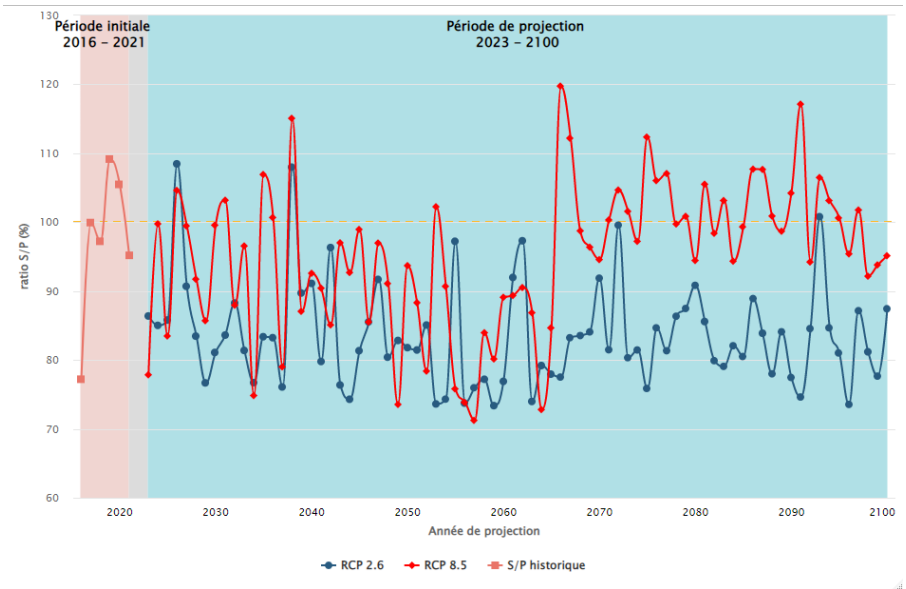


Figure 5 – S/P projetés selon les scénarios du GIEC et selon l'approche par modélisation.

Statistiques	*S/P historiques*	*Approche*			
		Déterministe		*Modélisation*	
		RCP 2.6	RCP 8.5	RCP 2.6	RCP 8.5
E(S/P)	97.4%	87.3%	91.9%	83.6%	95.1%
σ(S/P)	11.2%	19.1%	19.3%	7.5%	10.7%

Table 2 – Moyenne et écart-type des ratios S/P projetés et historiques.

Le tableau 2 met en évidence des S/P projetés moins volatiles dans l'approche par modélisation par rapport à l'approche déterministe.

On note une tendance linéaire à la hausse des réserves, quel que soit le scénario et l'approche considérée. Toutefois, les réserves dans le scénario RCP 8.5 sont inférieures à celles du scénario RCP 2.6 avec l'approche déterministe. Par contre, avec l'approche par modélisation, cette infériorité n'est plus observé au delà de 2085. Les réserves sont données par la formule :

$$R_n = R_{n-1} + \text{Résultat}(n) = R_0 + \sum_{i=2023}^{n} \text{Résultat}(i) \tag{8}$$

avec $\text{Résultat}(n) = P(n) + FG(n) - S(n)$, dont $P(n)$, $FG(n) = 0.1 * P(n)$ et $S(n)$ représentent respectivement la prime pure totale, les frais globaux et le sinistre total projetés pour l'année n et R_0 les réserves de départ égale à 55501059€ (somme des résultats de 2016 à 2021).

Conclusion

De manière générale, nous avons observé des conclusions assez similaires entre les deux approches, notamment en ce qui concerne les tendances des indicateurs retenus, à savoir le ratio S/P et les réserves. Dans les deux approches, les ratios S/P projetés présentent une forte irrégularité, principalement due à la variabilité climatique des données, tandis que les réserves affichent une tendance plutôt linéaire.

En termes de différences, nous avons remarqué que les ratios S/P étaient plutôt sous-estimés avec l'approche par modélisation, tandis qu'avec l'approche déterministe, nous avons observé une forte dégradation du ratio S/P en début de période de projection (2023 - 2028). De plus, les résultats indiquent une situation globalement défavorable pour les données climatiques issues du scénario RCP 8.5 (pessimiste). Cette constatation est encore plus marquée avec l'approche par modélisation.

Il faut tout de même nuancer ces résultats. En effet, nos deux méthodes de projection ont des limitations importantes. Elles ne prennent pas en compte des variables explicatives essentielles, ce qui peut influencer considérablement nos résultats. De plus, la résolution spatiale et temporelle choisie pourrait ne pas être suffisante pour représenter de manière précise les phénomènes liés aux risques couverts par le régime, tels que la température, le vent et la pluie. Enfin, l'utilisation d'une seule trajectoire du GIEC pour chaque scénario introduit une source d'incertitude, car nous n'avons pas accès à plusieurs trajectoires pour un même modèle climatique via le portail DRIAS, ce qui limite notre capacité à effectuer des projections plus précises.

Executive summary

Introduction

Tourism is essential for the French economy, contributing 7.13% of the GDP in 2016, with 4.85% from domestic visitor spending and 2.28% from foreign visitor spending. Tourism encompasses various activities during trips to places outside one's usual environment, for leisure, business, and more. Campgrounds are popular for family vacations.

Climate change, marked by an increase in extreme weather events, has an impact on tourism. The IPCC predicts more abnormally hot days in the near future. Campers suffer from these extreme weather conditions. Some campgrounds offer "Sunshine guarantees" to counteract rainy holidays. However, these guarantees are limited, covering only certain weather risks, primarily rain. Risks such as heatwaves and strong winds can also affect the appeal of campgrounds.

The objective of this dissertation is to present a scheme, which we will call the « **good weather** » scheme that offers an index-based insurance solution against climatic uncertainties (temperature, rainfall, and wind), providing compensation to a victim of bad weather while camping. This dissertation will also aim to assess the impact of climate change on this scheme.

To achieve this objective, the studies in this dissertation will be divided into two parts : first, the establishment of the fictional scheme, including its operational description and the creation of a tariff schedule. Subsequently, the implementation of two distinct approaches will allow for the assessment of the impact of climate change on key scheme indicators (S/P loss ratio and reserves).

Operation of the scheme, pricing, and analysis

The « good weather » scheme allows someone to be cover against three climatic risks : temperature, rain and wind. For the temperature risk, the scheme adopts two temperature indicators : one for a decrease (Tb) and the other for an increase (Th). The first one is related to heat waves, and the second one to cold waves. These indicators for day i are :

$$Th(i) = max(T_{mean}(i) - T_{saison}(i), 0) \quad et \quad Tb(i) = -min(T_{mean}(i) - T_{season}(i), 0) \tag{9}$$

with $T_{mean}(i)$: Average daily temperature of day i ; $T_{season}(i)$: Average temperature of the season of day i over the previous decade. For example, for the days of January 2020, $T_{season}(i)$ will represent the average temperature in winter between 2001 and 2011.

For wind risk, the associated indicator is denoted Ph is given for day i by :

$$Vh(i) = V(i) - V_{season}(i) \tag{10}$$

with $V(i)$: Average wind speed (m/s) on day i ; $V_{season}(i)$: Average wind speed for the season of day i over the previous decade.

the associated indicator to the risk of rain is denoted Ph and is given for day i by :

$$Ph(i) = P(i) - P_{season}(i) \tag{11}$$

with $P(i)$: Average precipitation of day i ; $P_{season}(i)$: Average precipitation for the season of day i over the previous decade.

On the model of parametric insurance, the scheme would offer a reimbursement of 50 €, in the event of exceeding one or more indices associated with the risks of temperature, wind and rain for an individual, and per night, having subscribed to the contract. The threshold for a particular index is its 95% quantile observed over the past ten years. To subscribe to the contract of this plan, the individual will pay a premium which will cover part or all of their camping stay.

If we denote by $\mathbb{1}_T(i)$ the indicator function of excess (1 in case of excess 0 otherwise) associated with the temperature risk, $\mathbb{1}_P(i)$ that associated with the rain risk and $\mathbb{1}_V(i)$ that associated with the wind risk, the total compensation function for an individual, and per night, having subscribed to the contract is written as the following way :

$$I_{total}(i) = 50 \times max(\mathbb{1}_T(i), \mathbb{1}_V(i), \mathbb{1}_P(i)) \tag{12}$$

The establishment of a tariff grid adapted to the description of the functioning of the « good weather » scheme required the use of data on overnight stays at campsites from surveys on attendance in outdoor hotels carried out by INSEE and historical data from Météo France (SYNOP). Several reprocessing steps made it possible to obtain daily data for each region, including the climatic variables of interest and the number of overnight stays. The SYNOP data covers the years 2001 to 2021, while the data on the number of nights covers the years 2011 to 2021.

The tariff grid, defined by region and by month, was obtained using the expected average cost approach weighted by nights over the period 2011 to 2015. This involves averaging the potential historical compensation that would have been paid by the structure of the contract during the period 2011 to 2015, by region and by month weighted by the number of camping nights (Nu). The pure premium for month m and region is given by the following formula :

$$\pi_{month,region} = \frac{\sum_j I(m,r,j) * Nu(m,r,j)}{\sum_j Nu(m,r,j)} \tag{13}$$

Figure 6 shows the tariff grid for the « **good weather** » scheme.

Tariff grid													
	Ile-de-France	Centre-Val de Loire	Bourgogne-Franche-Comte	Normandie	Hauts-de-France	Grand Est	Pays de la Loire	Bretagne	Nouvelle-Aquitaine	Occitanie	Auvergne-Rhone-Alpes	Provence-Alpes-Cote d Azur	Corse
January	3.32 €	2.68 €	10.31 €	6.03 €	12.35 €	8.74 €	5.11 €	11.61 €	2.29 €	2.26 €	7.43 €	5.06 €	4.91 €
February	4.93 €	7.19 €	11.22 €	7.25 €	7.88 €	9.30 €	6.93 €	10.82 €	6.85 €	5.40 €	9.00 €	5.25 €	4.67 €
March	11.95 €	13.14 €	13.14 €	20.47 €	21.66 €	23.11 €	10.64 €	13.93 €	6.89 €	8.16 €	11.42 €	10.97 €	6.48 €
April	4.13 €	6.83 €	5.08 €	7.17 €	5.55 €	2.92 €	5.65 €	9.80 €	3.46 €	4.83 €	4.45 €	10.39 €	2.55 €
May	2.99 €	4.46 €	4.84 €	2.92 €	6.25 €	3.59 €	4.60 €	8.04 €	2.16 €	3.07 €	4.39 €	7.58 €	4.00 €
June	9.84 €	4.35 €	7.65 €	9.78 €	9.10 €	3.11 €	4.66 €	12.22 €	6.55 €	5.69 €	3.84 €	8.43 €	1.94 €
July	8.05 €	6.30 €	10.03 €	9.41 €	10.43 €	5.84 €	5.12 €	10.50 €	4.02 €	6.48 €	8.60 €	18.69 €	11.87 €
August	6.79 €	4.86 €	6.46 €	7.38 €	7.78 €	4.22 €	5.18 €	10.12 €	6.66 €	5.70 €	6.38 €	13.17 €	8.39 €
September	2.36 €	2.47 €	3.62 €	3.52 €	4.53 €	0.94 €	2.35 €	4.55 €	2.35 €	1.69 €	4.87 €	5.59 €	3.05 €
October	3.38 €	3.49 €	7.52 €	8.81 €	5.13 €	6.33 €	4.42 €	7.96 €	3.07 €	3.50 €	7.40 €	8.91 €	5.84 €
November	8.36 €	7.72 €	16.81 €	15.34 €	12.53 €	14.22 €	7.10 €	15.64 €	7.99 €	9.54 €	11.42 €	14.46 €	5.02 €
December	3.85 €	3.40 €	6.89 €	12.07 €	10.03 €	5.92 €	7.04 €	14.22 €	4.81 €	1.25 €	2.71 €	2.93 €	1.55 €

Figure 6 – Pure premium of the « **good weather** » scheme.

Using this tariff grid, and assuming that the number of policyholders who would purchase the product (market share) is a variable proportion of the number of nights, we obtain the S/P ratios for the years 2016 to 2021 (figure below).

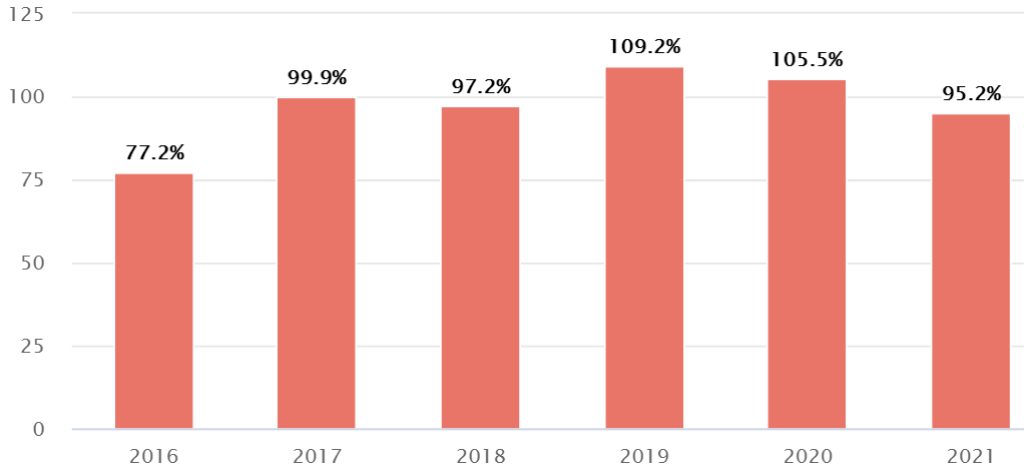


Figure 7 – Evolution of the scheme's S/P ratio (2016 - 2021).

The S/P ratio of the scheme remains satisfactory during the first three years (2016, 2017, 2018), staying below 100%. However, a deterioration in the S/P ratio is observed for the years 2019 and 2020, with S/P ratios of 109.2% and 105.5%, respectively. This trend encourages us to consider a tariff revision to account for the claims experience in more recent years. This tariff revision has led to a 10% increase in rates starting from the year 2021.

The analysis of the sinistrality resulting from the implementation of the scheme, using tools from extreme value theory and statistics, highlights extreme claims falling within the Fréchet domain of attraction. The table below presents the results of various methods for determining the extreme threshold, denoted as u.

	Mean excess plot	Hill Plot	Gestengarbe plot	AMSE Minimization
threshold u	125 000€ 180 000€	125 000€ 170 000€ 260 000€	132 200€	131 000€

Table 3 – Summary of thresholds by determination method

most of the thresholds obtained are close to 130,000€ . Based on expert opinion, we concluded that the threshold of **130,000€** appears consistently in most of the results obtained, making it a appropriate choice to distinguish extreme losses from non-extreme losses. With this threshold, extreme losses represent 5.7% of the total loss experience of the fair weather regime, including zero values.

Given the extreme nature of the regime's sinistrality, the generalized Pareto regression tree model appears to be an interesting model for designing more homogeneous classes of claims based on explanatory variables of interest for a better projection of the extreme claims experience according to a given horizon. The figure below shows the tree obtained from the generalized pareto regression (GP CART) procedure.

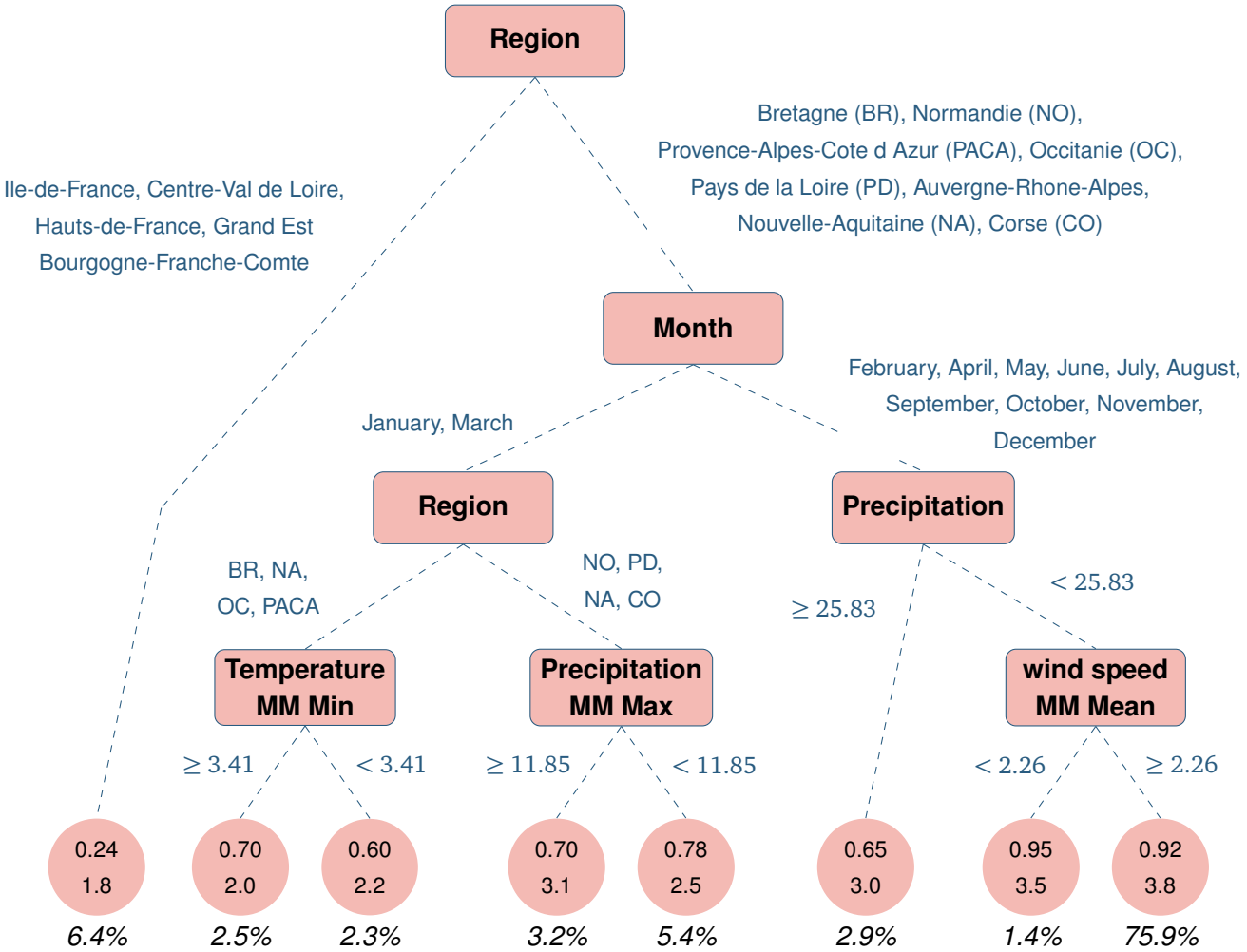


Figure 8 – Generalized Pareto Regression Tree (GP CART). For each leaf, the value of the shape parameter ξ (first line) and the scale parameter σ at 10^{-5} (second line) are given. The percentage of observations assigned to each sheet is mentioned.

The tree is composed of 8 leaves with four (4) *splits* according to only 6 variables : the region, the month, the daily precipitation, the minimum of the daily temperature on a rolling of 7 days (*MM Min*), the maximum daily precipitation over a 7-day sliding period (*MM Max*) and the average wind speed over a 7-day sliding period (*MM Mean*).

Projection data, approaches and results

For the projection of the scheme until 2100, we used data provided by DRIAS. DRIAS makes available the projection results from climate models developed in different climate modeling laboratories. It is important to note that all DRIAS data, including those available for past periods, are not observed weather data, but rather data simulated by these climate models. For data processing, we aggregated daily data from the DRIAS database over the period from 2023 to 2100, using the RCP 2.6 and 8.5 scenarios, so that they correspond to the spatial and temporal resolution chosen for our study.

This thesis adopts two approaches for projecting the regime until 2100 : *a deterministic approach* and *a modeling-based approach*. In both approaches, it is essential to make a number of realistic assumptions. These assumptions mainly relate to the evolution of the insured population and the tariff grid. Regarding the hypothesis regarding the evolution of the insured population, it is assumed to represent 25% of total nights, which is assumed to follow geometric growth. As for the hypothesis on the price scale, we assume that it remains constant over the projection period and corresponds to the last price scale which underwent an adjustment.

The first approach, called *deterministic*, simply consists of applying the rules of the regime mechanism « **good weather** » to the data from the DRIAS database according to the RCP 2.6 and RCP 8.5 scenarios. For example, using future data from RCP 2.6, we recalculate the temperature (Th), precipitation (Ph) and wind speed (Vh) indices. If one of these indices exceeds its defined threshold, the scheme will pay compensation of 50 euros to policyholders. The total premium and total loss projected for a given year are then determined using the assumptions made concerning the evolution of the insured population and the price list.

The *modeling* approach, on the other hand, assumes that the process which governs the operation of the scheme (triggering of payments, compensation, etc.) is random. Let S be the distribution of the payment (either 0 in the event of non-triggering or Y strictly positive).

Then the mathematical expectation of S is written :

$$E(S) = P(S > 0)[P(Y \geq u)E(Y|Y \geq u) + (1 - P(Y \geq u))E(Y|Y < u)] \tag{14}$$

Regarding this expectation of the distribution of S, we consider that the distribution of S is a mixture variable with the same distribution as $\tau(x) * [\delta(x)Z_1 + (1 - \delta(x))Z_2]$, with :

— $\tau(x)$ and $\delta(x)$ are Bernoulli random variables depending on X. $p_i(x) = P(\tau(x) = 1)$ and $m_i(x) = P(\delta(x) = 1)$ represent respectively the triggering probability ($S \neq 0$) and

the probability of exceeding the extreme threshold u ;

— $Z_1 = Y|Y \geq u$, has a distribution which is estimated using *the generalized Pareto tree* (GP CART) of figure 8 with $u = 130\,000$€ ;

— $Z_2 = Y|Y < u$, will be modeled using a classic GLM for severity. We consider $\tau \perp\!\!\!\perp (\delta, Z_1, Z_2)$ and $\delta \perp\!\!\!\perp (Z_1, Z_2)$.

τ is calibrated using a random forest (RF) model with maximum tree depth 15 trained on the balanced data using the resampling technique (SMOTE+ENN). The SMOTE+ENN technique aims to improve the generalization capacity of a model while reducing class imbalance. It first applies SMOTE to generate synthetic examples from the minority class, then it applies ENN to eliminate misclassified examples from the majority class. δ is calibrated using a random forest model of an approach that we adopted as part of this dissertation. This procedure consists of estimating $\hat{y}$ using a random forest model. Then, the exceedance probabilities p_i are estimated using a normal distribution with mean $\hat{y}$ and standard deviation $\sigma_y : \mathcal{N}(\hat{y}, \sigma_y)$, with σ_y equal to the standard deviation of y on the training data.

$$p_i = 1 - CDF_{\mathcal{N}\left(\hat{y}_i, \sigma_y^2\right)}(u) \tag{15}$$

As for the attritional sinistrality $Y|Y < u$, it is calibrated using a GLM Gamma model with a logarithmic link function. Finally, the extreme sinistrality is projected using the generalized Pareto tree (GP CART) and regression tree with a quadratic loss function that is calibrated on each of the 8 leaves of the GP CART. This procedure allows a significant gain in terms of RMSE of around 15% compared to the procedure which would consist of applying a regression tree with quadratic loss function directly on the extreme data.

Losses are then estimated by calibrating each component of this mixture of models. Premiums, for their part, are calculated based on projected losses and the hypothesis relating to the tariff grid. The advantage of this approach is to free itself from the hypothesis on the evolution of the insured population. A second advantage is to take into account, in the projection, the extreme aspect of the sinistrality of the scheme.

Figures 9 and 10 present the evolution of the projected S/P ratios according to the IPCC scenarios (RCP 2.6 and RCP 8.5) respectively according to the deterministic approach and according to the by modeling.

Generally speaking, we saw fairly similar findings between the two approaches, especially regarding trends. In both approaches, the projected S/P ratios show strong irregularity, mainly due to variability in climate data. We notice a clear deterioration of the S/P ratio at the start of the projection period (2025 - 2028) in the deterministic approach, while this strong deterioration is not observed in the approach based on modeling. This significant degradation could be explained mainly by an abrupt change in the climate trajectory currently observed. Furthermore, beyond the year 2063, in scenario 8.5, the situation becomes very unfavorable, with ratios which often exceed 100% and which are higher than those of the

optimistic scenario 2.6 in the case of the approach by modeling. Finally, the results indicate an overall unfavorable situation for the climate data from the RCP 8.5 scenario (pessimistic) and is much more marked with the modeling approach.

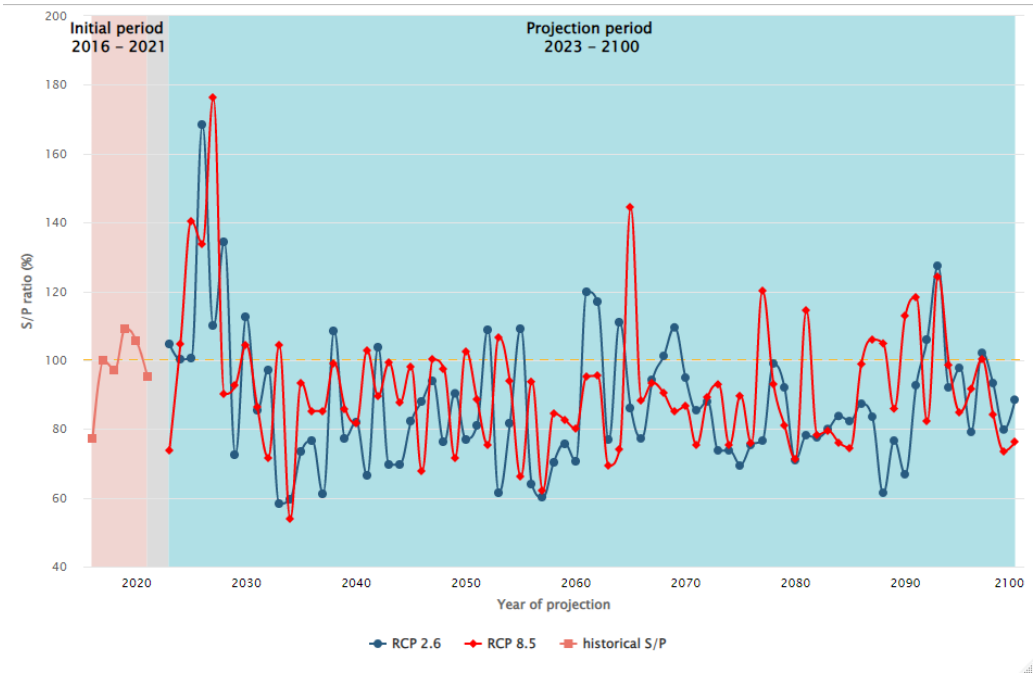


Figure 9 – S/P projected according to IPCC scenarios and according to the deterministic approach.

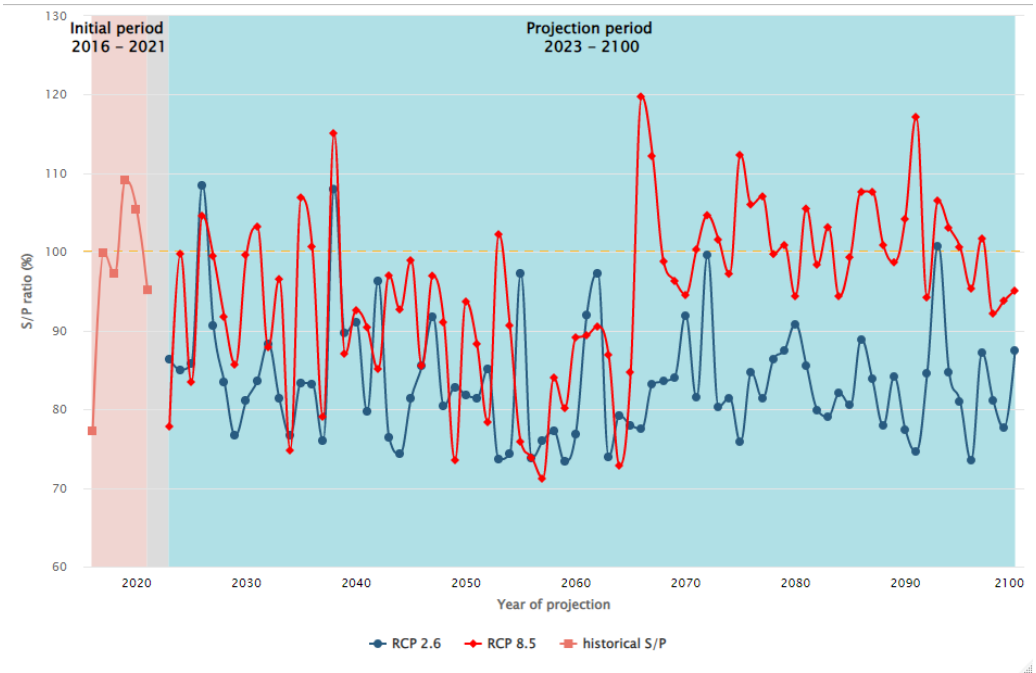


Figure 10 – S/P projected according to the IPCC scenarios and according to the modeling approach.

Table below highlights less volatile projected S/Ps in the modeling approach compared to the deterministic approach.

Statistics	*Historical S/P*	*Approach*			
		Deterministic		*Modeling*	
		RCP 2.6	RCP 8.5	RCP 2.6	RCP 8.5
E(S/P)	97.4%	87.3%	91.9%	83.6%	95.1%
σ(S/P)	11.2%	19.1%	19.3%	7.5%	10.7%

Table 4 – Mean and standard deviation of projected and historical S/P ratios.

There is a linear upward trend in reserves, whatever the scenario and approach considered. However, the reserves in the RCP 8.5 scenario are lower than those in the RCP 2.6 scenario with the deterministic approach. On the other hand, with the modeling approach, this inferiority is no longer observed beyond 2085. The reserves are given by the formula :

$$R_n = R_{n-1} + \text{Result}(n) = R_0 + \sum_{i=2023}^{n} \text{Result}(i) \tag{16}$$

with $\text{Result}(n) = P(n) + FG(n) - S(n)$, of which $P(n)$, $FG(n) = 0.1 * P(n)$ and $S(n)$ represent respectively the total pure premium, the overall costs and the total loss projected for the year n and R_0 the starting reserves equal to 55501059€ (sum of results from 2016 to 2021).

Conclusion

Generally speaking, we observed fairly similar conclusions between the two approaches, particularly with regard to the trends in the indicators used, namely the S/P ratio and reserves. In both approaches, the projected S/P ratios show strong irregularity, mainly due to the climatic variability of the data, while the reserves show a rather linear trend.

In terms of differences, we noticed that the S/P ratios were rather underestimated with the modeling approach, while with the deterministic approach, we observed a strong deterioration of the S/P ratio at the start of the period. projection (2023 - 2028). In addition, the results indicate an overall unfavorable situation for the climate data from the RCP 8.5 (pessimistic) scenario. This observation is even more marked with the modeling approach.

However, these results must be qualified. Indeed, our two projection methods have significant limitations. They do not take into account essential explanatory variables, which can considerably influence our results. Furthermore, the chosen spatial and temporal resolution may not be sufficient to accurately represent the phenomena related to the risks covered by the regime, such as temperature, wind and rain. Finally, the use of a single IPCC trajectory for each scenario introduces a source of uncertainty, because we do not have access to several trajectories for the same climate model via the DRIAS portal, which limits our ability to carry out more precise projections.

Table des matières

Sigles et Abréviations

ACPR	Autorité de contrôle prudentiel et de résolution
ADASYN	Echantillonnage Synthétique Adaptatif (*Adaptive Synthetic Sampling*)
AUC	Aire Sous la Courbe (*Area Under the Curve*)
CA	Chiffre d'Affaire
DGE	Direction Générale des Entreprises
DT	Arbre de décision (*Decision Tree*)
ENN	Voisins les plus proches corrigés (*Edited Nearest Neighbors*)
HPA	Hôtellerie de Plein Air
GAM	Modèles additifs généralisés (*Generalized Additive Model*)
GLM	Modèle Linéaire Généralisé (*Generalized Linear Model*)
GIEC	Groupe d'experts Intergouvernemental sur l'Evolution du Climat
GPD	Generalized Pareto Distribution
GP CART	Arbre de régression pareto généralisée
IA	Intelligence Artificielle
INSEE	Institut National de Statistiques et des Etudes Economiques
LIME	Explication Locale et Interprétable, Indépendante du Modèle (Local Interpretable Model-Agnostic Explanations)
MAE	Erreur Moyenne Absolue (*Mean Absolute Error*)
ML	Apprentissage Automatique (*Machine Learning*)
MSE	Erreur Quadratique Moyenne (*Mean Square Error*)
NN	Réseau de neurones (*neural network*)
OMM	Organisation Mondiale de la Météorologie
PDP	Graphique de Dépendance Partielle (*Partial Dependance Plot*)
RCP	Trajectoires représentatives de concentration (*Representative Concentration Pathway*)
RF	Forêt Aléatoire (*Random Forest*)
RGPD	Règlement général sur la protection des données
RMSE	Erreur Quadratique Moyenne (*Mean Square Error*)
SMOTE	Suréchantillonnage des Observations Minoritaires (*Synthetic Minority Oversampling TEchnique*)
SYNOP	Observation météorologique historiques France
SVM	Support Vector Machine
TVE	Théorie des Valeurs Extrêmes
XGBoost	eXtreme Gradient Boosting

Introduction

Le tourisme constitue une activité cruciale pour l'économie de la France. La consommation touristique intérieure représentait 7.13 % du PIB de l'année 2016, dont 4.85% pour la consommation des visiteurs français et 2.28% pour celle des visiteurs étrangers. Selon l'INSEE le tourisme comprend les activités déployées par les personnes au cours de leurs voyages et séjours dans des lieux situés en dehors de leur environnement habituel pour une période consécutive qui ne dépasse pas une année, à des fins de loisirs, pour affaires et autres motifs non liés à l'exercice d'une activité rémunérée dans le lieu visité. A ce type, on y retrouve les campings qui constituent une destination de vacances privilégiée pour de nombreux français et étrangers souhaitant passer des vacances en famille.

Le camping en vacance est un secteur important du tourisme français. En effet, selon les statistiques fournies par l'INSEE, la fréquentation de campings s'est accrue de 25% entre 2008 et 2017 (figure 11).

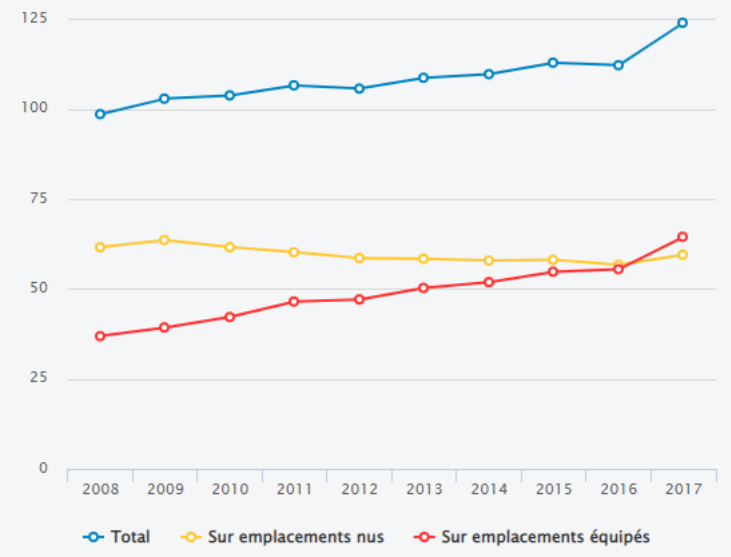


Figure 11 – Evolution de la fréquentation des campings français entre 2008 et 2017 (source : INSEE)

D'autre part, la variabilité climatique observée au cours des deux dernières décennies met en évidence un changement climatique fort et sans précédent. Une des conséquences de ce changement climatique est l'accroissement de la fréquence et de l'intensité des phénomènes météorologiques extrêmes (vagues de chaleur, sécheresses, inondations, cyclones, tempêtes, pluies...). Par exemple, le GIEC prévoit que le nombre de jours « anormalement chauds » (la température maximale supérieure de plus de 5°C à la normale 1981-2010) va continuer à augmenter, et on aura de 20 à 40 jours anormalement chauds supplémentaires dans un horizon proche (2021-2050). L'intensification des phénomènes climatiques cause des

séjours difficiles aux campeurs. Le climat est aujourd'hui un choix dominant pour les vacances en camping. Ainsi, pour garder le niveau d'attractivité, plusieurs campings et clubs français ont adapté leur offre en proposant des "garanties Soleil" contre l'angoisse des vacances pluvieuses. C'est le cas du groupe Sunelia qui inclut gratuitement et automatiquement son offre Soleil pour tout séjour de 7 nuits minimum entre avril et octobre, réservé au plus tard 72 heures avant la date d'arrivée.

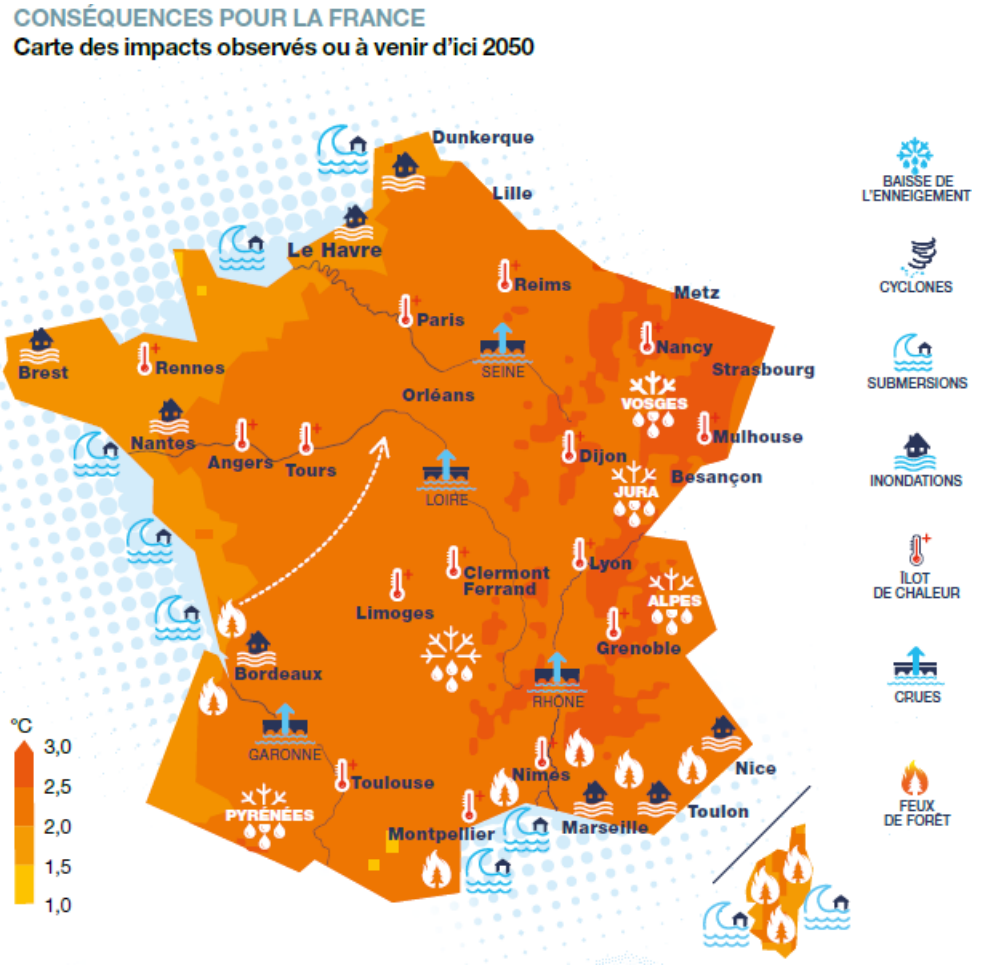


Figure 12 – Conséquence du réchauffement climatique en France (source : Météo France)

Toutefois, ces solutions comportent peu de risques météorologiques (principalement de pluie) et restent limitées. Pourtant, d'autres risques climatiques tels qu'une vague de chaleur et un vent trop fort peuvent affecter le séjour des personnes en camping et mettre un frein à l'attractivité d'un camping.

L'objectif de ce mémoire est de présenter un régime que l'on nommera "**régime beau temps**" qui proposera une solution d'assurance indicielle contre les aléas climatiques (température, pluie et vent) permettant d'indemniser une victime du mauvais temps en camping. Ce mémoire aura également pour objectif d'évaluer l'impact du changement climatique sur ce régime.

Pour atteindre les objectifs cités ci-dessus, ce mémoire est structuré en deux grandes parties. La première partie, composée de deux chapitres, portera sur le cadre conceptuel, mise

en place du régime fictif et aspect théorique. La deuxième partie, composé d'un chapitre, portera sur l'étude actuarielle de l'impact du changement climatique sur le régime. De manière spécifique, ce document est organisé en trois chapitres :

- dans le chapitre 1, il sera question de la phase de conception et de développement du produit ;
- le deuxième chapitre traitera de l'arbre de régression Pareto généralisée qui est l'outil de modélisation spécifique qui en intégrant les prévisions météorologiques permettra de faire la projection du régime. Dans ce chapitre il sera question également de parler de l'intérêt de cette modélisation spécifique pour la projection du régime ;
- enfin, le troisième chapitre sera consacré à la projection du régime à l'horizon 2100 à partir des scénarios du GIEC et à réaliser des études de sensibilité sur les principaux indicateurs du régime.

Le mémoire proposé ici constitue, en grande partie, un travail de recherche pour développer de nouvelles solutions répondant à des besoins réels. Il peut être vu comme un travail préliminaire à la mise en place de couvertures contre les aléas climatiques lors des séjours de campings.

PARTIE A

Cadre conceptuel : mise en place du régime d'assurance indicielle et aspect théorique

« On a toujours l'impression d'être plus victime du mauvais temps en camping. »

Marc Canavaggia, directeur de Sunelia (2015).

« Les Actuaires sont un peu comme les épidémiologistes : ils ont devant eux de nouvelles sources de données qu'il vont devoir prendre en considération. »

Antoine Lissowski, directeur général de CNP Assurances (2021).

Chapitre

I

Assurance indicielle beau temps : conception et tarification

Ce premier chapitre présente la phase de conception et de lancement du régime fictif « beau temps ». D'abord, nous présenterons la situation des trois risques couverts en France métropolitaine : la température, de la pluie et du vent . Ensuite, un accent sera porté sur la tarification du produit. Enfin, nous exposerons le ratio sinistres/-primes du régime après sa période de tarification pour s'assurer de la viabilité du régime fictif « **beau temps** ».

1 Température, pluie et vent en France métropolitaine

Le changement climatique est constatable à partir des données météorologique, dans cette partie, nous évoquerons l'évolution de la température, pluie et du vent à partir des données en provenance de Météo France.

1.1 Température (vague froids et chaudes)

1.1.1 Définition de la température, vague froids et chaudes

De manière courante, on définit la température comme une sensation de chaleur ou de froid éprouvée par le corps en un lieu, mesurée en degrés par rapport à une échelle connue (Celsius °C, Fahrenheit °F, Kelvin K). Elle est étroitement liée à l'énergie interne et l'enthalpie d'un système. Ainsi, une température est une mesure numérique d'une chaleur. Sa détermination se fait par détection de rayonnement thermique, la vitesse des particules, l'énergie cinétique, ou par le comportement de la masse d'un matériau thermométrique.

En fonction des seuils standards on distingue des vagues de chaleur et de froid. Les seuils de tolérances diffèrent en fonction des températures usuelles d'une région et de la capacité d'adaptation des populations.

Selon Météo France, les vagues de chaleur correspondent à des températures anormalement élevées, observées pendant plusieurs jours consécutifs. Mais il n'existe pas de définition universelle du phénomène : les niveaux de température et la durée de l'épisode qui permettent de caractériser une vague de chaleur varient selon les régions du monde et

les domaines considérés (caractérisation d'un point de vue climatologique, activité de recherche, dispositif de vigilance météorologique). Tout comme l'ensemble de la planète, la France connaît un réchauffement marqué depuis 1900. Les quatre années les plus chaudes ont d'ailleurs été observées au XXIème siècle : respectivement 2014, 2011, 2015 et 2018. Sur la période 1959-2009, la tendance observée est d'environ + 0,3°C par décennie. Ce réchauffement est à l'origine de vagues de chaleur plus intenses et plus fréquentes, et particulièrement marquées dans les villes. Par exemple les villes de Paris et Chartres ont une localisation assez proche et pourtant des seuils de température largement différents, cet effet est appelé « îlot de chaleur urbain ». Le nombre de jours « anormalement chauds » (jours pendant lesquels la température maximale est supérieure de plus de 5°C à la normale 1981-2010) va continuer à augmenter, avec possiblement, selon le scénario intermédiaire du GIEC : 20 à 40 jours anormalement chauds supplémentaires dans un horizon proche (2021- 2050) et 100 jours supplémentaires en 2100.

Une vague de froid est définie par Météo France comme un épisode de froid intense pendant plusieurs jours consécutifs sur une large étendue géographique. Ainsi, une vague de froid est un épisode durable et étendu de froid (au moins 3 jours). Pour qu'un épisode soit identifié à l'échelle nationale, il faut que la température moyenne nationale descende au moins une journée sous un certain seuil (-2 °C). Les vagues de froid sont aussi caractérisées à l'échelle d'une région lorsque l'épisode dure au moins deux jours et que les températures atteignent des valeurs nettement inférieures aux normales saisonnières de la région concernée. Les périodes de grand froid sont propices à la survenue d'autres phénomènes météorologiques potentiellement dangereux, comme la neige et le verglas, qui peuvent affecter gravement la vie quotidienne en interrompant la circulation routière, ferroviaire ou le trafic aérien. En France métropolitaine, les températures les plus basses de l'hiver surviennent habituellement en janvier ou février sur l'ensemble du pays. Mais des épisodes précoces (en novembre ou décembre) ou tardifs (en mars) sont également possibles.

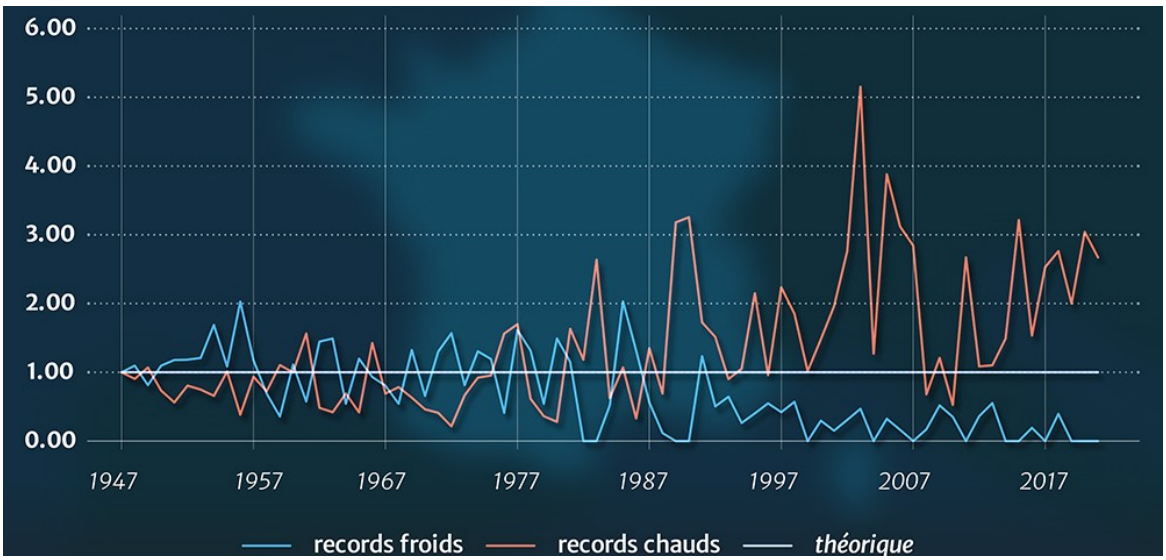


Figure I.1 – Évolution des records chauds et froids de l'indicateur thermique France de température moyenne quotidienne sur la période 1951-2018 (source : Météo France)

Le graphique I.1 représente le nombre de records quotidiens chauds ou froids qui ont été battus en tenant compte de la longueur de la série. Cette série est basée sur la série de température moyenne quotidienne en France (moyenne de 30 stations) depuis 1947. Dans un climat stationnaire, le nombre de records chauds (courbe rouge) et froids (courbe bleue) devraient être équivalents et identique à la valeur théorique (courbe noire, identique chaque année, égale à 1). Ce n'est pas ce qui est observé : depuis le milieu des années 1980, les records chauds (courbe rouge) sont systématiquement plus nombreux que les records froids (courbe bleue). Egalement, on observe que depuis le début du XXIe siècle, on a en moyenne deux fois plus de records chauds que la normale mais quatre fois moins de records froids.

1.1.2 Historique des températures en France

Evolution de la température moyenne

La température moyenne annuelle en France métropolitaine se situe entre 11.5 et 13.5°C (Figure I.2). La tendance observée qui s'en dégage est claire et correspond aux bilans de Météo France. Depuis 2000, les températures augmentent. Alors que le climat mondial s'est réchauffé d'environ 1°C à la surface du globe sur le siècle dernier, la France métropolitaine quant à elle a subi une augmentation de plus de 1.5°C. Avec 0.1°C d'augmentation par décennie, la tendance s'accélère depuis le XXI ème siècle avec +0.31°C par décennie (Météo France). Malgré la présence remarquable de l'année 2003 avec un pic à 13.21°C en moyenne sur l'année (année de la canicule historique), on note une augmentation continue avec +0.3°C jusqu'en 2020 avec une recrudescence des périodes caniculaires. Les variations annuelles exceptionnelles observées (fortes intempéries, vagues de chaleur extrême) attestent d'une évolution climatique globale et de son instabilité grandissante.

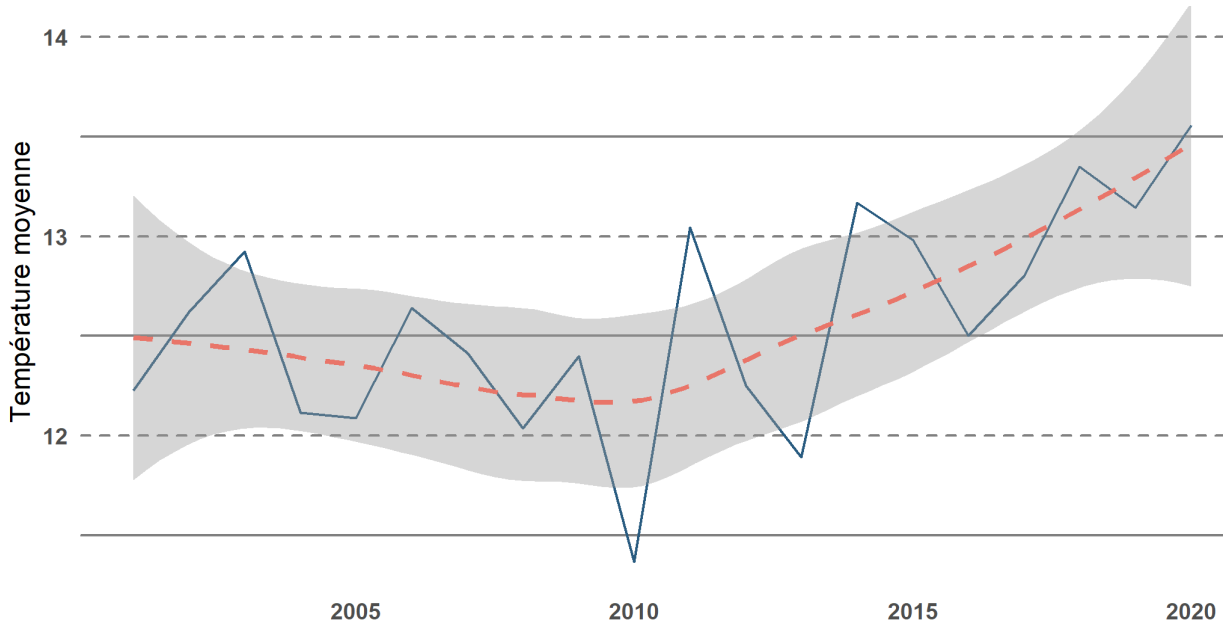


Figure I.2 – Température moyenne annuelle en France métropolitaine de 2001 à 2020 (données SYNOP Météo France).

De l'observation de la figure I.2, on identifie des pics annuels qui sont majoritairement le

solde d'une présence de vague de froid ou de chaleur durant l'année.

En analysant ces observations région par région, la présence de la mer semble atténuer la vitesse de réchauffement dans certaines régions (Bretagne, Normandie...), mais globalement la hausse des températures est visible partout et à vitesse comparable (figure I.3).

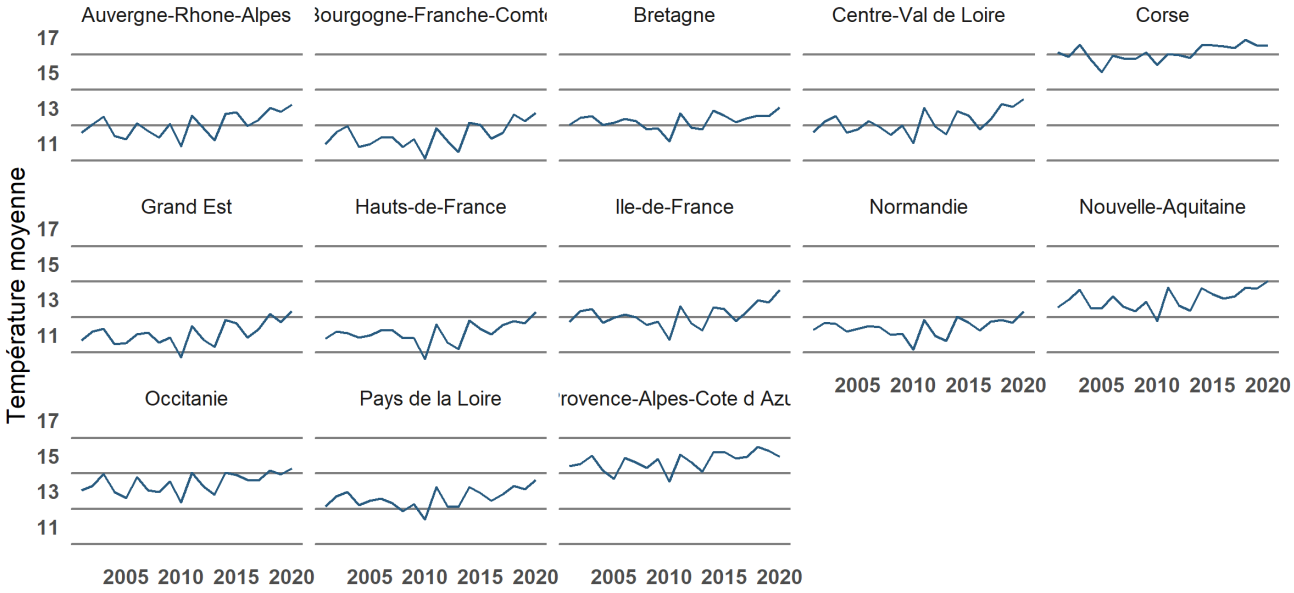


Figure I.3 – Températures moyennes par région en France métropolitaine entre 2001 et 2020 (données SYNOP Météo France).

Dépassement en température extrême

Dans ce mémoire nous définissons deux indicateurs de température à la baisse et à la hausse qui permettront d'analyser les dépassements de température en seuil extrême. Le premier est relatif aux vagues de chaleur et le deuxième aux périodes de vagues de froid.

— **Indicateur de température à la hausse** Th

$$Th(i) = max(T_{moyenne}(i) - T_{saison}(i), 0) \tag{I.1}$$

— **Indicateur de température à la baisse** Tb

$$Tb(i) = -min(T_{moyenne}(i) - T_{saison}(i), 0) \tag{I.2}$$

avec

- i : Jour de l'année sur lequel on effectue le calcul ;
- $T_{moyenne}(i)$: Température moyenne du jour i ;
- $T_{saison}(i)$: Température moyenne de la saison du jour i sur la décennie précédente. Par exemple, pour les jours du mois de janvier 2020, $T_{saison}(i)$ représentera la moyenne des température en période hivernale entre 2001 et 2011.

Avec les définitions des indicateurs (température à la baisse et température à la hausse) présentées ci-dessous, et en utilisant un seuil de dépassement à l'échelle nationale, les jours de fortes baisses ou de fortes hausses de température sont identifiés. Pour l'indicateur de

baisse de température (respectivement de hausse de température), ce seuil correspond au quantile 95% de l'indicateur Tb (respectivement Th) observé au cours de la décennie 2001 à 2010 et de la saison correspondante.

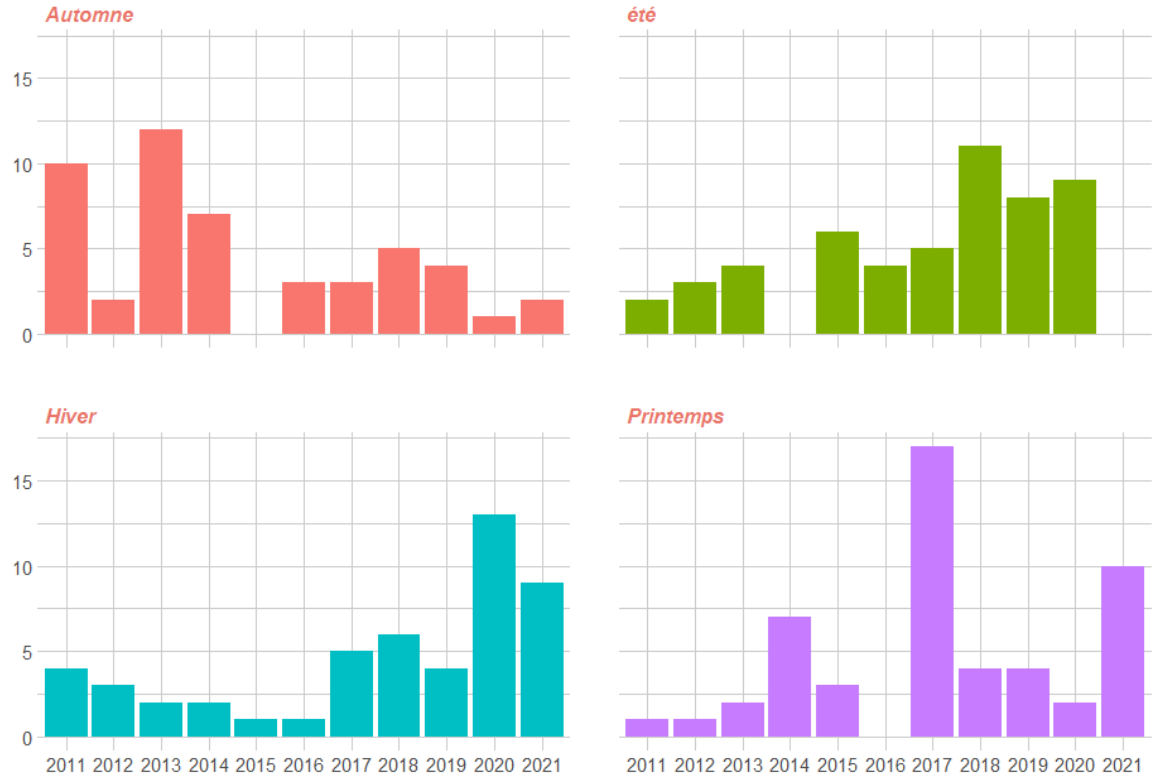


Figure I.4 – Nombre de dépassements de température à la hausse par saison en France métropolitaine de 2011 à 2021 (données SYNOP Météo France).

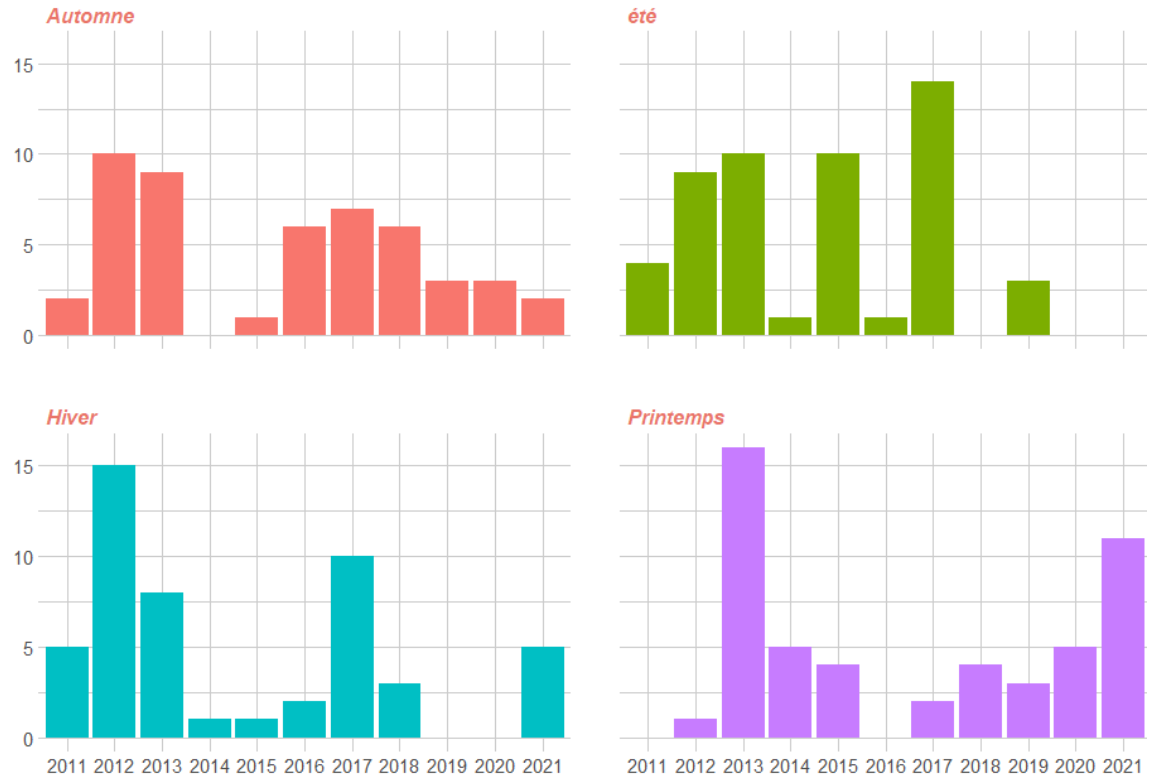


Figure I.5 – Nombre de dépassements de température à la baisse par saison en France métropolitaine de 2011 à 2021 (données SYNOP Météo France).

De façon global on constate une accélération des très fortes chaleurs ou des très fortes fraîcheurs (figure I.4 et I.5).

1.2 Pluie extrême en France métropolitaine

1.2.1 Définition

Le **centre d'information sur l'eau** (C.I.eau)[1] définit la pluie comme un phénomène naturel, qui apparaît sous forme de gouttes d'eau provenant des nuages et tombant vers le sol. C'est l'une des formes les plus courantes de précipitations sur Terre.

Il existe deux types de précipitations :

— **Les précipitations stratiformes** : elles couvrent une grande étendue, durent longtemps mais sont de faible intensité. Elles se produisent dans les zones de basse pression, les creux et sont associées à des nuages de types « stratus ». Il s'agit par exemple de la bruine ou encore de la pluie légère.

— **Les précipitations convectives** : elles couvrent de petites surfaces, ne durent pas longtemps mais sont de forte intensité. Elles sont très localisées et produites par l'instabilité convective de l'air. Ces précipitations sont associées à des nuages de types « cumulus ». C'est le cas notamment des orages, averses, cyclones...

Ainsi les précipitations peuvent prendre une forme liquide (pluie, bruine, pluie verglaçante, bruine verglaçante) ou solide (neige, neige en grains, neige roulée, grésil, grêle, granules de glace, cristaux de glace). Les précipitations contribuent à la fertilité et à l'habitabilité des zones tempérées ou tropicales ; dans les zones polaires, elles aident au maintien des calottes glaciaires.

Il n'existe pas de correspondance officielle entre l'appréciation "qualitative" d'une précipitation ("faible", "modérée" ou "forte") et son intensité chiffrée, qui peut s'exprimer en millimètres par minute ou millimètres par heure (1mm = 1 litre/m2).

Le caractère des précipitations dépend de la climatologie locale. Toutefois, en plaine et pour la France métropolitaine, Météo France suggère d'adopter les équivalences suivantes :

Pluie faible continue	1 à 3 mm par heure
Pluie modérée	4 à 7 mm par heure
Pluie forte	8 mm par heure et plus

Table I.1 – Caractérisation des précipitations (Météo France)

Les précipitations se mesurent en hauteur d'eau tombée au sol rapportée à une unité de surface. L'unité utilisée est le millimètre de précipitation par mètre carré. En supposant une répartition homogène des précipitations sur cette surface, 1 millimètre de pluie représente 1 litre d'eau par mètre carré. Selon **Météo France**, le cumul annuel moyen à échelle du pays

1. https ://www.cieau.com/

est passé de 934,8 mm sur la période 1981-2010 à un cumul de 934,7 mm sur la période 1991-2020, soit une baisse d'un dixième de millimètre.

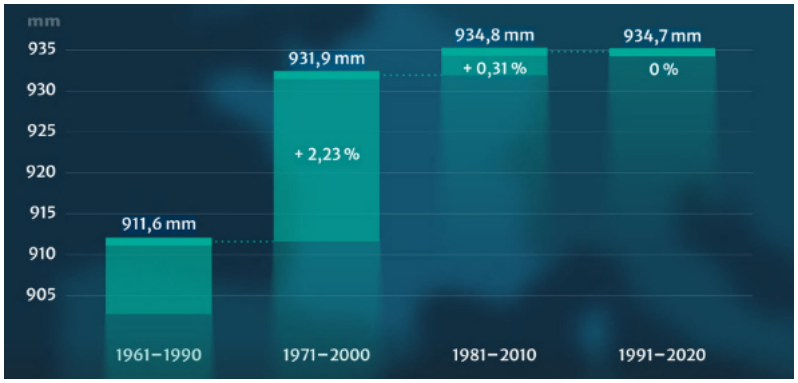


Figure I.6 – Évolution du cumul annuel moyen de précipitations en France au fil des décennies (source : Météo France)

1.2.2 Historique des précipitations en France

Evolution du volume d'eau en France métropolitaine

Le cumul annuel de précipitations en France métropolitaine entre 2001 et 2020 se situe entre 650 et 850 mm (Figure I.7). On note que le cumul de précipitations sur la période 2001-2020 présente une tendance annuelle erratique sur le pays. On identifie des pics annuels qui sont majoritairement le solde d'une présence de forte pluie durant l'année.

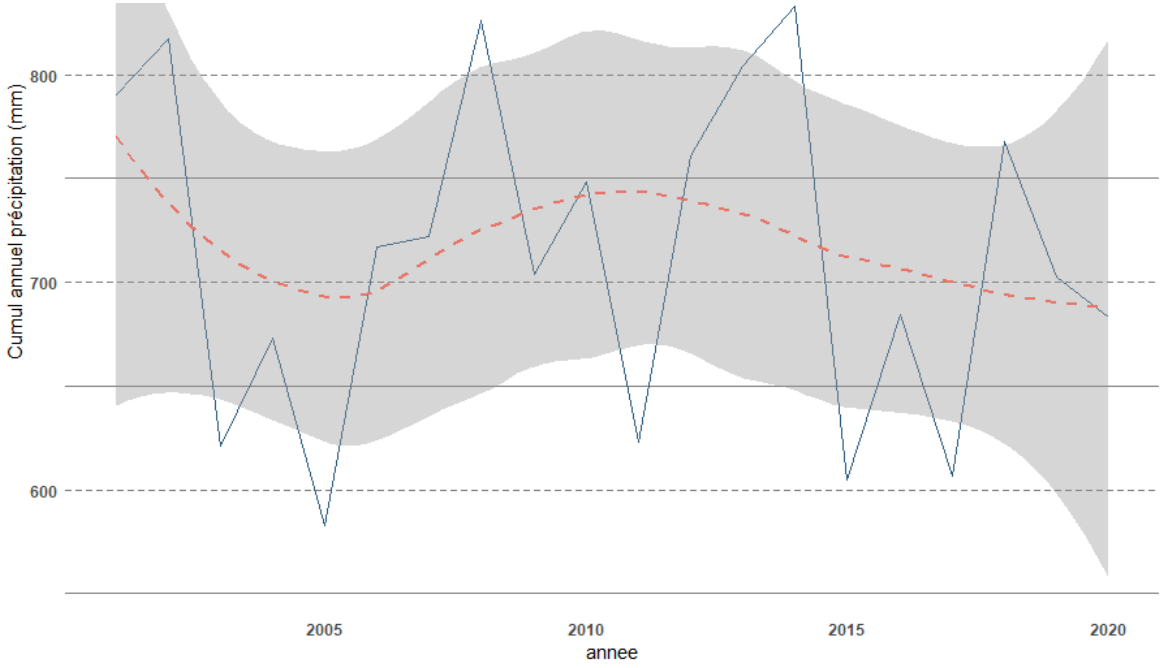


Figure I.7 – Évolution du cumul annuel moyen de précipitations en France Métropolitaine de 2001 à 2020 (données SYNOP Météo France).

En analysant ces observations région par région (figure I.8), le Grand-Est et le Centre-Val de Loire sont les régions qui s'assèchent le plus. À l'inverse, on observe une hausse de la pluviométrie annuelle dans les régions de la Nouvelle-Aquitaine et de la Corse.

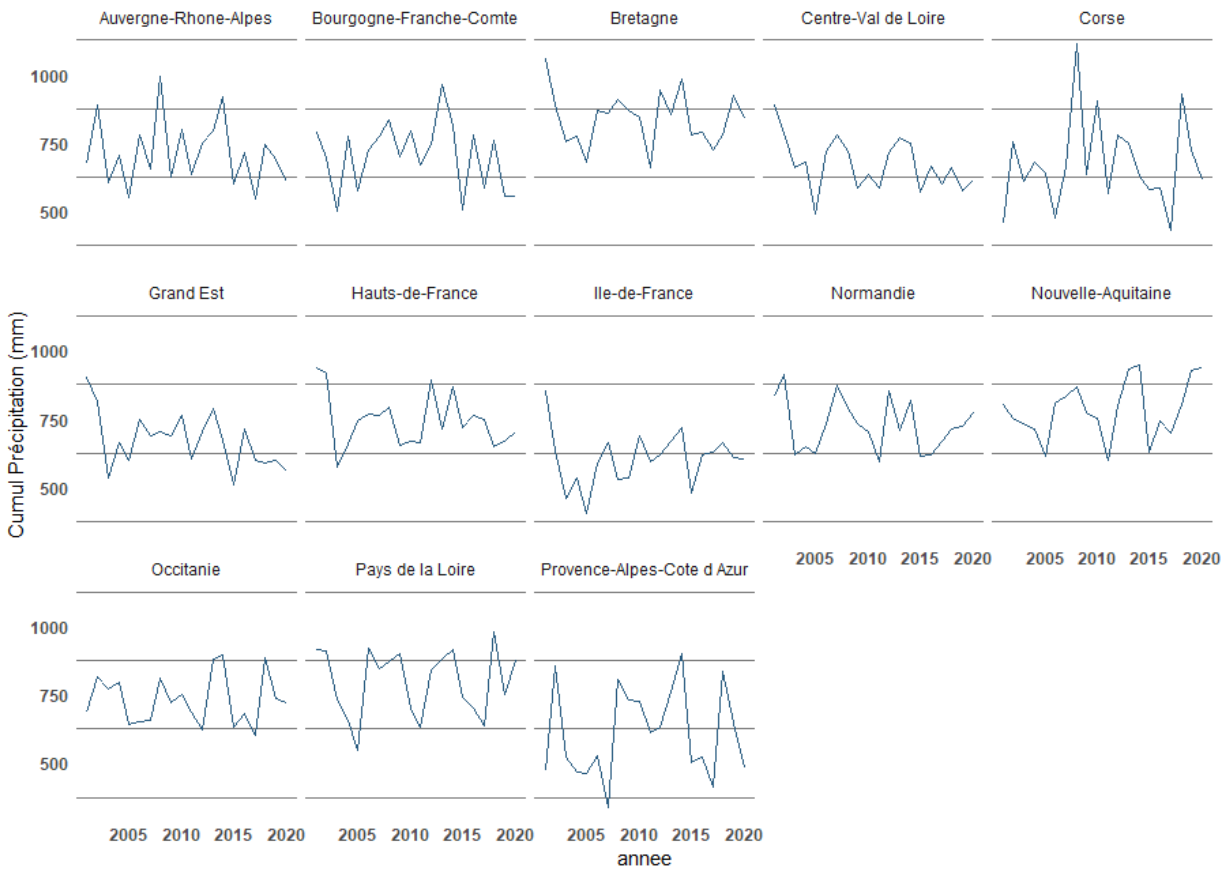


Figure I.8 – Cumul annuel moyen de précipitations par région en France métropolitaine de 2001 à 2020 (données SYNOP Météo France).

Dépassement en volume d'eau extrême

Au cours des dernières décennies, des preuves montrent que les épisodes pluvieux sont devenus plus intenses en France métropolitaine. Dans le cadre de ce mémoire nous avons défini l'indicateur permettant d'évaluer le caractère extrême d'une précipitation journalière. Ce indicateur se définit de la manière suivante :

$$Ph(i) = P(i) - P_{saison}(i) \tag{I.3}$$

avec

- i : Jour de l'année sur lequel on effectue le calcul ;
- $P(i)$: Précipitation moyenne du jour i ;
- $P_{saison}(i)$: Précipitation moyenne de la saison du jour i sur la décennie précédente. Par exemple, pour les jours du mois de janvier 2020, $P_{saison}(i)$ représentera la moyenne des précipitations en période hivernale entre 2001 et 2011.

Nous considérons qu'un jour présente un caractère extrême en terme de précipitation si l'indicateur de précipitation extrême de ce jour est supérieur à un seuil, au quantile 95% de l'indicateur Ph observé au cours de la décennie 2001 à 2010 et de la saison correspondante.

Le graphique I.9 montre le nombre de jours extrêmes en précipitation par saison en

France Métropolitaine.

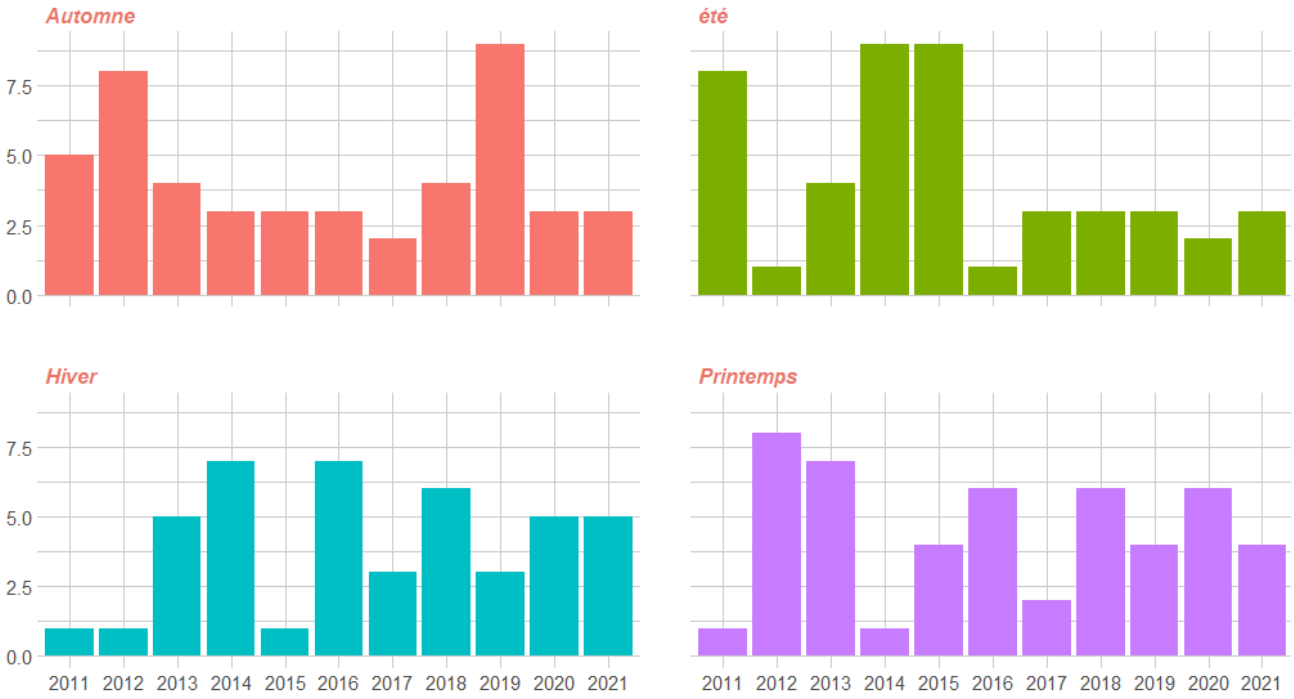


Figure I.9 – Nombre de jours extrêmes en précipitation par saison en France métropolitaine de 2011 à 2021 (données SYNOP Météo France).

L'évolution des précipitations extrêmes est contrastée selon la saison considérée, avec des précipitations hivernales en augmentation et des précipitations estivales plutôt en baisse.

1.3 Vent violent en France métropolitaine

1.3.1 Définition

Le vent naît sous l'effet des différences de températures et de pression. La pression sur la terre est haute si de l'air lourd et froid descend et basse si de l'air chaud et léger monte. La vitesse des vents s'exprime en kilomètres à l'heure ou en noeuds (environ 1,85 km/h par noeud).

La vitesse du vent est toujours mesurée sur un intervalle de temps, on distingue le vent instantané, qui est généralement mesuré sur 3 secondes et le vent moyen, généralement mesuré sur 10 minutes. Une rafale est une brusque augmentation de la vitesse instantanée du vent.

Les observations de vents violents ne coïncident pas toujours avec des dates de tempête car des vents violents peuvent aussi être observés sous forme de rafales lors d'orages violents.

Par convention internationale, les avis de tempête sont déclarés à partir d'une force Beaufort de 10 à 11, soit lorsque le vent instantané atteint 89 à 117 km/h ; au-delà de cette vitesse, on parle d'ouragan. Le tableau suivant présente les caractéristiques des différents degrés de

l'échelle de Beaufort qui permet de mesurer la force du vent.

Code	Vitesse (km/h)	Description	Effets sur terre
0	Inférieure à 1	Calme	La fumée s'élève à la verticale.
1	1 – 5	Très légère brise	La fumée est très légèrement déviée.
2	6 – 11	Légère brise	Les feuilles frémissent, les girouettes bougent.
3	12 – 19	Petite brise	Les feuilles et les petites branches bougent.
4	20 – 28	Jolie brise	Le vent soulève la poussière.
5	29 – 38	Bonne brise	Les arbustes commencent à se balancer.
6	39 – 49	Vent frais	Les câbles électriques sifflent.
7	50 – 61	Grand frais	Il devient assez difficile de marcher contre le vent.
8	62 – 74	Coup de vent	Les petites branches se cassent.
9	75 – 88	Fort coup de vent	Des branches se cassent. Faibles dégâts.
10	89 – 102	Tempête	Des arbres sont déracinés. Importants dégâts possibles aux habitations.
11	103 – 117	Violente tempête	Très gros ravages possibles.
12	118 et plus	Ouragan	Ravages catastrophiques.

Table I.2 – Les différents degrés de l'échelle de Beaufort (source : meteolor.fr)

La France a connu au total 41 tempêtes majeures, dont une majorité se sont produites en période hivernale au cours des mois de janvier, février et mars. Les tempêtes Lothar et Martin ont été les plus dramatiques de ces dernières dizaines d'années en termes de pertes humaines et de coût de dommages.

Selon les chiffres de Météo France, l'ensemble des 2 tempêtes (Lothar et Martin) a fait 92 victimes en France et 140 en Europe. Sur le plan financier, on note 15 milliards d'euros de dommages. Les indemnisations ont été évaluées à 7 milliards d'euros, au titre des contrats aux biens et automobiles selon une étude de la direction des études économiques et de l'évaluation environnementale.

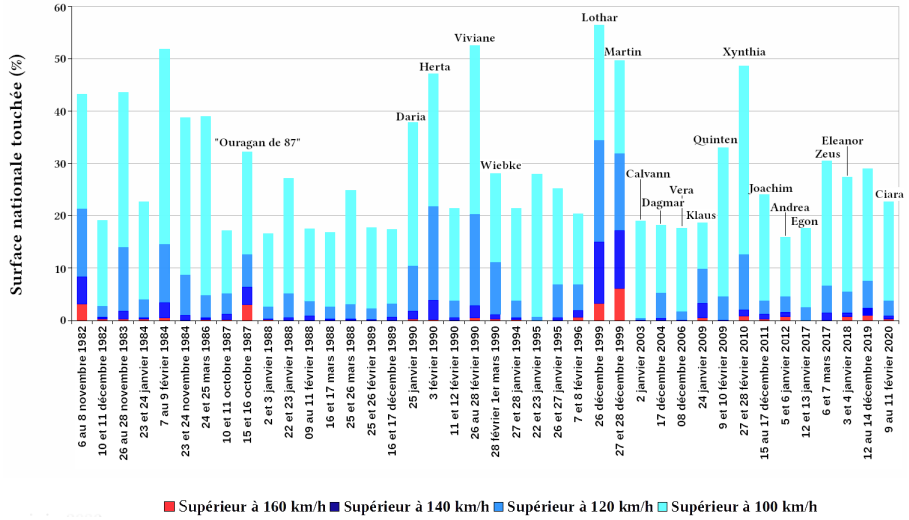


Figure I.10 – Tempêtes remarquables en France métropolitaine (Source : Météo France).

Le graphique I.10 montre les 40 tempêtes majeures en France métropolitaine de 1980 à juin 2020 en % du territoire national touché par de rafales supérieurs à 100 km/h. On remarque que les tempêtes les plus remarquables surviennent de manière assez irrégulière.

1.3.2 Historique de la vitesse de vent en France

Evolution de la vitesse de vent moyenne

La vitesse de vent moyenne annuelle en France métropolitaine entre 2001 et 2020 se situe entre 3 et 4.1 m/s (Figure I.11).

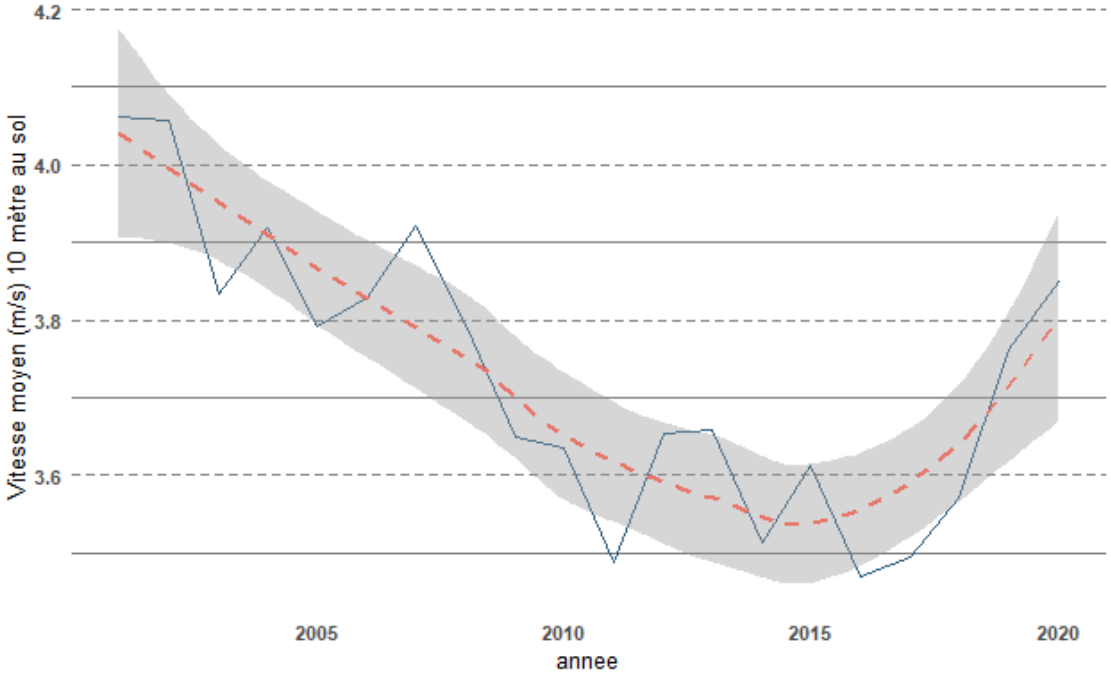


Figure I.11 – Évolution de la vitesse de vent moyen (m/s) en France métropolitaine de 2001 à 2020 (données SYNOP Météo France).

Dans l'ensemble, les vents ont soufflé un peu plus fort au cours de l'année 2020, au regard de la dernière décennie (tendance à la baisse entre 2001 et 2011).

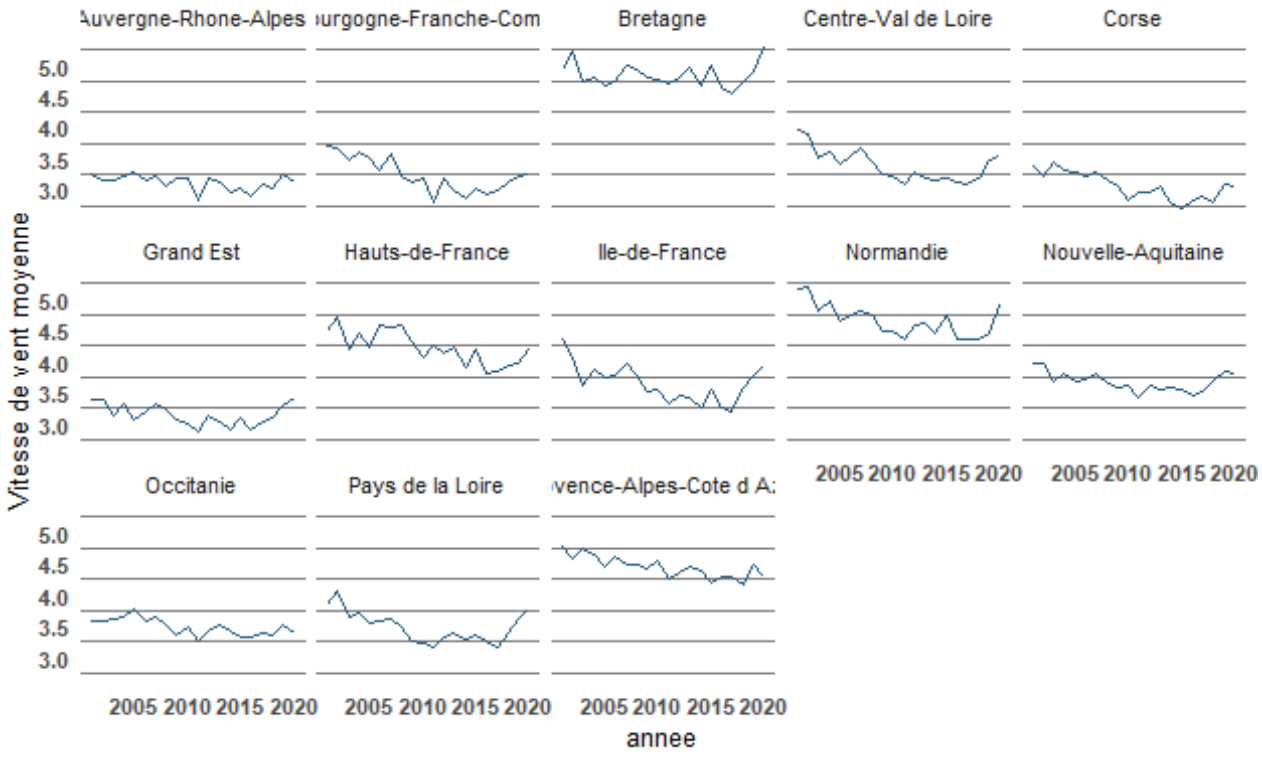


Figure I.12 – Vitesse de vent moyenne (m/s) par région en France métropolitaine de 2001 à 2020 (données SYNOP Météo France).

Une analyse par région (figure I.12), montre que les régions les plus venteuses de France sont les régions du nord-ouest (**Bretagne**, **Normandie**, **Haut-de-France**) et la région de **Provence-Alpes-Cote d'Azur** au sud du pays.

Dépassement en vitesse de vent

Les épisodes de vents forts sont devenus de plus en plus intenses en France métropolitaine au cours des dernières décennies. Dans le cadre de ce mémoire, nous avons défini l'indicateur permettant d'évaluer le caractère extrême d'une journée en terme de vitesse de vent. Cet indicateur se définit de la manière suivante :

$$Vh(i) = V(i) - V_{saison}(i) \tag{I.4}$$

avec

- i : Jour de l'année sur lequel on effectue le calcul ;
- $V(i)$: Vitesse de vent moyenne (m/s) qu'il a fait je jour i ;
- $V_{saison}(i)$: Vitesse de vent moyenne de la saison du jour i sur la décennie précédente. Par exemple, pour les jours du mois de janvier 2020, $V_{saison}(i)$ représentera la moyenne des vitesses de vent en période hivernale entre 2001 et 2010.

Nous considérons qu'un jour présente un caractère extrême en terme de vitesse de vent si l'indicateur de vitesse de vent extrême de ce jour (Vh) est supérieur au quantile 95% de l'indicateur Vh observé au cours de la décennie 2001 à 2010 et de la saison correspondante.

Le graphique I.13 montre le nombre de jours extrêmes en vent par saison en France Métropolitaine.

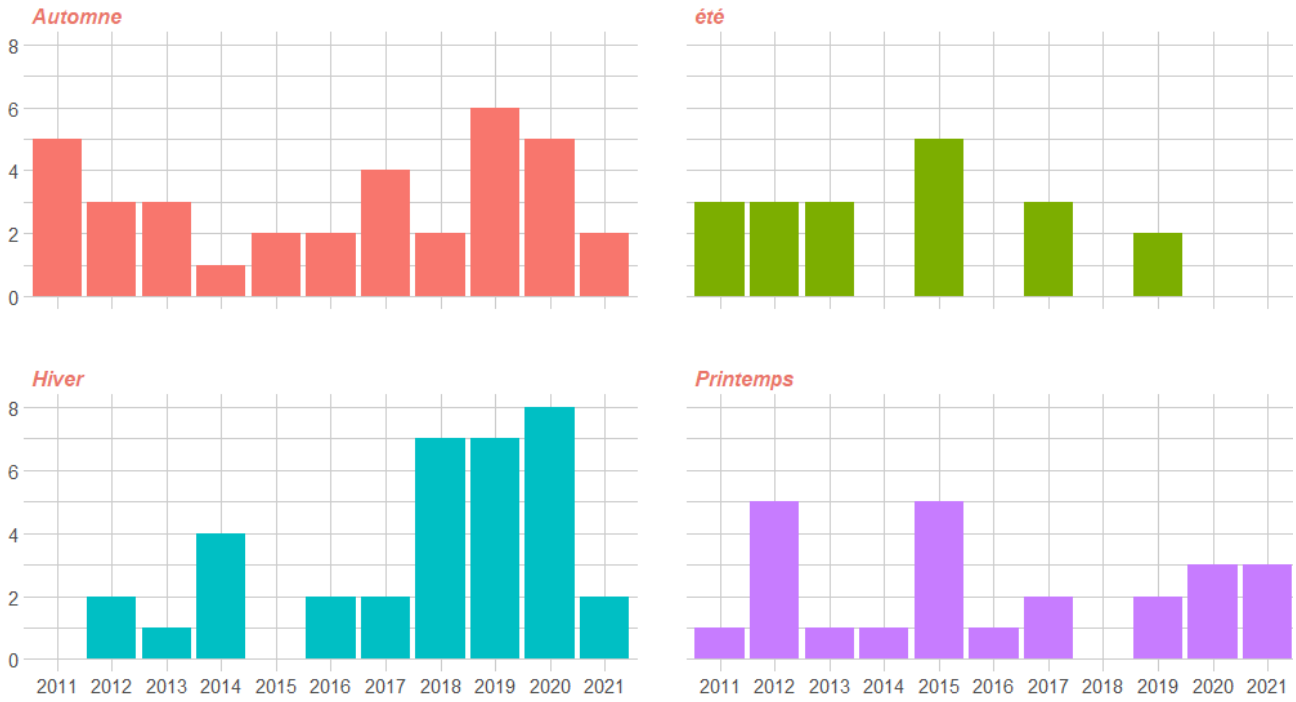


Figure I.13 – Nombre de jours extrêmes en vitesse de vent par saison en France métropolitaine de 2011 à 2021 (données SYNOP Météo France).

Dans l'ensemble on observe des vents de plus en plus extrêmes en période hivernale et en automne.

2 Campings en France métropolitaine

Le secteur de camping français affiche une santé éclatante depuis quelques années. Dans cette partie, il sera question de présenter les chiffres clés en matière de chiffre d'affaires, de capacité d'accueil et de fréquentation faisant de lui le principal mode d'hébergement touristique marchand en France.

2.1 Définition et organisation du secteur

Le camping, appelé également **l'hôtellerie de plein air** (HPA), est un secteur économique de l'hébergement touristique qui concerne l'activité de camping-caravaning et de parc résidentiel de loisirs, aménagé plus spécifiquement pour les habitats de types tentes, caravanes, maisons mobiles ou habitats légers de loisirs (source : INSEE [2]). Les campings sont constitués d'emplacements nus ou équipés de l'une de ces installations, ainsi que d'équipements communs. Un camping est ainsi un terrain aménagé pour camper, mais il existe également d'autres types de camping, comme le camping sauvage ou le camping chez l'habitant.

Le dictionnaire *Larousse* [3] définit le camping comme étant une activité de plein air consistant à vivre sous la tente avec un matériel adéquat.

Le secteur français du camping est organisé autour de grandes structures associatives et professionnelles. On y retrouve notamment **la Fédération française de camping caravaning** (FFCC) créée en 1938 et reconnue d'utilité publique en 1973, qui comprend environ 120 000 adhérents. Selon Eurostat, il y a près de 28500 campings en Europe et 27% d'entre eux sont en France. La France dépasse ainsi le Royaune-Uni (16%) et l'Allemagne (10%).

Les campings peuvent être classés, à la demande de l'exploitant et par un organisme accrédité, en 5 catégories allant de 1 à 5 étoiles. Les hébergements classés sont évalués selon trois grands axes : la qualité de confort, la qualité des services, les bonnes pratiques en matière de respect de l'environnement et d'accueil des clientèles en situation de handicap.

Selon la direction générale des entreprises (DGE), en 2018, les campings 1 et 2 étoiles représentaient 38% des campings classés, les campings 3 étoiles 40%, les campings 4 étoiles 18% et enfin les campings 5 étoiles, 4%. Ils sont ensuite classés en fonction de leur usage :

- **Tourisme** : si plus de la moitié des emplacements est destinée à la location à la nuitée, à la semaine ou au mois ;
- **Loisirs** : si plus de la moitié des emplacements est destinée à la location pour une durée supérieure à un mois ;

2. https ://www.insee.fr/
3. https ://www.larousse.fr/

— **Aire naturelle** : si le terrain de camping est destiné à l'accueil de tentes, de caravanes et d'autocaravanes, pendant une période d'exploitation n'excédant pas 6 mois par an (continus ou pas), sur des emplacements nus non desservis individuellement en eau ou en électricité et non raccordés au système d'assainissement.

2.2 Le marché du camping vacances en France

2.2.1 Évolution du chiffre d'affaires du secteur

Le secteur de l'hôtellerie de plein air connaît une dynamique exceptionnelle depuis de nombreuses années (figure I.14) avec un chiffre d'affaires qui est passé de 800 millions d'euros (M€) il y a 30 ans à presque 3 milliards d'euros (Md€) aujourd'hui. Malgré la crise sanitaire de 2020, on note une reprise de la fréquentation des campings qui a même franchi un nouveau record en 2022 avec plus de 135 millions de nuitées. Les nuitées correspondent au nombre total de nuits passées par les clients dans un camping. Un couple séjournant trois nuits consécutives dans un camping compte ainsi pour six nuitées, de même que six personnes ne séjournant qu'une nuit (source : INSEE).

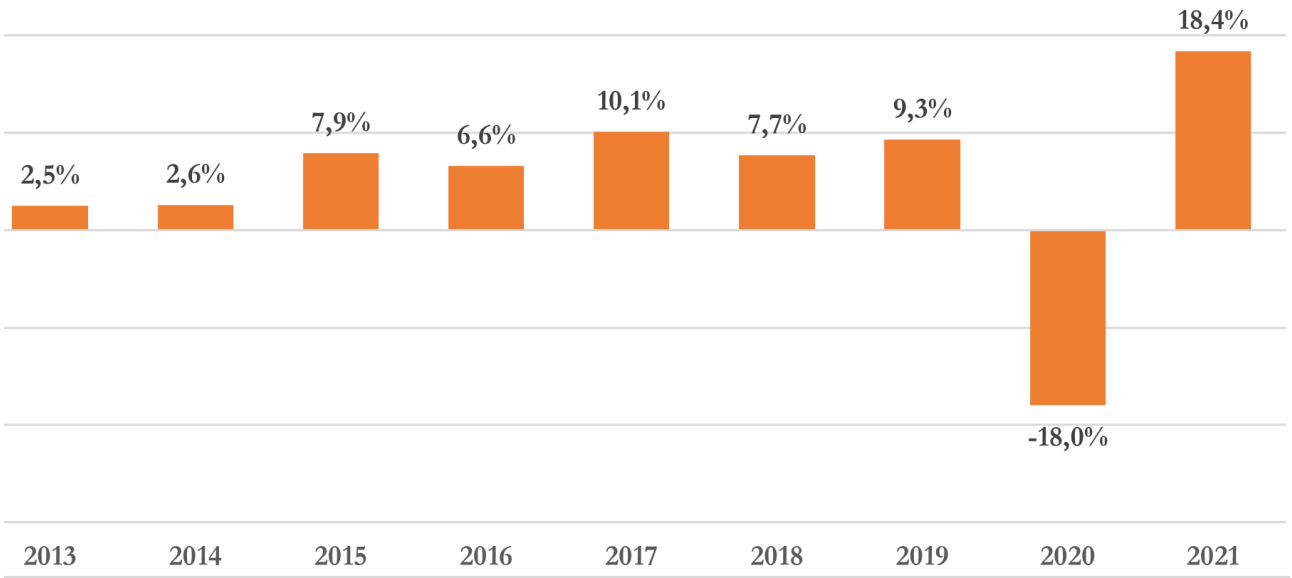


Figure I.14 – Evolution en variation annuelle du CA camping de 2013 à 2021 (source :DGE).

La bonne dynamique du secteur du camping s'explique principalement par sa résilience au contexte inflationniste. Par rapport à d'autres formes d'hébergement touristiques, les campings souffrent nettement moins des baisses du pouvoir d'achat des vacanciers induites par la flambée des prix. Ils tirent en effet parti d'un meilleur rapport qualité/prix avec un niveau de prestations proche des autres modes d'hébergement collectif, mais à des prix sensiblement moins élevés. Ceci malgré les hausses de tarifs pratiquées chaque année pour refléter leur montée en gamme. Ensuite, le secteur du camping en France présente un positionnement « nature ». Les campings constituent une offre unique dans le paysage touristique, proche de la nature et propice aux activités de plein air. Ils s'inscrivent de fait particulièrement bien dans les besoins des vacanciers en matière de « tourisme vert », par ailleurs

amplifiés par la crise sanitaire. Bon nombre d'entre eux prennent à bras-le-corps les problématiques de développement durable pour rendre leurs campings encore plus « verts ». On note des adhésions de plus en plus nombreuses des réseaux à des labels écolos (la clef verte, Refuge LPO, écolabel européen,...). Enfin, le succès du secteur du camping s'explique par sa capacité à élargir sa clientèle et à la fidéliser. L'hôtellerie de plein air (HPA) continue de séduire de nouvelles catégories socio-professionnelles (CSP) habituées à fréquenter d'autres hébergements touristiques grâce à la généralisation des mobil-homes et la multiplication des infrastructures de loisirs. À cela s'ajoutent les néo-campeurs qui ont franchi pour la première fois la porte des campings en 2020 et ont renouvelé l'expérience depuis.

2.2.2 Capacités d'accueil en camping et taux d'occupation

L'offre et les capacités d'accueil des terrains de camping en France ce sont renforcées au cours des dix dernières années avec le développement de nouvelles infrastructures pour améliorer les séjours des campeurs. Selon l'INSEE, l'hôtellerie de plein air constitue la première offre d'hébergement touristique marchand en France avec environ 50% des lits marchands et plus de 7 800 campings.

Le graphique I.15 présente au niveau régional le nombre d'emplacement et taux d'occupation de ces emplacements en 2019. Le taux d'occupation est le rapport entre le nombre d'emplacements occupés et le nombre d'emplacements effectivement offerts.

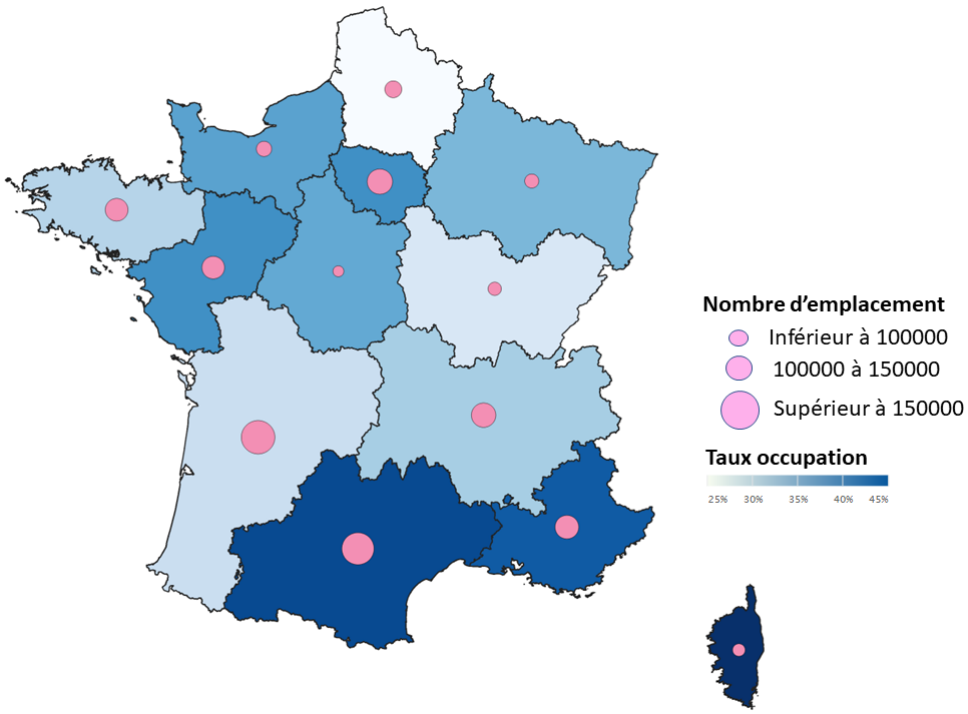


Figure I.15 – Taux d'occupation et nombre d'emplacements camping en 2019 (données INSEE).

On peut observer qu'au 1er janvier 2019, les campings en France possédaient plus de 860000 emplacements répartis dans tous les campings de l'hexagone. Ce sont les régions du sud sur le littoral qui possèdent le plus d'emplacement camping. Il s'agit notamment de la

Nouvelle-Aquitaine qui représente 20% des emplacements et de la région de **l'Occitanie** avec 18% des emplacements. Les régions du sud présentent les taux d'occupation les plus élevés. La **Corse** est la région avec le taux d'occupation le plus élevé (43,7%) suivie de la région **d'Occitanie** (42,4%) et de la région de **Provence-Alpes-Cote d Azur** (41,4%).

Le taux d'occupation moyen des campings de l'hexagone en 2019 atteint 38,0%. Ce taux d'occupation augmente avec le classement des campings : de 31,2% en camping 1 et 2 étoiles, le taux atteint 42,9% en 3,4 et 5 étoiles. Il tombe à 24,6% pour les campings non classés.

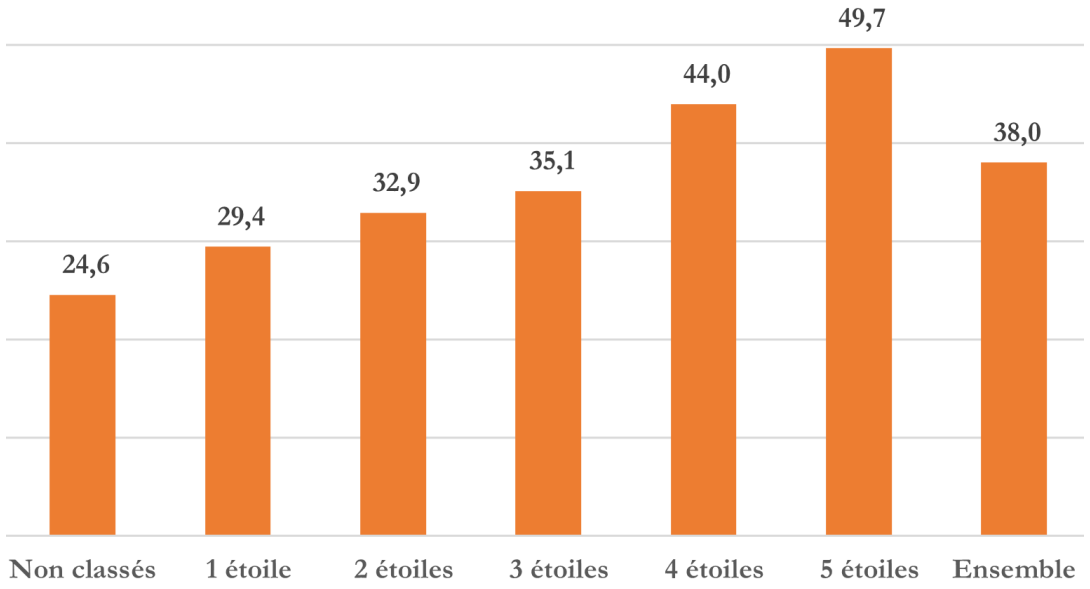


Figure I.16 – Taux d'occupation (%) par type de camping en 2019 (données INSEE).

Le tableau I.3 montre l'évolution du taux d'occupation en période estivale (sauf octobre) de 2014 à 2019 et pour l'année 2021. Il ressort de l'observation de ce tableau qu'en août 2019, les campings de l'hexagone enregistrent un taux d'occupation moyen maximal de 60,8%. En 2019, le taux d'occupation a progressé de 3,1 points par rapport à 2018, pour s'établir à 44,8%. Cette occupation accrue des campings s'observe pour chacun des mois de la saison de 2014 à 2019. C'est en juillet et en août qu'on a les taux les plus élevés en raison d'une météo particulièrement favorable et que les flux de vacances sont plus élevés dans ces deux mois. On note que le secteur du camping retrouve peu à peu son niveau d'avant la crise sanitaire.

	Mai	Juin	Juillet	Août	Septembre	Total
2021	18,9	24,1	42,1	49,3	32,9	34,5
2019	32,7	43,0	57,4	60,8	32,8	44,8
2018	39,4	38,7	53,5	58,8	38,0	41,7
2017	30,5	35,5	45,9	50,2	26,2	37,7
2016	31,1	34,2	44,7	35,3	26,5	34,5
2015	33,7	35,9	52,4	49,7	25,9	39,7
2014	35,1	35,1	49,7	52,6	32,1	41,1

Table I.3 – Evolution du taux d'occupation dans les campings de l'hexagone (en %) (Source : INSEE/DGE)

2.2.3 Fréquentation globale des campings et dépenses en camping

La France est le pays qui abrite le plus de camping, d'ailleurs elle est le premier parc en Europe puisqu'elle dispose de 27% des sites de camping existants. Le nombre de campings enregistrés pour la France la positionne au deuxième rang au niveau mondial. Le graphique I.17 montre l'évolution du nombre de nuitées annuel de 2010 à 2021. Le nombre de nuitées correspond au nombre total de nuits passées par les clients dans un établissement : deux personnes séjournant trois nuits dans un hôtel comptent ainsi pour six nuitées de même que six personnes ne séjournant qu'une nuit.

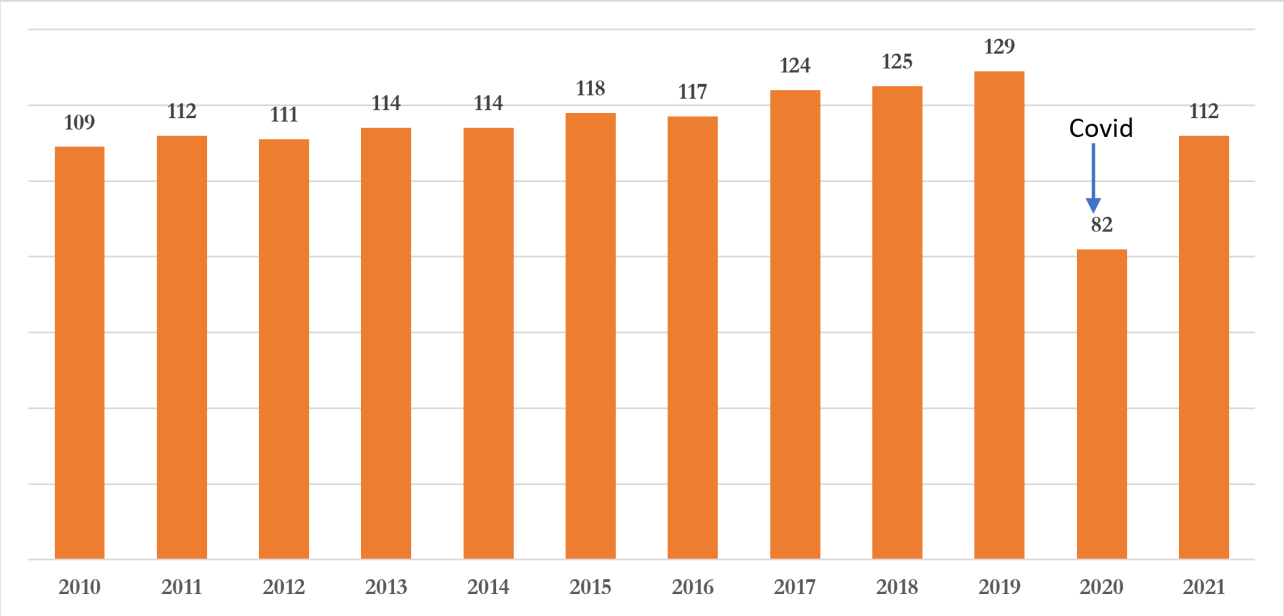


Figure I.17 – Evolution du nombre de nuitées annuel en camping (Données INSEE).

Le secteur du camping en France a généré 129 millions de nuitées en 2019 (graphique I.17), ce qui représente une croissance de 3% par rapport à 2018. Cette hausse est en parti due à la hausse de 2,5% des réservations issues de clients étrangers. Avec la crise sanitaire, on note une baisse d'environ 36% des nuitées. La baisse due à la crise sanitaire a été amorti en particulier par les nuitées nationales. En effet, plus de 88% des nuitées en hôtellerie de plein air ont été réalisées par des touristes français en 2020.

Selon la **Fédération Française de Camping et de Caravaning** (FFCC) les ménages touristiques séjournant en campings en France sont en moyenne de 3.5 personnes, ils résident 6 jours et dépensent en moyenne 45 €/personne/jour soit un budget de séjour de 957 €.

Ces chiffres varient en fonction de la période et de la région. Les Français sont prêts à dépenser beaucoup pour être au bord de la Méditerranée ou de l'Océan Atlantique, et encore davantage pour les eaux cristallines de la Corse. En effet, ils y consacrent des budgets élevés : 1133 € en **Corse**, 996 € en **Provence-Alpes-Cote d Azur** et 927 € en **Aquitaine**. A contrario, **l'Ile de France**, La **Champagne-Ardenne** et la **Bourgogne** restent très économiques avec des paniers moyens oscillant entre 509 et 555 €.

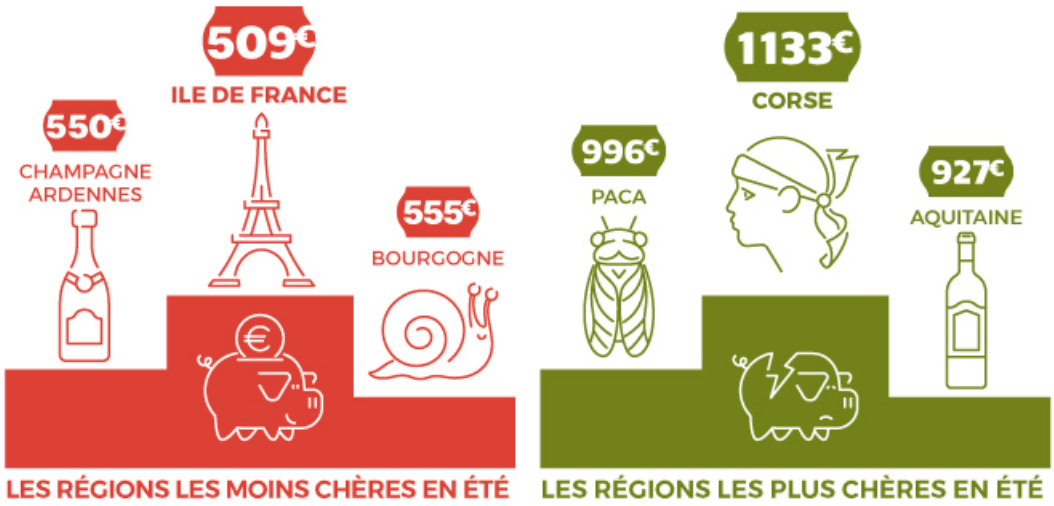


Figure I.18 – Les régions les moins chères et les plus chères en camping (période estivale).

Le tableau I.4 montre l'évolution du nombre de nuitées en camping entre 2017 et 2021 par région. On note qu'en 2021 les nuitées sont concentrées dans les régions du littoral (Nouvelle-Acquitaine (21,9%), Occitanie (21,5%) en raison des fortes chaleurs qui ont marqué l'été 2021. On observe une reprise de fréquentation des campings dans toutes les régions à l'exception de l'Île de France qui peine à retrouver son niveau précédent la crise.

Région	2017	2018	2019	2020	2021
Ile-de-France	1370980	1855912	1900000	1207752	829000
Centre-Val de Loire	1873972	2054102	2200000	1398450	1721000
Bourgogne-Franche-Comte	2891637	2939172	3100000	1970543	2418000
Normandie	3349469	3969358	4400000	2796899	3693000
Hauts-de-France	2136611	2499578	3000000	1906977	1986000
Grand Est	2712765	3152141	3100000	1970543	2153000
Pays de la Loire	11049810	12162722	12900000	8200000	10911000
Bretagne	10580775	11869844	12400000	7882171	10956000
Nouvelle-Aquitaine	24591705	25981669	27100000	17226357	24192000
Occitanie	24969837	26142977	26600000	16908527	24654000
Auvergne-Rhone-Alpes	11769696	12231893	12300000	7818605	10990000
Provence-Alpes-Cote d Azur	15105433	15690914	15500000	9852713	14169000
Corse	4216911	4457056	4100000	2606202	3649000

Table I.4 – Evolution du nombre de nuitées de 2017 à 2021 par région

2.3 Camping, aléa climatique et assurance : les enjeux

Le changement climatique est une réalité qui impacte déjà les territoires urbains et ruraux. L'été 2022 a été celui de tous les records de fréquentations des campings mais aussi celui des événements climatiques extrêmes : incendies en Gironde et dans le Gard, tempête en Corse, sécheresse exceptionnelle, entre autres. Ces sinistres, de plus en plus fréquents et intenses, pèsent sur l'offre de produit d'assurances sur le marché de l'hôtellerie en plein air. Au point que certains campings très exposés se trouvent sans couverture.

Le réchauffement climatique se fait ressentir notamment sur les côtes du littoral français. Les habitations, les commerces, les campings sont donc menacés par la montée du niveau

de la mer. La variabilité météorologique et climatique ont vraisemblablement des impacts directs sur l'occupation et les conditions normales de campings. Le climat devient de plus en plus un critère de choix pour la destination touristique en camping pour les campeurs. Ainsi, face aux épisodes climatiques qui sont de plus en plus violents, le secteur du camping français a connu une transformation au cours des 10 dernières années. Cette transformation a porté notamment sur l'adaptation des emplacements de camping et la protection des personnes. Aujourd'hui, de nombreux défis existent notamment en ce qui concerne l'adaptation des bâtiments et des campings dans l'ensemble. Cette dernière question présente des difficultés car les contraintes sont importantes, notamment sur le plan végétal et légal.

Il existe plusieurs garanties proposées par les assureurs aux clubs de Campings. Ces garanties peuvent être incluses dans un contrat d'assurance multirisques habitation ou faire l'objet d'un contrat d'assurance spécifique. Par exemple, l'assurance de la responsabilité civile destinée à prendre en charge les conséquences financières des dommages que les campeurs peuvent causer à autrui à l'occasion du camping. D'autres garanties peuvent être proposées, notamment l'assurance des dommages au matériel de camping ou encore l'assurance des accidents corporels pendant le camping pour faire face à des dépenses non prises en charge par l'organisme social. Cependant, très peu de garanties sont proposées pour faire face aux effets du changement climatique sur le secteur du camping.

Toutefois, pour garder le niveau d'attractivité de leur camping face aux aléas climatiques, les clubs de vacances proposent de plus en plus des offres de remboursement en cas de conditions climatiques capricieuses en particulier durant l'été. C'est le cas de la garantie nommée "garantie soleil" dont le principe consiste à un remboursement de quelques centaines d'euros en cas d'ensoleillement insuffisant. Par ailleurs, "la garantie soleil" comporte peu de risques météorologiques et est limitée pour atténuer les effets de pertes de la clientèle due au changement climatique. Ainsi, il existe de nombreux défis en matière d'assurance contre les aléas climatiques dans le secteur du camping.

3 Généralités sur l'assurance paramétrique

Depuis quelques années, le digital et les nouvelles technologies permettent de concevoir des produits d'assurance novateurs, dont notamment l'assurance paramétrique qui s'est particulièrement développé dans le secteur agricole. Dans cette partie, nous évoquerons les concepts clés de l'assurance paramétrique en particulier dans le domaine de l'agriculture.

3.1 Définition et principe de l'assurance indicielle

L'assurance paramétrique, nommée également assurance indicielle ou indexée, se base en majeure partie sur des données météorologiques : dès lors qu'une anomalie météorologique est constatée (sur la base d'un indice de pluviométrie, d'une température ou d'autres critères sélectionnés), l'indemnisation du sinistre est enclenchée. Le principe de l'assurance

paramétrique est la construction d'un indice corrélé avec la variable d'intérêt (c'est-à-dire les ventes, rendements agricoles, qualité des cultures, etc.). Cet indice doit cependant respecter certaines propriétés et se doit d'être :

- observable et facilement mesurable ;
- objectif ;
- vérifiable indépendamment ;
- communiqué dans un délai convenable ;
- cohérent dans le temps.

L'essor des technologies d'imagerie spatiale, d'analyse météorologique et des possibilités d'analyse fine par le *big data* permettent de déterminer les primes au plus juste pour les assurés. Ce procédé permet finalement d'aboutir à un produit d'assurance sur-mesure. L'assurance paramétrique présente plusieurs avantages sur une assurance dite traditionnelle : la phase d'indemnisation se voit simplifiée et moins coûteuse, puisqu'il n'est plus nécessaire de faire déplacer de coûteux experts et le processus peut se dérouler en quelques jours.

Néanmoins, la mise en place d'une assurance paramétrique reste complexe et coûteuse car elle nécessite d'une grande expertise technique avant de pouvoir commercialiser le produit. Le calcul de l'indice est la source principale de risque dans l'assurance paramétrique : en cas de corrélation incorrecte au risque à protéger, l'assuré peut être indemnisé sans préjudice. Inversement, si l'indice ne reflète pas bien les dommages subis, l'assuré peut ne pas être remboursé à la hauteur de ses attentes.

3.2 Différents formes d'assurance indicielle

Il existe trois formes d'assurance indicielle dans le secteur agricole :

- assurance indicielle avec indice rendement ;
- assurance indicielle avec indice climatique (précipitation, température, vitesse du vent...) ;
- assurance indicielle agricole satellitaire.

3.2.1 Assurance indicielle avec indice rendement

C'est une assurance basée sur les rendements agricoles. Elle implique la détermination d'un rendement moyen par zone considérée comme l'indice de référence. La détermination de cet indice par les concepteurs implique la collaboration des agriculteurs et est réalisée à partir des données historiques des productions des années précédentes. Si au cours d'une année, il y a une perte du rendement moyen dans la zone, chaque agriculteur ayant souscrit au contrat sera dédommagé d'un montant défini dans le contrat.

Une des limites de ce type d'assurance est que l'assureur doit toujours contrôler si le rendement global de la zone est bien juste pour l'année. S'agissant des coûts de conception, elles peuvent s'avérer très onéreuses. Elles nécessitent des études et expertises coûteuses et

peuvent reposer sur des moyens technologiques importants. Elles nécessitent très souvent des données statistiques précises et couvrant une longue période de temps.

3.2.2 Assurance indicielle avec indice climatique

Dans l'assurance indicielle climatique, le processus d'indemnisation repose sur la variation d'un indice lié à des facteurs météorologiques censés être corrélés aux rendements agricole des assurés d'une même zone. Les différents facteurs météorologiques peuvent être le taux d'humidité dans l'air, le niveau de précipitation, la température par exemple.

L'intérêt de ce type d'assurance dépend de la corrélation entre l'indice choisi et le rendement de l'agriculteur. La plus faible corrélation entre l'indice et le rendement individuel implique un risque de base plus important pour l'agriculteur. Ces indices proviennent le plus souvent de deux sources :

— stations météorologiques au sol ;
— sonnées satellitaires.

3.2.3 Assurance indicielle satellitaire

Dans l'assurance indicielle satellitaire, les principes sont similaires à ceux de l'assurance indicielle climatique à la différence que les indices sont construits sur la base de facteurs observables par imagerie satellitaire telle que l'évapotranspiration ou l'indice de végétation par différence normalisée (IVDN). L'IVDN met en relief la variation de l'absorption de l'humidité des plantes mesurée par leur capacité à réaliser de la photosynthèse.

Contrairement aux autres sources de données, l'imagerie satellite fournit des données détaillées pour des continents sur de nombreuses années. Ces images servent à construire des indices corrélés aux cycles de vie des cultures des régions à assurer. Le risque de base spatial dépendra de la résolution du satellite.

3.3 Risques de bases, aléa moral et anti-sélection

3.3.1 Risques de bases

Le risque de base dans l'assurance indicielle survient lorsque les mesures de l'indice ne correspondent pas aux pertes réelles d'un assuré individuel. Il existe deux principales sources de risque de base dans l'assurance indicielle. Une source de risque de base provient de produits mal conçus et l'autre d'éléments géographiques.

Le risque de base de conception du produit est minimisé grâce à une conception de produits robuste et soutenu par des tests des paramètres du contrat. Le risque de base géographique est un facteur de distance entre le lieu de mesure de l'indice et le champ de production. Plus la distance entre l'instrument de mesure et le terrain est grande, plus le risque de

base est élevé. Certains assurés qui subissent des pertes peuvent ne pas recevoir d'indemnisation tandis que d'autres qui ne subissent aucune perte peuvent recevoir des paiements. Ce risque de base est réduit à mesure que la densité des stations météorologiques et des pixels satellites augmente.

3.3.2 Aléa moral

L'aléa moral, en économie de l'assurance, désigne une situation où l'assureur ne peut anticiper certaines actions entreprises par l'assuré, qui peuvent conduire à une aggravation du risque. Par exemple, un agriculteur peut effectuer de fausses déclarations sur ses récoltes sans que l'assureur puisse le vérifier à moindre coût. Il n'est pas possible pour l'assureur d'inclure une éventuelle condition sur l'effort fourni par l'assuré dans la clause du contrat, dans la mesure où cet effort est inobservable et, ainsi, non quantifiable.

L'assurance basée sur un indice supprime cet aléa puisque tous les assurés se voient proposer une indemnisation en échange du même taux de primes et qu'un bénéficiaire ne peut à priori pas influencer une température moyenne d'une station météorologique.

3.3.3 Anti-sélection

L'anti-sélection désigne le phénomène d'asymétrie d'information en faveur de l'assuré, qui peut détenir une connaissance de son propre risque que l'assureur ignore. Dans la plupart des contrats d'assurance traditionnelle, l'assureur est confronté au problème d'anti-sélection. L'assuré connaît mieux son risque que l'assureur et pourra souscrire à une police d'assurance d'autant plus facilement qu'il sera susceptible d'avoir un sinistre.

En assurance indicielle se risque d'anti sélection se voit diminuer. Causé par l'asymétrie d'information, ce risque peut voir un assureur mal informé proposer un contrat qui repousse les personnes informées (qu'il voudrait attirer pour contractualiser) et n'intéresser que celles avec qui il souhaite le moins développer commercialement. Dans le cadre de l'assurance paramétrique, lorsque le contrat d'assurance est signé suffisamment en avance, l'assuré ne détient pas plus d'informations sur le risque couvert par le contrat d'assurance que la compagnie à laquelle il a transféré le risque.

3.4 Indemnisation, prime pure et prime commercial

L'indemnité d'assurance correspond à la somme d'argent versée par l'assureur à l'assuré ou à un tiers et visant à réparer un dommage résultant d'un sinistre garanti au titre du contrat d'assurance. En assurance indicielle, les assurés reçoivent les indemnités à la suite du déclenchement d'un indice. La figure I.19 montre un exemple de fonction d'indemnisation en assurance paramétrique basée sur l'indice de précipitation.

La prime pure c'est le montant du sinistre moyen auquel devra faire face l'assureur pour le risque. Mathématiquement, la prime pure est égale à l'espérance mathématique

des pertes. En assurance indicielle la prime d'assurance sera fixée en fonction de certains paramètres dans le contrat comme le niveau de couverture qui définit le seuil de l'indice.

La prime commerciale finale doit être gonflée pour tenir compte de l'incertitude entourant les données, du coût de réassurance, des marges de l'assureur (y compris les frais de distribution et frais généraux), et de tout autre coût lié aux pratiques commerciales.

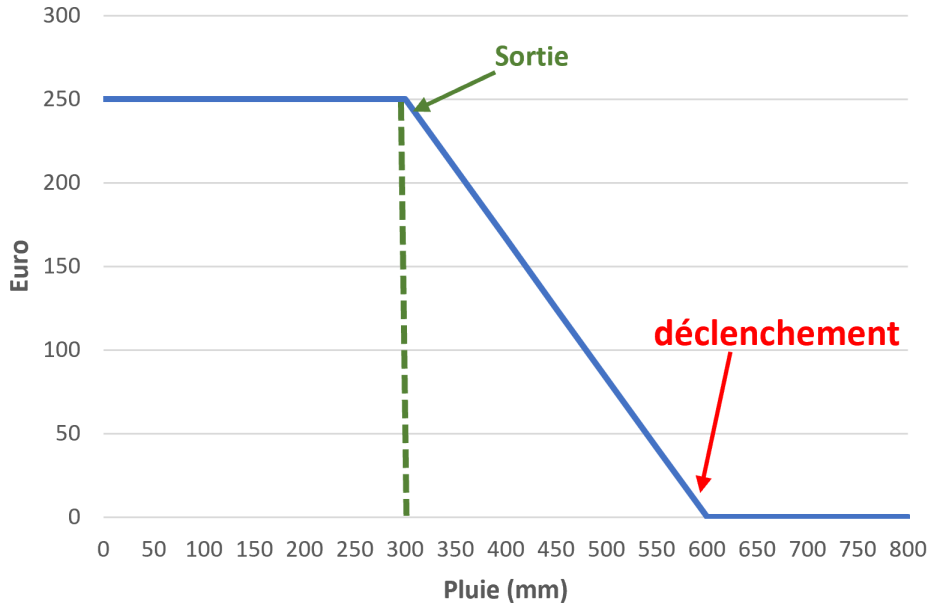


Figure I.19 – Schéma classique d'assurance paramétrique

3.5 Méthode de *pricing* en assurance indicielle climatique

Deux méthodes de *pricing* sont les plus utilisées en assurance paramètre. Il s'agit de l'approche par le coût moyen attendu et l'approche *Monte Carlo*.

3.5.1 Approche *Burning Cost* ou coût moyen attendu

C'est l'approche la plus simple pour le calcul de la prime. Il s'agit d'évaluer les indemnisations pour chacune des années de la chronique de données disponibles. Le coût moyen attendu, également connu comme prime de risque pure, peut être calculé en faisant la moyenne des indemnités historiques potentielles qui auraient été versées par la structure du contrat au cours de la période en question. Un produit d'assurance ayant un coût moyen attendu supérieur serait plus coûteux, car il octroierait des indemnités plus élevées et plus fréquentes. Le coût moyen du risque est simplement exprimé par :

$$\text{Coût Moyen} = \frac{1}{N}\sum_{i=1}^{N} I(i) \tag{I.5}$$

avec $I(i)$ l'indemnité historique.

Cette première approche a plusieurs limites. En effet, elle ne tient pas compte du caractère aléatoire du risque, de l'aspect solvabilité et probabilité de ruine. Aussi cette approche

est très limitée dans un contexte de profondeur d'historique très réduite, ce qui en réalité n'est pas propice pour des estimations non biaisées.

3.5.2 Approche actuarielle et *Monte Carlo*

Une alternative à l'approche par le coût moyen est l'approche stochastique. Cette approche consiste à calibrer une loi de distribution statistique continue avec celle des valeurs historiques de l'indemnité, il est possible ensuite d'utiliser la loi de distribution pour prévoir les réalisations possibles de cette indemnité. Même si la simulation *Monte Carlo* paraît moins aisée à mettre en place, elle est plus sophistiquée. L'approche *Monte Carlo* repose sur la loi forte des grands nombres, et le fait de répéter un grand nombre de tirages aléatoires permet de converger presque sûrement vers la valeur de la prime pure la plus juste possible. Le coût moyen dans de cas est exprimé par :

$$\text{Coût Moyen} = \int I(x) d\mu(x) \tag{I.6}$$

où $\mu(x)$ est la fonction de distribution de l'indemnité à déterminer.

4 Le régime « beau temps » : conception et tarification

Les clubs de camping français font face à la baisse du niveau d'attractivité de leur camping. Cette baisse est imputable aux changements climatiques. Ainsi, pour garder le niveau d'attractivité de leur camping face aux aléas climatiques, les clubs ont besoin d'intégrer dans leur offre des garanties qui puissent indemniser leur client. Dans cette section il sera question de présenter la phase de conception et de tarification du régime « **beau temps** » dont l'objectif est d'indemniser un campeur en cas de mauvais temps.

4.1 Fonctionnement et schéma conceptionel du produit

Sur le modèle d'une assurance paramétrique, le régime proposerait un remboursement de 50 €, en cas de dépassement d'un ou plusieurs indices associés aux risques de température, vent et pluie pour un individu, et par nuitée, ayant souscrit au contrat. Pour souscrire au contrat de ce régime, l'individu paiera une prime qui couvrira une partie ou la totalité de son séjour en camping. La grille tarifaire du régime sera exprimée en prix unitaire par jour ou encore par nuitée. Par exemple, pour une personne souhaitant être couvert pendant une semaine, la prime serait multiplié par 7 fois le prix unitaire. De même pour un couple souhaitant être couvert, ayant passée une nuitée dans un camping, la prime serait multipliée par deux fois le prix unitaire.

La conception de notre régime « **beau temps** » est fait en deux grandes phases. La première phase consiste principalement à la tarification du produit sur la période 2011 à 2015 et la deuxième consiste au lancement du produit sur la période 2016 à 2021. Pour la tarification,

nous avons choisi une fonction d'indemnisation binaire qui consiste à indemniser le campeur de 50 €. Le choix des 50 € se justifie principalement par le fait que selon la Fédération Française de Camping et de Caravaning (FFCC) les dépensent moyennes journalière par personne en camping est d'environ 45 €. Le mécanisme d'indemnisation du régime « **beau temps** » est décrit pour chaque risque dans le tableau suivant :

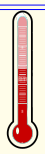
Résumé du mecanisme de declenchement de l'indemnisation		
Risque/Intempérie	**Seuil de déclenchement**	**Indemnisation**
Température 	La Température T_{max} du jour a été supérieure au quantile 95% de la distribution de T_{max} en été ou la température T_{min} du jour a été supérieure au quantile 95% de la distribution de T_{min} pour les autres saisons	50 €
Pluie 	Le volume d'eau du jour (mm) supérieure au quantile 95% de la distribution du volumne d'eau observée	
Vent 	La vitesse moyenne du jour supérieure au quantile 95% de la distribution de la vitesse de vent journalière observée	

Table I.5 – Mécanisme d'indemnisation du régime « **beau temps** »

Si on désigne par $\mathbb{1}_T(i)$ la fonction indicatrice de dépassement (1 en cas de dépassement 0 sinon) associée au risque température, $\mathbb{1}_P(i)$ celle associée au risque pluie et $\mathbb{1}_V(i)$ celle associée au risque vent, la fonction d'indemnisation totale pour pour un individu, et par nuitée, ayant souscrit au contrat s'écrit de la manière suivante :

$$I_{totale}(i) = 50 \times max(\mathbb{1}_T(i), \mathbb{1}_V(i), \mathbb{1}_P(i)) \tag{I.7}$$

4.2 Présentation des données

4.2.1 Données météorologiques SYNOP (Météo France)

Dans le cadre de ce mémoire on travaillera sur les données ponctuelles SYNOP (données d'observations issues des messages internationaux d'observation en surface) issues des stations météorologiques de Météo France.

Météo France est le service officiel météorologique et climatologique national. Affilié au ministère de la transition écologique et solidaire, sa mission première est d'assurer la sécurité météorologique des personnes et des biens. Elle est notamment en charge de la prévision et l'étude des phénomènes météorologiques pour les territoires français de métropole et d'outre-mer. Météo France s'assure de l'élaboration de cartes de vigilance météorologique dans le but de prévenir les phénomènes dangereux, leurs conséquences et les précautions à prendre pour se protéger. Les principales missions de Météo France sont les suivantes :

— le développement et la maintenance d'un réseau d'observation ;
— la collecte et le traitement de données climatologiques ;
— la prévision du temps et l'élaboration de projections climatiques ;
— la recherche dans les domaines de la météorologie et du climat.

Dans le cadre de ses missions de service public, Météo-France produit et diffuse quotidiennement un très grand volume d'informations. Un grand nombre d'entre elles peuvent être réutilisées. Parmi elles, les données d'observations issues des messages internationaux d'observation en surface (SYNOP) circulant sur le système mondial de télécommunication (SMT) de l'organisation météorologique mondiale (OMM). Les données SYNOP sont des données journalières pour 42 stations réparties en France métropolitaine. Elle contiennent :

— les paramètres atmosphériques mesurés tels que la **température**, l'humidité, la direction et **force du vent**, la pression atmosphérique et la **hauteur de précipitations** ;
— les paramètres atmosphériques observés tel que le temps sensible, la description des nuages et la visibilité.

D'autres paramètres peuvent être disponibles tels que la hauteur de neige ou l'état du sol en fonction de l'instrument de mesure utilisé.

Dans les fichiers SYNOP nous nous sommes intéressés à la variable « température », la variable « vitesse du vent moyen 10 mètre », les variables de « Précipitations dans les N dernières heures ». C'est la variable précipitation dans les 24 heures qui sera utilisée dans cette étude (voir tableau D.1 en annexe pour la liste des variables présentent dans les données SYNOP).

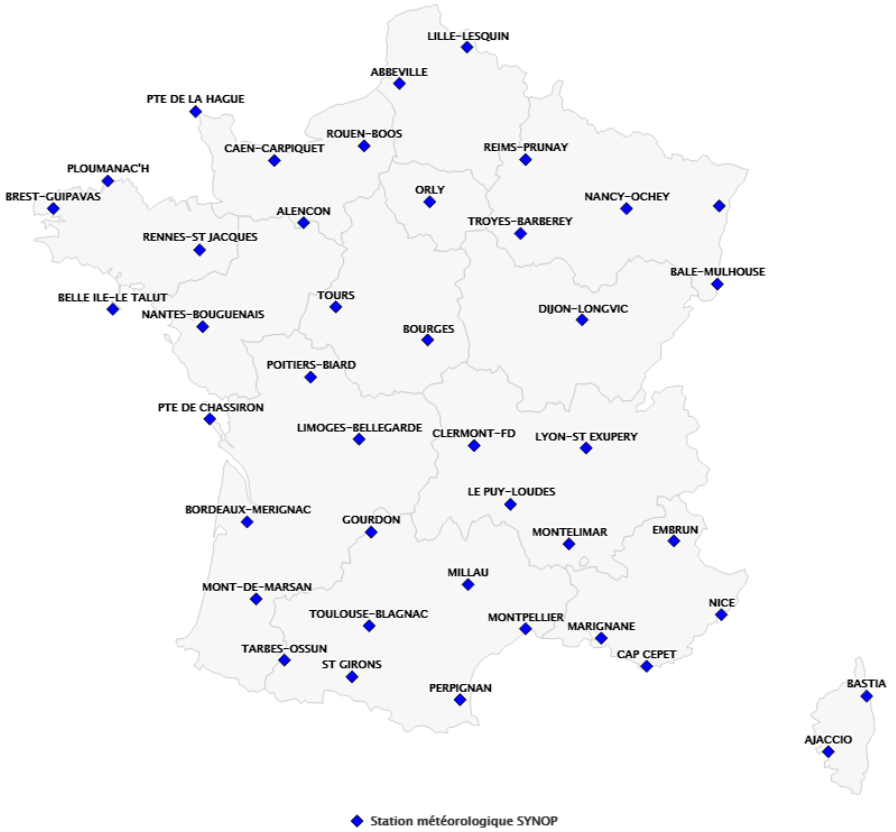


Figure I.20 – Localisation des stations météorologiques en France métropolitaine

Dans la figure I.21, est représentée la liste des 42 stations météorologiques en France métropolitaine.

Nous disposons des données SYNOP en série journalière entre janvier 2001 et décembre 2021, soit 21 années d'historiques.

4.2.2 Données sur les campings

Pour les données de camping on dispose dans un premier temps des données sur le nombre de nuitées de camping par département en 2016 ainsi que leur taux d'évolution entre 2010 et 2016. Ensuite, nous avons récupérer les données sur le nombre de camping de nuitée totale par région de 2017 à 2021 dans les rapports annuels sur la fréquentation des campings publié par **CRT normandie tourisme**[4]. Enfin, on dispose des nuitées dans l'hôtellerie en France métropolitaine par mois en 2011 à 2019. La structure temporelle des nuitées dans l'hôtellerie nous permettra de donner un caractère temporel aux données de nuitées en camping.

Ces données sont issues de l'enquête de fréquentation dans l'hôtellerie de plein air réalisée par INSEE. Cette enquête vise à étudier la fréquentation dans l'hôtellerie de plein air en France, tant en volume qu'en termes de structure de la clientèle, notamment géographique. Elle est administrée chaque mois de mai à septembre (de avril à septembre depuis 2017) auprès d'un échantillon représentatif de campings possédant au moins un emplacement de passage. Les résultats sont ensuite extrapolés à l'ensemble des campings. Elle permet également la connaissance exhaustive du parc (y compris les campings classés n'offrant aucun emplacement de passage) et de la fréquentation des seuls emplacements de passage, destinés à une clientèle touristique, par opposition aux emplacements loués à l'année destinés à une pratique plus résidentielle.

Au fil des années, l'enquête de fréquentation dans l'hôtellerie de plein air a connu plusieurs révisions dont :

- 2006 : rénovation de la méthode de redressement pour mieux traiter les non-réponses. Introduction d'une nomenclature des pays d'origine des touristes étrangers ;
- 2011 :, mise en place du nouveau classement Atout France, de 1 à 5 étoiles ;
- 2013 : révision de la méthode de redressement pour prendre en compte la nouvelle stratification géographique (Espaces Touristiques Nationaux) ;
- 2014 : prise en compte du nouveau classement Atout France. Les données sont rétropolées depuis 2010 ;
- 2017 : l'enquête de fréquentation dans les campings a été étendue au mois d'avril sur l'ensemble du territoire de France métropolitaine.

Afin d'autoriser des comparaisons dans le temps, les nuitées d'avril 2010 à 2016 de France métropolitaine et dans les régions ont été estimées. Le total annuel des années 2010 à 2016 a également été recalculé pour prendre en compte le mois d'avril.

4. https ://pronormandietourisme.fr/

4.2.3 Les retraitements sur les données

Le niveau de granularité de notre étude étant la région et le jour, nous devons avoir une série journalière pour toutes nos variables par région. Ainsi, nous avons procédé à des retraitements sur la base SYNOP et les bases de Campings.

Sur les données de SYNOP

Sur les données SYNOP nous avons regrouper les données en région. Pour cela nous avons agrégé les données des stations se trouvant dans la même région. Pour la température nous avons appliqué la moyenne des températures moyenne journalière, pour le vitesse du vent nous avons également appliqué la moyenne de la vitesse de vent journalière, et pour la précipitation nous avons calculer le volume d'eau total journalière moyenne.

Sur les données de Campings

Sur les données de Campings nous avons appliquer plusieurs traitements dont :

— **Etape 1** : **Passage à une dimension régionale de 2011 à 2021**.

Comme nous l'avons spécifié, nous disposons dans un premier temps du nombre de nuitées par département en 2016 et de leurs taux d'évolution global entre 2016 et 2010. Le graphique suivant met en exergue ce taux d'évolution global.

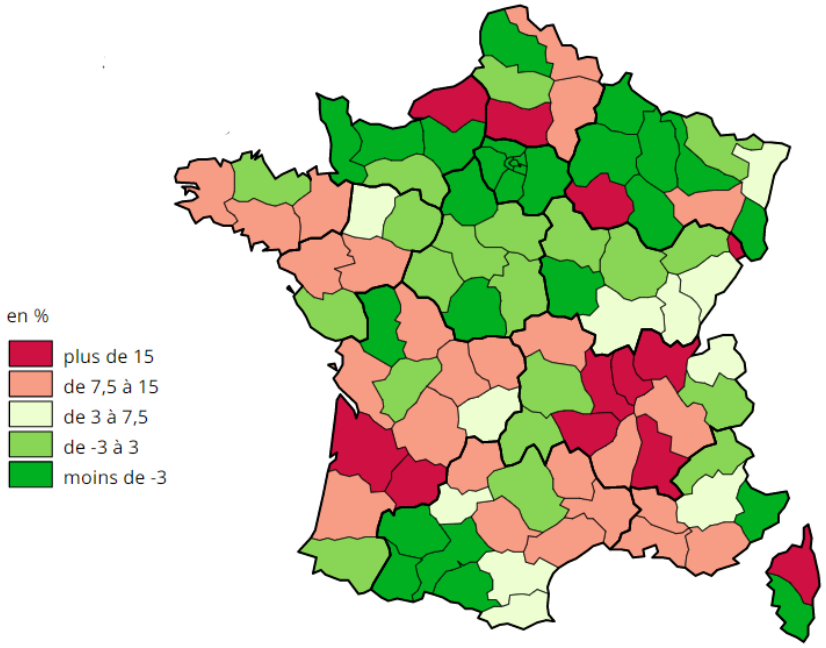


Figure I.21 – Évolution du nombre de nuitées dans les campings par département entre 2010 et 2016 (source : INSEE)

Soit $Tx(d)$ le taux d'évolution global entre 2016 et 2010 du département d. Le taux d'évolution moyen annuel entre 2016 et 2010 pour le département d est le suivant :

$$Tx_{moyen}(d) = (1 + Tx(d))^{\frac{1}{6}} - 1 \qquad \text{(I.8)}$$

On désigne par $Nu(annee, d)$ le nombre de nuitées de l'année $annee$ pour le département d. Avec les taux moyens annuels obtenus, on calcul les nuitées par région pour les années 2011 à 2015 grâce à la formule suivante :

$$Nu(annee, d) = \frac{Nu(2016, d)}{(1 + Tx(d))^{2016-annee}} \tag{I.9}$$

Ensuite on groupe les nuitées des départements pour avoir des nuitées au niveau régional.
Soit $Nu(annee, R)$ le nombre de nuitées pour l'année $annee$ et pour la région R. Ce nombre est calculé pour les années 2011 à 2016 par la formule suivante :

$$Nu(annee, R) = \sum_{d \in R} Nu(annee, d) \tag{I.10}$$

On complète la base avec le nombre de nuitées par région entre 2017 et 2021.

— **Etape 2** : **Passage de la base de région à région et mois de 2011 à 2021**

L'étape précédente a permise d'avoir une approximation du nombre de nuitées en camping par région de 2011 à 2021. Pour avoir une approximation de ce nombre par mois et par région nous avons utilisée la structure par mois du nombre de nuitées en hôtellerie de 2011 à 2021. Cette hypothèse est forte mais semble acceptable en particulier en période estivale où la tendance de fréquentation en hôtellerie est comparable à celle en camping.

Soit $Nh(mois, annee, R)$ le nombre de nuitées en hôtellerie pour un mois, une région et une année donnée. On obtient ainsi, le nombre de nuitées en camping pour un mois, une région et une année donnée par la formule suivante :

$$Nu(annee, mois, R) = \frac{Nu(annee, R) * Nh(mois, annee, R)}{\sum_{mois} Nh(mois, annee, R)} \tag{I.11}$$

— **Etape 3** : **Passage de la base de région mois à région et jour de 2011 à 2021**

Le passage des nuitées d'une granularité mensuelle à une granularité journalière est beaucoup plus délicat. En effet, la période de campings est plus concentrée en période de vacance estivale. Ainsi, pour un mois donnée les jours de vacances scolaires, les jours de week-end et les jours en semaine hors week-end n'ont pas les mêmes niveaux de fréquentation en nuitées. Pour se faire nous définissons la notion **d'équivalence en nuitée** notée EQ qui pour un jour donnée représente le poids relatif du jour j dans le nombre de nuitées total mensuel. La formule de **l'équivalence en nuitée** pour le jour j est la suivante :

$$EQ(j) = (1 - we(j)) + \alpha_{we} * we(j) + \alpha_{vacs} * vacs(j) \tag{I.12}$$

Avec :

$$we(j) = \begin{cases} 1 & \text{si } j \text{ est un jour week-end,} \\ 0 & \text{sinon,} \end{cases}$$

$$vacs(j) = \begin{cases} 1 & \text{si } j \text{ est un jour de vacance scolaire,} \\ 0 & \text{sinon,} \end{cases}$$

Les vacances scolaires comprennent les vacances de la Toussaint, de Noël, d'hiver, de printemps et d'été et diffères selon 3 zones géographiques (voir figure D.1 en annexe) et la Corse. Nous avons utilisé la fonction *is_holiday* du package R vacancesscolr (2019) pour savoir si un jour XX/XX/XXXX est un jour de vacance scolaire ou non.

α_{we} et α_{vacs} représentent respectivement la fréquentation supplémentaire en week-end versus hors week-end et la fréquentation supplémentaire en vacance versus hors vacance. Les paramètres α_{we} et α_{vacs} ont été déterminé via *la valeur cible* sur Excel (voir annexe B). On obtient $\alpha_{we} = 1.33$ et $\alpha_{vacs} = 2.74$. Le nombre de nuitées journalières est :

$$Nu(j) = \frac{EQ(j) * Nu(annee, mois, R)}{\sum_{j \in mois} EQ(j)} \tag{I.13}$$

4.3 Tarification du produit

4.3.1 Principe de Tarification en assurance IARD

Avant de passer au calcul de la prime pure, nous allons tout d'abord introduire les fondements théoriques de l'assurance non-vie, La tarification constitue l'un des cœurs de métier de l'actuaire. C'est une étape qui permet aux compagnies d'assurance d'évaluer les risques auxquelles elles doivent faire face. Elle se base sur l'estimation les flux futurs des prestations à verser afin de déterminer le montant des engagements probables au début d'un exercice. Dans le contexte du modèle collectif, on note la charge financière totale pour la période considérée par S tel que :

$$S = \sum_{i=1}^{N} X_i \tag{I.14}$$

Avec : - N est une variable aléatoire discrète représentant le nombre de sinistres. - $(X_i)_{i \in \mathbb{N}}$ une suite de variables aléatoires réelles.

La variable aléatoire S suit une loi composée avec la convention S = 0 lorsque N = 0. La prime pure est donnée par l'espérance de la charge totale des sinistres auxquels l'assureur devra faire face, c'est un montant déterministe noté mathématiquement par $E[S]$. Elle s'agit d'une prime pure relatif aux coûts des sinistres payés par la compagnie et n'intègre pas les autres coûts d'administration ; frais de gestion et autres frais que la compagnie va débourser afin d'assurer son activité. Pour simplifier le calcul de la prime pure, nous considérons que

le nombre d'événements est indépendant des paiements. De même, nous supposons que les paiements par événement sont indépendants et identiquement distribués. Dans ce cas la prime pure s'écrit :

$$E(S) = E(N)E(X_1)$$

en effet

$$E[S] = E[E[S \mid N]]$$

$$E[S] = E\left[E\left[\sum_{i=1}^{N} X_i \mid N\right]\right]$$

par indépendance :

$$E[S] = E[N]E[X_1]$$

Dans notre cas, N est le nombre de déclenchements annuels, $E(N)$ peut être estimé à partir des données historiques.

4.3.2 Prime d'assurance indicielle du régime "beau temps" et grille tarifaire

La prime pure correspond au montant du sinistre moyen auquel devra faire face l'assureur pour le risque.

$$\pi_{mois,region} = \frac{\sum_j I(m,r,j) * Nu(m,r,j)}{\sum_j Nu(m,r,j)} \tag{I.15}$$

Pour s'assurer du bon fonctionnement du régime, il est nécessaire de prélever des chargements.

Les chargements se présentent comme suit :

— **Chargement de gestion** : ce chargement est destiné à couvrir les frais de l'assureur dans la gestion du nouveau produit : Informatique, direction générale, fonctions centrales, masse salariale, etc.

— **Chargement de distribution** : il est destiné à couvrir les frais de distribution du nouveau produit (publicité, commerciaux, etc.), il est fixé tout simplement en ligne avec l'expérience du produit existant.

— **Chargement généraux** : ce chargement a pour objectif de couvrir les erreurs liées au modèle, les erreurs dans les données...

Chargement	Valeurs
Charges de gestion	λ des prime pure
Chargement de distribution	α des prime pure
Frais généraux	β des prime pure

Table I.6 – Chargement de la prime pure

La prime chargée totale par région et mois s'obtient par :

$$\pi'_{mois,region} = (1 + \lambda + \alpha + \beta) * \pi_{mois,region} \tag{I.16}$$

Pour calculer la prime commerciale, les charges de gestion et les frais généraux sont de l'ordre de 5% des primes, ces charges comme vont jusqu'à atteindre 20% en assurance classique et sont de l'ordre de 5% voire nulles en assurance paramétrique.

Quant aux chargements de distribution ils sont en principe fixés par rapport à l'expérience du marché du régime. Dans le cadre de cette étude nous le fixons également à 5% de la prime pure.

La figure I.22 ci-dessous représente la grille tarifaire du régime « beau temps ». Cette grille donne le prix unitaire que doit payer le souscripteur pour se couvrir durant une nuitée. Ainsi, pour une personne souhaitant effectuer deux jours de camping en **Normandie** durant le mois Août, la prime serait de 14,76€ soit 2 fois le prix unitaire de 7.38 € sans chargement éventuel.

Les primes les plus élevées et dépassant les 12€ sont les tarifs des mois de mars et de novembre dans les régions suivantes : **Normandie**, **Hauts-de-France** et **Grand Est**. Ces primes sont particulièrement plus élevées pour ces mois et ces régions parce que les indices que nous avons défini sont plus susceptible de se réaliser durant ces mois et pour ces régions. A contrario, les mois de septembre, octobre et mai sont les mois qui présentent les tarifs les plus faibles sur la grille tarifaire et cela quelle que soit la région considérée (prime inférieure à 10 €).

Grille tarifaire

	Ile-de-France	Centre-Val de Loire	Bourgogne-Franche-Comte	Normandie	Hauts-de-France	Grand Est	Pays de la Loire	Bretagne	Nouvelle-Aquitaine	Occitanie	Auvergne-Rhone-Alpes	Provence-Alpes-Cote d Azur	Corse
Janvier	3.32 €	2.68 €	10.31 €	6.03 €	12.35 €	8.74 €	5.11 €	11.61 €	2.29 €	2.26 €	7.43 €	5.06 €	4.91 €
Février	4.93 €	7.19 €	11.22 €	7.25 €	7.88 €	9.30 €	6.93 €	10.82 €	6.85 €	5.40 €	9.00 €	5.25 €	4.67 €
Mars	11.95 €	13.14 €	13.14 €	20.47 €	21.66 €	23.11 €	10.64 €	13.93 €	6.89 €	8.16 €	11.42 €	10.97 €	6.48 €
Avril	4.13 €	6.83 €	5.08 €	7.17 €	5.55 €	2.92 €	5.65 €	9.80 €	3.46 €	4.83 €	4.45 €	10.39 €	2.55 €
Mai	2.99 €	4.46 €	4.84 €	2.92 €	6.25 €	3.59 €	4.60 €	8.04 €	2.16 €	3.07 €	4.39 €	7.58 €	4.00 €
Juin	9.84 €	4.35 €	7.65 €	9.78 €	9.10 €	3.11 €	4.66 €	12.22 €	6.55 €	5.69 €	3.84 €	8.43 €	1.94 €
Juillet	8.05 €	6.30 €	10.03 €	9.41 €	10.43 €	5.84 €	5.12 €	10.50 €	4.02 €	6.48 €	8.60 €	18.69 €	11.87 €
Août	6.79 €	4.86 €	6.46 €	7.38 €	7.78 €	4.22 €	5.18 €	10.12 €	6.66 €	5.70 €	6.38 €	13.17 €	8.39 €
Septembre	2.36 €	2.47 €	3.62 €	3.52 €	4.53 €	0.94 €	2.35 €	4.55 €	2.35 €	1.69 €	4.87 €	5.59 €	3.05 €
Octobre	3.38 €	3.49 €	7.52 €	8.81 €	5.13 €	6.33 €	4.42 €	7.96 €	3.07 €	3.50 €	7.40 €	8.91 €	5.84 €
Novembre	8.36 €	7.72 €	16.81 €	15.34 €	12.53 €	14.22 €	7.10 €	15.64 €	7.99 €	9.54 €	11.42 €	14.46 €	5.02 €
Décembre	3.85 €	3.40 €	6.89 €	12.07 €	10.03 €	5.92 €	7.04 €	14.22 €	4.81 €	1.25 €	2.71 €	2.93 €	1.55 €

Figure I.22 – Prime pure du régime « **beau temps** ».

En observant la contribution relative de chaque risque à la formation de la prime pure (graphique I.23), le risque température est le risque qui contribue à la prime pure pour les

régions et mois où les tarifs sont les plus élevés. Par ailleurs, pour les régions les régions du littéral (par exemple **Bretagne**), c'est le risque vent qui contribue le plus à la formation de la prime pure. Par contre, dans la région de **Corse** la prime pure est portée par le risque de pluie car beaucoup plus fréquent dans cette région à l'exception des mois de juillet et d'août où il fait beaucoup chaud.

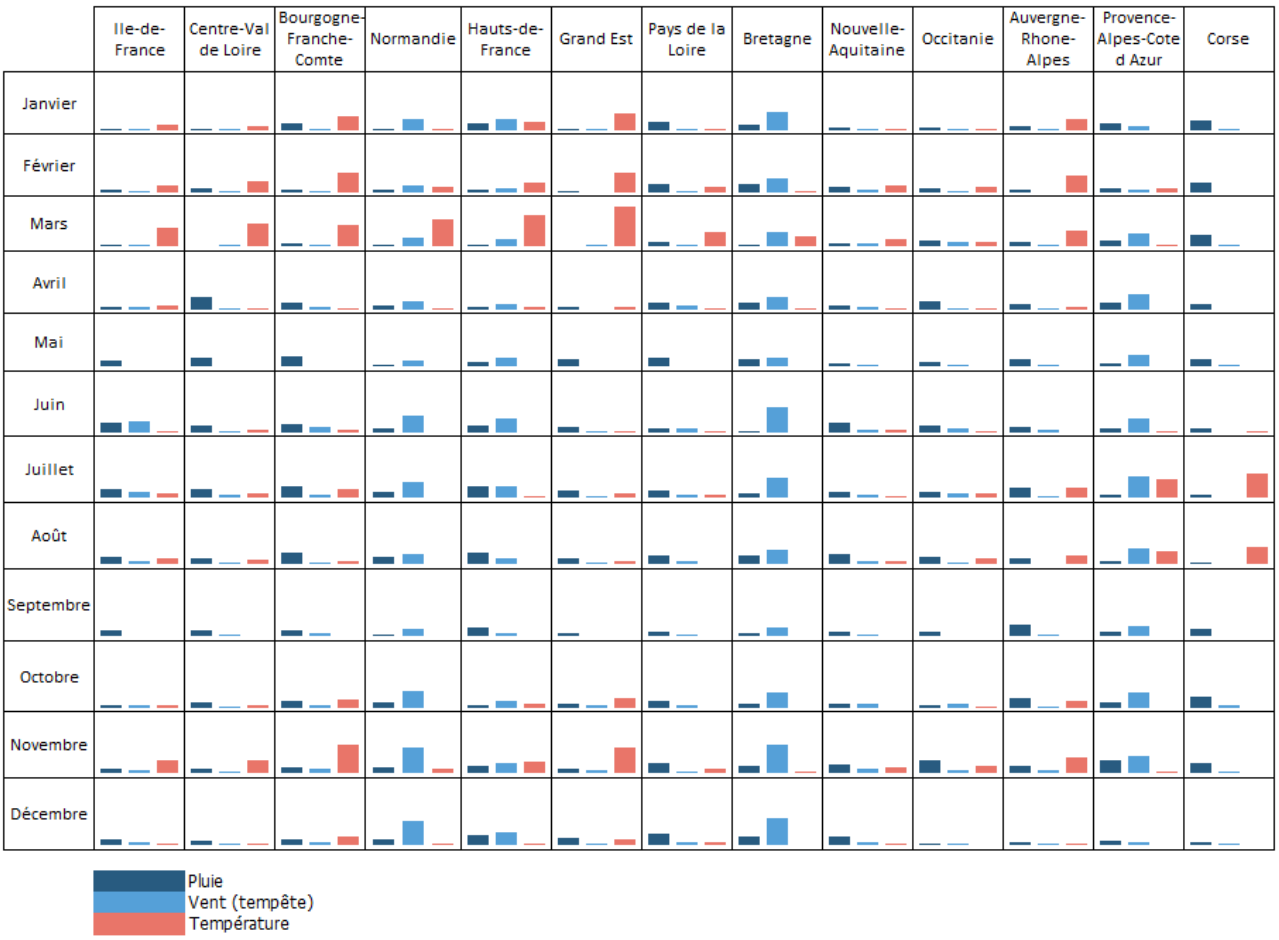


Figure I.23 – Contribution relative de chaque risque à la formation de la prime pure.

5 Résultat du régime et suivi du ratio S/P entre 2016-2021

Dans la partie précédente il était question d'établir la grille tarifaire du régime « **beau temps** ». Une fois, cette grille tarifaire mise en place, nous passons à la phase de lancement du produit entre 2016 et 2021. Dans cette partie, nous analyserons la fiabilité du régime à travers le suivi du compte de résultat technique et du ratio de sinistre sur prime (ratio S/P) entre 2016 et 2021.

Les résultats techniques correspondent à la différence entre les revenus de l'assureur, principalement les primes et les charges constituées par les prestations et les différents frais de gestion et de commercialisation des contrats. Les ratios S/P appelés aussi taux de sinistres permettent d'évaluer la santé et la rentabilité d'une compagnie d'assurance. Il se calcule

ainsi : coût des sinistres divisé par les primes perçues. Une entreprise perçoit des primes plus élevées que les montants payés en cas de sinistre, par conséquent, des ratios de sinistres élevés peuvent indiquer qu'une entreprise est en difficulté financière.

Soit $n(j,a)$ le nombre d'assuré que le régime couvre le jour j pour l'année a. Cette quantité est une proportion du nombre de nuitée journalière. Pour la suite, on suppose que :

$$n(j,a) \sim \mathcal{N}(\mu(a), \sigma(a)) * Nu(j) \tag{I.17}$$

Avec :

— $\mu(a)$ le taux de souscription pour l'année a

— $\sigma(a)$ l'incertitude autour du nombre d'assuré pour l'année a

On suppose que $\mu(a)$ est croissante avec l'année entre 2016 et 2021 avec une incertitude également $\sigma(a)$ croissante. Les valeurs supposées de ces paramètres sont présentes dans le tableau suivant :

Année	μ	σ
2016	0.15	0.03
2017	0.17	0.034
2018	0.18	0.036
2019	0.2	0.04
2020	0.23	0.046
2021	0.25	0.05

Table I.7 – paramètres du taux de souscription par année

Pour une année donnée, le résultat technique du régime s'écrit de la manière suivante :

$$\begin{aligned} R(a) &= \sum_{j,m,r} \pi'_{m,r} * n(j,a) - \sum_{j} I(j) * n(j,a) \\ &= (1+\lambda+\alpha+\beta) * \sum_{j,m,r} \pi_{m,r} * n(j,a) - \sum_{j} I(j) * n(j,a) \\ R(a) &\sim \mathcal{N}(\mu(a), \sigma(a)) \left[(1+\phi) * \sum_{j,m,r} \pi_{m,r} * Nu(j) - 50 \sum_{j} max(\mathbb{1}_T(j), \mathbb{1}_V(j), \mathbb{1}_P(j)) * Nu(j) \right] \end{aligned}$$

avec $\phi = \lambda + \alpha + \beta$; j, m et r désignent respectivement le jour, le mois et la région.

Il s'ensuit :

$$R(a) \sim \mathcal{N}(E(R(a)), \sigma(R(a))$$

La probabilité de perte est :

$$P(R(a) < 0) \approx 0.5$$

et la probabilité de ruine s'écrit : $P(R(a) < -FP)$ avec FP fonds propres ou capitaux propres.

$$\Phi = \frac{-FP + \mathbb{E}(R(a))}{\sigma(R(a))} \tag{I.18}$$

Φ est appelé coefficient de sécurité.

La ruine sera quasi impossible pour un coefficient de sécurité $\Phi > 3,1$.

Le graphique suivant montre l'évolution année par année du ratio S/P du régime « **beau temps** » entre 2016 et 2021 (P la prime pure).

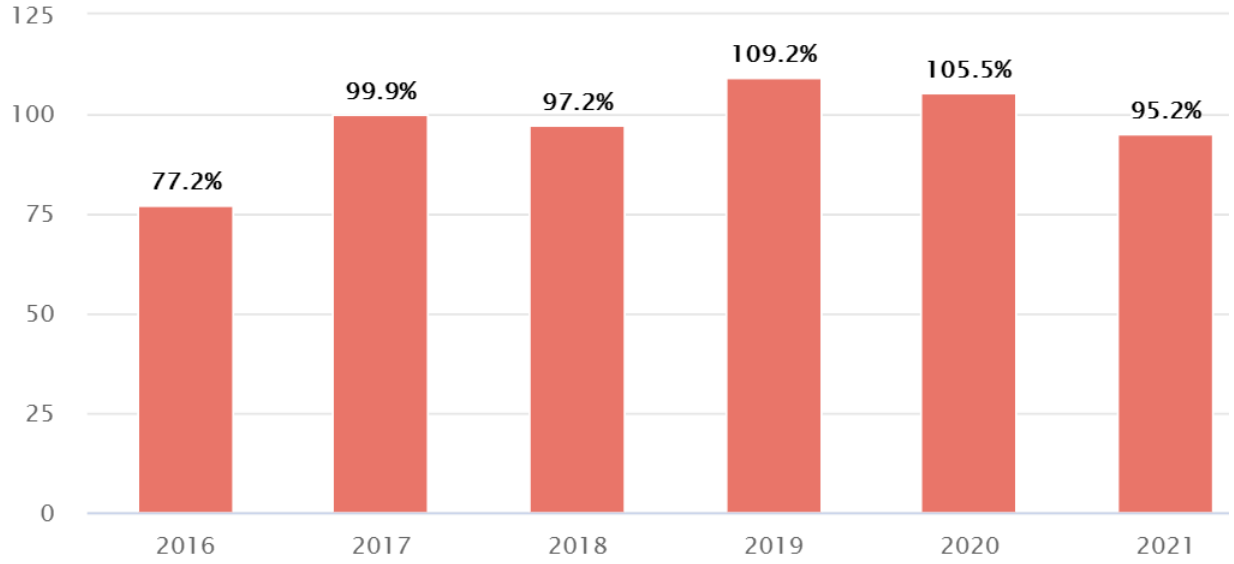


Figure I.24 – Évolution du ratio S/P du régime (2016 - 2021).

Le ratio S/P du régime « **beau temps** » est acceptable pour les 3 première années (2016, 2017, 2018) car inférieur à 100%. Par contre, on note une dégradation du ratio S/P pour l'année 2019 et 2020 avec des ratios S/P respectifs de 109.2% et de 105.5%. Cette situation nous conduire à appliquer une revue tarifaire pour prendre en considération la sinistralité des années plus récentes.

Pour se faire, on désigne par $Taux\_1$ le pourcentage de redressement tarifaire et par $Taux\_2$ le pourcentage de redressement tarifaire corrigé, $Evol\_Tarif$ le taux d'évolution du tarif annuelle qu'on limite arbitrairement à 5% et par P' la prime redressée. Pour une année donnée a, on a :

$$\text{Taux\_1(a)} = max(0, S(a)/P(a)) \quad \text{Evol\_Tarif}(a) = min(\text{Taux\_1}(a), 5\%)$$

$$\text{Taux\_2}(a) = (1 + Taux\_2(a-1))(1 + \text{Evol\_Tarif}(a)) - 1 \qquad P'(a) = P(a) * (1 + \text{Taux\_2}(a))$$

Le tableau I.8 suivant montre la revue tarifaire du régime beau temps.

année	Sinistre (S)	Prime (P)	S/P	Taux_1	Evol_tarif	Taux_2	Sinistre (S)	Prime (P')	S/P'
2016	83040921	107555669	77.2	0%	0%	0%	83 040 921	107 555 669	77.2
2017	108415082	108480365	99.9	0%	0%	0%	108 415 082	108 480 365	99.9
2018	120578573	124015726	97.2	0%	0%	0%	120 578 573	124 015 726	97.2
2019	180966456	165779975	109.2	0%	0%	0%	180 966 456	165 779 975	109.2
2020	130158651	123390702	105.5	9%	5%	5%	130 158 651	129 560 238	100.5
2021	173436368	182207443	95.2	5%	5%	10%	173 436 368	200 883 706	86.3

Table I.8 – Revue tarifaire du régime « **Beau temps** »

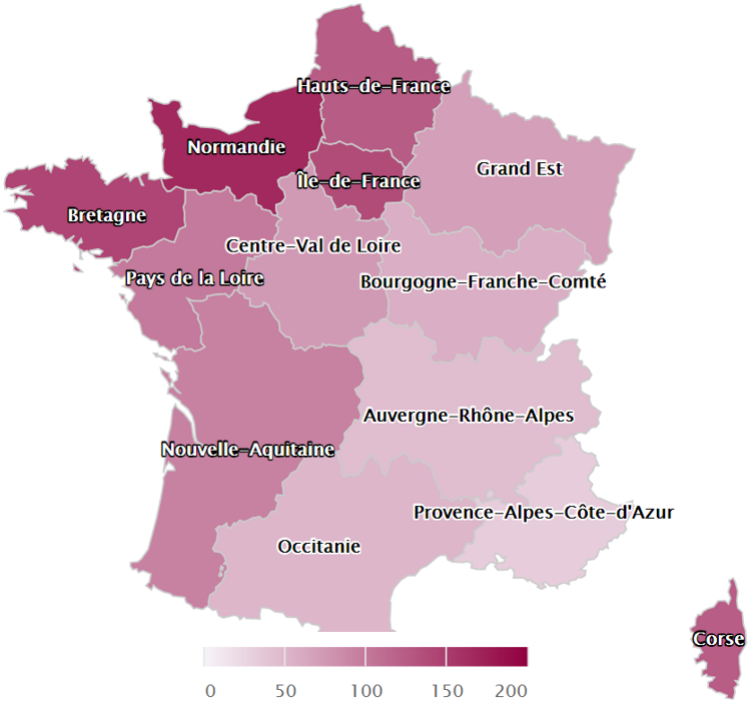


Figure I.25 – Ratio S/P corrigé par région pour l'année 2016

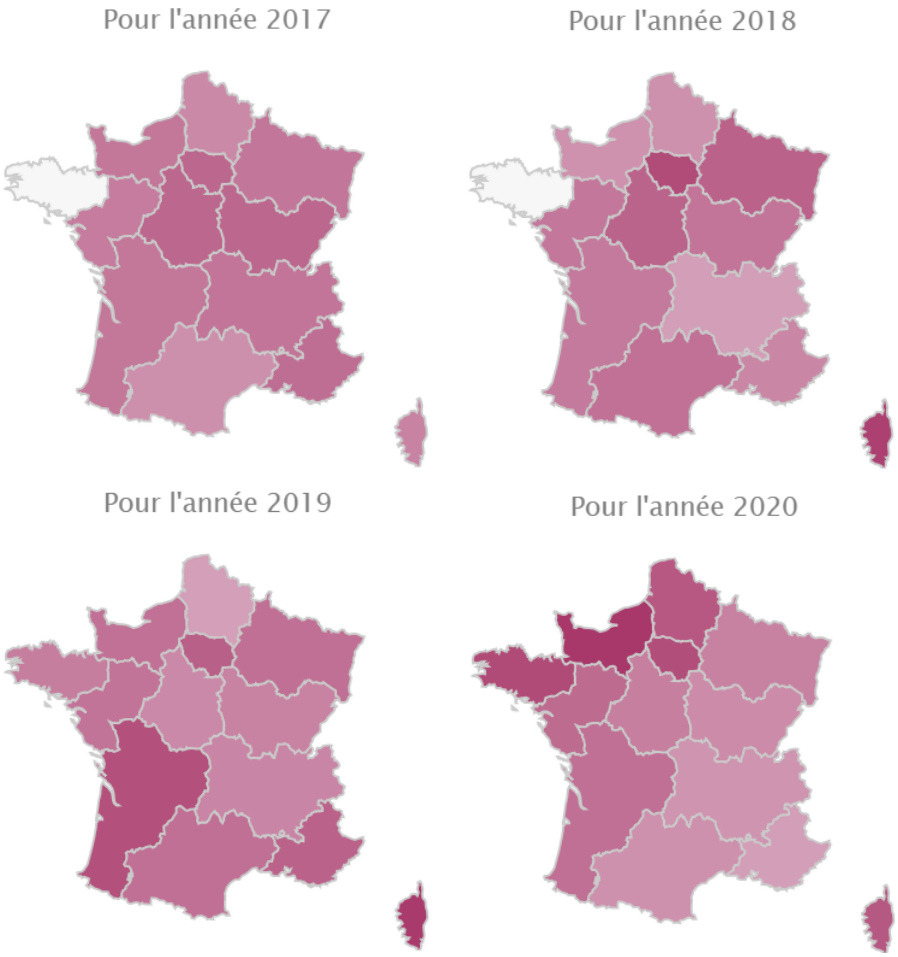


Figure I.26 – Ratio S/P corrigé par région de 2017 à 2021

Chapitre

II

Outil de projection de la sinistralité : arbre de régression Pareto généralisée

Dans ce chapitre, nous allons introduire dans un premier temps les concepts généraux relatifs à la Théorie des Valeurs Extrêmes (TVE) et de l'apprentissage statistique qui seront à la base de nos travaux. Cette première partie nous permettra d'analyser le caractère extrême de la sinistralité du régime "beau temps" au travers d'outils statistiques et de la théorie des valeurs extrêmes (TVE). Ensuite, nous allons présenter le modèle de l'*Arbre de régression pareto généralisée* développé par Farkas et al. (2021a), modèle qui est basé sur la théorie des valeurs extrêmes et l'apprentissage statistique pour une meilleure prise en compte des valeurs extrêmes dans un portefeuille d'assurance. Le modèle combine deux étapes : premièrement, une phase de « croissance » qui correspond à l'algorithme CART basé sur la *log-vraisemblance* Pareto Généralisé, et deuxièmement, une étape de « pruning » qui consiste en l'extraction d'un sous-arbre de la décomposition obtenue dans la phase initiale. Le modèle de l'arbre de régression pareto généralisée est illustrée dans un première temps sur des données simulées. Enfin, nous présenterons une application de ce modèle dans le cadre de la modélisation de la sinistralité extrêmes de notre régime « beau temps ».

1 Théorie des valeurs extrêmes (TVE) et analyse de la sinistralité du régime « beau temps »

Dans cette section, nous souhaitons présenter l'utilisation de la Théorie des valeurs extrêmes et faire une analyse des extrêmes de la sinistralité du régime.

1.1 Approche générale de la TVE, notations et loi du maximum

Les catastrophes naturelles sont des évènements extrêmes qui entraînent des pertes financière significatives pour les organismes de (ré)assurance. En, décembre 1999, les tempêtes Lothar et Martin frappent l'Europe. Ces tempêtes qui se classe parmi les cinq tempêtes les plus importantes des cinquante dernières années a causé une perte d'environ 6,9

milliards d'euros. Ces évènements extrêmes comme tant d'autre ont fait comprendre aux analystes qu'il est essentiel de tenir compte des évènements extrêmes dans la modélisation. En effet, bien que ces événements aient une occurrence faible, leur ampleur est telle que les dommages sont catastrophiques pour les organismes d'assurance.

La théorie des valeurs extrêmes (TVE), contrairement à la plupart des modèles qui ne se concentrent que sur les propriétés moyennes de la distribution, a été développée pour l'estimation de la probabilité d'occurrence d'évènements extrêmes. Elle permet également l'extrapolation du comportement de la queue de distribution des données à partir des plus grandes observations. Dans ce contexte, la TVE apporte une procédure rationnelle et scientifique pour l'estimation des phénomènes dans les queues des distributions. Ainsi, l'on s'intéressera à la modélisation de la loi du maximum.

La théorie des valeurs extrêmes fournit une base mathématique et probabiliste rigoureuse sur laquelle est construit des modèles statistiques pour prédire l'intensité et la fréquence des phénomènes rares (krach boursier, épidémie, attentat) mais aussi des évènements ayant une valeur importante (Christian (2022)).

Dans cette théorie, on utilise les notations suivantes : soit n observations $(x_1, \ldots, x_n)$ qui sont des réalisations d'un n-échantillon $(X_1 \ldots, X_n)$ de variables aléatoires (v.a) indépendantes et identiquement distribuées (iid), de fonction de répartition (f.d.r) F continue. On a : $F(x) = \mathbb{P}(X \leq x)$. Dans la suite X représente la charge d'un sinistre à coût large. On définit le point terminal de F noté x^F de la façon suivante : $x^F = sup\{x : F(x) < 1\}$. On note également $M_n = max(X_1, \ldots, X_n)$. On démontre facilement que la fonction de répartition de la loi de M_n est la puissance n-ième F^n de F.

$$\begin{aligned} \mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \times \mathbb{P}(X_2 \leq x) \times \cdots \times \mathbb{P}(X_n \leq x) \\ \mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x)^n = F^n(x) \end{aligned}$$

car les variables sont indépendantes et elles ont la même loi.

Si F n'est pas connue, cette formule est peu utile. De plus, nous nous intéressons souvent à M_n, lorsque la taille de l'échantillon est importante, et nous souhaitons avoir des approximations asymptotiques.

Le point de départ de la TVE est l'étude du maximum d'un échantillon de variable aléatoire, qui contient plusieurs informations sur les autres types d'extrêmes. L'idée est donc d'étudier le comportement de ce maximum et l'approximiser.

On peut énoncer les deux propriétés suivantes sur le point extrêmal de F.

Proposition 1. *Si $F(x) < 1$, alors $P(M_n \leq x) \to 0$ quand $n \to \infty$.*

Proposition 2. *Si x^F est le point extrêmal de F, alors $M_n \to x^F$ en probabilité quand $n \to \infty$ (et même presque sûrement comme suite croissante).*

La distribution asymptotique de M_n est donc dégénérée. Pour éviter d'avoir une distribution limite dégénérée lorsqu'on regarde le comportement asymptotique du maximum, le théorème fondamentale de la TVE fournit une caractéristique du comportement asymptotique du maximum. Ce résultat est similaire au théorème central limite. D'après le théorème central limite, la moyenne de n variables aléatoires iid converge en loi vers une loi normale.

Théorème 1 (Théorème Central Limite)**.** *Soit n variables aléatoires* $(X_1, X_2, \cdots, X_n)$ iid *telles que* $\mathbb{E}[X_1^2] < \infty$ *et* $Var(X_1) = \sigma^2$*. Alors,*

$$\frac{\sqrt{n}(X_n - \mu)}{\sqrt{\sigma^2/n}} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1)$$

Où $\mu = E(X_1)$

Le théorème fondamental de la TVE est le théorème de Fisher-Typpett. Il s'agit d'un résultat général de la t des valeurs extrêmes relatif à la distribution asymptotique des statistiques d'ordre extrêmes. Ce théorème fait appel aux notions de fonctions à variation régulière et normalisée.

Définition 1 (Variation régulière)**.** Une fonction mesurable $G : \mathbb{R} \to [0, \infty[$ est dite à variation régulière (à l'infinie) d'indice $\rho \in \mathbb{R}$ si pour tout $t > 0$,

$$\lim_{x\to\infty} \frac{G(tx)}{G(x)} = t^\rho$$

Dans le cas particulier où $\rho = 0$, on dit que G est une fonction à variation lente.

En remarquant que si G est à variation régulière d'indice ρ alors $\frac{G(x)}{x^\rho}$ est à variation lente. On démontre qu'une fonction à variation régulière d'indice ρ peut toujours s'écrire sous la forme $x^\rho l(x)$ où l est une fonction à variation lente.

Théorème 2. *l est une fonction à variation lente si et seulement si pour tout $x > 0$*

$$l(x) = c(x)e^{\int_1^x t^{-1}\epsilon(t)}$$

où c et ϵ sont des fonctions positives telles que :

$$\lim_{x\to\infty} c(x) = c \in]0, +\infty[\quad et \quad \lim_{t\to\infty} \epsilon(t) = 0$$

Pour trouver une distribution limite non dégénérée, il nous est nécessaire de transformer M_n. Pour se faire, il est assez naturel de se tourner, comme pour le TCL, vers une normalisation de la variable aléatoire.

La variable M_n est ajustée à l'aide des suites c_n (supposées positives) et d_n. Nous supposons l'existence d'une telle séquence de coefficients (c_n, $n > 0$). Gnedenko (1943) fournit la version définitive du théorème des valeurs extrêmes qui spécifie la forme de la loi limite F_Y quand la longueur de la période sur laquelle on observe les extrêmes croît indéfiniment.

Théorème 3 (**Fisher-Typpett-Gnedenko**). *Soient* $(X_1, X_2, \cdots, X_n)$ *une suite de variable aléatoires indépendantes avec pour fonction de répartition F, et soit* $M_n = max(X_1, \cdots, X_n)$. *S'il existe des réels* $c_n > 0$ *et* d_n *telles que :*

$$\lim_{n\to\infty} P\left(\frac{M_n - d_n}{c_n} \leq x\right) = \lim_{n\to\infty} (F(c_n x + d_n))^n = G(x) \tag{II.1}$$

Où G est une fonction de répartition non dégénérée. Alors G appartient à l'un des trois types suivants :

$$\begin{cases} Fréchet(\alpha > 0): & \Phi_\alpha(x) = exp(-x^{-\alpha})\mathbb{1}_{x>0} \\ Weibull(\alpha > 0): & \Psi_\alpha(x) = \mathbb{1}_{x>0} + exp(-(-x)^{-\alpha})\mathbb{1}_{x<0} \\ Gumbel: & \Lambda = exp(-e^{-x}), \quad \forall x \in \mathbb{R} \end{cases}$$

Ainsi le maximum d'un échantillon de variables aléatoires iid après renormalisation ne peut converger en loi que vers 3 types de loi : la loi de Gumbel, la loi de Fréchet ou la loi de Weibull.

Une fois le théorème sur la loi des valeurs extrêmes énoncé, il convient de définir la notion de domaine d'attraction.

Définition 2 (Domaine d'attraction)**.** On dit que F appartient au domaine d'attraction de G ($F \in D(G)$) s'il existe deux suites (c_n) et (d_n) telles que la convergence (II.1) ait lieu.

Jenkinson (1955) et Von Mises (1954) montrent que ces trois lois peuvent être regroupées sous une forme générale qu'ils appellent la distribution $GEV(\mu, \sigma, \xi)$ pour *Generalized Extreme Value Distribution* :

$$\text{GEV}: \qquad G(x, \mu, \sigma, \xi) = \begin{cases} \exp\left(-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right), & \text{si } \xi \neq 0, \\ \exp\left(-\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right), & \text{si } \xi = 0, \end{cases} \tag{II.2}$$

Le paramètre ξ qui apparaît dans la formule (II.2) est appelé l'indice de queue ou paramètre de forme ; μ est le paramètre de position et σ est le paramètre d'échelle.

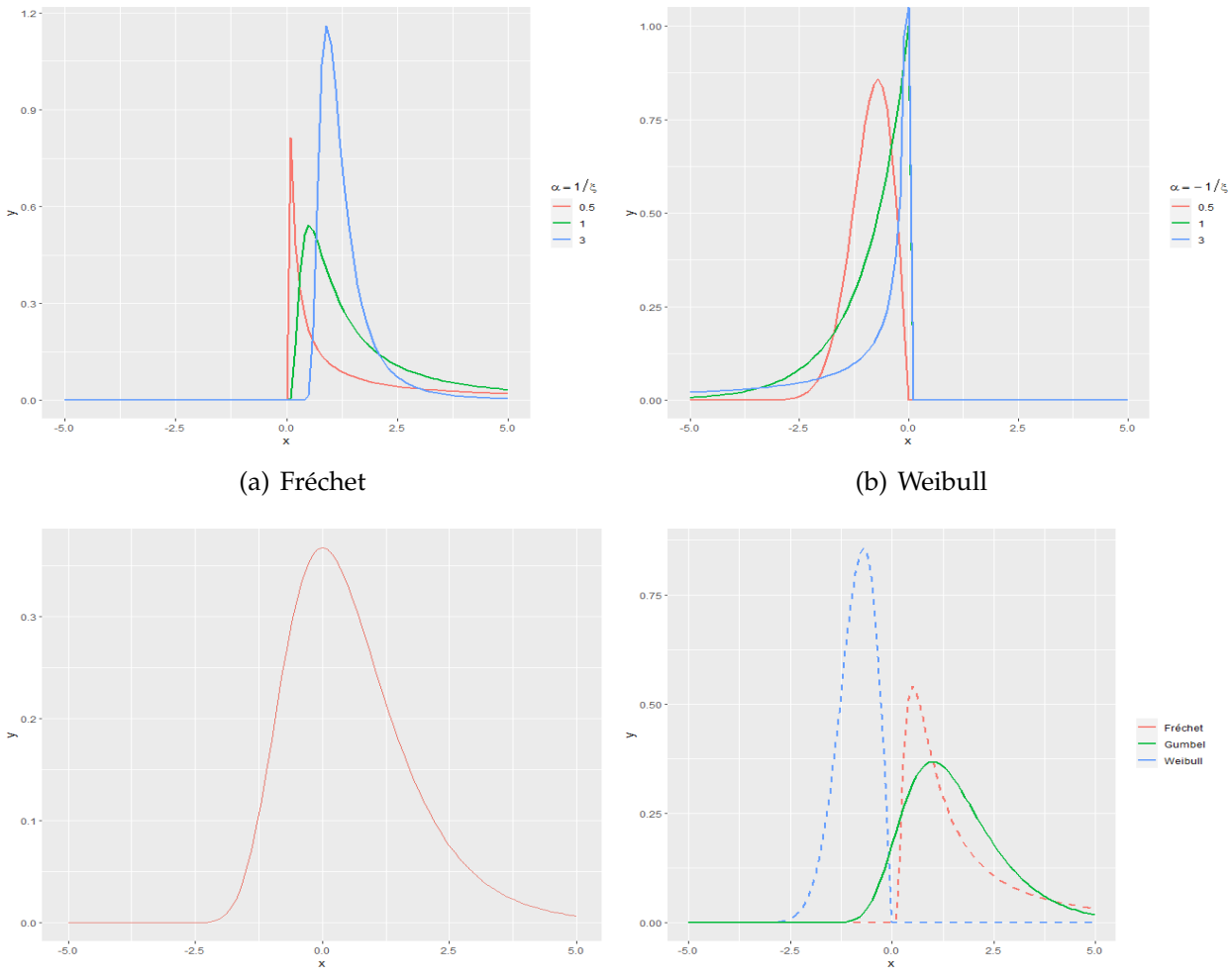
- $\xi = 0$: domaine d'attraction de Gumbel. Le point extrêmal est fini ou infini. La fonction de survie converge vers 0 à une vitesse exponentielle. Ce sont les distributions à queue légère. $GEV(0, 1, 0)$ = Gumbel ;
- $\xi > 0$: domaine d'attraction de Fréchet. Le point terminal est infini. La fonction de survie converge vers 0 à une vitesse polynomiale. Ce sont les distributions dites à queue épaisse. $GEV(1, \alpha^{-1}, \alpha^{-1})$ = Fréchet(α) ;
- $\xi < 0$: domaine d'attraction de Weibull. Le point terminal est fini. La fonction de survie converge vers 0 à une vitesses polynomiale. Ce sont les distributions à queue fine. $GEV(-1, \alpha^{-1}, -\alpha^{-1})$ = Weibull(α)

Le tableau suivant regroupe quelques lois usuelles en fonction de leur domaine d'attraction.

Fréchet	Gumbel	Weibull
Pareto	Normale	
Log-gamma	Exponentielle	Uniforme
Student	Gamma	Beta
Burr	Log-normale	

Table II.1 – Domaine d'attraction des lois usuelles

Ainsi, il ressort de la caractérisation des distributions des trois lois vu précédemment que la loi de Weibull permet de modéliser les distributions dont le support de F est borné, s'il ne l'est pas, on utilisera la loi de Gumbel ou de Fréchet. Les coûts des sinistres extrêmes étant considérés comme à support non borné, la charge n'admettant pas un point terminal fini, nous pouvons déjà exclure le domaine de Weibull. Le graphique suivant illustre les densités des lois de Gumbel et de Fréchet pour différentes valeurs du paramètre de queue et comparaison des densités des lois de reverse Weibull, Fréchet et Gumbel.



(a) Fréchet (b) Weibull

(c) Gumbel (d) Comparaison

Figure II.1 – Exemples de densités GEV et comparaison des densités des lois

En réalité, il y a un seuil pour lequel on décide que toute valeur qui dépasse ce seuil

apparaît comme une valeur extrême. Ainsi, on peut définir un extrême comme une valeur dépassant un certain seuil. Premièrement, une attention est portée sur l'intensité de dépassement de ce seuil et deuxièmement on s'intéresse à la dynamique de dépassement du seuil. On pourra se demander avec quelle régularité temporelle et avec quelle fréquence ces évènements de dépassement du seuil surviennent. Dans la théorie des valeurs extrêmes, deux approches principales sont utilisées : la première, dénommée *blocs maxima method* (BM) et la seconde approche est désignée sous le terme *Peaks-Over-Threshold method* (POT). Selon le contexte, l'une ou l'autre des approches peut se révéler mieux adaptée, mais il est le plus souvent utile de les mettre toutes deux en œuvre et d'en comparer les résultats.

La méthode du *blocs maxima* (BM) est celle présentée jusqu'à présent. La modélisation des queues de distribution par la méthode des *blocs maxima* (Coles (2001)) s'appuie sur le théorème de Fisher-Tippet (page 44). Elle consiste à s'intéresser au maximum dans un échantillon et à effectuer dans un premier temps le choix de la taille des blocs. Dans cette approche, le maximum est choisi périodiquement (par exemple annuellement). Elle consiste à un découpage des données en blocs, dont les *maxima* sont supposés distribués selon une loi d'une famille connue *GEV* (forumule II.2,page 44). Cette approche classique de la théorie des valeurs extrêmes qui s'appuie sur l'étude du maximum n'est pas assez générale et est peu exploitable en assurance du fait de la rareté des sinistres extrêmes. C'est pour ces problématiques qu'une seconde approche de la théorie des valeurs extrêmes a été développée. C'est l'approche par dépassement de seuil (*Peak Over Threshold*). Elle n'est pas sans lien avec l'étude de la distribution du maximum et fait l'objet de la sous-section suivante.

1.2 La loi des excès et la loi de Pareto Généralisée

L'approche *Peak Over Threshold* (POT) pour méthode des excès-au-delà en français, est une méthode qui consiste à analyser la distribution des dépassements au-dessus des niveaux élevés afin d'estimer les quantiles extrêmes. Cette méthode est largement utilisée depuis 1990 (voir Coles (2001),Davison and Smith (1990)) et vise à répondre à la question suivante : « Etant donné un sinistre extrême, à quel point ce sinistre est-il extrême ? ». Elle est basée sur les travaux de Pickands (1975), Davison and Smith (1990) qui ont observé que la queue extrême d'une distribution a souvent une forme plutôt simple et standardisée, quelle que soit la forme des parties les plus centrales de la distribution.

Le niveau du seuil qu'on note u doit être choisi suffisamment haut pour que la queue ait approximativement la forme normalisée, mais pas si haut qu'il en reste trop peu d'observations au-dessus. Ainsi, le choix du seuil u implique un équilibre entre biais et variance. Un seuil trop bas est susceptible de ne pas respecter la base asymptotique du modèle, conduisant à un biais ; un seuil trop élevé générera peu d'excès avec lesquels le modèle pourra être estimé, conduisant à une forte variance. La pratique courante consiste à choisir un seuil aussi bas que possible, sous réserve que le modèle limite fournisse une approximation raisonnable.

Dans l'approche POT, au lieu de considérer le maximum M_n de l'échantillon $X_1, X_2, \ldots, X_n$, on s'intéresse à N_n qui représente le nombre de dépassement du seuil u_n, avec $N_n = \sum_{i=1}^{n} \mathbb{1}_{X_i > u_n}$ c'est à dire aux observations $(X_i - u_n)_+$ qui sont strictement positives. Ces observations au-dessus du seuil u_n, sont réparties selon une *distribution généralisée de Pareto (GPD)*. Cette distribution de paramètre (σ, ξ) se définit comme suit :

$$\text{GPD :} \qquad G^p(x,\sigma,\xi) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x}{\sigma}\right)\right]_+^{-1/\xi}, & \text{si } \xi \neq 0, \\ 1 - e^{-x/\sigma}, & \text{si } \xi = 0, \end{cases} \tag{II.3}$$

Où :

$$\begin{array}{ll} x \geq 0 & \text{si } \xi \geq 0 \\ 0 \leq x \leq -\frac{\sigma}{\xi} & \text{si } \xi < 0 \end{array}$$

L'espérance d'une Pareto généralisée est donnée par :

$$\forall \xi < 1, \qquad \mathbb{E}(X) = \frac{\sigma}{1-\xi}$$

La variance d'une Pareto généralisée est donnée par :

$$\forall \xi < \frac{1}{2}, \qquad \mathbb{V}(X) = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}$$

La Figure II.2 illustre les observations étudiées dans le cadre de dépassements d'un seuil u_n.

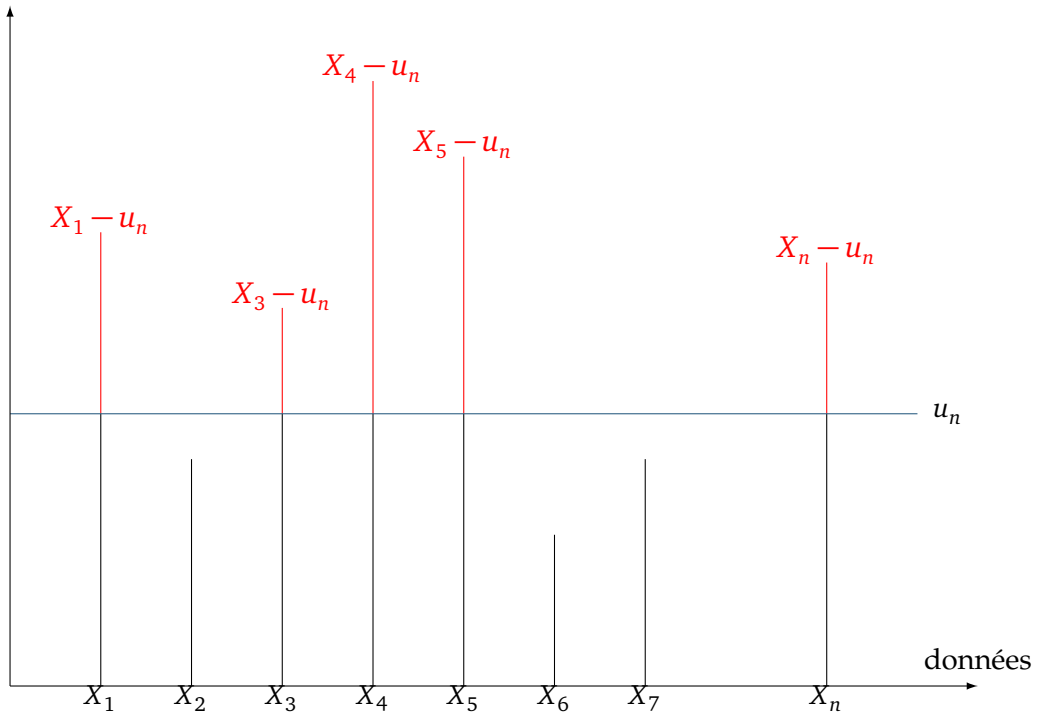


Figure II.2 – Dépassement de seuil - POT

La loi de Pareto Généralisée peut donc être vue comme la distribution asymptotique des excédents au-delà d'un certain seuil qui tend vers l'infini. Le théorème de Pickands, Balkema, De Hann est utile lorsqu'on travaille avec des observations dépassant un seuil fixé puisqu'il assure que la loi des excès puisse être approchée par une GPD.

Théorème 4 (Pickands, Balkema, De Hann). *Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes, alors il existe* $\sigma(u)$ *et* $\xi \in \mathbb{R}$ *tels que :*

$$\lim_{u \to x_F} \sup_{0 \geqslant y \geqslant x_F - u} \left| F_u(y) - G_{\xi,\sigma(u)}(y) \right| = 0$$

où $G_{\xi,\sigma(u)}(y)$ *est la fonction de répartition de la loi de Pareto généralisée et* F_u *f.d.r. des excès au-delà du seuil u.*

De plus, quand u est suffisamment grand, $F_u \approx G_{\xi,\sigma(u)}$.
La loi des excès ainsi que la GPD reposent sur l'utilisation d'un seuil noté u. Nous allons présenter dans les parties qui suivent des outils permettant de trouver la valeur de u en particulier pour la sinistralité du régime "beau temps".

Le graphique II.3 ci-dessous présente les distributions de trois lois de la GPD : *GDP(1,1/3)*, *GDP(2,1/2)* et *GDP(10,1)*

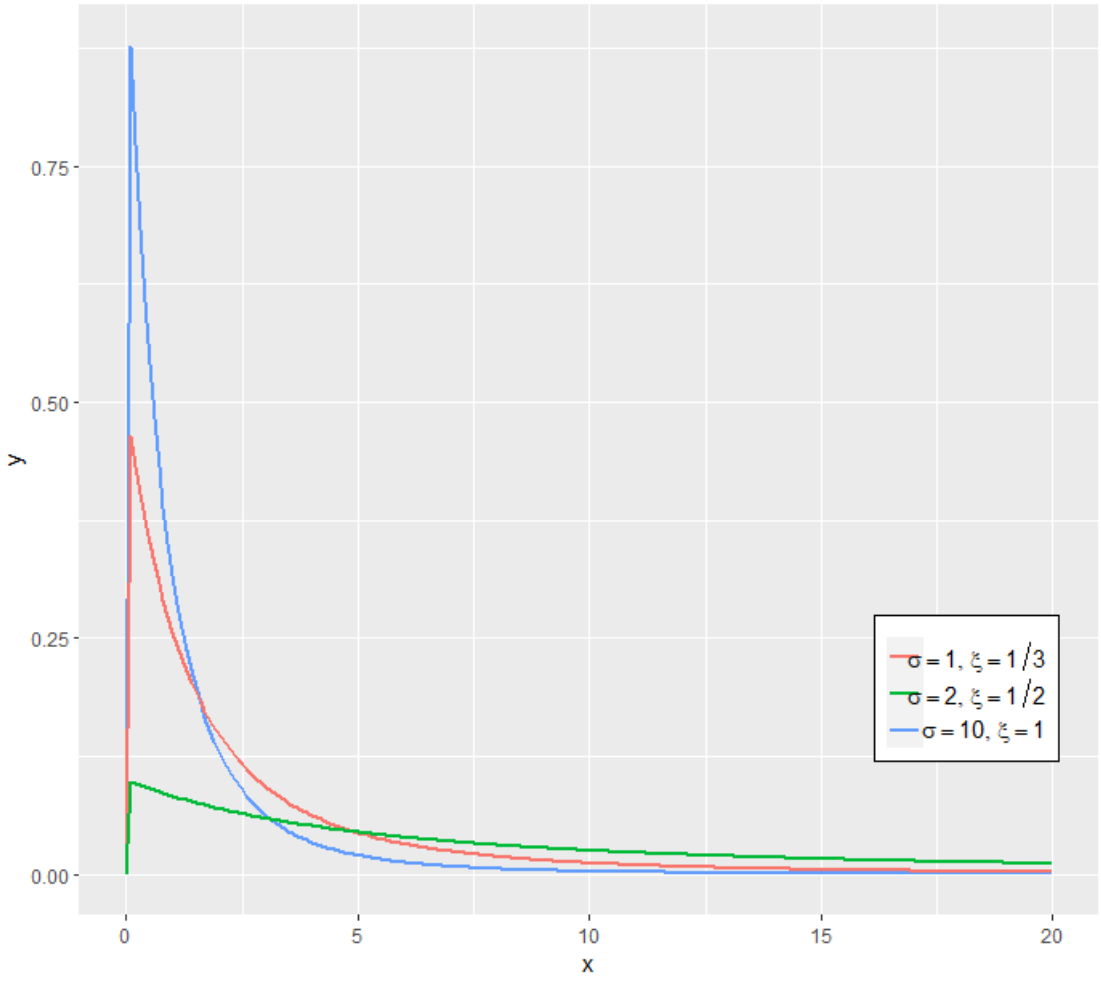


Figure II.3 – Exemple de densités GPD

1.3 Analyse de la sinistralité du régime : sinistres extrêmes et non-extrêmes

Il existe un grand nombre d'outils graphiques et statistiques qui permettent d'étudier une distribution. Au nombre desquels on retrouve la boîte à moustache et le QQPlot. Dans la suite nous avons souhaité tester la stabilité dans le temps.

Boîte à moustache

Une boîte à moustache (*boxplot* en anglais) est un graphique simple composé d'un rectangle duquel deux droites sortent afin de représenter certains éléments des données. Il permet notamment de détecter les valeurs aberrantes.

Le graphique suivant présente la boîte à moustache de la sinistralité du régime par année de 2016 à 2021. Il apparaît un seuil stable au cours des trois première années (2016, 2017 et 2018). Pour cette période on a un seuil qui semble stable aux alentours des 450000€. Par contre à partir de 2019 on constate une instabilité du seuil extrême et qui augmente au cours du temps. Cette situation met en évidence le caractère de plus en plus extrêmes des conditions climatiques et qui se répercute sur la sinistralité.

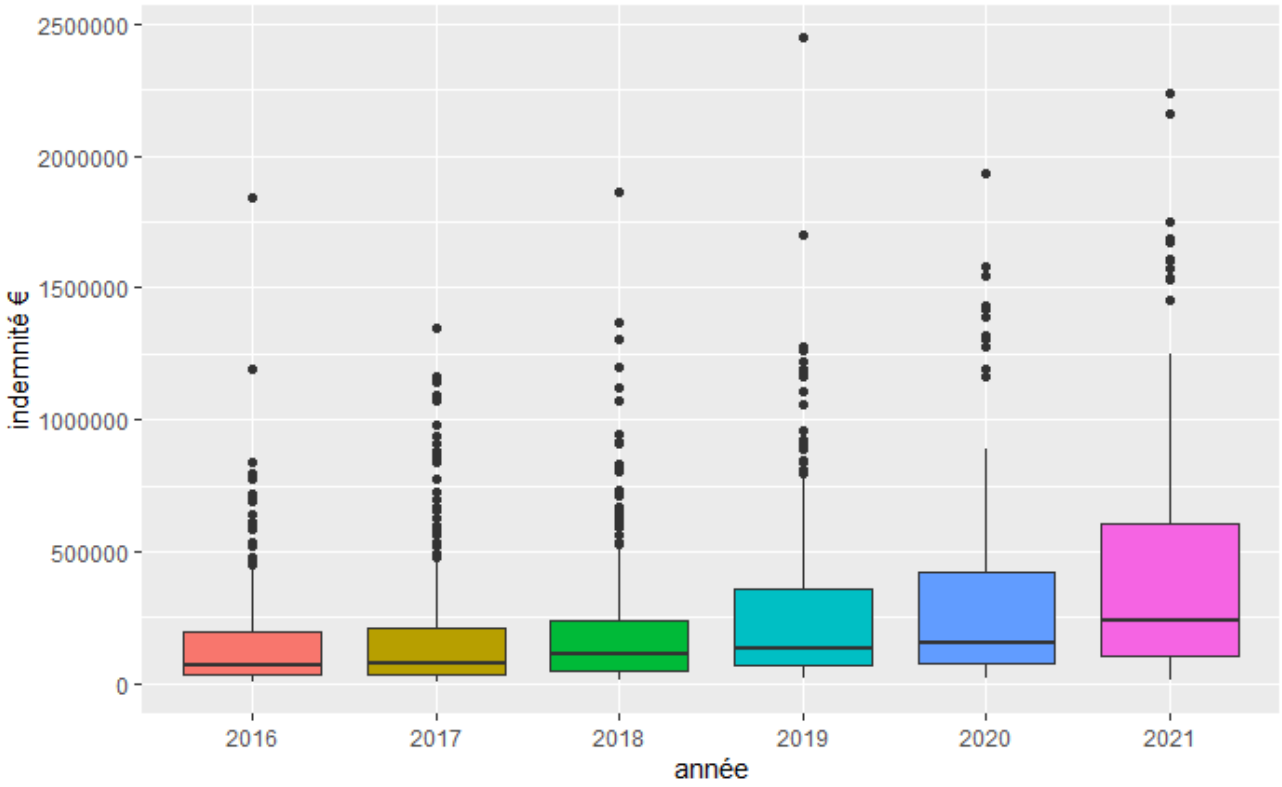


Figure II.4 – Boîte à moustache de la sinistralité du régime par année

Avec le seuil de 450000€ , on obtient 2% des sinistres au-delà de ce seuil sur la période de 2016 à 2021.

Bien que la boîte à moustache soit très utilisé pour faire de l'analyse descriptive, il se trouve très limités en dehors d'une distribution Normale. Nous ne pouvons donc pas nous fier à cet outil dans l'optique de séparer les sinistres extrêmes des non-extrêmes de la distribution de la sinistralité.

Graphique quantile-quantile (QQ-plot)

Le *graphique quantile-quantile* (QQ-plot) est un outil qui permet d'évaluer la pertinence de l'ajustement d'une distribution donnée à un modèle théorique. Pour ce faire, le graphique quantile-quantile consiste à représenter les quantiles empiriques en fonction des quantiles théoriques afin d'étudier visuellement la linéarité entre ces deux quantités.

On souhaite donc déterminer la forme de la queue de distribution et avoir une idée du domaine d'attraction de la pour la distribution des sinistres du régime. Le QQ-Plot est l'un des indicateurs graphiques (avec la *Mean Excess Function* qu'on verra plus tard) qui donnent une indication de l'appartenance probable à l'un des domaines d'attraction possibles.

Le graphique quantile-quantile est une méthode nous permettant de mesurer graphiquement l'adéquation d'une variable observée à une loi théorique de fonction de répartition F continue. Lorsque X est à fonction de répartition F continue, $F(X) \sim U[0;1]$. En notant $X_{(1)} \leqslant \ldots \leqslant X_{(n)}$ la statistique d'ordre associée à une variable, on a donc

$$\left(F\left(X_{(i)}\right)\right)_{i=1,\ldots,n} = \left(U_{(i)}\right)_{i=1,\ldots,n} \tag{II.4}$$

et donc en utilisant F^{-1}, on a l'équivalence suivante :

$$\left(X_{(i)}\right)_{i=1,\ldots,n} = \left(F^{-1}\left(U_{(i)}\right)\right)_{i=1,\ldots,n} \tag{II.5}$$

Le graphique quantile-quantile est donc donné par le couple

$$\left\{X_{(i)}, F^{-1}\left(1-\frac{i}{n}\right), i=1,\ldots,n\right\} \tag{II.6}$$

L'adéquation des observations à la loi de fonction de répartition F se traduit par la linéarité du nuage de points obtenu. C'est-à-dire que le graphique quantile-quantile peut être adapté à de nombreuses distributions de probabilité afin d'effectuer un premier test graphique d'adéquation de loi

— Pour la loi exponentielle, le graphique consiste à tracer les points :

$$\left\{\left(x_{(i)}, -\ln\left(1-\frac{i}{n+1}\right)\right), \quad \text{pour tout } i \in [\![1;n]\!]\right\} \tag{II.7}$$

— Pour la loi de Pareto, le graphique consiste à tracer les points :

$$\left\{\left(\ln\left(x_{(i)}\right), -\ln\left(1-\frac{i}{n+1}\right)\right), \quad \text{pour tout } i \in [\![1;n]\!]\right\} \tag{II.8}$$

— Pour la loi Log-Normale, en utilisant le lien avec la loi Normale centrée réduite et sa fonction de répartition φ le graphique consiste à tracer les points :

$$\left\{\left(\ln\left(x_{(i)}\right), \varphi^{-1}\left(\frac{i}{n+1}\right)\right), \text{ pour tout } i \in [\![1;n]\!]\right\} \tag{II.9}$$

— Pour la loi de Weibull, le graphique consiste à tracer les points :

$$\left\{\left(\ln\left(x_{(i)}\right), \ln\left(-\ln\left(1-\frac{i}{n+1}\right)\right)\right), \quad \text{pour tout } i \in [1;n]\right\} \tag{II.10}$$

Si l'échantillon est identiquement distribué selon la même loi que X, alors les points doivent être alignés sur une droite. Dans ce cas les quantiles empiriques sont égaux aux quantiles théorique de la loi analysée. C'est une condition nécessaire pour estimer que la distribution des données disponibles peut être ajustée par la loi théorique en question. En revanche, si le graphique du QQ-plot est concave, cela indique que l'échantillon est issu d'une distribution à queue plus épaisse, et inversement, un graphe convexe indique que l'échantillon est issu d'une loi à queue plus fine.

Dans le cas de la loi exponentielle, on supposera u suffisamment grand et on distinguera trois cas en fonction de la loi dont sont issues les observations et du domaine d'attraction.

— **Gumbel** : Les points du graphique sont approximativement alignés.

— **Fréchet** : On observe une forme concave.

— **Weibull** : On observe une forme convexe.

En représentant les quantiles de la loi exponentielle contre les quantiles empiriques des excès ordonnés de la sinistralité du régime par année de 2016 à 2021, on obtient des graphiques comme suit :

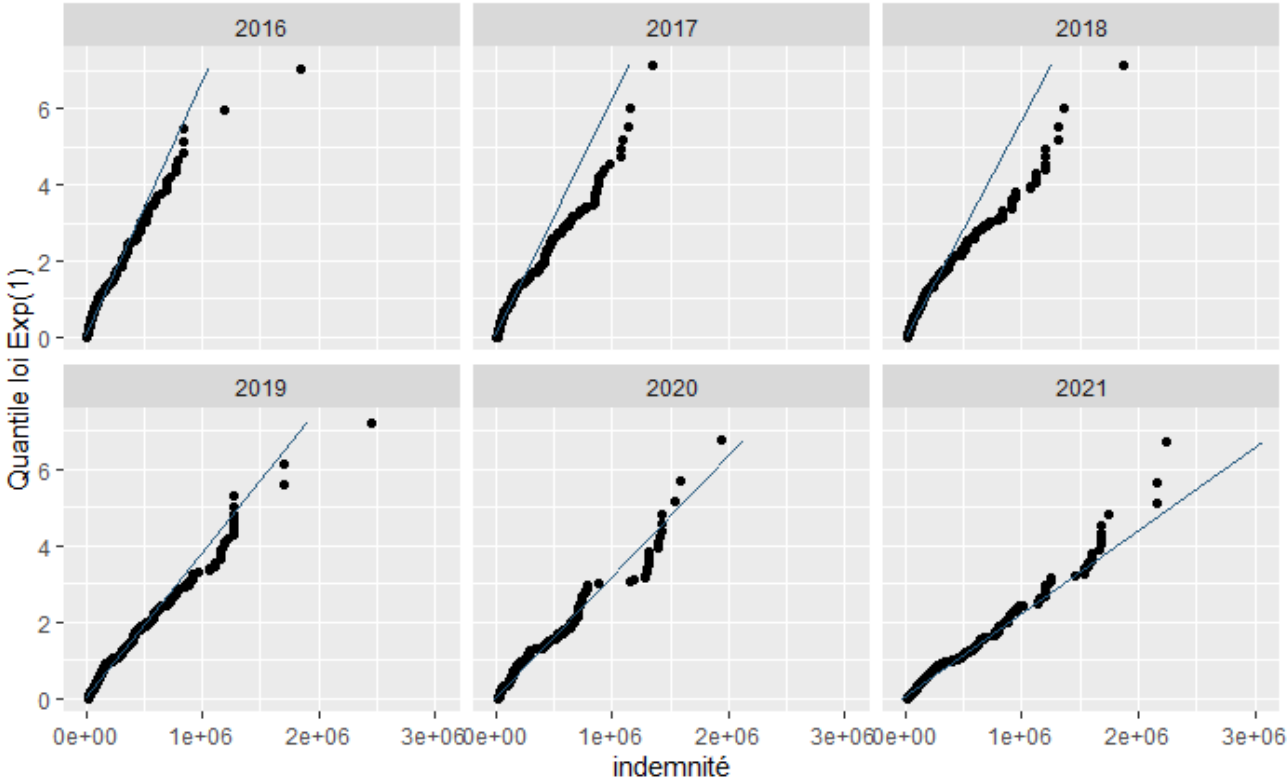


Figure II.5 – QQ-Plot des observations de la sinistralité par année

Pour tous les années les points ont tendance à diverger vers le haut. Ainsi, le QQ-plot suggère que nos extrêmes appartiennent à une loi de Fréchet ce qui nous indique une distri-

bution de queue épaisse et une convergence lente. Ainsi, il est possible d'utiliser l'estimateur de Hill afin de trouver l'estimateur de ξ.

1.4 Détermination du seuil des extrêmes u

Dans la partie précédente, nous avons justifié que la sinistralité du régime fictif a bien une distribution à queue lourde. Ainsi, nous pouvons à présent déterminer le seuil optimal qui garantit l'existence de la distribution de Pareto généralisée (GPD).

En pratique, de nombreuses méthodes efficaces existent pour sélectionner ce seuil optimal. Ces méthodes peuvent être classées en méthodes graphiques et numériques que nous allons présenter dans cette partie. Les méthodes graphiques nécessitent un contrôle de la part de l'utilisateur et une part de subjectivité contrairement aux méthodes numériques qui fournissent une valeur pour l'estimation du seuil.

Mean Excess Plot (ME-plot)

La technique la plus utilisée est la fonction d'excès en moyenne (*mean excess function* ou *mean residual life plot* (ME-plot) en anglais). La *mean excess function* est un outil graphique important pour étudier les extrêmes. La *mean excess function* aussi appelée fonction de dépassement moyen ou durée de vie résiduelle se définit comme suit :

Définition 3 (*Mean excess-function*)**.** La *mean excess-function* est définie de la manière suivante :

$$\{(u, e_n(u), X_{(1)} < u < X_{(n)}\}$$

où $X_{(1)}$ et $X_{(n)}$ sont respectivement les maximum et minimum de l'échantillon. Et $e_n(u)$ est définie par :

$$e_n(u) = \frac{\sum_{i=1}^{n}(X_i - u)_+}{\sum_{i=1}^{n} \mathbb{1}_{X_i > u}} \tag{II.11}$$

Théoriquement, la *sample mean excess function* $e_n(u)$ est l'estimateur empirique de la *mean excess function* :

$$e(u) = \mathbb{E}[X - u | X > u] \tag{II.12}$$

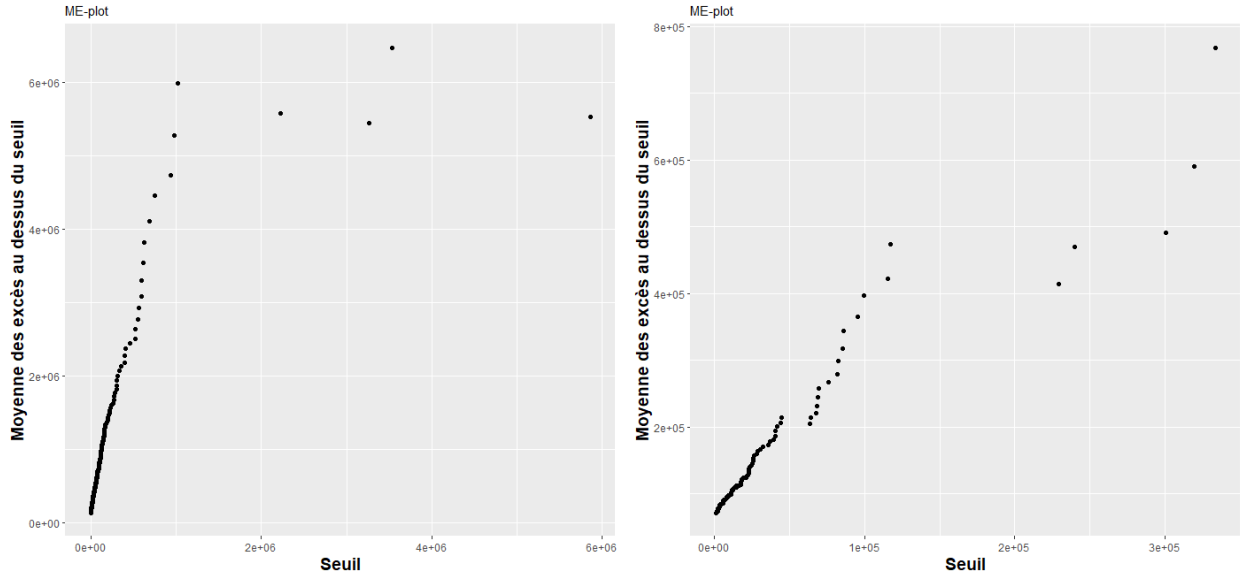
Le *mean excess plot* (ME-plot) est un outil graphique qui permet de déterminer le seuil des valeurs extrêmes u à partir duquel les observations se comportent comme une distribution de Pareto généralisée. La proposition suivante permet de faire le lien entre l'allure du *mean excess plot* et le comportement d'une loi de Pareto généralisée. En effet, le graphe des excès moyens d'une distribution de Pareto généralisée est linéaire en u.

Proposition 3. *Lorsque X suit la GPD, la mean excess function est linéaire en u :*

$$e(u) = \frac{\sigma + \xi u}{1 - \xi} \tag{II.13}$$

Si le *ME-plot* semble avoir un comportement linéaire à partir d'un certain seuil positif u, cela signifie que les données supérieures à ce seuil suivent une loi de Pareto généralisée (GPD). Les propriétés et l'estimation des paramètres de la GDP sont présentés dans l'annexe. Ainsi, si la fonction empirique des excès moyens se comporte linéairement il convient alors de déterminer le seuil à partir duquel le mean excess plot est approximativement linéaire.

Pour illustrer cette propriété de linéarité, deux échantillons sont simulés. Le premier comporte 600 valeurs issues d'une loi de Pareto Généralisée de paramètres $\xi = 0,9$ et $\sigma = 15000$ et le deuxième de 100 valeurs issues de cette même loi. Le *mean excess* plot est tracé pour ces deux échantillons :



(a) 600 observations (b) 100 observations

Figure II.6 – *Mean excess* plot d'échantillons issues d'une GPD(0.9,15000)

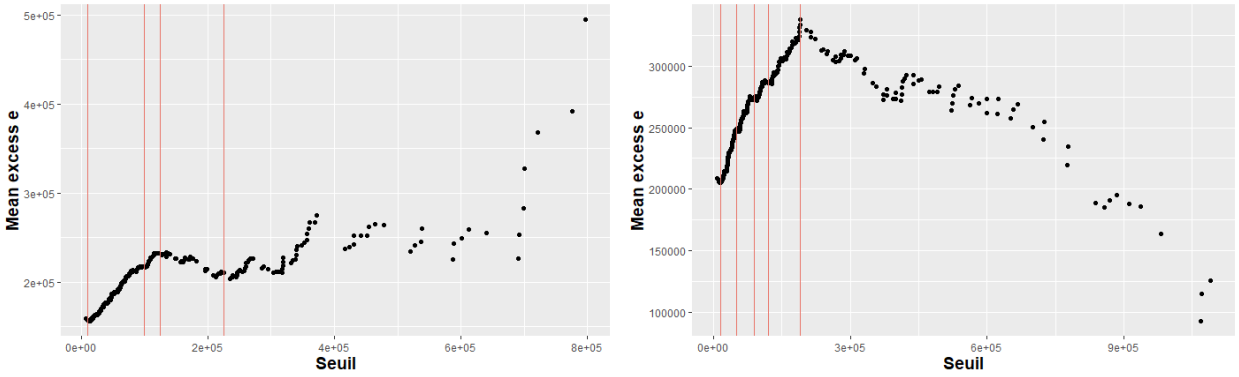
Pour les graphiques d'échantillons des lois de distribution, nous pouvons voir que l'estimation du *mean excess* est trop volatile pour les seuils pris dans les quantiles supérieurs, cet effet est dû à la faible quantité de données que la fonction prend pour faire une estimation, plus le seuil est élevé, moins il y a de points pour estimer l'espérance de vie résiduelle.

Comme la distribution de la sinistralité du régime appartient au domaine d'attraction de Fréchet, une plage de la *Mean excess plot* sera acceptable lorsque la ME-plot est croissante et linéaire en u sur cette plage. A partir du graphique II.7 qui représente la *Mean excess plot* de la sinistralité du régime par année on observe :

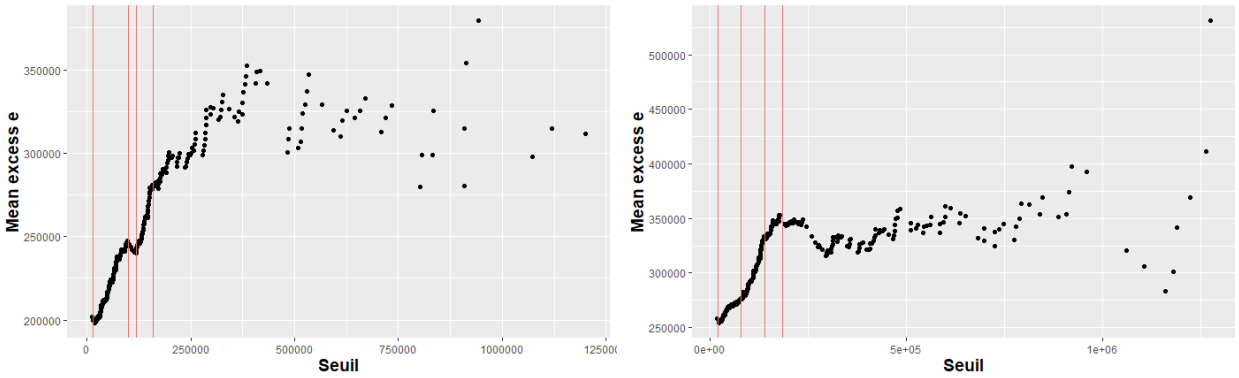
— Pour l'année 2016, on distingue principalement cinq (5) plages à analyser. la plage 1 correspond à $u \leq 10\,000€$. Dans cette plage le biais du modèle est maximal : elle n'est donc pas acceptable. Les plages 2 et 3 correspondent respectivement à $u \in]10\,000€, 100\,000€]$ et $u \in]100\,000€, 125\,000€]$. Sur ces plages, la courbe de la *Mean excess plot* est approximativement une droite avec une pente positive. Elles sont donc acceptables et cohérentes avec le domaine d'attraction auquel la distribution des données appartient. Dans ces conditions nous identifions $100\,000€$ et $125\,000€$ comme

potentiels seuils. Quand aux plages 4 ($u \in]125\,000€, 225\,000€]$) et 5 ($u > 225\,000€$), on remarque que la *Mean excess plot* est non-croissante strictement. Ces plages ne sont donc pas acceptable car on observe un changement de pente ce qui n'est pas cohérente avec le domaine d'attraction auquel les données appartiennent.

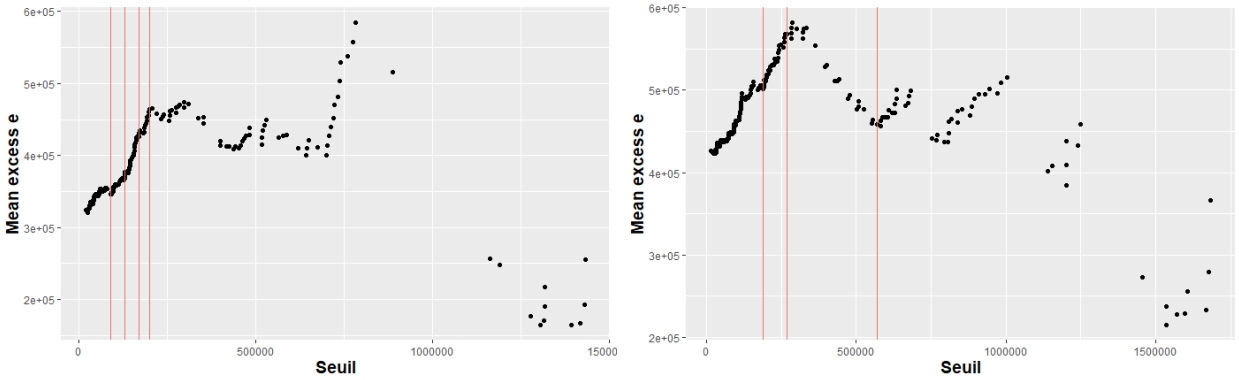
— En suivant la même analyse que pour l'année 2016, on obtient pour les années 2017 à 2021, les seuils acceptables suivants :
  — 2017 : 90 000€, 125 000€ et 185 000€ ;
  — 2018 : 125 000€ et 175 000€ ;
  — 2019 : 130 000€ et 180 000€ ;
  — 2020 : 125 000€, 180 000€ et 200 000€ ;
  — 2021 : 120 000€, 180 000€ ;



(a) 2016 (b) 2017



(c) 2018 (d) 2019



(e) 2020 (f) 2021

Figure II.7 – *Mean excess* plot de la sinistralité du régime par année

En résumé avec la méthode graphique ME-plot, nous avons identifié l'existence de deux seuils stables au cours du temps proche de **125 000**€ et **180 000**€ qui sont cohérents avec le domaine d'attraction auquel la distribution des données de la sinistralité du régime appartient.

Les résultats obtenus via cette méthode peuvent parfois susciter des débats, car l'identification des plages demeure essentiellement graphique et pourrait varier d'un observateur à un autre.

Le graphique de Hill (*Hill plot*)

Le graphique de Hill de détermination du seuil repose sur la propriété de stabilité de la loi de Pareto généralisée. Ce graphique consiste à tracer la valeur de l'estimateur de Hill $\hat{\xi}_n^{(H)}$ en fonction du nombre d'excès k_n considérés. Le seuil à sélectionner correspond, par équivalence, au plus petit nombre d'excès pour lequel l'estimateur se stabilise.

Soit $X_1, \cdots, X_n$ un échantillon de variables aléatoires indépendantes et identiquement distribuées. L'échantillon ordonné par ordre croissant est noté $X_{(1)}, \cdots, X_{(n)}$.

Le graphe du Hill Plot consiste à représenter les points :

$$\left\{ k_n, \hat{\xi}_n^{(H)}(k_n) = \frac{1}{k_n} \sum_{i=1}^{k_n} \ln\left(X_{(n-i+1)}\right) - \ln\left(X_{(n-k_n)}\right) \right\} \tag{II.14}$$

L'inconvénient du Hill Plot est qu'il est valable uniquement pour les distributions de probabilités appartenant au domaine d'attraction de Fréchet, c'est-à-dire celles ayant un indice des valeurs extrêmes strictement positif. L'estimateur de Hill est donc l'un des estimateurs particulièrement adapté aux données de la sinistralité du régime « beau temps « vu le domaine d'attraction auquel elles appartiennent.

L'estimateur de Hill assure un bon équilibre entre le biais et la variance. L'idée ici est d'identifier les intervalles sur lesquels la courbe de l'indice de queue de Hill est approximativement une droite horizontale. Il faut sélectionner un nombre d'excès k_n pas trop petit, au risque de disposer de trop peu d'observations pour estimer convenablement les paramètres de la GPD. Le nombre d'excès k_n ne doit pas non plus être trop grand au risque que l'approximation par une loi GPD ne soit pas vérifiée.

En théorie, le nombre d'excès optimal k_n à choisir devrait vérifier $\lim_{n\to+\infty} k_n = +\infty$ (pas trop petit) et $\lim_{n\to+\infty} \frac{k_n}{n} = 0$ (pas trop grand).

Graphiquement sur le Hill plot, il faut sélectionner la plus petite valeur de k_n pour laquelle l'estimateur de Hill est stable.

Le graphique II.8 ci-dessous représente le Hill plot de la sinistralité du régime fictif tout année confondu (2016 - 2021).

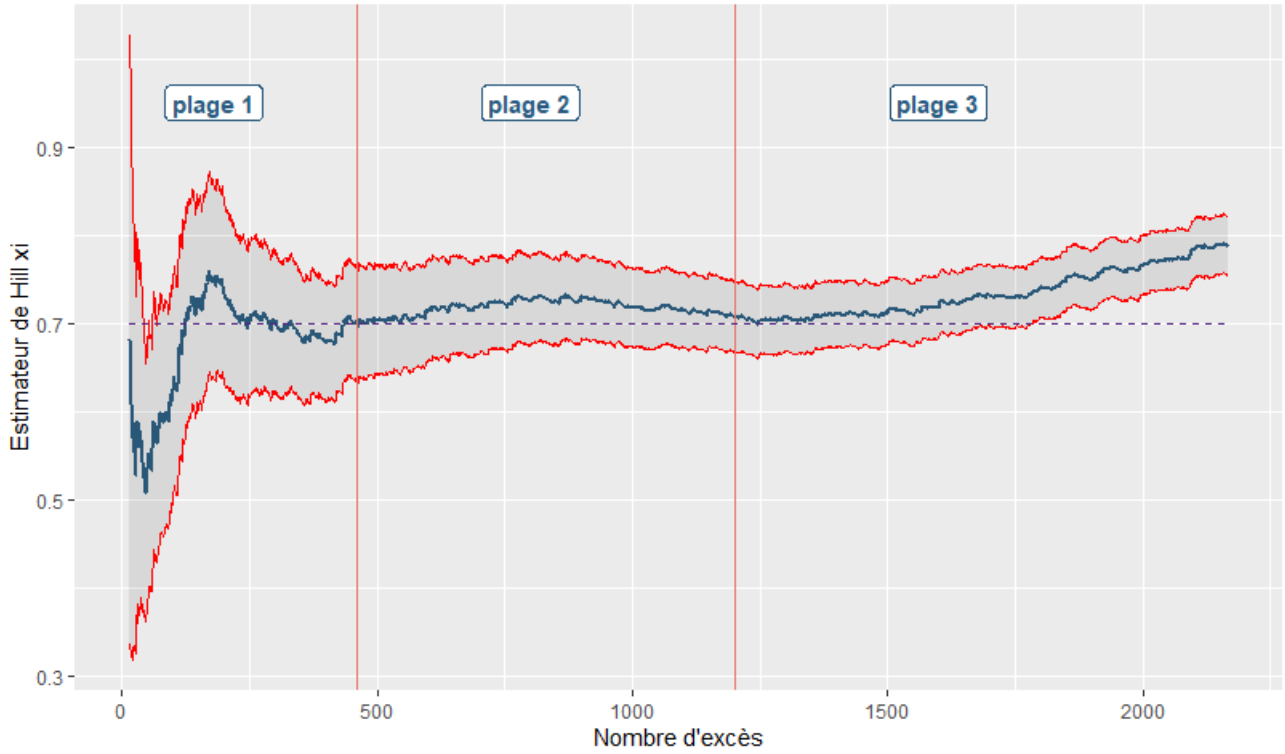


Figure II.8 – Hill Plot sur les données de la sinistralité du régime fictif (2016 - 2021)

Afin de bien analyser la courbe de l'indice de queue de Hill et d'identifier les seuils optimaux u, nous découpons le Hill Plot en trois plages selon les valeurs de k :

- la plage 1 correspond à $k \leq 460$. Dans cette plage, la courbe de l'indice de Hill de la sinistralité du régime est volatile et la variance du modèle est maximale : elle n'est donc pas acceptable ;
- la plage 2 corresond à $k \in]460, 1200]$. Dans cette plage la courbe de l'indice de Hill est à plusieurs endroits stable. Cette plage semble être adéquate pour la recherche des seuils optimaux. Une analyse approfondie de cette plage nous permet d'identifier trois intervalles sur lesquels la courbe de l'indice de queue de Hill est approximativement une droite. Ces intervalles correspondent aux intervalles de montants : [100 000€ , 125 000€], [135 000€ , 170 000€] et [190 000€ , 260 000€] ;
- quant à la plage 3, elle correspond à $k > 1200$. Dans cette plage, nous identifions également des intervalles où la courbe de l'indice de Hill est stable. Toutefois cette plage n'est pas acceptable car le nombre d'observations au delà des seuils de cette zone est trop élevé ce qui induit un grand biais dans le modèle.

En résumé, avec l'estimateur de Hill, nous pouvons choisir comme seuils : **125 000€** , **170 000€** et **260 000€** en tenant compte du comportement de la courbe de l'indice de queue de Hill et du dilemme biais-variance.

Tout comme la méthode du *mean excess plot* (ME-plot), l'estimation de Hill suscite parfois des interrogations. Cela est principalement dû au fait que cette méthode repose essentiellement sur des éléments graphiques, ce qui signifie que l'identification des intervalles appropriés ou de l'indice de queue de Hill peut varier d'un observateur à l'autre.

Graphique de Gerstengarbe

Le graphique de Gerstengarbe (*Gerstengarbe plot*) est issu des travaux de Gerstengarbe and Werner (1989). Cette méthode repose sur le test non-paramétrique de Mann-Kendall qui sert à déterminer si une tendance est identifiable dans une série temporelle. En considérant les différences de coût entre deux sinistres successifs, l'idée est qu'il semble raisonnable de s'attendre à un changement de comportement de ces écarts entre ceux issus des données extrêmes et ceux issus des données non-extrêmes. Ainsi, il existe un point de changement dans le comportement des écarts et ce point de changement est considéré comme le point de départ des données extrêmes. Le seuil des extrêmes est alors estimé par ce point de changement.

Mathématiquement, soit $X_1, \cdots, X_n$ l'échantillon des montants sinistres et $X_{(1)} \leq \cdots \leq X_{(n)}$ ce même échantillon ordonné par ordre croissant. La série des différences entre deux montants de sinistres est notée $\Delta_i = X_{(i)} - X_{(i-1)}$ pour tout $i \in [\![2; n]\!]$. On s'attend à un changement de comportement de cette série de différences, ce point de changement est le point d'entrée dans la zone des données extrêmes. Pour identifier ce point d'entrée, la version séquentielle du test de Mann-Kendall est utilisée. Pour $i = 1, \cdots, n-1$, la série suivante est calculée :

$$U_i = \frac{\sum_{k=2}^{i} n_k - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i-1)(i+5)}{72}}} \tag{II.15}$$

où $n_k = \sum_{j=2}^{k} \mathbb{1}_{\Delta_j \leq \Delta_k}$, c'est-à-dire le nombre de valeurs dans $\Delta_2, \cdots, \Delta_k$ inférieures à Δ_k. Une seconde série $\tilde{U}_i$ est calculée de la même manière mais en utilisant la série décroissante des différences $\Delta_n, \cdots, \Delta_2$. Le Gerstengarbe plot consiste à tracer les deux séries $(U_i)_{2 \leq i \leq n}$ et $\left(\tilde{U}_i\right)_{2 \leq i \leq n}$. Le point de départ des données extrêmes correspond alors au point d'intersection des deux séries.

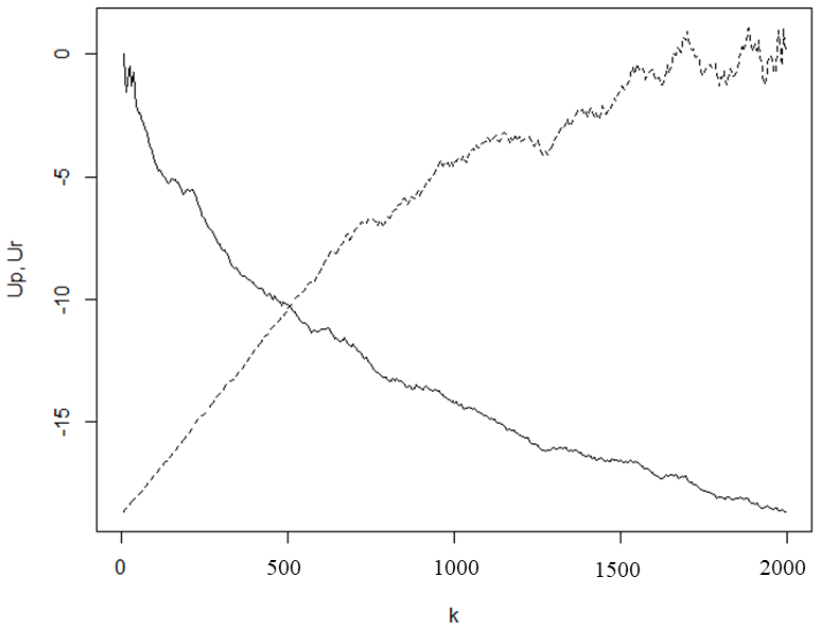


Figure II.9 – Gerstengarbe plot sur sur les données de la sinistralité du régime fictif (2016 - 2021)

Le graphique II.9 ci-dessous représente le Gerstengarbe plot de la sinistralité du régime fictif (2016 - 2021).

Avec la méthode graphique de *Gerstengarbe*, nous obtenons un seuil de **132 200**€ . Le seuil u ainsi déterminé correspond à la 525ème moins grande observation.

L'avantage du Gerstengarbe plot est de fournir une valeur fixe pour le seuil d'entrée dans la zone extrême, ce que ne permet pas le *mean excess plot* ou le Hill plot. Cependant, cette méthode n'est pas issue des travaux de la théorie des valeurs extrêmes, le seuil fourni par cette méthode est difficile à vérifier et valider.

Méthode de minimisation de l'AMSE

Cette méthode numérique de sélection d'un seuil optimal est issue des travaux de Caeiro and Gomes (2016). L'idée générale de cette méthode est de sélectionner un nombre d'excès k_0 qui minimise l'erreur quadratique moyenne de l'estimateur de Hill. On rappelle que pour un échantillon $X_1, \cdots, X_n$ de variables aléatoires indépendantes et de même loi appartenant au domaine d'attraction de Fréchet, l'estimateur de Hill de l'indice des valeurs extrêmes γ est donné par :

$$\hat{\xi}_n^{(H)}(k_n) = \frac{1}{k_n}\sum_{i=1}^{k_n} \ln\left(X_{(n-i+1)}\right) - \ln\left(X_{(n-k_n)}\right) \tag{II.16}$$

où $X_{(n-k_n)}$ est une estimation du seuil des extrêmes et k_n un entier compris entre 1 et n. Cette méthode consiste alors à sélectionner un seuil $\hat{u} = X_{(n-\hat{k}_0)}$ où $\hat{k}_0$ est une estimation de la valeur k_0 de k_n qui minimise l'erreur quadratique moyenne asymptotique (AMSE) de l'estimateur de Hill, c'est-à-dire la somme du biais au carré et de la variance de la distribution asymptotique de $\hat{\xi}_n^{(H)}(k_n)$. L'objectif est de sélectionner un seuil qui représente un bon compromis entre le biais et la variance de l'estimateur. Sous certaines conditions de régularité qui ne sont pas présentées ici, il est possible d'exprimer l'AMSE par :

$$\text{AMSE}\left(\hat{\xi}_n^{(H)}(k_n)\right) = \xi^2\left(\frac{1}{k_n} + \frac{\lambda^2}{(1-\rho)^2}\left(\frac{n}{k_n}\right)^{2\rho}\right) \tag{II.17}$$

On peut montrer que cette quantité est minimisée pour la valeur :

$$k_0 = \left\lfloor \left(\frac{(1-\rho)^2 n^{-2\rho}}{-2\rho\lambda^2}\right)^{\frac{1}{(1-2\rho)}} \right\rfloor \tag{II.18}$$

Avec $\lfloor . \rfloor$ la partie entière et ρ et λ des paramètres du second ordre qui contrôlent la vitesse de convergence de l'estimateur de Hill. Ces paramètres peuvent être estimés comme décrit par Caeiro and Gomes (2016), et ainsi obtenir une estimation du nombre d'excès à considérer pour estimer le seuil des extrêmes :

$$\hat{k}_0 = \left\lfloor \left(\frac{(1-\hat{\rho})^2 n^{-2\hat{\rho}}}{-2\hat{\rho}\hat{\lambda}^2}\right)^{\frac{1}{(1-2\rho)}} \right\rfloor \tag{II.19}$$

Finalement, cette méthode fournit le nombre d'excès optimal qui minimise l'AMSE de l'estimateur de Hill. Il existe une correspondance naturelle entre le nombre d'excès de l'échantillon et la valeur du seuil correspondant. En appliquant la méthode de minimisation de l'AMSE sur les données de la sinistralité du régime « beau temps » de 2016 à 2021, on obtient un seuil de **131 000€** qui correspond à la 518ème moins grande observation.

Cette méthode de détermination de seuil présente l'avantage de fournir une valeur pour l'estimation du seuil contrairement aux méthodes graphiques qui peuvent être difficiles à interpréter.

Résumé des différentes approches de détermination du seuil

Les résultats des quatre approches utilisées dans la démarche de la détermination du seuil des extrêmes u sont récapitulés dans le tableau ci-après :

	Mean excess plot	Hill Plot	Gestengarbe plot	Minimisation AMSE
Seuil u	125 000€ 180 000€	125 000€ 170 000€ 260 000€	132 200€	131 000€

Table II.2 – Synthèse des seuils par méthode de détermination

Les différents seuils obtenus par les quatre méthodes utilisées sont compris entre 125 000€ et 260 000€. La plupart des seuils obtenus sont proches de 130 000€ . En nous basant sur l'opinion d'experts, nous avons conclu que le seuil de **130 000€** apparaît de manière consistante dans la plupart des résultats obtenus, ce qui en fait un choix approprié pour distinguer les sinistres extrêmes des sinistres non-extrêmes. Avec ce seuil, les sinistres extrêmes représentent 5.7% de l'ensemble de la sinistralité du régime « beau temps ».

1.5 Ajustement de la GPD

Dans la sous-section précédente nous avons utilisé différentes méthodes de détermination du seuil dans le but d'aboutir à un seuil consistant de $u = 130\,000$€. Il nous est donc possible d'ajuster une loi de Pareto Généralisée aux observations extrêmes c'est-à-dire dont la valeur est au dessus de 130 000€.

Nous avons déjà présenté la fonction de répartition de la distribution GPD (formule II.3). La densité de probabilité correspondante est la suivante :

$$\text{densité GPD :} \qquad g^p(x,\sigma,\xi) = \begin{cases} \frac{1}{\sigma}\left(1+\xi\frac{x}{\sigma}\right)_+^{-\left(\frac{1}{\xi}+1\right)}, & \text{si } \xi \neq 0, \\ \frac{1}{\sigma}e^{\frac{-x}{\sigma}}, & \text{si } \xi = 0, \end{cases} \tag{II.20}$$

on peut poser $x = y - u$, u étant le seuil de nos valeurs extrêmes. L'estimation des paramètres peut être effectuée par le maximum de vraisemblance (voir annexe A). La *log-vaisemblance* s'écrit comme suit :

$$-\log g^{p}(y,\sigma,\xi,u) = \begin{cases} -\log(\sigma) - \left(\frac{1}{\xi}+1\right)\log\left(1+\xi\frac{y-u}{\sigma}\right), & \text{si } \xi \neq 0, \\ -log(\sigma) + \frac{y-u}{\sigma}, & \text{si } \xi = 0, \end{cases} \tag{II.21}$$

Pour résoudre la minimisation de la log-vraissemblance, nous avons utilisé la fonction *gpdFit* du package R fExtremes (2009). Les paramètres σ et ξ sont estimés sur la période 2016 - 2021. Les résultats obtenus sont les suivants :

Estimateur	Valeur	Déviance	log-vraisemblance
$\hat{\xi}$	6.987528e-01	3.920776e-02	22265.35
$\hat{\sigma}$	1.689524e+05	8.280392e+02	

Table II.3 – Estimation des paramètres de la GPD ajustée aux données

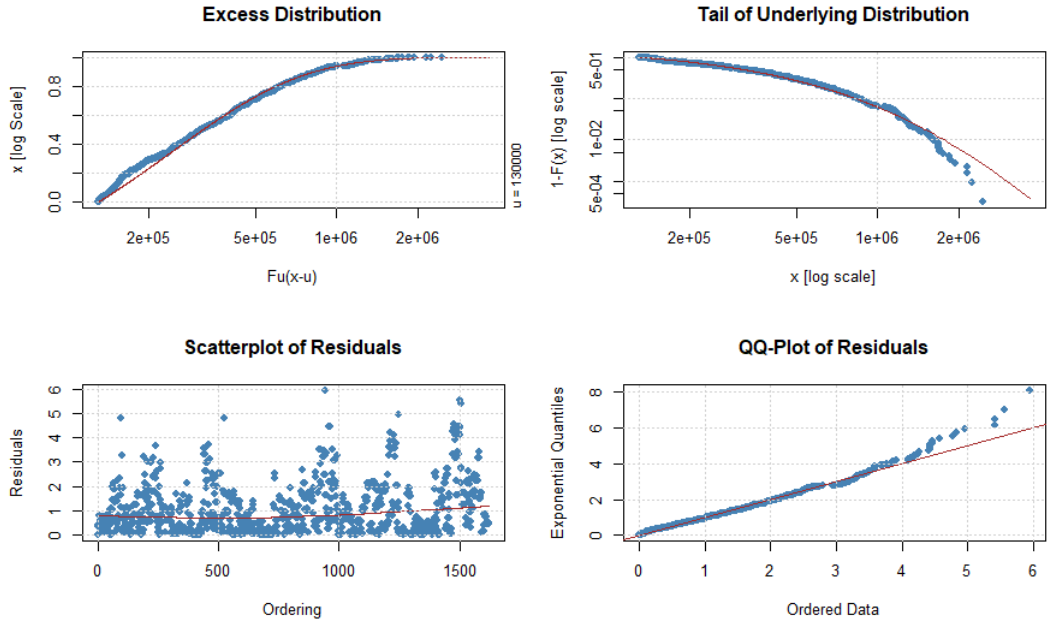


Figure II.10 – Distribution de la loi de dépassement (en haut à gauche), queue de la distribution du sous-jacente (en haut à droite), nuage de points des résidus (en bas à gauche) et QQ plot de la GDP (en bas à droite)

Il est observable que la loi que nous avons adaptée montre une correspondance satisfaisante avec une GPD dont les paramètres sont présentés dans le tableau II.3.

Pour s'assurer que statistiquement la distribution de la sinistralité du régime au-delà de 130 000€ est bien une pareto généralée, nous allons réaliser des validations à l'aide des tests d'adéquation (test de Kolmogorov-Smirnov, test d'Anderson-Darling et test de Cramer Von Mises). La description théorique de ces tests se trouve à la section 3 de l'annexe A.

Test	Distance (D_n)	Seuils critiques	p-val	décison
Kolmogorov-Smirnov	0.1322	0.41	0.885	*Acceptée*
Anderson-Darling	0.341	0.75	0.95	*Acceptée*
Cramer Von Mises	2.512	2.65	0.35	*Acceptée*

Table II.4 – Résultats des tests d'adéquation

Le tableau II.4 présente notamment les P-value des test de Kolmogorov-Smirnov, d'Anderson-Darling et de Cramer Von Mises. Les tests d'adéquation, avec un niveau de confiance de 95%, confirment la bonne ajustement de la loi GDP dont les paramètres se trouvent dans le tableau II.3 avec les données de la sinistralité au dessus du seuil $u = 130\,000€$. Cependant, ces tests permettent de vérifier l'ajustement à une loi exacte et ponctuelle, et non à une famille de lois. Pour nous assurer de l'ajustement à la famille GDP, nous avons effectué le test du khi-2 (sous-section 3.4 de l'annexe A). Nous obtenons une valeur p (P-value) de 0.15, supérieure à 5%, confirmant ainsi le bon ajustement à la famille GDP.

2 Apprentissage statistique (machine learning)

2.1 Généralités sur l'apprentissage

L'apprentissage statistique ou encore *machine learning* en anglais, est un champ d'étude de l'intelligence artificielle (IA) qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d' « apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes. Arthur Samuel (1959), l'un des pionniers de l'intelligence artificielle et de l'apprentissage statistique, en donne une définition plus générale et moins technique :

« *L'apprentissage statistique est la discipline donnant aux ordinateurs la capacité d'apprendre sans qu'ils soient explicitement programmés* »

Un autre pionnier dans ce domaine est Vapnik, notamment grâce à ses apports fournis dans l'ouvrage Vapnik (2013), a posé les idées fondamentales qui sous-tendent la théorie statistique de l'apprentissage et de sa généralisation.

Aujourd'hui, les techniques d'apprentissage statistique (*machine learning*) prennent une place prépondérante dans de nombreux secteurs allant de la recherche médicale, de l'informatique à la l'industrie en passant par l'assurance. Grâce notamment à leurs performances, les modèles de *machine learning* fournissent une alternative aux méthodes plus classiques utilisées en actuariat. Cependant, ce gain de précision à un coût : il se fait au détriment de l'interprétabilité et de la transparence. L'interprétabilité désigne généralement la capacité d'expliquer ou de présenter une information dans des termes humainement compréhensibles. Par ailleurs, le RGPD (règlement général sur la protection des données) impose une nouvelle exigence sur la transparence, obligeant à justifier l'usage de modèle avancé dite de "boîte noire". Aussi, l'ACPR (Autorité de contrôle prudentiel et de résolution) se doit d'assurer la protection des clients : l'assureur doit être capable de justifier toutes les décisions prises de manière détaillée et ne pas transférer la responsabilité à la machine ou à la "boîte noire".

L'objectif général de l'apprentissage statistiques (*machine learning*) est d'élaborer des procédures automatiques qui permettent de mettre en évidence des règles générales à partir d'exemples. Il s'agit donc d'imiter le fonctionnement inductif du cerveau humain dans le but de développer des systèmes d'intelligence artificielle. Ce domaine est considéré aujourd'hui comme une branche de l'informatique (ou, plus précisément, de *computer science*). Cependant, les liens avec la statistique sont étroits, notamment avec la théorie non-paramétrique. Pour schématiser, la différence principale entre la statistique et l'apprentissage statistique, le concept central en statistique est le modèle, tandis qu'en apprentissage statistique, c'est l'algorithme (Dalalyan (2018)).

Le point de départ en apprentissage est l'échantillon $Z_1, \ldots, Z_n$, que l'on veut utiliser, par exemple, pour faire des prédictions ou encore détecter des régularités (*patterns*). En apprentissage statistique, il existe plusieurs types de tâches (techniques) caractérisées par la nature de l'échantillon et l'objectif poursuivi.

Dans les méthodes d'**apprentissage supervisé** chaque observation $Z_i = (X_i, Y_i)$ de l'ensemble de données est composé d'une variable d'entrée $X_i \in \mathbb{R}^d$, souvent appelée prédicteur ou *feature*, et est associée à une étiquette ou à une valeur de sortie correspondante $Y_i \in \mathbb{R}$ (encore appelée *label*). L'objectif est donc d'entraîner l'algorithme à prédire les étiquettes ou *label*.

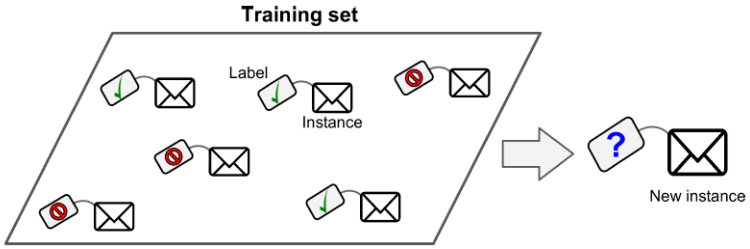
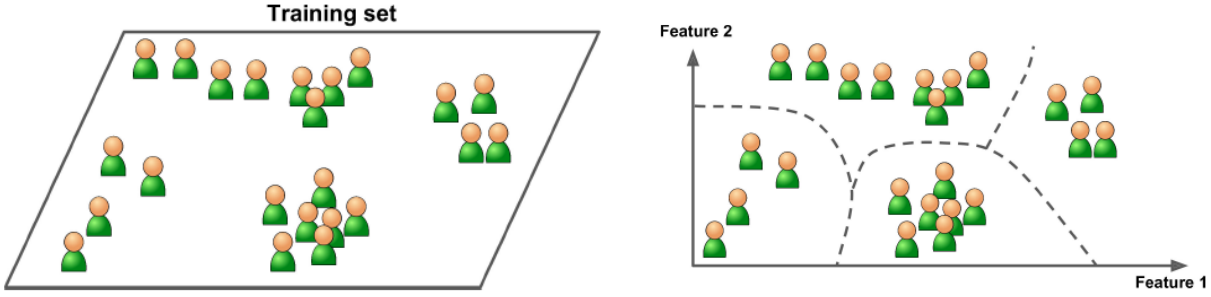


Figure II.11 – Illustration de l'apprentissage supervisé extrait du livre de *Deep Learning avec Keras et TensorFlow* de Géron (2020)

Dans l'**apprentissage non supervisé**, il n'y a pas d'étiquettes pour les données d'apprentissage Z_i. Le plus souvent, $Z_i \in \mathbb{R}^d$ pour $d \in \mathbb{N}$ assez grand, le but est de caractériser la loi de probabilité ayant engendré ces observations. Le *clustering* ou encore l'estimation de la densité sont les problèmes les plus étudiés en apprentissage non supervisé.



(a) Non clusterisé (b) Clusterisé

Figure II.12 – Illustration de l'apprentissage non supervisé extrait du livre de *Deep Learning avec Keras et TensorFlow* de Géron (2020)

Dans la méthode **semi-supervisée**, seule une faible proportion n_1 des observations est étiquetée. On a donc $Z_i = (X_i, Y_i)$ pour $i \in [\![1, n_1]\!]$ et $Z_i = X_i$ pour $i \in [\![n_1 + 1, n]\!]$. Le but est le même qu'en apprentissage supervisé. Ce cadre est intéressant pour de nombreuses applications où le coût d'étiquetage est très élevé.

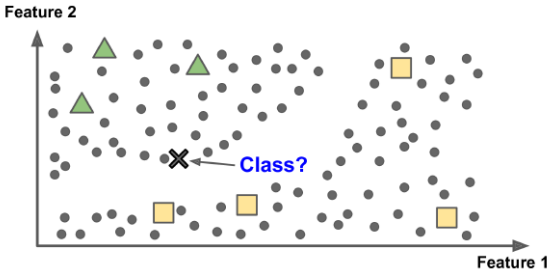


Figure II.13 – Illustration de l'apprentissage semi-supervisé extrait du livre de *Deep Learning avec Keras et TensorFlow* de Géron (2020)

Dans l'**apprentissage par renforcement**, l'algorithme détermine les actions à entreprendre dans une situation pour maximiser une récompense (sous la forme d'un nombre) dans l'optique d'atteindre un objectif spécifique. Elle est très populaire pour les robots et les agents en intelligence artificielle (IA).

2.2 Apprentissage supervisé : régression et classification

Supposons que l'on observe une base de données composées de n couples $Z_i = (X_i, Y_i)$ que nous supposons être des réalisations indépendantes d'une même loi P inconnue. On écrira :

$$Z_i = (X_i, Y_i) \approx P \tag{II.22}$$

Les X_i appartiennent à un espace $\mathcal{X}$ et s'appellent les entrées ou les *features*. Typiquement, $\mathcal{X} = \mathbb{R}^d$ pour un grand entier d. Les Y_i appartiennent à un espace $\mathcal{Y}$, et s'appellent les sorties ou les étiquettes. Typiquement, $\mathcal{Y}$ est fini ou $\mathcal{Y}$ est un sous-ensemble de $\mathbb{R}$.

L'objectif de l'apprentissage supervisé est de prévoir l'étiquette Y associée à toute nouvelle entrée X, où il est sous-entendu que la paire (X,Y) est une nouvelle réalisation de la loi P, cette réalisation étant indépendante des réalisations précédemment observées.

Une fonction de prédiction est une fonction mesurable de $\mathcal{X}$ dans $\mathcal{Y}$. Dans ce qui suit, nous supposons que toutes les quantités que nous manipulons sont mesurables. L'ensemble de toutes les fonctions de prédiction est noté $\mathcal{F}(\mathcal{X}, \mathcal{Y})$. La base de données $(Z_1, \ldots, Z_n)$ est appelée ensemble d'apprentissage.

Un algorithme d'apprentissage est une fonction qui à tout ensemble d'apprentissage renvoie une fonction de prédiction, c'est-à-dire une fonction de la réunion :

$$\bigcup_{k \in \mathbb{N}^*} \mathcal{Z}^k$$

dans l'ensemble $\mathscr{F}(\mathscr{X}, \mathscr{Y})$, où $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$. C'est un estimateur de « meilleure » fonction de prédiction, où le terme « meilleure » sera précisé ultérieurement. Soit $l(y, y')$ la perte encourue lorsque la sortie réelle est y et la sortie prédite est y'. La fonction $l : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$ est appelée fonction de perte. Les deux exemples les plus fréquemment utilisés sont :

- **La classification** : $l(y, y') = \mathbb{1}_{\{y \neq y'\}}$, c'est-à-dire $l(y, y') = 1$ si $y \neq y'$ et $l(y, y') = 0$ sinon. Un problème d'apprentissage pour lequel cette fonction de perte est utilisée est appelé problème de classification. L'ensemble $\mathscr{Y}$ considéré en classification est le plus souvent fini, voire même de cardinal deux en classification binaire.
- **La régression** L_p : $\mathscr{Y} = \mathbb{R}$ et $l(y, y') = |y - y'|^p$ où $p \geq 1$ est un réel fixé. Dans ce cas, on parle de régression L_p. La tâche d'apprentissage lorsque $p = 2$ est aussi appelée régression aux moindres carrés (MCO).

La qualité d'une fonction de prédiction $f : \mathscr{X} \to \mathscr{Y}$ est mesurée par son risque ou erreur de généralisation :

$$R_P(f) = \mathbb{E}_P\left[\ell(Y, f(X))\right] \tag{II.23}$$

Le risque est donc l'espérance par rapport à loi P de la perte encourue sur la donnée (X, Y) par la fonction de prédiction f. La « meilleure » fonction de prédiction est une fonction de $\mathscr{F}(\mathscr{X}, \mathscr{Y})$ minimisant le risque R_P :

$$f_P^* \in \underset{f \in \mathscr{F}(\mathscr{X}, \mathscr{Y})}{\arg\min}\, R_P(f) \tag{II.24}$$

Une telle fonction f_P^* n'existe pas nécessairement mais existe pour les fonctions de pertes usuelles. Cette « meilleure » fonction est appelée fonction oracle ou prédicteur de Bayes. Elle dépend de la probabilité inconnue P et, par conséquent, est inconnue.

2.3 Théorie de minimisation du risque empirique

Le but d'un algorithme d'apprentissage est de trouver une fonction de prédiction dont le risque est aussi faible que possible (autrement dit aussi proche que possible du risque des prédicteurs oracles).

La distribution $\mathbb{P}$ générant les données étant inconnue, le risque $\mathscr{R}_P$ et les prédicteurs bayes sont inconnus. Néanmoins, le risque $R_{\mathbb{P}}(g)$ peut être estimé par son équivalent empirique :

$$\hat{R}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) \tag{II.25}$$

Si on suppose que $\mathbb{E}_{\mathbb{P}}[\ell(X_i, g(X_i))^2] < +\infty$, alors nous déduisons de la loi forte des grandes nombres et du théorème central limite que :

$$\hat{R}_n(f) \xrightarrow[n \to +\infty]{p.s} R_{\mathbb{P}}(f) \tag{II.26}$$

$$\sqrt{n}(\hat{R}_n(f) - R_{\mathbb{P}}(f)) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \mathbb{V}(\ell(Y, f(X))) \tag{II.27}$$

Pour toute fonction de prédiction f, la variable aléatoire $\hat{R}_n(f)$ effectue donc des déviation en $\mathcal{O}(\frac{1}{\sqrt{n}})$ autour de sa moyenne $R_{\mathbb{P}}(f)$. Puisque nous cherchons une fonction qui minimise le risque $R_{\mathbb{P}}$ et puisque ce risque est approché par le risque empirique $\hat{R}_n$, ul est naturel de considérer l'algorithme d'apprentissage, dit de minimisation du risque empirique, défini par :

$$\hat{f}_{n,\mathcal{G}} = \arg\min_{f\in\mathcal{G}} \hat{R}_n \tag{II.28}$$

Où $\mathcal{G}$ est un sous-ensemble de $\mathcal{F}(\mathcal{X}, \mathcal{Y})$. Prendre $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$ n'est pas une bonne idée. Tout d'abord, cela entraîne un problème de choix puisqu'en général, pour tout ensemble d'apprentissage, il existe une infinité de fonctions de prédiction minimisant le risque empirique. D'autre part, si l'on prend l'algorithme du plus proche voisin comme minimiseur du risque empirique, alors on peut montrer que cet algorithme est loin d'être universellement consistant (Dalalyan (2018)).

Prendre $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$ mène en général à un sur-apprentissage (**overfitting**) dans la mesure o'u l'algorithme qui en résulte a un risque empirique qui peut être très inférieur à son risque réel. En pratique, il faut prendre $\mathcal{G}$ suffisamment grand pour pouvoir raisonnablement approcher toute fonction, tout en ne le prenant pas trop grand afin d'éviter que l'algorithme sur-apprenne. La "grandeur" de l'ensemble $\mathcal{G}$ est appelée capacité ou complexité. Un autre point de vue consiste à rajouter à $R_n(f)$ une pénalisation, quand par exemple, la fonction f est trop irrégulière. Ces deux approches sont en fait proches l'une de l'autre. En notant $f^*_{\mathbb{P},G} = \arg\min_{f\in\mathcal{G}} R_{\mathbb{P}}(f)$ (prédicteur de Bayes) et

$$R_{\mathbb{P}}\left(\hat{g}_{n,G}\right) \geq R_{\mathbb{P}}\left(g^*_{p,G}\right) \geq R_{\mathbb{P}}\left(g^*_{\mathbb{P}}\right)$$

On peut alors décomposer l'excès de risque de $\hat{g}_{n,G}$ par rapport au prédicteur de Bayes $g^*_{\mathbb{P}}$, en deux termes :

$$R_{\mathbb{P}}\left(\hat{g}_{n,G}\right) - R_{\mathbb{P}}\left(g^*_{\mathbb{P}}\right) = \underbrace{\left(R_{\mathbb{P}}\left(\hat{g}_{n,G}\right) - R_{\mathbb{P}}\left(g^*_{p,G}\right)\right)}_{\text{erreur d'estimation}} + \underbrace{\left(R_{\mathbb{P}}\left(g^*_{p,G}\right) - R_{\mathbb{P}}\left(g^*_{\mathbb{P}}\right)\right)}_{\text{erreur d'approximation}}$$

Le premier terme $R_{\mathbb{P}}\left(\hat{g}_{n,G}\right) - R_{\mathbb{P}}\left(g^*_{\mathbb{E}_p,G}\right)$ est appelé erreur stochastique ou erreur d'estimation, tandis que le deuxième $R_{\mathbb{P}}\left(g^*_{\mathbb{E}_p,G}\right) - R_{\mathbb{P}}\left(g^*_{\mathbb{P}}\right)$ s'appelle erreur systématique ou erreur d'approximation ou encore biais. Il découle que plus G va être grand plus l'erreur d'approximation sera faible (biais fiable), mais plus l'erreur d'estimation sera grande (variance élevée). Plus G sera petit, nous aurons alors les conclusions inverses, c'est-à-dire plus le biais sera élevé mais plus la variance sera faible. Nous avons donc un compromis à réaliser, appelé dilemme "biais-variance", représenté sur la figure . Nous pouvons également définir ce di-

lemme biais-variance d'une autre manière. Considérons notre ensemble d'apprentissage :

$$\{x_1, \ldots, x_n\} \text{ et } \{y_1, \ldots y_n\}, \text{ avec } \forall i \in \{1, \ldots, n\} x_i \in \mathbb{R}^p \text{ et } y_i \in \mathbb{R}$$

On suppose que l'on peut écrire :

$$\forall i \in \{1, \ldots, n\}, y_i = f(x_i) + \epsilon_i$$

Avec ϵ_i centré et de variance $\sigma^2 (\sigma > 0)$. Soit $\hat{f}$ la fonction associée au modèle qu'on utilise, alors l'erreur attendue est : $\mathbb{E}_{\mathbb{P}}\left[(Y - \hat{f}(X))^2\right]$. On peut alors montrer que cette erreur se décompose en trois termes, à savoir :

$$\mathbb{E}_{\mathbb{P}}\left[(Y - \hat{f}(X))^2\right] = B[\hat{f}(X)]^2 + \text{Var}[\hat{f}(X)] + \sigma^2$$

Les différents termes présents dans cette équation sont :

— Le biais :

$$B[\hat{f}(X)] = \mathbb{E}_{\mathbb{P}}[\hat{f}(X) - f(X)]$$

Il peut être interprété comme l'erreur due au modèle simplifié utilisé.

— La variance :

$$\text{Var}[\hat{f}(X)] = \mathbb{E}_{\mathbf{P}}\left[\left(\hat{f}(X) - \mathbb{E}_{\mathbf{P}}[\hat{f}(X)]\right)^2\right]$$

Elle peut être vue comme l'erreur due à la sensibilité aux petites fluctuations de l'échantillon d'apprentissage.

— L'erreur irréductible σ^2, résultant du bruit lui-même.

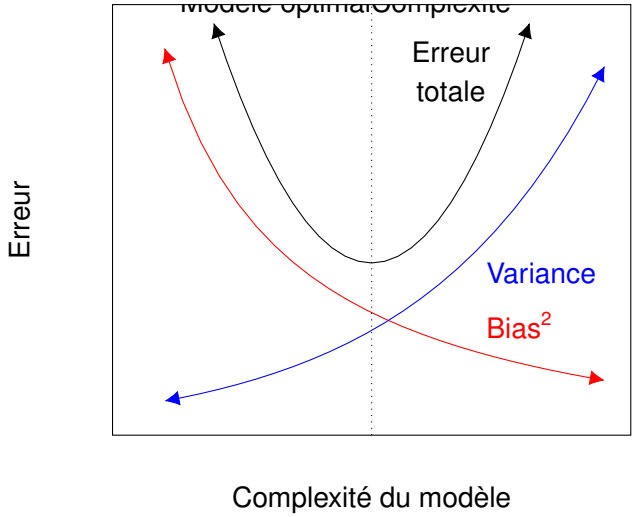


Figure II.14 – Compromis biais-variance

Afin d'obtenir une estimation plus robuste, avec biais et variance, de la performance de validation du modèle on a recourt à la technique de validation croisée.

Il existe différentes méthodes de validation croisée, les trois les plus utilisées sont :

— La validation croisée à k blocs, « k-fold cross-validation » : on divise l'échantillon original en k échantillons (ou « blocs »), puis on sélectionne un des k échantillons

comme ensemble de validation pendant que les $k-1$ autres échantillons constituent l'ensemble d'apprentissage. Après apprentissage, on peut calculer une performance de validation. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les blocs prédéfinis. À l'issue de la procédure nous obtenons ainsi k scores de performances, un par bloc. La moyenne et l'écart type des k scores de performances peuvent être calculés pour estimer le biais et la variance de la performance de validation.

— La validation croisée d'un contre tous, « leave-one-out cross-validation » (LOOCV) : il s'agit d'un cas particulier de la validation croisée à k blocs où $k = n$. C'est-à-dire qu'à chaque itération d'apprentissage-validation, l'apprentissage se fait sur $n-1$ observations et la validation sur l'unique observation restante.

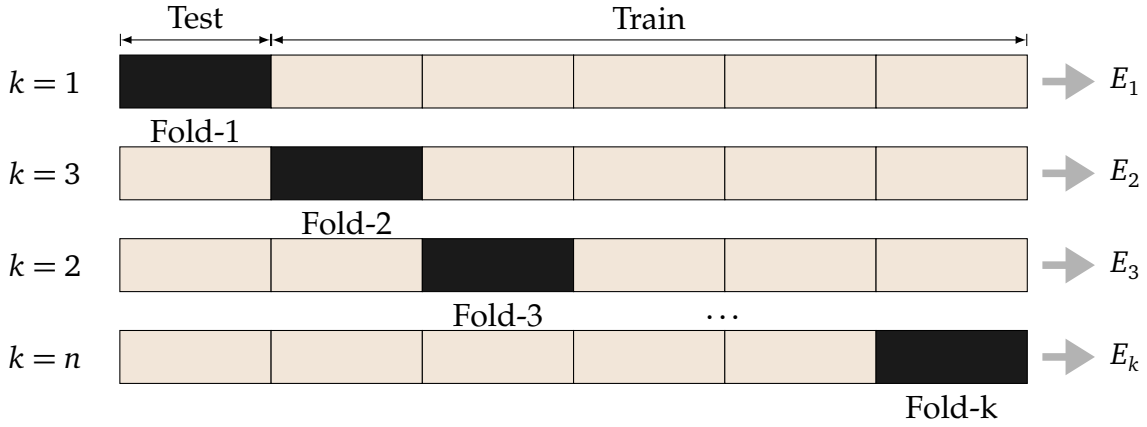


Figure II.15 – Illustration de la validation croisée à k blocs « k-fold cross-validation »

2.4 Application combinant la TVE et le machine learning

Dans la littérature sur la modélisation des sinistres à coût large, les approches combinant apprentissage automatique et Théorie des Valeurs Extrêmes sont très peu présentes et récentes.

Les travaux de Farkas et al. (2021b) ont montré que l'apprentissage statistique est utile pour évaluer et améliorer la modélisation des sinistres extrêmes avec prise en compte de variable explicative dans la queue de distribution. Pour ce faire, ils ont proposé une méthode d'arbre de régression dans laquelle la perte quadratique utilisée dans la phase "de croissance" de l'arbre a été remplacée par une perte de log-vraisemblance basée sur la vraisemblance des distributions de Pareto généralisées. Ils ont illustré leur approche méthodologique sur une base de données publique portant sur le risque cyber aux Etats-Unis qui est la *Privacy Rights Clearinghouse* (PRC [1]). Cette base rassemble des informations sur chaque événement cyber (son type, le nombre d'enregistrements concernés par la violation, une description de l'événement) et sa victime (le nom de l'entreprise ciblée, ses activités, sa localisation). Leur approche a permis d'identifier des critères de classification et d'évaluation

1. The Privacy Rights Clearinghouse, https://privacyrights.org/data-breaches

des sinistres pour le risque cyber avec prise en compte de variable exogène telles que le type d'organisation, la *source des données*, le *type de domaine* et l'*année de survenance* de l'évènement.

Cependant, Arthur and Christian (2022) trouvent que ces estimateurs non paramétriques précédents de la fonction d'indice de queue ne sont pas en mesure de prendre en compte l'hypothèse selon laquelle cette fonction ne prend qu'un nombre fini de valeurs sur l'espace des covariables. En outre, ils ont proposé une méthode pour estimer à la fois les sous-ensembles de partition de l'espace des covariables ainsi que les valeurs de la fonction d'indice de queue. Leur méthode combine deux étapes : premièrement, un ensemble d'arbres additif basé sur la déviance Gamma est ajusté (qui inclut une forêt aléatoire et un renforcement d'arbre de gradient), deuxièmement, un clustering hiérarchique avec des contraintes spatiales est utilisé pour estimer les sous-ensembles de la partition. Cette procédure a été illustrée sur des données simulées. Une étude de cas réel sur les dommages matériels causés par les tornades a été également présentée.

Pour d'autres applications non assurantielle combinant l'apprentissage automatique et du Machine Learning, , le lecteur pourra se référer à la thèse de Sabourin (2021) institulé *Extreme Value Theory and Machine Learning* dont les travaux ont contribué à combler le fossé entre la théorie des valeurs extrêmes et l'apprentissage statistique d'un point de vue théorique ainsi que dans les applications.

3 Arbres de régression et analyse des valeurs extrêmes

Dans cette section, on désigne par Y une variable de réponse (typiquement une variable coût avec des valeurs extrêmes), et $\mathbf{X} \in \mathbb{R}^d$ des covariables. Les arbres de régression sont des outils pratiques lorsque l'on veut simultanément prédire une réponse et filtrer l'hétérogénéité en déterminant des clusters parmi les données. La sortie d'un arbre de régression est très facile à comprendre et ne nécessite aucune connaissance statistique pour la lire et l'interpréter. L'arbre de régression est l'un des moyens les plus rapides d'identifier les variables les plus significatives et les relations entre deux variables ou plus. Les arbres de régression visent donc à déterminer des « règles » pour rassembler des observations dans des classes de risque en fonction des valeurs de leurs caractéristiques X_i. Ils sont donc particulièrement adaptés aux situations où la variété des profils de X_i induit une certaine hétérogénéité. Ainsi, dans cette section nous presenterons l'algoritheme de CART (Classification and Regression Trees) utilisé pour contruire l'arbre de régression et nous allons introduit l'*arbre de regression Pareto Généralisé (GP)*.

3.1 Arbres de régréssion

Les arbres de régression sont des outils de modélisation qui permettent d'introduire une modélisation de l'hétérogénéité (non linéaire) entre les observations, en les divisant en classes sur lesquelles différents modèles de régression sont ajustés. Le but est de rétrouver

une fonction de régression g^* qui minimise le risuqe empirque $\mathbb{E}[\ell(Y, g(X))]$, avec $g \in \mathcal{G}$ où $\mathcal{G}$ est une classe de fonction sur $\mathbb{R}^d$ et ℓ est une fonction de perte. Farkas et al. (2021b) dans leur étude ont considéré trois types de fonctions différents :

- **la perte quadratique** $\ell(y, g(x)) = (y - g(x))^2$ correspondant à la situation où l'on s'intéresse à la moyenne conditionnelle $g(x) = \mathbb{E}[Y|X = x]$ et $\mathcal{G}$ est l'ensemble des fonctions de x à moment fini du second ordre ;
- **la perte absolue** $\ell(y, g(x)) = |y - g(x)|$ où g^* est la médiane conditionnelle ;
- **la perte log-likelihood** $\ell(y, g(x)) = -log f_{g(x}(y)$ où $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ est une famille paramétrique de densités. Cela correspond au cas où l'on suppose que la loi conditionnelle de $Y|X = x$ appartient à la famille paramétrique $\mathcal{F}$ pour tout x, de paramètre $g(x)$ dépendant de x.

La procédure d'obtention de l'arbre de régression comprend deux phases : une phase de « croissance » qui correspond à l'algorithme CART, et une étape de « pruning » qui consiste en l'extraction d'un sous-arbre de la décomposition obtenue dans la phase initiale. L'élagage (*pruning*) peut donc être compris comme une procédure de sélection de modèle.

3.1.1 Construction de l'arbre maximal : algorithme du CART

Les arbres de classification et de régression (CART) [Breiman et al. (1984)] sont couramment utilisés dans toutes sortes d'application de la data science parce qu'ils sont considérés comme l'une des meilleures méthodes d'apprentissage supervisé. Les CART constituent une classe de modèles prédictifs avec une précision, une stabilité et une capacité d'interprétation élevées. Il crée un arbre en divisant l'échantillon en deux ensembles homogènes ou plus en fonction du séparateur le plus significatif dans les variables explicatives.

Cependant, considérer l'ensemble des partitionnements possibles de l'espace des prédicteurs nécessiterait des calculs excessivement lourds. Pour cette raison, un algorithme *glouton* et *top-down* (appelé *recursive binary splitting*) est utilisé afin de construire l'arbre de façon récursive. On parle d'algorithme *glouton* car à chaque étape de la construction de l'arbre, on construit la meilleure division possible du nœud en deux sous-nœuds. Par ailleurs, on parle d'algorithme *top-down* car l'algorithme démarre à la racine de l'arbre (au point où toutes les observations appartiennent à une seule région) et sépare ensuite l'espace des prédicteurs en ajoutant progressivement des nœuds.

L'algorithme CART consiste donc à déterminer itérativement un ensembe de règles $\mathcal{R} = (R_j)_{j \in [\![1,J]\!]}$ qui divise l'espace des prédicteurs $\mathcal{X}$ en J régions distinctes, disjointes deux à deux. Plus précisément, pour chaque valeur possible des variables explicatives $x = (x_1, \ldots, x_d)$, $R_j(x) = 1$ ou 0 selon que certaines conditions soient satisfaites par x, avec $R_j(x)R_{j'}(x) = 0$ pour $j \neq j'$ et $\sum_j R_j(x) = 1$. En thérie, les régions pourraient avoir n'importe quelle forme. Néanmoins, il est choisi de diviser l'espace des prédicteurs en rectangles ou boîtes de grande dimension, pour des raisons de simplicité et de facilité l'interprétation du modèle prédictif résultant. Ainsi, dans le cas d'arbres de régression, si d = 1, les règles peuvent être identifiées

comme des segments de partitionnement, si d = 2 ce sont des rectangles (hyper-rectangles dans le cas général). La détermination de ces règles d'une étape à l'autre peut être représenté sous la forme d'un arbre binaire, puisque chaque règle R_j à l'étape k génère deux règles R_{j1} et R_{j2} (avec $R_{j1}(x)+R_{j2}(x)=0$ si $R_j(x)=0$) à l'étape $k+1$. L'algorithme peut être résumé comme suit :

Encadré 1 : Résumé algortithme CART

Etape 1 : Cette étape consiste à fixer la racine de l'arbre. Ainsi, $R_1(x)=1$ pour tout x et $n_1=1$

Etape k +1 : Soit $(R_1,R_2,\ldots,R_{n_k})$ les règles de partition de $\mathscr{X}$ obtenu à l'étape k.

Pour j allant de 1 à n_k,

- Si toutes les observations telles que $R_j(x)=1$ ont les mêmes caractéristiques, alors on conserve la règle R_j car il n'est plus possible de segmenter la population ;
- Sinon, la règle R_j est remplacée par deux nouvelles règles R_{j1} et R_{j2} qui s'obtiennent de la façon suivante : pour chaque variable X^d , on cherche à définir le meilleur seuil S_{X^k}, pour diviser les données, de sorte que $S_{X^d} = \arg\min_{x^d} L(R_j,x^d)$, avec

$$\begin{aligned} L(R_j,x^d) &= \sum_{i=1}^n \ell(Y_i,\hat{g}(R_j))R_j(x) - \sum_{i=1}^n \ell(Y_i,g_{d-}(X_i,R_j))\mathbb{1}_{X_i\leq x^d}R_j(x) \\ &\quad - \sum_{i=1}^n \ell(Y_i,g_{d+}(X_i,R_j))\mathbb{1}_{X_i>x^d}R_j(x) \end{aligned} \tag{II.29}$$

Où

$$\begin{aligned} \hat{g}(R_j) &= \arg\min_{g\in\mathscr{G}} \sum_{i=1}^n \ell(Y_o,g(X_i))R_j(X_i) \\ g_{d-}(X,R_j) &= \arg\min_{g\in\mathscr{G}} \sum_{i=1}^n \ell(Y_i,g(X_i))\mathbb{1}_{X_i\leq x^d}R_j(X_i) \\ g_{d+}(X,R_j) &= \arg\min_{g\in\mathscr{G}} \sum_{i=1}^n \ell(Y_i,g(X_i))\mathbb{1}_{X_i>x^d}R_j(x) \end{aligned} \tag{II.30}$$

Ensuite, on selectionne la meilleure variable à prendre en compte : $\hat{d} = \arg\min_d L(R_j,S_{X^d})$ et on définit deux nouvelles règles $R_{j1}(x)=R_j(x)\mathbb{1}_{X^{(\hat{d})}\leq S_{X^{(\hat{d})}}}$ et $R_{j2}(x)=R_j(x)\mathbb{1}_{X^{(\hat{d})}>S_{X^{(\hat{d})}}}$

- à la fin de l'étape, on note n_{k+1} le nouveau nombre de règles.

Règle d'arrêt : L'agorithme prend fin lorsqu'on obtient pas de nouvelle règle c'est-à-dire $n_{k+1}=n_k$

♠

Remarque 1. *Dans cette version de l'algorithme CART, toutes les variables sont continues ou de valeur 0, 1. Pour les variables qualitatives à plus de deux modalités, elles doivent être transformées en variables binaires, ou l'algorithme doit être légèrement modifié pour que l'étape de découpage de chaque R_j se fasse en trouvant la meilleure partition en deux groupes sur les valeurs des modalités qui minimise la fonction de perte. Cela peut se faire en ordonnant les modalités par rapport à la valeur moyenne – ou la valeur médiane – de la réponse pour les observations associées à cette modalité. La règle d'arrêt peut également être légèrement modifiée pour s'assurer qu'il y a un nombre minimal de points des données d'origine dans chaque feuille de l'arbre à chaque étape.*

Pour passer de l'arbre définit par un ensemble de règles $\mathscr{R} = (R_j)_{j=1,\dots,J}$ à la fonction de régression, un estimateur $\hat{g}^{\mathscr{R}}$ de la fonction g est le suivant :

$$\hat{g}(x) = \sum_{j=1}^{n} \hat{g}(R_j) R_j(x) \tag{II.31}$$

Le dernier ensemble de règles $\mathscr{R}^M$ obtenu à partir de l'algorithme CART est appelé l'arbre maximal T_{max}. Cela conduit à un estimateur trivial de g, puisque soit le nombre d'observations dans une feuille est unique, soit toutes les observations dans cette feuille ont les mêmes caractéristiques x. L'arbre T_{max} possède un faible biais mais obtiendra une grande erreur de généralisation du fait de sa forte variance (un léger changement des données induit un arbre T_{max} très différent). Pour diminuer cette variance il est nécessaire d'ajouter une pénalisation au critère $L(R_j, x^d)$, pénalisation croissante en la complexité de l'arbre (le nombre de feuille). L'étape d'élagage consiste à extraire un sous-arbre du maximum arbre, réalisant un compromis entre simplicité et bon ajustement.

3.1.2 Élagage de l'arbre maximal : algorithme du pruning

De la suite de l'arbre maximal T_{max} précédemment obtenue par l'algorithme du CART, l'étape de *pruning* consiste à réduire la taille de l'arbre de décision en supprimant les sections de l'arbre qui ne sont pas critiques. En réduisant la complexité, l'élagage (*pruning*) améliore la précision prédictive et atténue le sur-apprentissage. une manière standard de procéder au *pruning* consiste à utiliser une approche pénalisée pour sélectionner le sous-arbre approprié (Breiman et al. (1984),Gey and Nédélec (2005)). Pour se faire, on extrait une sous-suite par minimisation, pour $\alpha \geq 0$, du critère pénalisé

$$Crit_\alpha(T) = \sum_{i=1}^{n} \ell^{\mathscr{R}^T}(Y_i, g(X_i)) + \alpha n_T \tag{II.32}$$

Avec T un sous-arbre de l'arbre maximal associé à un ensemble de règles $\mathscr{R}^T = (R_1^T, \dots, R_{n_T}^T)$ de cardinalité n_T.

Ainsi, les arbres avec un grand nombre de feuilles (c'est-à-dire de règles) sont pénalisés par rapport aux plus petits. Pour déterminer cet arbre $\hat{T}(\alpha)$, il n'est pas nécessaire de calculer tous les sous-arbres à partir de l'arbre maximal (T_{max}). Il suffit de déterminer, pour tout $K \geq 0$, le sous-arbre T_K qui minimise le critère (II.32) parmi tous les sous-arbres T avec $n_T = K$, puis de choisir l'arbre T_K qui minimise le critère par rapport à K. Breiman et al. (1984) montrent que ces T_K sont faciles à déterminer, puisque T_K est obtenu en enlevant une feuille à T_{K+1}.

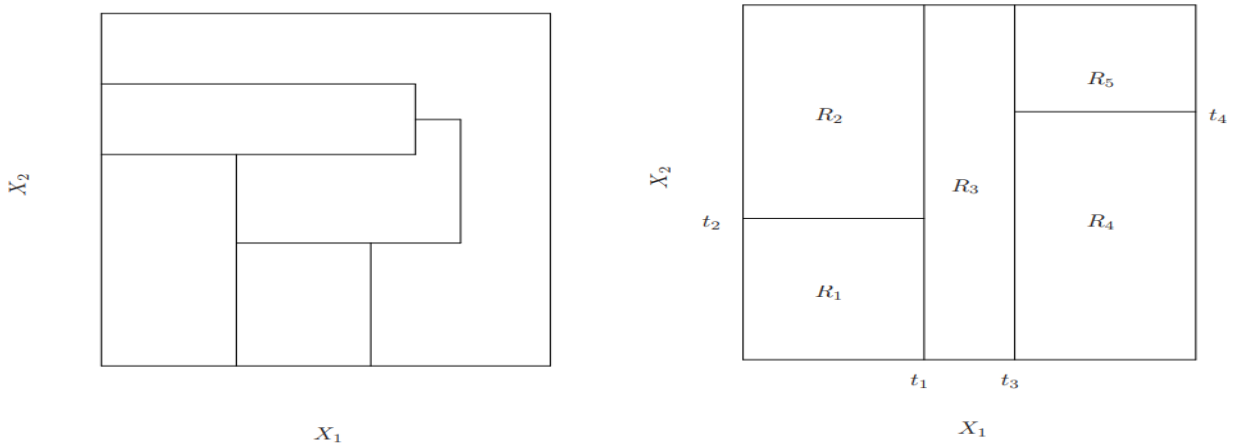
La constante de pénalisation α est choisie à l'aide d'un échantillon test ou à l'aide d'une validation croisée (voir sous-section 2.3).

Soit $\hat{\alpha}$ la constante de pénalisation calibrée à l'aide de l'approche de validation croisée

k-fold, l'estimateur final est donc :

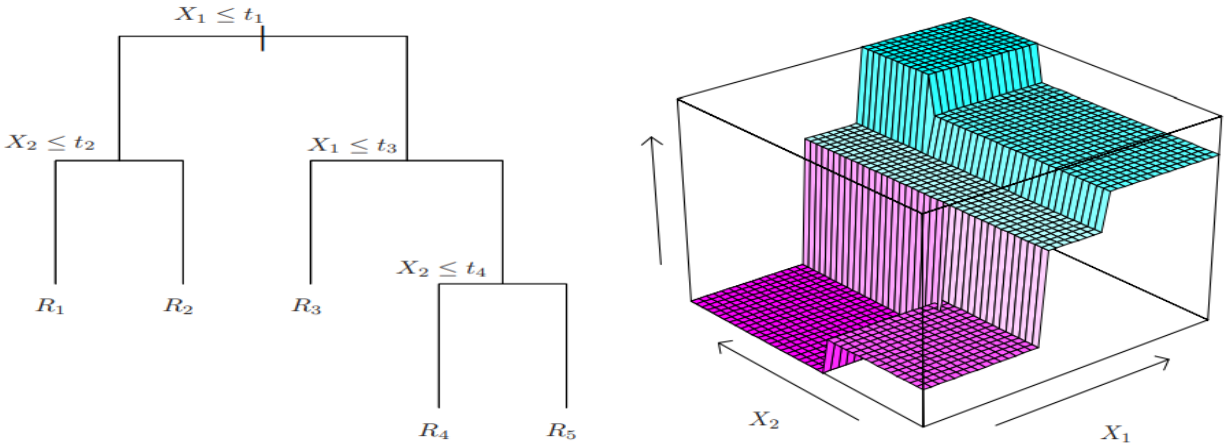
$$\hat{\alpha} = g^{\hat{T}(\hat{\alpha})}(x) \tag{II.33}$$

La figure II.16 ci-dessous illustre l'approche CART dans le cas de deux variables explicatives (X_1, X_2) et de cinq régions $R_1, \ldots, R_5$.



(a) Une partition d'espace de caractéristiques bidimensionnel qui ne peut pas résulter de l'algorithme *recursive binary splitting recursive binary splitting récursif*.

(b) Partition d'un espace de prédicteurs bidimensionnel par l'algorithme *recursive binary splitting*.



(c) Arbre de régression qui correspond à la partition obtenue en figure (b)

(d) Surface de prédiction qui correspond à l'arbre obtenu en figure (c)

Figure II.16 – Illustration de l'approche dans le cas de deux prédicteurs X1, X2 et cinq régions $R_1, \ldots, R_5$.

3.2 Arbres de régression Pareto Généralisée

Dans cette sous-section, on suppose que notre variable extrême Y est une variable aléatoire positive, à valeurs réelles et à queue lourde. Nous supposons également que sa distribution conditionnelle étant donné X satisfait les conditions suivantes :

$$\lim_{t \to \infty} \frac{\bar{F}(ty \mid \mathbf{x})}{\bar{F}(y \mid \mathbf{x})} = y^{-1/\xi_0(\mathbf{x})}, \forall y > 0 \tag{II.34}$$

Où $\bar{F}(y \mid \mathbf{x}) = \mathbb{P}(Y \geq y \mid \mathbf{X} = \mathbf{x})$, et

$$\lim_{u(\mathbf{x})\to\infty} \sup_{z>0} \left| \bar{F}_{u(\mathbf{x})}(z \mid \mathbf{x}) - \bar{H}_{\sigma_{0u(\mathbf{x})}(\mathbf{x}),\xi_0(\mathbf{x})}(z) \right| = 0. \tag{II.35}$$

On note $\sigma(x) = \sigma_{u(x)}(x)$ pour simplifier la notation.

L'idée de l'Arbre de régression Pareto Généralisée (GP regression tree) développé par Farkas et al. (2021a) est alors d'appliquer la procédure de l'arbre de régression présenté précédemment aux observations $(Y_i - u(X_i), X_i)$ pour lesquelles $Y_i \geq u(X_i)$, en utilisant la *log-vraisemblance* Pareto Généralisée en tant que critère de division (*split criterion*) , c'est-à-dire :

$$\ell(y, g(x)) = -\log(\sigma(x)) - \left(\frac{1}{\xi(x)} + 1\right) \log\left(1 + \frac{y\xi(x)}{\sigma(x)}\right) \tag{II.36}$$

où $g(x) = (\sigma(x), \xi(x))$. La fonction $u(x)$ doit être prise pour que l'ajustement de la distribution GP semble approprié pour toutes les valeurs des variables explicatives considérées.

Une possibilité serait d'adapter l'algorithme CART pour sélectionner, à chaque étape, un choix de seuil qui pourrait être différent dans chaque feuille. Cependant, cela complexifie considérablement la technique.

Au final, les feuilles de l'arbre identifient des classes, chacune correspondant à des comportements de queue différents (c'est-à-dire avec des valeurs différentes de $g(x) = (\sigma(x), \xi(x))$, la fonction g étant constante sur chaque feuille.

Farkas et al. (2021a) ont étudié la consistance d'un arbre ajusté $T(u)$, un sous-arbre de l'arbre maximal $T_{\max}(u)$, avec K feuilles $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$,. Ils ont comparé cet arbre ajusté à $T^*(u \mid T)$, qui est l'arbre basé sur la même subdivision, mais où, dans chaque feuille ℓ, le paramètre est $\theta^*_\ell(u)$ (au lieu de $\widehat{\theta}_\ell(u)$ dans $T(u)$).

Soit $\|(a,b)\|_\infty = \max(|a|,|b|)$, et pour deux arbres T et S,

$$\|T - S\|_2 = \left(\int \|T(\mathbf{x}) - S(\mathbf{x})\|^2_\infty \mathrm{dP}(\mathbf{x}) \right)^{1/2} \tag{II.37}$$

Le théorème suivant montre la consistance de l'algorithme en montrant que $\|T(u) - T^*(u \mid T)\|_2$ est bornée.

Théorème 5 (Consistance de l'algorithme). *Sous certaines hypothèses, notamment sur l'espace des paramètres, sur le seuil u et des conditions de régularité :*

$$\begin{aligned}
&\mathbb{P}\left(\sup_{u_{\min} \leq u \leq u_{\max}} \|T(u) - T^*(u \mid T)\|_2^2 \geq t \right) \\
&\quad \leq 2\left(\exp\left(-\frac{\mathscr{C}_1 k_n t}{K\beta^2 (\log k_n)^2} \right) + \exp\left(-\frac{\mathscr{C}_2 k_n t^{1/2}}{K^{1/2}\beta \log k_n} \right) \right) + \frac{\mathscr{C}_3 K}{k_n^{5/2} t^{3/2}},
\end{aligned}$$

où $\mathscr{C}_1, \mathscr{C}_2$ et $\mathscr{C}_3$ sont des constantes positives.

Sachant que la log-vraisemblance $L_n(T_K, u)$ associé à un arbre $T_K(u)$ avec K feuilles $\left(\mathcal{T}_\ell^K\right)_{\ell=1,\dots,K}$ avec les paramètres $\widehat{\theta}^K(u) = \left(\widehat{\theta}_\ell^K(u)\right)_{\ell=1,\dots,K}$, est définie comme suite :

$$L_n(T_K, u) = \sum_{\ell=1}^{K} L_n^\ell\left(\widehat{\theta}_\ell^K, u\right). \tag{II.38}$$

On a :

$$L(T_K, u) = \mathbb{E}\left[L_n(T_K, u)\right]. \tag{II.39}$$

Ainsi, pour deux arbres T et $S, \Delta L_n(T,S) = L_n(T,u) - L_n(S,u)$ et de manière similaire, $\Delta L(T,S) = L(T,u) - L(S,u)$.

Le théorème 5 suivant montre que la méthodologie d'élagage sélectionne un arbre $\widehat{T}(u)$ qui atteint approximativement le même taux que $T_{K_0}(u)$, même si $K_0(u)$ est inconnu , à condition que la constante de pénalité λ appartienne à un intervalle raisonnable.

Théorème 6 (Consistance de l'étape d'élagage). *Soit* $\mathfrak{D} = \inf_u \inf_{K<K_0(u)} \Delta L\left(T^*(u), T_K^*(u)\right)$ *et supposons qu'il existe une constante* $c_2 > 0$ *tel que la constante de pénalisation* λ *vérifie*

$$c_2\left\{\log k_n\right\}^{1/2} k_n^{-1/2} \leq \lambda \leq \left(\mathfrak{D} - 2c_2\left\{\log(k_n)\right\}^{1/2} k_n^{-1/2}\right) k_n^{-1},$$

pour tout $u \in [u_{\min}, u_{\max}]$,

$$\mathbb{E}\left[\left\|\widehat{T}(u) - T^*(u)\right\|_2^2\right] \leq \frac{\mathscr{C}_5 K_0(u)\left(\log k_n\right)^2}{k_n},$$

Où $\mathscr{C}_5$ *est une constante dépendant de* $T^*(u)$.

4 Application aux données réelles et simulées

Dans cette section, nous allons explorer la performance du modèle de régression pareto généralisée sur les données simulées par rapport au modèle additif généralisé (GAM) pour la prise en compte du caractère extrêmes de la variables d'intérêt. A la fin de cette section, nous appliquerons le modèle aux données réelles pour une illustration concrète de leur efficacité et de leur pertinence dans un contexte de données extrêmes.

4.1 Régression pareto généralisée sur données simulées

Cette partie est consacrée à l'illustration de la méthodologie de *l'arbre de régression Pareto Généralisée (GP)*. Nous analyserons les performances de cette approche sur des données simulées que nous comparerons avec celle de l'approche GAM Pareto Généralisée dont l'approche a été développée par Chavez-Demoulin et al. (2016) (sous-section 4.2 de l'annexe A).

Pour cela, nous considérons deux variables aléatoires ($d = 2$, où $\mathscr{X} = [0;1] \times [0;1]$), X_1 et X_2 indépendantes et uniformément réparties sur $[0;1]$. La variable réponse Y , conditionnel-

lement à $X = x$, est distribuée selon une distribution de Burr de paramètres $(\sigma, \mu_0(x))$ dont la fonction de survie est donnée par :

$$\overline{F}(y|x) = \frac{1}{1 + \left(\frac{y}{\sigma}\right)^{1/\mu_0(x)}} \tag{II.40}$$

avec $\sigma > 0$ et $\mu_0(x)$ pour tout x. On note que $\overline{F}(.|x)$ satisfait la propriété

Dans la suite, nous considérons deux cas :

— (i) $\mu_0(x)$ est comme fonction *step-wise* sur $\mathcal{X}$ et

— (ii) μ_0 comme une fonction continue (*smooth function*).

Dans les deux cas, on fixe le paramètre d'échelle σ égal à 1.

(i) fonction step-wise : Dans ce cas, la fonction μ_0 s'exprime comme

$$\mu_0(x_1, x_2) = \begin{cases} 0.5 & \text{si } 0 \leq x_1 < 0.5 \text{ et } 0 \leq x_2 < 0.5, \\ 1 & \text{si } 0 \leq x_1 < 0.5 \text{ et } 0.5 \leq x_2 \leq 1, \\ 1.5 & \text{si } 0.5 \leq x_1 \leq 1 \text{ et } 0 \leq x_2 < 0.5, \\ 2 & \text{si } 0.5 \leq x_1 \leq 1 \text{ et } 0.5 \leq x_2 \leq 1, \end{cases} \tag{II.41}$$

Le graphique suivant présente la représentation graphique de la variable Y catégorisée (target_cut) en fonction de X_1 et X_2 sur un échantillon de taille 5000.

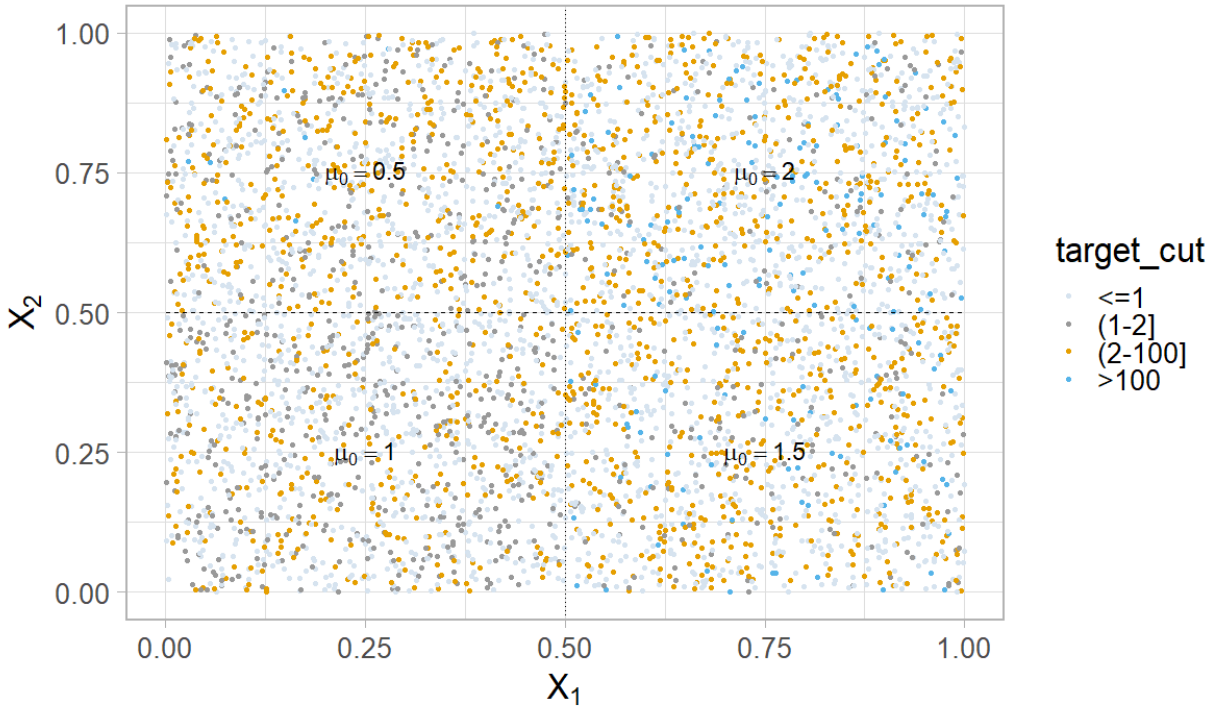


Figure II.17 – Partitions de l'espace $\mathcal{X}$ avec leurs valeurs d'index de queue μ_0 associées

(ii) fonction continue : Dans ce cas, la fonction μ_0 s'exprime comme, pour $x_1 \in [0,1]$ et $x_2 \in [0,1]$

$$\mu_0(x_1, x_2) = 1 + \frac{\tanh(10(x_1 + x_2 - \frac{1}{4}))}{4} + \frac{\tanh(10(x_1 + x_2 - \frac{3}{4}))}{4} \tag{II.42}$$

Le graphique suivant présente la représentation graphique de la variable Y catégorisée (*target_cut*) en fonction de X1, X2 et μ_0 sur un échantillon de taille 5000.

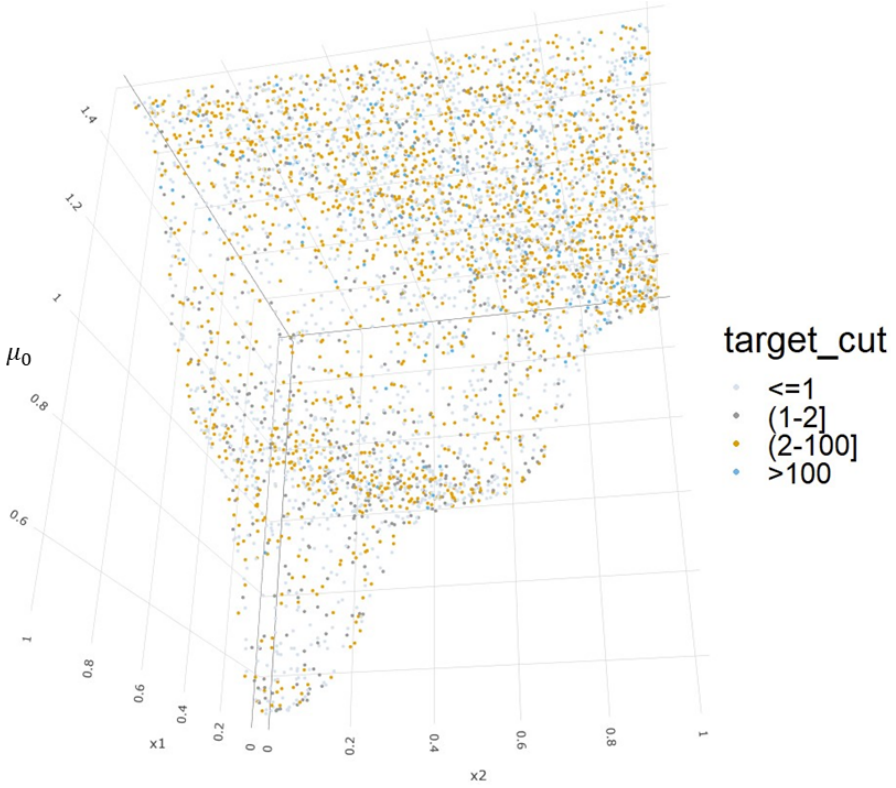


Figure II.18 – Surface $\mathscr{X}$ avec leurs valeurs d'index de queue μ_0 en continue

Nous avons simulons 1000 répétitions pour différentes tailles d'échantillon (n = 1000, 2500, 5000, 10 000 et 20000) selon le cadre décrit précédemment. Pour chaque échantillon, on considère les dépassements au-dessus du quantile empirique 90%, qui correspond à $k_n =$ 100, 250, 500, 1000 et 2000.

Pour chaque échantillon simulé, nous exécutons la procédure d'arbre de régression (CART) Pareto Généralisée, et la méthode basée sur le modèle additif généralisé (GAM) proposé par Chavez-Demoulin (2005). Ensuite, pour chaque estimateur nous avons calculé $\int_O^1 \int_O^1 (\mu_O(x) - \mu_0(x))^2 \, dx_1 \, dx_2$. L'erreur quadratique moyenne empirique est alors obtenue par en faisant la moyenne de ces erreurs sur les 1000 répétitions. Les résultats sont présentés dans le tableau suivant :

	k_n	100	250	500	1000	2000
(i)	GP CART	0.310	0.135	0.117	0.090	0.070
	GAM	0.326	0.210	0.122	0.091	0.068
(ii)	GP CART	0.210	0.115	0.107	0.055	0.035
	GAM	0.229	0.132	0.063	0.031	0.012

Table II.5 – Erreurs quadratiques moyennes empiriques pour la procédure d'arbre de régression GP et le modèle GAM pour différentes tailles d'échantillon dans le cas (i) et (ii).

Notons que l'approche GAM n'est pas conçue pour capturer des fonctions non lisses comme dans la fonction définir dans notre cas. Néanmoins, nous voyons que cette technique parvient à s'ajuster relativement correctement même dans ce cas lorsque la taille de l'échantillon est grande. Pour $k_n = 1000$ et 2000, les résultats de l'approche GAM sont similaires voire légèrement meilleurs que la méthode de l'arbre de régression. D'autre part, nous observons que les arbres de régression conduisent à un meilleur ajustement pour les petites tailles d'échantillon, même dans le cas de la fonction continue où il n'est pas conçu pour prendre en compte la régularité de $\mu_0(x)$.

Nous observons que nos conclusions sont en accord avec celles obtenues par Farkas et al. (2021a), qui ont analysé les performances de la méthode de l'arbre de régression Pareto généralisée par rapport au modèle GAM en utilisant des données simulées. Cependant, leur approche diffère de la nôtre en termes de dimensionnalité des variables explicatives, car ils se sont concentrés sur une seule variable explicative $X_1 \in [0, 1]$.

La figure II.19 ci-dessous montre un exemple d'un arbre de régression obtenu lors des 1000 simulations sur un échantillon de taille 5000 dans le cas (i). le tableau II.6 présente, pour chaque feuille de l'arbre, l'estimation du paramètre de Burr μ_O et son intervalle de confiance à 95% associé.

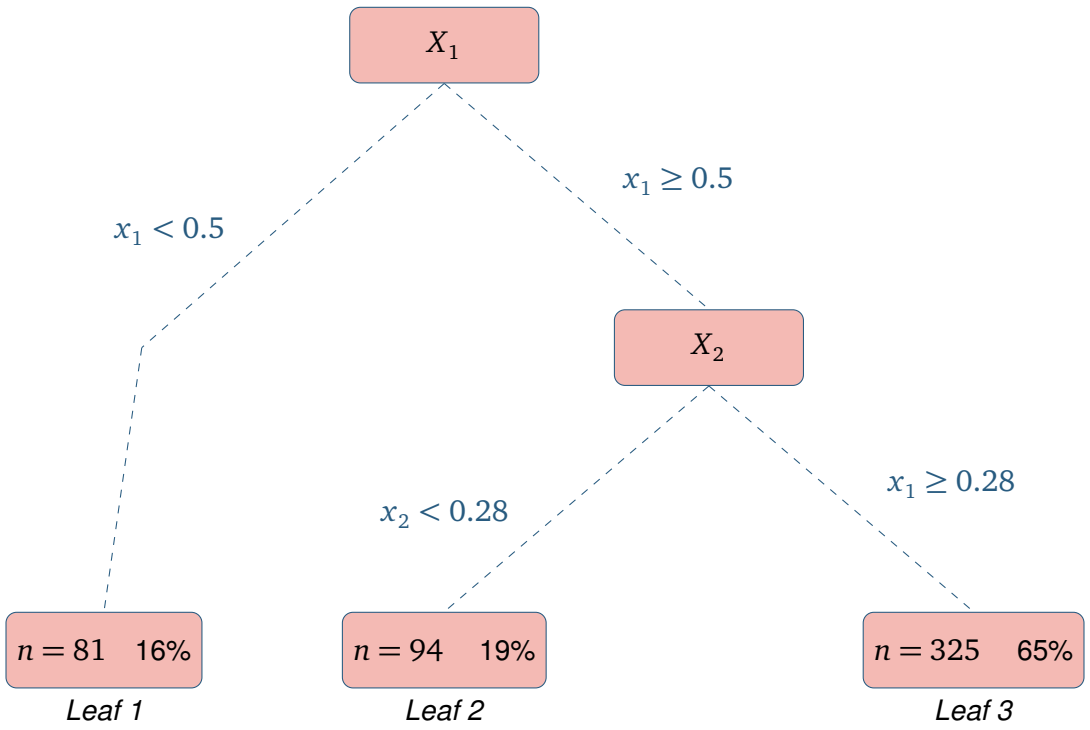


Figure II.19 – Exemple d'arbre de régression pareto généralisée pour un échantillon de taille 5000 et sur les observations dont la variable cible Y est supérieure au quantile 90%

Feuille	*Leaf 1*	*Leaf 2*	*Leaf 3*
$\hat{\mu}_0$	0.52	1.47	2.08
IC(95%)	[0.35 - 0.63]	[1.34 - 1.68]	[1.91 - 2.25]

Table II.6 – Estimation paramètre de Burr $\mu_0(x)$ dans chaque feuille

4.2 Application en assurance : sinistralité extrême du régime « beau temps »

Dans cette sous-section, nous illustrons la procédure de l'*arbre de régression pareto généralisée* (GP CART) sur la sinistralité du régime « beau temps » pour mieux comprendre le caractère extrême de cette sinistralité dont la modélisation revêt un intérêt important en assurance et en réassurance. La capacité de la procédure GP CART à concevoir des classes de sinistres plus homogènes en fonction de variables explicatives est une propriété intéressante pour une meilleure projection de la sinistralité extrême selon un horizon donné.

La sinistralité du régime « beau temps » est très volatile avec une variance empirique égale à 85.3e+09. Une brève analyse descriptive de cette sinistralité est présentée dans le tableau D.2 en annexe.

La figure II.20 montre la moyenne des sinistres supérieurs au quantile 95% de la sévérité au sein de chaque région de la France Métropolitaine. On observe une hétérogénéité avec grande différence entre les régions du Sud-ouest (**Occitanie et Nouvelle-Aquitaine**) et les autres régions.

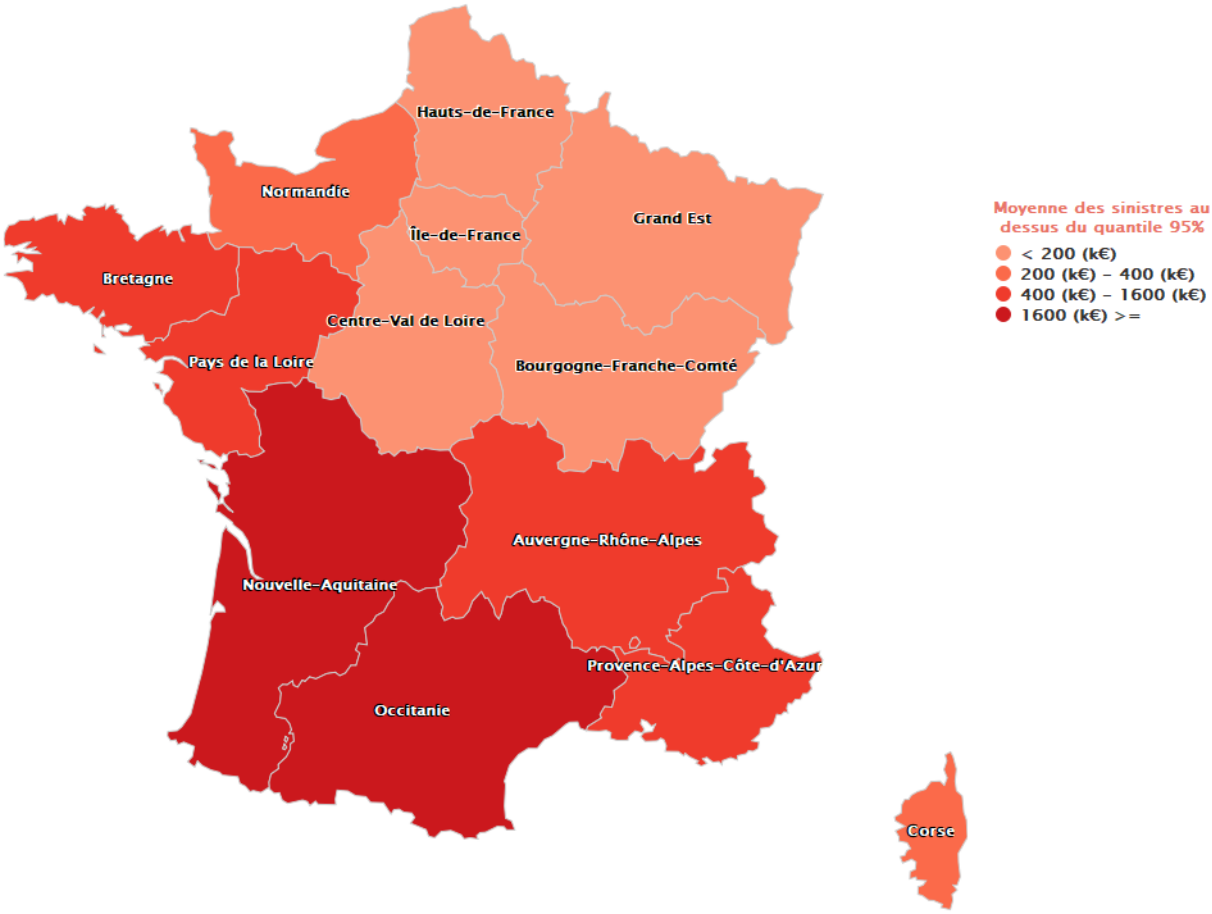


Figure II.20 – Moyenne des sinistres au dessus du seuil quantile 95% de la sinistralité pour chaque région (2016 - 2021)

La sous-section 1.4 de ce chapitre nous a permis de déterminer le seuil idéal $u = 130\,000€$, qui peut être perçu comme un compromis entre le biais et la variance. On obtient 1622 observations extrêmes, c'est-à-dire dont le coût est supérieur à 130 000€.

Dans le but de comprendre l'hétérogénéité du caractère extrême de la sinistralité du régime « beau temps », nous avons appliqué l'arbre de régression pareto généralisée à la base de données correspondant aux observations extraits de la base de données originale pour lesquels le sinistre est supérieur à $u = 130\,000$€. Dans cette base de donnée, on retrouve cinq (5) variables explicatives qualitatives et douze (12) variables quantitatives. Les variables qualitatives sont : **région**, **saison**, **mois**, **vacance** (1 si le jour est un jour de vacance scolaire 0 sinon) et **week-end** (1 si le jour est soit samedi ou dimanche 0 sinon). La répartition des catégories pour chacune de ces variables qualitatives se trouve en annexe (tableau D.4). Concernant les variables quantitatives nous avons retenu la **température** journalière en degrés celsius (°C), **la précipitation** en millimètre (mm) et la **vitesse de vent** en mètre par seconde (m/s). Ces trois variables sont associées aux risques couverts par le régime. Aussi, pour chacune de ces variables nous avons retenu la moyenne, le minimum et le maximum en moyenne mobile (glissante) sur 7 jours. Une brève analyse descriptive de ces variables quantitatives est présentée dans le tableau D.3 en annexe.

La figure II.21 montre l'arbre obtenu à partir de la procédure de régression pareto généralisée (GP CART).

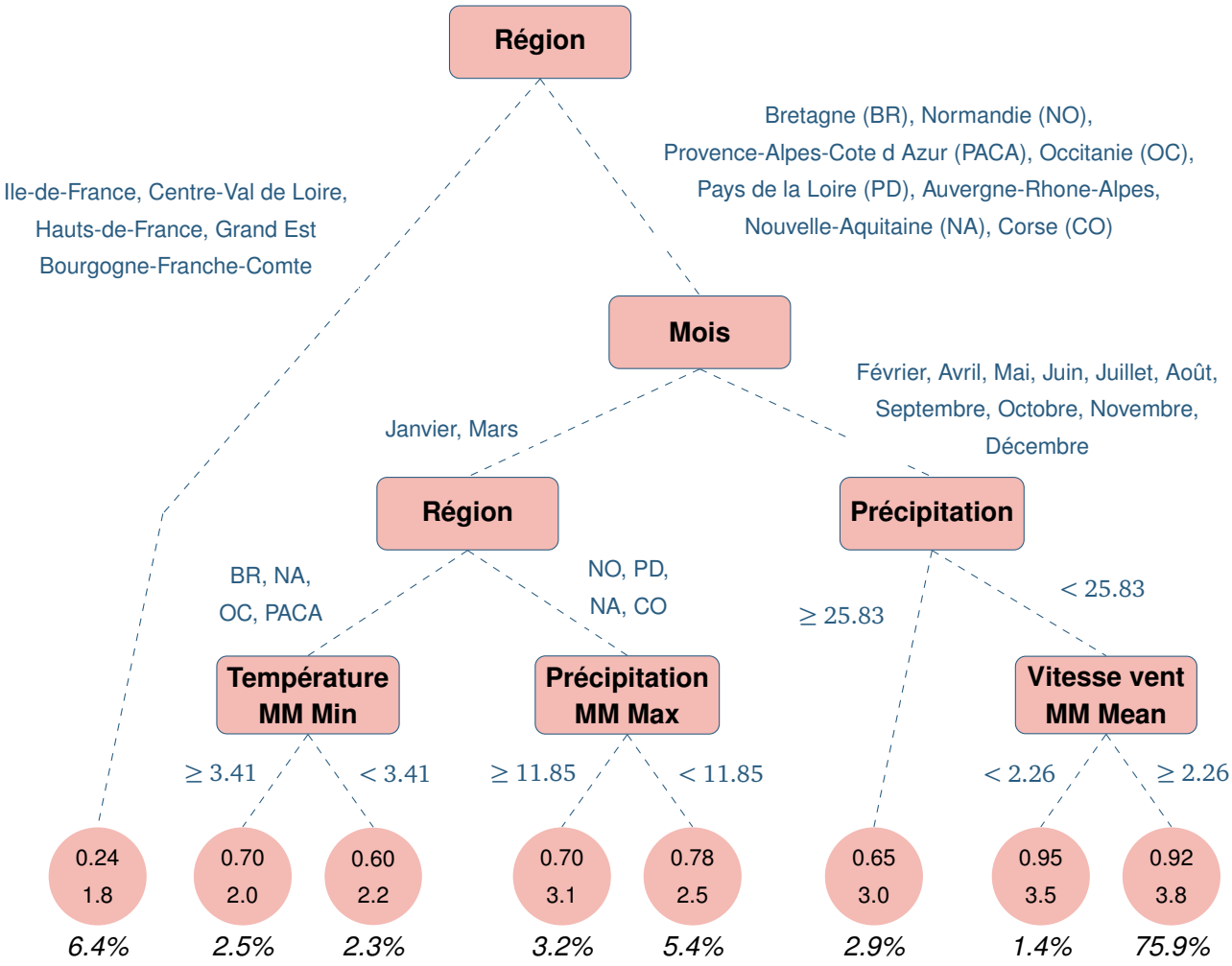


Figure II.21 – Arbre de régression pareto généralisée (GP CART). Pour chaque feuille, la valeur du paramètre de forme ξ (première ligne) et le paramètre d'échelle σ à 10^{-5} (deuxième ligne) sont donnés. Le pourcentage d'observations affectées à chaque feuille est mentionné.

L'arbre est composé (*leafs*) de 8 feuilles avec quatre (4) *splits* selon seulement 6 variables : la région, le mois, la précipitation journalière, le minimum de la température journalière sur un glissement de 7 jours (*MM Min*), le maximum de la précipitation journalière sur un glissement de 7 jours (*MM Max*) et la moyenne de la vitesse du vent en glissement de 7 jours (*MM Mean*). La présence des variables *région* et *mois* dans l'arbre GP CART semble raisonnable dans la mesure où l'activité de camping diffère énormément entre les régions et varie selon le mois. Ces deux variables expliques fortement les sorties de l'arbre (voir figure II.22 pour l'importance des variables pour l'arbre GP CART). Parmi les variables climatiques, on retrouve la variable *t_mean_rol_min*, c'est-à-dire la température minimale journalière sur une période glissante de 7 jours, ainsi que la variable *vent_rol_min* (vitesse minimale du vent journalier sur une fenêtre glissante de 7 jours) qui présentent une importance plus élevée que les autres variables climatiques.

Dans chaque feuille, sont donnés les paramètres de forme et d'échelle. Le pire des cas correspond aux deux feuilles situés à l'extrême droite, avec des paramètres de forme qui sont proches de 1 (0.95 pour la 7ème feuille et 0.92 pour la 8ème feuille) et représentent la quasi majorité des sinistres extrêmes du régime (77.3%). Ces deux feuilles correspondent aux sinistres pour lesquels 8 régions de la France métropolitaine (les régions du sud et les régions du nord-est) et 10 mois de l'année (sauf le mois de janvier et mars) sont concernées. Le cas le moins sévère correspond à la première feuille en partant de la gauche, avec un paramètre de forme égal à 0,24 et contenant 6.4% des sinistres.

Soit Y une distribution pareto généralisée avec un paramètre d'échelle σ et un paramètre de forme ξ, la médiane théorique $Med(Y)$ et la moyenne théorique $E(Y)$ sont données par :

$$\begin{cases} Med(Y) & = \frac{\sigma(2^{\xi}-1)}{\xi}, \\ E(Y) & = \frac{\sigma}{1-\xi}, \quad si \quad \xi < 1 \quad et \quad \infty \quad si \quad \xi \geq 1 \end{cases} \tag{II.43}$$

Le tableau II.7 présente pour chaque feuille la médiane et la moyenne empirique des sinistres extrêmes ainsi que la médiane et la moyenne théorique de la distribution pareto généralisée (GP) correspondante.

Leaf (feuille)	Paramètres		Médiane €		Moyenne €	
	ξ	σ	Empirique	Théorique	Empirique	Théorique
1	0.24	1.8	144 674	135 744	156 206	236 842
2	0.70	2.0	186 628	178 430	202 233	666 667
3	0.60	2.2	192 363	189 096	246 057	550 000
4	0.70	3.1	288 169	276 566	370 032	1 033 333
5	0.78	2.5	278 586	229850	338 650	1 136 364
6	0.65	3.0	281 304	262 693	414 860	857 143
7	0.95	3.5	357 125	343 322	483 201	7 000 000
8	0.92	3.8	379 426	368 482	471 044	4 750 000

Table II.7 – Médiane et moyenne empirique et médiane et moyenne théorique pour chaque feuille.

Tout d'abord, pour chaque feuille, la médiane est bien plus petite que la moyenne, ce qui suggère que nous avons effectivement affaire à des événements extrêmes. Ensuite, les médianes empiriques et théoriques sont du même ordre pour chaque feuille tandis que les moyennes empiriques et théoriques (à sa sortie) ne sont comparables que pour la feuille 1 dont le paramètre de forme est significativement différent de 1.

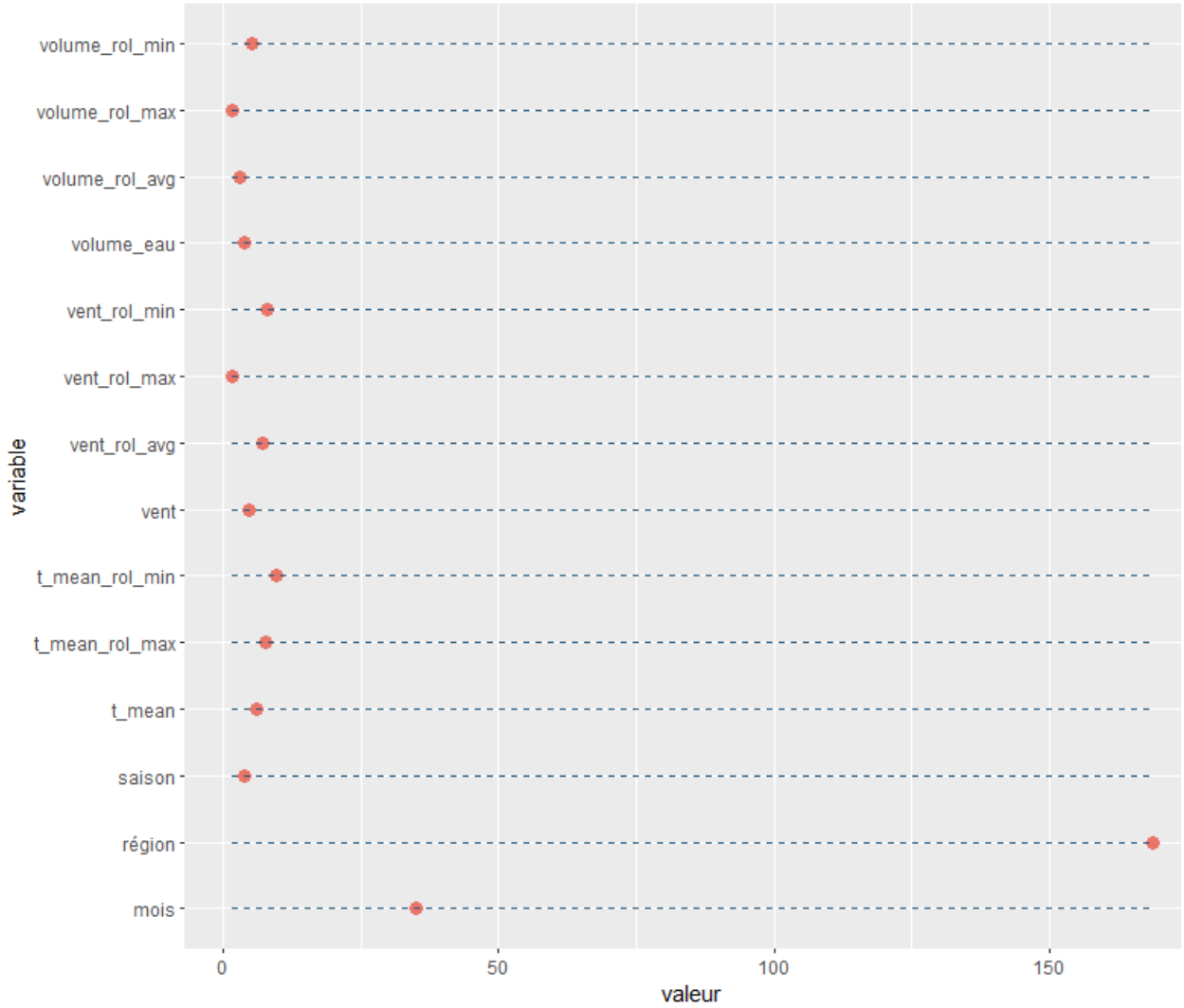


Figure II.22 – Importance des variables pour la régression pareto généralisée

Synthèse partielle

Dans cette première partie, nous avons abordé la conception d'un régime fictif baptisé « **beau temps** », qui intervient en indemnisant une victime dès lors qu'au moins l'un des indicateurs du régime liés à la température, au vent ou aux précipitations dépasse un seuil. La mise en place de la grille tarifaire associée à ce régime a requis l'utilisation de données relatives aux séjours en camping en France métropolitaine, issues des enquêtes menées par l'INSEE, ainsi que des données fournies par Météo France. Cette grille tarifaire a été élaborée en fonction du mois et de la région.

Par la suite, lors de l'analyse de la sinistralité découlant de la mise en oeuvre du régime, nous avons utilisé des outils de la théorie des valeurs extrêmes et des statistiques

pour mettre en évidence la nature extrême de la sinistralité du régime. Cette analyse nous a également permis de segmenter les données en deux catégories : les données extrêmes et celles qui ne le sont pas, en utilisant un seuil fixé à 130 000 €. Ce seuil a été obtenu à l'aide des méthodes de détermination du seuil des extrêmes.

Nous avons ensuite exploré le modèle de l'arbre de régression généralisée de Pareto, qui combine la théorie des valeurs extrêmes et l'apprentissage statistique pour une meilleure prise en compte des valeurs extrêmes.

Enfin, en utilisant des données simulées et en analysant la sinistralité extrême du régime, nous avons évalué les performances pratiques de l'arbre de régression pareto généralisée (GP CART).

PARTIE B

Etude actuarielle de l'impact du changement climatique sur le régime « beau temps »

« Seulement ceux qui prendront le risque d'aller trop loin découvriront jusqu'où on peut aller. »

Thomas Stearns Eliot.

« Il est extrêmement probable que l'influence de l'homme a été la cause principale du réchauffement observé depuis la moitié du XXe siècle. Les preuves s'en sont multipliées grâce à l'amélioration et à la prolifération des observations, à une meilleure compréhension des réactions du système climatique et à l'amélioration des modèles du climat. Le réchauffement du système climatique est sans équivoque et, depuis 1950, on observe dans ce système de nombreux changements sans précédent à une échelle temporelle allant de quelques décennies à plusieurs millénaires. »

5ème rapport du GIEC (GIEC, 2014)

Chapitre

III

Projection du régime à l'horizon 2100 via les scénarios du GIEC et étude de sensibilité

La partie précédente a été consacrée à la mise en place du régime « **beau temps** », dont l'objectif est de fournir une indemnisation en cas de conditions météorologiques défavorables lors d'un séjour en camping. Dans cette première partie, nous nous sommes particulièrement concentrés sur le ratio de rentabilité S/P, qui garantit la fiabilité du régime. Dans ce nouveau chapitre, nous allons étudier l'impact du changement climatique sur ce régime en projetant le ratio S/P selon deux scénarios du GIEC : le scénario le plus pessimiste, *RCP 2.6*, et le scénario le plus optimiste, *RCP 8.5*. Pour ce faire, nous commencerons par présenter les données qui serviront à la projection jusqu'à l'horizon 2100. Ce choix de projection à l'horizon 2100 s'explique par notre souhait d'avoir une perspective à moyen et long terme sur le ratio S/P. Tout au long de ce chapitre, nous détaillerons les approches de projection retenues et présenterons les hypothèses que nous avons adoptées. En outre, nous examinerons et interpréterons les résultats obtenus. Enfin, la fin de ce chapitre sera consacrée à une étude de sensibilité.

1 Contexte de la projection du régime « beau temps »

Le but de la projection du régime « **beau temps** » est d'estimer l'évolution du compte de résultat et de certains indicateurs (ratio S/P, réserve) de 2023 à 2100 selon les deux scénarios du GIEC (RCP 8.5 et 2.6). Nous avons adopté un compte de résultat simplifié qui contient principalement trois (3) postes. Au niveau des produits (côté gauche du compte de résultat), nous avons la prime pure totale (P) et les frais globaux (FG) qui représentent 10% de la prime pure totale. Au niveau des charges (côté droit du compte de résultat), on retrouve seulement les sinistre (S). Pour rappel, la prime pure totale pour une année spécifique sont calculées en utilisant une grille tarifaire qui dépend du mois du séjour en camping et de la région où celui-ci a lieu. Concernant les sinistres, ils sont issus du mécanisme d'indemnisation du régime « **beau temps** » consistant à indemniser 50€, lorsque au moins l'un des indices de température (Th,Tb), de précipitation (Ph) ou de vitesse de vent (Vh) dépasse le seuil défini pour ce indice.

En ce qui concerne les réserves, on suppose que le régime dispose d'une réserve initiale

R_0 de 55 500 059 € au début de la projection (2023), ce montant étant la somme des résultats sur la période historique (2016 à 2021). L'évolution des réserves est uniquement conditionnée par les résultats du compte de résultats, et pour une année de projection donnée n, elle est exprimée par la formule suivante :

$$R_n = R_{n-1} + \text{Résultat}(n) = R_0 + \sum_{i=2023}^{n} \text{Résultat}(i) \tag{III.1}$$

avec Résultat$(n) = P(n) + FG(n) - S(n)$, dont $P(n)$, $FG(n)$ et $S(n)$ représentent respectivement la prime pure totale, les frais globaux et le sinistre total projetés pour l'année n.

Le tableau III.1 ci-dessous présente l'historique des principaux postes du compte de résultats du régime, ainsi que les indicateurs de performance que nous avons retenus.

POSTE\ANNEE	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*
PRIME (P)	107 555 669 €	108 480 365 €	124015 726 €	165 779 975 €	129 560 238 €	200 883 706 €
SINISTRE (S)	83 040 921 €	108 415 082 €	120 578 573 €	180 966 456 €	130 158 651 €	173 436 368 €
FRAIS GLOBAUX (FG)	10 755 567 €	957 970 €	1 062 034 €	1 475 494 €	720 467 €	849 901 €
RESULTATS (P+FG - S)	35 270 314 €	1 023 253 €	4 499 187 €	-13 710 987 €	122 054 €	28 297 239 €
Indicateurs de performance technique						
RESERVES						***55 501 059 €***
RATIO S/P	77.2%	99.9%	97.2%	109.2%	100.5%	86.3%
RATIO S/(P+FG)	70.2%	99.1%	96.4%	108.2%	99.9%	86.0%

Table III.1 – Compte de résultats et indicateurs historiques du régime « **Beau temps** »

Dans le cadre de ce mémoire, nous avons adopté deux approches pour la projection du régime à l'horizon 2100 : ***une approche déterministe*** et ***une approche par modélisation***. Dans ces deux approches, il est impératif que nous formulions un certain nombre d'hypothèses réalistes. Les hypothèses concernent principalement l'évolution de la population assurée, la grille tarifaire, ainsi que diverses considérations telles que le mécanisme d'indemnisation, entre autres. Ces hypothèses sont détaillées dans la section 3 de ce chapitre.

La première approche, appelée *déterministe*, consiste simplement à appliquer les règles de fonctionnement du régime « **beau temps** » aux données de la base DRIAS en fonction des scénarios RCP 2.6 et RCP 8.5. Par exemple, en utilisant les données futures du RCP 2.6, nous recalculons les indices de température (Th,Tb), de précipitation (Ph) et de vitesse du vent (Vh). Si l'un de ces indices dépasse son seuil défini, le régime paiera une indemnité de 50 € aux assurés. La prime totale et le sinistre total projetés pour une année donnée sont alors déterminés en utilisant les hypothèses formulées concernant l'évolution de la population assurée et la grille tarifaire.

L'approche par la *modélisation*, en revanche, suppose que le processus qui régit le fonctionnement du régime (déclenchement des paiements, indemnité,...) est aléatoire. Soit S la distribution du paiement (soit 0 en cas de non déclenchement ou Y positive strictement).

Alors l'espérance mathématique de S s'écrit :

$$E(S) = P(S > 0)\,[P(Y \geq u)E(Y|Y \geq u) + (1 - P(Y \geq u))E(Y|Y < u)] \tag{III.2}$$

Au vu de cette espérance de la distribution de S, nous considérons que la distribution de S est une variable de mélange de même distribution que $\tau(x) * [\delta(x)Z_1 + (1-\delta(x))Z_2]$, avec :

- $\tau(x)$ et $\delta(x)$ sont des variables aléatoires de Bernoulli dépendant de X. $p_i(x) = P(\tau(x) = 1)$ et $m_i(x) = P(\delta(x) = 1)$ représentent respectivement la probabilité de déclenchement ($S \neq 0$) et la probabilité de dépassement du seuil des extrêmes u. Ces deux paramètres seront modélisés à l'aide des techniques de *machine learning* (ML) ;
- $Z_1 = Y|Y \geq u$, sera calibré à l'aide de *l'arbre de pareto generalisée* (GP CART) de la figure II.21 avec $u = 130\,000$€ ;
- $Z_2 = Y|Y < u$, sera modélisé à l'aide d'un GLM classique pour la sévérité. On considère $\tau \perp\!\!\!\perp (\delta, Z_1, Z_2)$ et $\delta \perp\!\!\!\perp (Z_1, Z_2)$.

Les sinistres sont ensuite estimés en calibrant chaque composante de ce mélange de modèles. Les primes, quant à elles, sont calculées en se basant sur les sinistres projetés et l'hypothèse relative à la grille tarifaire (plus de détails seront fournis dans la suite).

L'intérêt de cette approche est de s'affranchir de l'hypothèse sur l'évolution de la population assurée. Un deuxième avantage est de prendre en compte, dans la projection, l'aspect extrême de la sinistralité du régime mis en évidence dans le chapitre 2.

Les approches adoptées pour la projection du régime « **beau temps** » à l'horizon 2100 peuvent se résumer de la façon suivante (Figure III.1).

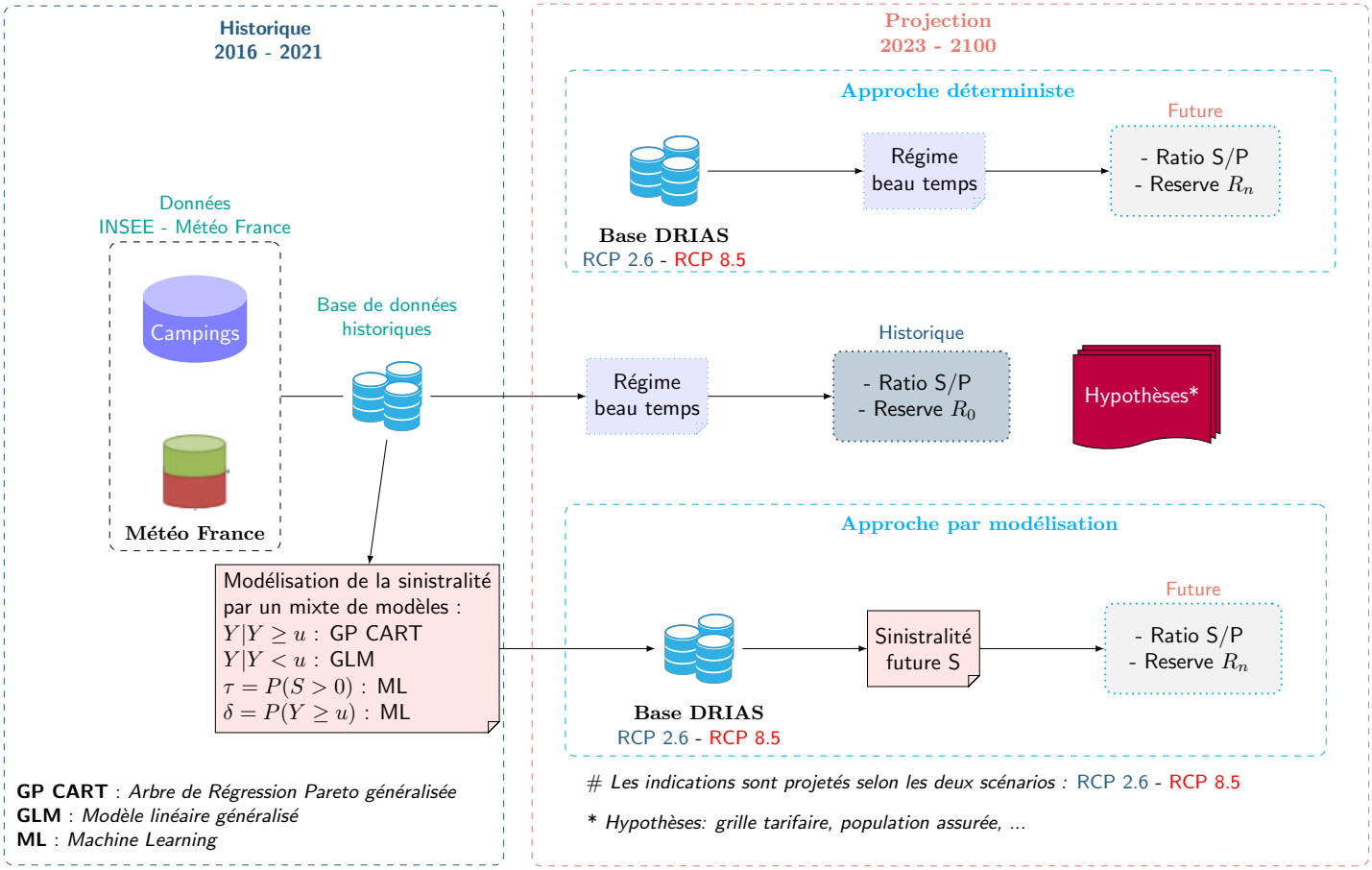


Figure III.1 – Diagramme résumant la stratégie de projection du régime « **beau temps** ».

2 Les données DRIAS et les scénarios du GIEC

Cette section est dédiée aux données DRIAS utilisées pour la projection du régime à l'horizon 2100, ainsi qu'aux scénarios du GIEC. Dans un premier temps, nous aborderons le 6ème rapport du GIEC. Ensuite, nous présenterons les perspectives des modèles climatiques, ainsi que la disponibilité des données DRIAS. Enfin, nous conclurons cette section en analysant la concordance entre les données historiques de Météo France (SYNOP) et les données DRIAS.

2.1 Contexte et le 6ème rapport du GIEC

Le 6ème rapport du Groupe d'Experts Intergouvernemental sur l'Évolution du Climat (GIEC), publié en 2022, offre une vue d'ensemble des connaissances actuelles en matière de climat. Fruit d'une analyse approfondie de milliers d'études mondiales, de nouvelles recherches et de simulations climatiques futures, ce rapport vise à informer le grand public et les décideurs. Sa conclusion majeure met en évidence la responsabilité de plus en plus établie de l'activité humaine dans le réchauffement climatique.

"le changement climatique est une menace pour le bien-être humain et la santé de la planète. Nous devons donc accélérer nos actions pour anticiper au maximum les risques que le changement climatique fait peser sur nos territoires."

6ème rapport du GIEC (GIEC, 2022)

Parmi les conclusions majeures de ce rapport, celles d'un réchauffement sans précédent provoqué par les activités humaines. L'intensité du réchauffement climatique est particulièrement forte : en 2022, les températures étaient supérieures d'1.2 degrés environ par rapport aux moyennes pré-industrielles. Concernant le rythme du réchauffement climatique, il est même supérieur aux projections précédentes : les températures augmentent de façon très rapide, trop rapide pour que les écosystèmes puissent s'y adapter. Les records de chaleur s'enchaînent et l'apparition des canicules est de plus en plus fréquente, particulièrement en Europe, en Asie et en Australie. Alors qu'en Europe les températures ont déjà augmenté de plus d'1°C, la vitesse de réchauffement a été plus rapide que pour le reste du globe et l'exposition des populations au risque caniculaire est grandissante. L'augmentation des températures atmosphériques influe également sur la présence des pollens allergéniques et des maladies vectorielles.

2.2 La modélisation du climat

Un modèle climatique vise ainsi à représenter, de manière formelle, le climat et son évolution. Basés sur les équations de la physique, de la chimie et de de la biogéochimie, ces modèles couplant atmosphère, océans et continents, constituent une réplique numérique du monde réel. Les projections climatiques issues de ces modèles permettent l'étude d'un

réchauffement climatique en cours.

Les modèles régionaux couvrent des zones plus restreintes. Ils permettent une résolution spatiale très fine (10 à 20 km) prenant en compte l'atmosphère et la végétation. Les caractéristiques océaniques n'y sont pas représentées mais extraites des modèles globaux les entourant. Les limites aux bords des RCM (Regional Climate Model) sont forcées par les modèles globaux.

Le modèle CNRM-ALADIN63 est un modèle de climat régional, c'est-à-dire qu'il est utilisé pour simuler le climat à une échelle régionale plutôt qu'à une échelle globale. Il est basé sur la dynamique des fluides et les lois de la thermodynamique pour simuler les interactions entre l'atmosphère, l'océan, la surface terrestre et les êtres vivants. Il est particulièrement adapté pour la modélisation des événements météorologiques extrêmes tels que les tempêtes, les précipitations intenses et les températures fortes.

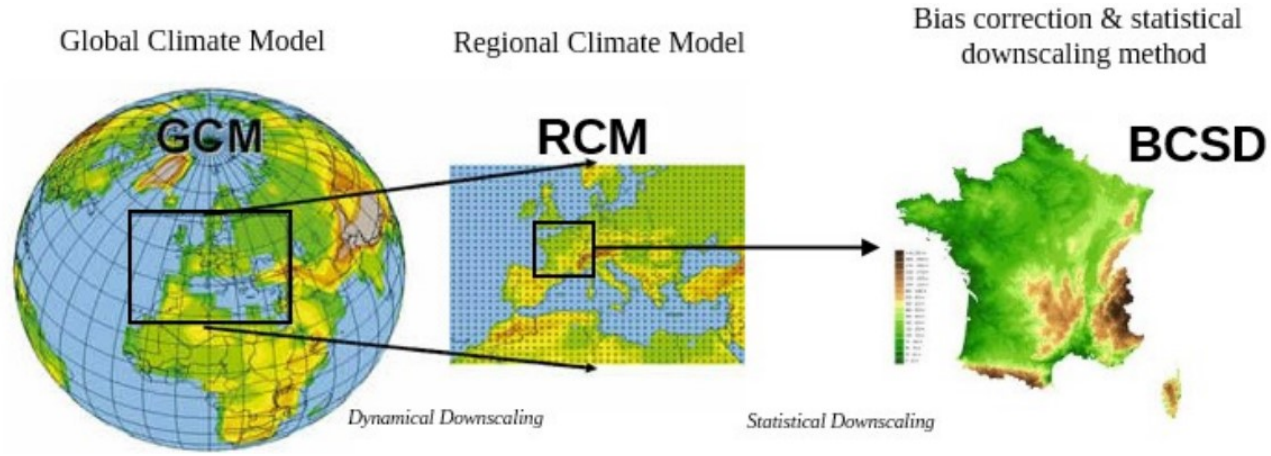


Figure III.2 – Architecture des modèles climatiques (DRIAS, 2020)

DRIAS utilise 4 profils représentatifs d'évolutions des concentrations de gaz à effet de serre RCP (Representative Concentration Pathways) publiés par le GIEC (Groupe d'experts intergouvernemental sur l'évolution du climat) dans son 5^{me} rapport.

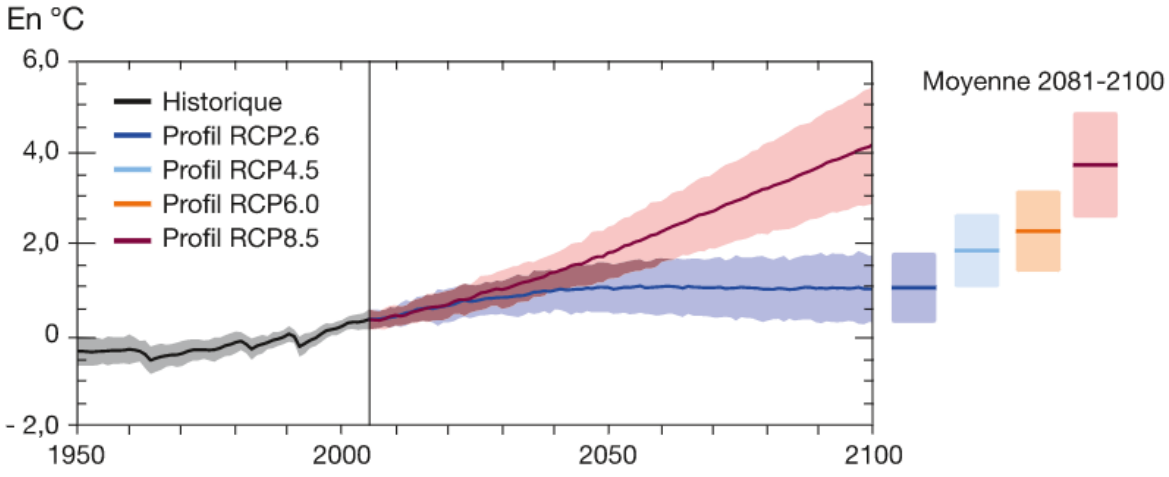


Figure III.3 – Projection de la variation de température moyenne mondiale suivant différents scénarios (Source : GIEC)

Le forçage radiatif est la différence entre l'énergie radiative reçue (rayonnement solaire) et celle émise par le système atmosphère/Terre dû à des facteurs d'évolution du climat tels que la concentration des gaz à effet de serre. Un forçage positif tend à réchauffer la surface. II se mesure en $W/m2$. Le RCP 8.5 , le plus pessimiste, correspond à un forçage radiatif de l'ordre de 8, 5 W/m2, ce qui se traduit par une élévation de la température moyenne de +4°C environ en 2100. Le RCP 2.6, le plus favorable, permettrait quant à lui de respecter l'Accord de Paris sur le Climat en maintenant le réchauffement bien en dessous de +2°C.

Dans son 6eme et dernier rapport d'évaluation en date, le GIEC présente les évolutions possibles du climat selon 5 nouveaux scénarios : les SSP (Shared Socioeconomic Pathways). On les note SSPx - y avec « x " le scénario socio-économique utilisé et " y " la valeur du forçage radiatif. Ces nouveaux scénarios sont plus représentatifs des trajectoires socio-économiques potentielles et tiennent compte notamment d'hypothèses sur le développement humain, l'éducation, la santé, la croissance économique et démographique, la sécurité et les inégalités sociales.

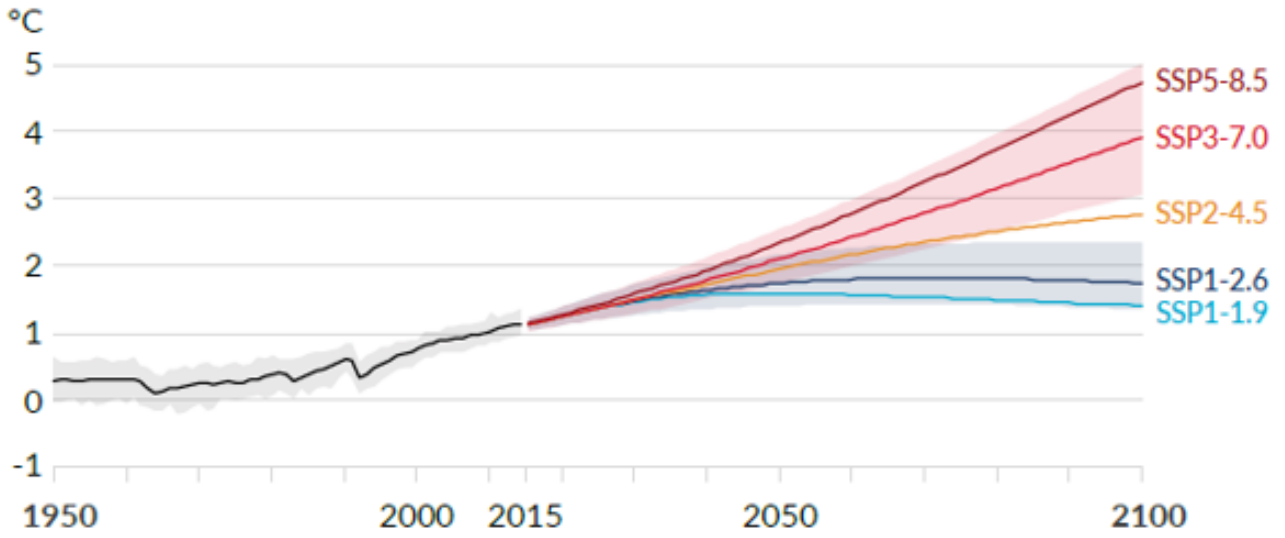


Figure III.4 – Augmentation de la température de surface dans chacun des scénarios par rapport aux niveau de 1850-1900 (Source : GIEC)

A titre d'exemple, le SSP5 prévoit des investissements considérables en matière d'éducation et de santé et une croissance économique rapide mais à forte intensité énergétique non renouvelable.

2.3 Les données disponibles

DRIAS met à disposition les résultats de projections issues de modèles climatiques, obtenus dans différents laboratoires de modélisation du climat. Il est important de souligner que l'ensemble des données DRIAS, y compris les données disponibles sur les périodes passées, ne sont pas des données météo observées mais des données simulées par ces modèles climatiques. Ainsi, les situations météorologiques simulées sont virtuelles et n'ont pas pour objectif de reproduire exactement la situation réellement observée ou qui sera observée, mais

de décrire les potentielles évolutions et tendances climatiques.

Les données sont principalement disponibles pour la France métropolitaine, les ressources étant plus limitées pour l'outre-mer.

L'espace « Découverte » du site permet d'obtenir de manière simple, rapide et sous forme graphique les résultats des projections. Par exemple, la figure 3 rend compte de l'évolution du nombre de jours de gel par an, de l'écart du cumul de précipitations d'avril à octobre et du nombre de jours dont la température maximale dépasse les 35°C.

Pour des études plus approfondies, il est nécessaire de passer par les données numériques, téléchargeables en fichier csv. Pour ceci, l'utilisateur est d'abord invité à sélectionner un modèle, puis à choisir le scénario d'intérêt : « historique », RCP2.6, RCP4.5 ou RCP8.5 (le scénario RCP6.0 n'est pas disponible pour le moment).

Les données « historiques » mises à disposition pour la période 1950-2005 sont également issues de simulations. Elles ne sont donc pas adaptées à l'analyse fine de régimes où la sinistralité est directement dépendante des conditions météorologiques.

Pour ce qui est des projections, elles sont généralement disponibles de 2006 à 2100. Une sélection par saison ou par mois peut être réalisée. Concernant les variables d'intérêt, elles sont réparties en 6 catégories : températures, précipitations, humidité, rayonnement, vent et évapo-transpiration potentielle.

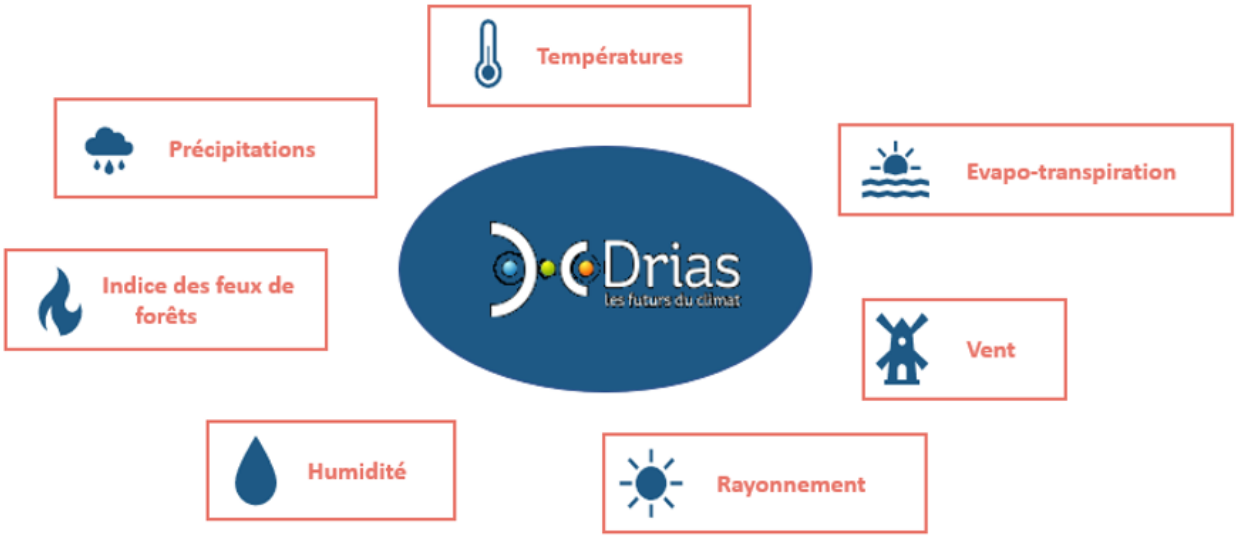


Figure III.5 – Catégories de variables disponibles

Enfin, il convient de sélectionner la localisation des projections parmi les 8 981 points proposés en France métropolitaine. Les projections disponibles sont en effet issues de modèles climatiques globaux dont la maille varie de 50 à 300km, couplés à des modèles régionaux avec une maille de l'ordre d'une dizaine de kilomètres. Les données subissent encore une étape de correction visant à enlever le biais qui peut exister en raison des observations locales et à affiner de nouveau l'échelle spatiale.

Avec le portail DRIAS, nous avons récupéré les données journalières des températures moyennes, des précipitations et des vitesses de vent de 2023 à 2100 selon le modèle ALADIN63. Deux scénarios sélectionnés pour la suite de l'étude sont le RCP 8.5 (le scénario le plus pessimiste) et le RCP 2.6 (le plus optimiste). Ce choix nous permet d'obtenir deux évaluations de l'impact du changement climatique sur le régime « beau temps », l'une dans le meilleur des cas et l'autre dans le pire des cas. Une autre raison pratique a motivé ce choix : le portail DRIAS ne met pas à disposition le scénario intermédiaire 6.0.

La figure III.6 présente le positionnement des points DRIAS que nous avons sélectionnés par rapport aux stations météorologiques de Météo France (SYNOP).

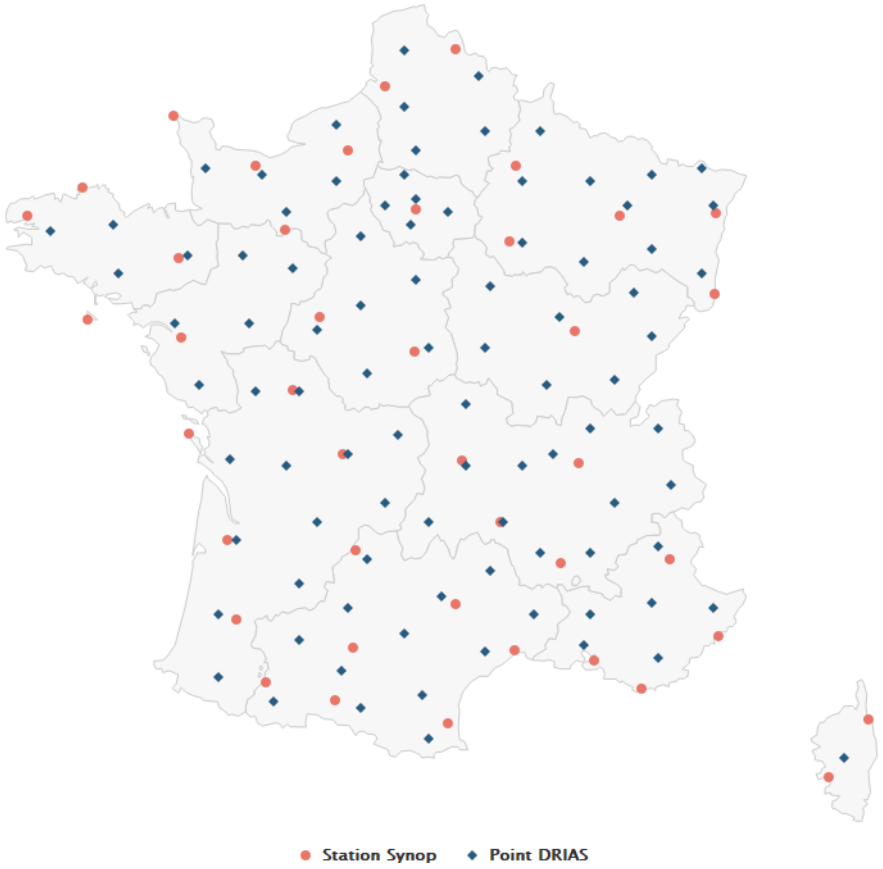


Figure III.6 – Positionnement des points DRIAS sélectionnés et des stations SYNOP

2.4 Adéquation entre les données DRIAS et les données Synop

Les données produites par ces modèles ont ensuite été corrigées pour représenter au mieux les évolutions attendues de température, de précipitations et de vitesses de vent en France, en utilisant la méthode ADAMONT. Cette méthode de correction est basée sur une analyse statistique des erreurs de simulation des modèles climatiques régionaux par rapport aux observations réelles, ce qui permet de réduire les biais dans les projections climatiques. les données journalières de débit retenues sur la période de 2023 à 2100, pour les scénarios RCP 2.6 et 8.5, ont été agrégées par jour et par région pour correspondre à la résolution spatiale et temporelle choisie pour l'étude. Lors de cette agrégation, nous avons choisi de

sélectionner la moyenne pour la température et la vitesse de vent ainsi que la somme pour les précipitations. Ce traitement permet d'être en adéquation avec les traitements que nous avons effectuer sur données historiques SYNOP.

Les figures III.7 et III.8 mettent en évidence l'évolution de la température, précipitation et vitesse de vent moyenne en France métropolitaine suivant les données à notre disposition. De manière générale on note une tendance globalement cohérent sur ces trois variables entre les données historiques et les données futures DRIAS.

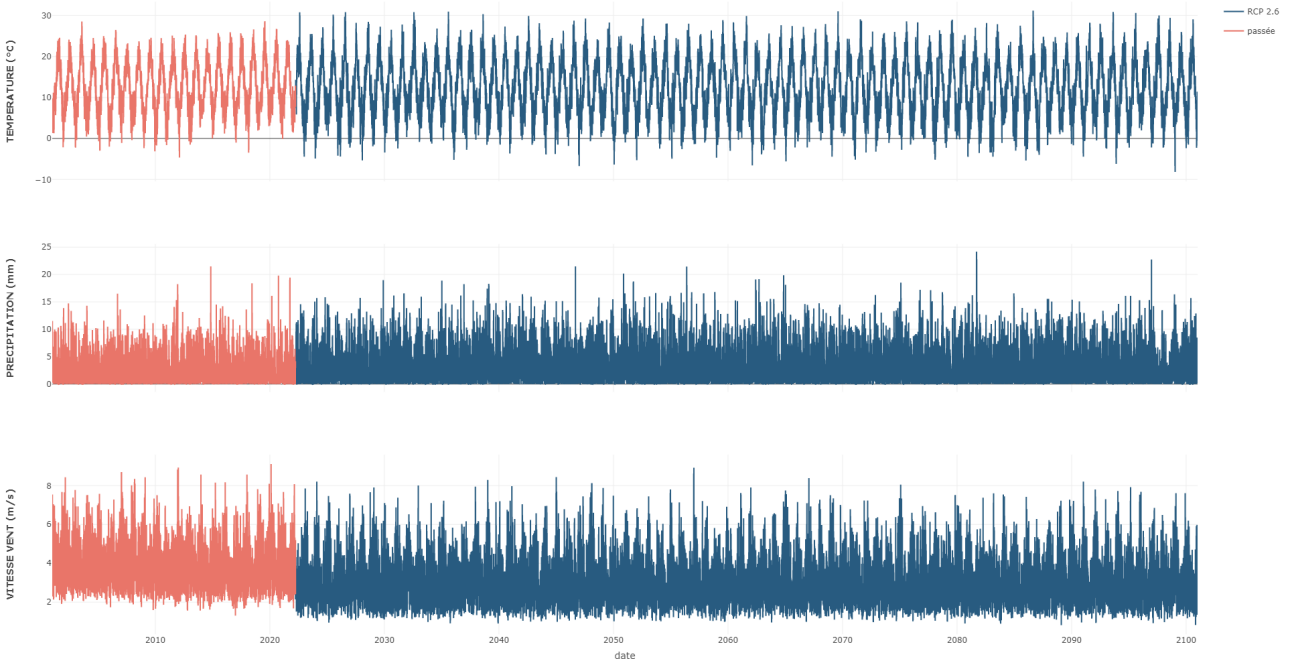


Figure III.7 – Température, précipitation et vitesse de vent moyenne en France métropolitaine : Données historiques Météo France vs DRIAS RCP 2.6

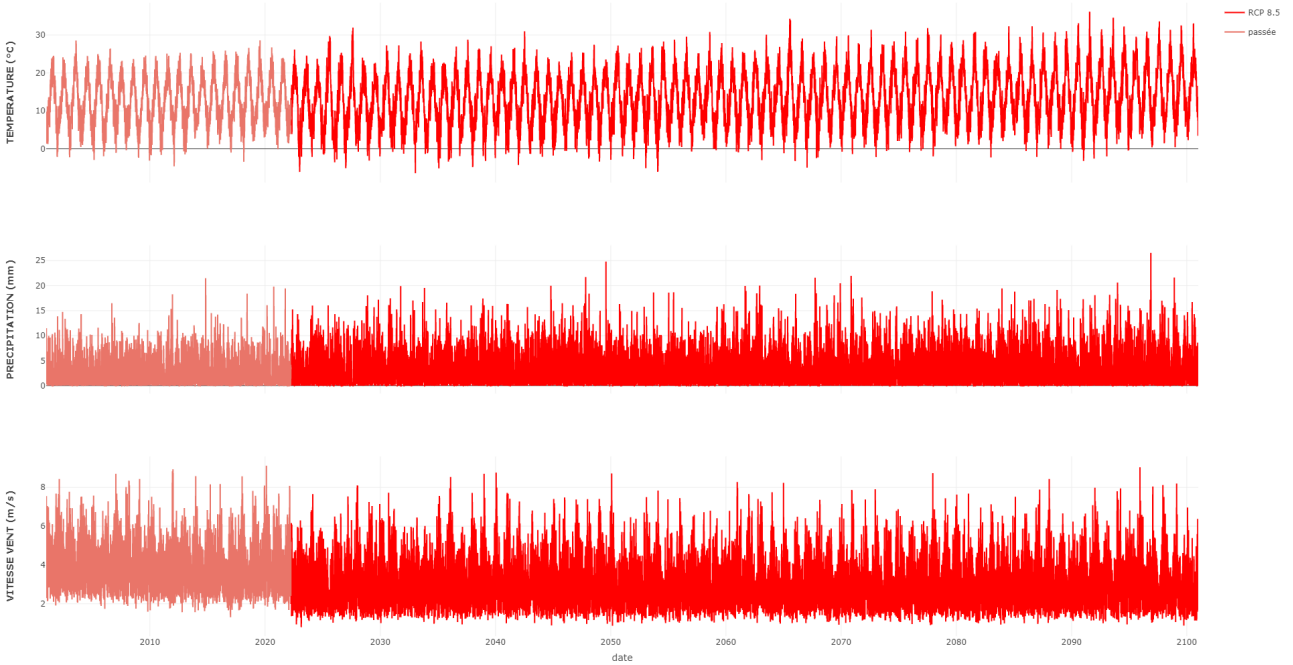


Figure III.8 – Température, précipitation et vitesse de vent moyenne en France métropolitaine : Données historiques Météo France vs DRIAS RCP 8.5

3 Hypothèses de projection du régime

Dans cette section nous examinerons en détail les hypothèses fondamentales qui sous-tendent les projections futures du régime fictif « **beau temps** ». Il s'agit principalement des hypothèses sur l'évolution de la population assurée ainsi que celles de la grille tarifaire.

3.1 Hypothèses sur la population assurée

Les hypothèses sur la population assurée fournissent des estimations sur l'évolution probable de cette population à l'horizon 2100. Pour rappel, le nombre de personnes assurées pour une journée donnée dans un mois et une région était un pourcentage variable du nombre de nuitées journalières correspondantes. Nous avons estimé ce nombre à l'aide de la formule suivante :

$$Nu(j) = \frac{EQ(j) * Nu(annee, mois, R)}{\sum_{j \in mois} EQ(j)} \quad \text{(III.3)}$$

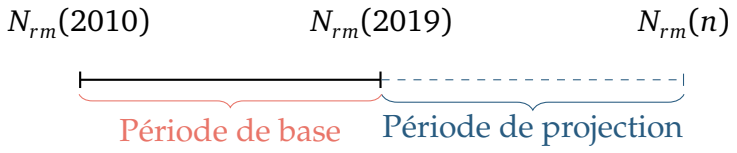
$EQ(j)$ est **l'équivalence en nuitée** du jour j, représente le poids relatif du jour j dans le nombre de nuitées total mensuel et est fonction de la variable week-end (we) indiquant si le jour j est un samedi ou un dimanche et de la variable vacance ($vacs$) indiquant si le jour j est un jour de vacance scolaire ou non. Sa formule est donnée par :

$$EQ(j) = (1 - we(j)) + \alpha_{we} * we(j) + \alpha_{vacs} * vacs(j) \quad \text{(III.4)}$$

À partir de la formule III.3, nous pouvons projeter le nombre de nuitées journalières en projetant, dans un premier temps, le nombre total de nuitées mensuelles pour un mois et une région donnés. Pour ce faire, et en tenant compte de l'observation des tendances historiques des nuitées mensuelles entre 2010 et 2019 pour chaque région et mois, nous avons opté pour une méthode basée sur l'extrapolation des tendances. Les méthodes d'extrapolation des tendances consistent à prolonger les tendances historiques observées et sont couramment utilisées en démographie de la population.

Parmi les méthodes basées sur l'extrapolation des tendances, nous avons utilisé celle relative à la croissance géométrique pour la projection du nombre de nuitées mensuelles à l'horizon 2100. Cette méthode suppose que le nombre total de nuitées pour une région donnée et un mois donné change au même taux (en pourcentage) par unité de temps (par exemple, une année) et que ces changements se produisent à intervalles distincts.

Pour simplifier les notations, on désigne par $N_{r,m}(n)$ le nombre total de nuitées du mois m pour la région r et l'année de projection n. La situation peut être représentée de la manière suivante :



Avec :

$N_{rm}(2019)$: Nombre de nuitées de départ

$N_{rm}(2010)$: Nombre initial de nuitées (période de base)

Le choix de la période de base (2010 - 2019) se justifie par le fait que l'activité de camping a été impactée par la pandémie du COVID-19 en 2019 et 2020, ce qui a affecté le nombre de nuitées. Ainsi, l'année 2019 constitue une année avec des données fiables et récentes. L'année 2010 est justifiée par le fait que les taux de croissance sont positifs pour la période de 2010 à 2019. Il est tout à fait logique de penser qu'il y aura de plus en plus de nuitées, en particulier en raison de la croissance de la population.

Soit T_{rm} le taux géométrique de variation annuelle moyenne pour la région r et le mois m. Sa formule est la suivante :

$$T_{rm} = \left[\left(\frac{N_{rm}(2019)}{N_{rm}(2010)}\right)^{1/9}\right] - 1 \qquad \text{(III.5)}$$

Ainsi, le nombre total de nuitées pour la région r et le mois m pour l'année de projection n ($n \geq 2023$) est donné par la formule suivante :

$$N_{rm}(n) = N_{rm}(2019)[1 + T_{rm}]^{(n-2019)} \qquad \text{(III.6)}$$

Le graphique III.9 montre la matrice des taux moyens T_{rm} d'évolution du nombre de nuitées par région et par mois. La région des **Hauts-de-France** présente les taux d'évolution les plus élevés, tandis que la région de la **Corse** présente les taux les plus faibles.

	Auvergne-Rhone-Alpes	Bourgogne-Franche-Comte	Bretagne	Centre-Val de Loire	Corse	Grand Est	Hauts-de-France	Ile-de-France	Normandie	Nouvelle-Aquitaine	Occitanie	Pays de la Loire	PACA
janvier	2.6	2.4	4	2.7	0.8	2.9	5.6	3.4	3.6	3	2.2	2.7	1.7
février	2.7	2.6	4.2	2.9	1	3.1	5.8	3.5	3.8	3.1	2.3	2.9	1.9
mars	2.4	2.2	3.8	2.5	0.6	2.7	5.4	3.2	3.4	2.8	2	2.5	1.5
avril	2.3	2.2	3.8	2.5	0.6	2.7	5.4	3.1	3.4	2.7	1.9	2.5	1.5
mai	1.8	1.6	3.2	1.9	0.1	2.1	4.8	2.6	2.8	2.2	1.4	1.9	1
juin	2.2	2	3.6	2.3	0.4	2.5	5.2	3	3.2	2.6	1.8	2.3	1.3
juillet	1.7	1.5	3.1	1.9	0	2	4.8	2.5	2.8	2.1	1.3	1.9	0.9
août	2.2	2	3.7	2.4	0.5	2.6	5.3	3	3.3	2.6	1.8	2.4	1.4
septembre	2	1.8	3.4	2.2	0.3	2.3	5.1	2.8	3.1	2.4	1.6	2.2	1.2
octobre	2.2	2.1	3.7	2.4	0.5	2.6	5.3	3	3.3	2.7	1.9	2.4	1.4
novembre	3.1	2.9	4.5	3.2	1.3	3.4	6.2	3.9	4.1	3.5	2.7	3.2	2.2
décembre	2.7	2.5	4.2	2.9	1	3.1	5.8	3.5	3.8	3.1	2.3	2.9	1.9

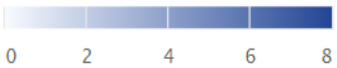


Figure III.9 – Matrice des taux d'évolution du nombre de nuitées par région et par mois.

Une fois que les nuitées mensuelles sont projetées à l'aide de la matrice des taux d'évolution, nous devons passer aux nuitées journalières en utilisant la formule III.4 pour calculer l'équivalence en nuitées EQ par jour pour les années de 2023 à 2100. À cet effet, nous considérons que les paramètres α_{vacs} (= 2.74) et α_{we} (= 1.33) sont constants sur toute la période de projection, quel que soit le scénario du RCP. La variable week-end (we), qui indique si le jour j est un samedi ou un dimanche, est connue à l'avance pour toute la période de projection. En revanche, la variable vacance ($vacs$), qui indique si le jour j est un jour de vacances scolaires ou non, n'est pas connue pour toutes les années de la période de projection.

L'arrêté [1] du 7 décembre 2022 fixe le calendrier scolaire des années 2023-2024, 2024-2025 et 2025-2026. Pour déterminer les jours de vacances pour les années de 2027 à 2100, nous les avons réparties en fonction des années qui partagent le même calendrier que celles dont les jours de vacances sont connus. En effet, il existe 14 calendriers annuels en tout, comprenant sept (7) calendriers pour les années ordinaires et sept (7) pour les années bissextiles. Dans les deux cas, ces calendriers débutent chaque année avec l'un des sept jours de la semaine. Par conséquent, si le calendrier d'une année est identique à celui d'une autre année, nous supposons que leurs jours de vacances scolaires sont également identiques. Cette hypothèse nous semble plausible, car depuis l'année 2000, le nombre de jours de vacances scolaires annuels est pratiquement similaire, à un ou deux jours près, totalisant environ 118 jours par an en moyenne. Le tableau III.2 présente la répartition des années de 2027 à 2100 en fonction des 14 calendriers et des années dont les jours de vacances sont déjà connus.

Calendrier	Années futures	identiques à
années non-bissextiles		
1	2034, 2045, 2051, 2062, 2073, 2079, 2090	2026
2	2031, 2042, 2053, 2059, 2070, 2081, 2087, 2098	2025
3	2034, 2045, 2051, 2062, 2073, 2079, 2090	2023
4	2033, 2039, 2050, 2061, 2067, 2078, 2089, 2095	2022
5	2027, 2038, 2049, 2055, 2066, 2077, 2083, 2094	2021
6	2030, 2041, 2047, 2058, 2069, 2075, 2096, 2097	2019
7	2029, 2035, 2046, 2057, 2063, 2074, 2085, 2091	2018
années bissextiles		
8	2052	2024
9	2048,2076	2020
10	2044,2072,2100	2016
11	2040,2068,2096	2012
12	2036,2064,2092	2008
13	2032,2060,2088	2004
14	2028,2056,2084	2000

Table III.2 – Répartition des années futures de projection selon les 14 calendriers.

Une fois que les jours de vacances scolaires ont été déterminés pour toute la période de projection, il est possible de calculer les nuitées journalières en appliquant la formule III.3.

1. https ://www.legifrance.gouv.fr/jorf/id/JORFTEXT000046704476

Le graphique présenté dans la figure D.3 en annexe illustre l'évolution annuelle des nuitées et met en évidence la tendance géométrique que nous avons préalablement supposée.

Nous faisons l'hypothèse que le nombre d'assurés couverts par le régime le jour j, au cours des années de projection à venir, équivaut à 25% du nombre de nuitées journalières dans un marché caractérisé par une concurrence forte. Par conséquent, nous supposons qu'il n'y aura aucune variation de la part de marché sur la période de projection, ni entre les différentes régions de la France métropolitaine, ni entre les scénarios du GIEC (optimiste RCP 2.6 et pessimiste RCP 8.5).

3.2 Hypothèses sur la grille tarifaire

Tout comme les hypothèses formulées concernant la population assurée, la formulation d'hypothèses sur la structure tarifaire revêt une importance cruciale pour la projection du régime, car cette grille sert de base au calcul des primes annuelles totales.

Nous supposons que la grille tarifaire restera inchangée sur toute la période de projection, sans aucune adaptation tarifaire (redressement tarifaire), même en cas de détérioration du ratio S/P, quel que soit le scénario du GIEC (RCP 2.6 et RCP 8.5). Cette grille tarifaire correspond à celle de la dernière révision tarifaire du régime, comme détaillé dans la section 5 du chapitre 1. La figure III.10 suivante présente cette grille tarifaire qui sera utilisée dans les deux approches que nous avons définies précédemment (approche déterministe et par modélisation) pour la projection du régime.

Dernière grille tarifaire révisée sur la période historique

	Ile-de-France	Centre-Val de Loire	Bourgogne-Franche-Comte	Normandie	Hauts-de-France	Grand Est	Pays de la Loire	Bretagne	Nouvelle-Aquitaine	Occitanie	Auvergne-Rhone-Alpes	Provence-Alpes-Cote d Azur	Corse
Janvier	3.65 €	2.95 €	11.34 €	6.64 €	13.58 €	9.61 €	5.62 €	12.77 €	2.52 €	2.49 €	8.17 €	5.57 €	5.41 €
Février	5.42 €	7.91 €	12.34 €	7.98 €	8.67 €	10.23 €	7.62 €	11.90 €	7.53 €	5.94 €	9.90 €	5.78 €	5.14 €
Mars	13.14 €	14.45 €	14.56 €	22.52 €	23.83 €	25.42 €	11.70 €	15.33 €	7.58 €	8.97 €	12.56 €	12.07 €	7.13 €
Avril	4.55 €	7.51 €	5.59 €	7.89 €	6.11 €	3.21 €	6.21 €	10.78 €	3.80 €	5.31 €	4.90 €	11.43 €	2.80 €
Mai	3.29 €	4.91 €	5.32 €	3.22 €	6.88 €	3.95 €	5.06 €	8.84 €	2.37 €	3.38 €	4.83 €	8.34 €	4.40 €
Juin	10.82 €	4.78 €	8.42 €	10.76 €	10.01 €	3.42 €	5.13 €	13.44 €	7.20 €	6.26 €	4.22 €	9.28 €	2.13 €
Juillet	8.86 €	6.93 €	11.04 €	10.35 €	11.47 €	6.43 €	5.64 €	11.55 €	4.42 €	7.13 €	9.46 €	20.56 €	13.05 €
Août	7.47 €	5.35 €	7.10 €	8.12 €	8.55 €	4.64 €	5.69 €	11.13 €	7.33 €	6.27 €	7.02 €	14.49 €	9.23 €
Septembre	2.60 €	2.72 €	3.98 €	3.87 €	4.98 €	1.04 €	2.58 €	5.01 €	2.58 €	1.86 €	5.35 €	6.15 €	3.35 €
Octobre	3.72 €	3.84 €	8.28 €	9.70 €	5.65 €	6.96 €	4.87 €	8.75 €	3.38 €	3.86 €	8.13 €	9.80 €	6.43 €
Novembre	9.20 €	8.50 €	18.49 €	16.87 €	13.78 €	15.64 €	7.81 €	17.20 €	8.79 €	10.49 €	12.56 €	15.91 €	5.52 €
Décembre	4.24 €	3.74 €	7.58 €	13.28 €	11.03 €	6.51 €	7.74 €	15.64 €	5.30 €	1.38 €	2.98 €	3.22 €	1.71 €

Figure III.10 – Grille tarifaire retenue pour la projection du régime « **beau temps** ».

3.3 Autres hypothèses de projection

Outre les hypothèses concernant la population assurée et la grille tarifaire, nous pouvons établir des hypothèses plus intrinsèques au fonctionnement du régime. Par exemple, dans l'approche déterministe, nous supposons que le mécanisme de changement restera identique quel que soit le scénario du GIEC. En d'autres termes, le régime continuera de couvrir les trois risques - la température, la pluie et le vent - et les définitions des formules liées à ces risques et aux seuils resteront inchangées. De plus, nous présumons que l'indemnité de 50€ restera fixe.

De plus, dans les deux approches que nous avons adoptées, nous supposons qu'il n'y aura pas d'événements de type pandémique susceptibles d'affecter le régime, et nous présumons qu'il n'y aura pas d'effet d'inflation sur les tarifs et les paiements forfaitaires.

4 Résultats de la projection

Dans cette section, nous présenterons les résultats de la projection du régime selon les deux scénarios du GIEC retenus (RCP 2.6 et 8.5) et selon les deux approches adoptées : l'approche déterministe et l'approche par modélisation. Nous accorderons une attention particulière aux indicateurs de performance, notamment le ratio S/P et l'évolution des réserves.

4.1 Approche déterministe

4.1.1 Présentation et formalisation

L'approche déterministe consiste simplement à appliquer le mécanisme de fonctionnement du régime aux données du GIEC. Pour rappel, le régime « beau temps » est constitué de quatre indicateurs associés aux risques de température (Th et Tb), de vent (Vh) et de pluie (Ph). Ainsi, on calcule, pour les deux scénarios du GIEC (RCP 8.5 et 2.6), les indicateurs journaliers suivants pour les années 2023 à 2100 :

$$\begin{cases} \text{Température :} & Th(i) = max(T_{moyenne}(i) - T_{saison}(i), 0) \text{ et} \\ & Tb(i) = -min(T_{moyenne}(i) - T_{saison}(i), 0) \\ \text{Vent :} & Vh(i) = V(i) - V_{saison}(i) \\ \text{Pluie :} & Ph(i) = P(i) - P_{saison}(i) \end{cases}$$

Avec :

— i : jour de l'année (2023 à 2100) sur lequel on effectue le calcul ;
— $T_{moyenne}(i)$, $V(i)$ et $P(i)$ respectivement la température moyenne (°C), la vitesse de vent moyenne (m/s) et la précipitation (mm) journalière du jour i ;
— $T_{saison}(i)$, $V_{saison}(i)$ et $P_{saison}(i)$ respectivement la température moyenne, la vitesse de vent et la précipitation saisonnière du jour i sur la décennie précédente.

Sur la période de 2001 à 2100, on compte au total 10 périodes décennales, à savoir : 2001 à 2010, 2011 à 2020, 2021 à 2030, 2031 à 2040, 2041 à 2050, 2051 à 2060, 2061 à 2070, 2071 à 2080, 2081 à 2090, et 2091 à 2100. Ainsi, pour les jours du mois de janvier 2050, $T_{saison}(i)$, $V_{saison}(i)$ et $P_{saison}(i)$ représenteront respectivement la moyenne des températures, la moyenne des vitesses de vent et la moyenne des précipitations en période hivernale entre 2031 et 2040.

Le tableau III.3 donne les valeurs de T_{saison}, V_{saison} et P_{saison} pour les 4 saisons et les dix (10) périodes décennales couvrant les années 2001 à 2100.

Saison	Moyenne	Température		Vitesse vent		Précipitation	
	Période	RCP 2.6	RCP 8.5	RCP 2.6	RCP 8.5	RCP 2.6	RCP 8.5
Eté	2001 - 2010	19.5		3.52		1.65	
	2011 - 2020	19.95		3.27		1.43	
	2021 - 2030	19.92	19.51	2.59	2.64	1.74	2.08
	2031 - 2040	19.92	19.23	2.48	2.55	1.89	2.04
	2041 - 2050	20.13	19.49	2.51	2.55	1.85	2.19
	2051 - 2060	19.69	20.34	2.45	2.57	2.12	1.79
	2061 - 2070	19.88	21.19	2.46	2.44	2.04	1.86
	2071 - 2080	19.83	21.59	2.47	2.45	2.05	1.87
	2081 - 2090	19.11	22.52	2.53	2.49	2.04	1.77
	2091 - 2100	19.89	23.86	2.52	2.48	1.92	1.58
Hiver	2001 - 2010	5.65		4.24		1.96	
	2011 - 2020	6.52		4.04		2.00	
	2021 - 2030	5.34	5.22	3.43	3.43	2.28	2.50
	2031 - 2040	5.32	5.13	3.24	3.22	2.93	2.66
	2041 - 2050	5.21	5.75	3.31	3.44	2.92	2.90
	2051 - 2060	5.09	5.97	3.28	3.16	2.74	2.68
	2061 - 2070	5.30	6.54	3.30	3.27	2.93	2.97
	2071 - 2080	4.48	7.03	3.17	3.14	2.51	2.81
	2081 - 2090	4.70	7.53	3.25	3.20	2.68	2.78
	2091 - 2100	5.08	8.20	3.36	3.40	2.86	3.08
Autome	2001 - 2010	9.97		3.82		2.39	
	2011 - 2020	10.84		3.56		2.44	
	2021 - 2030	9.52	9.03	3.00	2.89	2.98	2.84
	2031 - 2040	9.88	9.43	2.96	2.90	3.13	3.23
	2041 - 2050	9.97	9.91	2.85	2.78	2.96	3.03
	2051 - 2060	10.32	10.98	2.87	2.82	3.05	2.79
	2061 - 2070	9.45	11.10	2.87	2.82	2.85	3.35
	2071 - 2080	9.59	11.61	2.85	2.81	3.13	3.07
	2081 - 2090	9.81	11.87	2.80	2.77	3.16	3.28
	2091 - 2100	9.31	12.70	2.83	2.90	3.04	3.29
Printemps	2001 - 2010	13.7		3.83		1.93	
	2011 - 2020	13.84		3.57		1.90	
	2021 - 2030	12.95	12.88	2.97	2.90	2.31	2.47
	2031 - 2040	12.74	13.24	2.81	2.82	2.44	2.39
	2041 - 2050	13.11	13.24	2.91	2.83	2.62	2.44
	2051 - 2060	13.46	13.42	2.83	2.82	2.47	2.43
	2061 - 2070	12.83	14.23	2.89	2.86	2.61	2.30
	2071 - 2080	12.91	14.94	2.89	2.85	2.57	2.40
	2081 - 2090	13.33	14.89	2.75	2.89	2.44	2.50
	2091 - 2100	13.65	16.12	2.82	2.76	1.86	2.38

Table III.3 – Température (°C), Vitesse de vent (m/s) et précipitation (mm) moyenne décennale selon les scénarios du GIEC (2.6 et 8.5) de 2001 à 2100.

L'analyse de ce tableau révèle que les températures saisonnières sont plus élevées dans le scénario RCP 8.5, quelle que soit la saison. En ce qui concerne les précipitations, les moyennes saisonnières montrent une tendance à la baisse sur la période de projection pour le scénario RCP 8.5, tandis qu'elles affichent plutôt une tendance à la hausse dans le scénario RCP 2.6. Cependant, la tendance des vitesses de vent saisonnières semble plutôt erratique.

Le régime indemnisera de 50€ lorsque au moins l'un des indicateurs de baisse de température Tb, de hausse de température Th, de vitesse de vent Vh et de précipitation Ph sera supérieur à son quantile 95% observé sur une période décennale. Les quantiles décennales sont présentés dans le tableau D.5 en annexe.

La figure ci-dessous montre l'évolution selon les scénarios du GIEC (2.6 et 8.5) du nombre de jours moyens annuels où au moins l'un des quatre indicateurs est dépassé par son seuil. La ligne horizontale en pointillés (couleur corail rouge) représente le nombre de dépassements moyen historique, qui est égal à 52.

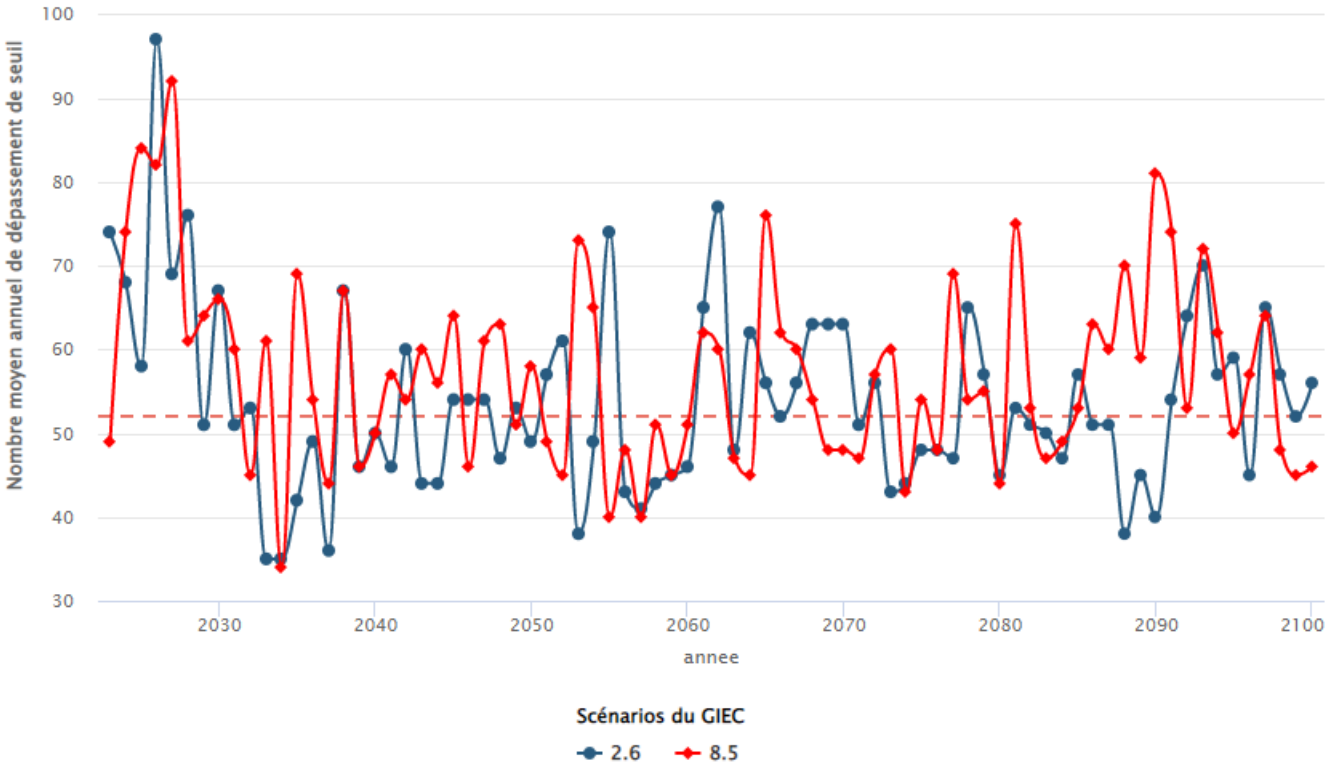


Figure III.11 – Nombre de dépassement annuel selon les scénarios du GIEC.

On constate principalement une grande variabilité dans le nombre annuel de dépassements, principalement en raison des variations climatiques. De plus, on observe un nombre élevé de dépassements de seuil en début de période de projection (2023 - 2030), en partie en raison du changement de trajectoire. Cette tendance imprévisible aura un impact sur les réserves et le ratio S/P en début de projection.

Pour calculer les sinistres et les primes annuelles, l'approche déterministe intègre les hypothèses concernant la population assurée et la grille tarifaire. Les sinistres (S(n)) et les primes annuelles (P(n)) pour une année de projection n sont obtenus à l'aide des formules suivantes :

$$P(n) = \sum_{j,m,r} \pi^{h}_{m,r} * (0.25 * N_{rm,j}(n))$$

$$S(n) = \sum_{j} 50 * (0.25 * N_{rm,j}(n)) * max(\mathbb{1}_T(j), \mathbb{1}_V(j), \mathbb{1}_P(j))$$

Avec :

— $\pi^{h}_{m,r}$ prime issue des hypothèses sur la grille tarifaire pour le mois m et la région r ;

— $0.25 * N_{rm,j}(n)$ issu des hypothèses de projection sur la population assurée et représente le nombre d'assurée pour le jour j ($N_{rm,j}$: nuitée projetée) ;

— $max(\mathbb{1}_T(j), \mathbb{1}_V(j), \mathbb{1}_P(j))$ indicateur indiquant si au moins l'un des quatre indicateurs est dépassé par son seuil.

4.1.2 Résultats de la projection du régime par l'approche direct

La figure III.12 montre la projection du ratio sinistre à prime (pure) du régime « beau temps » pour chaque scénario du GIEC retenu et selon l'approche déterministe de 2023 à 2100.

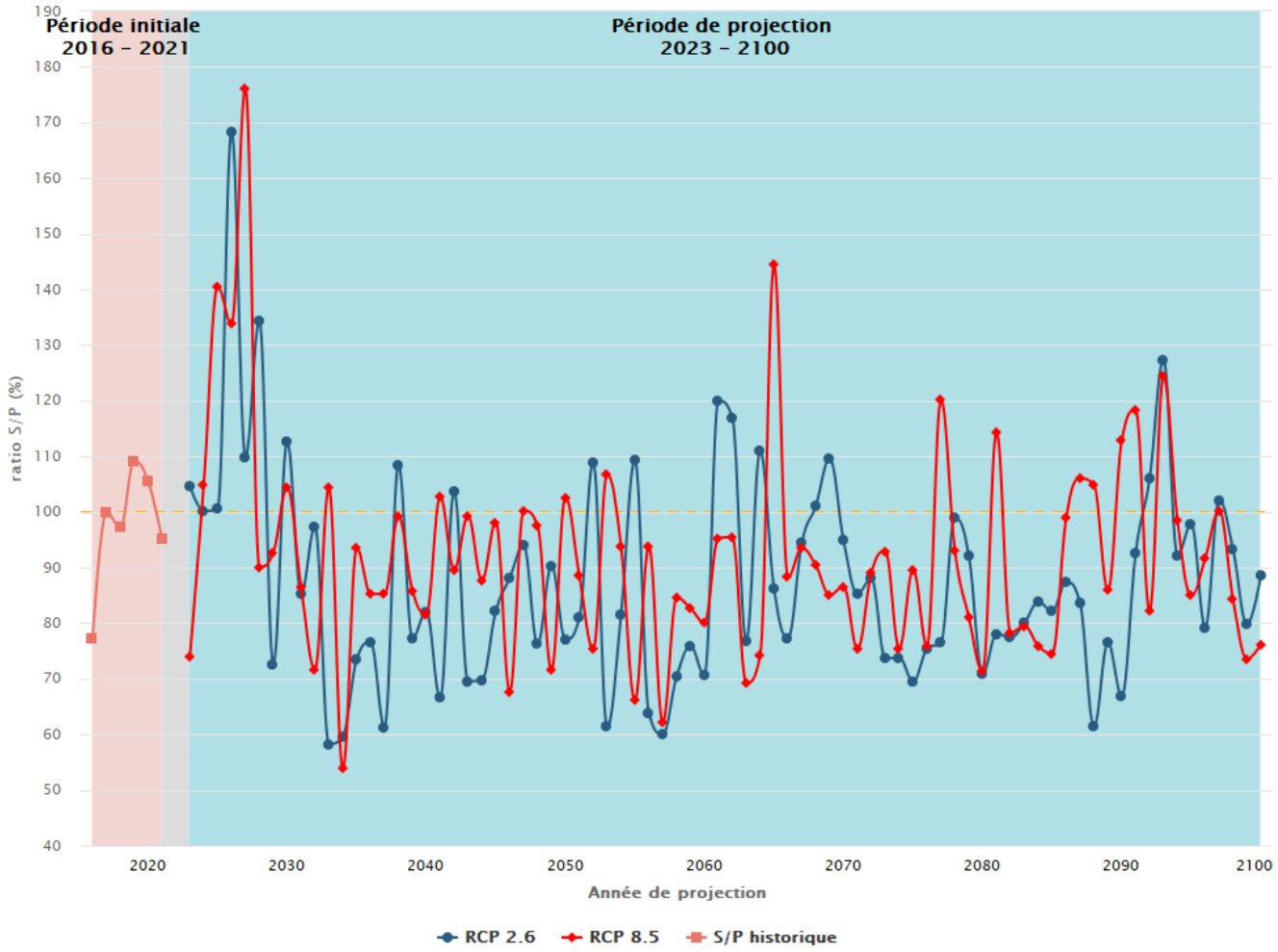


Figure III.12 – S/P projetés selon les scénarios du GIEC et selon l'approche déterministe.

Les premières analyses de ce graphique révèlent que la tendance des S/P est très erratique, quel que soit le scénario. Cette observation initiale n'est pas surprenante, étant donné la variabilité observée dans le climat. Un autre constat important est la forte détérioration du ratio sinistre à prime en début de période de projection (entre 2023 et 2028). Cette dégradation significative s'explique principalement par un changement brutal de la trajectoire climatique actuellement observée.

Le scénario RCP 8.5 est le scénario le plus pessimiste du GIEC, prévoyant l'absence de politiques ou d'actions visant à réduire les émissions de gaz à effet de serre d'ici à 2100. Selon ce scénario, on prévoit une augmentation des températures en France d'environ 5°C d'ici à la fin du siècle. Sous ce scénario, la moyenne des ratios S/P sur l'ensemble de la période de projection est de 91.9%, ce qui est inférieur à la moyenne historique de 97.4%. Cependant, cette moyenne est accompagnée d'une variabilité beaucoup plus élevée, avec un écart-type de 19.3% par rapport à un écart-type de 11.2% sur la période historique. Cette situation découle du risque accru de températures élevées, en raison de l'augmentation des températures prévue dans le cadre du scénario RCP 8.5.

Dans le scénario le plus optimiste, correspondant à une réduction significative et rapide des émissions de gaz à effet de serre, on constate une amélioration du ratio S/P. Cependant, on n'observe pas de réduction significative de la volatilité du ratio S/P par rapport au scénario RCP 8.5. En effet, la moyenne des ratios S/P dans ce scénario est de 87.3%, avec une volatilité de 19.1%, ce qui est assez proche de celle observée dans le scénario RCP 8.5.

On observe généralement une situation moins favorable dans le scénario 8.5 par rapport au scénario 2.6, en particulier au-delà de 2065. En effet, plusieurs années de projection présentent un ratio S/P supérieur à 100%, ce qui n'est pas le cas dans le scénario 2.6. La moyenne des ratios S/P sur la période après 2065 est de 92.1% pour le RCP 8.5 et de 86.1% pour le RCP 2.6, avec des volatilités respectives de 16.9% et de 13.3%.

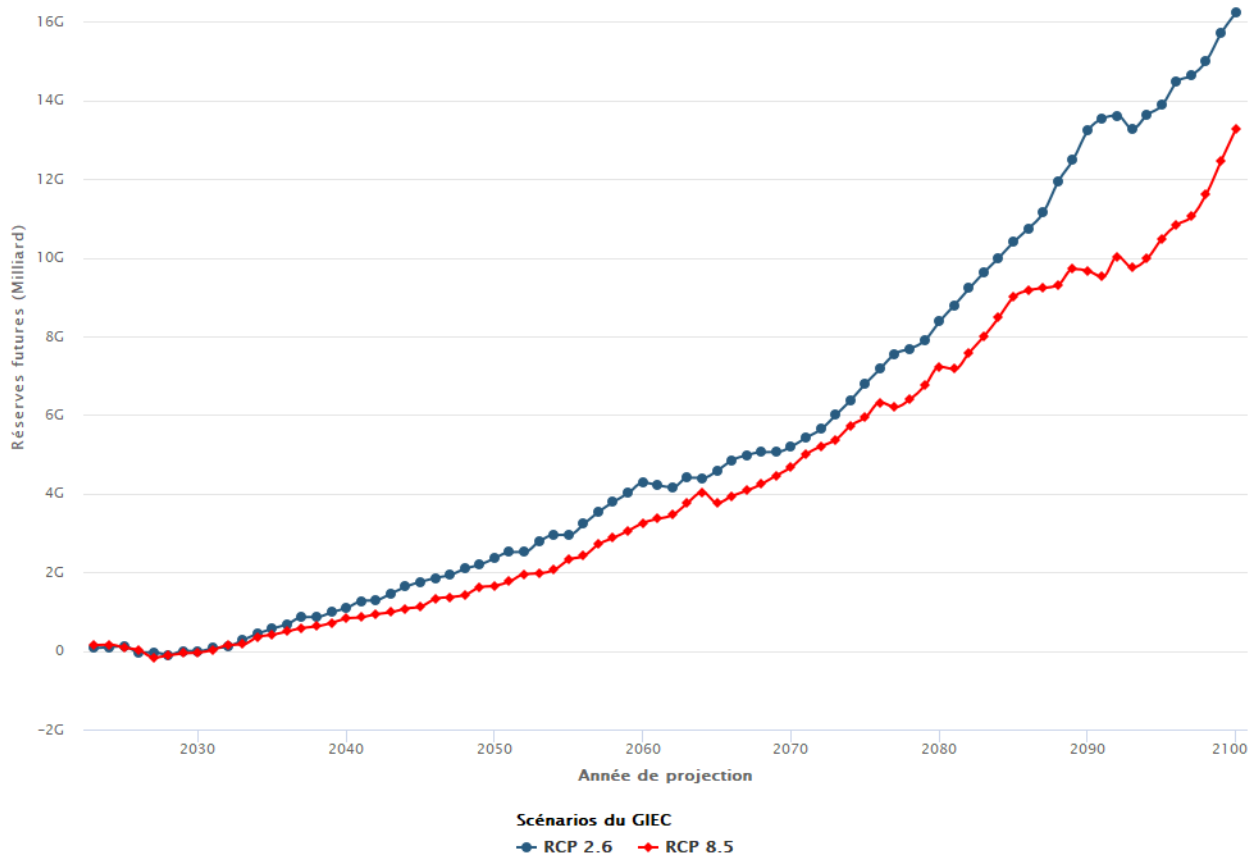


Figure III.13 – Réserves projetées selon les scénarios du GIEC par approche déterministe.

La figure III.13 présente l'évolution des réserves de 2023 à 2100 selon les scénarios du GIEC RCP 8.5 et 2.6. On observe une tendance linéaire à la hausse des réserves, quel que soit le scénario. Toutefois, les réserves dans le scénario RCP 8.5 sont inférieures à celles du scénario RCP 2.6. On constate une augmente progressivement de l'écart entre les réserves des deux scénarios sur tout le long de la période de projection.

La figure III.14 présente l'évolution du ratio S/P par région en 2100 selon les deux scénarios retenus. On observe que dans le scénario RCP2.6, les ratios les plus élevés se trouvent dans les régions du centre et du nord-ouest, tandis que dans le scénario RCP8.5, ce sont les régions du sud-est qui affichent les ratios les plus élevés.

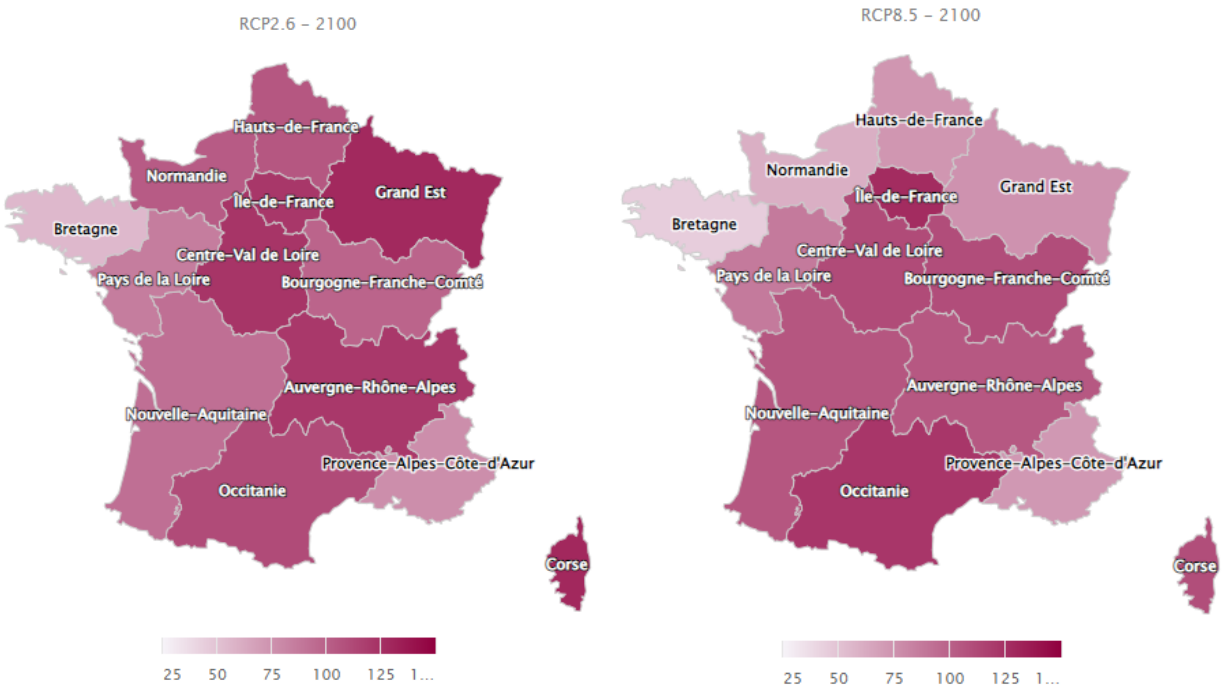


Figure III.14 – Ratio S/P par région en 2100 selon les deux scénarios.

4.2 Approche par modélisation

4.2.1 Résultats de la modélisation et formalisation

Comme nous l'avons décrit dans la section 1 de ce chapitre, l'approche par la *modélisation* suppose que le processus qui régit le fonctionnement du régime (déclenchement des paiements, indemnité,...) est aléatoire. Ainsi, dans cette approche, on projette dans un premier temps la sinistralité du régime sur la période de projection à l'aide d'un mixte de modèles qui consiste à modéliser les 4 composantes de l'espérance de la sinistralité S (voir formule III.2) que sont : la probabilité de déclenchement de paiements(τ), la probabilité de dépassement du seuil des extrêmes u (δ), la sinistralité attritionnelle ($Y|Y < u$) et la sinistralité extrême ($Y|Y \geq u$). Cette approche suppose l'indépendance entre d'une part la probabilité de déclenchement et le couple formé par les autres composantes et d'une autre part entre la probabilité de dépassement du seuil des extrêmes et le couple formé par la sinistralité

extrême ($Y|Y \geq u$) et la sinistralité attritionnelle ($Y|Y \geq u$). Nous n'avons pas testé ces indépendances dans le cadre de ce mémoire. Cependant, elles nous semblent acceptables dans la mesure où la sinistralité du régime dépend du nombre de nuitées, qui est indépendant du fonctionnement du régime. Les démarches et les résultats finaux de la modélisation de chaque composante sont les suivants :

Modélisation de la probabilité de déclenchement

Pour la modélisation de la probabilité de déclenchement nous avons suivi la démarche méthodologie présenté dans la figure C.1 de la section 1 en annexe C. Cette méthodologie est constituée de cinq (5) phases dont : *preprocessing* (Feature engineering, sélection des variables et rééchantillonnage), classification, comparaison, choix du seuil optimal et interprétation. Le lecteur pourra se référer à la section 1 de l'annexe C pour la description complète de ces différentes étapes. On y retrouve également les statistiques descriptives, les résultats et les interprétations.

Nous avons testé six modèles de machine learning : la Régression logistique (LR), l'Arbre de décision (DT), XGBoost (XBG), le Support Vector Machine (SVM), le Random Forest (RF), et les Réseaux de neurones (NN), avec six méthodes de rééchantillonnage (RUS, ROS, SMOTE, ADASYN, SMOTE+ENN, et SMOTE+TOMEK). Tous les modèles ont été hyperparamétrés à l'aide de la méthode GridSearchCV (CV = 5). La meilleure combinaison a été choisie en utilisant la métrique de performance Recall (Rappel), qui mesure la capacité d'un modèle de machine learning à identifier tous les exemples positifs, c'est-à-dire les déclenchements dans notre cas. Nous avons choisi le Recall car il est préférable de faire une prédiction erronée de déclenchement dans le scénario le plus pessimiste RCP 8.5 que de faire une prédiction erronée en indiquant qu'il n'y aura pas de déclenchement.

Le graphique ci-dessous présente la métrique Recall pour l'ensemble des combinaisons ('No' : modèle sans GridSearchCV et 'Yes' : modèle avec GridSearchCV).

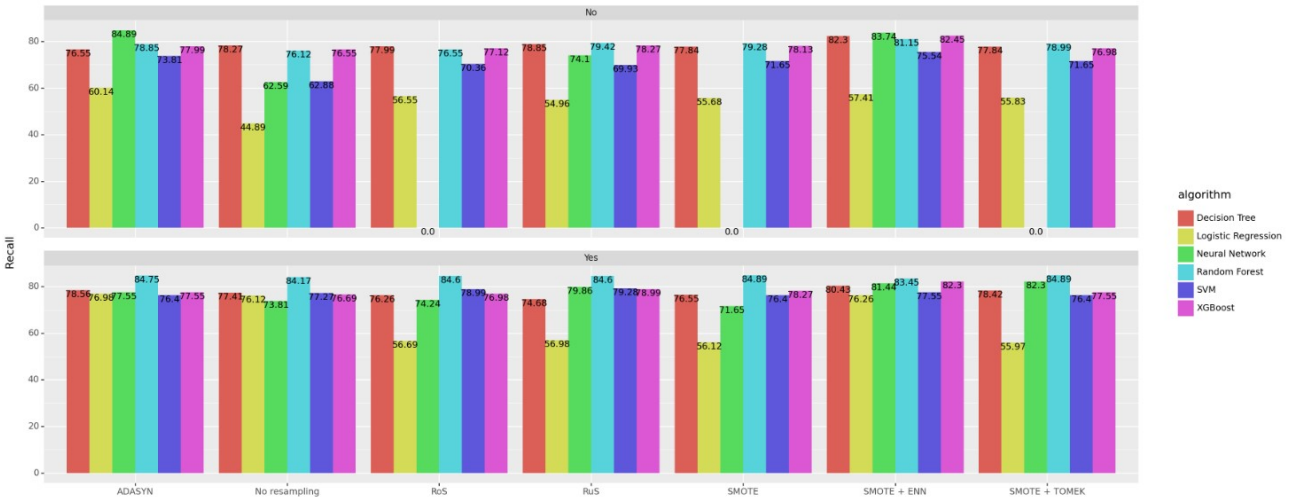


Figure III.15 – Sélection des meilleurs modèles pour la probabilité de déclenchement à l'aide du Recall (%).

La valeur la plus grande du Recall est de 84.89% et les combinaisons qui dont la valeur du Recall correspond à 84.89% : SMOTE + Forêt aléatoire (RF) + GridSearchCV ; (SMOTE+ENN) +Forêt aléatoire (RF) + GridSearchCV et ADASYN + Réseau de neurone (NN).

En se servant des autres mesures de performances (figure III.16) que sont l'extactitude, la précision, le F1, le G-mean et l'AUC ont abouti au choix de la combinaison *(SMOTE+ENN) +Forêt aléatoire (RF) + GridSearchCV* pour la prédiction des déclenchements.

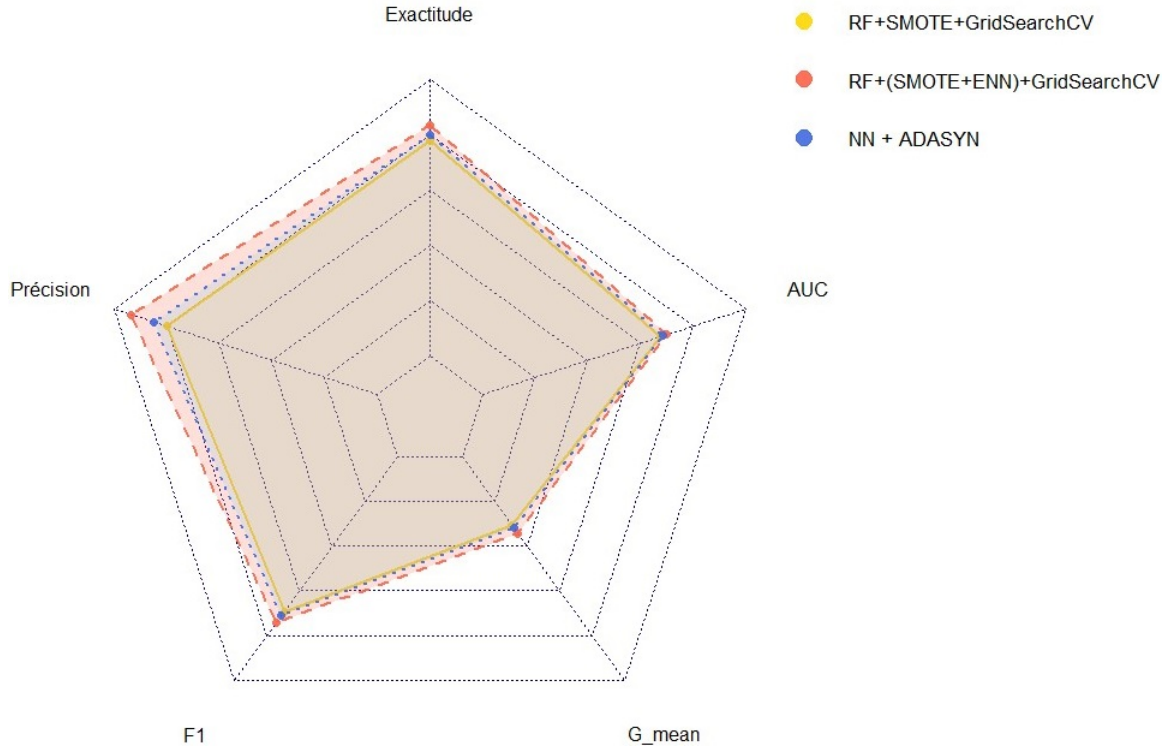


Figure III.16 – Sélection des meilleurs modèles pour la probabilité de déclenchement à l'aide des autres mesures.

Les paramètres du modèle final, hyperparamétré par GridSearchCV, sont les suivants :

— nombre d'arbres dans la forêt (*n_estimators*) : 100 ;
— nombre maximum de nœuds feuilles (*max_leaf_nodes*) : 20 ;
— poids associés aux classes (*class_weight*) : "balanced" ;
— fonction pour mesurer la qualité d'un split (*criterion*) : 'gini' ;
— profondeur maximale de l'arbre (*max_depth*) : 15 ;
— autres paramètres : défaut (scikit-learn).

Modélisation de la probabilité de dépassement du seuil extrême u

Tout comme dans la modélisation de la probabilité de déclenchement, la modélisation de la probabilité de dépassement du seuil extrême *u* se déroulera en cinq étapes distinctes (voir C.10 pour la démarche méthodologique). Ces étapes comprennent la phase de prétraitement (*preprocessing*), la classification, la comparaison, le choix du seuil optimal, et enfin, l'interprétation du meilleur modèle sélectionné. Cependant, contrairement à la phase de *prétraitement* de la modélisation de la probabilité de déclenchement, nous n'appliquons pas de rééchantillonnage car nous n'avons plus affaire à des classes déséquilibrées. De plus, dans la phase

de classification, nous incluons notre approche spécifique pour modéliser la probabilité de dépassement du seuil des extrêmes $u = 130\,000€$.

Soit $\hat{y}$ la prédiction future. Notre procédure consiste à supposer que $\hat{y}$ peut être modélisé selon une distribution normale de moyenne $\hat{y}$ et d'écart type σ_y : $\mathcal{N}(\hat{y}, \sigma_y)$, avec σ_y est calculé en fonction de l'écart type des données d'entraînement. Nous pouvons ainsi estimer $p_i(x)$, la probabilité de dépassement pour un individu avec les caractéristiques x, en utilisant la fonction de distribution cumulative (CDF) de $\mathcal{N}\left(\hat{y}_i, \sigma_y^2\right)$:

$$p_i(x) = 1 - CDF_{\mathcal{N}\left(\hat{y}_i, \sigma_y^2\right)}(u) \tag{III.7}$$

Avec $u = 130\,000€$.

Dans un premier temps, nous avons cherché à estimer $\hat{y}$ à l'aide d'un modèle de régression. Pour cela, nous avons utilisé la régression des moindres carrés (OLS), la régression Ridge, la régression Lasso, la régression Elastic Net, la régression Gradient Boosting, les forêts aléatoires (RF) et les réseaux de neurones (NN). En utilisant le *Root Mean Squared Error* (RMSE), une mesure d'évaluation couramment utilisée pour évaluer la précision d'une régression, qui quantifie la racine de l'écart moyen entre les valeurs prédites par le modèle et les valeurs réelles de la variable cible (ou de la variable de réponse) dans un ensemble de données, nous avons sélectionné le meilleur modèle. Les résultats indiquent que le modèle RF présente le RMSE le plus faible et le R2 le plus élevé (voir figure III.17).

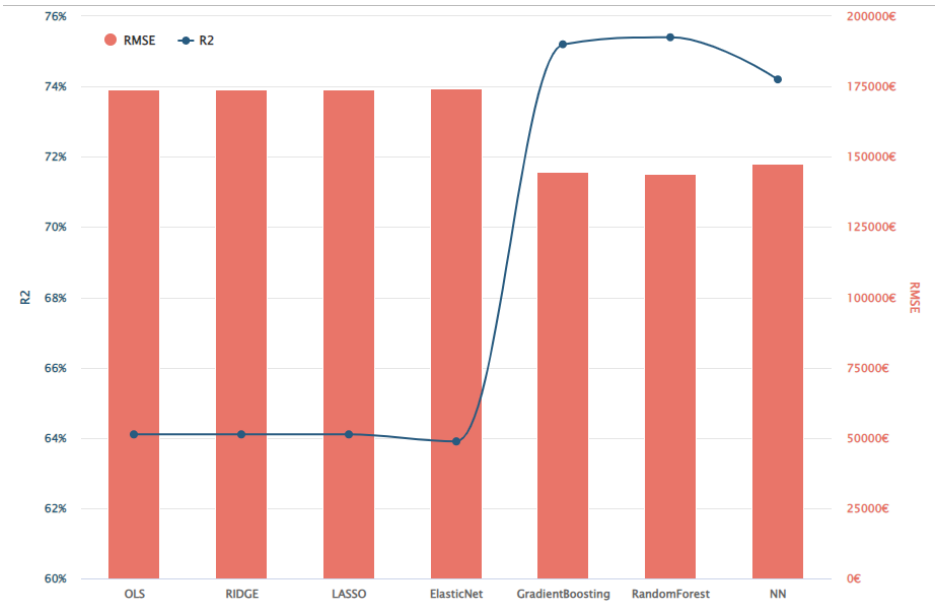


Figure III.17 – R2 et MSE pour le choix du modèle de régression.

Ensuite, nous avons hyperparametré le modèle de forêt aléatoire (RF) à l'aide du GridSearchCV avec CV=5 et calculer les probabilités p_i à l'aide de la formule III.7 (RF+CDFNormale).

Pour comparer notre procédure (RF+CDF Normale) avec d'autres modèles de classification (LR, DT, XG, SVM, NN, RF) on utilise la métrique de l'exactitude (*accuracy* en anglais). L'exactitude quantifie la proportion de prédictions correctes faites par le modèle

parmi toutes les prédictions effectuées. En d'autres termes, l'exactitude mesure la capacité du modèle à classer correctement les exemples d'un ensemble de données. Notre approche (RF+CDF Normale) présente la meilleure exactitude (87.91%) par rapport aux autres classificateurs (figure III.19).

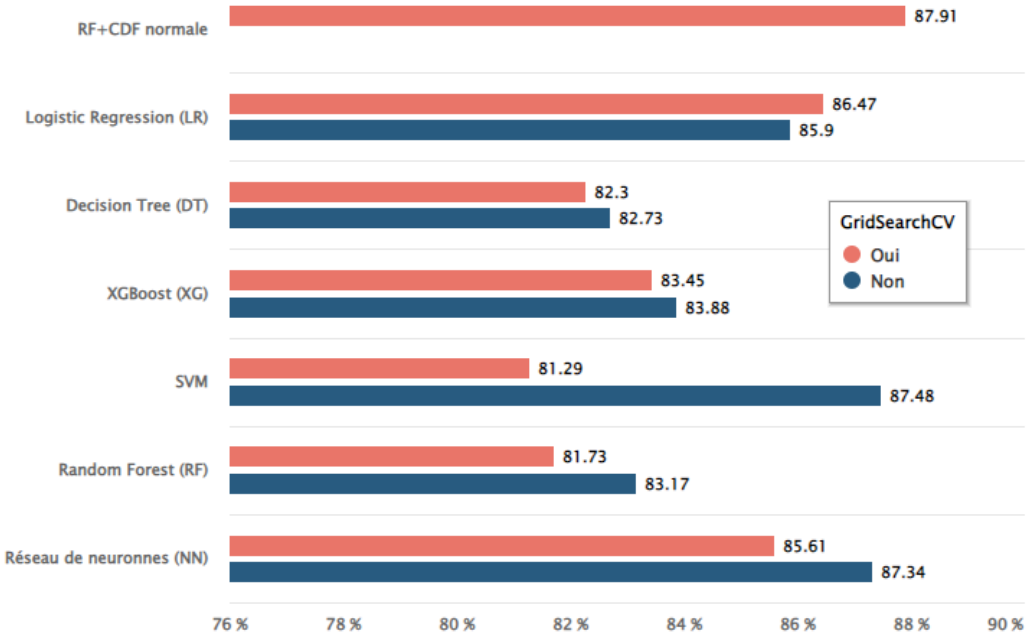


Figure III.18 – Exactitude de chaque modèle pour le choix du modèle de probabilité de dépassement du seuil u.

Ainsi, dans la projection du régime, nous utiliserons le modèle de forêt aléatoire (RF) hyperparamétré par GridSearchCV, combiné avec la fonction de répartition de la loi normale pour prédire les dépassements du seuil des extrêmes. Les paramètres du modèle final, hyperparamétré par GridSearchCV, sont les suivants : nombre d'arbres dans la forêt (*n_estimators*) : 20 ; profondeur maximale de l'arbre (*max_depth*) : 13 ; autres paramètres : défaut (scikit-learn).

Pour obtenir davantage de résultats concernant la modélisation de la probabilité de dépassement du seuil extrême u, notamment la description des modèles, la sélection des variables, les statistiques descriptives, le choix du seuil, et l'interprétation, le lecteur peut se reporter à la section 2 de l'annexe C.

Modélisation de la sinistralité attritionnelle

Nous avons utilisé un modèle de régression linéaire généralisé (GLM) pour modéliser la sinistralité attritionnelle $Y|Y < u$, en testant deux distributions potentielles : la loi gamma et la loi inverse gaussienne, toutes deux utilisant une fonction de lien logarithmique. Le tableau ci-dessous présente les critères AIC, BIC, et la déviance associés à chaque modèle :

Modèle	AIC	BIC	Déviance
Gamma	23774	24039	69.33
Inverse Gaussienne	24150	24350	0.0023

Table III.4 – Critères de comparaison des modèles pour la sinistralité attritionnelle

Les résultats du tableau ci-dessus laissent entrevoir que le modèle gamma est plus approprié pour la projection de la sinistralité attritionnelle. (Voir section 3 de l'annexe C pour le détail des résultats).

Modélisation de la sinistralité extrêmes

Pour les projections de la sinistralité extrême $Y|Y \leq$ nous avons utilisé l'arbre de pareto généralisée de la figure II.21 du chapitre 2. Ensuite dans chacune des 8 feuilles de cet arbre nous avons entraîner un modèle d'arbre de décision avec critère de split la perte quadratique en retenant que les variables présentes dans le GP CART. La RMSE dans chaque feuille et la RMSE totale sont présentés dans le tableau suivant :

Feuille (Leaf)	Leaf 1	Leaf 2	Leaf 3	Leaf 4	Leaf 5	Leaf 6	Leaf 7	Leaf 8	RMSE TOTAL
RMSE	19205.74	71329.66	90387.57	185564	90216.03	230299	272686	241770	209900.7
Poids	6.4%	2.5%	2.3%	3.2%	5.4%	2.9%	1.4%	75.9%	

Table III.5 – RMSE pour chaque feuille de l'arbre de pareto généralisée

Cette procédure permet d'obtenir une précision accrue par rapport à une méthode consistant à appliquer directement un arbre de décision avec une perte quadratique aux données extrêmes. En effet, l'utilisation de l'arbre de décision avec une perte quadratique directement sur les données extrêmes produit une RMSE (Root Mean Square Error) de 241388. Par conséquent, l'approche GP CART, suivie d'un arbre de décision avec une perte quadratique sur chaque feuille, offre une amélioration de précision d'environ 15% par rapport à la méthode consistant à appliquer directement un arbre de décision avec une perte quadratique aux sinistres extrêmes.

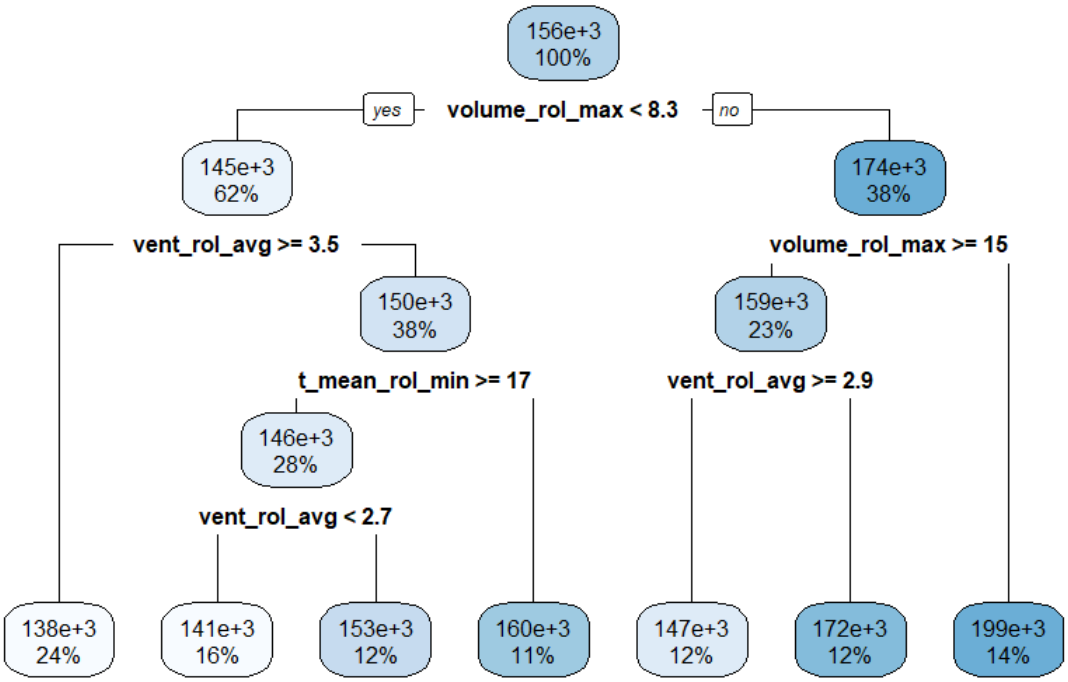


Figure III.19 – Arbre de décision avec perte quadratique sur les données de la feuille 8 du GP CART.

Une fois que tous les modèles sont entraînés et bien calibrés, on projette la sinistralité journalière pour chaque région à l'aide de la formule suivante (j : jour futur) :

$$\hat{S}(j) = \hat{\tau}(j) * [\hat{\delta} * \hat{Y}|Y < u + (1-\hat{\delta}) * \hat{Y}|Y \leq u] \tag{III.8}$$

Lorsque la sinistralité projetée (j) est strictement positive, le nombre d'assurée correspond à $\hat{S}(j)/50$. Pour obtenir le nombre d'assurés pour les jours où la sinistralité projetée est nulle, nous avons effectué une régression linéaire multiple sur les données historiques (dont la sinistralité est nulle) entre le nombre d'assurés journaliers et les variables suivantes : *saison*, *région*, *week-end*, *vacances*.

Le tableau III.6 ci-dessous présente les résultats de cette régression.

Résultats régression				
term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	4670.58***	114.02	40.96	0
Saison (ref = Automne)				
saisonété	1217.33***	84.30	14.44	4.45E-47
saisonHiver	-1376.11***	80.99	-16.99	2.22E-64
saisonPrintemps	-645.40***	79.89	-8.08	6.87E-16
Région (ref = Auvergne-Rhone-Alpes)				
name_regBourgogne-Franche-Comte	-4497.34***	142.69	-31.52	9.3E-214
name_regBretagne	-315.16*	161.11	-1.96	0.050457
name_regCentre-Val de Loire	-4897.96***	140.96	-34.75	3E-258
name_regCorse	-3802.71***	140.71	-27.03	1.6E-158
name_regGrand Est	-4595.01***	141.66	-32.44	5.5E-226
name_regHauts-de-France	-4794.27***	142.31	-33.69	3.1E-243
name_regIle-de-France	-5185.74***	141.67	-36.60	1.3E-285
name_regNormandie	-4040.50***	143.14	-28.23	1.6E-172
name_regNouvelle-Aquitaine	6660.85***	140.27	47.49	0
name_regOccitanie	7518.33***	139.76	53.79	0
name_regPays de la Loire	-249.19*	140.82	-1.77	0.07681
name_regProvence-Alpes-Cote d Azur	1084.95***	143.23	7.57	3.72E-14
Vacances (ref = non)				
vacance_scolaireoui	3331.66***	63.81	52.21	0
Week-end (ref = non)				
weekendoui	1069.94***	62.72	17.06	7.09E-65

(***p<0.01,**p<0.05,*p<0.1)

Table III.6 – Résultats de la régression linéaire multiple entre le nombre d'assurés journaliers et les variables suivantes : *saison*, *région*, *week-end*, *vacances*.

Les sinistres annuels sont obtenus en faisant la somme des sinistres projetés. En ce qui concerne les primes annuelles (P(n)) projetées, nous les obtenons en utilisant l'hypothèse concernant la grille tarifaire et le nombre d'assurés.

4.2.2 Résultats de la projection par l'approche modélisation

La figure III.20 montre la projection du ratio sinistre à prime (pure) du régime « beau temps » pour chaque scénario du GIEC retenu et selon l'approche par modélisation de 2023 à 2100.

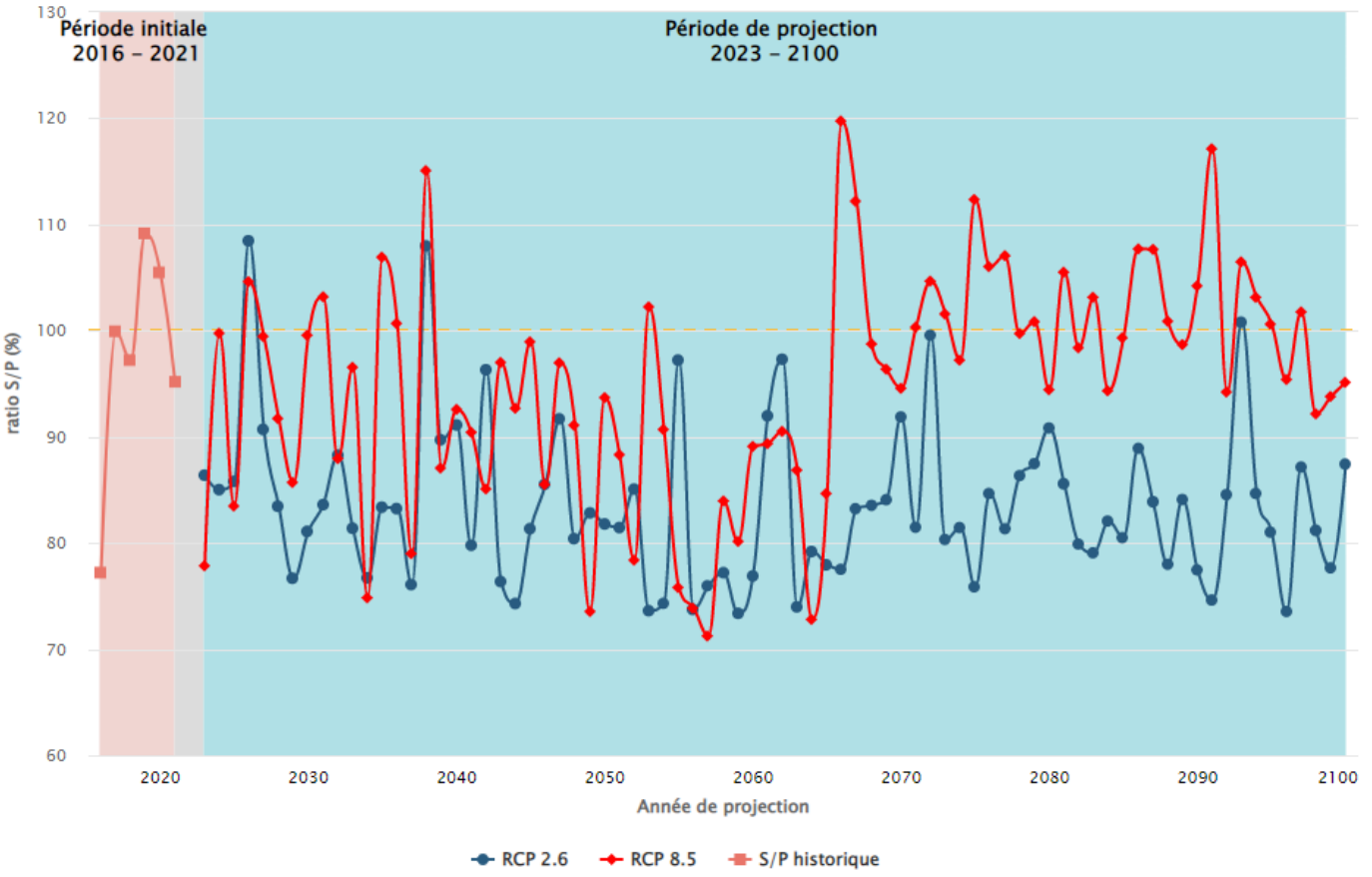


Figure III.20 – S/P projetés selon les scénarios du GIEC et selon l'approche par modélisation.

Comme c'est le cas pour les ratios S/P projetés par l'approche déterministe, ceux projetés par la méthode de modélisation présentent également une grande variabilité. Cette observation confirme le constat que nous avons fait : la forte volatilité des ratios S/P est en grande partie due à la variabilité climatique. Cependant, il est important de noter que quelle que soit la scénario du GIEC considéré, la volatilité historique semble être réduite. En effet, pour le scénario RCP 2.6, la moyenne des S/P est de 83.6% avec une volatilité de 7.48% (comparée à 11.2% pour les ratios historiques). En ce qui concerne le scénario RCP 8.5, la moyenne des S/P est de 95.1% avec une volatilité de 10.7%.

Un autre constat que nous pouvons tirer de l'analyse des ratios S/P projetés par l'approche de modélisation est qu'il n'y a plus de forte dégradation du ratio S/P en début de période, comme c'était le cas avec l'approche déterministe. Cette observation s'expliquerait principalement par le fait que l'approche de modélisation cherche à reproduire les tendances historiques observées, contrairement à l'approche déterministe. De plus, il est important de noter qu'au-delà de l'année 2063, dans le scénario 8.5, la situation devient très défavorable,

avec des ratios qui dépassent souvent 100% et qui sont plus élevés que ceux du scénario optimiste 2.6.

Sur l'évolution des réserves (figure III.21), on observe un résultat assez contre-intuitif. En effet, les réserves dans le scénario RCP 8.5 dépassent celles du scénario 2.6 à partir de 2085. Pour rappel, les réserves sont constituées par la différence entre la prime pure commerciale et le total des sinistres, soit $1.1 \times P(n) - S(n)$. Du fait de l'approche par modélisation, les primes totales annuelles dépendent de la sinistralité projetée. Ainsi, des sinistres élevés entraîneront des primes commerciales élevées ($1.1 \times P(n)$), et l'effet de cette hausse des primes commerciales semble dominer sur celui des sinistres.

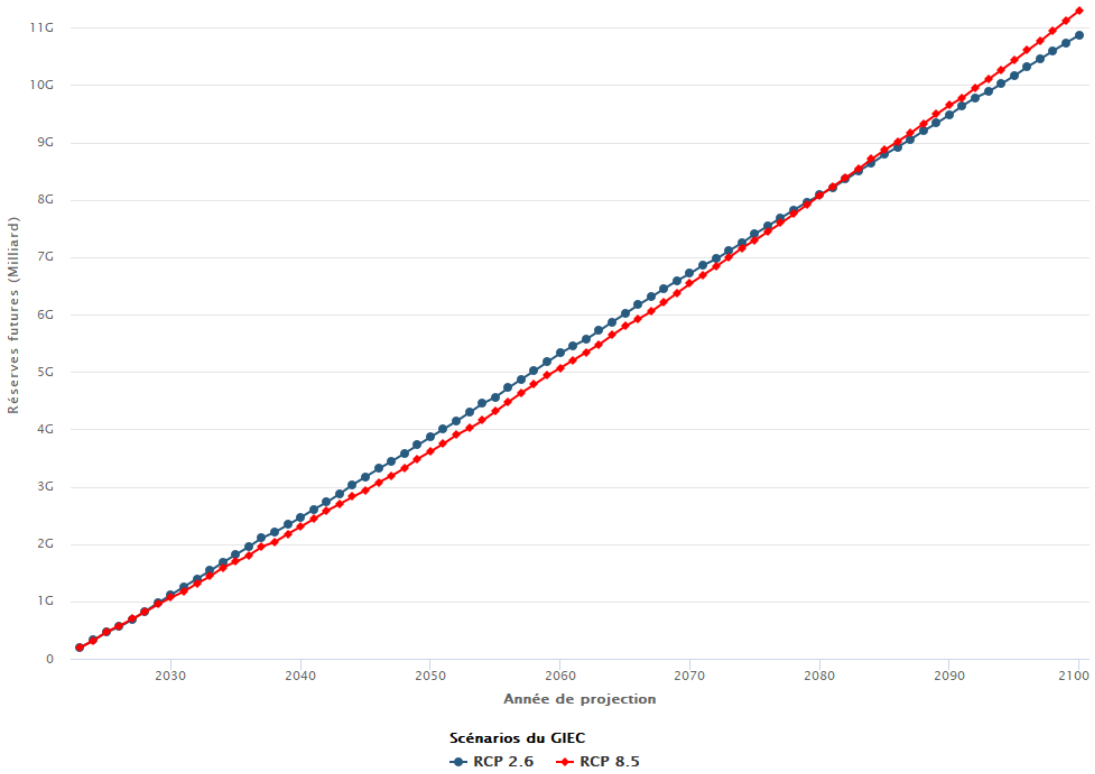


Figure III.21 – Réserves projetées selon les scénarios du GIEC par approche modélisation.

5 Étude de sensibilité et limites de l'étude

Dans cette dernière section du chapitre 3, nous examinerons la sensibilité des résultats présentés dans la section précédente par rapport aux hypothèses formulées. Comme toute étude, celle-ci présente des limites et points d'amélioration qui nécessitent d'être pris en compte pour une analyse plus précise et complète. Dans cette section, nous détaillons aussi les principales limites relevées et proposons des pistes d'amélioration possibles.

5.1 Sensibilité

Nous évaluons l'impact d'une modification des hypothèses d'entrée sur les résultats des projections selon les deux approches. Nous nous intéressons à l'impact de ces modifications

sur la moyenne des S/P et sur leur volatilité (écart-type).

Par rapport aux hypothèses sur la population assurée

On considère trois scénarios sur la population assurée. Le scénario 0 correspond à notre hypothèse de base concernant la population assurée. Pour rappel, cette hypothèse suppose une évolution géométrique des nuitées et considère que le nombre d'assurés est égal à 25% (part de marché) du nombre de nuitées et est utilisée seulement dans l'approche déterministe. On considère deux autres situations où la part de marché diffère, le premier scénario (scénario 1) prend une part de marché de 20%, tandis que le deuxième scénario (scénario 2) utilise une part de marché de 30%. Les résultats de ces différents scénarios pour la moyenne et l'écart-type du ratio S/P pour chaque scénario du GIEC sont présentés dans le tableau suivant :

Scénario sensibilité	**Scénario 0**		**Scénario 1**		**Scénario 2**	
Scénario du GIEC	***RCP 2.6***	***RCP 8.5***	***RCP 2.6***	***RCP 8.5***	***RCP 2.6***	***RCP 8.5***
E(S/P)	87.3%	91.9%	87.1%	91.7%	90.3%	105.3%
σ(S/P)	19.1%	19.3%	17.5%	17.9	20.9%	25.3%

Table III.7 – Impact de la modification de la part de marché sur les ratios S/P

On constat un gain assez faible sur la moyenne ratio S/P dans le scénario 1 (baisse de la part de marché), par rapport au scénario de base (scénario 0). Ce gain est de 0.2% dans les deux scénarios du GIEC (RCP 2.6 et 8.5). Par contre on note un gain assez significatif en terme de volatilité du ratio S/P (1.6% dans le RCP 2.6 et 1.4% dans le 8.5).

Contrairement au scénario 1, le scénario 2 (hausse de la part de marché de 25% à 30%) entraîne une perte de solvabilité et donc la probabilité de ruine du régime devient élevée par rapport au scénario 0. La hausse de la moyenne du ratio des S/P est plus élevée dans le scénario RCP 8.5 (13.4%) que dans le scénario 2.6 (3%). Les volatilités augmentent également dans les deux scénarios du GIEC. La perte de volatilité est de 1.8% dans le scénario RCP 2.6 et de 6% dans le scénario 8.5. Ainsi, une hausse de la part de marché entraîne une dégradation du ratio S/P accompagnée d'une forte volatilité des S/P.

Par rapport aux hypothèses sur la grille tarifaire

La grille tarifaire sert de base au calcul des primes annuelles totales. L'hypothèse sur la grille tarifaire est utilisée dans les deux approches de projection pour calculer les primes annuelles. Pour analyser l'impact d'une modification de cette grille tarifaire sur les résultats de la projection, nous considérons trois scénarios. Le scénario 0 correspond à l'hypothèse de base formulée sur la grille tarifaire (grille tarifaire de la figure III.10). Le scénario 1 correspond à la situation où la grille tarifaire du scénario 0 subit une baisse de 10% et le scénario 2 au cas où elle subit une hausse de 10%. Les résultats de ces différents scénarios pour la moyenne et l'écart-type du ratio S/P pour chaque scénario du GIEC sont présentés dans le tableau suivant :

Scénario sensibilité	Scénario 0		Scénario 1		Scénario 2	
Scénario du GIEC	*RCP 2.6*	*RCP 8.5*	*RCP 2.6*	*RCP 8.5*	*RCP 2.6*	*RCP 8.5*
Approche déterministe						
E(S/P)	87.3%	91.9%	97%	102.1%	79.4%	83.5%
σ(S/P)	19.1%	19.3%	21.2%	21.4%	17.4%	17.5%
Approche par modélisation						
E(S/P)	83.6%	95.1%	92.9%	105.7%	76%	86.5%
σ(S/P)	7.5%	11.2%	8.3%	12.4%	6.8%	10.2%

Table III.8 – Impact de la modification de la grille tarifaire sur les projections du ratio S/P

Une réduction d'un pourcentage de la grille tarifaire entraîne systématiquement une augmentation de la moyenne des ratios S/P ainsi qu'une augmentation de leur volatilité, quel que soit le scénario du GIEC ou l'approche considérés. En revanche, une augmentation de la grille tarifaire entraîne une diminution à la fois de la moyenne et de la volatilité des S/P. Cependant, une augmentation des tarifs peut être plus complexe à mettre en œuvre en pratique, car cela peut entraîner une perte de compétitivité.

5.2 Limites de l'étude

Comme toute étude, celle-ci présente certaines limites principales :

Tout d'abord, les résultats des projections doivent être interprétés avec précaution. En effet, nos deux approches retenues pour la projection ne tiennent pas compte d'un certain nombre de facteurs explicatifs pertinents. L'omission de ces variables pertinentes peut avoir des conséquences significatives sur les résultats de notre étude. Par exemple, l'information sur l'évolution de l'offre de camping en termes de capacité d'accueil, de nombre de lits, de nombre d'emplacements de camping et l'émergence de nouvelles offres de camping ne sont pas prises en compte dans notre étude. Une solution potentielle pour atténuer cette limite serait de collaborer avec des experts de l'industrie du camping ou de recourir à des données provenant d'organismes de tourisme ou d'associations professionnelles pour obtenir des informations à jour sur l'évolution de l'offre de camping. Ces données pourraient ensuite être intégrées dans nos modèles de projection pour améliorer la précision de nos projection.

Ensuite, la résolution spatiale et temporelle choisie peut être inadéquate pour une représentation précise des phénomènes associés aux risques couverts par le régime (température, vent et pluie). Pour une analyse efficace des phénomènes météorologiques extrêmes tels que les vagues de chaleur, les tempêtes et les vents violents, il est essentiel de sélectionner une granularité spatiale et temporelle adaptée. Dans le cadre de notre étude, nous avons opté pour une résolution régionale et un intervalle de temps hebdomadaire, qui pourrait ne pas suffire à une compréhension approfondie du phénomène. Au niveau spatial, la résolution régionale pourrait ne pas être assez fine pour prendre en compte les variations locales du climat, ce qui pourrait entraîner une sous-estimation ou une négligence des zones les plus exposées aux risques, impactant ainsi la fiabilité des prévisions. De plus, l'intervalle

de temps hebdomadaire pourrait ne pas être suffisamment précis pour saisir les variations à court terme du climat. Cela pourrait avoir un impact significatif sur nos projections des risques couverts, notamment en ce qui concerne les variations de température, de précipitations et de vent. En effet, les événements météorologiques extrêmes, qui peuvent se produire en quelques heures seulement, pourraient ne pas être pleinement pris en compte dans notre approche actuelle.

Enfin, une source d'incertitude découle de l'utilisation d'une seule trajectoire du GIEC pour chaque scénario. Étant donné que le portail DRIAS ne permet pas d'accéder à plusieurs trajectoires pour un même modèle climatique, notre capacité à effectuer des projections plus précises est limitée. Les projections que nous avons réalisées sont donc fortement dépendantes des trajectoires spécifiques choisies pour chaque scénario. Une amélioration de la précision des projections pourrait être obtenue en ayant accès à un ensemble plus large de sorties générées par le modèle climatique. Il serait également intéressant d'examiner les impacts du changement climatique sur le régime en utilisant des données simulées à partir d'autres modèles climatiques, ce qui permettrait de mieux évaluer la robustesse de nos résultats.

Conclusion

Somme toute, l'objectif principal de ce mémoire était de présenter un régime fictif nommé « beau temps » qui intervient en indemnisant une victime de mauvais condition météorologique du à la température, au vent ou encore à la pluie lors de son séjour de camping et d'évaluer les impacts éventuels du changement climatique sur ce régime

Dans le but d'atteindre cet objectif, nous avons tout d'abord initié la phase de conception du régime, ce qui a impliqué la création d'une grille tarifaire en conformité avec les règles du régime. L'élaboration de cette grille tarifaire a exigé l'utilisation de données concernant les séjours en camping en France métropolitaine, issues des enquêtes de l'INSEE, ainsi que des données fournies par Météo France. Pour garantir la fiabilité du régime, nous avons effectué une analyse du ratio S/P historique.

Par la suite, en examinant la sinistralité résultant de la mise en œuvre du régime, nous avons employé des outils de la théorie des valeurs extrêmes et des méthodes statistiques pour mettre en évidence la nature exceptionnelle de cette sinistralité, dépassant un seuil fixé à 130 000 euros. Cette nature exceptionnelle nous a conduit à adopter le modèle de l'arbre de régression de Pareto généralisée, une approche combinant la théorie des valeurs extrêmes et l'apprentissage statistique, permettant ainsi une meilleure prise en compte des valeurs extrêmes lors de la projection du régime "beau temps".

Dans un second temps, nous avons présenté deux approches (l'approche déterministe et l'approche par modélisation) qui nous ont permis d'évaluer l'impact du changement climatique sur le régime à l'horizon 2100, en fonction des scénarios du GIEC (RCP 2.6 et 8.5). Pour ce faire, nous avons utilisé les données de projection climatique disponibles sur le portail DRIAS.

De manière générale, on note des conclusions assez similaires entre les deux approches, notamment en ce qui concerne les tendances des indicateurs retenus, à savoir le ratio S/P et les réserves. Dans les deux approches, les ratios S/P projetés présentent une forte irrégularité, principalement due à la variabilité climatique des données, tandis que les réserves affichent une tendance plutôt linéaire.

En termes de différences, les ratios S/P sont plutôt sous-estimés avec l'approche par modélisation, tandis qu'avec l'approche déterministe, nous avons observé une forte dégradation du ratio S/P en début de période de projection (2023 - 2028). De plus, les résultats indiquent une situation globalement défavorable pour les données climatiques issues du scénario RCP 8.5 (pessimiste). Cette constatation est encore plus marquée avec l'approche par modélisation.

On pourrait explorer la potentialité pour une véritable compagnie d'assurance, cherchant à pénétrer le marché de l'assurance paramétrique pour le camping, d'utiliser nos résultats en tant que point de départ. Cette compagnie disposerait d'un capital initial avec un profil d'investissement défini. Dans ce contexte, des questions se posent quant à la structure de réassurance optimale à adopter, ainsi que sur le potentiel retour sur investissement que cela pourrait générer.

Bibliographie

M. Arthur et R. Christian. Tail-index partition-based rules extraction with application to tornado damage insurance. 2022. doi : Workingpaper.

C.-A. Azencott. Cours sur openclassrooms : Utilisez des modèles supervisés non linéaires.

N. Belaidi. Xgboost : Tout savoir sur le boosting. URL https://blent.ai/blog/a/xgboost-tout-comprendre.

N. V. Bowyer, Chawla et all. Smote : Synthetic minority over-sampling technique.

L. Breiman, J. Friedman, R. Olshen, et C. J. Stone. Classification and regression trees. 1984. doi : 10.2307/2530946.

F. Caeiro et M. I. Gomes. Threshold Selection in Extreme Value Analysis. In Extreme Value Modeling and Risk Analysis, pages 69–86. Chapman and Hall/CRC 2007, Jan. 2016. ISBN 978-1-4987-0129-7. doi : 10.1201/b19721-5.

A. C. Chavez-Demoulin, V. et Davison. Generalized additive modelling of sample extremes. Journal of the Royal Statistical Society : Series C (Applied Statistics), 54 (1) :207–222, 2005. ISSN 1467-9876. doi : 10.1111/j.1467-9876.2005.00479.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2005.00479.x. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2005.00479.x.

V. Chavez-Demoulin, P. Embrechts, et M. Hofert. An Extreme Value Approach for Modeling Operational Risk Losses Depending on Covariates. Journal of Risk and Insurance, 83(3) :735–776, 2016. ISSN 1539-6975. doi : 10.1111/jori.12059. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jori.12059. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/jori.12059.

N. V. Chawla. C4.5 and imbalanced data sets : Investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In Workshop on Learning from Imbalanced Data Sets II.

R. Christian. Cours de Théorie des Valeurs Extrêmes 2021-2022. Notes de cours ENSAE Paris, 2022.

S. Coles. An Introduction to Statistical Modeling of Extreme Values. Springer Science & Business Media, Aug. 2001. ISBN 978-1-85233-459-8. Google-Books-ID : 2nugUEaKqFEC.

A. Dalalyan. Cours d'Apprentissage statistique 2018-2019. Notes de cours ENSAE Paris, 2018.

A. C. Davison et R. L. Smith. Models for Exceedances Over High Thresholds. Journal of the Royal Statistical Society : Series B (Methodological), 52(3) :393–425, 1990. ISSN 2517-6161. doi : 10.1111/j.2517-6161.1990.tb01796.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1990.tb01796.x. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1990.tb01796.x.

e. a. Drummond. Class imbalance, and cost sensitivity : Why under-sampling beats over-sampling. Workshop on Learning from Imbalanced Data Sets II.

S. Farkas, A. Heranval, O. Lopez, et M. Thomas. Generalized Pareto Regression Trees for extreme events analysis. Dec. 2021a. URL https://hal.archives-ouvertes.fr/hal-03486564.

S. Farkas, O. Lopez, et M. Thomas. Cyber claim analysis using generalized pareto regression trees with applications to insurance. 98 :92–105, 2021b. ISSN 0167-6687. doi : 10.1016/j.insmatheco.2021.02.009. URL https://www.sciencedirect.com/science/article/pii/S0167668721000330.

fExtremes. Diethelm wuertz, tobias setz, yohan chalabi, maintainer tobias setz, and suggests runit. package 'fextremes'. 2009. URL https://github.com/cran/fExtremes.

F. W. Gerstengarbe et P. C. Werner. A method for the statistical definition of extreme-value regions and their application to meteorological time series. Zeitschrift fuer Meteorologie; (German Democratic Republic), 39 :4, Jan. 1989. URL https://www.osti.gov/etdeweb/biblio/5445028.

S. Gey et E. Nédélec. Model Selection for CART Regression Trees. HAL, 2005(0), 2005. URL http://dml.mathdoc.fr/item/hal-00326549/.

K. H. Gong J. Rhsboost : Improving classification performance in imbalance data. Computational Statistics and Data Analysis.

A. Géron. Deep Learning avec Keras et TensorFlow - 2e éd. - Mise en oeuvre et cas concrets : Mise en oeuvre et cas concrets. DUNOD, Paris, 2020. ISBN 978-2-10-079066-1.

H. He, Y. Bai, E. A. Garcia, et S. Li. Adasyn : Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pages 1322–1328, June 2008. doi : 10.1109/IJCNN.2008.4633969.

J. Laurikkala. Improving identification of difficult small classes by balancing class distribution. Springer Berlin Heidelberg.

V. Perchet. Cours de fondements mathématiques du machine learning.

J. Pickands. Statistical Inference Using Extreme Order Statistics. The Annals of Statistics, 3 (1) :119–131, 1975. ISSN 0090-5364. URL https://www.jstor.org/stable/2958083. Publisher : Institute of Mathematical Statistics.

G. E. B. R. C. Prati et M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, vol. 6, no. 1,.

A. Sabourin. Extreme Value Theory and Machine Learning. thesis, Institut polytechnique de Paris, Oct. 2021. URL https://hal.archives-ouvertes.fr/tel-03408958.

I. Tomek. Two modifications of cnn. IEEE Transactions on Systems Man and Communications SMC-6.

vacancesscolr. Pierre formont, antoine augusti et colin fay : Package vacances scolaires en france pour r. 2019. URL https://github.com/Tutuchan/vacancesscolr.

V. Vapnik. The Nature of Statistical Learning Theory. Springer Science & Business Media, June 2013. ISBN 978-1-4757-3264-1. Google-Books-ID : EqgACAAAQBAJ.

G. M. Weiss. Foundations of Imbalanced Learning. URL http://tools.ietf.org/html/draft-ietf-lisp-alt-07.

Annexe

A

Complément Théorique

1 Estimation des paramètres de la GPD

Les paramètres de la loi de Pareto Généralisée doivent être estimés non pas à partir de l'échantillon initial $X_1, \cdots, X_n$ mais à partir de l'échantillon des données supérieures à un seuil u correspondant à la k_n valeurs de l'échantillon initial. L'échantillon des excès sur lequel les paramètres sont estimés est noté :

$$\forall i \in [\![1; k_n]\!], \quad Y_i = X_{(n-k_n+i)} - X_{(n-k_n)}$$

Cette section présente trois méthodes pour l'estimation des paramètres de la loi de Pareto Généralisée à partir de l'échantillon des excès $Y_1, \cdots, Y_{k_n}$ supposés indépendants et de même loi $\mathscr{GPDPD}(\gamma, \sigma)$.

1.1 Méthode du maximum de vraisemblance

Considérons que les variables aléatoires $Y_1, \cdots, Y_{k_n}$ soient indépendantes et identiquement distribuées. Le théorème de Pickands, Balkema et De Hann permet de supposer que ces excès suivent une loi de Pareto Généralisée. La méthode du maximum de vraisemblance permet d'estimer la valeur des paramètres γ et σ qui maximise la fonction de log-vraisemblance suivante :

$$(y_1, \cdots, y_{k_n}) = \begin{cases} -k_n \ln(\sigma) - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{k_n} \ln\left(1 + \gamma \frac{y_i}{\sigma}\right) & \text{si } \gamma \neq 0 \\ -k_n \ln(\sigma) - \frac{1}{\sigma} \sum_{i=1}^{k_n} y_i & \text{si } \gamma = 0 \end{cases}$$

Les estimateurs du maximum de vraisemblance sont obtenus en déterminant la valeur des paramètres qui annule les dérivées partielles de la fonction de log-vraisemblance. Dans le cas où $\gamma = 0$, l'estimation du paramètre σ est donné par :

$$\hat{\sigma} = \frac{\sum_{i=1}^{k_n} y_i}{k_n}$$

Dans le cas où $\gamma \neq 0$, l'estimation des paramètres $\hat{\gamma}$ et $\hat{\sigma}$ s'obtient en résolvant le système d'équation suivant :

$$\begin{cases} -\frac{k_n}{\sigma} + \sum_{i=1}^{k_n} \frac{y_i(1+\gamma)}{\sigma^2+\sigma\gamma y_i} = 0 \\ \sum_{i=1}^{k_n} \ln\left(1+\frac{\gamma y_i}{\sigma}\right) + \gamma(1-\gamma)\sum_{i=1}^{k_n} \frac{y_i}{\sigma+\gamma y_i} = 0 \end{cases}$$

Les solutions de ce système ne sont pas explicites, des méthodes numériques itératives sont utilisées pour approcher la solution.

De plus, Smith (1984) a montré la normalité asymptotique de ces estimateurs dans le cas où $\gamma > -\frac{1}{2}$. La loi asymptotique des estimateurs est donnée par :

$$\sqrt{k_n}\begin{pmatrix} \hat{\sigma} \\ \hat{\gamma} \end{pmatrix} \xrightarrow{\mathscr{L}} \mathcal{N}\left(\begin{pmatrix} \sigma \\ \gamma \end{pmatrix}, \begin{pmatrix} 2\sigma^2(1+\gamma) & -\sigma(1+\gamma) \\ -\sigma(1+\gamma) & (1+\gamma)^2 \end{pmatrix}\right)$$

Ce résultat permettra dans la suite de construire des intervalles de confiance pour l'estimation des quantiles. Dans notre étude, c'est cette méthode d'estimation qui est retenue. Cependant deux autres méthodes régulièrement utilisées sont présentées. C'est la méthode des moments et la méthode des moments pondérés.

1.2 Méthode des moments

Cette méthode est valable uniquement dans les cas où $\gamma < \frac{1}{2}$. La méthode des moments consiste à égaler les deux premiers moments théoriques de la loi de Pareto Généralisée avec les deux premiers moments empiriques. Il suffit ensuite de résoudre le système de deux équations à deux inconnues pour exprimer les paramètres de la loi en fonction des moments empiriques. Dans le cas où une variable aléatoire X suit une loi de Pareto Généralisée de paramètres γ et σ, les deux premiers moments théoriques s'expriment pour tout $\gamma < \frac{1}{2}$ par :

$$m_1 = \mathbb{E}(X) = \frac{\sigma}{1-\gamma}$$
$$m_2 = \mathbb{E}\left(X^2\right) = \frac{2\sigma^2}{(1-\gamma)(1-2\gamma)}$$

En faisant coïncider ces moments théoriques avec les moments empiriques suivant :

$$\hat{m}_1 = \frac{1}{k_n}\sum_{i=1}^{k_n} y_i = \bar{Y}$$
$$\hat{m}_2 = \frac{1}{k_n}\sum_{i=1}^{k_n} y_i^2 = \hat{s}^2 + \bar{Y}^2$$

où $\hat{s}$ est l'écart-type de l'échantillon des excès. Il vient comme estimateur des moments des paramètres de la distribution de Pareto Généralisée, sous réserve que $\gamma < \frac{1}{2}$:

$$\hat{\gamma} = \frac{1}{2}\left[1 - \frac{\bar{Y}^2}{\hat{s}^2}\right]$$
$$\hat{\sigma} = \bar{Y}\left[\frac{1 + \frac{\bar{Y}^2}{\hat{s}^2}}{2}\right]$$

1.3 Méthode des moments pondérés

Cette méthode, présentée par Hosking et al. (1985) est valable uniquement dans les cas où $\gamma < 1$. Dans un premier temps, la notion de moments pondérés est définie.

Définition 4. Définition 5 (Moments pondérés). Soit Z une variable aléatoire de fonction de répartition F. Si Z est intégrable, le moment pondéré d'ordre $r \in \mathbb{N}$ et $s \in \mathbb{N}$ de Z est :

$$WM_Z(r,s) = \mathbb{E}[ZF^r(Z)(1-F(Z))^s]$$

Dans la suite, on considère $r = 0$. Le résultat suivant fournit la valeur du moment pondéré d'une variable aléatoire suivant une loi de Pareto Généralisée de paramètres γ et σ.

Théorème 3 (Moments pondérés d'une Pareto Généralisée). Soit Z une variable aléatoire de loi de Pareto Généralisée de paramètres γ et σ. Dans le cas où $\gamma < 1$, on a pour tout $s \in \mathbb{N}$:

$$WM_Z(0,s) = v_s = \frac{\sigma}{(s+1)(s+1-\gamma)}$$

Les paramètres de la loi de Pareto Généralisée s'expriment en fonction des moments pondérés d'ordre 0 et 1 par :

$$\sigma = \frac{2v_0v_1}{v_0 - 2v_1} \quad \text{et } \gamma = \frac{4v_1 - v_0}{2v_1 - v_0}$$

De la même façon que pour la méthode des moments classique, il est possible d'estimer v_0 et v_1. De manière générale, le moment pondéré d'ordre s, noté v_s peut être estimé par la quantité suivante :

$$\hat{v}_s = \frac{1}{k_n}\sum_{i=1}^{k_n} Z_{i,k_n}\frac{(k_n - i + s)!k_n!}{(k_n - i)!\,(k_n + s)!}$$

Une estimation des paramètres γ et σ par la méthode des moments pondérés s'obtient en remplaçant v_0 et v_1 par la valeur de leur estimateur $\hat{v_0}$ et $\hat{v_1}$.

2 Estimation semi-paramétrique de l'indice des valeurs extrêmes

Les trois méthodes décrites précédemment supposent que les variables aléatoires $(Y_i)_{1\leq i\leq k_n}$ soient distribuées selon H_θ. La famille de lois H_θ est complètement paramétrique.

Les estimateurs présentés dans cette section s'appliquent à des échantillons dont la distribution n'est plus H_θ, mais qui appartient à un certain domaine d'attraction de H. C'est pourquoi cette approche est qualifiée de semi-paramétrique. Les trois estimateurs présentés sont basés sur la statistique d'ordre $X_{(1)} \leq \cdots \leq X_{(n)}$, obtenue à partir de l'échantillon de sinistres initiaux en considérant les k_n plus grandes valeurs.

2.1 Estimateur de Hill

L'estimateur de Hill est l'estimateur le plus populaire pour estimer la valeur de l'indice des valeurs extrêmes γ. Cependant, il est valable uniquement pour les distributions appartenant au domaine d'attraction de Fréchet, c'est-à-dire dans le cas où $\gamma > 0$. Il se définit à partir d'un nombre d'excès considéré k_n par :

$$\hat{\gamma}_n^{(H)}(k_n) = \frac{1}{k_n}\sum_{i=1}^{k_n} \ln\left(X_{(n-i+1)}\right) - \ln\left(X_{(n-k_n)}\right)$$

Si k_n est choisi de sorte que $\lim_{\substack{\to+\infty \\ n\to+\infty}} k_n = +\infty$ et $\lim_{n\to+\infty} \frac{k_n}{n} = 0$, on peut montrer que l'estimateur de Hill est asymptotiquement gaussien :

De plus, on peut montrer que l'estimateur de Hill est construit à partir de l'estimateur du maximum de vraisemblance d'une loi de Pareto de paramètre α de sorte que la relation $\alpha = \frac{1}{\gamma}$ est vérifiée.

2.2 Estimateur de Pickands

L'estimateur de Pickands présente l'avantage d'être valable pour les trois domaines d'attraction, il n'y a pas de restriction sur la valeur de l'indice des valeurs extrêmes γ. Il se définit à partir d'un certain nombre d'excès considéré k_n par :

$$\hat{\gamma}_n^{(P)}(k_n) = \frac{1}{\ln(2)} \ln\left(\frac{X_{(n-k_n)} - X_{(n-2k_n)}}{X_{(n-2k_n)} - X_{(n-4k_n)}}\right)$$

3 Test d'adéquation

Il existe plusieurs tests statistiques permettant d'effectuer une adéquation (goodness of fit) des données à une loi usuelle. Dans cette sous-section nous nous limiterons au test de Kolmogorov-Smirnov, de Anderson –Darling, celui de Cramer Von et du khi-2.

3.1 Test de Kolmogorov-Smirnov (KS)

Le test de Kolmogorov-Smirnov consiste à mesurer, pour une variable aléatoire continue, la plus grande distance entre la distribution théorique F_0 et la distribution empirique F_n. Soit $X_1, \ldots, X_n$, un échantillon de loi F_X et soit F_0 une loi continue. On souhaite tester :

$$\begin{cases} H_0 : & F_X = F_0 \\ H_1 : & F_X \neq F_0 \end{cases}$$

La statistique de Kolmogorov-Smirnov est définie par :

$$KS_n = \sup_x |F_n(x) - F_0(x)|$$

qui peut aussi s'écrire :

$$KS_n = \max\left\{KS_n^+, KS_n^-\right\}$$

avec :

— $KS_n^+ = \max\left(\frac{i}{n} - F_0(X_i), 0\right)$

— $KS_n^- = \max\left(F_0(X_i) - \frac{i-1}{n}, 0\right)$

Où X_i est la i-ème observation de l'échantillon et n le nombre total d'observations.

Le test au seuil α associé à cette statistique est défini par la région critique de la forme :

$$\{KS_n \geq c_\alpha\}$$

où c_α est le quantile $(1-\alpha)$ de la table de Kolmogorov-Smirnov, appelé aussi la valeur critique. Quand la statistique KS_n est supérieure à la valeur critique, nous rejetons H_0 au niveau de α. Dans le cas contraire, nous ne pouvons pas rejeter H_0. La valeur critique c_α pour un test avec F_0 continue et bien spécifié se calcule de la façon suivante :

$$\frac{\sqrt{-0.5 \times \ln\left(\frac{\alpha}{2}\right)}}{\sqrt{n}}$$

où n est la taille de l'échantillon testé. En pratique, α est souvent spécifié à 5%. La valeur critique ne dépend donc pas de la distribution théorique spécifiée, ce qui est un grand avantage du test. Cependant, elle dépend de la taille de l'échantillon, c'est-à-dire que plus l'échantillon est petit, plus la valeur critique est grande, qui permet d'accepter H_0 avec une statistique KS_n (distance) plus grande.

3.2 Test d'Anderson-Darling

Le test d'Anderson Darling est une alternative du test de Kolmogorov-Smirnov, il repose également sur le calcul d'une distance entre la fonction de répartition empirique de l'échantillon et la fonction de répartition théorique testée. Il peut également être vu comme un cas particulier du test de Cramér-von Mises. En effet, ce test pondère plus les queues de distribution, comme nous le verrons plus bas. On veut tester si la loi P des observations a la même fonction de répartition F qu'une loi continue donnée. H_0 = {la distribution observée n'est pas significativement différente de la distribution théorique } contre H_1 = {la distribution observée est significativement différente de la distribution théorique }

La statistique du test d'Anderson Darling s'exprime comme :

$$A_n^2 = \int_{-\infty}^{+\infty} (F_n(x) - F_0(x))^2 \, w(X) dF_0(x)$$

Avec w une fonction de pondération telle que $w(X) = \frac{1}{F_0(x)(1-F_0(x))}$. Cette fonction permet de pondérer davantage les queues de distribution.

Cette statistique s'exprime aussi à l'aide de l'échantillon ordonné $X_{(1)} \leq \cdots \leq X_{(n)}$, par la formule suivante :

$$A_n^2 = -n + \sum_{i=1}^{n} \frac{(2i-1)}{n} \left(\ln\left(F_0\left(X_{(i)}\right)\right) + \ln\left(1 - F_0\left(X_{(n+1-i)}\right)\right)\right)$$

La région critique du test au seuil α est définie par : $\left\{A_n^2 \geq c_\alpha\right\}$ où c_α est le quantile $(1-\alpha)$ de la table d'Anderson-Darling. Comme pour la statistique KS_n du test de Kolmogorov-Smirnov, la loi de A_n^2 est indépendante de F_0 sous l'hypothèse nulle H_0, ce qui permet de la tabuler.

La statistique d'Anderson Darling prend en compte l'étendue de la distribution et notamment la queue de distribution relative aux montants des sinistres les plus importants. Le test d'Anderson-Darling accorde une importance plus grande aux queues de distribution, il est naturellement plus approprié pour notre problématique de modélisation de la sévérité. En cas de conclusion contradictoire entre le test de Kolmogorov-Smirnov et Anderson Darling, c'est la conclusion de ce dernier qui est préférée.

3.3 Test de Cramer Von Mises

Le test de Cramér-von Mises peut être vu comme une version plus puissante du test de KolmogorovSmirnov. Les deux tests se basent sur l'écart entre les deux fonctions de répartition que nous souhaitons comparer, la seule différence étant que le second se base sur la valeur maximale de cette différence tandis que le premier se base sur la somme des différences. Ainsi, si le test de Kolmogorov-Smirnov est sensible aux valeurs aberrantes, le second l'est beaucoup moins.
Hypothèses : On veut tester si la loi P des observations a la même fonction de répartition F qu'une loi continue donnée. $H_0 =$ { la distribution observée n'est pas significativement différente de la distribution théorique contre $H_1 =$ {la distribution observée est significativement différente de la distribution théorique}

Soient $X_1 < \cdots < X_n$ n observations ordonnées. Alors leur fonction de répartition empirique est de la forme :

$$F_{emp}(x) = \begin{cases} 0 & \text{si } x < X_1 \\ \frac{i}{n} & \text{si } X_i \leq x \leq X_{i+1} \\ 1 & \text{sinon} \end{cases}$$

Statistique de test : On utilise la statistique :

$$t_n = \sqrt{n \int_{-\infty}^{\infty} \left[F_{emp}(X) - F(X)\right]^2 dF(X)}$$

Graphiquement, cette statistique peut être vue comme l'aire entre les deux fonctions de répartitions. Dans notre cas, nous utilisons la forme discrète simplifiée suivante :

$$T_n = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{F(X_i) - (2i-1)}{2n}\right]^2$$

Détermination du seuil critique :

$$\alpha = \mathbb{P}\left[\text{ rejeter } H_0 \text{ quand } H_0 \text{ vraie }\right] = \mathbb{P}\left[T_n > \text{ h quand } H_0 \text{ vraie }\right] = \mathbb{P}\left[T_n > h\right]$$

$\boldsymbol{\alpha}$ étant donné, on en déduit la valeur de h. Décision : Les valeurs X_i ayant été observées sur l'échantillon, et les valeurs $F(X_i)$ calculées, il ne reste plus qu'à comparer d avec h pour prendre une décision en fonction de la règle de décision.

- $T_n > h$ on rejette H_0 avec un risque α de se tromper
- $T_n < h$ on ne rejette pas H_0

3.4 Test du khi-deux d'ajustement à une loi donnée

Soit une variable X observée sur n individus, soit Ω l'espace des observations divisé en K catégories.Est il vraisemblable de considérer que ces observations constituent un échantillon tiré d'une population de distribution donnée ? En d'autres termes, la loi empirique constatée de X peut elle être assimilée à une loi théorique connue ?

Soit $P_{i,\ i=1,2,...,k}$ les probabilités de classe obtenues à partir de la loi théorique et n_i les effectifs de classe. La distance du khi-deux entre la loi théorique et la loi empirique est donnée par :

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - nP_i)^2}{nP_i} = n \sum_{i=1}^{K} \frac{(\hat{P}_i - P_i)^2}{P_i}, \hat{P}_i = \frac{n_i}{n}$$

De plus, $\chi^2 \mapsto \chi^2(K-1)$

Le test d'ajustement se présente comme suit :

$$\begin{cases} Ho : \text{La loi empirique est la loi théorique} \\ H1 : \text{La loi empirique est différente de la loi théorique} \end{cases}$$

La région critique W du test est donnée par :

$\{W = (n_1, n_2, ..., n_k) \setminus \chi^2 > \chi^2_{lu}\}$

4 Modèles additifs généralisés (GAM)

Le modèle additif généralisé est un modèle statistique développé par Trevor Hastie et Rob Tibshirani pour fusionner les propriétés du modèle linéaire généralisé avec celles du modèle additif. Tout comme le GLM, ce modèle permet d'adapter la variable à expliquer aux lois de la famille exponentielle. Cependant, à l'instar du modèle additif, il ne présuppose pas que la relation entre les variables explicatives et la réponse soit linéaire. Dans cette section, nous présenterons dans un premier temps, les fondements théoriques du modèle additif et dans un second temps, le modèle additif généralisé dans le cadre de la loi de pareto (GAM GP).

4.1 Brève présentation du fondement théorique du GAM

Les modèles additifs sont une classe de modèles statistiques qui reposent sur le principe fondamental de l'additivité. Ils sont largement utilisés pour modéliser des relations complexes entre une variable à expliquer (la réponse) et un ensemble de variables explicatives (les prédicteurs) en tenant compte de la non-linéarité et de l'interaction entre les prédicteurs. Plus formellement, les GAM les plus simples peuvent s'écrire de la manière suivante :

$$g\left(\mathbb{E}\left[Y_i \mid X_i\right]\right) = \sum_{j=1}^{p} f_j\left(X_{ij}\right) \tag{A.1}$$

où la fonction de lien g et la distribution de probabilité sous-jacente des observations $(Y_i)_i$ sont les deux caractéristiques inchangées de l'équation de la forme générale d'un GLM. Seule la structure linéaire est modifiée pour remplacer les coefficients β_j par des fonctions non linéaires f_j spécifiques à chaque prédicteur, ce qui rend les GAM semi-paramétriques. Sous condition que les fonctions f_j aient une forme paramétrique spécifié e^8, comme ce sera le cas dans cette étude, la méthode de maximisation de la vraisemblance convient toujours dans cette situation.

Néanmoins, le choix des fonctions f_j, qui déterminent le caractère non linéaire du modèle, constitue une véritable problématique. Ces fonctions peuvent être sélectionnées parmi un ensemble de transformations classiques : les fonctions polynomiales, les fonctions en escaliers et les splines.

L'estimation des paramètres dans les modèles additifs repose souvent sur des méthodes de moindres carrés pondérés ou des méthodes de maximisation de vraisemblance. L'une des principales forces des modèles additifs réside dans leur capacité à interpréter les effets de chaque prédicteur de manière indépendante. Cela signifie que vous pouvez examiner l'effet de chaque prédicteur sur la réponse tout en maintenant les autres prédicteurs constants. Les modèles additifs sont particulièrement utiles lorsque la relation entre les prédicteurs et la réponse est non linéaire. Les fonctions lisses permettent de capturer ces relations complexes.

Les modèles additifs peuvent être adaptés à une grande variété de situations, notamment la régression, la classification, et la modélisation de survie. Ils peuvent également gérer des prédicteurs catégoriels et numériques. Comme pour tous les modèles statistiques, il est essentiel de valider un modèle additif pour s'assurer qu'il est approprié pour les données. Cela peut être fait en utilisant des techniques de validation croisée, des graphiques de diagnostic, et d'autres méthodes.

4.2 Modèle additif généralisé pareto généralisée (GAM GP)

Chavez-Demoulin et al. (2016) ont proposé un cadre de modèle additif généralisé (GAM) pour les paramètres de distribution de valeurs extrêmes qui pourrait s'adapter aux termes paramétriques et lisses.

Le cadre du GAM pareto généralisée proposé par Chavez-Demoulin et al. (2016) prédit les paramètres de pareto généralisé (ξ, σ) par la modélisation de chaque paramètre par un modèle additif généralisé (GAM). Ils ont étudié en détail $\theta_{GAMGP}(x) = (\xi(x), \nu(x))$, qui est une reparamétrisation d'une loi de pareto généralisée, avec ξ le paramètre de forme et $\nu = log((1+\xi)\sigma)$. Il suppose ξ et ν on la forme suivante :

$$\xi = f_\xi(x) + h_\xi(t)$$

$$\nu = f_\nu(x) + h_\nu(t)$$

où f_ξ, f_ν représentent les fonctions dans les niveaux de facteurs de la variable x et $h_\xi, h_\nu : [0, T] \to \mathbb{R}$ sont des fonctions mesurables générales.

Annexe

B

Paramètres $\alpha_{we} = 1.33$ et $\alpha_{vacs} = 2.74$.

Le dispositif valeur cible permet de trouver la valeur d'un paramètre servant à réaliser une formule. La procédure de la valeur cible sur *Excel* se résumer de la manière suivante :

— on accède à l'option « Analyse de scénario » dans le menu « Données » et on sélectionne la valeur cible ;
— on choisit la cellule à définir et on indique la valeur à atteindre ;
— on sélectionne la cellule à modifier et on obtient le résultat recherché.

Pour rappel, α_{we} et α_{vacs} représentent respectivement la fréquentation supplémentaire en week-end versus hors week-end et la fréquentation supplémentaire en vacance versus hors vacance.

Pour obtenir le paramètre α_{vacs}, nous avons trouvé deux années (en occurrence 2015 et 2016) et un mois (octobre) pour lequel le nombre de jours vacances de ce mois est supérieur dans une des années (2015) mais le nombre de jours hors vacances est supérieur pour l'autre année (2016).

La figure B.1 présente les paramètres d'entrée pour la valeur cible. Comme valeur d'entrée, nous avons le nombre de nuitée (en milliers) du mois, le nombre de jours dans le mois, le nombre de jours vacances et le nombre de jours hors vacances scolaires.

	A	B	C	D
1				
2		2015 (1)	2016 (2)	(2)/(1)
3	**Nuitées octobre**	5866	5443	93%
4	**Nb jours octobre**	31	31	100%
5	**Nb jours vacances oct**	10	8	80%
6	**Nb jours hors vacances oct**	21	23	110%
7				
8	*Nuitées/jours vacs*	332,401643	*Valeur cible pour 1 en D12*	
9				
10	**Nuitées vacances**	3324,01643	2659,21315	
11	**Nuitées hors vacances**	2541,98357	2783,78685	
12	**Nuitées/jours hors vacs**	121,046837	121,034211	0,9998957
13				
14	**Fréquentation supplémentaire en vacs vs hors vacs**	2,74605808		

Figure B.1 – Paramètres d'entrée, de sortie et résultat de la valeur cible pour le paramètre α_{vacs}

Identifier deux années où, pour un mois donné, le nombre de jours de vacances est supérieur dans l'une des années tandis que le nombre de jours hors vacances est supérieur dans l'autre permet de garantir l'obtention d'un paramètre α_{vacs} supérieur à 1. En effet, il est parfaitement raisonnable de supposer qu'au cours d'un mois donné, il y aura plus de nuitées pendant les jours de vacances et moins pendant les jours sans vacances. Par conséquent, il est justifié d'accorder plus de poids aux jours de vacances par rapport aux jours hors vacances scolaires dans le calcul de l'équivalence en nuitée.

Comme paramètres de sortie, nous avons le nombre de nuitées par jour de vacances (B8), le nombre de nuitées en période de vacances (produit du nombre de nuitées par jour de vacances et du nombre de jours de vacances scolaires), le nombre de nuitées en période de hors vacances scolaires (différence entre le nombre de nuitées et le nombre en période de vacances scolaires), ainsi que le nombre de nuitées par jour hors vacances scolaires.

La valeur cible est définie par le rapport du nombre de nuitées par jour hors vacances scolaires des deux années (cellule D12). La valeur à atteindre est de 1. La valeur à modifier est la cellule B8, qui correspond aux nuitées par jour de vacances scolaires. Ainsi, en appliquant la procédure de la valeur cible, on obtient le nombre de nuitées par jour de vacances qui permettra d'atteindre la cible. Le paramètre α_{vacs} est alors obtenu en calculant le rapport entre le nombre de nuitées par jour de vacances et le nombre de nuitées par jour hors vacances scolaires de l'année qui présente le nombre de jours de vacances scolaires le plus élevé pour le mois considéré. On obtient donc $\alpha_{vacs} = 2.74$.

En appliquant au paramètre α_{we} la même démarche que pour le paramètre α_{vacs}, nous avons obtenu $\alpha_{we} = 1.33$.

	A	B	C	D
19				
20		2014	2015	(1)/(2)
21	**Nuitées juin**	12450	12324	99%
22	**Nb jours juin**	30	30	100%
23	**Nb jours we juin**	9	8	89%
24	**Nb jours hors we**	21	22	105%
25				
26	*Nuitées/jours we*	502,921535	*Valeur cible pour 1 en D31*	
27				
28	**Nuitées we**	4526,29381	4023,37228	
29	**Nuitées hors we**	7923,70619	8300,62772	
30	**Nuitées/jours hors we**	377,319342	377,30126	0,99995208
31				
32	**Fréquentation supplémentaire en we vs hors we**	1,33288034		

Figure B.2 – Paramètres d'entrée, de sortie et résultat de la valeur cible pour le paramètre α_{we}

Annexe

C

Résultats de la modélisation

1 Prédiction de la survenance du déclenchement τ

Dans cette section, nous allons examiner en détail la modélisation de la survenue de déclenchements τ. Il est important de noter qu'il y a très peu de déclenchements dans la base de données initiale (environ 12%), ce qui crée un déséquilibre significatif entre les classes. Pour commencer, nous ferons une brève revue de la littérature concernant les problèmes de déséquilibre de classes dans le contexte de la classification binaire. Ensuite, nous présenterons les résultats de la modélisation de τ et enfin, nous interpréterons le meilleur modèle que nous aurons sélectionné.

1.1 Problèmes de classification déséquilibrée

Les algorithmes de classification fonctionnent mieux lorsque le nombre d'individus de chaque classe est à peu près égal. Lorsque le nombre d'instances d'une classe dépasse de loin celui de l'autre, des problèmes surviennent. Dans les cas déséquilibrés, les algorithmes de classificateur standard ont un biais en faveur des classes qui ont un grand nombre d'instances (Weiss, Drummond et Chawla). Ils ont tendance à prédire uniquement les données de la classe majoritaire. Les caractéristiques de la classe minoritaire sont traitées comme du bruit et sont souvent ignorées. Il existe donc une forte probabilité de classification erronée de la classe minoritaire par rapport à la classe majoritaire.

Dans la littérature, de nombreuses méthodes ont été développées pour surmonter le problème du déséquilibre des classes (Weiss). Ces méthodes peuvent être classées en deux catégories générales, à savoir les **méthodes d'échantillonnage** et les **classificateurs *skew-insensitive*** (Weiss et ,Drummond). Dans le cadre de ce mémoire, nous nous sommes intéressé uniquement aux méthodes d'échantillonnage.

Les méthodes d'échantillonnage peuvent être classées en trois groupes, à savoir :

— **sous-échantillonnage de la classe majoritaire** : la principale méthode de ce groupe est le *random undersampling* (RUS). Le but de cette méthode est d'équilibrer la répartition des classes grâce à l'élimination aléatoire des observations de la classe majoritaire. L'inconvénient majeur de la méthode du *random undersampling* est qu'elle peut éliminer des données potentiellement utiles qui pourraient être importantes pour la prédiction (Weiss et Chawla). Afin de résoudre ce problème, diverses techniques ont été développées, qui visent à conserver toutes les informations utiles présentes dans

la classe majoritaire en supprimant les instances redondantes ou bruyantes. Une des premières méthodes alternatives à RUS est due à *Ivan Tomek* (Tomek). Cette méthode est connue sous le nom de *Tomek Link* (T-Link). Elle est considérée comme une amélioration de la règle du plus proche voisin (KNN). *Tomek Link* peut être défini comme suit : étant donné deux observations E_i et E_j appartenant à des classes différentes, et $d(E_i, E_j)$ est la distance entre E_i et E_j. Une paire (E_i, E_j) est appelée un *Tomek Link* s'il n'y a pas d'observation E_l, tel que $d(E_i, E_l) < d(E_i, E_j)$ ou $d(E_j, E_l) < d(E_i, E_j)$. Si deux observations forment un *Tomek Link*, alors l'un de ces observations est du bruit ou les deux. Une autre méthode de sous-échantillonnage est appelée *Neighborhood Cleaning Rule* (NCR) et a été développé par Laurikkala. Cette technique utilise la règle *Edited Nearest Neighbor* (ENN) de Wilson pour sélectionner les instances de la classe majoritaire à supprimer de l'ensemble des données.

— **sur-échantillonnage de la classe minoritaire** : dans cette catégorie , trois méthodes sont fréquemment mentionnées dans la littérature (Weiss, Drummond). Il s'agit de la méthode du *random oversampling* (ROS), de la méthode de Sur-échantillonnage Synthétique de la Classe Minoritaire (SMOTE [1])(Bowyer and all.), et de l'Echantillonnage Synthétique Adaptatif (ADASYN [2])(He et al. (2008)). Le *random oversampling* est une méthode visant à équilibrer la distribution des classes en répliquant aléatoirement des observations de la classe minoritaire. Cette technique souffre de la perte d'informations potentiellement utiles et du problème de sur-ajustement. Pour remédier à ce problème, Chawla et al. ont développé une méthode pour créer des instances synthétiques au lieu de simplement copier les instances existantes dans l'ensemble de données. Cette technique est la Technique de Sur-échantillonnage Synthétique de la Classe Minoritaire (SMOTE). La méthode SMOTE utilise les k-plus proches voisins pour générer de nouvelles instances en fonction de la distance entre les données de la classe minoritaire et certains voisins les plus proches sélectionnés au hasard. L'algorithme de SMOTE est décrit par [1]. ADASYN est basé sur l'idée de générer de manière adaptative des échantillons de données de la classe minoritaire en fonction de leurs distributions : davantage de données synthétiques sont générées pour les échantillons de la classe minoritaire plus difficiles à entraîner par rapport à ceux qui sont plus faciles à entraîner. Ainsi, dans la méthode ADASYN, nous tenons compte d'une distribution de densité qui détermine le nombre d'échantillons synthétiques à générer pour un point donné, tandis que dans SMOTE, il y a un poids uniforme pour tous les points de la classe minoritaire. [2] fait la description de l'algorithme d'ADASYN.

— **Méthodes hybrides** : Des techniques combinant le sur-échantillonnage et le sous-échantillonnage ont été développées. Le principal avantage cette combinaison est que l'ensemble de données peut être équilibré en ne perdant pas trop d'informations (c'est-à-dire en sous-échantillonnant trop d'instances de classe majoritaire), ni

1. Anglais : Synthetic Minority Oversampling Technique
2. Anglais : Adaptive Synthetic sampling

en souffrant d'un sur-échantillonnage (c'est-à-dire trop de sur-échantillonnage). Ces techniques sont appelées **méthode hybride**. Deux exemples de techniques hybrides qui ont été développées incluent *SMOTE+Tomek* et *SMOTE+ENN*(Prati and Monard.), dans lesquels SMOTE est utilisé pour sur-échantillonner la classe minoritaire, tandis que Tomek et ENN, respectivement, sont utilisés pour le sous-échantillonner la classe majoritaire.

L'inconvénient le plus important des méthodes d'échantillonnage est qu'elles n'ont aucun effet sur les cas de rareté absolue impliquant à la fois des classes rares et des cas rares. Il a également été suggéré que, puisque les méthodes d'échantillonnage équilibrent artificiellement les données et créent une nouvelle distribution, la distribution sous-jacente reste toujours déséquilibrée. Une autre critique est que les approches d'échantillonnage ne fonctionnent que sur le déséquilibre entre les classes et non sur le déséquilibre au sein des classes.

1.2 Cadre méthodologique

1.2.1 Description de la méthodologie

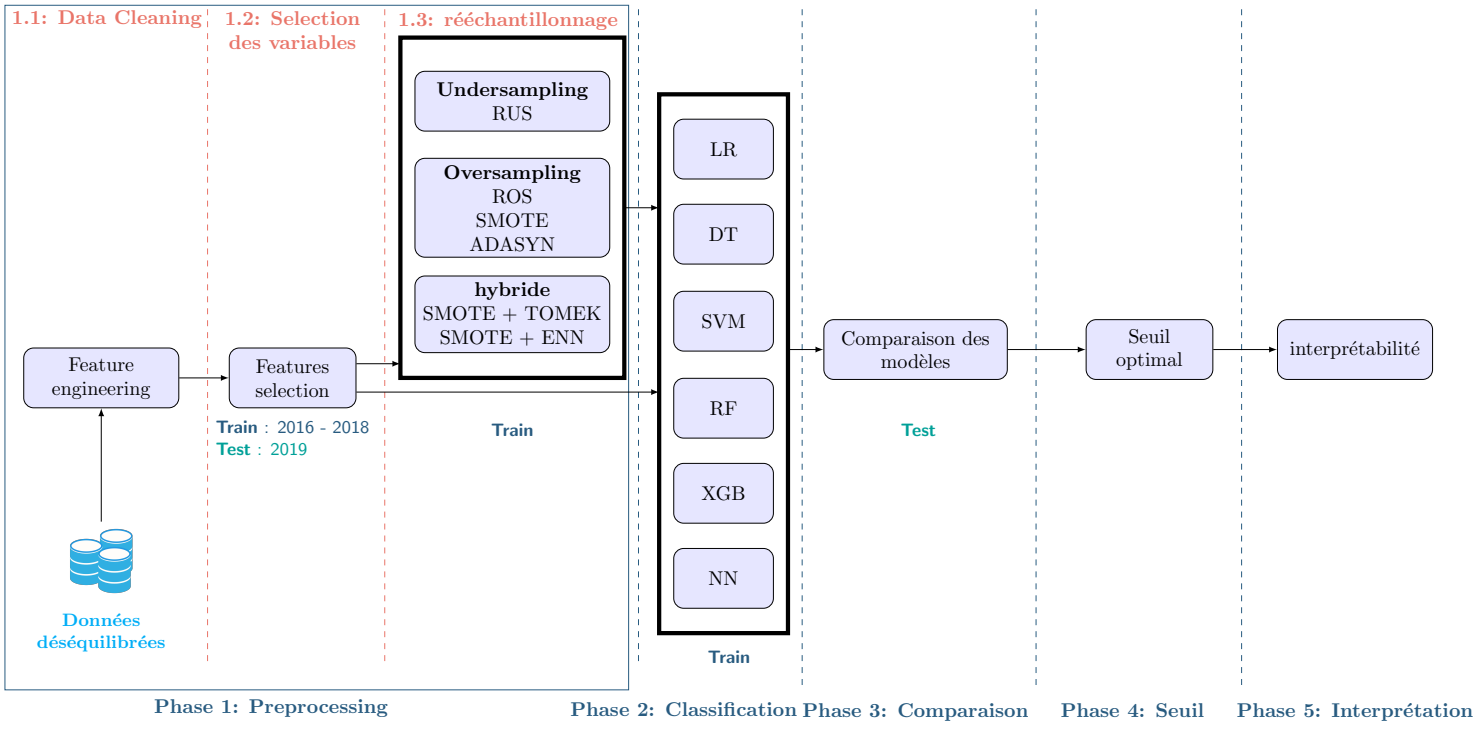


Figure C.1 – Diagramme résumant l'approche de modélisation du déclenchement d'un paiement τ par le régime « beau temps ».

La figure C.1 ci-dessous illustre les différentes étapes de la modélisation pour parvenir à prédire la survenance d'un paiement. La première phase de la méthodologie est le **prétraitement** ou encore *preprocessing* en anglais. Nous prenons d'abord les données brutes déséquilibrées et les transmettons à travers une étape de *feature engineering* ou d'ingénierie des fonctionnalités en français pour nettoyer les données. Après cela, les données traitées

sont divisées en données d'apprentissage (*Train*) et en données de test. Nous prenons les données des années 2016 à 2018 comme données d'apprentissage et les données de l'année 2019 comme données test. Nous n'avons pas intégré les années 2020 et 2021 en raison de la crise de la pandémie de Covid-19 qui a eu un impact sur le régime. Ensuite, nous passerons à la sélection des variables afin de choisir le nombre approprié de variables et de retenir les plus pertinentes. Enfin, nous terminons cette première phase en appliquant les approches de ré-échantillonnage pour la prise en compte du problème de déséquilibre des classes.

La deuxième phase est appelée **Classification** dans laquelle nous appliquerons plusieurs algorithmes de *machine learning* sur les données équilibrées par la phase de rééchantillonnage mais aussi sur les données originales. Six (6) algorithmes de classification ont été principalement retenu pour la phase 2. Il s'agit, de la régression logistique (LR), de l'arbre de décision (DT), du machine à vecteurs de support (SVM), du forêt aléatoire (RF), de l' eXtreme gradient boosting (XGB) et des réseaux de neurones (NN). Nous présenterons tous ces algorithmes dans la suite de cette section. Ensuite, à l'aide des métriques de performances de la classification, nous déterminerons la combinaison la plus efficace dans la **phase 3**. Enfin, avec ce modèle final nous tenterons de trouver un seuil optimal en **phase 4** et de faire une interprétabilité de ce modèle final retenu en **phase 5**.

1.2.2 Feature engineering

L'objectif principal du *features engineering* est d'améliorer la fiabilité des données en nettoyant les données et en sélectionnant le sous-ensemble de variables pertinentes. Dans la prédiction du déclenchement de paiement, ignorer les caractéristiques non pertinentes peut augmenter la précision de la classification et réduire les coûts de calcul liés à l'exécution de plusieurs modèles d'apprentissage automatique.

Les données disponibles sont de bonne qualité, ce qui nous a conduits, lors de cette phase de *features engineering*, à appliquer la méthode d'encodage *one-hot* et à effectuer la normalisation des données. L'encodage *one-hot* ou encodage 1 parmi n consiste à encoder une variable à n états sur n bits dont un seul prend la valeur 1, le numéro du bit valant 1 étant le numéro de l'état pris par la variable. Pour la standardisation, nous utilisons la normalisation min-max :

$$X_{nouveau} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{C.1}$$

1.2.3 Application des techniques de rééchantillonnage

Comme mentionné dans la revue sur le déséquilibre des classes, les algorithmes de classification fonctionnent mieux lorsque le nombre d'instances de chaque classe est à peu près égal. Lorsque le nombre d'instances d'une classe dépasse de loin celui de l'autre, des problèmes surviennent. Dans le cas de la modélisation du déclenchement, la classe Label 0 [Pas de déclenchement] constitue environ 87.8% et la classe Label 1 [Déclenchement d'un

paiement] constitue environ 12.2%. L'ensemble de données utilisé pour la prédiction du déclenchement est déséquilibré avec un rapport de 7.2 proportionnel à Label 0 : Label 1. Et, en raison de ce déséquilibre, il est nécessaire d'équilibrer la classe pour obtenir des prédictions appropriées.

Notre étude utilise l'approche de rééchantillonnage pour résoudre le problème de déséquilibre. Dans notre travail, nous avons recours à trois catégories d'approches de rééchantillonnage, à savoir les méthodes de sous-échantillonnage, de suréchantillonnage, et hybrides. Dans l'approche de sous-échantillonnage, nous avons principalement utilisé la méthode du *random undersampling* (RUS). Pour l'approche de suréchantillonnage, le *random oversampling* (ROS), la technique de suréchantillonnage synthétique minoritaire (SMOTE) et l'échantillonnage synthétique adaptatif (ADASYN) sont étudiés. Enfin, le SMOTE + Tomek (SMOTE-TOMEK) et le plus proche voisin SMOTE+ ENN (SMOTE-ENN) sont utilisés dans l'approche hybride.

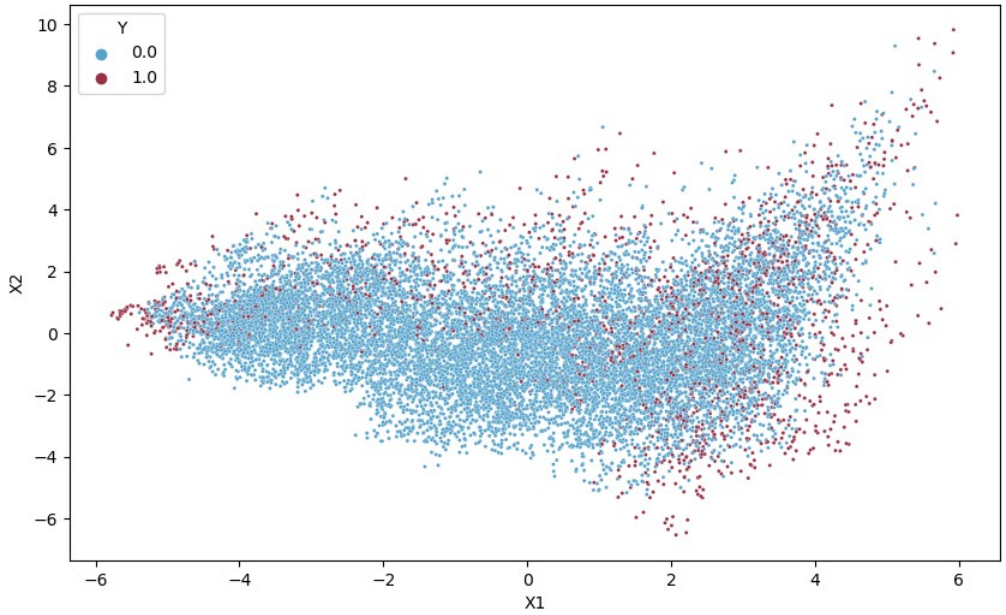


Figure C.2 – Visualisation à l'aide de l'ACP (Analyse en Composantes Principales) des données avant application des méthodes de rééchantillonnage.

1.2.4 Théorie de la sélection des variables (features selection)

En apprentissage automatique et en science des données en général, la sélection des variables ou *features selection* en anglais (également appelée sélection de variables, sélection d'attributs ou sélection de sous-ensembles) est le processus par lequel un data scientist sélectionne automatiquement ou manuellement un sous-ensemble de variables pertinentes à utiliser dans la construction d'un modèle d'apprentissage automatique.

En réalité, il s'agit de l'un des concepts fondamentaux en apprentissage automatique qui a un impact considérable sur les performances de vos modèles, car il est essentiel pour créer des modèles d'apprentissage automatique fiables. Le processus de *features selection* sé-

lectionnera le meilleur sous-ensemble d'attributs qui sont les plus importants et ont une contribution élevée au moment de la prise de décision.

Les données peuvent parfois contenir des caractéristiques non pertinentes qui n'améliorent pas les prédictions, ainsi que des caractéristiques redondantes qui perdent leur pertinence en présence d'autres, rendant le processus d'apprentissage complexe et susceptible de causer un sur-apprentissage. C'est pourquoi nous avons besoin de techniques visant à éliminer toute caractéristique susceptible de perturber l'apprentissage.

Il existe plusieurs autres raisons justifiant la finalisation de la sélection des caractéristiques, notamment :

Les modèles simples offrent des avantages significatifs :

— **Facilité d'interprétation** : Il est bien plus aisé de comprendre les résultats d'un modèle utilisant 10 variables que ceux d'un modèle utilisant 100 variables ;

— **Temps d'apprentissage réduit** : La réduction du nombre de variables diminue les coûts de calcul, accélère la formation du modèle, et, de manière cruciale, entraîne généralement des temps de prédiction plus rapides ;

— **Amélioration de la généralisation en réduisant le sur-apprentissage** : Souvent, de nombreuses variables ne sont que du bruit et ont peu de valeur prédictive. Cependant, le modèle d'apprentissage automatique apprend de ce bruit, ce qui peut provoquer un sur-apprentissage tout en réduisant la capacité de généralisation. En éliminant ces caractéristiques inutiles et bruyantes, nous pouvons significativement améliorer la capacité de généralisation des modèles d'apprentissage automatique.

— **Élimination de la redondance des variables** : Les caractéristiques d'un ensemble de données sont souvent fortement corrélées, et des caractéristiques fortement corrélées fournissent essentiellement les mêmes informations, ce qui les rend redondantes. Dans de tels cas, nous pouvons conserver une seule caractéristique tout en supprimant les caractéristiques redondantes, sans perdre d'informations. Moins de redondance signifie moins d'opportunités pour le modèle de faire des prédictions basées sur le bruit.

En général, les méthodes de sélection de caractéristiques peuvent être divisées en trois principales catégories :

1. **Méthodes de filtrage** (*filter methods*) : Elles se basent sur les variables elles-mêmes sans utiliser d'algorithme d'apprentissage automatique. Très adaptées pour un "tri et élimination" rapide des variables non pertinentes.

2. **Méthodes de type wrapper** (*wrapper methods*) : Elles considèrent la sélection d'un ensemble de caractéristiques comme un problème de recherche, puis utilisent un algorithme d'apprentissage automatique prédictif pour sélectionner le meilleur sous-ensemble de caractéristiques. En essence, ces méthodes entraînent un nouveau modèle sur chaque sous-ensemble de caractéristiques, ce qui les rend évidemment très coûteuses en termes de calcul. Cependant, elles fournissent le sous-ensemble de caractéristiques offrant les meilleures performances pour un algorithme d'apprentissage automatique donné.

3. **Méthodes intégrées** (*Embedded methods*) : Tout comme les méthodes de type wrapper, les méthodes intégrées (*Embedded methods*) prennent en compte l'interaction entre les caractéristiques et les modèles. Elles effectuent également la sélection de caractéristiques dans le cadre du processus de construction du modèle, et elles sont moins coûteuses en termes de calcul.

1.2.5 Modèles de prédiction du déclenchement

La modélisation de la probabilité de déclenchement est un problème de classification binaire. Puisque chaque algorithme a ses spécificités, ses forces et ses faiblesses, nous appliquerons plusieurs méthodes, à savoir la régression logistique (LR), l'arbre de décision (DT), les réseaux de neurones (NN), les Machine à vecteurs de support (SVM) et les méthodes d'ensembles telles que le XGBoost (XGB) et les forêts aléatoires (RF). Ainsi, après une étape de comparaison, nous retiendrons les modèles qui présentent les meilleures performances en termes de prédiction.

Régression logistique (LR)

La régression logistique est un modèle de classification linéaire qui est le pendant de la régression linéaire. Elle permet d'évaluer et de caractériser la relation entre la survenance d'un évènement, la variable expliquée qualitative, et un ou plusieurs facteurs, les variables explicatives, susceptibles de l'influencer. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.

Dans la régression logistique, ce n'est pas la réponse binaire (pas de déclenchement / déclenchement) qui est directement modélisée, mais la probabilité de réalisation d'une des deux modalités sachant les *features*.

Étant une probabilité, elle ne peut pas être modélisée par une droite car celle-ci conduirait à des valeurs en-dehors de l'intervalle $[0, 1]$. Elle est plutôt modélisée par une courbe sigmoïde, bornée par 0 et 1, qui est définie par la fonction logistique d'équation :

$$s(x) = \frac{\exp(x)}{1 + \exp(x)} \tag{C.2}$$

La fonction logistique est d'abord ajustée à des données observées, à partir de l'optimisation des coefficients de régression. La probabilité de réalisation s'écrit alors comme :

$$P(x) = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)} = \frac{\exp(\beta X)}{1 + \exp(\beta X)} \tag{C.3}$$

Les coefficients du modèle fournissent des informations sur l'importance relative de chaque variable d'entrée et sont estimés par la méthode du maximum de vraisemblance.

Nous le rappelons, la probabilité de survenance peut prendre n'importe quelle valeur entre 0 et 1. Afin de prédire la variable y, l'on se fixe ensuite un seuil tel que : y = 1 si

$P(x) \geq seuil$ et 0 si $P(x) < seuil$. De manière générale le seuil est fixé à 0.5.

Arbre de décision (DT)

Nous avons largement présenté ce modèle dans le chapitre II de ce mémoire. Pour rappel, un arbre de décision est un algorithme de Machine Learning qui est utilisé pour faire une prédiction ou de la classification. Un arbre de décision est composé de noeuds qui possèdent chacun une condition qui amène aux autres noeuds avec une structure descendante. On appelle noeud terminal, le noeud qui donne une réponse. Par analogie à un arbre il s'agit de la *feuille*.

Dans le cadre de la classification binaire, la *feuille* représente la prédiction de $\mathbb{Y}$ et les noeuds sont construits sous des conditions $x_j \leq a$ contre $x_j \geq a$, où x_j est la j-ème composante de $\mathbb{X}$ et a un seuil (Perchet).

Il est beaucoup utilisé dans les domaines d'aide à la décision de par sa représentabilité et son interprétabilité.

Forêts aléatoires

Les forêts d'arbres aléatoires (ou *random forests*) sont l'association de plusieurs arbres décisionnels. L'idée est de générer à partir de l'échantillon d'apprentissage B nouvelles bases de données sur le principe de tirage avec remise. Ensuite, l'algorithme apprend et construit sur chaque base un prédicteur qui est un arbre décisionnel ; on obtient finalement B prédicteurs correspondant à B arbres décisionnels. Le prédicteur par forêts aléatoires, n'est autre que la moyenne de ces B prédicteurs. Formellement, le principe est de :

— générer à partir de $\mathscr{D}_n$, par un tirage avec remise, B nouveaux sous-échantillons $\mathscr{D}_n^1, ..., \mathscr{D}_n^B$, avec une sélection aléatoire des features pour chaque nouveau sous-échantillon ;
— apprendre ensuite B prédicteurs différents $\hat{f}(\mathscr{D}_n^1, .), ..., \hat{f}(\mathscr{D}_n^B, .)$;
— prendre enfin la moyenne de ces prédicteurs

$$\hat{f}(.) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}(\mathscr{D}_n^b, .)$$

XGBoost

L'algorithme XGBoost (eXtreme Gradient Boosting) est un modèle de Machine Learning faisant partie de la famille des modèles ensemblistes qui consistent en une combinaison de modèles individuels en vue de créer un modèle plus fort et plus puissant. Plus encore, comme indiqué par Belaidi Belaidi, XGBoost est un modèle ensembliste séquentiel, c'est-à-dire que les modèles individuels créés sont très dépendants les uns des autres. Chaque modèle individuel entraîné vise à corriger les erreurs commises sur le modèle individuel entraîné précédemment. On parle d'apprentissage par boosting. Dans le cadre de l'algorithme XGBoost, les modèles individuels entraînés sont particulièrement des arbres de décision, et

la fonction de perte est minimisée au moyen de l'algorithme de descente de gradient.

Considérons un classifieur faible initial f_0. Après calibration, la méthode de boosting cherche à construire un nouveau classifieur faible f_1 à partir de f_0 en introduction un terme de résidu h de sorte que f_1 soit plus performant que f_0 :

$$f_1(x) = f_0(x) + h(x)$$

L'opération est répétée un certain nombre n de fois, pour construire un classifieur final F complexe qui est une somme pondérée des f_i entraînés avec des poids α_i associés :

$$F(x) = \sum_{i=1}^{n} \alpha_i f_i(x)$$

Contrairement aux forêts aléatoires où les arbres entraînés sont indépendants, l'algorithme XGBoost permet d'avoir plus rapidement des résultats plus précis à travers à la dépendance des arbres de décisions calibrés. En effet, l'algorithme se concentre donc uniquement sur les mauvaises prédictions de l'itération précédente au lieu d'entraîner de nouveau toute la base.

Support Vector Machine (SVM)

L'algorithme SVM appartient à la catégorie des classificateurs linéaires, c'est-à-dire qui utilisent une séparation linéaire des données. Un SVM cherchera simplement à trouver un hyperplan qui sépare les deux modalités de la variable à expliquer.

Pendant la phase d'apprentissage, l'algorithme détermine le vecteur de poids réels $w = (w_1, \ldots, w_n)$ et le biais $b \in \mathbb{R}$, pour former l'hyperplan vectoriel d'équation :

$$w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n = 0$$

Ainsi, pour classer une nouvelle entrée $x = (a_1, \ldots, a_n) \in \mathbb{R}^n$, le SVM regardera :

$$h(x) = \omega_1 a_1 + \ldots + \omega_n a_n + b = \sum_{i=1}^{n} \omega_i a_i + b = \omega^T \cdot x + b.$$

Si $h(x)$ est positif ou nul, alors x est d'un côté de l'hyperplan affine et appartient à la première catégorie, sinon x est de l'autre côté de l'hyperplan, et donc appartient à la seconde catégorie.

$$\begin{cases} h(x) \geq 0 \implies x \in \text{modalité 1} \\ h(x) < 0 \implies x \in \text{modalité 2} \end{cases}$$

Réseaux de neurones

Les réseaux de neurones constituent un sous-ensemble de l'apprentissage automatique qui ont la capacité de comprendre des ensembles de données complexes. Dans les problèmes

de prédiction impliquant des données non structurées (images, texte, etc.), les réseaux de neurones artificiels ont tendance à surpasser tous les autres algorithmes ou cadres.

Il s'agit d'une succession de trois types de couches de neurones à savoir :

— la couche d'entrée composée des neurones d'entrée qui lisent les données de l'extérieur ;
— les couches intermédiaires composées des neurones intermédiaires ou de traitement ;
— la couche de sortie constituée des neurones de sorties.

Un poids est associé à chaque neurone et un biais peut être appliqué à chaque couche du réseau. À l'instar d'un réseau de neurones biologique, les neurones des couches intermédiaires et de sortie s'activent et transmettent l'information aux couches suivantes uniquement lorsque le signal reçu dépasse un seuil donné. En effet, ces neurones sont dotés de fonctions d'activation (des relations mathématiques simples) qui font intervenir les données en sortie des couches précédentes. Pour illustrer le fonctionnement d'un réseau de neurones, nous présentons le perceptron, un réseau de neurones simplifié.

Le perceptron est formée d'une première couche ou neurones qui permettent de lire les données : chaque neurone de cette couche correspond à une des variables d'entrée. On peut rajouter un neurone de biais w_0 qui est toujours activé (il transmet 1 quelles que soient les données). Comme expliqué par Azencott, les neurones sont reliés à un seul et unique neurone de sortie, qui reçoit la somme des neurones qui lui sont reliés, pondérée par des poids de connexion.

Pour p variables $x_1, x_2, \ldots, x_p$, la sortie reçoit donc $w_0 + \sum_{j=1}^{p} w_j \cdot x_j$. L'unité de sortie applique alors à cette sortie une fonction d'activation a. Un perceptron prédit donc grâce à une fonction de décision f définie par :

$$f(x) = a\left(\sum_{j=1}^{p} w_j \cdot x_j + w_0\right)$$

Une fois qu'ils sont correctement entraînés, les réseaux de neurones artificiels peuvent apprendre d'eux-mêmes et se mettre à jour en permanence afin de fournir des données de sortie de plus en plus précises.

1.2.6 Métriques d'évaluation des modèles

Une méthode courante pour déterminer les performances d'un classificateur consiste à utiliser une matrice de confusion. Une matrice de confusion est un tableau qui classe les prédictions selon qu'elles correspondent ou non à la valeur réelle. Dans une matrice de confusion, TN est le nombre d'instances négatives correctement classées (Vrais Négatifs), FP est le nombre d'instances négatives incorrectement classées comme positives (Faux Positif), FN est le nombre d'instances positives incorrectement classées comme négatives (Faux Négatifs) et TP est le nombre d'instances positives correctement classées comme positives (True Positives). Le tableau ci-dessous montre cette matrice de confusion :

Confuxion Matrix	Predicted	
Actual	Positive (1)	Negative (0)
Positive (1)	TP	FN
Negative (0)	FP	TN

Table C.1 – Matrice de confusion

À partir de la matrice de confusion, de nombreuses métriques d'évaluation standard peuvent être définies. Traditionnellement, l'exactitude (*accuracy*) est la métrique de performance la plus courante dans un problème de classification binaire. Il s'agit simplement d'un rapport entre les observations correctement prédites et l'ensemble des observations :

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Cependant, lors de l'évaluation de jeux de données déséquilibrés, l'exactitude a tendance à mettre l'accent sur la classe majoritaire, ce qui rend difficile la performance du classificateur sur la classe minoritaire. Dans la littérature, nous avons quelques métriques alternatives utilisées pour évaluer les performances des classificateurs sur des jeux de données déséquilibrés. Il s'agit : de la précision, du rappel (*recall*), de l'AUC, de la mesure F_1, et de la mesure G-mean.

Précision et le rappel (*recall*)

La précision et le rappel peuvent être calculés à partir de la matrice de confusion comme suit :

$$\text{Précision} = \frac{TP}{TP + FP} \qquad Rappel = \frac{TP}{TP + FN}$$

À partir des équations, nous voyons que la précision est le rapport entre les observations positives correctement prédites et le total des observations positives prédites. une valeur élevée de la précision est liée au faible taux de faux positifs. Le rappel peut être défini comme le rapport entre le nombre total d'observations positifs correctement classés et le nombre total d'observations positifs. Un rappel élevé indique que la classe est correctement prédite (un petit nombre de FN). Comme nos données sont déséquilibrées, l'un de nos objectifs est d'améliorer le rappel sans nuire à la précision. Ces objectifs sont cependant souvent contradictoires, car pour augmenter le TP pour la classe minoritaire, le nombre de FP est également souvent augmenté, ce qui entraîne une précision réduite.

Mesure AUC : la courbe de ROC

Une courbe ROC (Receiver Operating Characteristic) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs. L'aire sous la courbe ROC correspond à AUC (Area Under Curve).

Une façon de calculer AUC est : étant donné n_0 points de classe 0, n_1 points de classe 1 et S_0 comme la somme des rangs des observations de la classe 0 alors,

$$AUC = \frac{S_0 - n_0(n_0+1)}{2n_0n_1}$$

Mesure F_1

La mesure F_1 donne la moyenne harmonique de la précision et du rappel. La mesure F_1 tente ainsi de mesurer les compromis entre précision et rappel en produisant une valeur unique qui reflète la qualité d'un classificateur. La formule du F_1 est la suivante :

$$F_1 = 2 \times \frac{\text{Précision} \times Rappel}{\text{Précision} + Rappel}$$

Mesure G-mean

La mesure G-mean prend en compte à la fois la sensibilité et la spécificité des deux classes dans le calcul de ses scores et constitue donc une mesure efficace pour les ensembles de données déséquilibrés (Gong J.). La formule G-mean est la suivante

$$\text{G-mean} = \sqrt{\text{Précision} \times Rappel}$$

Une G-Mean faible indique une mauvaise performance dans la classification des cas positifs même si les cas négatifs sont correctement classés comme tels. Cette mesure est importante pour éviter un sur-ajustement de la classe négative et un sous-ajustement de la classe positive.

Dans le contexte de la prédiction des déclenchements, la métrique privilégiée est le rappel. En effet, le rappel mesure la capacité du modèle à détecter de manière efficace les déclenchements de paiements. De plus, il est préférable de faire une prédiction erronée de déclenchement dans le scénario le plus pessimiste RCP 8.5 que de faire une prédiction erronée en indiquant qu'il n'y aura pas de déclenchement.

1.2.7 Méthodes d'interprétabilité des modèles

Dans l'étape de sélection des modèles il se peut que nous le meilleur modèle soit un modèle de type boite noire (*black box model*) tel que les réseaux de neurones, l'XGBoost ou encore les forêts aléatoires. L'interprétation des modèles boîte noire (algorithmes d'apprentissage automatique complexes) est la plupart du temps difficile, car nous ne pouvons pas calculer les effets marginaux comme d'habitude. Ainsi, nous avons recours à des méthodes d'interprétation des modèles de type boîte-noire qui nous permettrons d'interpréter un modèle tel que l'XGBoost.

Il existe de nombreuses approches d'interprétation, et elles peuvent être regroupées en

deux grandes catégories exposées dans la littérature : les méthodes globales et les méthodes locales. Les méthodes locales se concentrent sur la compréhension de la prédiction générée par une observation spécifique de la boîte noire, tandis que l'approche globale cherche à comprendre le modèle dans son ensemble. Comme méthode globale, nous allons nous intéresser au graphique de dépendance partielle (PDP)[3], à l'importance des variables et au graphique des effets locaux accumulés (ALE). Concernant les méthodes locales, nous utiliserons LIME (*Local Interpretable Model-agnostic Explanations Local Interpretable Model-agnostic Explanations*) et la valeur de Shapley (SHAP).

Graphique de dépendance partielle (PDP)

L'objectif du PDP est de montrer l'effet marginal d'une ou plusieurs variables explicatives (généralement pas plus de deux) sur la prédiction faite par un modèle. La PDP est une méthode d'interprétation visuelle qui montre l'effet marginal d'une ou plusieurs variables sur la prédiction faite par le modèle.

Pour obtenir la fonction de dépendance partielle, on calcule :

$$\hat{f}_{x_S} = \mathbb{E}_{X_C}[\hat{f}(x_S, X_C)] = \int \hat{f}(x_S, x_C) d\mathbb{P}_{X_C}(x_c) \tag{C.4}$$

où $\hat{f}$ est le modèle formé, X_S est les variables pour lesquelles nous calculons l'effet et X_C est les variables de contrôle. Cette formule qui est différente de l'espérance conditionnelle de base de X_S et pour avoir son estimation, nous utilisons l'algorithme de Monte Carlo. Le PDP est le moyen le plus simple d'obtenir l'effet d'une variable sur les performances d'un modèle. En fait, si le graphique produit par le PDP est principalement plat, nous pouvons conclure que la variable n'a pas d'effet énorme sur la prédiction de Y.

les avantages de la PDP sont : simplicité d'interprétation, facilité d'implémentation et fonctionne pour les variables catégorielles. Comme inconvénient majeur, temps d'exécution conséquent. Aussi, la PDP repose sur une hypothèse d'indépendance entre les variables, qui est rarement vérifiée en pratique. Enfin, la PDP ne tient pas compte de la distribution des variables explicatives et masque les effets hétérogènes entre les variables, du fait de la moyenne réalisée. Nous avons besoin d'outils plus robustes pour obtenir un outil d'interprétation juste et fiable.

Graphique d'effets locaux accumulés (ALE)

Le graphique des effets locaux accumulés (Accumulated Local Effects Plot) a pour objectif de corriger le PDP, notamment lorsque l'on possède des variables explicatives corrélées entre elles. Pour neutraliser l'effet des corrélations, cette méthode calcule l'effet moyen des différences de prédiction. En d'autres termes, si nous souhaitons connaître l'effet d'une valeur x de la variable X, nous commençons par définir un intervalle autour de x dans lequel

3. Friedman, *Greedy Function Approximation : A Gradient Boosting Machine*

nous calculons l'effet de toutes les valeurs prises par x dans cet intervalle. En fin de compte, nous prenons la moyenne de ces effets. Nous obtenons ainsi :

$$\hat{f}_{x_S,ALE(x_S)} = \int_{x_S}^{z} \mathbb{E}_{X_C|X_S}[\hat{f}^{(S)}(X_S, X_C)|X_S = x_S]dz_S = \int_{x_S}^{z}\int_{x_C} \hat{f}^{(S)}(x_S, x_C)\mathbb{P}(x_C|x_S)dx_C dz_S \quad \text{(C.5)}$$

où z est la limite utilisée pour la délimitation de l'intervalle et $\hat{f}^{(S)}(x_S, x_C) = \frac{\partial \hat{f}(x_S,x_C)}{\partial x_S}$.

Cette méthode présente plusieurs avantages significatifs. Tout d'abord, elle est remarquablement simple à interpréter, ce qui la rend accessible même pour les personnes qui ne sont pas expertes en apprentissage automatique. De plus, elle offre un avantage en termes de vitesse de calcul par rapport à d'autres méthodes telles que le graphique de la Dépendance Partielle (PDP). Enfin, un autre atout majeur est sa capacité à fonctionner efficacement même lorsque les variables présentent des corrélations entre elles, ce qui est un avantage précieux dans des scénarios de modélisation de données réelles où les relations entre les variables peuvent être complexes.

Comme inconvénients de la ALE, on peut citer entre autre :

— Le choix du nombre d'intervalles est souvent un compromis délicat à trouver ;
— Cette méthode a tendance à masquer les effets hétérogènes qui peuvent exister entre les variables, car elle se fonde sur une moyenne globale, ce qui peut limiter la capacité à identifier des tendances spécifiques à certaines parties des données. ;
— l'interprétation est toujours compliquée en cas de variables très corrélées.

Importance des variables

La notion d'importance des variables dans un modèle a suscité de nombreuses définitions différentes dans la littérature. Nous adoptons a définition de l'importance des variables, qui est calculée de manière indépendante du modèle considéré, et que nous notons PFI (Permutation Feature Importance).

Cette méthode prend l'erreur de prédiction d'un modèle comme une fonction de la variabilité des variables. En d'autres termes, le PFI considère une variable comme plus importante si une permutation de ses valeurs entraîne une augmentation de l'erreur de prédiction du modèle. En fait, si la prédiction d'un modèle est fortement modifiée par le mélange des valeurs de la variable, cela signifie que le modèle est sensible aux variations de cette variable. Bien que cette méthode soit facile à comprendre et à calculer, le principal problème réside dans le caractère aléatoire de la permutation. L'importance des variables dépend de la manière dont les permutations sont effectuées.

$$PFI_S = \mathbb{E}[L(\hat{f}(\tilde{X}_S, X_C), Y)] - \mathbb{E}[L(\hat{f}(X), Y)] \quad \text{(C.6)}$$

où $L(y, f(x)) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ est la fonction d'erreur, $\tilde{X}_S$ est une permutation indépendante de X_S.

Cette méthode d'importance des variables présente de nombreux avantages. Tout d'abord, elle offre une interprétation facile : plus l'erreur de prédiction est grande lorsqu'on modifie la valeur d'une variable, plus l'information qu'elle véhicule est détériorée. De plus, elle permet d'obtenir une vue d'ensemble du comportement global du modèle, similaire à l'interprétation des coefficients dans un modèle de régression linéaire. Un autre avantage majeur est que cette méthode fournit un critère de comparaison entre différents modèles. De plus, elle tient compte à la fois des effets de la variable en question et des interactions entre cette variable et les autres variables, bien que cela puisse également être considéré comme un inconvénient dans certaines situations. Enfin, le calcul de l'importance des variables ne nécessite pas de ré-entraîner le modèle, ce qui permet de gagner du temps par rapport à d'autres méthodes.

Néanmoins, il existe également des inconvénients associés à cette méthode, notamment les suivants :

— Le choix des ensembles d'apprentissage et de test peut parfois ne pas être très clair, ce qui peut entraîner une certaine incertitude dans les résultats.
— Les résultats fournis par l'algorithme peuvent varier considérablement en raison du hasard introduit par les permutations, ce qui peut rendre les évaluations moins stables.
— L'ajout d'une variable corrélée à une autre peut réduire l'importance de la variable considérée, ce qui peut parfois être contre-intuitif.
— Les permutations peuvent créer des instances fictives. En effet, lorsqu'une variable est permutée au sein d'une instance, on ne prend pas en compte le fait que la nouvelle instance ainsi générée puisse être réellement observée dans les données réelles, ce qui peut poser des problèmes de réalisme dans l'interprétation des résultats.

Explications locales interprétables, agnostiques au modèle (LIME)

La méthode LIME est l'un des modèles de substitution locale. Dans la littérature, nous avons des modèles de substitution globale et des modèles de substitution locale. La première veut aborder le modèle de boîte noire par un modèle le plus simple comme les modèles d'arbre de décision ou les modèles linéaires. Les seconds, font de même mais juste sur une sélection d'observations.

Cette méthode utilise un modèle de substitution (SM) pour approcher le modèle boîte noire (BM). Le modèle de substitution qui est un arbre de décision dans notre cas de classification, s'obtient en résolvant le problème :

$$\hat{g} = \underset{g \in G}{argmin}[J(f, g, \pi_x) + \Omega(g)] \tag{C.7}$$

où J est la fonction de coût, f est une fonction de BM, g est une fonction de SM, π_x est la mesure du voisinage de x, et Ω représente la complexité du modèle. Pour la mise en œuvre de la méthode, nous commençons par générer un nouvel ensemble de données ($\tilde{X}_S$) dans le voisinage de l'individu que nous souhaitons expliquer. Ensuite, nous entraînons le SM sur les prédictions du BM ($\hat{Y} = \hat{f}(\tilde{X}_S)$). Enfin, le SM ainsi obtenu est utilisé pour expliquer

localement le BM.

Cette méthode présente plusieurs avantages notables. Tout d'abord, elle utilise l'algorithme Lasso, ce qui permet d'obtenir des explications de manière parcimonieuse, concise et claire. Cette approche est particulièrement adaptée pour expliquer des modèles dans un large éventail de domaines, que ce soient des tableaux de données, des textes ou des images. De plus, elle se distingue par sa rapidité d'exécution, ce qui la rend pratique pour une utilisation en temps réel ou pour de grands ensembles de données.

Cependant, il existe également des inconvénients à prendre en compte. Tout d'abord, les instances générées ne tiennent pas compte des corrélations entre les variables, ce qui peut parfois conduire à des explications simplifiées ou biaisées. Des variantes de la méthode existent pour remédier à ce problème, mais elles peuvent ajouter de la complexité au processus. De plus, les explications fournies par cette méthode peuvent parfois être instables, ce qui signifie qu'elles peuvent varier d'une exécution à l'autre. Enfin, la définition du "bon" voisinage n'est pas toujours claire, ce qui peut être un défi pour certains scénarios.

Valeur de Shapley (SHAP)

La valeur de Shapley consiste à calculer la contribution de chaque variable dans la prédiction faîte par le modèle, en attribuant un score "juste". Cette valeur trouve son origine dans la théorie des jeux. Si l'on considère un jeu $G = (1, ..., p, v)$ avec p joueurs et v la fonction caractéristique qui donne l'importance de chaque coalition (sous-ensemble de $P = (1, ..., p, p \in \mathbb{N}^*)$). Le but du jeu est de trouver la juste importance ($\phi = (\phi_1, ..., \phi_p)$) de chaque joueur dans le gain. Shapley montre que :

$$\phi_i(v) = \sum_{S \in S(1,...,p) \setminus i} \frac{(p - |S| - 1)!|S|!}{p!} (v(S \cup i) - v(S)) \tag{C.8}$$

Dans le cas de notre modèle de prédiction (boîte noire), le jeu est la prédiction d'une permutation $\tilde{x}$ de x, le gain est la prédiction réelle moins la prédiction moyenne de toutes les permutations et les joueurs sont les variables x_j. Dans ce cas général, la contribution d'une variable x_j, est définie par sa valeur shapley :

$$\phi_j(\Delta^{\tilde{x}}) = \sum_{S \in S(x_1,...,x_p) \setminus x_j} \frac{(p - |S| - 1)!|S|!}{p!} (\Delta^{\tilde{x}}(S \cup x_j) - \Delta^{\tilde{x}}(S)) \tag{C.9}$$

Cette méthode présente plusieurs avantages significatifs. Tout d'abord, elle possède la propriété de distribution des contributions qui est considérée comme "juste", ce qui signifie que les explications fournies sont équilibrées et cohérentes. De plus, elle est la seule méthode d'interprétation reposant sur une véritable théorie mathématique, ce qui renforce sa robustesse et sa fiabilité. Cette méthode pourrait également trouver une application pertinente dans le cadre du "droit à l'explication" du Règlement général sur la protection des données (RGPD), ce qui souligne son importance dans un contexte juridique. Enfin, elle se distingue

par sa simplicité d'interprétation, ce qui la rend accessible même pour les utilisateurs non experts en statistiques.

Cependant, il existe également des inconvénients à prendre en compte. Tout d'abord, le temps de calcul nécessaire pour obtenir une valeur exacte peut être très important, voire impossible à calculer en pratique, ce qui peut limiter son utilisation dans des applications en temps réel ou avec de grands ensembles de données. De plus, l'instabilité des résultats peut être introduite en raison de l'échantillonnage réalisé, ce qui peut rendre les explications moins fiables. Tout comme certaines autres méthodes, les instances générées ne tiennent pas compte des corrélations entre les variables, bien qu'il existe des variantes pour atténuer ce problème. Enfin, cette méthode ne permet pas de réaliser une sélection parcimonieuse des variables, car toutes les variables sont utilisées pour l'explication.

1.3 Exploration des données

base de donnée pour la modélisation de la probabilité de déclenchement contient 28946 observations et 19 variables. Au niveau des variables on a la variable *fréquence* qui est la variable de déclenchement ou non. Initialement la base contient 5 variables qualitatives dont la région, la saison, le mois, l'indicateur de vacance scolaire et l'indicateur de week-end. Concernant les variables quantitatives, la base contient la **température** journalière en degrés celsius (°C), **la précipitation** en millimètre (mm) et la **vitesse de vent** en mètre par seconde (m/s). Ces trois variables sont associées aux risques couverts par le régime. Aussi, pour chacune de ces variables nous avons retenu la moyenne, le minimum et le maximum en moyenne mobile (glissante) sur 7 jours.

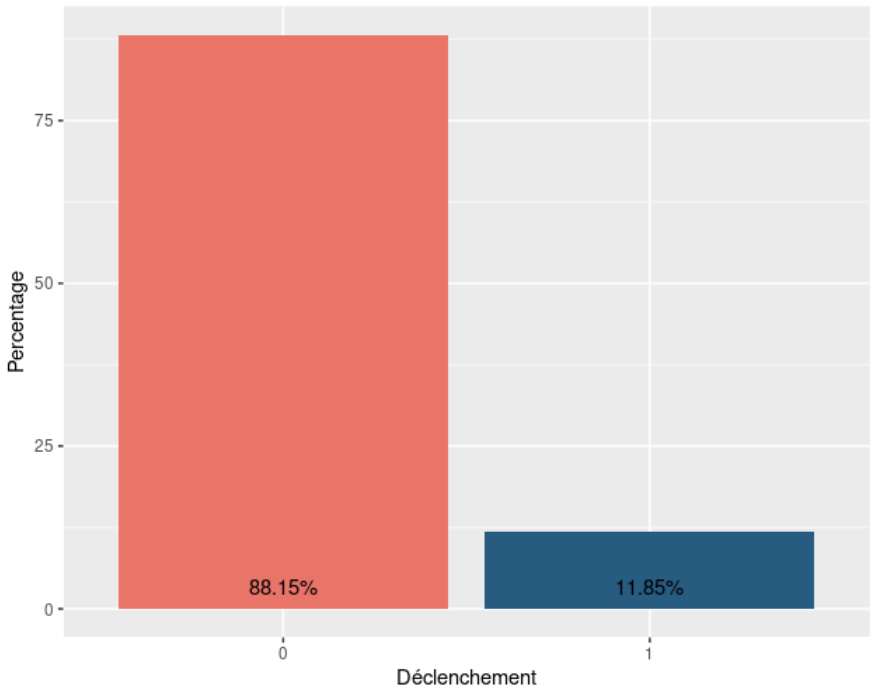


Figure C.3 – Distribution du déclenchement.

La figure ci-dessus montre la répartition des déclenchements dans la base de données. Il n'y a 11.85% d'observation où il y a déclenchement. Cette situation montre le déséquilibre qui existe dans l'ensemble des données entre les observations de déclenchements et les observations ou il n'y a pas eu de déclenchement.

Avec la variable *année* présente dans la base originale, nous avons séparé les données en un ensemble d'entraînement (années 2016 à 2018) et un ensemble de test (année 2019). Ensuite, les méthodes de rééchantillonnage retenues ont été appliquées à l'ensemble d'entraînement avec une stratégie d'échantillonnage à 30%. Cette stratégie d'échantillonnage à 30% a pour résultat une proportion de 30% entre la classe de non-déclenchement et la classe de déclenchement après l'application des techniques de rééchantillonnage. La figure C.5 ci-dessous montre comment les classes sont réparties après application de ces techniques. Les coordonnées (X1, X2) correspondent aux points sur les deux premiers axes de l'ACP (Analyse en Composantes Principales), qui est une technique de réduction de la dimensionnalité.

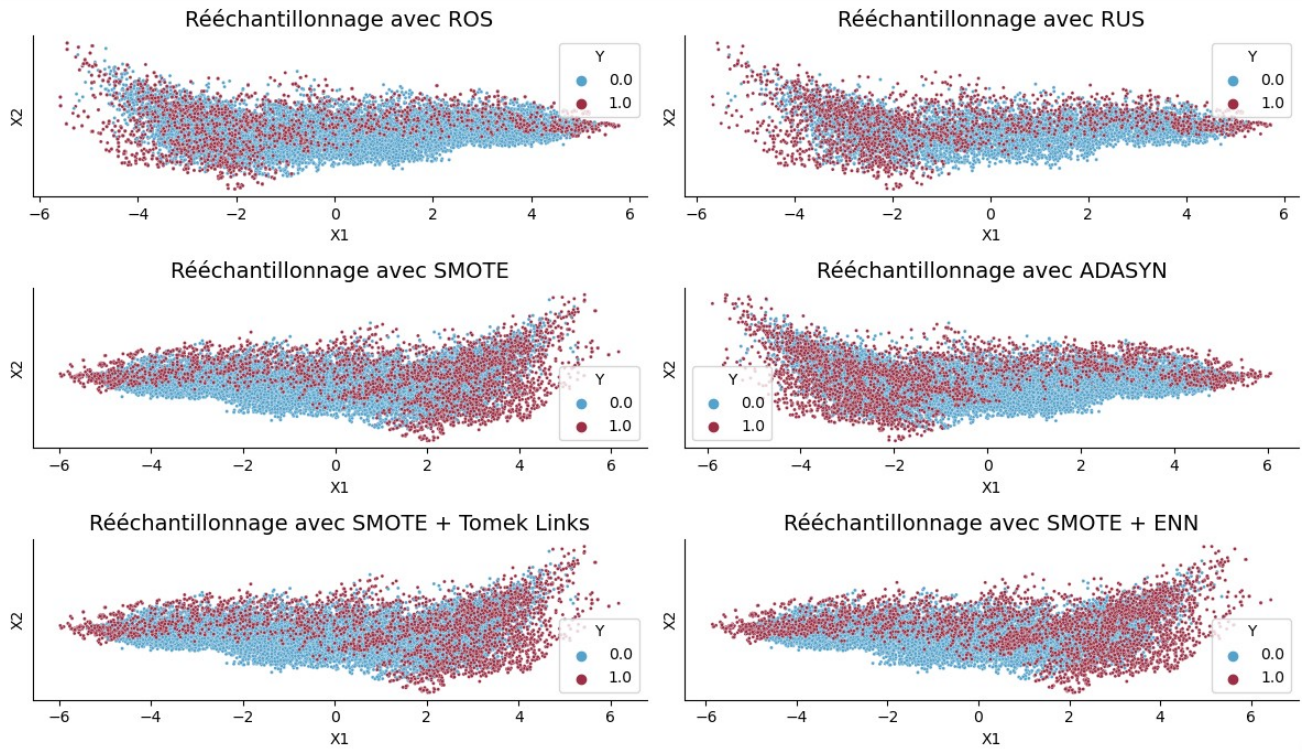


Figure C.4 – Visualisation à l'aide de l'ACP (Analyse en Composantes Principales) des données après application des méthodes de rééchantillonnage.

Après application du *features engineering* (one-hot-encoding, standardisation), nous avons sélection les features les plus pertinentes pour la modélisation selon les méthodes de *filter method*, *wrapper method* et *embedded method*. 12 features ont été retenu que sont : *t_mean* (température moyenne journalière), *vent* (vitesse de vent journalière), *volume_eau* (précipitation journalière), *t_mean_rol_min* (minimum de la température journalière en glissement de 7 jours), *t_mean_rol_max* (maximum de la température journalière en glissement de 7 jours), *vent_rol_min* (minimum de la vitesse de vent journalière en glissement de 7 jours), *vent_rol_max* (maximum de la vitesse de vent journalière en glissement de 7 jours), *code_insee_reg_53*

(indicateur indiquant si la région est Bretagne ou non), *saison_Automne* (indicateur indiquant si le jour est en automne ou non), *saison_Hiver* (indicateur indiquant si le jour est en hiver ou non), *saison_Printemps* (indicateur indiquant si le jour est au printemps ou non) et *saison_ete* (indicateur indiquant si le jour est en été ou non).

La figure ci-dessous montre matrice de corrélation de pearson entre les variables retenues.

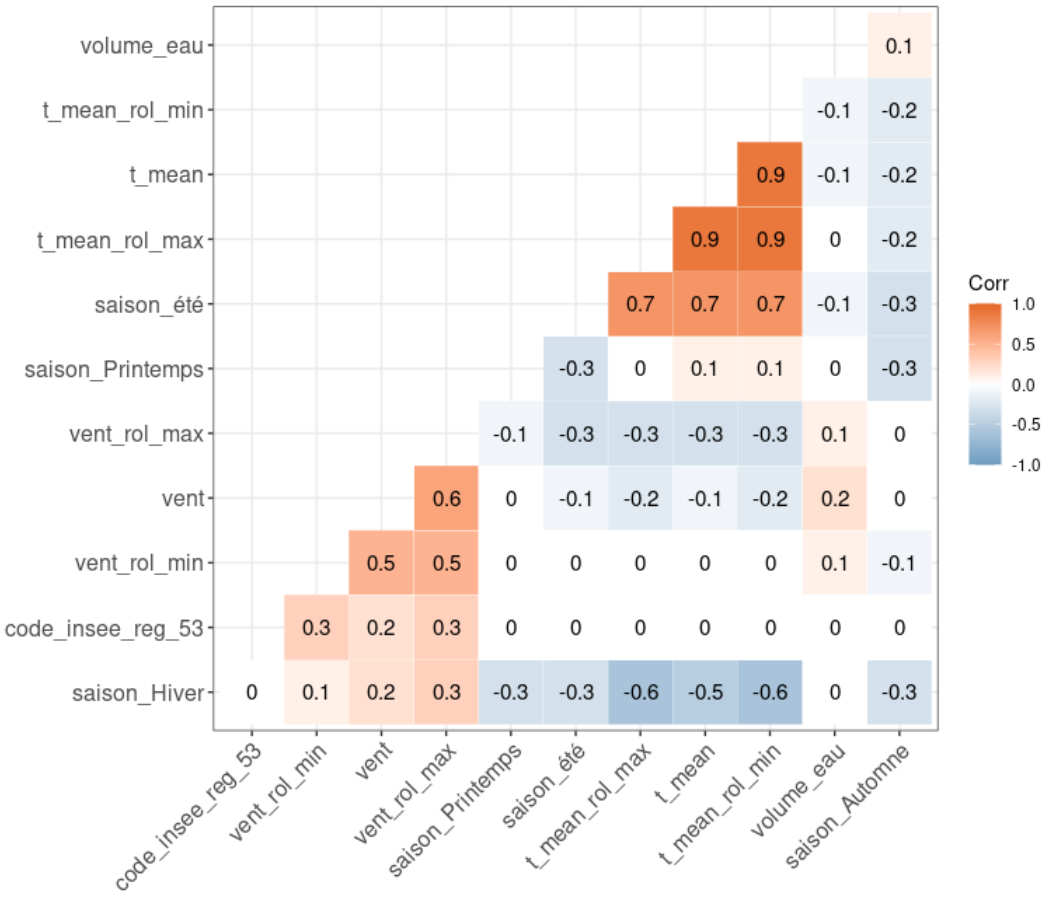


Figure C.5 – Matrice de corrélation des variables retenues.

1.4 Choix du modèle et seuil optimal

1.4.1 Choix du modèle

Pour choisir le meilleur modèle, nous avons évalué les performances des classificateurs sélectionnés en les combinant avec diverses méthodes de rééchantillonnage. Les algorithmes de classification retenus, tels que la Régression logistique (LR), l'Arbre de décision (DT), XGBoost (XBG), le Support Vector Machine (SVM), le Random Forest (RF) et les Réseaux de neurones (NN), ont été testés avec chaque type de méthode de rééchantillonnage. Nous avons également testé ces classificateurs sur la base de données originale, ce qui nous a permis de déterminer s'il était utile d'avoir recours au rééchantillonnage.

Pour chaque combinaison, nous avons faire le *tuning* des paramètres du l'aide de GridSearchCV avec un CV = 5. Le GridSearchCV est une méthode d'optimisation des hyperparamètres qui automatise le processus de recherche de la meilleure configuration d'hyperparamètres pour un modèle de machine learning donné en utilisant une validation croisée pour évaluer les performances. Cela permet de sélectionner les hyperparamètres qui donnent les meilleures performances sur un ensemble de données spécifique.

Le tableau C.2 ci-dessous présente les valeurs de la mesure de performance Recall pour l'ensemble des combinaisons constituées d'un classificateur (DT, LR, XGB, SVM, RF, ou encore NN), d'un algorithme de rééchantillonnage, ainsi que du fait que le modèle a été configuré à l'aide du GridSearchCV ou non. On y retrouve également les cas d'application directe sur les données d'origine, c'est-à-dire sans rééchantillonnage.

	Originale		RUS		ROS		SMOTE		ADASYN		SMOTE + TOMEK		SMOTE + ENN	
	Grid Search CV-5													
	Non	*Oui*	*Non*	*Oui*	*Non*	*Oui*	*Non*	*Oui*	*Non*	*Oui*	*Non*	*Oui*	*Non*	*Oui*
LR	44.89	76.12	54.96	56.98	56.55	56.69	55.68	56.12	60.14	76.98	55.83	55.97	57.41	76.26
DT	78.27	77.41	78.85	74.68	77.99	76.26	77.84	76.55	76.55	78.56	77.84	78.42	82.3	80.43
XGB	76.55	76.69	78.27	78.99	77.12	76.98	78.13	78.27	77.99	77.55	76.98	77.55	82.45	82.3
SVM	62.88	77.27	69.93	79.28	70.36	78.99	71.65	76.4	73.81	76.4	71.65	76.4	75.54	77.55
RF	76.12	84.17	79.42	84.6	76.55	84.6	79.28	**84.89**	78.85	84.75	78.99	**84.89**	81.15	83.45
NN	62.59	73.81	74.1	79.86	0	74.24	0	71.65	**84.89**	77.55	0	82.3	83.74	81.44

Table C.2 – Sélection des meilleurs modèles à l'aide du Recall (en %)

Les combinaisons les moins performantes se caractérisent par un Recall de 0, ce qui signifie qu'elles ne parviennent pas du tout à bien classifier la classe 1, correspondant aux déclenchements. En revanche, les combinaisons les plus performantes atteignent un Recall de 84,89%. Trois de ces combinaisons se distinguent particulièrement selon le Recall. Il s'agit de :

— SMOTE + Forêt aléatoire (RF) + GridSearchCV ;
— (SMOTE+ENN) +Forêt aléatoire (RF) + GridSearchCV ;
— ADASYN + Réseau de neurone (NN).

On constate qu'aucun modèle entraîné sur le jeu de données d'origine n'a un meilleur Recall. Cela met en évidence l'importance de l'amélioration apportée par le rééchantillonnage dans la prédiction du déclenchement.

Pour départager ces trois modèles, nous les comparons en utilisant les autres mesures de performance telles que l'Exactitude, la Précision, le F1, le G_mean et l'AUC. Le tableau C.3 présente ces mesures de performance pour les trois modèles que nous avons retenu selon le Recall.

Classificateur retenu	Exactitude	Précision	F1	G_mean	AUC
RF+SMOTE+GridSearch	82.34	44.6	58.47	61.53	83.4
RF+SMOTE+ ENN+GridSearch	**82.51**	**44.87**	**58.71**	**61.72**	**83.5**
NN + ADASYN	82.4	44.7	58.56	61.6	83.43

Table C.3 – Sélection du modèle final à l'aide des autres mesures de performance (en %)

La combinaison *RF+SMOTE+ENN+GridSearch* domine les deux autres combinaisons sur l'ensemble de ces métriques.

Ainsi, *SMOTE+ENN+RF+GridSearch* s'est révélé être la meilleure méthode pour prédire le déclenchement. Notre approche indique que l'utilisation d'une technique de rééchantillonnage SMOTE+ENN (approche hybride) pour résoudre les problèmes de déséquilibre de classe, suivie de l'utilisation d'un classificateur RF hyerparamétré par GridSearchCV et entrainer sur l'ensemble d'entraînement rééchantillonnage , surpasse les autres méthodes examinées dans l'étude. Les paramètres du modèle final, hyperparamétré par GridSearchCV, sont les suivants :

— nombre d'arbres dans la forêt (*n_estimators*) : 100 ;
— nombre maximum de nœuds feuilles (*max_leaf_nodes*) : 30 ;
— poids associés aux classes (*class_weight*) : "balanced" ;
— fonction pour mesurer la qualité d'un split (*criterion*) : 'gini' ;
— profondeur maximale de l'arbre (*max_depth*) : 50 ;
— autres paramètres : défaut (scikit-learn).

1.4.2 Seuil optimal

En pratique, les prédictions des observations sont classées en utilisant un seuil de 0.5. Ainsi, si la probabilité prédite est supérieure ou égale à 0.5, l'observation est classée dans la classe de référence. Cependant, en ajustant ce seuil, il est possible d'observer une variation des résultats. Habituellement, pour déterminer le seuil optimal, on se réfère à la courbe ROC. Cependant, en raison du déséquilibre de notre ensemble de données, nous préférons utiliser la métrique F1 comme mesure d'évaluation impartiale. Cela nous conduit à sélectionner un seuil optimal de classification binaire de 0.53, qui est très proche du seuil classique de 0.5. La figure C.6 montre la position sur la courbe ROC du seuil optimal qui maximise le F1.

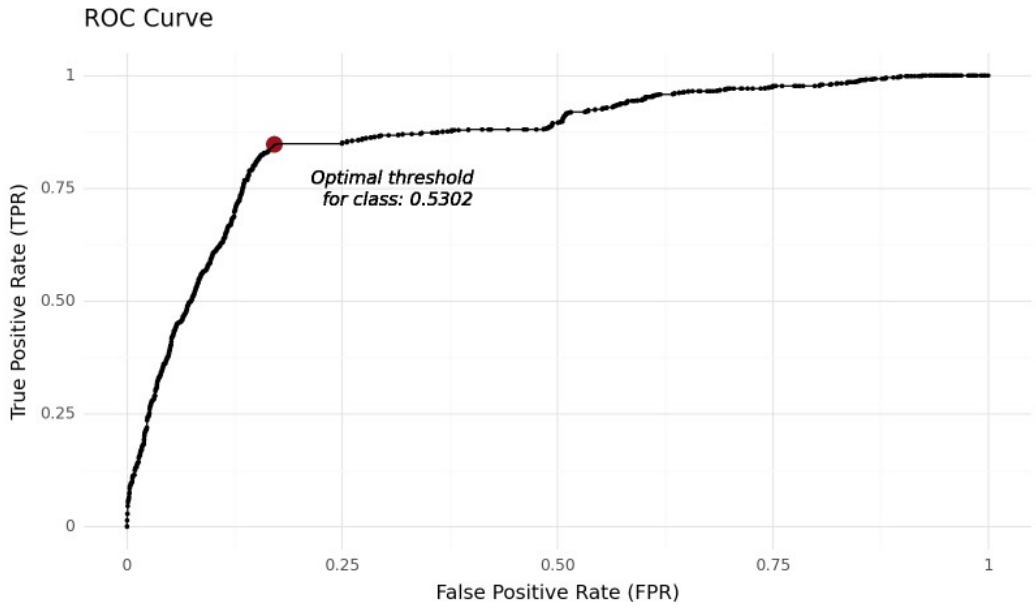


Figure C.6 – Sélection du seuil (*threshold*) pour le modèle *SMOTE+ENN+RF+GridSearch*.

1.5 Explicatiblité du meilleur modèle

Notre meilleur combinaison pour la prédiction des déclenchements comprend le modèle du Random Forest avec 100 arbres pour faire la prédictions. Ce modèle est souvent considéré comme une "boîte noire" en raison de sa complexité intrinsèque. Pour rappel, le RF est un ensemble d'arbres de décision qui travaillent ensemble pour effectuer des prédictions. Chacun de ces arbres individuels est déjà relativement difficile à interpréter en soi, car il divise l'espace des caractéristiques en une série de décisions binaires complexes. Lorsqu'on combine de nombreux arbres dans un Random Forest, la complexité augmente considérablement, ce qui rend la compréhension du modèle dans son ensemble encore plus difficile. Pour rendre ce modèle plus compréhensible et interprétable, il est souvent nécessaire de faire appel à des techniques d'interprétabilité de modèle.

Pour l'interprétation du meilleur modèle, nous avons utilisé les techniques suivantes : Importance des variables, ALE, les graphiques de dépendance partielle (PDP) et le LIME et la valeur de Shapley.

Comme nous l'avons déjà expliqué, l'importance des caractéristiques fait référence à des techniques qui attribuent un score aux caractéristiques d'entrée en fonction de leur utilité pour prédire une variable cible. Le graphiques ci-dessous montre l'importance des variables pour notre modèle final.

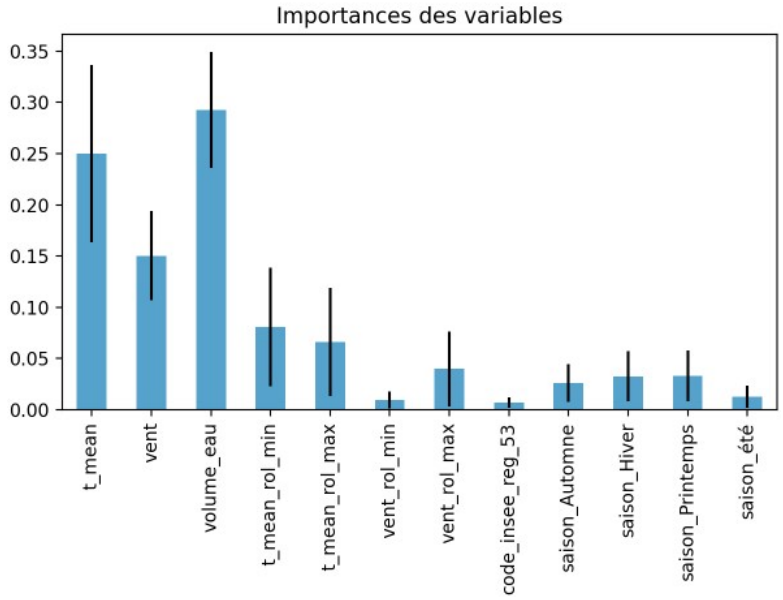


Figure C.7 – Importances des variables pour le RF.

L'analyse du graphique d'importance des variables montre que les trois variables (*t_mean*, *vent*,*volume_eau*) en lien direct avec les risques couverts (température, vent, pluie) par le régime sont les variables qui expliquent le plus les prédictions du Forêt aléatoire. La précipitation journalière (*volume_eau*) est la variable qui a le plus d'impact sur la variabilité des prédictions de la probabilité de déclenchement. Avec le graphique de l'importance des variables, nous ne pouvons pas savoir si l'effet de la variable *volume_eau* est positif ou non,

nous pouvons seulement classer les variables en fonction de la force de leur corrélation avec la prédiction de la survenance du déclenchement.

La figure ci-dessous représente le PDP de toute les features retenues pour la modélisation (à gauche) et les graphiques d'ALE de la température moyenne (*t_mean*) ainsi que de la variable vitesse de vent (vent). On remarque que dans la plage inférieur à 0.5°C plus la température moyenne est élevée plus la probabilité de déclenchement est faible. A partir de 0.5°C la température moyenne à un impact moyen nul sur les prédiction du modèle.

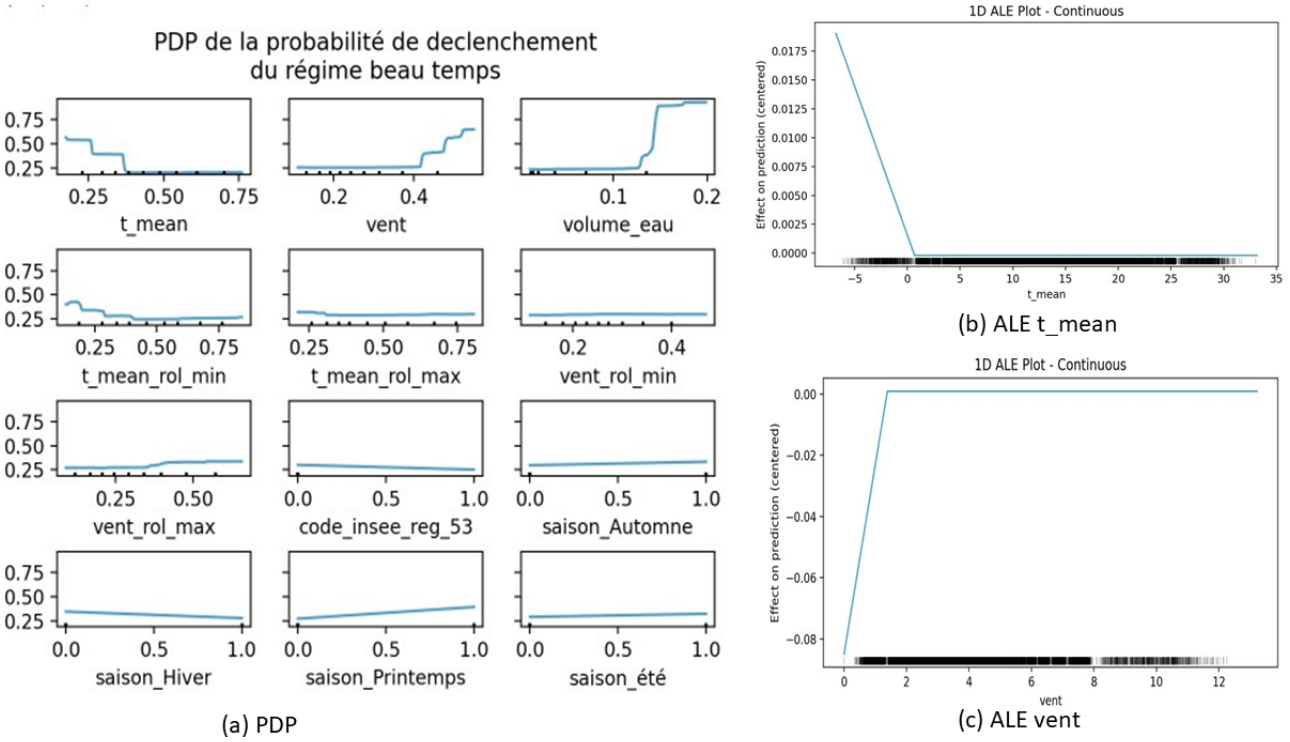


Figure C.8 – PDP pour le RF, ALE *t_mean* et ALE *t_vent* .

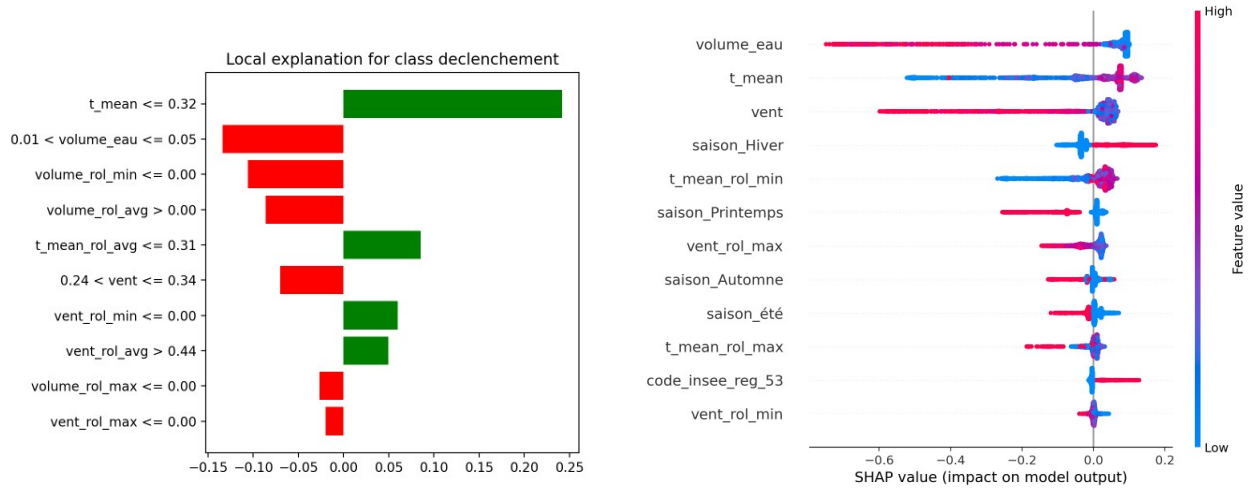


Figure C.9 – LIME (à gauche) et valeur de Shapley (à droite) pour le RF

2 Prédiction du dépassement du seuil (δ)

Dans cette section, nous présenterons les différentes étapes de la modélisation concernant la probabilité de dépassement du seuil des extrêmes, fixé à 130 000€, par la sévérité (Y). Pour ce faire, nous commencerons par présenter le cadre méthodologique adopté pour cette modélisation. Ensuite, nous présenterons les résultats de la modélisation de δ, puis nous interpréterons le meilleur modèle que nous aurons sélectionné.

2.1 Cadre méthodologique

2.1.1 Description de la méthodologie

Tout comme dans la modélisation de la probabilité de déclenchement (voir section 1), la modélisation de la probabilité de dépassement du seuil extrême u se déroulera en cinq étapes distinctes. Ces étapes incluent la phase de prétraitement (*preprocesing*), la classification, la comparaison, le choix du seuil optimal, et enfin, l'interprétation du meilleur modèle sélectionné. A la différence que nous n'adaptons pas le réechantillonnage dans la phase de *préprocessing* car nous ne sommes plus en présence de classe déséquilibrée et nous incluons dans la phase de classification, notre approche spécifique pour la modélisation de la probabilité de dépassement du seuil des extrêmes u.

Notre approche s'appuie dans un premier temps sur les prédictions numériques générées par un modèle de régression concernant la valeur de la sinistralité future. Fondamentalement, l'approche que nous décrivons est indépendante du modèle de régression qui effectue les prédictions.

Soit $\hat{y}$ la prédiction future. Nous supposons que $\hat{y}$ peut être modélisé selon une distribution normale de moyenne $\hat{y}$ et d'écart type σ_y : $\mathcal{N}(\hat{y}, \sigma_y)$, avec σ_y est calculé en fonction de l'écart type des données d'entraînement. Nous pouvons ainsi estimer $p_i(x)$, la probabilité de dépassement pour un individu avec les caractéristiques x, en utilisant la fonction de distribution cumulative (CDF) de $\mathcal{N}\left(\hat{y}_i, \sigma_y^2\right)$:

$$p_i(x) = 1 - CDF_{\mathcal{N}\left(\hat{y}_i, \sigma_y^2\right)}(u) \tag{C.10}$$

Lorsqu'elle est évaluée au seuil $u = 130\,000$€, la CDF représente la probabilité que la variable aléatoire respective prenne une valeur inférieure ou égale à u. En effet, nous prenons la valeur complémentaire pour obtenir la probabilité que la variable aléatoire dépasse u.

Pour la prédiction de la variable y, nous avons exploré plusieurs modèles de régression, notamment les Moindres Carrés Ordinaires (OLS), la Régression LASSO, la Régression Ridge, la Régression ElasticNet, le Gradient Boosting (GB), la Forêt Aléatoire (RF) et les Réseaux de Neurones (NN). Après avoir évalué la performance de ces modèles à l'aide de différentes métriques de régression, nous avons sélectionné le meilleur modèle, puis nous

l'avons affiné en utilisant GridSearchCV. Une fois que nous avons obtenu le modèle de régression optimal, nous avons comparé les performances de notre approche avec d'autres méthodes de classification.

En ce qui concerne la classification, nous avons examiné plusieurs modèles, notamment la Régression Logistique (LR), l'Arbre de Décision (DT), la Machine à Vecteurs de Support (SVM), la Forêt Aléatoire (RF), le Gradient Boosting Extrême (XGB) et les Réseaux de Neurones (NN). Nous avons ensuite procédé à une phase de comparaison en utilisant diverses métriques d'évaluation des modèles de classification et avons déterminé le seuil optimal pour la classification. Enfin, nous avons conclu notre démarche en interprétant le meilleur modèle sélectionné. Le schéma méthodologique de cette approche est résumé dans la Figure C.10.

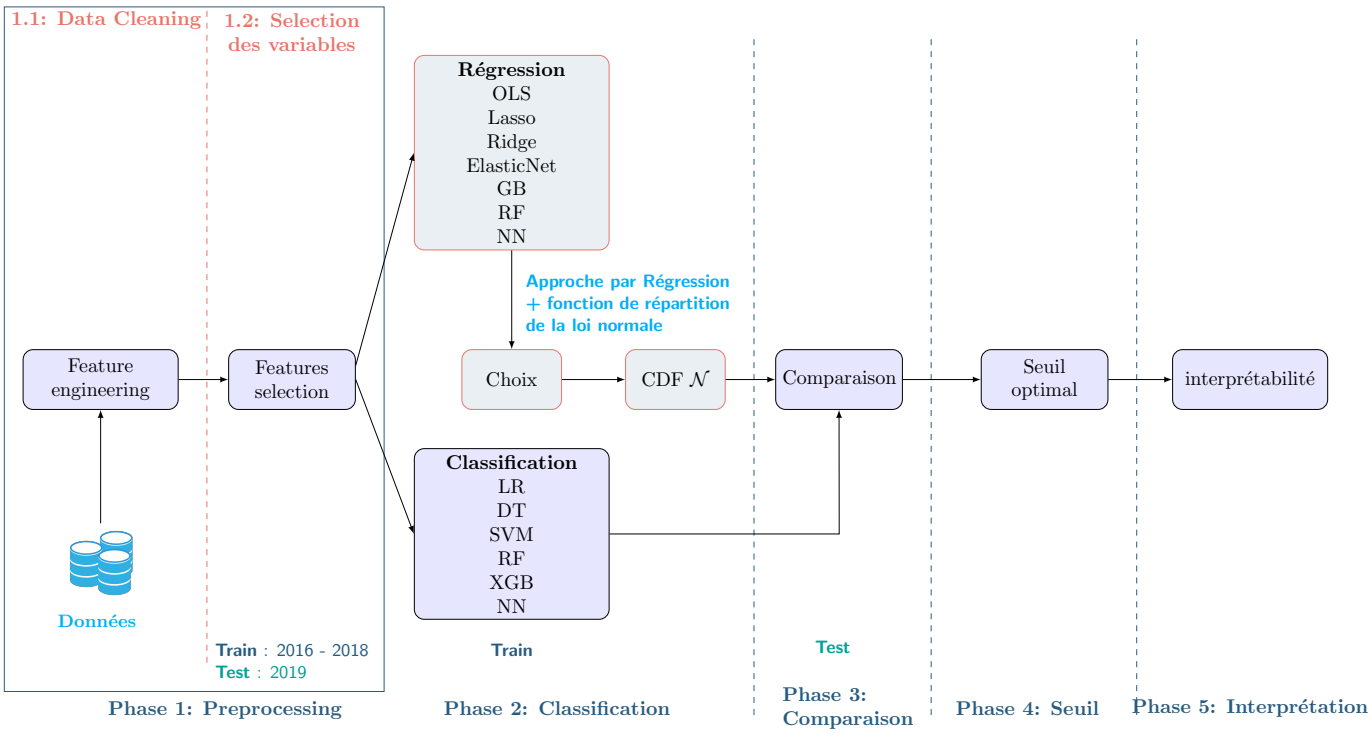


Figure C.10 – Diagramme résumant l'approche de modélisation de dépassement du seuil 130 000€ par la sévérité Y.

2.1.2 Modèles de régression adoptée

Régression linéaire multiple (OLS)

La régression OLS (moindres carrés ordinaires) est une technique pour estimer les coefficients d'une régression linéaire qui décrivent les relations entre une ou plusieurs variables indépendantes $(X_1, X_2, \ldots, X_p)$ et une variable dépendante (Y). Les moindres carrés désignent l'erreur quadratique minimale.

Le modèle de régression linéaire multiple peut être représenté par l'équation suivante :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i + \ldots \beta_p X_p + \varepsilon \tag{C.11}$$

où :

— Y est la variable dépendante à prédire ;

— X_i est la variable indépendante i ;

— β_0 est l'intercept du modèle ;

— β_i est le coefficient associé à la variable i ;

— ε est le terme d'erreur (résidu) qui capture les variations non expliquées.

L'objectif de la régression OLS est de trouver les valeurs des coefficients $\beta_0, \ldots \beta_1$ qui minimisent la somme des carrés des résidus, c'est-à-dire la somme des carrés des différences entre les valeurs observées Y_i et les valeurs prédites $\hat{Y}_i$:

$$\text{minimiser} \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{C.12}$$

On montre que cette minimisation conduit aux estimateurs des paramètres du modèle suivants : $= (X'X)^{-1}X'Y$ où β désigne le vecteur des estimateurs des paramètres β_i, X est la matrice des variables explicatives précédées d'un vecteur de 1, Y est le vecteur des n valeurs observées pour la variable dépendante. Une fois que les coefficients sont estimés, le modèle peut être utilisé pour faire des prédictions sur de nouvelles données.

Régression Lasso

La régression Lasso est une technique de régression linéaire régularisée qui pénalise les coefficients des variables moins importantes en ajoutant une pénalité de type L1 à la fonction de coût. Le modèle Lasso peut être formulé comme suit :

$$\min_{\beta_0, \beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + X_i^T \beta))^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{C.13}$$

où :

— Y_i est la variable cible ;

— X_i sont les prédicteurs ;

— β_0 est l'intercept ;

— β sont les coefficients des prédicteurs ;

— λ est le paramètre de régularisation.

La régularisation L1 encourage la parcimonie en forçant de nombreux coefficients β_j à être exactement nuls, ce qui équivaut à une sélection automatique des variables. Cela permet de gérer efficacement les problèmes de surajustement.

La figure ci-dessous illustre comment la régularisation L1 pousse les coefficients β_j vers zéro à mesure que λ augmente, ce qui conduit à la sélection des variables les plus importantes.

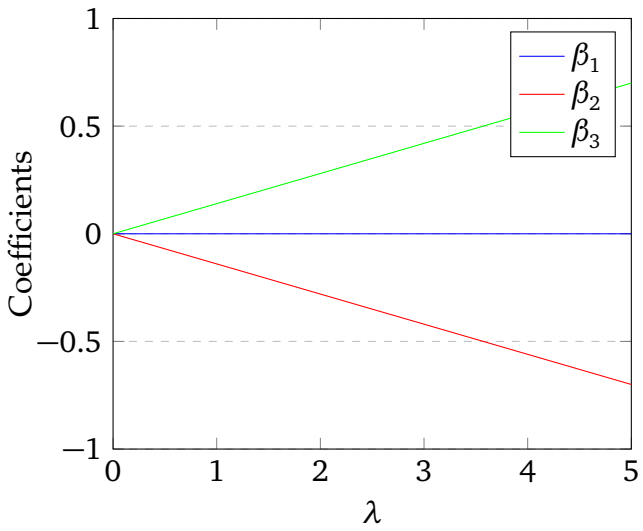


Figure C.11 – Effet de la régularisation L1 sur les coefficients.

Régression Ridge

La régression Ridge est une technique de régression linéaire régularisée qui pénalise les coefficients des variables moins importantes en ajoutant une pénalité de type L2 à la fonction de coût. Le modèle Ridge peut être formulé comme suit :

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + X_i^T \beta))^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{C.14}$$

où :

— Y_i est la variable cible ;

— X_i sont les prédicteurs ;

— β_0 est l'intercept ;

— β sont les coefficients des prédicteurs ;

— λ est le paramètre de régularisation.

La régularisation L2 pénalise les coefficients β_j en les forçant à être proches de zéro sans nécessairement les rendre exactement nuls. Cela permet de gérer efficacement les problèmes de surajustement.

Régression Elastic Net

La régularisation de type "Elastic net" consiste à combiner les deux régularisations précédentes (Ridge et Lasso) afin d'éviter la sélectivité trop forte que peut proposer Lasso tout en conservant possiblement des variables fortement corrélées.

Le modèle Elastic Net peut être formulé comme suit :

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + X_i^T \beta))^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\} \tag{C.15}$$

où :

— Y_i est la variable cible ;
— X_i sont les prédicteurs ;
— β_0 est l'intercept ;
— β sont les coefficients des prédicteurs ;
— λ_1 est le paramètre de régularisation L1 ;
— λ_2 est le paramètre de régularisation L2.

La régularisation Elastic Net combine les avantages de Lasso (sélection de variables) et Ridge (réduction de la multicollinéarité) en un seul modèle.

Régression Gradient Boosting (GB)

La Régression Gradient Boosting est une technique d'apprentissage automatique ensembliste qui combine plusieurs modèles de régression faibles pour créer un modèle plus puissant. Le modèle final est une somme pondérée des modèles faibles, où chaque modèle contribue en fonction de son erreur résiduelle. Le modèle de Régression Gradient Boosting peut être formulé comme suit :

$$\hat{Y} = \sum_{i=1}^{M} \alpha_i h_i(X) \tag{C.16}$$

où :
— $\hat{Y}$ est la prédiction finale ;
— M est le nombre de modèles faibles ;
— α_i sont les poids attribués à chaque modèle faible ;
— $h_i(X)$ sont les prédictions des modèles faibles.

La Régression Gradient Boosting fonctionne en ajustant itérativement les modèles faibles pour minimiser l'erreur résiduelle. Chaque modèle est ajusté en fonction de la perte résiduelle, ce qui permet d'obtenir des prédictions plus précises à chaque itération.

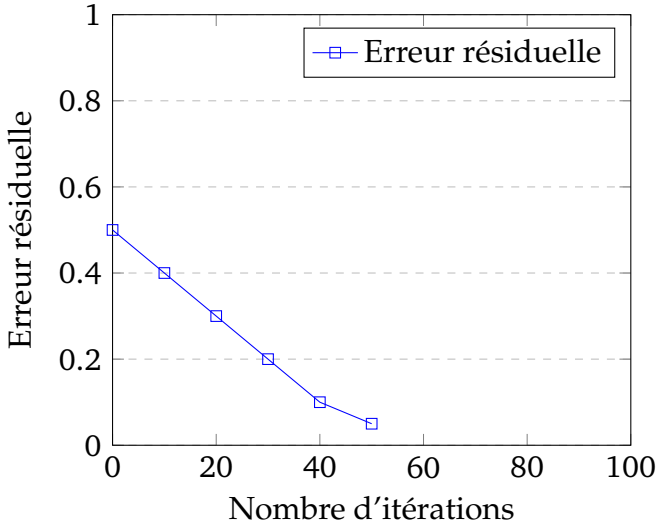


Figure C.12 – Réduction de l'erreur résiduelle par itération dans la Régression GB.

La figure ci-dessus montre comment l'erreur résiduelle diminue à chaque itération de la Régression Gradient Boosting, ce qui conduit à une amélioration de la prédiction finale.

2.1.3 Métriques d'évaluation de la régression

Les métriques d'évaluation sont utilisées pour mesurer la performance d'un modèle de régression en comparant ses prédictions aux valeurs réelles. Dans le cadre de ce mémoire nous nous limité aux trois métriques couramment utilisées que sont : le RMSE, MSE et le MAE.

Le RMSE (**Root Mean Squared Error**) est une mesure de l'erreur qui calcule la racine carrée de la moyenne des carrés des écarts entre les prédictions du modèle ($\hat{y}_i$) et les valeurs réelles (y_i) pour toutes les observations. Plus le RMSE est bas, plus les prédictions du modèle sont proches des valeurs réelles. L'avantage du RMSE est qu'il donne une indication de la précision absolue du modèle et les erreurs sont pondérées en fonction de leur magnitude. Comme inconvénient majeur du RMSE est qu'il est sensible aux valeurs atypiques.

Le MSE (**Mean squared error**) est similaire au RMSE, mais il ne prend pas la racine carrée de la moyenne des carrés des écarts. Il mesure la moyenne des carrés des erreurs et est donc plus sensible aux valeurs aberrantes.

Le MAE (**Mean Absolute Error**) mesure la moyenne des valeurs absolues des écarts entre les prédictions du modèle et les valeurs réelles. Contrairement au RMSE et au MSE, il ne pénalise pas fortement les valeurs aberrantes.

La formule du MSE, RMSE et du MAE sont les suivantes :

$$MSE = RMSE^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{C.17}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{C.18}$$

2.2 Exploration des données

base de donnée pour la modélisation de la probabilité de dépassement du seuil des extrêmes u, contient 2508 observations et 19 variables. Au niveau des variables on a la variable *dépassement* qui est la variable de dépassement du seuil extrême u ou non. Initialement la base contient 5 variables qualitatives dont la région, la saison, le mois, l'indicateur de vacance scolaire et l'indicateur de week-end. Concernant les variables quantitatives, la base contient la **température** journalière en degrés celsius (°C), **la précipitation** en millimètre (mm) et la **vitesse de vent** en mètre par seconde (m/s). Ces trois variables sont associées aux risques couverts par le régime. Aussi, pour chacune de ces variables nous avons retenu la moyenne, le minimum et le maximum en moyenne mobile (glissante) sur 7 jours.

La figure C.13 ci-dessous montre la répartition des dépassements dans la base de données. La classe de dépassement du seuil extrême u représente 42.78% des observations. Ainsi, contrairement aux données utilisées pour la modélisation de la probabilité de dé-

clenchement la base de données pour la modélisation de la probabilité de dépassement est non déséquilibrée.

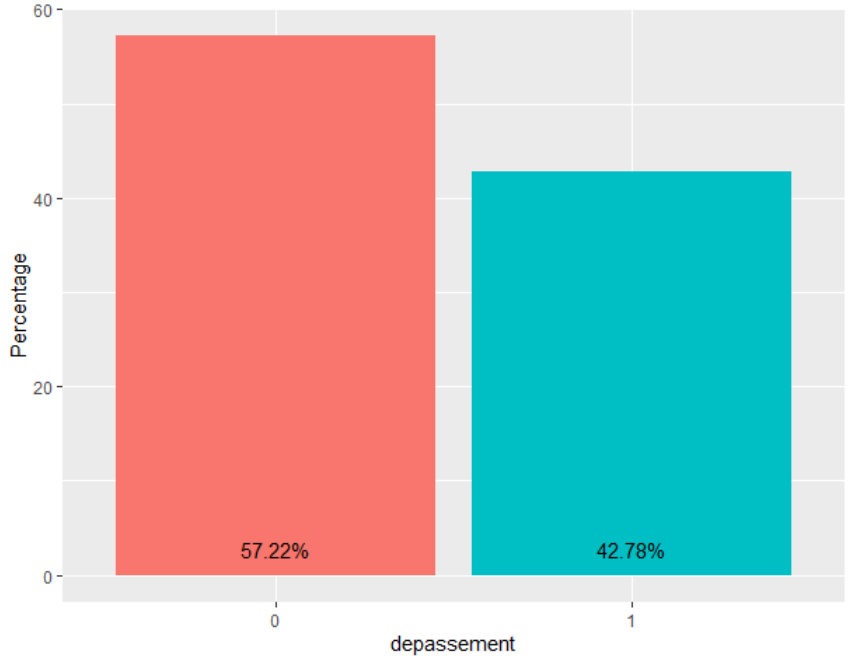


Figure C.13 – Distribution du dépassement.

Avec la variable *année* présente dans la base originale, nous avons séparé les données en un ensemble d'entraînement (années 2016 à 2018) et un ensemble de test (année 2019).

Après application du *features engineering* (one-hot-encoding, standardisation), nous avons sélection les features les plus pertinentes pour la modélisation selon les méthodes de *filter method*, *wrapper method* et *embedded method*. 21 features ont été retenu que sont :

— *t_mean* : température moyenne journalière ;
— *volume_eau* : précipitation journalière ;
— *t_mean_rol_min* : minimum de la température journalière en glissement de 7 jours ;
— *t_mean_rol_max* : maximum de la température journalière en glissement de 7 jours ;
— *t_mean_rol_avg* : moyenne de la température journalière en glissement de 7 jours ;
— *vent_rol_min* : minimum de la vitesse de vent journalière en glissement de 7 jours ;
— *vent_rol_max* : maximum de la vitesse de vent journalière en glissement de 7 jours ;
— *vent_rol_avg* : moyenne de la vitesse de vent journalière en glissement de 7 jours ;
— *volume_rol_avg* : moyenne de la précipitation journalière en glissement de 7 jours ;
— *volume_rol_min* : minimum de la précipitation journalière en glissement de 7 jours ;
— *code_insee_reg_53* : indicateur binaire relatif à la Bretagne ;
— *code_insee_reg_52* : indicateur binaire relatif à la Bretagne ;
— *code_insee_reg_75* : indicateur binaire relatif à la Bretagne ;
— *code_insee_reg_76* : indicateur binaire relatif à la Bretagne ;
— *code_insee_reg_84* : indicateur binaire relatif à la Bretagne ;
— *code_insee_reg_93* : indicateur binaire relatif à la Bretagne ;
— *mois_5* : indicateur binaire relatif mois de mai ;

— *mois_8* : indicateur binaire relatif mois d'août ;

— *mois_5* : indicateur binaire relatif mois de septembre ;

— *mois_5* : indicateur binaire relatif mois de novembre ;

— *vacance_scolaire_oui* : indicateur indiquant si le jour est un jour de vacance.

La figure ci-dessous représente la matrice de corrélation de pearson entre les variables retenues.

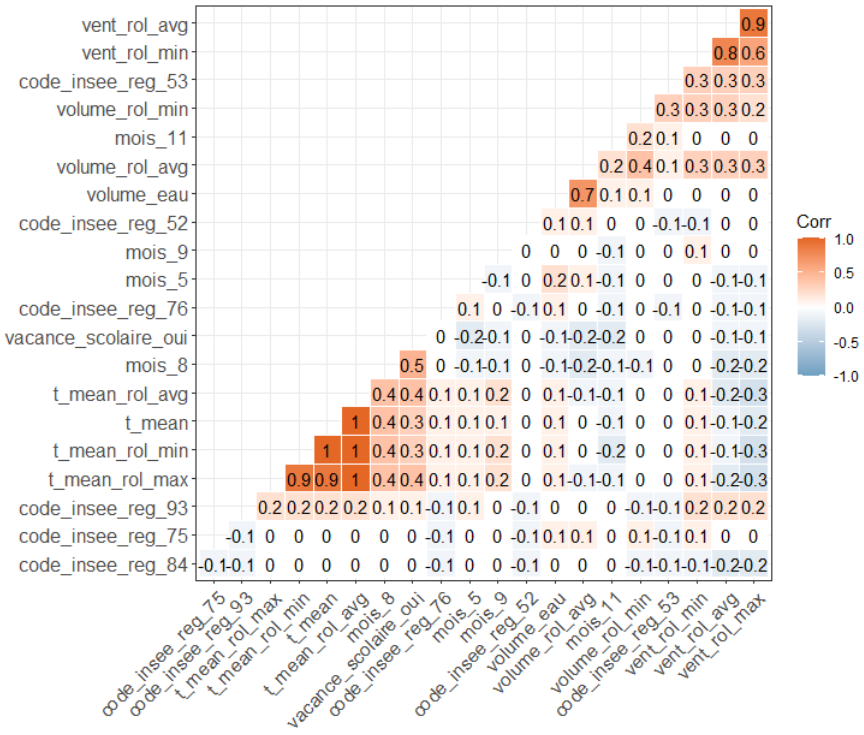


Figure C.14 – Matrice de corrélation des variables retenues pour la modélisation de la probabilité de dépassement du seuil u.

2.3 Choix du modèle et seuil optimal

Pour le choix du modèle de régression, le meilleur modèle selon les métriques de performances a été la forêt aléatoire (RF). Le tableau ci-dessous présente l'ensemble des résultats.

Régression	R2	MSE	MAE
OLS	0.64	173893	109716
RIDGE	0.64	173893	109716
LASSO	0.64	173893	109716
ElasticNet	0.64	174350	109565
GradientBoosting	0.75	144498	83245
RandomForest	0.75	144012	81809
NN	0.74	147378	89996

Table C.4 – Mesures de performance pour la régression

Ensuite, nous avons effectué l'hyperparamétrage du modèle de Forêt Aléatoire (RF) à l'aide du GridSearchCV avec CV=5 et calculer les probabilités p_i à l'aide de la formule C.10

(RF+CDFNormale).

Pour comparer notre procédure (RF+CDF Normale) avec d'autres modèles de classification (LR, DT, XG, SVM, NN, RF) on utilise la métrique de l'exactitude (*accuracy* en anglais). L'exactitude quantifie la proportion de prédictions correctes faites par le modèle parmi toutes les prédictions effectuées. En d'autres termes, l'exactitude mesure la capacité du modèle à classer correctement les exemples d'un ensemble de données. Notre approche (RF+CDF Normale) présente la meilleure exactitude (87.91%) par rapport aux autres classificateurs (voir le tableau ci-dessous).

Modèle	GridSearchCV	Exactitude	Recall	Précision	F1	G_mean	AUC
RF+CDF normale	*Oui*	87.91	82.69	93.48	87.76	87.92	88.17
Logistic Regression (LR)	*Non*	85.9	75.55	96.83	84.88	85.53	86.42
Logistic Regression (LR)	*Oui*	86.47	81.32	91.93	86.3	86.46	86.73
Decision Tree (DT)	*Non*	82.73	75.0	90.4	81.98	82.34	83.12
Decision Tree (DT)	*Oui*	82.3	75.55	89.0	81.72	82.0	82.64
XGBoost (XG)	*Non*	83.88	78.3	89.62	83.58	83.77	84.16
XGBoost (XG)	*Oui*	83.45	74.45	92.49	82.5	82.98	83.9
SVM	*Non*	87.48	78.57	96.95	86.8	87.28	87.93
SVM	*Oui*	81.29	66.76	96.43	78.9	80.23	82.02
Random Forest (RF)	*Non*	83.17	77.47	88.96	82.82	83.02	83.45
Random Forest (RF)	*Oui*	81.73	76.92	86.69	81.51	81.66	81.97
Réseau de neuronnes (NN)	*Non*	87.34	80.77	94.23	86.98	87.24	87.67
Réseau de neuronnes (NN)	*Oui*	85.61	75.0	96.81	84.52	85.21	86.14

Table C.5 – Mesures de performance pour le choix du meilleur modèle pour la modélisation du dépassement du seuil extrême u.

Les paramètres du modèle de RF sont : nombre d'arbres dans la forêt (*n_estimators*) : 20 ; profondeur maximale de l'arbre (*max_depth*) : 13 ; autres paramètres : défaut (scikit-learn).

Pour sélectionner un seuil optimal de classification binaire, nous avons cherché le seuil qui maximise l'exactitude. On trouve un seuil optimal de 0.47. La figure C.15 montre l'exactitude (taux de bon classement) en fonction du seuil.

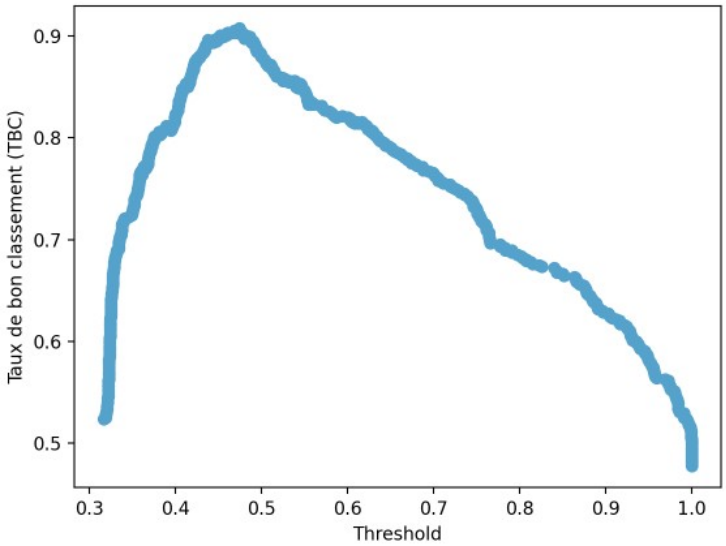


Figure C.15 – Sélection du seuil (*threshold*) pour le modèle *RF+GridSearch+CDF Normale*.

2.4 Importances prédictives des variables

La figure C.16 présente la contribution de chacune de nos variables au pouvoir prédictif global du modèle final. Cette analyse a pour but d'identifier les variables les plus pertinentes pour la prédiction des dépassements de seuil. Ce sont les variables binaires associées aux régions (Occitanie, Nouvelle-Aquitaine,Provence-Alpes-Cote d Azur) et la variable *vacance_scolaire_ou* qui explique le plus les prédictions du modèle.

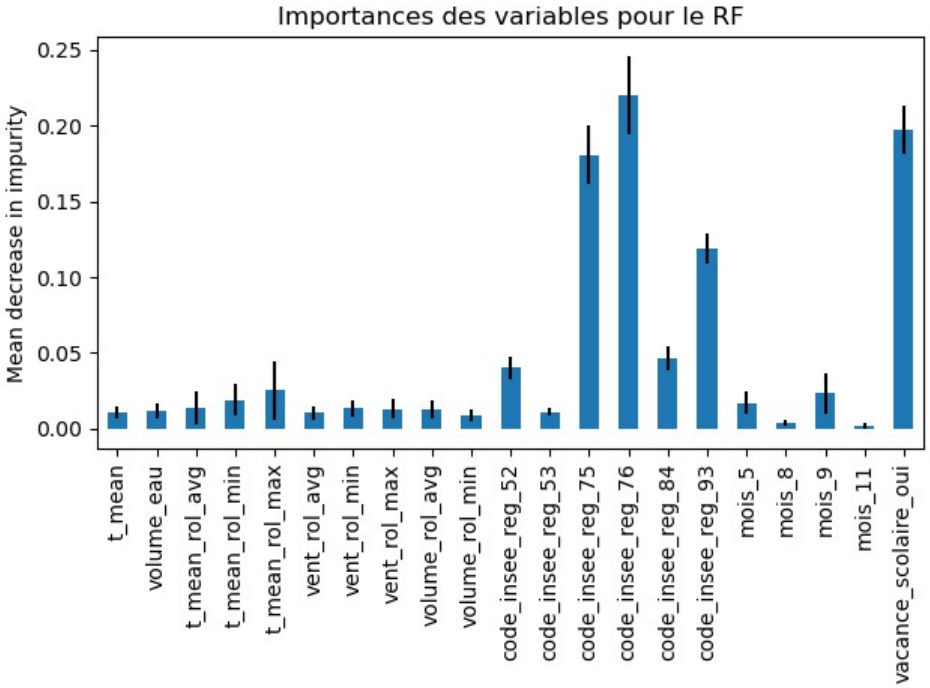


Figure C.16 – Importances des variables pour le RF + CDF Normale.

3 Modélisation de la sévérité attritionnelle $Y|Y < u$

Dans cette section, nous nous intéressons à la modélisation de la sinistralité attritionnelle qui correspond aux observations pour lesquels $Y|Y < u$ à l'aide du GLM avec $u = 130\,000$€. Les modèles GLM permettent d'exprimer l'espérance d'une variable à expliquer en fonction des variables explicatives.

3.1 Modèle linéaire généralisé (GLM)

3.1.1 Fondements théoriques

Nous effectuons une présentation synthétique du modèle linéaire généralisé. Nous nous plaçons dans un contexte de régression où nous cherchons à expliquer une variable Y par p variables explicatives $X_1, \ldots, X_p$. Nous disposons d'un n-échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ où les $x_i = (X_{i,1}, \ldots, X_{i,p})$ sont supposées fixés et les Y_i sont des variables aléatoires réelles indépendantes avec i nombre d'observations.

Famille exponentielle

Les variables aléatoires $Y_1, \ldots, Y_n$ ont une densité de probabilité exponentielle de la forme :

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \tag{C.19}$$

On dit alors que Y suit une loi $\mathscr{F}_{\text{exp}}(\theta_i, \phi_i, a, b, c)$.

θ est appelé le paramètre naturel et ϕ le paramètre de dispersion (ou paramètre de nuisance), $b(.)$ et $c(.)$ sont des fonctions réelles qui dépendent de la loi des Y_i. On montre que l'espérance et la variance des Y_i sont données par

$$\mathbb{E}(Y) = b'(\theta) \text{ et } \text{Var}(Y) = b''(\theta)\phi$$

Pour la modélisation des sinistres attritionnels du régime, nous allons estimer et comparer deux distributions standard, à savoir la loi gamma et la loi inverse gaussienne. La tableau C.6 ci-dessous présente les paramètres de la famille exponentielle et les fonctions de liens de ces deux lois.

Loi	f.m.p./densité	θ	ϕ	$a(x)$	$b(x)$	Espérance	Var. fonction	Support
Gamma(α, β)	$\frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x}$	$-\frac{\beta}{\alpha} = \frac{1}{\mu}$	$\frac{1}{\alpha}$	x	$-\ln(-x)$	$\mu = -\frac{1}{\theta}$	μ^2	$\mathbb{R}_+$
Loi inverse gaussienne(μ, λ)	$\sqrt{\frac{\lambda}{2\pi x^3}} e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}$	$-\frac{1}{2\mu^2}$	$\frac{1}{\lambda}$	x	$-\sqrt{-2x}$	$\mu = (-2\theta)^{-\frac{1}{2}}$	μ^3	$\mathbb{R}_+$

Table C.6 – Famille et fonctions liens pour la loi Gamma et inverse gaussienne

Définition du GLM

Un modèle linéaire généralisé (GLM) est caractérisé par trois hypothèses
1. une loi de probabilité : $(Y_i)_i$ sont indépendant et Y_i suit une loi $\mathscr{F}_{\text{exp}}(\theta_i, \phi_i, a, b, c)$,
2. une fonction déterministe : le vecteur de variables explicatives X_i donne le prédicteur linéaire $\eta_i = X_i^T \beta$
3. une fonction lien $g : \mathbb{R} \mapsto \overline{\mathbb{X}}$ monotone, différentiable et inversible telle que $E(Y_i) = g^{-1}(\eta_i)$, pour $i \in \{1, \ldots, n\}$, où θ_i est le paramètre d'échelle, ϕ le paramètre de dispersion et b, c trois fonctions.

Notons que les paramètres θ_i sont liés au prédicteur linéaire par la relation

$$\mu_i = E(Y_i) = b'(\theta_i) = g^{-1}(\eta_i)$$

Si $\theta_i = \eta_i$, alors la fonction lien est dit canonique. C'est à dire $\theta = g(b'(\theta)) \Leftrightarrow g(x) = (b')^{-1}(x)$. Le paramètre de dispersion ϕ_i est en général commun à toutes les observations à un transformation connue près. On suppose généralement que $\phi_i = \phi/\omega_i$ pour ω_i un poids connu.

3.1.2 Estimateur du modèle, choix du modèle et Mesure d'erreur

Estimateur du modèle

Notons $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_p, \phi\right)'$. En supposant que les variables $Y_1, \ldots, Y_n$ sont indépendantes. L'estimation du vecteur de paramètres β et du paramètre ϕ se fait la méthode de maximum de vraisemblance. La log-vraisemblance est

$$\ln \mathscr{L}(\beta, \phi) = \sum_{i=1}^{n} \frac{\theta_i(\beta) y_i - b\left(\theta_i(\beta)\right)}{a\left(\phi_i\right)} + \sum_{i=1}^{n} c\left(y_i, \phi_i\right),$$

avec $\theta_i = b^{-1}\left(g^{-1}\left(X_i'\beta\right)\right)$.

On veut maximiser cette vraisemblance $L(\beta)$. Maximiser $L(\beta)$ revient à maximiser la fonction de log-vraisemblance : $I(\beta) = \log(L(\beta))$.

Il n'y a pas de solution algébrique à la maximisation de cette vraisemblance. on utilise l'algorithme de Newton-Raphson on cherche à trouver numériquement la solution de

$$S(\beta) = \frac{\partial}{\partial \beta} \log(L(\beta)) = 0$$

Une propriété des distributions de la famille exponentielle (example loi de Gamma) est que leur fonction de vraisemblance ou de log-vraisemblance est strictement concave. En conséquence, il existe une seule solution à $S(\beta) = 0$ et le point lequel $S(\beta)$ est nécessairement un maximum.

Choix des variables

L'objectif est de sélectionner le meilleur modèle c'est à dire sélectionner les variables optimisant le modèle. Pour ce faire, nous comparons les modèles entre eux en fonction de deux critères. Le critère d'information d'Akaike et le critère d'information bayésien.

Critère AIC

Le critère d'information d'Akaike appelé AIC permet de comparer les différents modèles. Ce critère s'appuie sur le maximum de vraisemblance et le nombre de paramètres du modèle. Il s'écrit :

$$AIC = 2 \times (p - \mathscr{L}(\hat{\mu} \mid Y)),$$

où p et $\mathscr{L}$ sont respectivement le nombre de variables du modèle et la log-vraisemblance du modèle. Le meilleur modèle est celui présentant l'AIC le plus faible.

Critère BIC

Le critère d'information bayésien appelé BIC permet de comparer les différentes modèles. Il s'écrit :

$$BIC = -2 \times \mathscr{L}(\hat{\mu} \mid Y) + p \times \ln(n)$$

où

— p est le nombre de variables du modèle.
— $\mathscr{L}$ est la log-vraisemblance du modèle.
— n est le nombre d'observations.

Le meilleur modèle est celui présentant le BIC le plus faible. Ce critère pénalise de manière plus importante le nombre de variables du modèle.

Sélection des variables

Dans notre démarche, certaines variables peuvent ne pas être significatives. Une sélection individuelle des variables ne peut pas être envisagée du fait de l'interaction entre les variables. En effet, dans un modèle une variable peut ne pas être significative avec certaines variables et peut le devenir en ajoutant ou supprimant une variable du modèle. Ces contraintes techniques nous amènent à considérer trois méthodes afin de sélectionner une combinaison de variables intéressantes.

Forward Selection

Cette méthode consiste à utiliser un modèle avec une seule variable explicative. Cette variable peut être la variable la plus significative du modèle. Nous ajoutons la variable améliorant le plus le modèle. Ainsi de suite, jusqu'à ce que le modèle ne s'améliore plus. Dans notre cas, l'amélioration du modèle signifie une baisse de l'AIC ou BIC.

Backward Selection

Cette méthode consiste à utiliser un modèle avec toutes les variables explicatives. Nous enlevons la variable diminuant le plus l'AIC ou BIC. Ainsi de suite, jusqu'à ce que l'AIC augmente lors de la suppression d'une variable.

Stepwise Selection

Cette méthode consiste à utiliser un modèle avec les variables jugées les plus significatives du modèle. À chaque étape, nous enlèverons la variable la moins significative et nous rajouterons la variable la plus significative. Une variable peut être significative avec une combinaison de variables et ne plus l'être avec une autre combinaison de variables. Cette méthode est une combinaison des deux méthodes précédentes.

Qualité d'ajustement du modèle

L'objectif de cette partie est de présenter des outils de validation des différents modèles. Posons $\mu_i = E(Y_i \mid X_i), \sigma_i^2 = \text{var}(Y_i \mid X_i)$ et soient $\hat{\mu}_i, \hat{\sigma}_i^2$ les estimateurs du maximum de vraisemblance de μ_i et σ_i^2. Pour la validation des modèles, nous pouvons analyser les résidus. En général nous considérons trois types de résidus :

— Les résidus bruts :

$$\varepsilon_i = Y_i - \hat{\mu}_i$$

— Les résidus de déviance :

$$\varepsilon_{i,D} = \text{signe}(d_i) * \sqrt{|d_i|},$$

avec la déviance résiduelle

$$D = \sum_{i=1}^{n} d_i,$$

l'expression de d_i dépend de la loi choisie.

— Les résidus normalisés de Pearson :

$$\varepsilon_{i,P} = \frac{Y_i - \hat{\mu}_i}{\hat{\sigma}_i}.$$

Les tests d'adéquation du modèle sont basés sur ces deux derniers résidus.

Déviance résiduelle

Il s'agit d'évaluer la qualité du modèle en considérant toutes les variables explicatives et en se basant sur les écarts entre observations et estimations. Le modèle estimé est comparé au modèle saturé (c'est à dire au modèle possédant autant de paramètres que d'observations. Cette comparaison est basée sur la différence des log-vraisemblances, qu'on appelle déviance résiduelle et est donnée par

$$D = 2(\mathscr{L}(Y,Y) - \mathscr{L}(Y,\hat{\beta})),$$

où $\mathscr{L}(Y,Y)$ et $\mathscr{L}(Y,\hat{\beta})$ sont respectivement les log-vraisemblances du modèle saturé et estimé. Nous montrons qu'asymptotiquement D suit une loi de Khi-deux à $n-(p+1)$ degrés de liberté ($(p+1)$ est le nombre de paramètres inconnus du modèle). On peut alors effectuer un test de rejet ou d'acceptation du modèle selon la valeur de la déviance. Nous acceptons le modèle avec un risque α si $D \leq x_\alpha$, avec $P\left(\mathscr{X}^2(n-(p+1)) > x_\alpha\right) = \alpha$.

Test de Pearson

Le test de Pearson est un test de type Khi-deux dont le but est de comparer les valeurs observées Y_i aux valeurs estimées par le modèle. Posons $\mu_i = E(Y_i \mid X_i), \sigma_i^2 = \text{var}(Y_i \mid X_i)$ et soient $\hat{\mu}_i, \hat{\sigma}_i^2$ les estimateurs du maximum de vraisemblance de μ_i et σ_i^2. La statistique de test est donnée par :

$$\mathscr{X}^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2},$$

On montre qu'asymptotiquement $\mathscr{X}^2$ suit une loi de Khi-deux à $n-(p+1)$ degrés de liberté. En résumé, en appliquant le test de déviance et/ou le test de Pearson, le modèle est généralement accepté si $D \leq n-(p+1)$ et/ou $\mathscr{X}^2 \leq n-(p+1)$. Les résidus de deviance et de Pearson peuvent aussi être validés par des analyses graphiques.

3.2 Exploration des données

base de donnée pour la modélisation de la modélisation de la sinistralité attritonnelle, contient 1755 observations et 19 variables. Au niveau des variables on a la variable *dépassement* qui est la variable de dépassement du seuil extrême u ou non. Initialement la base contient 5 variables qualitatives dont la région, la saison, le mois, l'indicateur de vacance scolaire et l'indicateur de week-end. Concernant les variables quantitatives, la base contient la **température** journalière en degrés celsius (°C), **la précipitation** en millimètre (mm) et la **vitesse de vent** en mètre par seconde (m/s). Ces trois variables sont associées aux risques couverts par le régime. Aussi, pour chacune de ces variables nous avons retenu la moyenne, le minimum et le maximum en moyenne mobile (glissante) sur 7 jours.

La figure C.17 représente la sinistralité moyenne par région. On remarquera qu'il n'y a pas de sinistralité attritionnelle dans la région de Bretagne.

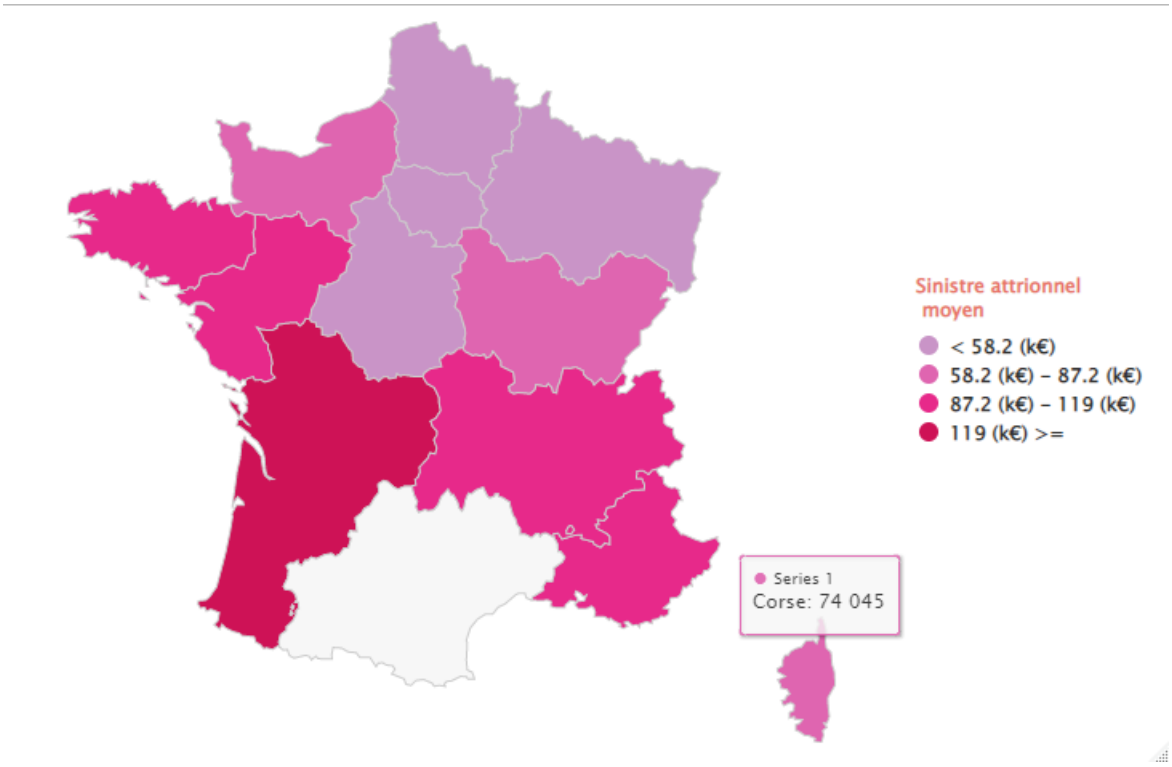


Figure C.17 – Moyenne des sinistres attritionnels par région

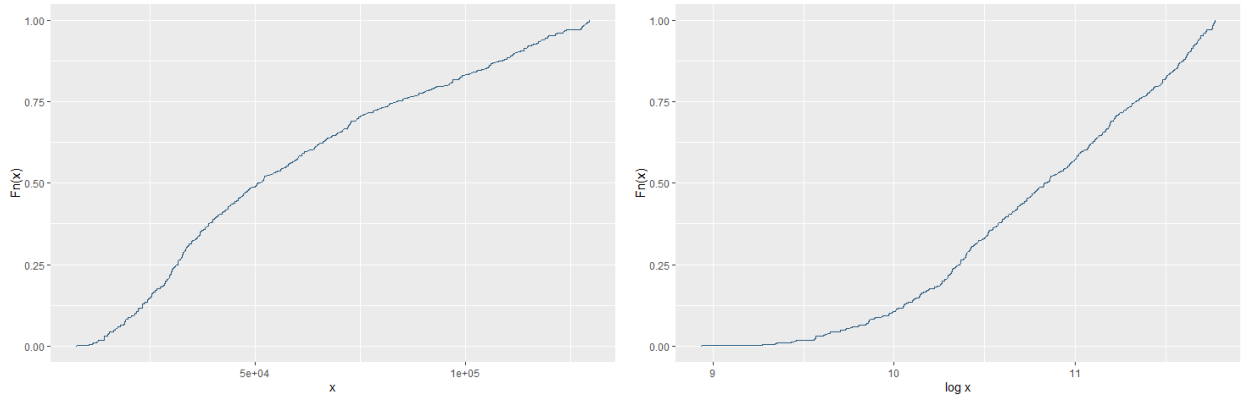


Figure C.18 – Fonction de répartition empirique des sinistres et log des sinistres attritionnels

Une première analyse a consisté à discrétiser nos variables afin d'avoir des coefficients interprétables plus facilement. Pour cela nous utilisons les modèles additifs généralisés (GAM) qui ont l'avantage de représenter la relation linéaire avec plus de flexibilité entre deux variables. Ensuite nous utilisons une approximation par des fonctions en morceaux pour créer des classes de modalités.

La figure ci-dessous présente le graphique de l'impact des variables explicatives sur les prédictions du GAM.

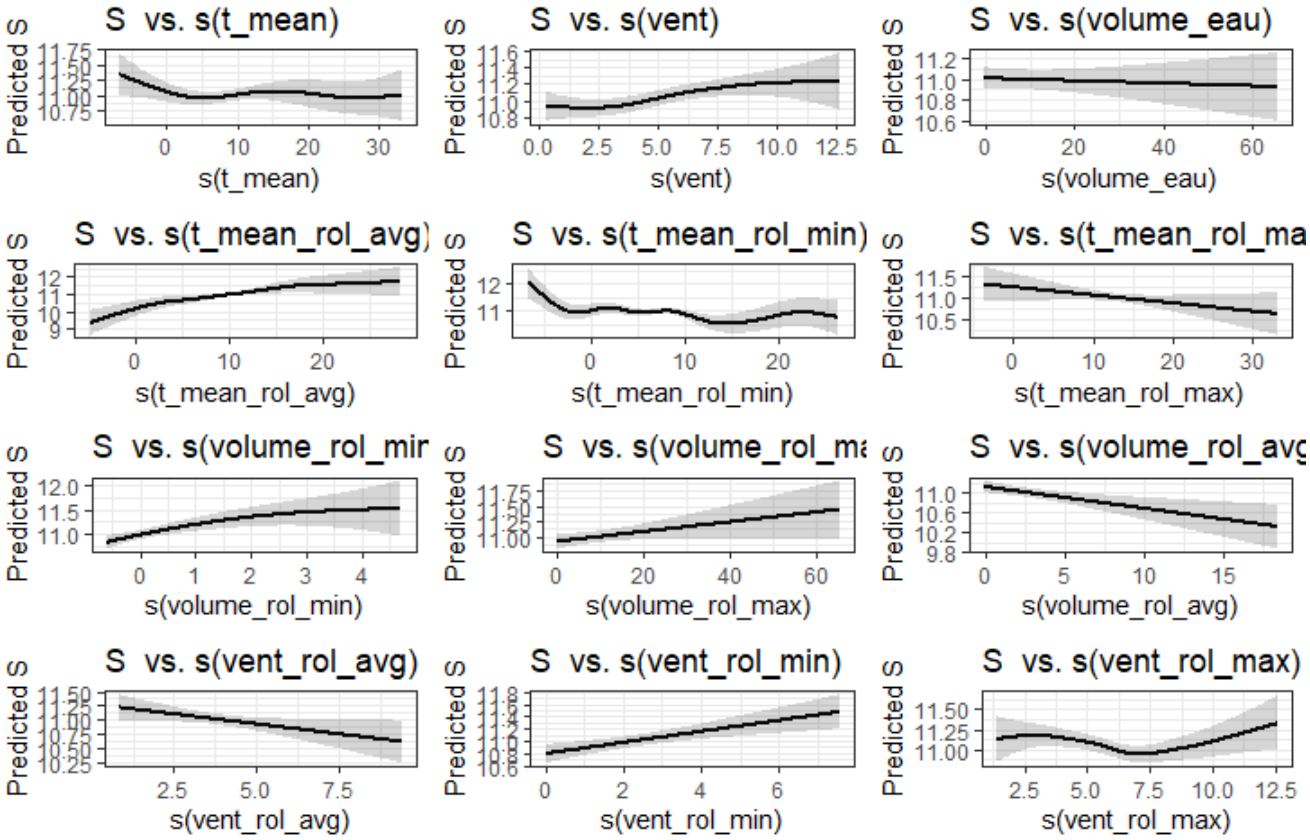


Figure C.19 – Graphique du GAM pour la sinistralité attritionnelle

La figure ci-dessous montre la boite à moustache des sinistres en fonction de la température catégorisée (gauche) et de la variable saison (droite).

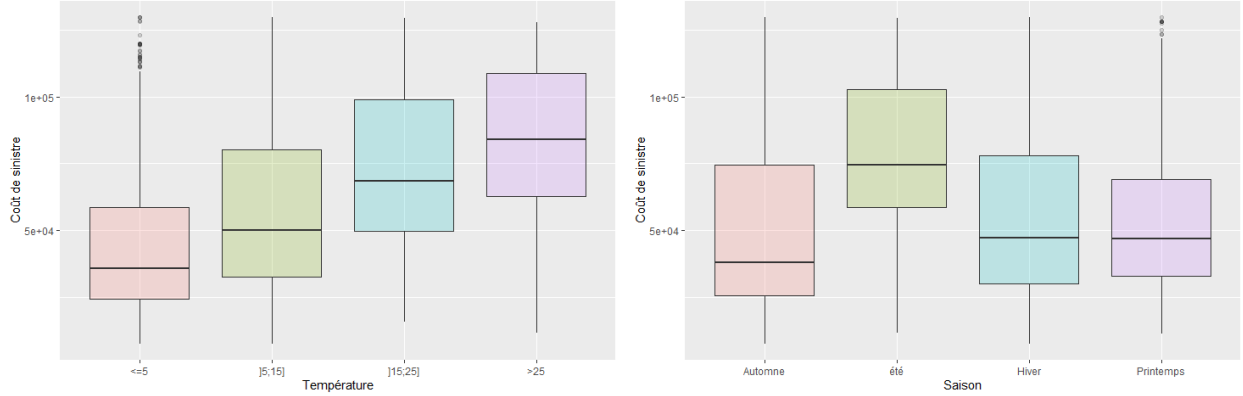


Figure C.20 – A gauche : boite à moustache des sinistres en fonction de la température catégorisée, A droite : boite à moustache des sinistres en fonction des saisons.

La figure ci-dessous montre le V de cramer entre les variables catégorielles explicatives retenues. Le V de Cramer, également appelé le coefficient de contingence, est une mesure statistique utilisée pour évaluer la force de l'association entre deux variables catégorielles dans une table de contingence (un tableau croisé).

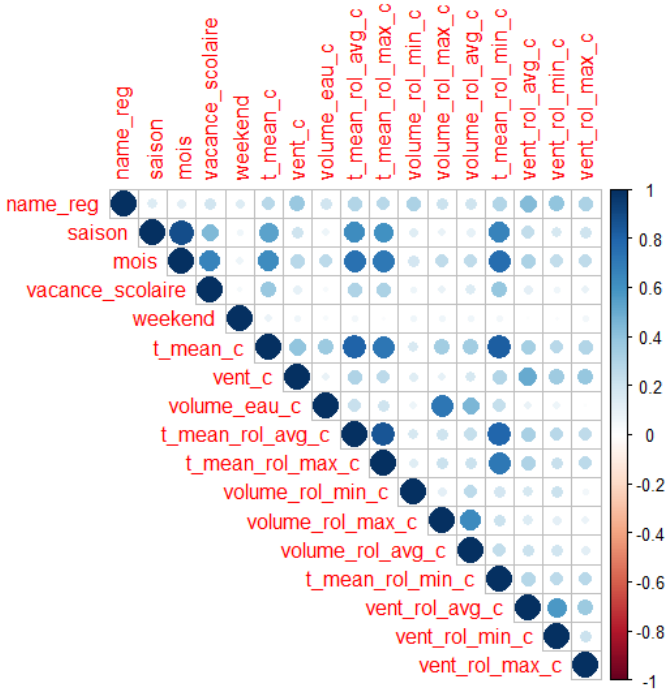


Figure C.21 – V-cramer entre les variables retenues pour la modélisation de la sinistralité attritionnelle

3.3 Modèle final pour la sévérité attritionnelle

Nous avons utilisé un modèle de régression linéaire généralisé (GLM) pour modéliser la sinistralité attritionnelle $Y|Y < u$, en testant deux distributions potentielles : la loi gamma et la loi inverse gaussienne, toutes deux utilisant une fonction de lien logarithmique. Nous avons utilisé la procédure forward pour la selection automatique des variables. Le tableau ci-dessous présente les critères AIC, BIC, et la déviance associés à chaque modèle :

Modèle	AIC	BIC	Déviance
Gamma	23774	24039	69.33
Inverse Gaussienne	24150	24350	0.0023

Les résultats du tableau ci-dessus laissent entrevoir que le modèle gamma est plus approprié pour la projection de la sinistralité attritionnelle. Le tableau suivant montre les mesures de performances du RMSE, du MAE et du MSE pour le modèle gamma :

Test	*Train*	*Métrique*
16862	13890	**RMSE**
12238	9980	**MAE**
284329680	192953808	**MSE**

Le tableau suivant présente les coefficients issus du résultat de la modélisation de la sinistralité attritionnelle par le modèle GLM gamma.

	Model 1
(Intercept)	11.74***
regionBourgogne-Franche-Comte	-1.35***
regionBretagne	-0.21**
regionCentre-Val de Loire	-1.71***
regionCorse	-1.00***
regionGrand Est	-1.37***
regionHauts-de-France	-1.60***
regionIle-de-France	-2.00***
regionNormandie	-1.20***
regionNouvelle-Aquitaine	0.30
regionPays de la Loire	-0.03
regionProvence-Alpes-Cote d Azur	-0.07
saisonété	-0.16*
saisonHiver	0.04
saisonPrintemps	0.01
mois2	-0.32***
mois3	0.28***
mois4	-0.09
mois5	0.73***
mois6	0.84***
mois7	-0.06
mois8	-0.18
mois9	0.85***
mois10	0.06
mois11	0.07
mois12	-0.38***
vacance_scolaireoui	1.15***
weekendoui	0.24***
t_mean_c]5;15]	-0.10**
t_mean_c]15;25]	-0.17**
t_mean_c>25	0.04
vent_c]3.62;6.46]	0.06**
vent_c>6.46	0.11***
volume_eau_c]6.87;21.3]	-0.02
volume_eau_c>21.3	0.05
t_mean_rol_avg_c]5.82;14.8]	0.06
t_mean_rol_avg_c>14.8	0.08
t_mean_rol_max_c]8.26;17.6]	-0.07*
t_mean_rol_max_c>17.6	-0.17*
volume_rol_min_c]-0.06;0.94]	-0.00
volume_rol_min_c>0.94	0.02
volume_rol_max_c]7.69;21.1]	0.07*
volume_rol_max_c>21.1	0.03
volume_rol_avg_c]2.09;5.63]	-0.02
volume_rol_avg_c>5.63	0.01
t_mean_rol_min_c]3;15]	-0.10*
t_mean_rol_min_c>15	0.18**
vent_rol_avg_c]3.64;5.52]	0.06*
vent_rol_avg_c>5.52	0.02
vent_rol_min_c]2;3.47]	0.01
vent_rol_min_c>3.47	-0.00
vent_rol_max_c>3.5	0.04
AIC	23773.66
BIC	24039.01
Log Likelihood	-11833.83
Deviance	69.33
Num. obs.	1104

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table C.7 – Gamma avec lien log

Annexe D

Tableaux, figures complémentaires

Descriptif	Mnémonique	Type *	Unité
indicatif OMM station	numernsta	car	
date (UTC)	date	car	AAAAMMDDHHMISS
pression au niveau mer	pmer	int	Pa
variation de pression en 3 heures	tend	int	Pa
type de tendance barométrique	cod tend	int	code (0200)
direction du vent moyen 10mn	dd	int	degré
vitesse du vent moyen 10mn	ff	réel	m/s
température	t	réel	K
point de rosée	td	réel	K
humidité	u	int	%
visibilité horizontale	vv	réel	mètre
temps présent	ww	int	code (4677)
temps passé 1	w1	int	code (4561)
temps passé 2	w2	int	code (4561)
nébulosité totale	n	réel	%
nébulosité des nuages de l'étage inférieur	nbas	int	octa
hauteur de la base des nuages de l'étage inférieur	hbas	int	mètre
type des nuages de l'étage inférieur	cl	int	code (0513)
type des nuages de l'étage moyen	cm	int	code (0515)
type des nuages de l'étage supérieur	ch	int	code (0509)
pression station	pres	int	Pa
niveau barométrique	niv bar	int	Pa
géopotentiel	geop	int	m^2/s^2
variation de pression en 24 heures	tend24	int	Pa
température minimale sur N heures	tnN	réel	K
température maximale sur N heures	txN	réel	K
température minimale du sol sur 12 heures	tminsol	réel	K
méthode mesure tw	sw	int	code (3855)
température du thermomètre mouillé	tw	réel	K
rafales sur les 10 dernières minutes	raf10	réel	m/s
rafales sur une période	rafper	réel	m/s
période de mesure de la rafale	per	réel	minute
état du sol	etat_sol	int	code (0901)
hauteur totale de la couche de neige, glace, au sol	ht_neige	réel	mètre
hauteur de la neige fraîche	ssfrai	réel	mètre
Période de mesure de la neige fraîche	perssfrai	réel	1/10 heure
Précipitations dans les N dernières heures	rrN	réel	mm
Phénomène spécial	phenspeN	réel	code (3778)
Nébulosité couche nuageuse N	nnuageN	int	octa
Type de nuage N	ctypeN	int	code(0500)
Hauteur de base de nuage N	hnuageN	int	mètre

Table D.1 – liste des variables de la base SYNOP

* **car** : caractère ASCII, **int** : nombre entier, **réel** : nombre réel (avec décimale).
Les nombres entre parenthèses après le mot « code » sont les numéros de table de code de l'OMM

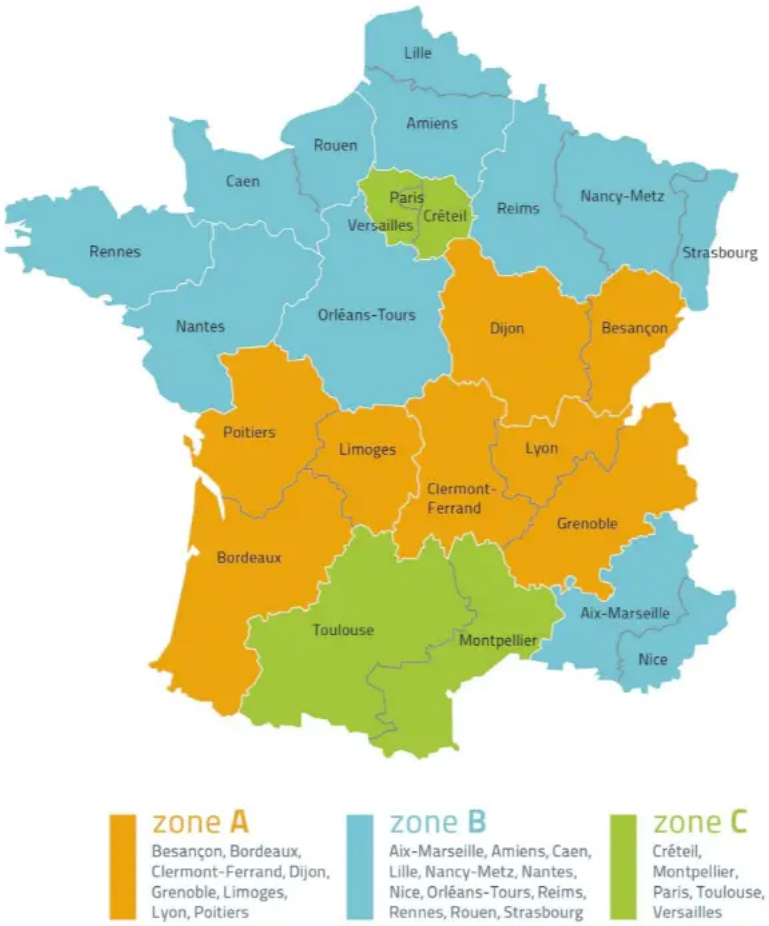


Figure D.1 – Les 3 zones de vacance scolaire en 2022

Région	Nombre	Moyenne	$q_{0.25}$	Médiane	$q_{0.75}$	$q_{0.95}$	max
Ile-de-France	269	40 966	19 844	30 980	55 956	96 403	154 401
Centre-Val de Loire	230	52 405	28 550	41 851	72 332	117 318	192 516
Bourgogne-Franche-Comte	322	75 058	33 068	61 127	105 055	160 131	240 088
Normandie	346	106 644	47 145	78 070	145 108	284 072	335 372
Hauts-de-France	303	61 427	31 753	49 016	83 747	137 770	154 773
Grand Est	269	66 679	32 996	51 427	90 462	144 674	258 249
Pays de la Loire	222	292 147	151 585	234 142	384 835	673 115	929 652
Bretagne	248	355 663	156 045	267 242	506 724	781 897	972 866
Nouvelle-Aquitaine	191	673 581	317 860	538 019	888 640	1 495 853	1 933 490
Occitanie	161	813 248	400 000	726 572	1 190 149	1 669 073	2 448 415
Auvergne-Rhone-Alpes	251	325 721	152 799	280 588	475 559	672 929	945 004
Provence-Alpes-Cote d Azur	350	431 297	235 579	377 118	593 324	878 714	1 865 121
Corse	215	142 087	70 782	131 752	179 305	297 161	362 038
France	**3 377**	**235 889**	**486 46**	**123 286**	**303 169**	**838 401**	**2 448 415**

Table D.2 – Statistiques descriptives de la sévérité selon la région (2016 - 2019).

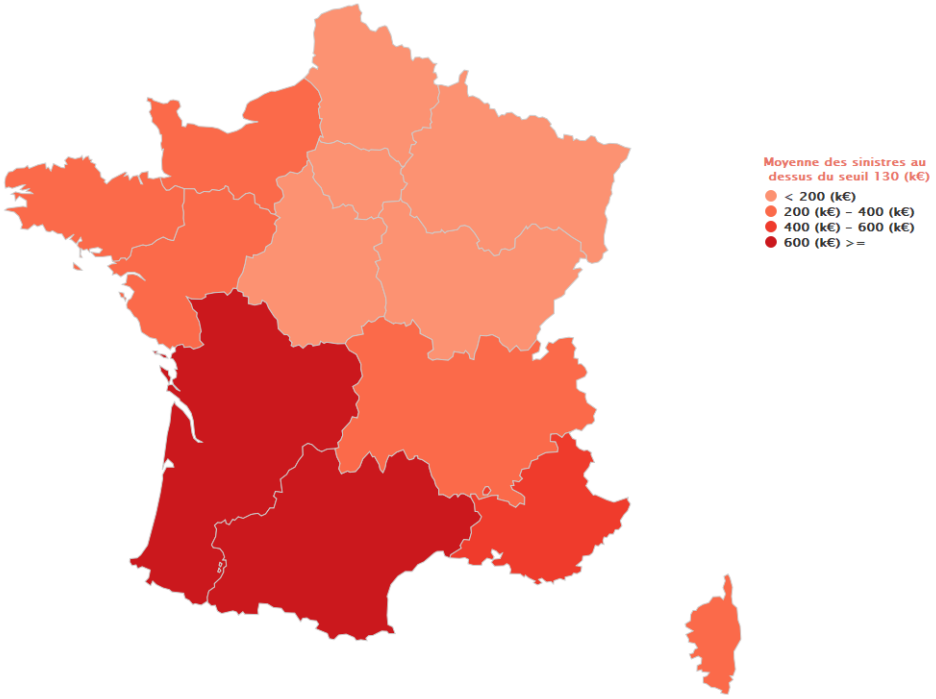


Figure D.2 – Moyenne des sinistres au dessus du seuil $u = 130000$ pour chaque région (2016 - 2021)

variable		min	$q_{.25}$	Médiane	Moyenne	$q_{.75}$	max
Sinistre extrême au dessus du seuil u = 130000		130065	186628	313271	427138	550312	2448415
Température	*Température journalière (t_mean)*	-4.6	9.8	14.2	14.9	20	31.4
	Moyenne en glissement sur 7 jours de t_mean (t_mean_rol_avg)	-3.1	9.4	14.7	15.1	21.5	28.3
	Minimum en moyenne mobile sur 7 jours de t_mean (t_mean_rol_min)	-6.8	7.1	12.5	12.6	18.3	27.3
	Maximum en moyenne mobile sur 7 jours de t_mean (t_mean_rol_max)	-1.7	11.8	17	17.5	24.2	31.4
Vitesse du vent	*Vitesse de vent journalière (vent)*	0	3	5.2	5.3	7.1	13.2
	Moyenne en glissement sur 7 jours de vent (vent_rol_avg)	1.3	3.1	4	4.2	5	9.6
	Minimum en moyenne mobile sur 7 jours de vent (vent_rol_min)	0	2.1	2.5	2.7	3.1	7.6
	Maximum en moyenne mobile sur 7 jours de t_mean (vent_rol_max)	1.6	4.3	6.2	6.2	7.8	13.2
Précipitation	*Précipitation moyenne journalière (volume_eau)*	-0.3	0	3.1	7	12.2	69.7
	Moyenne en glissement sur 7 jours de volume_eau (volume_eau_rol_avg)	-0.1	0.4	2.2	2.8	4.2	18.9
	Minimum en moyenne mobile sur 7 jours de volume_eau (t_mean_rol_min)	-0.5	0	0	0.1	0	6
	Maximum en moyenne mobile sur 7 jours de volume_eau (volume_eau_rol_max)	0	2.1	9.2	9.9	14.9	69.7

Table D.3 – Statistiques descriptives des variables quantitatives pour les observations dont le sinistre est supérieur à 130 000€

Variable	Catégorie	Nombre d'observation	Pourcentage
Région	*Ile-de-France*	2	0.1%
	Centre-Val de Loire	6	0.4%
	Bourgogne-Franche-Comte	46	2.8%
	Normandie	101	6.2%
	Hauts-de-France	24	1.5%
	Grand Est	26	1.6%
	Pays de la Loire	186	11.5%
	Bretagne	210	12.9%
	Nouvelle-Aquitaine	188	11.6%
	Occitanie	161	9.9%
	Auvergne-Rhone-Alpes	211	13.0%
	Provence-Alpes-Cote d Azur	346	21.3%
	Corse	115	7.1%
Mois	*Janvier*	97	6.0%
	Février	99	6.1%
	Mars	131	8.1%
	Avril	84	5.2%
	Mai	123	7.6%
	Juin	110	6.8%
	Juillet	196	12.1%
	Août	282	17.4%
	Septembre	116	7.2%
	Octobre	119	7.3%
	Novembre	124	7.6%
	Décembre	141	8.7%
Saison	*Automne*	373	23.0%
	Eté	626	38.6%
	Hiver	308	19.0%
	Printemps	315	19.4%
Vacances	*Oui*	778	48.0%
	Non	844	52.0%
Week-end	*Oui*	542	33.4%
	Non	1080	66.6%

Table D.4 – Statistiques descriptives des variables qualitatives pour les observations dont le sinistre est supérieur à 130 000€ .

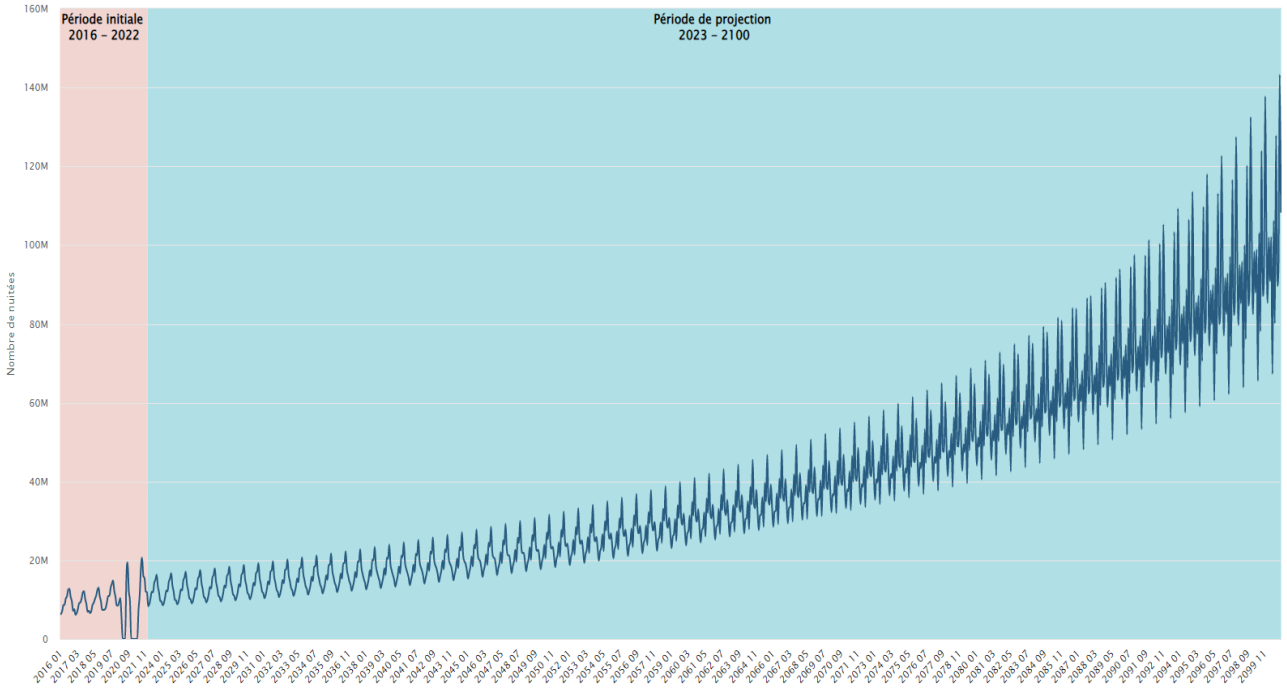


Figure D.3 – Évolution du nombre de nuitées totale mensuel en France métropolitaine.

Saison	Quantile	Th (°C)		Tb (°C)		Vh (m/s)		Ph (mm)	
	Période	RCP 2.6	RCP 8.5	RCP 2.6	RCP 8.5	RCP 2.6	RCP 8.5	RCP 2.6	RCP 8.5
Eté	2001 - 2010								
	2011 - 2020	6.445		4.94		2.02		6.55	
	2021 - 2030	7.27	6.47	5.78	5.95	1.73	1.87	7.99	9.05
	2031 - 2040	6.68	5.63	5.58	5.29	2.25	2.28	8.04	8.17
	2041 - 2050	6.23	6.17	4.93	4.66	2.45	2.47	7.67	8.91
	2051 - 2060	5.51	7.00	5.61	4.30	2.18	2.44	8.68	6.86
	2061 - 2070	6.43	7.27	5.21	4.51	2.23	2.06	8.21	7.65
	2071 - 2080	6.04	7.10	5.10	5.04	2.28	2.17	8.40	7.89
	2081 - 2090	5.52	8.06	6.22	4.44	2.35	2.33	7.92	7.99
	2091 - 2100	8.05	8.62	5.03	4.60	2.32	2.26	7.88	7.23
Hiver	2001 - 2010								
	2011 - 2020	6.02		6.02		3.52		7.64	
	2021 - 2030	5.22	4.41	8.15	8.44	3.29	3.18	8.43	8.77
	2031 - 2040	6.21	6.42	6.54	7.07	3.53	3.76	10.16	8.99
	2041 - 2050	6.04	6.71	6.61	6.11	3.76	4.28	9.44	9.46
	2051 - 2060	6.04	6.70	6.60	6.36	3.63	3.30	8.75	8.84
	2061 - 2070	6.48	7.41	6.68	6.22	3.90	3.78	9.36	9.83
	2071 - 2080	5.82	6.35	7.25	5.10	3.55	3.49	8.54	9.08
	2081 - 2090	6.64	6.41	6.54	6.04	3.59	3.97	8.56	9.76
	2091 - 2100	6.65	6.43	6.09	5.50	4.03	4.01	9.44	10.64
Autome	2001 - 2010								
	2011 - 2020	8.26		7.02		3.08		9.45	
	2021 - 2030	5.82	6.27	8.74	10.10	2.70	2.50	10.34	10.42
	2031 - 2040	7.63	8.37	8.10	8.50	3.28	3.15	10.76	10.83
	2041 - 2050	7.73	7.82	8.10	8.19	3.08	2.93	9.80	10.32
	2051 - 2060	8.22	8.37	8.14	7.14	3.18	3.15	10.02	9.71
	2061 - 2070	5.57	7.59	9.08	8.80	3.33	3.17	9.87	12.03
	2071 - 2080	7.59	8.45	8.30	8.14	3.20	3.23	10.99	10.61
	2081 - 2090	8.02	7.54	8.71	8.35	3.11	2.96	10.72	12.47
	2091 - 2100	7.71	8.46	9.40	7.82	3.16	3.43	10.42	11.48
Printemps	2001 - 2010								
	2011 - 2020	7.07		6.68		2.46		7.77	
	2021 - 2030	6.86	6.38	8.40	8.30	2.36	2.12	8.31	8.94
	2031 - 2040	7.95	7.87	8.19	6.73	2.31	2.66	8.27	7.85
	2041 - 2050	7.82	7.23	6.56	7.54	3.05	2.82	8.31	8.49
	2051 - 2060	8.38	8.56	7.02	7.10	2.64	2.60	8.22	8.03
	2061 - 2070	7.24	8.75	7.83	6.99	2.77	2.62	8.77	8.01
	2071 - 2080	8.08	8.61	7.65	7.15	2.81	2.74	8.19	8.57
	2081 - 2090	8.45	7.58	7.40	8.16	2.47	2.70	8.31	8.35
	2091 - 2100	8.28	8.74	6.78	6.44	2.65	2.53	6.70	8.63

Table D.5 – Quantile décennale saisonnière des indicateurs des indicateurs selon les scénarios du GIEC (2.6 et 8.5) de 2011 à 2100.

Algorithme 1: SMOTE(T, N, k)

Input: Nombre d'échantillons de la classe minoritaire T ; Taux de SMOTE N% ; Nombre de plus proches voisins k

Output: $(T/100) \times T$ échantillons synthétiques de la classe minoritaire

(* Si N est inférieur à 100, randomisez les échantillons de la classe minoritaire car seulement un pourcentage aléatoire d'entre eux sera SMOTEd. *)

if $N < 100$ **then** /* Randomiser les T échantillons de la classe minoritaire */

  $T = (N/100) \times T$

  $N = 100$

end if

N = (int)(N/100)(* On suppose que la quantité de SMOTE est un multiple entier de 100 *)

k = Nombre de plus proches voisins

$numattrs$ = Nombre d'attributs

$Sample$[][] : tableau pour les échantillons originaux de la classe minoritaire

$newindex$: garde un compte du nombre d'échantillons synthétiques générés, initialisé à 0

$Synthetic$[][] : tableau pour les échantillons synthétiques

/* Calcul des k plus proches voisins pour chaque échantillon de la classe minoritaire uniquement */

for $i \leftarrow 1$ **to** T **do**

  Calculer les k plus proches voisins pour i, et enregistrer les indices dans le $nnarray$

  Populate($N, i, nnarray$)

end for

/* Populate($N, i, nnarray$) : Fonction pour générer les échantillons synthétiques */

while $N \neq 0$ **do**

  Choisir un nombre aléatoire entre 1 et k, l'appeler nn. /* Cette étape choisit l'un des k plus proches voisins de i. */

  for $attr \leftarrow 1$ **to** $numattrs$ **do**

    Calculer : $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$

    Calculer : gap = nombre aléatoire entre 0 et 1

    $Synthetic[newindex][attr] = Sample[i][attr] + gap \times dif$

  end for

  $newindex++$

  $N = N - 1$

end while

return /* Fin de Populate */

Algorithme 2: ADASYN

Input: Ensemble de données d'entraînement D_{tr} avec m échantillons $\{x_i, y_i\},, i = 1, \ldots, m$ où x_i est une instance dans l'espace des caractéristiques de dimension n X et $y_i \in Y = \{1, -1\}$ est l'étiquette de classe associée à x_i. On définit m_s et m_l comme le nombre d'exemples de la classe minoritaire et le nombre d'exemples de la classe majoritaire, respectivement. Par conséquent, $m_s \leq m_l$ et $m_s + m_l = m$

Output: (X_{res}, Y_{res})

Calculer : $d = m_s/m_l$ /* le degré du déséquilibre de classe */

if $d < d_{th}$ **then** /* d_{th} est un seuil prédéfini pour le taux maximum toléré de déséquilibre de classe */

  Calculer : $G = (m_l - m_s) \times \beta$ /* G : le nombre d'exemples de données synthétiques à générer pour la classe minoritaire, $\beta \in [0,1]$ est un paramètre utilisé pour spécifier le niveau d'équilibre souhaité après la génération des données synthétiques. */

  foreach $x_i \in classeminoritaire$ **do**

    Trouver les K plus proches voisins en fonction de la distance euclidienne dans l'espace de dimension n

    Calculer : $r_i = \Delta_i / K$

    /* Δ_i est le nombre d'exemples parmi les K plus proches voisins de x_i qui appartiennent à la classe majoritaire, donc $r_i \in [0,1]$ */

  end foreach

  foreach $x_i \in classeminoritaire$ **do**

    Calculer : $\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i$ /* Normalisation de r_i */

    Calculer : $g_i = \hat{r}_i \times G$ /* le nombre d'exemples de données synthétiques à générer pour chaque x_i de la classe minoritaire */

    for $z \leftarrow 0$ **to** g_i **do**

      Choisir aléatoirement un exemple de données de la classe minoritaire, x_{zi}, parmi les K plus proches voisins de la donnée x_i.

      Générer l'exemple de données synthétiques :

    end for

  end foreach

end if

Liste des graphiques

Liste des tableaux